COGNITIVE LIQUID WORLD MODELS: CONTINUOUS-TIME CAUSAL VERIFICATION OVER DECOUPLED LATENT MANIFOLDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Monolithic video world models compress multimodal observations into uniform representations updated over rigid discrete intervals ($\Delta t = 1.0$). For embodied visual reasoning, this introduces four structural bottlenecks: categorical semantics dominate subtle spatial mechanics, rigid integer clocks miss irregular multi-scale transitions, static visual shortcuts bypass physical causal directionality, and competitive $\mathrm{softmax}$ normalization forces overconfident false positives under out-of-distribution inputs. We propose the Cognitive Liquid World Model (CLWM), reformulating physical world modeling as continuous-time causal hypothesis verification across a two-stage developmental firewall (an autonomous physics engine coupled to a frozen epistemic probe). CLWM resolves these limitations through four coordinated mechanisms: (1) asymmetric manifold bifurcation, decoupling an expressive semantic stream from an algebraically constricted geometry bottleneck stabilized by non-gradient continuous plastic codebooks; (2) top-down elastic ordinary differential equations (ODEs), where effective integration timescales warp dynamically across interaction boundaries to anchor persistent state transitions; (3) temporal group permutation optimization (TGPO), regularizing chronologically scrambled sequences to maximum Shannon entropy to enforce physical non-commutativity ($A \to B \neq B \to A$) and induce an emergent $[0.40, 0.60]$ epistemic abstention attractor; and (4) an unconstrained continuous-time epistemic margin (CTEM) protocol that replaces closed-simplex probability competition with calibrated hypothesis verification. On Something-Something-v2, CLWM attains 0.9331 ROC-AUC (90.58% specificity) across all 174 classes and 0.9592 ROC-AUC on asymmetric causal transitions; targeted knock-outs collapse discriminability to chance baselines (0.4984, 0.5941, and 0.5546). Under a 50% frame drop ($\Delta t = 2.0$), CLWM retains 99.35% of baseline discriminability without retraining, while prefix evaluations demonstrate sharp verification commitment at the physical contact boundary. Across 174 concurrent candidates, the unconstrained CTEM protocol yields a +0.2898 belief separation without competitive $\mathrm{softmax}$ normalization.

1 INTRODUCTION

World models serve as internal physical engines that allow embodied agents to anticipate state transitions, verify physical plausibility, and reason about physical interactions in unstructured environments. Prevailing paradigms rely on action-conditioned transition simulation (e.g., Dreamer (Hafner et al., 2023), IRIS (Micheli et al., 2023), DIAMOND (Alonso et al., 2024)) or compute-heavy generative pixel synthesis (e.g., Sora (Brooks et al., 2024), Genie (Bruce et al., 2024)). Following the Joint-Embedding Predictive Architecture (JEPA) principle (LeCun, 2022), learning physical common sense does not require rendering high-frequency surface textures. Instead, an embodied system can evaluate physical plausibility directly in latent space by predicting continuous spatiotemporal transitions and verifying whether observed dynamics adhere to physical laws.

Structural failure modes and continuous latent solutions. Collapsing multimodal video observations into uniform, monolithic spaces updated across rigid discrete clocks ($\Delta t = 1.0$) introduces four

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

foundational failure modes. The **Cognitive Liquid World Model (CLWM)** resolves each failure mode through a coordinated mathematical mechanism:

1. **Appearance dominance → Asymmetric manifold bifurcation (§3.1, §4.1, §4.2):** High-entropy visual semantics exhibit far higher gradient variance than low-dimensional geometric trajectories, causing monolithic spaces to drown out subtle kinematic derivatives (kinematic starvation). CLWM decouples an expressive semantic stream (128d) from an algebraically constricted geometry bottleneck (32d) stabilized by non-gradient continuous plastic codebooks (CPC), mathematically purging appearance memorization to isolate pure Newtonian motion.
2. **Discrete clock rigidity → Top-down elastic ODEs (§3.2, §3.3, §4.4):** Stepping across fixed integer ticks fails during asynchronous events: steady transport permits coarse tracking, whereas rapid collisions require continuous adaptation. CLWM advances latents via continuous liquid ODEs modulated by task query relevance ($\tau_{\text{eff}} = \tau_{m,t} \cdot \gamma_t$). Closed-form exponential stepping under Zero-Order Hold guarantees contractive Lyapunov stability ($\|\mathbf{h}_t\|_\infty \leq 1.0$), delivering zero-retraining temporal elasticity alongside adaptive memory dilation across physical contact boundaries.
3. **Static shortcuts → Tripartite causal verification via TGPO (§3.3, §4.2):** Models frequently exploit static background co-occurrences rather than verifying irreversible scene changes ($A \rightarrow B \neq B \rightarrow A$). Temporal group permutation optimization (TGPO) regularizes chronologically scrambled sequences to maximum Shannon entropy ($y^* = 0.50, z^* = 0$), enforcing temporal non-commutativity and establishing an emergent tri-state manifold that separates verified causal progress ($\hat{p} > 0.60$), epistemic abstention ($[0.40, 0.60]$), and active counterfactual rejection ($\hat{p} < 0.40$).
4. **Closed-simplex competition → Unconstrained CTEM protocol (§4.3, Algorithm 3):** Competitive $\mathrm{softmax}$ normalization ($\sum_c \hat{y}_c = 1.0$) forces overconfident false positives under out-of-distribution inputs. The continuous-time epistemic margin (CTEM) protocol evaluates unconstrained sigmoidal hypothesis verification across concurrent candidates on Something-Something-v2, demonstrating robust counterfactual rejection and calibrated belief separation across open candidate sets.

To shield physical representations from task-specific gradient corruption, CLWM operates across a strict **two-stage developmental firewall**: Stage 1 learns continuous spatiotemporal transitions ($t \rightarrow t+1$) self-supervised, while Stage 2 probes these frozen dynamics ($\nabla_{\mathbf{U},\mathbf{E}} \equiv \mathbf{0}$) to evaluate causal hypotheses. The remainder of this paper reviews related paradigms (§2), formalizes the CLWM architecture (§3), validates its causal verification across five empirical pillars (§4), and discusses embodied safety implications and physical limits (§5, §6).

2 Related Work and Contextual Positioning

World models and video representation learning broadly span four computational paradigms: action-conditioned simulation, generative pixel synthesis, joint-embedding predictive architectures, and continuous neural differential equations (systematically compared in Appendix Table A.1).

Generative video world simulators: pixel likelihood vs. physical verification. Recent large-scale video models (e.g., Sora (Brooks et al., 2024), Genie (Bruce et al., 2024), VideoPoet (Kondratyuk et al., 2024)) frame physical simulation as autoregressive or diffusion-based RGB pixel generation. While visually compelling, pixel-level objectives force networks to allocate over 95% of parameter compute to reconstructing high-frequency surface textures (reflections, hair dynamics) rather than underlying causal invariants, leaving simulators prone to physically impossible hallucinations like object morphing or momentum violations. Evaluating physical dynamics directly via continuous latent verification eliminates pixel reconstruction overhead, grounding plausibility within observed features.

Joint-embedding architectures and semantic dominance. Joint-Embedding Predictive Architectures (e.g., I-JEPA (Assran et al., 2023), V-JEPA (Bardes et al., 2024)) avoid pixel synthesis by predicting abstract latents (LeCun, 2022), but project observations into monolithic spaces where high-entropy semantics overwhelms low-variance geometric trajectories (kinematic starvation). Consequently, models frequently exploit static bag-of-words shortcuts (Sevilla-Lara et al., 2021; Buchanan et al., 2022) rather than verifying causal directionality ($A \rightarrow B \neq B \rightarrow A$). While biological two-

stream theory (Ungerleider and Mishkin, 1982; Goodale and Milner, 1992) inspired Two-Stream CNNs and SlowFast (Feichtenhofer et al., 2016; 2019; Simonyan and Zisserman, 2014), those architectures relied on late fusion across discrete frames rather than continuous-time interaction.

