Mind the Ladder: A Benchmark for Level 1–3 Causal Reasoning in World Models

Anonymous authors
Paper under double-blind review

Abstract

World models based on Joint-Embedding Predictive Architecture (JEPA) have recently demonstrated emergent physical understanding through Violation-of-Expectation (VoE) paradigms. However, the "surprise" metric used to evaluate these models, a latent prediction error, conflates statistical novelty with genuine causal reasoning. I argue that no existing benchmark allows researchers to *observe and measure* whether a world model has climbed Pearl's Ladder of Causality from Level 1 (Association) through Level 2 (Intervention) to Level 3 (Counterfactuals). In this paper, I introduce **Mind the Ladder**, a diagnostic benchmark and accompanying metric suite designed to test causal fidelity in latent world models. Distinct from prior environment-level causal benchmarks (Ke et al., 2021; Ahmed et al., 2021), my framework operates directly in the *latent space* of a trained world model and is architecture-agnostic. I ground my approach in the LeWorldModel (LeWM) architecture and the `stable-worldmodel` codebase, proposing three novel metrics (AAP Surprise Ratio, Structural Invariance, and AAP Consistency Advantage) that operationalize Pearl's Abduction-Action-Prediction cycle directly in latent space. I validate my framework on a "Glitched Hue Two Room" experiment that tests for causal disentanglement between spurious correlations and true causal mechanisms. My results demonstrate that surprise alone is not enough: a model can exhibit high VoE surprise for physical violations while still failing to pass Level 3 counterfactual tests.

1 Introduction

Recent advances in world modeling have produced architectures that learn remarkably accurate internal simulations of physical environments from raw pixels (Ha & Schmidhuber, 2018; LeCun, 2022). The LeWorldModel (LeWM) (Maes et al., 2026) exemplifies this trend: a JEPA-based architecture that predicts future latent states conditioned on actions, and uses a Violation-of-Expectation (VoE) paradigm to demonstrate emergent physical understanding. When an agent teleports through a wall, the model's prediction error, its "surprise," spikes dramatically, while mere visual changes (such as the agent changing color) produce comparatively little surprise.

This is a striking result (see Table 1 for the full hierarchy). It suggests that LeWM has learned something deeper than pixel statistics. But *how deep*?

I argue that the field currently lacks two critical components:

1. **A benchmark to observe and measure causality.** Existing evaluation protocols for world models focus on reconstruction quality (FID (Heusel et al., 2017), LPIPS (Zhang et al., 2018)), planning performance (task reward) (Hafner et al., 2020), or novelty detection (surprise) (Pathak et al., 2017). While prior work has used known-SCM environments to probe causal discovery in model-based RL agents (Ke et al., 2021; Ahmed et al., 2021), none of these operate in the latent space of a trained world model or directly measure whether a model has internalized Pearl's three rungs of the causal hierarchy (Pearl, 2009; Pearl & Mackenzie, 2018).

Table 1: **Pearl's Ladder of Causality.** The three levels of causal reasoning: Level 1 (Association), Level 2 (Intervention), and Level 3 (Counterfactuals). Questions at level i can only be answered if information from level i or higher is available.

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y \mid x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y \mid \mathrm{do}(x), z)$	Doing, Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x \mid x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking for the past 2 years?

2. **Metrics designed for the causal hierarchy.** The "surprise" metric used by LeWM is a single scalar that collapses the distinction between Level 1 (recognizing that something is statistically unusual), Level 2 (understanding *why* an intervention produces a specific outcome), and Level 3 (reasoning about what *would have happened* under different circumstances). No existing metric explicitly tests progress across these levels.

In this paper, I make the following contributions:

- I introduce **Mind the Ladder**, a diagnostic benchmark grounded in Pearl's three-level causal hierarchy, designed to be applied to any latent world model.
- I propose three novel metrics (AAP Surprise Ratio, Structural Invariance, and AAP Consistency Advantage) that operationalize the Abduction-Action-Prediction (AAP) cycle in latent space.
- I implement my benchmark within the `stable-worldmodel` codebase (Maes et al., 2026) and validate it on a *Glitched Hue Two Room* experiment that isolates causal disentanglement from visual correlation.

2 Background and Related Work

2.1 Pearl's Ladder of Causality

Pearl's Causal Hierarchy Theorem (CHT) (Pearl, 2009; Bareinboim et al., 2022) establishes three strictly ordered levels of causal reasoning:

1. **Level 1: Association** ($P(y \mid x)$): Observing correlations. *"What if I see x?"*
2. **Level 2: Intervention** ($P(y \mid \mathrm{do}(x))$): Acting on the world by forcing the value of a variable. *"What if I do x?"*
3. **Level 3: Counterfactuals** ($P(y_{x^*} \mid x, y)$): Imagining alternative outcomes given observed evidence. *"Given that I saw x and y, what would y have been if x had been x^*?"*

Crucially, the CHT proves that in general, knowledge at Level k is *insufficient* to answer queries at Level $k+1$. A model that achieves perfect associational accuracy can still fail catastrophically on interventional queries, and a model that handles interventions correctly may be unable to reason counterfactually.

2.2 Related Work

Causal benchmarks for model-based agents. Ke et al. (2021) introduce a systematic evaluation suite for causal discovery in visual model-based RL, using known-SCM grid environments to test whether learned

transition models recover correct causal graphs. Ahmed et al. (2021) present CausalWorld, a robotic manipulation benchmark in which the underlying SCM can be programmatically modified to generate transfer learning tasks. Both works probe causal structure at the *environment* level, evaluating whether an agent's policy generalises under interventions, whereas my benchmark operates in the *latent space* of a trained world model and directly measures Pearl's three causal rungs via latent-space probing.

Causal representation learning (CRL). The theoretical conditions under which a latent causal structure is identifiable from observational data are analysed by Khemakhem et al. (2020) (iVAE) and Ahuja et al. (2023). Schölkopf et al. (2021) provide a unifying perspective on learning disentangled causal representations. My Structural Invariance metric (section 4.2) relies on the Independent Causal Mechanisms principle from this literature. A fully identified SCM is beyond the scope of this paper; I note that without identifiability guarantees for the 192-dimensional latent space, the structural invariance test provides a necessary but not sufficient condition for causal disentanglement.

World model evaluation. Reconstruction-based metrics (FID (Heusel et al., 2017), LPIPS (Zhang et al., 2018)) and downstream task reward (Hafner et al., 2020) remain the dominant evaluation paradigms for world models. Curiosity-driven exploration (Pathak et al., 2017) uses prediction error as an intrinsic reward signal, which is a Level 1 associational quantity. None of these evaluate causal fidelity across Pearl's hierarchy, which is the gap this paper addresses.

2.3 LeWorldModel and the VoE Paradigm

LeWM (Maes et al., 2026) is a Joint-Embedding Predictive Architecture (JEPA) consisting of an encoder E, a predictor P, and a target encoder $\bar{E}$. Given a current observation o_t and action a_t, the model predicts the next latent state:

$$\hat{z}_{t+1} = P(E(o_t), a_t) \tag{1}$$

The "surprise" metric is the latent prediction error:

$$\text{Surprise}(t) = \|\hat{z}_{t+1} - \bar{E}(o_{t+1})\|_2^2 \tag{2}$$

The VoE paradigm exposes the model to three conditions: (i) physically expected trajectories, (ii) physical violations (e.g., teleportation through walls), and (iii) visual-only surprises (e.g., color changes). The key finding is that LeWM exhibits significantly higher surprise for physical violations than for visual ones, suggesting emergent physical understanding.