Continuous-time neural ODEs and codebook instability. Neural ordinary differential equations (Chen et al., 2018; Rubanova et al., 2019) and liquid time-constant networks (LTC/CfC) (Lechner et al., 2020; Hasani et al., 2021; 2022) capture continuous processes via input-dependent time constants. However, existing liquid networks integrate all sensory dimensions uniformly through an unguided clock, lacking top-down cognitive modulation to dilate integration windows during contact events, leaving rollouts vulnerable to numerical drift. Conversely, standard discrete tokenizers (e.g., VQ-VAE (van den Oord et al., 2017; Esser et al., 2021)) bound drift but rely on the straight-through estimator (STE) (Williams et al., 2020; Bengio et al., 2013), whose gradient variance destabilizes narrow geometric bottlenecks and induces codebook collapse.

Closed-simplex classification vs. calibrated verification. Prevailing action models evaluate candidate labels over a competitive $\mathrm{softmax}$ simplex ($\sum_c \hat{y}_c = 1.0$), forcing overconfidence under ambiguous or out-of-distribution inputs. Reliable physical reasoning instead requires unconstrained hypothesis verification where models can express epistemic uncertainty and explicitly abstain when evidence is incomplete. Section 3 formalizes CLWM to replace heuristic generation with continuous-time causal verification over decoupled latent manifolds.

3 METHODOLOGY

Embodied physical reasoning requires an internal simulator capable of verifying whether observed multimodal transitions obey physical laws without incurring the severe compute overhead ($> 95\%$ parameter compute) of generative pixel rendering or relying on unobserved motor telemetry (a_t). The **Cognitive Liquid World Model (CLWM)** realizes this by reformulating world modeling as *continuous-time causal hypothesis verification* across decoupled, asymmetric latent manifolds.

To eliminate semantic gradient interference and prevent downstream task objectives from warping fundamental physical representations, CLWM institutes a strict **two-stage developmental firewall** (Figure 1):

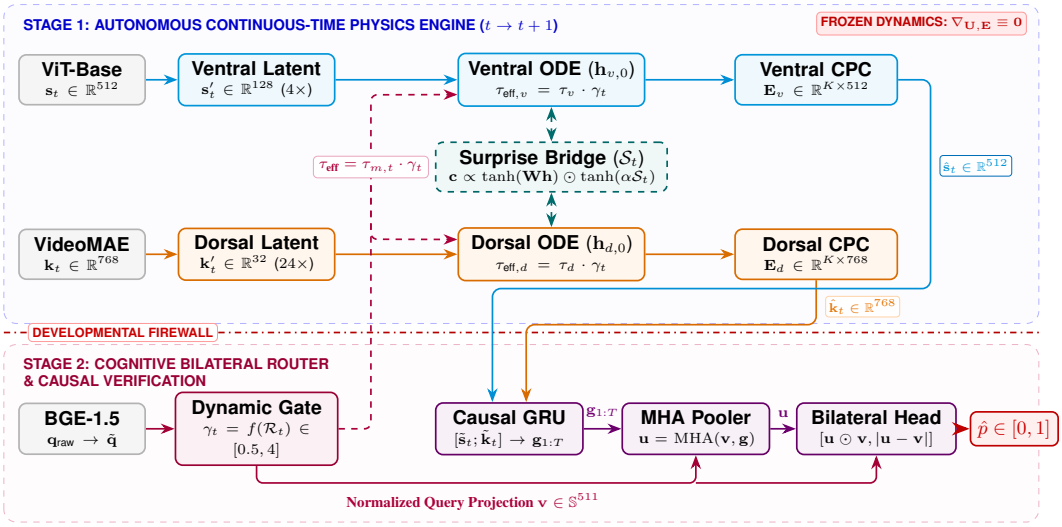


Figure 1: **Two-stage developmental pipeline and continuous causal verification. Stage 1 (top):** Upstream representations project to decoupled ventral (128d) and dorsal (32d) streams, advanced via liquid ODEs with surprise coupling and non-gradient CPC memory ($\nabla_{\mathbf{U},\mathbf{E}} \equiv \mathbf{0}$). **Stage 2 (bottom):** Queries modulate ODE integration rates ($\tau_{\text{eff}} = \tau \cdot \gamma_t$), while trajectories $[\tilde{\mathbf{s}}_t; \tilde{\mathbf{k}}_t]$ are fused via causal GRU, cross-attended against query $\mathbf{v}$, and verified on the bilateral manifold ($[\mathbf{u} \odot \mathbf{v}, |\mathbf{u} - \mathbf{v}|] \to \hat{p}$).

- **Stage 1 (autonomous continuous-time physics engine):** Optimized strictly self-supervised to extrapolate continuous forward spatiotemporal transitions ($t \to t+1$) over decoupled latent

manifolds (128d Ventral $\times$ 32d Dorsal). It learns physical momentum, geometric continuity, and occlusion dynamics entirely independent of language queries, semantic tasks, or extrinsic rewards.

- **Stage 2 (cognitive epistemic probe & bilateral router):** During downstream training, Stage 1 continuous physics dynamics—recurrence weights ($\mathbf{U}_s, \mathbf{U}_k, \mathbf{U}_h, \mathbf{W}_\tau$) and memory attractor banks ($\mathbf{E}_v, \mathbf{E}_d$)—remain strictly frozen ($\nabla_{\mathbf{U},\mathbf{E}}\mathcal{L}_{\text{router}} \equiv \mathbf{0}$), while linear sensory projection heads receive only gentle coordinate calibration (Appendix F.2). Stage 2 acts as a frozen epistemic probe that translates immutable continuous physical trajectories into calibrated symbolic verification verdicts under temporal group permutation optimization (TGPO).

3.1 Sensory Feature Extraction and Dimensional Asymmetry

Given video $\mathcal{V}$ sampled to $T = 16$ frames via temporal jittering ($t_i = i \cdot S + \delta_i$ with stride S and offset $\delta_i \sim \mathcal{U}(-1, 1)$), upstream frozen backbones extract semantic and kinematic features. Semantic vectors $\mathbf{s}_t \in \mathbb{R}^{512}$ (extracted via `ViT-Base/16` (Dosovitskiy et al., 2021; Oquab et al., 2023) with linear projection $\mathbf{W}_{\text{sem_init}}$) and dense kinematic tokens $\mathbf{k}_t \in \mathbb{R}^{768}$ (extracted via `VideoMAE-Base` (Tong et al., 2022)) project into decoupled latent manifolds:

$$\mathbf{s}'_t = \text{LN}\left(\mathbf{W}_s\mathbf{s}_t + \mathbf{b}_s\right) \in \mathbb{R}^{128}, \quad \mathbf{k}'_t = \text{LN}\left(\mathbf{W}_k\mathbf{k}_t + \mathbf{b}_k\right) \in \mathbb{R}^{32} \tag{1}$$

where $\text{LN}(\cdot)$ denotes Layer Normalization, $\mathbf{W}_s \in \mathbb{R}^{128\times512}$ parameterizes the ventral projection, and $\mathbf{W}_k \in \mathbb{R}^{32\times768}$ defines the dorsal bottleneck.

Algebraic constriction vs. semantic leakage. Expressive capacity is allocated to semantics ($d_v = 128$) to index diverse object categories, materials, and static scene context. In contrast, spatial mechanics are constricted to $d_d = 32$. Rigid-body physical interactions inherently occupy low intrinsic dimensionality—tracking 3D relative positions, linear velocities, angular rotations, and contact boundaries requires fewer than 20 active degrees of freedom (Appendix C.1). Upstream encoders such as VideoMAE produce dense 768d tokens that heavily entangle these trajectory statistics with high-entropy background textures. Allocating excessive dimensionality to kinematics (e.g., $d_d \geq 128$) triggers semantic leakage: because static visual textures exhibit strong temporal auto-correlation, gradient descent prioritizes memorizing visual appearance shortcuts over tracking subtle motion derivatives. Constricting kinematics strictly to $d_d = 32$ imposes an algebraic capacity barrier: the subspace possesses sufficient bandwidth to track continuous physical trajectories and contact events, but lacks the metric volume to store high-rank appearance textures. This structurally shields trajectory derivatives from appearance dominance, isolating invariant Newtonian motion. Learnable resting vectors $\mathbf{h}_{v,0} \in \mathbb{R}^{128}$ and $\mathbf{h}_{d,0} \in \mathbb{R}^{32}$ anchor continuous integration.