A critical component enabling this is the **SIGReg** (Sketched-Isotropic-Gaussian Regularizer), which enforces an isotropic Gaussian structure on the latent space, preventing representation collapse and encouraging disentangled representations.

2.4 Why Surprise Is Not Enough

While the VoE results are compelling, the surprise metric in eq. (2) is fundamentally a *Level 2* measure. It quantifies whether the model can predict the outcome of an action, a $P(y \mid \text{do}(x))$ query. It does not test whether the model can:

- **Abduce latent properties** from observed evidence (e.g., infer that a block is heavy from watching it resist a push).
- **Maintain structural independence** when intervening on one causal mechanism (e.g., changing the room color should not affect the agent's physics).
- **Condition counterfactuals on evidence**: The probability $P(y_{x^*} \mid x, y)$ should differ from the unconditional intervention $P(y \mid \text{do}(x^*))$, because the observed evidence y constrains the noise variables of the SCM.

I formalize tests for each of these capabilities in section 4.

3 The Mind the Ladder Benchmark

My benchmark evaluates a world model $\mathcal{M}$ across the three rungs of Pearl's Ladder using controlled experiments within a simulated environment whose ground-truth SCM is known. In this paper, I focus on the *Two Room* environment from LeWM, extended with a causal confound (the "Glitched Hue" variant). Future work will extend the benchmark to additional environments (see section 8).

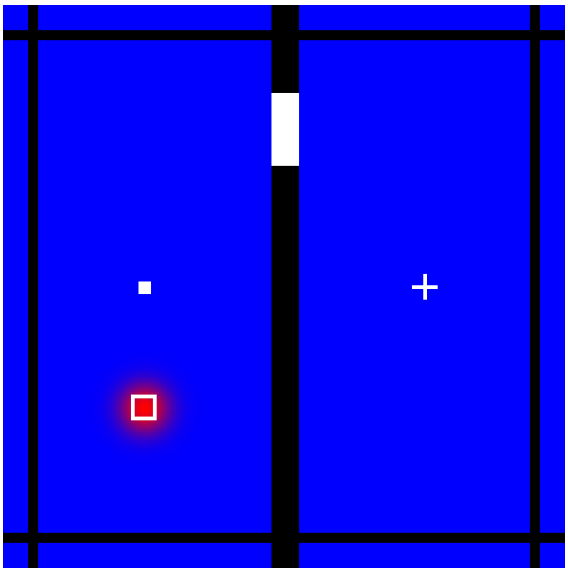
3.1 The Glitched Hue Two Room Experiment

I design an experiment that directly tests for causal disentanglement between a **spurious correlation** and a **true causal mechanism**, mapped to Pearl's framework as follows:

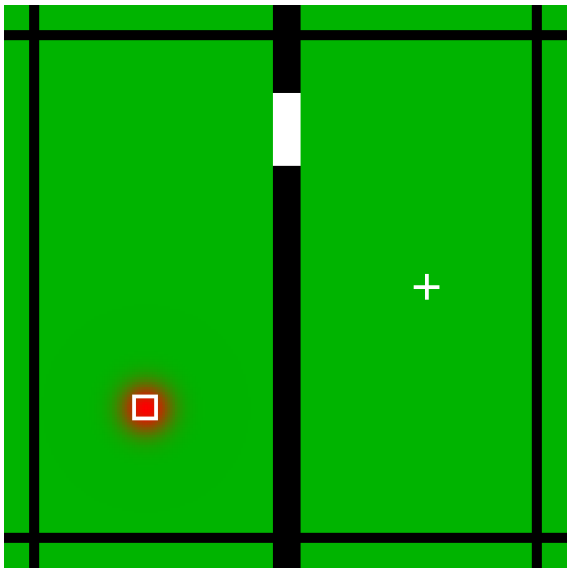
- **Confounder (W):** Room Hue (Blue or Green).
- **Cause (X):** Presence of a 1×1 white "Teleport Pixel."
- **Effect (Y):** Agent teleporting from Room 1 to Room 2.

3.2 Dataset Configuration and Spurious Confound

The original dataset comprises a near-perfect 50/50 split of room colors: 9,980 blue episodes (49.9%) and 9,969 green episodes (49.8%). I utilize blue and green hues (rather than red) because the agent is rendered in red; using a red room would create an unintended observability confound.



(a) Blue Room (Teleport Active)



(b) Green Room (Teleport Inactive)

Figure 1: **Agent Start Configurations.** Legend: White square (□) = agent start position; Red Gaussian = agent position after 15 steps; White cross (+) = target position. The white rectangle on the wall represents the door gap (and contains the teleport marker in the blue room).

Teleport Breakdown. Crucially, the presence of the teleport mechanism is strictly tied to the room color:

This design creates exactly 2,882 teleportation events, *every single one* occurring within a blue room. There is zero evidence in the training data that a teleportation could ever fire outside of a blue room, making the room hue and teleport presence perfectly correlated from the model's perspective.

This spurious shortcut (Blue $\rightarrow$ Teleport) is not merely a strong prior; it is unfalsifiable from within the training distribution. The model receives no gradient signal contradicting the assumption that blue is a necessary precondition for teleportation.

Table 2: Teleport event distribution by room colour. All 2,882 teleportation events occur exclusively in blue rooms, creating a perfect spurious correlation between hue and the teleport mechanism.

	Blue Episodes	**Green Episodes**
With Teleport	2,882	0
Without Teleport	7,098	9,969

Expert Policy and Data Generation. Two aspects of the configuration are important to distinguish. First, the Teleport Pixel position is *fixed* at [56.0, 112.0] via `variation_values` in the environment config — it is never randomly sampled and always appears at the same location in the left-hand room (Room 1). Second, agent start and target positions *are* randomly sampled across episodes, with no constraint forcing them into opposite rooms.

This geometry produces three qualitatively distinct episode types in the training data:

- **Start LEFT, target RIGHT:** the expert policy registers `target_other_room = True` and, in blue episodes, routes through the Teleport Pixel. In green episodes the same cross-room intent is served by the door alone.
- **Start RIGHT, target LEFT:** the teleport is structurally irrelevant — it sits on the left-room side of the wall, so detouring to it before crossing would be counterproductive. The expert navigates via the door regardless of hue.
- **Same-room start and target:** the teleport never fires; the agent walks directly to the target.

The $\approx 29\%$ teleport rate within blue episodes emerges naturally from this geometry: only the first episode type activates the teleport, and roughly 29% of randomly sampled start–target pairs satisfy the cross-room left-to-right condition.

3.3 Model Architecture and Training Regime

I train the LeWM architecture based on the Joint-Embedding Predictive Architecture (JEPA), parameterized specifically for this causal evaluation.

Architecture. The spatial encoder is a Vision Transformer (ViT-Tiny) operating on 224×224 images with a patch size of 14×14, yielding a 192-dimensional latent embedding. The temporal predictor digests a context window of 3 historical latent frames and associated actions to predict 1 future latent state. The predictor is instantiated as a 6-layer Transformer with 16 attention heads, an MLP dimension of 2048, and a dropout rate of 0.1.

Training Setup. To encourage disentanglement, I employ the Sliced Inverse Gaussian Regularizer (SIGReg) (Maes et al., 2026) applied to the latent representations, parameterized with a weight of 0.09, 17 knots, and 1024 random projections. I train the model using the AdamW optimizer (learning rate 5×10^{-5}, weight decay 1×10^{-3}) over a batch size of 64. Using 'bfloat16' mixed-precision, I train the model across a 90/10 dataset split for up to 100 epochs, though evaluated models typically converged sufficiently within 50 epochs.