3.2 Stage 1: Autonomous Continuous-Time Physics Engine

Stage 1 establishes an unsupervised latent physics simulator trained unconditionally to predict continuous forward transitions ($t \to t + 1$) without language tokens or extrinsic reward.

Surprise-gated cross-stream coupling. Let $\mathbf{x}'_{v,t} \triangleq \mathbf{s}'_t \in \mathbb{R}^{128}$ and $\mathbf{x}'_{d,t} \triangleq \mathbf{k}'_t \in \mathbb{R}^{32}$ denote stream inputs. Bilateral communication is modulated by Euclidean surprise $\mathcal{S}_{m,t} = \|\mathbf{x}'_{m,t} - \mathbf{h}_{m,t-1}\|_2 + 10^{-6}$ for $m \in \{v, d\}$. The cross-stream coupling vector $\mathbf{c}_{\bar{m}\to m,t} \in \mathbb{R}^{d_m}$ from opposing stream $\bar{m}$ ($\bar{v} \triangleq d, \bar{d} \triangleq v$) is formulated as:

$$\mathbf{c}_{\bar{m}\to m,t} = \tanh\left(\text{MLP}_{\bar{m}\to m}(\mathbf{h}_{\bar{m},t-1})\right) \odot \tanh\left(\alpha_{\bar{m}}\mathcal{S}_{\bar{m},t}\right) \tag{2}$$

where $\text{MLP}_{\bar{m}\to m} : \mathbb{R}^{d_{\bar{m}}} \to \mathbb{R}^{d_m}$, $\odot$ denotes the Hadamard product, and $\alpha_v = 0.40, \alpha_d = 0.05$ set sensitivities. Equation 2 isolates streams during laminar transport ($\mathcal{S}_t \approx 0$) and opens communication channels during mechanical impacts ($\mathcal{S}_t \gg 0$).

Autonomous liquid ODE integration. In Stage 1, in the absence of task queries ($\mathcal{R}_t \equiv 0$), base continuous timescales $\tau_{m,t} > 0$ adapt purely to sensory observations: $\tau_{m,t} = \text{Softplus}(\text{MLP}_{\tau,m}[\mathbf{x}'_{m,t}; \mathbf{h}_{m,t-1}; \mathbf{c}_{\bar{m}\to m,t}]) + 2.0$. Operating at baseline multiplier $\gamma_0 \equiv 4.0$, effective timescale is $\tau_{\text{eff},m} = \tau_{m,t} \cdot 4.0$. Given elapsed interval $\Delta t \geq 0$, states $\mathbf{h}_{m,t} \in \mathbb{R}^{d_m}$ advance via analytical exponential Euler stepping under Zero-Order Hold (Appendix B.1):

$$\mathbf{I}_{m,t} = \tanh\left(\mathbf{U}_m\mathbf{x}'_{m,t} + \mathbf{U}_{h,m}\mathbf{h}_{m,t-1} + \mathbf{c}_{\bar{m}\to m,t}\right) \in \mathbb{R}^{d_m}, \tag{3}$$

$$\mathbf{h}_{m,t} = \mathbf{h}_{m,t-1} \odot \exp\left(-\Delta t/\tau_{\text{eff},m}\right) + \mathbf{I}_{m,t} \odot \left(\mathbf{1} - \exp\left(-\Delta t/\tau_{\text{eff},m}\right)\right) \tag{4}$$

where $\mathbf{U}_m, \mathbf{U}_{h,m} \in \mathbb{R}^{d_m \times d_m}$ are learnable weights, and $\mathbf{I}_{m,t} \in [-1,1]^{d_m}$. Theorem 1 (Appendix B.2) proves that this update is an affine convex combination, guaranteeing global Lyapunov contractive stability ($\|\mathbf{h}_{m,t}\|_\infty \leq 1.0$) across arbitrary rollouts.

Continuous plastic codebooks (CPC) as attractor memory. To bound continuous trajectories against numerical drift without Straight-Through Estimator (STE) gradient pathologies, hidden states are soft-quantized over dual non-gradient attractor banks $\mathbf{E}_v \in \mathbb{R}^{K\times 512}$ and $\mathbf{E}_d \in \mathbb{R}^{K\times 768}$ ($K = 2048$ centroids):

$$\mathbf{Q}_{m,t} = \mathrm{LN}\left(\mathbf{W}_{\text{fuse},m}[\mathbf{x}_{m,t}; \mathbf{h}_{m,t}]\right), \quad \mathbf{A}_{m,t} = \mathrm{softmax}\left(\frac{\mathbf{Q}_{m,t}\mathbf{E}_m^\top}{\tau_{\text{temp}}\sqrt{D_m}}\right), \quad \hat{\mathbf{x}}_{m,t} = \mathbf{A}_{m,t}\mathbf{E}_m \tag{5}$$

where $\mathbf{x}_{m,t}$ represents raw input ($\mathbf{s}_t \in \mathbb{R}^{512}$ or $\mathbf{k}_t \in \mathbb{R}^{768}$) of native dimension D_m, $\mathbf{W}_{\text{fuse},m} \in \mathbb{R}^{D_m \times (D_m + d_m)}$, and τ_{temp} anneals from $1.50 \to 0.15$. Centroid updates follow continuous Hebbian target relaxation (Hebb, 1949; Miconi et al., 2018) outside the backward pass ($\nabla_{\mathbf{E}_m}\mathcal{L} \equiv \mathbf{0}$) with time constant $\tau_{\text{plastic}} = 10.0$ frames:

$$\Delta\mathbf{E}_{m,k} = \frac{\bar{A}_{m,k}}{\tau_{\text{plastic}}}\left(\frac{\sum_{b=1}^{B} A_{m,b,k}\mathbf{x}_{m,b,t+1}}{\sum_{b=1}^{B} A_{m,b,k} + 10^{-8}} - \mathbf{E}_{m,k}\right) \in \mathbb{R}^{D_m} \tag{6}$$

where $\bar{A}_{m,k} = \frac{1}{B}\sum_{b=1}^{B} A_{m,b,k}$. Under exact analytical stepping ($\Delta t = 1.0$), this continuous relaxation maps to an effective discrete momentum of $\beta = \exp(-\Delta t/\tau_{\text{plastic}}) = \exp(-0.10) \approx 0.9048 \approx 0.90$, deriving the standard $90/10$ EMA codebook update directly from continuous-time dynamics. Crucially, $\tau_{\text{plastic}} = 10.0$ establishes an order-of-magnitude timescale separation between fast sensory state integration ($\tau_{\text{eff}} \in [1.0, 4.0]$) and slow structural memory consolidation, preventing centroid drift from destabilizing ODE hidden states. Inactive centroids ($u_k < 0.05 \cdot \frac{1}{K}$) tracked via EMA ($\beta_{\text{ema}} = 0.995$) are stochastically revived (Zeghidour et al., 2021; Dhariwal et al., 2020) (Appendix C.2), preserving 100.0% codebook utilization throughout training without gradient noise.