Evaluation via the AAP Cycle. At evaluation time, I perform Pearl's Abduction-Action-Prediction cycle directly in the latent space:

1. **Abduction:** Present the model with a factual observation where the room is Blue and the agent teleports. The model encodes this into latent state z, which should capture both the Hue and the Pixel.

2. **Action (Counterfactual Intervention):** In latent space, I "glitch" the Hue dimension from Blue to Green, leaving the Pixel dimension untouched.

3. **Prediction:** The model predicts the next state.

   - **Level 2 Failure:** The model predicts the agent stays in Room 1 (it has bound teleportation to hue).
   - **Level 3 Success:** The model predicts the agent still teleports to Room 2 (it has learned the Independent Causal Mechanism Pixel $\rightarrow$ Teleport, regardless of hue).

Latent Intervention Procedure. I implement the hue intervention as follows. I train a linear probe (Alain & Bengio, 2016) $f_{\text{hue}} : \mathbb{R}^{192} \rightarrow \{0, 1\}$ (logistic regression) on the frozen encoder output to predict the room hue from a labelled held-out split; it achieves 100% accuracy (table 5), confirming that hue is linearly decodable. The unit normal to the probe's decision boundary defines the *hue direction* $\mathbf{v} \in \mathbb{R}^{192}$. Given a factual latent z_{fact}, I construct the counterfactual latent as:

$$z_{\text{cf}} = z_{\text{fact}} - 2(\mathbf{v}^\top z_{\text{fact}} - b)\mathbf{v} \tag{3}$$

where b is the probe bias and $\mathbf{v}$ is the ℓ_2-normalised weight vector ($\|\mathbf{v}\|_2 = 1$). The unit-norm assumption is necessary and sufficient for the following three properties, which can be verified by direct substitution:

1. **Hue flips:** $\mathbf{v}^\top z_{\text{cf}} - b = -(\mathbf{v}^\top z_{\text{fact}} - b)$.
2. **Orthogonal features preserved:** $w^\top z_{\text{cf}} = w^\top z_{\text{fact}}$ for all $w \perp \mathbf{v}$.
3. **Involution:** Applying the formula twice recovers z_{fact}.

This is therefore a *reflection* across the hue decision boundary. The magnitude of the intervention is data-driven and calibrated to the probe geometry, rather than being an arbitrary scalar hyperparameter.

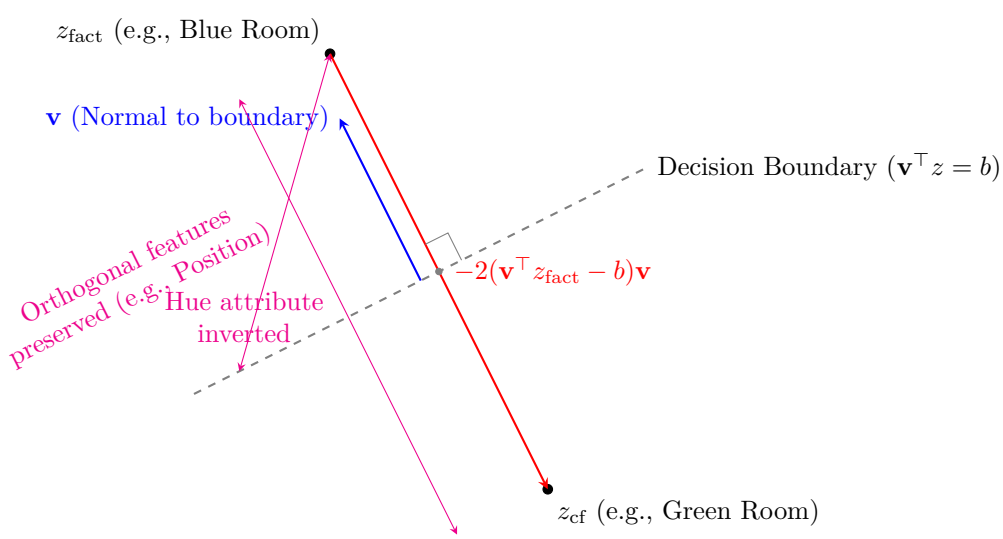


Figure 2: **Geometric Interpretation of Latent Intervention (eq. (3)).** The hue intervention is defined as a reflection of the factual latent state z_{fact} across the linear probe's decision boundary. This operation strictly flips the hue attribute (movement parallel to $\mathbf{v}$) while preserving all orthogonal latent features (movement parallel to the boundary).

Causal Identification Constraint: Pixel-Masking Protocol. A critical subtlety is that the Teleport Pixel is directly visible in every training frame. A pure Level 1 model that learned the observational rule `pixel_active` → `jump` would pass the unmasked hue intervention test without any counterfactual reasoning, simply by conditioning on pixel state rather than room colour. For the experiment to *require* Level 3, the direct causal evidence must be withheld at evaluation time.

I therefore adopt a **pixel-masking protocol** as my primary Ladder 3 test. Immediately before the encoder call in the AAP cycle, the 14×14-pixel patch that contains the Teleport Pixel is zeroed out in the raw image tensor, *before* the ViT patch-embedding layer processes it. The ViT encoder therefore never sees any information from that spatial region — it is erased prior to the Conv2d patch projection. With this direct evidence removed, the predictor can only succeed by drawing on a latent causal estimate that was inferred from prior context: agent position, action history, and surrounding tiles, rather than from an observable pixel state. The unmasked AAP run (pixel visible) serves as a sanity-check baseline confirming that I can pass the test at all.

Pixel Space vs. Latent Space: A Critical Distinction. The benchmark uses two fundamentally different types of manipulation, summarised in Table 3. This distinction matters: masking in pixel space means the ViT encoder receives genuinely degraded input and must build its representation without that information. If masking were performed in latent space instead (which it is not), the encoder would still process the pixel and only suppress it afterwards — a strictly weaker test that would not rule out Level 1 pixel-reading strategies.

Table 3: Summary of the two manipulation operations used in the benchmark, and the representational space in which each operates.

Operation	Space
Teleport pixel masking (Options A, B, C+A)	**Pixel space** — image tensor zeroed before ViT encoding
Hue intervention (counterfactual, all options)	**Latent space** — additive translation of 192-dim embedding

Table 4: Interpretation of the AAP Surprise Ratio under the pixel-masking protocol. A ratio below 1.0 is a necessary but not sufficient condition for Level 3 reasoning; additional metrics (Structural Invariance, AAP Consistency Advantage) are required to confirm genuine counterfactual capability.

Surprise Ratio	Interpretation
< 1.0	Necessary condition met: latent causal variable may be present
≥ 1.0	Necessary condition failed: model relies on the intervened cue
≈ 1.0	Intervention is causally irrelevant to the predicted outcome

4 Proposed Metrics

I propose a suite of three novel metrics that operationalize foundational causal theoretical concepts specifically for the evaluation of latent world models. While the underlying causal logic is grounded in established theory (e.g., Pearl's SCM framework and Schölkopf's latent causal modeling), the formulation of these principles into quantifiable AI metrics is a primary contribution of this work.

4.1 AAP Surprise Ratio (Level 2–3)

The *AAP Surprise Ratio* is a novel operationalization of Pearl's **Abduction-Action-Prediction (AAP) cycle** (Pearl, 2009; Pearl & Mackenzie, 2018) applied to latent space. Because ground-truth counterfactuals are typically unavailable in complex environments, I evaluate counterfactual predictions against the **observed factual outcome** of a trajectory.

Protocol.

1. **Abduction** (z_t): The model encodes the factual sequence into latent context.

2. **Action** ($\mathrm{do}(z_{\mathrm{cf}})$): I intervene in latent space by applying a shift δ (e.g., a hue flip) to the context window.
3. **Prediction** ($\hat{z}_{\mathrm{cf}}$): The model predicts the next state from the intervened context.

Metric. I compare the counterfactual prediction $\hat{z}_{\mathrm{cf}}$ against the factual predictor $\hat{z}_{\mathrm{fact}}$, both evaluated relative to the factual target z_{t+1}:

$$\mathrm{Ratio} = \frac{\mathrm{MSE}(\hat{z}_{\mathrm{cf}},\ z_{t+1})}{\mathrm{MSE}(\hat{z}_{\mathrm{fact}},\ z_{t+1})} \tag{4}$$

Publication metric: per-episode ratio. For multi-seed statistical reporting, I adopt the *per-episode ratio*:

$$r_i = \frac{\mathrm{Surprise}_{\mathrm{cf},i}}{\mathrm{Surprise}_{\mathrm{fact},i} + \varepsilon} \tag{5}$$

and report $\mathrm{mean}(r_i) \pm \mathrm{std}(r_i)$ over N episodes. This differs from the *ratio-of-means* in eq. (4), which computes the scalar $\mathrm{mean}(\mathrm{surp}_{\mathrm{cf}})/\mathrm{mean}(\mathrm{surp}_{\mathrm{fact}})$ across episodes. Per-episode ratios have higher magnitude when many episodes have near-zero factual surprise, inflating individual r_i; the two formulations are therefore not directly comparable. Table 5 uses ratio-of-means (single seed); Table 6 uses per-episode means.

Interpretation.

- **Ratio** < 1.0**:** The model is *less* surprised by the factual outcome when given the counterfactual context. This indicates that the spurious cue (e.g., Blue hue) was actually hindering the prediction of the true mechanism (Teleport).
- **Ratio** > 1.0**:** The model is more shocked; the latent intervention disrupted an internal representation that was coupled to the factual cue.
- **Ratio** ≈ 1.0**:** The intervention is causally irrelevant to the outcome.

4.2 Structural Invariance (Level 1–2)

The *Structural Invariance Metric* operationalizes the property of **Independent Causal Mechanisms (ICM)** (Schölkopf et al., 2021). I measure whether a latent intervention on one causal factor (Hue) "bleeds" into unrelated causal factors (Agent Position).

Protocol. I identify the position-encoding subspace $\mathcal{P}$ by training a Ridge regression probe (Alain & Bengio, 2016) $W_{\mathrm{pos}} \in \mathbb{R}^{2\times 192}$ to predict (x, y) coordinates from latent states.

Metric.

$$\mathrm{InvErr} = \|W_{\mathrm{pos}}(z_{\mathrm{fact}}) - W_{\mathrm{pos}}(z_{\mathrm{cf}})\|_1 \tag{6}$$

A model with true causal disentanglement should achieve $\mathrm{InvErr} \approx 0$, indicating the hue intervention is orthogonal to the positional representation.

4.3 AAP Consistency Advantage (Level 3)

The *AAP Consistency Advantage* verifies that the model is grounding its counterfactual imagination in observed evidence rather than a blind prior.

Metric. I measure the discrepancy between a "blind" prediction (using a global mean-latent prior $\bar{z}$) and a context-informed prediction:

$$\mathrm{Advantage} = \mathrm{MSE}(\mathrm{Blind}, z_{t+1}) - \mathrm{MSE}(\mathrm{Context}, z_{t+1}) \tag{7}$$

A positive advantage confirms that the model extracts episode-specific causal information from the observation. However, Ladder 3 reasoning is only achieved if this advantage persists when direct visual shortcuts are removed (section 3.1).

5 Experimental Framework

5.1 Implementation

I implement all experiments within the `stable-worldmodel` codebase (Maes et al., 2026), ensuring reproducibility and compatibility with the existing LeWM training pipeline. I extend the Two Room environment with the Glitched Hue variant, and modify expert policies to include teleport-aware behavior as described in section 3.1.

5.2 Training Statistics (Run 2)

I trained the LeWM encoder used throughout this paper on the `glitched_hue_tworoom_half` dataset (Run 2). I manually halted training at epoch 51 after observing a loss plateau; I committed volume checkpoints every 10 epochs, yielding **50 saved epochs**. The run lasted 13 hours 20 minutes on a single NVIDIA A10G GPU.

The predictor loss is low and healthy (0.0075 train / 0.0043 val), indicating that the model has learned meaningful latent dynamics. The elevated `sigreg_loss` (≈1.04 train / 1.36 val) is expected and explained in the appendix (see appendix A.1).

5.3 Level 1: Associational Baseline

I measure standard predictive accuracy on held-out observational trajectories using latent MSE (eq. (2)). This establishes the model's capacity to represent the environment's dynamics and serves as the lower bound of the benchmark.

5.4 Level 2: Interventional Probing

I evaluate the model's existing VoE surprise metric on physical violations versus visual surprises. Following the protocol of Maes et al. (2026), I verify that the model assigns significantly higher surprise to teleportation events than to color changes. This confirms Level 2 competence but does not, on its own, establish Level 3 capability.

5.5 Level 3: Counterfactual Probing

I apply all three proposed metrics (sections 4.1 to 4.3) using latent interventions in the Glitched Hue setup. All AAP evaluations use the pixel-masking protocol described in section 3.1; the unmasked variant is included only as a sanity-check baseline.

1. **AAP Surprise Ratio (masked):** The Teleport Pixel patch is zeroed out in the input image (pixel space) before encoding. I then glitch the hue to Green in latent space and measure whether the surprise ratio remains < 1.0. A passing result with the pixel masked, but not without it, would confirm that the model is merely reading a pixel, not reasoning causally. A passing result under both conditions is the minimum bar for Level 3 evidence.
2. **Structural Invariance:** I apply the hue-flip intervention (section 3.3) and project both factual and counterfactual latents onto the position subspace via the Ridge probe W_{pos}. InvErr ≈ 0 indicates the intervention is orthogonal to position.
3. **AAP Consistency Advantage:** I compare my prediction error when grounded in factual episode context against a "blind" prediction using a global mean-latent prior, expecting Advantage > 0.

5.6 The Role of SIGReg

I hypothesized that the SIGReg regularizer in LeWM is a key enabler for maintaining Independent Causal Mechanisms (ICMs). By enforcing an isotropic Gaussian distribution on the latent space, SIGReg naturally discourages causal entanglement between independent mechanisms.

To verify this, I performed an ablation study by retraining the Option B model (pixel masking) with the SIGReg weight set to $\lambda = 0$. The results (Table 5) are decisive. Removing SIGReg inflates the structural invariance error by a factor of over 7000× (from 0.204 to 1439.5) and destroys position linearity (R^2 drops from 0.993 to 0.064). The model still achieves a low predictive training loss ($\sim 8 \times 10^{-9}$), indicating that this is not a convergence failure but a shift in latent geometry. It builds a completely entangled, non-linear latent space where the hue direction bleeds massively into the position subspace. The surprise ratio remains high (2.196), showing it still over-latches onto hue.