Unsupervised forward predictive objective ($\mathcal{L}_{\text{Stage1}}$). Stage 1 parameters are trained unconditionally to extrapolate future sensory observations ($t \to t+1$): $\mathcal{L}_{\text{Stage1}} = \mathcal{L}_{\text{physics}} + 0.25\mathcal{L}_{\text{commit}} + 0.15\mathcal{L}_{\text{entropy}}$, where:

$$\begin{aligned}
\mathcal{L}_{\text{physics}} &= \frac{1}{T-1}\sum_{t=1}^{T-1}\sum_{m\in\{v,d\}}\left(1 - \frac{\hat{\mathbf{x}}_{m,t}^\top \mathbf{x}_{m,t+1}}{\|\hat{\mathbf{x}}_{m,t}\|_2\|\mathbf{x}_{m,t+1}\|_2}\right)\cdot M_{t+1}, \\
\mathcal{L}_{\text{commit}} &= \frac{1}{T-1}\sum_{t=1}^{T-1}\sum_{m\in\{v,d\}}\|\mathbf{Q}_{m,t} - \mathrm{sg}[\hat{\mathbf{x}}_{m,t}]\|_2^2 \cdot M_{t+1}, \\
\mathcal{L}_{\text{entropy}} &= \sum_{m\in\{v,d\}}\left[-\frac{1}{B}\sum_{b=1}^{B}\sum_{k=1}^{K} A_{m,b,k}\ln A_{m,b,k} + \sum_{k=1}^{K}\bar{A}_{m,k}\ln\bar{A}_{m,k}\right]
\end{aligned} \tag{7}$$

where $M_{t+1} \in \{0,1\}$ is a valid temporal frame mask, and $\mathrm{sg}[\cdot]$ is the stop-gradient operator.

Diagnostic decodability probing. Before downstream supervision, frozen Stage 1 representations ($\nabla_{\mathbf{U},\mathbf{E}} \equiv \mathbf{0}$) are audited in isolation via two diagnostic probes (Table 1, Appendix E.1): (1) **direct next-state tracking**, evaluating forward kinematic prediction ($\mathbf{k}_{t+1}$) over the 32d dorsal ODE (0.9667 cosine); and (2) **the liquid binding decoder**, which reconstructs dense visual features via cross-attention to verify spatial retention ($\rho_{\text{feature}} = 0.9800$) and first-order velocity derivative alignment ($\rho_{\text{velocity}} = 0.7454$ on $\Delta\hat{\mathbf{z}}_t$ vs. $\Delta\mathbf{k}_t$). These probes confirm that the self-supervised manifold captures continuous Newtonian momentum rather than static appearance shortcuts.

3.3 Stage 2: Syntax-Aware Interrogatives and Causal Bilateral Router

Stage 2 introduces language queries, couples top-down clock modulation, and optimizes causal verification.

Syntax-aware interrogatives and asymmetric hard negatives. We synthesize orthogonal query triplets (semantic $\mathbf{q}_{\text{sem}}$, kinematic $\mathbf{q}_{\text{kin}}$, and binary verification $\mathbf{q}_{\text{bin}}$) per video using unquantized `Qwen2.5-7B-Instruct` (Qwen Team, 2024) (Appendix D.2), embedded via `bge-base-en-v1.5` (Xiao et al., 2023) using sequence-level masked mean-pooling: $\tilde{\mathbf{q}} = \mathrm{LN}(\mathbf{W}_{\text{text_proj}}\mathbf{q}_{\text{raw}}) \in \mathbb{R}^{512}$ with

$\mathbf{W}_{\text{text_proj}} \in \mathbb{R}^{512\times768}$. To enforce physical non-commutativity ($A \to B \neq B \to A$) beyond heuristic classification (Wei et al., 2018), 20 asymmetric physical trigger verbs ($\mathcal{T}_{\text{asym}}$, Appendix D) pair with time-reversed sequences ($\text{Flip}(\mathbf{x}_{1:16})$) during training ($p_{\text{neg}} = 0.50$).

Top-down query-modulated integration ($\tau_{\text{eff}} = \tau_{m,t} \cdot \gamma_t$). Task queries modulate the ODE integration clock through a relevance-gated timescale:

$$\gamma_t = \text{clamp}\big(4.0 - 3.5\,\mathcal{R}_t,\ 0.5,\ 4.0\big), \quad \tau_{\text{eff},m} = \tau_{m,t} \cdot \gamma_t, \tag{8}$$

$$\mathcal{R}_t = \frac{1}{2}\left[\sigma\left(\frac{\mathbf{s}_t^\top(\mathbf{W}_{q\to s}\mathbf{q}_{\text{raw}})}{\|\mathbf{s}_t\|_2\,\|\mathbf{W}_{q\to s}\mathbf{q}_{\text{raw}}\|_2}\right) + \sigma\left(\frac{\mathbf{k}_t^\top(\mathbf{W}_{q\to k}\mathbf{q}_{\text{raw}})}{\|\mathbf{k}_t\|_2\,\|\mathbf{W}_{q\to k}\mathbf{q}_{\text{raw}}\|_2}\right)\right] \tag{9}$$

where $\mathbf{W}_{q\to s} \in \mathbb{R}^{512\times768}$ and $\mathbf{W}_{q\to k} \in \mathbb{R}^{768\times768}$. Intuitively, physically grounded queries (high relevance) accelerate integration, while ungrounded distractors dilate the clock, turning the liquid ODE into a heavily damped low-pass filter that suppresses false-positive activations. Dynamic dilation provides intrinsic habituation, dissipating ungrounded shocks ($\mathcal{R}_t \to 0$) to prevent rigid-clock latching (Figure E.2, Appendix E.5).

Sequential causal fusion and bilateral interaction manifold. Stage 1 latents project into router space via $\tilde{\mathbf{s}}_t = \text{LN}(\mathbf{W}_{\text{sem_align}}\hat{\mathbf{x}}_{v,t}) \in \mathbb{R}^{512}$ and $\tilde{\mathbf{k}}_t = \text{LN}(\mathbf{W}_{\text{kin_align}}\hat{\mathbf{x}}_{d,t}) \in \mathbb{R}^{512}$, with prefix frames stochastically zeroed ($\tilde{\mathbf{s}}_{1:3}, \tilde{\mathbf{k}}_{1:3} \leftarrow \mathbf{0}$, $p = 0.40$) to suppress static shortcuts. Trajectories are sequentially fused and cross-attended against the normalized query $\mathbf{v} = \tilde{\mathbf{q}}/\|\tilde{\mathbf{q}}\|_2$:

$$\mathbf{g}_t = \text{GRU}\left(\left[\tilde{\mathbf{s}}_t;\, \tilde{\mathbf{k}}_t\right],\, \mathbf{g}_{t-1}\right) \in \mathbb{R}^{512}, \quad \mathbf{v}^* = \text{MHA}\left(\mathbf{Q} = \text{LN}(\tilde{\mathbf{q}}),\, \mathbf{K} = \mathbf{g}_{1:T},\, \mathbf{V} = \mathbf{g}_{1:T}\right) \in \mathbb{R}^{512}, \tag{10}$$

$$\mathbf{f}_{\text{bin}} = \left[\mathbf{u} \odot \mathbf{v};\, |\mathbf{u} - \mathbf{v}|\right] \in \mathbb{R}^{1024}, \quad z = \mathbf{W}_{\text{bin},2}\ \text{GELU}\left(\mathbf{W}_{\text{bin},1}\mathbf{f}_{\text{bin}} + \mathbf{b}_1\right) + b_2, \quad \hat{p} = \sigma(z) \tag{11}$$

where $\mathbf{u} = \mathbf{v}^*/\|\mathbf{v}^*\|_2 \in \mathbb{S}^{511}$. Hadamard agreement ($\mathbf{u} \odot \mathbf{v}$) measures directional collinearity, while element-wise metric discrepancy ($|\mathbf{u} - \mathbf{v}|$) isolates trajectory deviations to produce scalar verification belief $\hat{p} \in [0, 1]$ (Appendix C.3).

Tripartite causal objective and auxiliary regularization ($\mathcal{L}_{\text{Stage2}}$). Stage 2 optimizes a composite objective over query branches ($\mathbf{q}_{\text{bin}}, \mathbf{q}_{\text{sem}}, \mathbf{q}_{\text{kin}}$), combining a **tripartite causal core** with **auxiliary geometric regularizers** (Algorithm 2):

$$\mathcal{L}_{\text{Stage2}} = \underbrace{\mathcal{L}_{\text{BCE}} + \lambda_{\text{tgpo}}\mathcal{L}_{\text{TGPO}} + \mathcal{L}_{\text{align}}}_{\text{tripartite causal core}} + \underbrace{\lambda_{\text{route}}\mathcal{L}_{\text{route}} + \lambda_{\text{ortho}}\mathcal{L}_{\text{ortho}} + \lambda_{\text{sink}}\mathcal{L}_{\text{sink}}}_{\text{auxiliary geometric guardrails}} \tag{12}$$

The tripartite core enforces verification and cross-modal alignment:

$$\mathcal{L}_{\text{BCE}} = \text{BCE}(\hat{p}_{\text{ord}},\, y), \tag{13}$$

$$\mathcal{L}_{\text{TGPO}} = \text{BCE}(\hat{p}_{\text{shuffled}},\, 0.50) \equiv D_{\text{KL}}\Big(\text{Bernoulli}(0.50) \parallel \text{Bernoulli}(\hat{p}_{\text{shuffled}})\Big), \tag{14}$$

$$\mathcal{L}_{\text{align}} = \frac{1}{2}\left[\text{CE}\left(\frac{\mathbf{S}_{\text{act}}\mathbf{S}_{\text{txt}}^\top}{\tau_{\text{align}}},\, \mathbf{I}\right) + \text{CE}\left(\frac{\mathbf{S}_{\text{txt}}\mathbf{S}_{\text{act}}^\top}{\tau_{\text{align}}},\, \mathbf{I}\right)\right], \tag{15}$$

TGPO pushes temporally scrambled sequences toward maximum entropy ($z^* = 0$, Proposition 1). Auxiliary terms encourage stream orthogonality ($\mathcal{L}_{\text{ortho}}$), reduce cross-stream leakage ($\mathcal{L}_{\text{route}}$), and prevent early-frame attention collapse ($\mathcal{L}_{\text{sink}}$):

$$\mathcal{L}_{\text{ortho}} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\tilde{\mathbf{s}}_t^\top\tilde{\mathbf{k}}_t}{\|\tilde{\mathbf{s}}_t\|_2\|\tilde{\mathbf{k}}_t\|_2}\right)^2, \tag{16}$$

$$\mathcal{L}_{\text{route}} = \frac{1}{T}\sum_{t=1}^{T}\left[\text{ReLU}\left(\frac{\tilde{\mathbf{k}}_t^\top\tilde{\mathbf{q}}_{\text{sem}}}{\|\tilde{\mathbf{k}}_t\|_2\|\tilde{\mathbf{q}}_{\text{sem}}\|_2}\right) + \text{ReLU}\left(\frac{\tilde{\mathbf{s}}_t^\top\tilde{\mathbf{q}}_{\text{kin}}}{\|\tilde{\mathbf{s}}_t\|_2\|\tilde{\mathbf{q}}_{\text{kin}}\|_2}\right)\right], \tag{17}$$

$$\mathcal{L}_{\text{sink}} = \frac{1}{B}\sum_{b=1}^{B}\text{ReLU}\left(\alpha_{b,t=0}^{\text{sem}} - 0.15\right). \tag{18}$$

Defaults: $\lambda_{\text{tgpo}} = 0.50$, $\lambda_{\text{route}} = 0.10$, $\lambda_{\text{ortho}} = 0.01$, and $\lambda_{\text{sink}} = 0.20$ ($\tau_{\text{align}} = 0.05$). Here, $\mathbf{S}_{\text{act}} \triangleq [\mathbf{u}_1, \ldots, \mathbf{u}_B]^\top \in \mathbb{R}^{B\times512}$ and $\mathbf{S}_{\text{txt}} \triangleq [\mathbf{v}_1, \ldots, \mathbf{v}_B]^\top \in \mathbb{R}^{B\times512}$.

Epistemic calibration and the tri-state attractor. The $[0.40, 0.60]$ deadzone emerges directly as an *intrinsic epistemic attractor*: TGPO provides a restoring force toward the maximum-entropy point

$z^* = 0 \implies \hat{p} = 0.50$ (Proposition 1), while ODE timescale dilation ($\gamma_t \to 4.0$) provides viscous damping. Mapping minimal Bayesian evidence bounds ($\text{Odds} \in [1/1.5, 1.5]$) into logit space yields the symmetric noise radius $|z| \leq \ln(1.5) \iff \hat{p} \in [0.40, 0.60]$ (Appendix B.4). This establishes a calibrated tri-state manifold: **(1) Active Contradictions** ($\hat{p} < 0.40$, penalizing counterfactuals via $|\mathbf{u} - \mathbf{v}|$), **(2) Epistemic Abstention** ($[0.40, 0.60]$), and **(3) Verified Causality** ($\hat{p} > 0.60$, confirming forward alignment $\mathbf{u} \odot \mathbf{v}$; Appendix B).

4 Experiments and Empirical Validation

We evaluate CLWM across five empirical pillars on **Something-Something-v2 (SSv2)** (Goyal et al., 2017) over $N = 24{,}673$ held-out causal verification samples (12,333 positive / 12,340 negative pairs; split curation and provenance detailed in Appendix D.4) within an ultra-compact trainable footprint of ≈ 4.72M parameters (Appendix Table F.1).

4.1 Pillars 1 & 2: Forward Physics Prediction and Velocity Alignment

Table 1: **Stage 1 physics engine: alignment and decodability probes.** Directional feature tracking ($t \to t+1$) and dynamic velocity alignment on frozen latents ($N = 24{,}673$).

Stage 1 physics engine: alignment and decodability probes		
Evaluation pillar	**Diagnostic metric**	**Score (↑)**
Pillar 1: Dynamics	Dorsal kinematic tracking (32d, $t \to t+1$)	**0.9667** *(Cosine)*
Pillar 1: Dynamics	Ventral semantic tracking (128d, $t \to t+1$)	**0.6017** *(Cosine)*
Pillar 2: Latent probe	Spatial feature alignment ($\mathbf{z}_{\text{MAE}}$ decodability)	**0.9800** *(Cosine)*
Pillar 2: Latent probe	Dynamic velocity alignment ($\cos(\Delta\hat{\mathbf{z}}, \Delta\mathbf{z})$ derivative)	**0.7454** *(Cosine)*

Pillar 1: Unsupervised dynamics (0.9667 kinematic vs. 0.6017 semantic). Under forward predictive extrapolation ($t \to t+1$), the 32d dorsal stream achieves directional trajectory tracking of **0.9667** cosine (Table 1). As derived in Section 3.1, the rank-32 bottleneck shields spatial velocity from high-rank appearance. Lower semantic tracking (0.6017) matches physical principles: spatial coordinates shift continuously, whereas categorical identities remain temporally invariant.

Pillar 2: Dynamic velocity alignment ($\rho_{\text{velocity}} = 0.7454$). Evaluating frozen Stage 1 representations via the auxiliary decoder probe (Appendix E.1) confirms near-lossless spatial retention ($\rho_{\text{feature}} = \mathbf{0.9800}$). Evaluating directional cosine similarity on first-order temporal derivatives ($\Delta\hat{\mathbf{z}}_t$ vs. $\Delta\mathbf{k}_t$) yields $\boldsymbol{\rho}_{\text{velocity}} = \mathbf{0.7454}$, verifying that the Liquid ODE captures Newtonian momentum rather than interpolating static backgrounds.

4.2 Pillars 3 & 4: Causal Verification and Double Dissociation Knock-outs

Table 2: **Stage 2 causal router: verification and functional Knock-outs.** Evaluated on the canonical asymmetric physical cohort of Something-Something-v2 ($N = 8{,}453$). Isolating irreversible asymmetric transitions (Appendix D) is physically required to audit arrow-of-time causality ($A \to B \neq B \to A$); on arbitrary symmetric actions (e.g., *"holding an object"*), time reversal presents no physical contradiction. Primary discriminability is quantified via threshold-independent ROC-AUC. Operational classification metrics evaluate affirmative hypothesis verification against $\theta_{\text{verify}} = 0.60$ ($z \geq \ln(1.5)$). Diagnostic sweep across $[0.40, 0.60]$ is detailed in Appendix Table E.2.