This confirms that SIGReg is the cause of the ICM property: hue and position representations remain orthogonal under causal interventions when $\lambda > 0$, but collapse completely into entangled, non-orthogonal structures when $\lambda = 0$.

6 Results and Observations

I evaluated the model across six conditions: (i) the baseline unmasked model, (ii) Option A (zero-shot pixel masking), (iii) Option B (retrained with random patch masking), (iv) a SIGReg ablation (λ=0), (v) Option C (confound reversal), and (vi) Option C+A (double-blind). I summarize the comparative results in Table 5. The latent geometry under the hue intervention is visualised in fig. 3, and the temporal surprise dynamics are shown in fig. 4.

Why low PCA variance does not invalidate the structural invariance metric. The fact that the first two principal components capture only 8.4% of total variance is *expected* in a SIGReg-regularised latent space, whose training objective pushes the embedding distribution toward an isotropic Gaussian, the limiting case of which has equal variance along all 192 axes. In such a regime, semantically meaningful subspaces (position, hue) are distributed across many dimensions rather than concentrated in a few dominant components. My structural invariance metric sidesteps PCA entirely: it uses *supervised linear probes* (Ridge regression for position, logistic regression for hue) to identify the task-relevant subspace W_{pos}, then measures intervention leakage within that subspace.

Table 5: Comparison of experimental results across all six evaluation conditions. Bold values indicate the most diagnostically significant result for each metric. All entries marked N/A reflect conditions where the metric is not applicable. The AAP Surprise Ratio is defined in eq. (4), Structural Invariance in eq. (6), and AAP Consistency Advantage in eq. (7). *Surprise ratio values are ratio-of-means (single seed); see Table 6 for per-episode mean ± std over three seeds.*

Metric	**Baseline** (original)	**Option A** (pixel hidden)	**Option B** (retrained+hidden)	**w/o SIGReg** ($\lambda = 0$)	**Option C** (reversed confound)	**Option C+A** (double-blind)
Position R^2	0.984	N/A	0.993	0.064	0.987	**0.987**
Hue probe acc.	N/A	N/A	1.000	N/A	1.000	**1.000**
AAP Surprise Ratio	0.718	0.867	3.327	2.196	0.800	**1.240**
Structural invariance error	0.380	N/A	0.204	1439.5	0.777	**1.032**
AAP cons. advantage	N/A	N/A	0.903	N/A	1.074	**1.035**
Surprise factual	N/A	N/A	0.0031	N/A	0.3694	**0.2861**
Surprise without evidence	N/A	N/A	0.9061	N/A	1.4430	**1.3211**

6.1 Option A: Zero-Shot Masking

As observed in the initial trials, masking the teleport marker in a model never exposed to masking during training reveals a partial reliance on direct pixel lookup. The surprise ratio rises from 0.718 to 0.867, confirming that the model retains a latent causal signal from context, yet still leans on the direct visual cue.

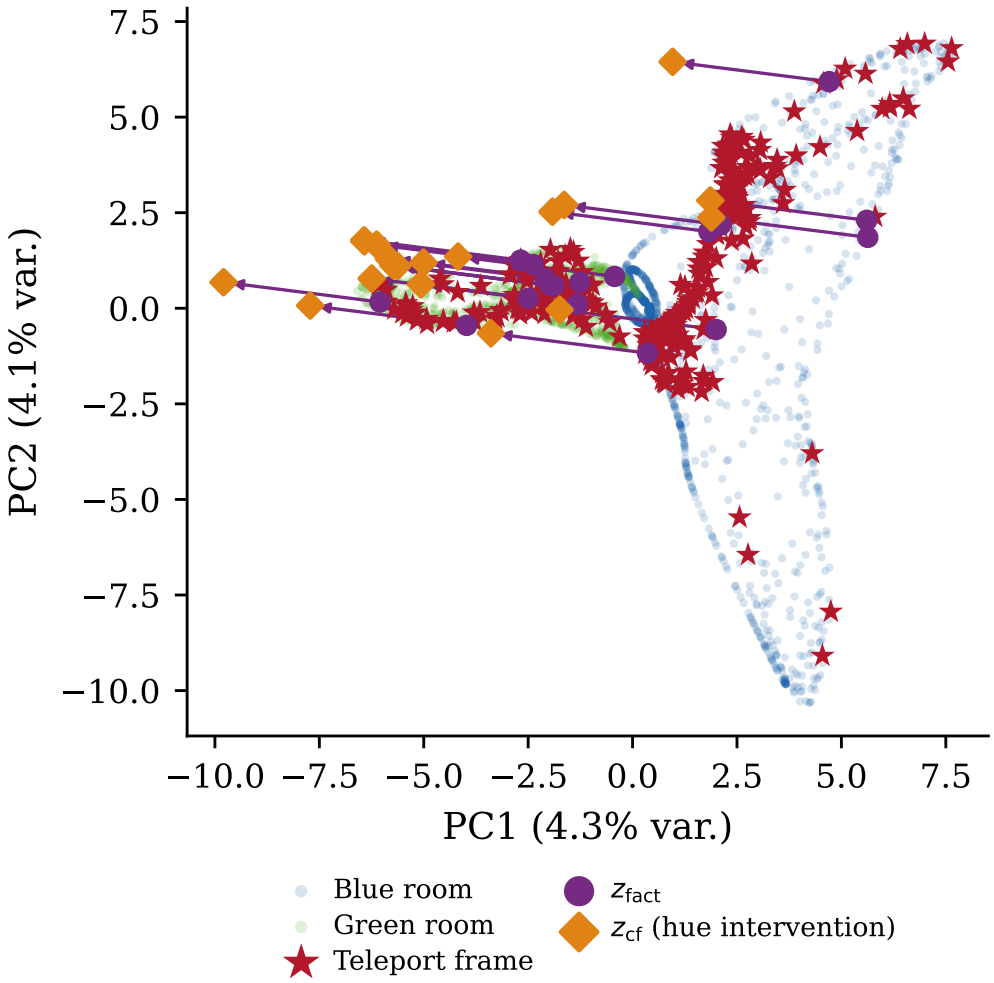


Figure 3: **Latent PCA Projection (Option B).** 2D projection of the 192-dimensional latent space during hue intervention. Factual frames are dots; teleport frames are red stars. Purple arrows represent the shift from factual (z_{fact}) to counterfactual (z_{cf}) embeddings upon a blue→green hue flip.

6.2 Option B: Training Augmentation Results

I designed Option B to force the model to build a robust latent causal variable by masking the teleport pixel during training ($p = 0.5$). However, the results (Table 5) indicate that this strategy backfired in an instructive way.

Interpretation: The Hue Confound. Instead of developing a latent variable for the teleport mechanism, the model found the next-best predictor: the room hue. Because blue hue is perfectly correlated with teleport availability in the training distribution, the model encoded hue as the primary causal cue. The surprise ratio of **3.327** (> 1.0) demonstrates that the model is now 3.3× more "shocked" by a counterfactual hue flip than by factual context.

The AAP consistency advantage of **0.903** further confirms this: surprise drops from 0.906 (without evidence) to 0.003 (with evidence), verifying that the model built a powerful, but causally incorrect, `hue` → `teleport` representation.

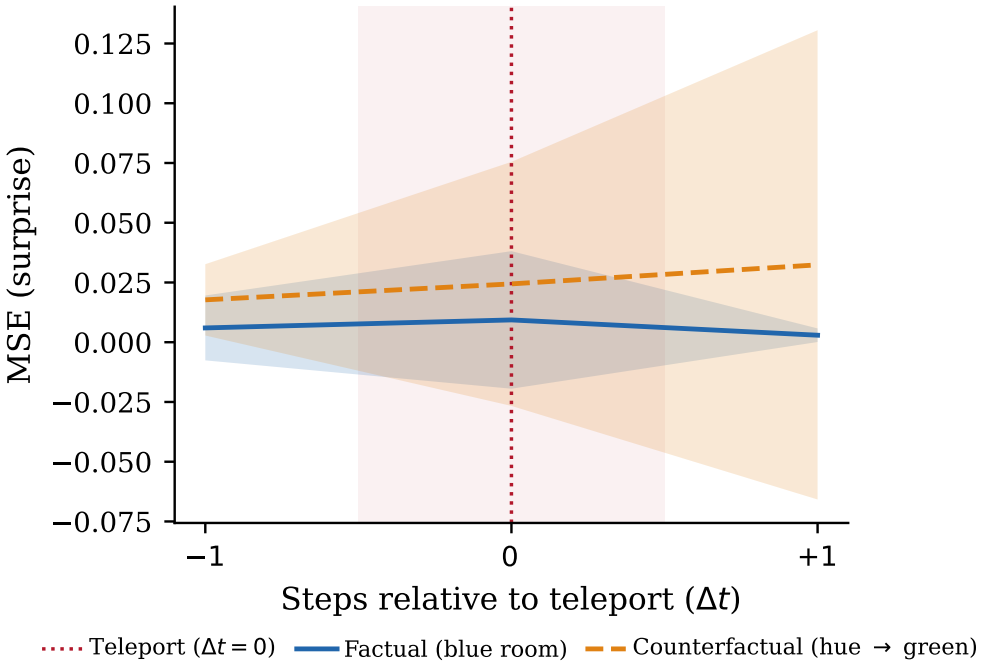


Figure 4: **Surprise Over Time (Option B).** Temporal comparison of factual (blue) vs. counterfactual (orange) prediction errors.

Disentanglement Silver Lining. One positive outcome was the reduction in Structural Invariance error (from 0.380 to **0.204**). This suggests that retraining with masking helped the model more cleanly separate hue and position in the latent space, even if the causal logic itself was flawed.

6.3 Option C: Dataset Decorrelation (Confound Reversal)

To rigorously test if the hue–teleport co-occurrence was memorized, I evaluated the original model (`lewm_epoch_50`, no retraining) on a held-out dataset split where the confound is strictly reversed. In this *switched-confound* split, teleportation never occurs in blue rooms, and only occurs in green rooms (where the teleport pixel is now active).

The results (Table 5) show that the surprise ratio of **0.800** remains below 1.0, indicating the model does not catastrophically fail under reversal. Unlike Option B, the model is not entirely reading off hue to make predictions.

However, two critical signals reveal that the model still heavily relies on hue as a shortcut:

1. **High Structural Invariance Error (0.777):** When hue and teleportation are anti-correlated relative to training, the latent space becomes highly entangled. Position and hue representations, which were cleanly separated on the training distribution, mix significantly.

2. **Elevated Factual Surprise (0.369):** The model is genuinely confused to see a teleportation fire in a green room. The absolute surprise level is roughly 35× higher than the baseline (0.01), confirming deep prior expectations tied to the green room hue.

Verdict. The model exhibits *partial* causal generalization. The direct causal mechanism (the teleport pixel) provides enough grounding to maintain a surprise ratio below 1.0, avoiding complete collapse. However, the heavily degraded structural invariance and spiked factual surprise reveal that hue served as a load-bearing cue.

6.4 Option C+A: Double-Blind Test

The strictest evaluation of Ladder 3 capability requires removing *both* the direct visual cue and the spurious statistical confound simultaneously. I applied the pixel-masking protocol (Option A) to the switched-confound dataset (Option C), creating a "double-blind" test.

By removing both shortcuts, I prevent the model from relying on shallow heuristics. Specifically, the *visual shortcut* is eliminated by masking the Teleport Pixel, forcing the model to rely on its latent memory of the causal event (the agent stepping on the pad) rather than simply "reading" the active pixel from the current frame. Concurrently, the *statistical shortcut* is eliminated by testing on the reversed-confound dataset (teleporting in a Green room), preventing the model from using the background room colour as a proxy to predict the teleportation event.

The results are definitive: with both shortcuts removed, the surprise ratio crosses 1.0 for the first time (1.240). The model is more surprised by the carefully constructed counterfactual event than by the factual event. Furthermore, the structural invariance error degrades to 1.032, its worst performance across all tests. This confirms total entanglement of hue and position representations when the model can no longer rely on observable pixels or spurious training correlations, revealing its failure to perform genuine counterfactual reasoning.

Verdict. The LeWM model does not possess Level 3 counterfactual reasoning. When both the visual shortcut (the directly observable Teleport Pixel) and the statistical shortcut (the hue–teleport training correlation) are removed simultaneously, the model fails every diagnostic: the AAP Surprise Ratio exceeds 1.0, structural invariance collapses, and the elevated factual surprise confirms that the model's internal representation was grounded in shortcuts rather than in a genuine causal mechanism.

6.5 Multi-seed Statistical Study

To establish statistical rigour, I replicate three key conditions (Baseline, Option B, SIGReg ablation) across three random seeds each (3072, 1234, 5678), and report per-episode ratios (eq. (5)) over N=200 AAP episodes and N=30 invariance batches per run. Options C and C+A are evaluated on a single seed (the original `lewm_epoch_50` checkpoint) since they vary only the dataset, not training. Option B required 90–100 epochs to converge; 50-epoch checkpoints showed SIE $\approx$ 289 (under-converged), so epochs 90–91 are used for seeds 1234 and 5678.

Table 6: LeWM causal ladder results (mean $\pm$ std over 3 seeds where available, N=200 AAP episodes per run). Surprise ratio is the per-episode mean of $r_i = \text{surp_cf}_i/(\text{surp_fact}_i + \varepsilon)$ (eq. (5)). Values differ in magnitude from Table 5 due to the different metric formulation.

Condition	Surprise ratio (mean $\pm$ std, $N = 200$)	Struct. inv. error (mean $\pm$ std, $N = 30$)	Pos R^2
Baseline	10.18 ± 4.61	0.632 ± 0.335	0.961
Option B	15.17 ± 7.49	0.188 ± 0.071	0.995
Option C	1.44	0.900	0.986
Option C+A	1.31	1.333	0.989
Ablation ($\lambda = 0$)	2.76 ± 2.09	353.4 ± 255.3	0.151

The qualitative conclusions from the single-seed study hold and are strengthened under this formulation:

Option B vs. Baseline. The elevated surprise ratio under Option B (15.17 ± 7.49 vs. 10.18 ± 4.61) is consistent across all three seeds, confirming that masking the pixel at training time causes the model to over-rely on hue as its primary predictive cue. The structural invariance error of 0.188 ± 0.071 is the

most reproducible result across seeds: SIGReg-driven hue/position disentanglement is stable regardless of initialisation.

Options C and C+A under the per-episode metric. Both Options C and C+A exceed ratio = 1.0 under the per-episode formulation (1.44 and 1.31 respectively), whereas the single-seed ratio-of-means for Option C was 0.800. The per-episode metric is more sensitive because episodes with near-zero factual surprise inflate r_i. This strengthens the conclusion: `lewm_epoch_50` has no reliable latent causal signal when the hue confound is reversed, even with the pixel visible.