Stage 2 causal router: verification and functional Knock-outs				
Condition	**ROC-AUC**	**Acc (@0.60)**	**TPR**	**TNR**
Pillars 3 & 4: Full pipeline	**0.9592**	**88.81%**	**86.54%**	**91.08%**
1. Ventral stream KO ($\mathbf{s}_t \to \mathbf{0}$)	0.4984	50.15%	0.23%	99.72%
2. Dorsal stream KO ($\mathbf{k}_t \to \mathbf{0}$)	0.5941	50.61%	3.22%	97.65%
3. Arrow-of-time KO (scrambled)	0.5546	50.19%	0.35%	99.68%

Pillar 3: Multimodal verification via frozen epistemic probing. Evaluating frozen Stage 1 dynamics ($\nabla_{\mathbf{U},\mathbf{E}} \equiv \mathbf{0}$), the router achieves **0.9331 ROC-AUC** (90.58% specificity at $\theta_{\text{contra}} = 0.40$, $N = 24{,}673$). On the asymmetric cohort ($N = 8{,}453$, Table 2), discriminability reaches **0.9592 ROC-AUC** (88.81% accuracy, 91.08% specificity at $\theta_{\text{verify}} = 0.60$; breakdown in Appendix Table D.1). The bilateral head projects states onto the tri-state manifold (Appendix C.3): metric

discrepancy $|\mathbf{u} - \mathbf{v}|$ drives counterfactual reversals into active contradiction ($\hat{p} < 0.40$), while collinearity $\mathbf{u} \odot \mathbf{v}$ verifies forward causality ($\hat{p} > 0.60$).

Pillar 4: Double dissociation Knock-outs and manifold purity. Ablating semantics ($\mathbf{s}_t \to \mathbf{0}$), kinematics ($\mathbf{k}_t \to \mathbf{0}$), or chronological order collapses discriminability to chance (ROC-AUC ≤ 0.5941, Acc $\approx 50\%$, Table 2). Specifically, the Ventral KO collapse (0.4984 ROC-AUC) confirms kinematic purity, proving the rank-32 dorsal bottleneck eliminates visual appearance leakage to isolate pure Newtonian motion. Furthermore, corrupted inputs settle into the $[0.40, 0.60]$ deadzone ($> 99\%$ mass; Appendix Table E.2), suppressing affirmative assertions at $\theta_{\text{verify}} = 0.60$ to near zero (TPR $\leq 3.22\%$, TNR $\geq 97.65\%$) to enforce safe epistemic abstention.

4.3 Pillar 5: Continuous-Time Epistemic Margin (CTEM) Protocol

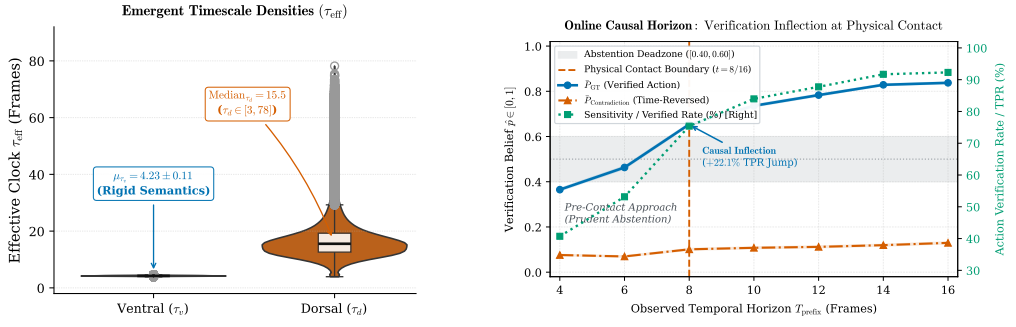
Bypassing competitive softmax normalization, CLWM evaluates $C = 174$ candidates via independent sigmoidal beliefs $\hat{p}_j = \sigma(z_j)$ in a single zero-copy parallel broadcast pass (Algorithm 3; Appendix C). To eliminate lexical shortcuts, all candidate queries share identical scene nouns (Appendix D.1), forcing discrimination to depend exclusively on temporal and kinematic mechanics. Across $N = 24{,}673$ sequences, ground-truth belief concentrates at $\bar{P}_{\text{GT}} = \mathbf{0.7876}$ while distractors anchor at maximum entropy ($\bar{P}_{\text{Dist}} = \mathbf{0.4978} \approx 0.50$), establishing an unconstrained margin of $\mathbf{\Delta}_{\text{CTEM}} = \mathbf{+0.2898}$ (Appendix Figure E.4). Clamping integration ($\gamma_t \equiv 1.0$) permits distractor drift to 0.5420, collapsing this margin by 26.9% to +0.2119 (Appendix Table E.1).

4.4 Dynamical Invariance and State Telemetry

Table 3: **Zero-shot dynamical stress tests (frozen weights).** Continuous Δt scaling and prefix sweeps on canonical $N = 8{,}453$ samples. Primary discriminability is evaluated via threshold-independent ROC-AUC, alongside operational classification at $\theta_{\text{contra}} = 0.40$.

Evaluation dimension	Configuration	ROC-AUC (↑)	Acc (@0.40)	TNR (spec.)	Retention / state
Temporal elasticity	$T = 16,\ \Delta t = 1.00$	**0.9597**	**89.65%**	**86.71%**	100.0% (Base)
	$T = 12,\ \Delta t = 1.33$	0.9577	89.14%	85.65%	99.79% AUC
	$T = 8,\ \Delta t = 2.00$	0.9535	88.57%	85.22%	**99.35% AUC**
	$T = 4,\ \Delta t = 4.00$	0.9072	83.30%	77.24%	94.53% AUC
Causal horizon	$T_{\text{pref}} = 4/16$ (25%)	0.8296	67.24%	**94.00%**	Pre-contact ($\bar{P}_{\text{GT}} = 0.3621$)
	$T_{\text{pref}} = 8/16$ (50%)	**0.9267**	**82.76%**	90.31%	Contact inflection ($\bar{P}_{\text{GT}} = 0.6523$)
	$T_{\text{pref}} = 14/16$ (87.5%)	**0.9589**	**89.31%**	87.11%	Saturated ($\bar{P}_{\text{GT}} = 0.8276$)

Continuous elasticity and early horizons. Closed-form exponential stepping enables zero-retraining evaluation across continuous intervals $\Delta t \geq 0$ (Table 3). Dropping 50% of frames ($T = 8, \Delta t = 2.00$) preserves **99.35% AUC** (0.9597 → 0.9535) at 85.22% specificity. In prefix sweeps, belief safely withholds during approach ($T_{\text{pref}} \leq 6, \bar{P}_{\text{GT}} = 0.4626$) before inflecting sharply at physical contact ($T_{\text{pref}} = 8$), jumping +22.1% in sensitivity (0.9267 AUC) and converging to a +0.8386 counterfactual margin (Appendix Table D.1).



(a) Emergent timescale densities (τ_{eff})

(b) Online causal horizon inflection at contact

Figure 2: **Continuous dynamical telemetry and causal verification commitment. (a)** Effective integration timescales (τ_{eff}), illustrating rigid semantic stability ($\tau_v \approx 4.18 \pm 0.08$) versus multi-scale kinematic elasticity ($\tau_d \in [4, 80]$ frames; median 16.0). **(b)** Online temporal prefix sweep ($N = 8{,}453$): verification belief $\hat{p}$ safely withholds within the $[0.40, 0.60]$ epistemic deadzone during object approach ($t \leq 6/16$), commits sharply at physical contact ($t = 8/16$), and suppresses time-reversed counterfactuals into active contradiction ($\hat{p} < 0.40$; t-SNE in Appendix Figure E.3).

Emergent timescale densities and peri-contact telemetry. Physical decoupling manifests directly in timescale differentiation (Figure 2a): ventral integration remains rigid ($\tau_v \approx 4.18 \pm 0.08$ frames) to anchor identity, while dorsal integration spans an elastic spectrum ($\tau_d \in [4, 80]$ frames; median 16.0). In event-locked telemetry ($N = 8{,}453$; Appendix E.4), dorsal memory dilates across the contact boundary ($18.34 \rightarrow 23.76 \rightarrow 28.52$ frames) as kinematic surprise spikes ($\mathcal{S}_d = 7.73 \pm 0.32$), confirming that verification commitment (Figure 2b) is driven by spatial state transitions rather than semantic drift. Furthermore, query modulation (γ_t) acts as a kinetic circuit breaker: while rigid clocks latch into false-positive saturation upon sudden impact ($\hat{p} = 0.9165$), dynamic dilation damps transient shocks safely into the $[0.40, 0.60]$ deadzone ($\hat{p} = 0.4304$; Figure E.2, Appendix E.5).

5 Discussion and Limitations

Physical verification over structural invariants. Prevailing world models allocate $> 95\%$ of capacity to rendering surface textures, frequently hallucinating physical violations. CLWM establishes that physical common sense requires *continuous-time causal hypothesis verification rather than generative rendering*, replacing iterative pixel synthesis with lightweight parallel verification. Sensitivity sweeps (Appendix Table E.3) confirm high discriminability even when auxiliary losses are ablated ($\lambda_{\text{route}} = 0, \lambda_{\text{ortho}} = 0$). Hence, physical decoupling is guaranteed by architectural geometry—the developmental firewall ($\nabla_{\mathbf{U},\mathbf{E}} \equiv \mathbf{0}$), rank-32 bottleneck, Lyapunov contraction, and non-gradient Hebbian memory—rather than delicate loss weighting.

Emergent attractors and deployment safety. Unlike closed-simplex models that yield overconfident out-of-distribution predictions, coupling liquid ODEs with TGPO induces an *intrinsic epistemic attractor well* ($z^* = 0$). Under corrupted or scrambled inputs, timescale dilation ($\gamma_t \rightarrow 4.0$) acts as viscous friction, dissipating kinetic updates into the $[0.40, 0.60]$ resting deadzone ($> 99\%$ empirical mass; Appendix Table E.2). This provides an autonomous circuit breaker, reliably withholding affirmative actions under ungrounded or physically counterfactual observations.

Limitations. CLWM currently targets primary agent-object manipulation dyads over passive video without explicit 3D mesh reconstruction or tactile feedback. Scaling beyond dyadic interactions to dense multi-body collisions and fluid continuum mechanics remains an open frontier.eractions to high-density multi-body collisions and complex fluid topologies remains an active frontier.

6 Conclusion

The **Cognitive Liquid World Model (CLWM)** reformulates physical world modeling as continuous-time causal hypothesis verification over frozen, decoupled latent manifolds. By coupling an autonomous liquid ODE foundation (128d Ventral $\times$ 32d Dorsal) with a downstream epistemic probe regularized by TGPO, CLWM replaces generative rendering with inductive structural invariants. The architecture resolves kinematic starvation (0.9667 trajectory tracking), attains robust causal discriminability (0.9331 ROC-AUC, 90.58% specificity), demonstrates continuous temporal elasticity (99.35% AUC retention under 50% frame loss), and establishes an unconstrained $+0.2898$ CTEM margin across 174 concurrent hypotheses while relaxing ungrounded dynamics into an emergent $[0.40, 0.60]$ resting attractor.

Looking forward, key future extensions include: (1) *continuous topological deformations*, extending the rank-32 rigid-body bottleneck (Appendix C.1) to non-rigid continuum mechanics such as high-viscosity fluids and cloth; (2) *multi-body relational topologies*, scaling ODE manifolds to graph-structured collision dynamics; and (3) *closed-loop embodied control*, embedding scalar verification belief ($\hat{p} \in [0, 1]$) directly as an online plausibility gate within interactive robotic policies.

Reproducibility statement. All analytical ODE stepping formulations (Eq. 4), Lyapunov contraction proofs (Theorem 1), and TGPO stationary attractor derivations (Proposition 1) are detailed in Appendices B and B. Algorithms 1, 2, and 3 formalize the two-stage developmental pipeline and zero-copy parallel evaluation (Appendix C). Layer dimensions, architectural specifications, and optimization hyperparameters are provided in Appendices F and F.2.

Author Disclosure of LLM and AI Usage

In compliance with the ICLR policy on LLM usage, we disclose the following assistance in the preparation of this manuscript:

- **Dataset Synthesis:** We utilized `Qwen2.5-7B-Instruct` offline as a data engine to construct orthogonal query triplets (semantic, kinematic, and binary verification) from video descriptions (Appendix D.2).
- **Literature Discovery:** We employed conversational AI (Gemini, Google) to assist in identifying relevant literature and related work across continuous-time dynamics and world models. All retrieved citations were subsequently read, verified, and vetted by us.
- **Manuscript Refinement & Layout:** We used conversational AI for prose compression, language copyediting to fit conference page limits, and LaTeX float optimization.

All architectural designs, formal mathematical derivations and proofs, implementations, and empirical analyses were conceived, verified, and executed directly by us, and we take full scientific responsibility for the contents of this work.

References

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 6306–6315, 2017.

Adrien Bardes, Quentin Duval, Alexander Kirillov, Alireza Fathi, Christoph Feichtenhofer, Yann LeCun, Xiaolong Wang, and Nicolas Ballas. V-jepa: Video joint-embedding predictive architecture. *arXiv preprint arXiv:2404.08471*, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1933–1941, 2016.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6202–6211, 2019.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Carl Doersch, Lu Jiang, et al. VideoPoet: A large language model for zero-shot video generation. In *International Conference on Machine Learning (ICML)*, PMLR volume 235, pages 24869–24888, 2024.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. In *International Conference on Learning Representations (ICLR)*, 2023.

Donald Olding Hebb. *The Organization of Behavior: A Neuropsychological Theory*. John Wiley & Sons, New York, 1949.

Donglai Wei, Joseph J. Lim, Andrew Zisserman, and William T. Freeman. Learning and using the arrow of time. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8052–8060, 2018.

Eloi Alonso, Danijar Hafner, Jan Robine, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 65866–65889, 2024.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuhui Shi, Edward Hughes, Matthew Lai, Volodymyr Mnih, Adrià Puigdomènech Badia, et al. Genie: Generative interactive environments. In *International Conference on Machine Learning (ICML)*, PMLR volume 235, pages 4467–4494, 2024.

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, pages 568–576, 2014.

Laura Sevilla-Lara, Mohammadreza Zolfaghari, Yashar Akhoundi, and Thomas Brox. Only time can tell: Discovering temporal data for temporal modeling. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 535–544, 2021.

Leslie G. Ungerleider and Mortimer Mishkin. Two cortical visual systems. In D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield, editors, *Analysis of Visual Behavior*, pages 549–586. MIT Press, Cambridge, MA, 1982.

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. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15619–15629, 2023.

Mathias Lechner, Ramin Hasani, Alexander Amini, Thomas A. Henzinger, Daniela Rus, and Radu Grosu. Neural circuit policies enabling auditable autonomy. *Nature Machine Intelligence*, 2(10): 642–652, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2023.

Melvyn A. Goodale and A. David Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, 1992.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, 2021.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller, et al. The "something something" video database for learning and evaluating visual common sense. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5842–5850, 2017.

Ramin Hasani, Mathias Lechner, Alexander Amini, Daniela Rus, and Radu Grosu. Liquid time-constant networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7657–7666, 2021.

Ramin Hasani, Mathias Lechner, Alexander Amini, Lucas Liebenwein, Aaron Ray, Max Tschaikowski, Gerald Teschl, and Daniela Rus. Closed-form continuous-time neural networks. *Nature Machine Intelligence*, 4(11):992–1003, 2022.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pages 6571–6583, 2018.