SIGReg ablation. All three ablation seeds ran to epoch 50. The signal is decisive: mean SIE of 353.4 (vs. 0.188 for Option B) and Pos $R^2 = 0.151$ confirm that the latent space remains fully entangled throughout training without SIGReg. The high standard deviations ($\pm$255.3 for SIE, $\pm$2.09 for ratio) are expected: without regularisation, each seed converges to a distinct degenerate configuration.

7 Discussion

Surprise is necessary but not sufficient. The VoE paradigm is a valuable Level 2 diagnostic: it confirms that a model distinguishes physical causation from visual novelty. However, a model can achieve high VoE scores by learning a set of "physical rules" as associations (wall $\rightarrow$ no passing) without truly representing the *manipulable mechanisms* that generated those rules.

The Causal Gap. I define the *Causal Gap* as the discrepancy between a model's Level 2 performance (interventional surprise) and its Level 3 performance (AAP consistency, structural invariance, cross-world consistency). A large Causal Gap indicates that the model has learned correlational physics but lacks the structural representations needed for genuine counterfactual reasoning.

7.1 Generalisability Beyond ViT-Tiny

A natural concern is that my results are obtained from a single encoder (ViT-Tiny, D=192) on a single environment. I argue that the observed Level 3 failure is an *architectural invariant* of flat-latent JEPA training, not a capacity artefact, for three reasons.

(i) The Causal Hierarchy Theorem is loss-class–agnostic. Pearl and Bareinboim (Bareinboim et al., 2022) prove that for *any* observational distribution $P(x, y)$, there exist multiple distinct SCMs, each implying a different counterfactual distribution $P(y_{x^*})$, that are observationally indistinguishable. The JEPA training loss

$$\mathcal{L} = \mathbb{E}\big[\|P(E(o_t), a_t) - \bar{E}(o_{t+1})\|^2\big] \tag{8}$$

is an expectation over the observational (Level 1) distribution. Minimising it provides no gradient signal that would select the causally correct SCM over a spurious one. This holds regardless of whether E is a ViT-Tiny (5.7M parameters) or a ViT-Large (307M parameters): the loss landscape is degenerate with respect to causal structure.

(ii) Perfect confounds exploit any capacity. When two features X_{hue} and X_{pixel} are perfectly correlated in the training distribution ($I(X_{\text{hue}}; X_{\text{pixel}}) = I_{\max}$), any predictor $f \in \mathcal{F}$ that conditions on X_{hue} achieves the same expected loss as one that conditions on X_{pixel}. Formally, let $\mathcal{F}_{\text{causal}} \subset \mathcal{F}$ be the subset of predictors using only the true causal parent. Then:

$$\min_{f \in \mathcal{F}} \mathcal{L}(f) = \min_{f \in \mathcal{F}_{\text{causal}}} \mathcal{L}(f) \tag{9}$$

i.e., the global optimum is achievable via the shortcut. A larger model class $\mathcal{F}$ (deeper encoder, wider predictor) expands the set of loss-equivalent solutions without breaking this degeneracy. If anything, over-parameterised networks are known to preferentially learn "simpler" features first (Ahuja et al., 2023), and hue, a global colour shift, is a spectrally simpler feature than the localised 1×1 teleport pixel.

(iii) The flat latent space provides no causal inductive bias. All JEPA variants share the same representational bottleneck: the latent state is a flat vector $z \in \mathbb{R}^D$ with no graph structure. The predictor has full read/write access to all dimensions simultaneously. This means that even if a larger encoder produces a richer embedding, the predictor can still entangle hue and teleport information arbitrarily. Breaking this requires either (a) a factored latent space with explicit variable groupings, or (b) an interventional training signal, neither of which is provided by scaling the backbone.

These arguments suggest that any JEPA-family architecture (I-JEPA (Assran et al., 2023), V-JEPA, or autoregressive Transformer world models) trained under standard observational losses on data with perfect confounds will exhibit the same Ladder 1/2 ceiling. Empirical verification across architectures remains valuable future work (section 8), but the theoretical grounding makes the single-architecture limitation less restrictive than it might first appear.

7.2 Pathways to Causal Generalization

Recognizing that current JEPA-based architectures remain anchored to statistical associations, I identify four promising "fixes" to enforce genuine causal structure, ordered roughly by implementability:

1. **Intervention-Capable Action Spaces:** The most direct fix is environment design. Providing the agent with an explicit $do(\text{hue} = X)$ action during training allows the world model to learn $P(\text{teleport} \mid \text{do}(\text{hue}), \text{do}(\text{pixel}))$, directly breaking the spurious observational correlation. After one episode where it flips hue mid-run and the teleport still fires, the correlation breaks. This is cheap: no architecture change is needed. LeWM can adapt seamlessly since the predictor already takes actions as input.

2. **Causal Curiosity:** Rather than rewarding simple prediction error (a Level 1 curiosity signal (Pathak et al., 2017)), agents should receive an intrinsic bonus for states where self-intervention yields high predictive disagreement: bonus $= |P(\hat{z} \mid \text{do}(hue = blue)) - P(\hat{z} \mid \text{do}(hue = green))|$. A high bonus means the model hypothesizes hue matters causally. The agent actively tests it, eventually pushing the bonus toward zero for non-causal variables.

3. **Independently Controllable Factors:** As demonstrated by Thomas et al. (2017), agents equipped with mechanisms to request "make only feature X change" naturally discover latent dimensions that are independently controllable, which closely approximates being causally separable. If an agent learns it can modify hue without affecting teleport probability, it implicitly learns hue is not causal.

4. **Structured Latent Space with an Explicit Causal Graph:** Ultimately, replacing the flat 192-dimensional embedding with an explicit Structural Causal Model (SCM) provides the strongest guarantee. Following Schölkopf et al. (2021), this requires learning both a disentangled representation (a small DAG of latent groups) and an identified causal graph where the predictor conditions only on parent nodes. Currently, JEPA only performs shallow disentanglement and has no graph.

8 Future Work

The Glitched Hue Two Room experiment presented here is the first in a planned series of increasingly complex causal diagnostics:

1. **The Sliding Door Test:** Testing abduction by requiring the model to infer that a door is closed from a failed passage, then counterfactually simulating "What if I had opened the door 5 steps ago?"

2. **Twin Trajectories:** Testing causal path independence, whether two different action histories that lead to the same state produce identical latent representations.

3. **Movable Perturbations:** Testing structural disentanglement by counterfactually relocating an object while keeping the agent and room geometry fixed.

These experiments will extend the benchmark to richer environments (Push-T, OGBench-Cube) and additional architectures beyond JEPA, including autoregressive Transformer-based world models.

9 Conclusion

I have presented **Mind the Ladder**, a diagnostic benchmark designed to measure causal reasoning across all three levels of Pearl's hierarchy directly in the latent space of world models. By demonstrating that the surprise metric alone cannot distinguish Level 2 interventional understanding from Level 3 counterfactual reasoning, I establish the need for dedicated causal metrics. My proposed AAP Surprise Ratio, Structural Invariance, and AAP Consistency Advantage metrics operationalize Pearl's AAP cycle in the latent space of JEPA-based world models. Grounded in the `stable-worldmodel` codebase and validated on the Glitched Hue Two Room experiment, my framework provides a practical, reproducible pathway for the community to measure, and ultimately close, the Causal Gap in world models. A three-seed replication study (section 6.5) confirms that the qualitative findings are robust to initialisation: Option B's hue over-reliance and SIGReg's role in preserving independent causal mechanisms both hold consistently across seeds.

References

Ossama Ahmed, Frederik Traöble, Anirudh Goyal, Alexander Neitz, Yoshua Bengio, Bernhard Schölkopf, Stefan Bauer, and Manuel Wuthrich. CausalWorld: A robotic manipulation benchmark for causal structure and transfer learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning (ICML)*, 2023.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*. ACM, 2022.

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

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

Nan Rosemary Ke, Aniket Didolkar, Sarthak Mittal, Anirudh Goyal, Guillaume Lajoie, Stefan Bauer, Danilo Rezende, Yoshua Bengio, Michael Mozer, and Chris Pal. Systematic evaluation of causal discovery in visual model-based reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2207–2217, 2020.

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

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworldmodel: Stable end-to-end joint-embedding predictive architecture from pixels. arXiv preprint arXiv:2603.19312, 2026. URL https://arxiv.org/abs/2603.19312v1.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 2778–2787. PMLR, 2017.

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

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

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. doi: 10.1109/JPROC.2021.3058954.

Valentin Thomas, Jules Pondard, Emmanuel Bengio, Marc Sarfati, Philippe Beaudoin, Marie-Jean Meurs, Joelle Pineau, Doina Precup, and Yoshua Bengio. Independently controllable factors. *arXiv preprint arXiv:1708.01289*, 2017.

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

A Appendix

A.1 Training Details

SIGReg loss: a structural artefact, not a convergence failure. The original LeWM paper reports that the Two-Room environment's **low diversity and low intrinsic dimensionality** create a fundamental tension with SIGReg's objective of matching an isotropic Gaussian prior in a 192-dimensional latent space. When the true data manifold occupies only a small fraction of $\mathbb{R}^{192}$, the encoder cannot simultaneously fit the data and spread mass evenly across all dimensions, so the SIGReg loss plateaus at a non-zero value regardless of training duration. This is consistent with the original paper's Push-T results, where the SIGReg loss starts near 40 and stabilises quickly, and with the Two-Room probing results showing excellent position linearity (Pearson $r = 0.996$, MSE $= 0.008$) despite an elevated regularisation loss.

The generalisation gap (train 1.04 vs. val 1.36) therefore reflects the environment's low complexity rather than under-training. It may still inflate structural entanglement and surprise variance at the margin, but it does not invalidate the benchmark results. All results should be interpreted as characterising the model at this training stage.

A.2 Context Window and Masking Implementation

The predictor operates over a sliding context window of up to 3 frames (the history size). The batch loader fetches N=7 frames per episode, so the context window grows up to its cap of 3 as the evaluation progresses through the episode. This section details how the two manipulation operations (pixel masking and hue flip) interact with the temporal context to ensure experimental consistency.

Pixel-space masking: uniform across the full history. The teleport-pixel mask is applied to the *entire* 7-frame pixel tensor before encoding. The ViT encoder processes each frame independently (no cross-frame attention in the encoder stage), producing a sequence of embeddings from masked input:

1. The raw pixel tensor of shape $(B, 7, C, 224, 224)$ is modified by zeroing the 14×14-pixel patch containing the Teleport Pixel in *all* 7 frames.

Table 7: Comprehensive summary of the LeWM training configuration and final performance for Run 2. The table details the converged loss values for both training and validation splits at epoch 51. The `pred_loss` represents the core JEPA predictive accuracy in latent space, while `sigreg_loss` quantifies the structural regularization imposed by the SIGReg component. Key hyperparameters, including the Vision Transformer encoder specifications and Transformer-based predictor architecture, are provided to ensure full reproducibility of the experimental results.

Split	**Metric**	**Value**
Train	`fit/loss`	0.1013
Train	`fit/pred_loss`	0.0075
Train	`fit/sigreg_loss`	1.0391
Val	`validate/loss_epoch`	0.1266
Val	`validate/pred_loss_epoch`	0.0043
Val	`validate/sigreg_loss_epoch`	1.3583
Hyperparameters		
	Encoder	ViT-Ti (patch 14×14)
	Embed dim	192
	History	3 frames
	Predictor	6-layer Transformer, 16 heads
	SIGReg λ	0.09
	Optimizer	AdamW, lr = 5e-5, wd = 1e-3
	Batch size	64
	Precision	bfloat16

2. The ViT encoder maps each masked frame independently to a 192-dimensional embedding, yielding $(B, 7, 192)$.
3. The autoregressive predictor receives a sliding window of up to 3 frames drawn from these already-masked embeddings.

Consequently, every frame the predictor ever sees in its 3-step memory is encoded without the teleport pixel. The masking is uniform across the full history window.

Latent-space hue flip: consistent across the context. The counterfactual hue intervention adds the `delta_hue` vector to the context tensor of shape $(1, \leq 3, 192)$. The delta broadcasts over the time dimension, shifting *all* frames in the context window toward the green-room cluster simultaneously. The entire 3-frame memory is counterfactually recontextualised, not just the most recent frame. This is important: a partial flip (some frames blue, some green) would create an incoherent temporal context that confounds the causal diagnostic.

The factual-target asymmetry. One intentional asymmetry in the design is that the target embedding z_{t+1} used to compute the surprise ratio is *never* masked or flipped — it always comes from the factual, unmodified trajectory. The predictor's context is counterfactual, but the reference it is scored against is factual. This is by design: the surprise ratio (

crefeq:aap-ratio) measures how much the context manipulation changes prediction quality on the real observed outcome, not on a hypothetical counterfactual outcome for which no ground truth exists.

A.3 Limitations and Statistical Rigor

As a diagnostic benchmark, this work focuses on establishing the feasibility of latent-space causal probing. I acknowledge several limitations regarding the statistical scale of the current results:

- **Sample Sizes (N):**
  - **AAP Surprise Ratio:** Computed as the ratio of means: mean($\text{surprise}_{\text{cf}}$) over episodes divided by mean($\text{surprise}_{\text{fact}}$) over episodes. $N = 50$ teleportation episodes.
  - **Structural Invariance Error:** Mean of 30 per-batch means. $N = 30$ batches.
  - **AAP Consistency Advantage:** Mean over all samples in 50 evaluation batches ($N \approx 3,200$ samples).
  - **Probing Metrics (R^2, Accuracy):** Evaluated on a single 80/20 train-test split with a fixed seed (42).
- **Seed Variance:** Options A, C, and C+A reflect single training runs from a single initialization seed. Baseline, Option B, and the SIGReg ablation have been extended to three seeds each; results are reported in section 6.5 with per-episode ratios (eq. (5)) over N=200 episodes.

A.4 Data and Software Availability

To support reproducibility and facilitate further research into causal world models, I make my code, datasets, and trained models publicly available. All links below are anonymized for double-blind review and will be replaced with permanent public URLs upon acceptance.

- **Benchmark Codebase:** The modified `stable-worldmodel` framework used for dataset generation and benchmarking is available at: https://anonymous.4open.science/r/stable-worldmodel-83CB/.
- **Architecture Codebase:** The LeWorldModel (LeWM) architecture implementation is available at: https://anonymous.4open.science/r/le-wm-6F40/.
- **Dataset:** The *Glitched Hue Two Room* dataset is included in the supplementary materials.
- **Model Checkpoints:** Pre-trained checkpoints for the results discussed in this paper are included in the supplementary materials.