Sam Buchanan, Andrew Stern, Jing Chen, and João Carreira. Space-time crop & resize: Are we learning temporal features in video models? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 31548–31560, 2022.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general Chinese and English embedding. *arXiv preprint arXiv:2309.07597*, 2023.

Thomas Miconi, Kenneth O. Stanley, and Jeff Clune. Differentiable plasticity: training plastic neural networks with backpropagation. In *International Conference on Machine Learning (ICML)*, pages 3559–3568, 2018.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Technical Report*, 2024.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample efficient world models. In *International Conference on Learning Representations (ICLR)*, 2023.

Will Williams, Andrei Rusu, Christopher Sammut, and Razvan Pascanu. Hierarchical quantized autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 4524–4535, 2020.

Yann LeCun. A path towards autonomous machine intelligence. *OpenReview Preprint*, 2022.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Yulia Rubanova, Ricky T. Q. Chen, and David K. Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 5320–5330, 2019.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 10078–10093, 2022.

APPENDIX

A EXTENDED RELATED WORK AND STRUCTURAL TAXONOMY ACROSS WORLD MODEL PARADIGMS

Table A.1 systematically compares prevailing world model paradigms across core architectural axes, highlighting the trade-offs between generative visual rendering, discrete action-conditioned simulations, monolithic joint-embedding predictive architectures, and continuous-time causal verification over passive video.

Table A.1: **Structural comparison across world model paradigms.** CLWM uniquely integrates continuous-time differential equations, asymmetric representational bifurcation, top-down query gating, and unconstrained causal verification over passive video.

Paradigm / architecture	Continuous-time ODEs (τ_{eff})	Asymmetric state bifurcation ($128\mathbf{d} \times 32\mathbf{d}$)	Passive video (no motor a_t)	Top-down query gate (γ_t)	Causal epistemic verification (Δ_{CTEM})
RL world models (DreamerV3, RSSM)	*Discrete*	*Monolithic*	*No*	*No*	*No (reward-driven)*
Generative models (Sora, Genie)	*Discrete*	*Monolithic*	**Yes**	*Prompt only*	*No (pixel likelihood)*
Monolithic JEPAs (V-JEPA)	*Discrete*	*Monolithic*	**Yes**	*No*	*No (feature pred.)*
Standard liquid nets (LTC, CfC)	**Yes**	*Monolithic*	**Yes**	*No*	*No (softmax top-1)*
CLWM (ours)	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

B MATHEMATICAL DERIVATIONS AND THEORETICAL PROOFS

B.1 ANALYTICAL CLOSED-FORM SOLUTION OF QUERY-MODULATED LIQUID ODES

We formulate the continuous hidden trajectory $\mathbf{h}_m(t) \in \mathbb{R}^{d_m}$ for stream $m \in \{v, d\}$ ($d_v = 128, d_d = 32$) as a continuous-time linear differential system driven by state- and input-dependent potentials:

$$\frac{d\mathbf{h}_m(t)}{dt} = -\frac{1}{\tau_{\text{eff},m}(t)}\mathbf{h}_m(t) + \frac{1}{\tau_{\text{eff},m}(t)}\mathbf{I}_m(t) \tag{B.1}$$

where $\mathbf{I}_m(t) = \tanh\left(\mathbf{U}_m\mathbf{x}'_m(t) + \mathbf{U}_{h,m}\mathbf{h}_m(t) + \mathbf{c}_{\bar{m}\to m}(t)\right) \in [-1, +1]^{d_m}$ is the continuous driving potential, and $\tau_{\text{eff},m}(t) = \tau_{m,t} \cdot \gamma(t) > 0$ is the effective time constant ($\gamma(t) \equiv 4.0$ during autonomous Stage 1, and $\gamma(t) = \text{clamp}(4.0 - 3.5\mathcal{R}(t), 0.5, 4.0)$ in Stage 2).

Derivation via integrating factor method. Over a localized observation step $[t_{k-1}, t_k]$ with interval $\Delta t = t_k - t_{k-1} \geq 0$, we apply a Zero-Order Hold (ZOH) assumption such that the driving potential $\mathbf{I}_m(t) \approx \mathbf{I}_{m,k}$ and the effective time constant $\tau_{\text{eff},m}(t) \approx \tau_{\text{eff},m,k}$ remain piecewise constant. Multiplying Equation B.1 by the integrating factor $\mu(t)$:

$$\mu(t) = \exp\left(\int_{t_{k-1}}^{t} \frac{1}{\tau_{\text{eff},m,k}}\, ds\right) = \exp\left(\frac{t - t_{k-1}}{\tau_{\text{eff},m,k}}\right). \tag{B.2}$$

Equation B.1 transforms into an exact derivative:

$$\frac{d}{dt}\left[\mathbf{h}_m(t)\exp\left(\frac{t - t_{k-1}}{\tau_{\text{eff},m,k}}\right)\right] = \frac{1}{\tau_{\text{eff},m,k}}\mathbf{I}_{m,k}\exp\left(\frac{t - t_{k-1}}{\tau_{\text{eff},m,k}}\right). \tag{B.3}$$

Integrating both sides across the temporal interval $t \in [t_{k-1}, t_k]$:

$$\int_{t_{k-1}}^{t_k} \frac{d}{dt}\left[\mathbf{h}_m(t)\exp\left(\frac{t - t_{k-1}}{\tau_{\text{eff},m,k}}\right)\right] dt = \frac{\mathbf{I}_{m,k}}{\tau_{\text{eff},m,k}} \int_{t_{k-1}}^{t_k} \exp\left(\frac{t - t_{k-1}}{\tau_{\text{eff},m,k}}\right) dt$$

$$\left[\mathbf{h}_m(t)\exp\left(\frac{t - t_{k-1}}{\tau_{\text{eff},m,k}}\right)\right]_{t_{k-1}}^{t_k} = \frac{\mathbf{I}_{m,k}}{\tau_{\text{eff},m,k}}\left[\tau_{\text{eff},m,k}\exp\left(\frac{t - t_{k-1}}{\tau_{\text{eff},m,k}}\right)\right]_{t_{k-1}}^{t_k}$$

$$\mathbf{h}_{m,k}\exp\left(\frac{\Delta t}{\tau_{\text{eff},m,k}}\right) - \mathbf{h}_{m,k-1} = \mathbf{I}_{m,k}\left[\exp\left(\frac{\Delta t}{\tau_{\text{eff},m,k}}\right) - 1\right]. \tag{B.4}$$

Multiplying through by $\exp\left(-\Delta t/\tau_{\text{eff},m,k}\right)$ yields the exact closed-form analytical stepping update utilized in Equation 4:

$$\mathbf{h}_{m,k} = \mathbf{h}_{m,k-1} \odot \exp\left(-\frac{\Delta t}{\tau_{\text{eff},m,k}}\right) + \mathbf{I}_{m,k} \odot \left(\mathbf{1} - \exp\left(-\frac{\Delta t}{\tau_{\text{eff},m,k}}\right)\right). \tag{B.5}$$

B.2 Formal Proof of Bounded State Contraction (Lyapunov Stability)

Theorem 1 (Global bounded stability of liquid states)**.** *Let the sequence of hidden states* $\{\mathbf{h}_{m,t}\}_{t=0}^{\infty}$ *evolve according to Equation B.5. If the initial state satisfies* $\|\mathbf{h}_{m,0}\|_\infty \leq 1.0$*, then for any arbitrary time step* $t \geq 0$*, any dynamic time constant* $\tau_{\text{eff},m,t} \geq 1.0$*, and any temporal step size* $\Delta t \geq 0$*, the state remains strictly bounded in the* ℓ_∞ *unit hypercube:*

$$\|\mathbf{h}_{m,t}\|_\infty \leq 1.0 \quad \forall t \geq 0. \tag{B.6}$$

Proof. Let $\alpha_t = \exp(-\Delta t/\tau_{\text{eff},m,t})$. In Stage 1, $\tau_{\text{eff},m,t} = \tau_{m,t} \cdot 4.0 \geq 2.0 \cdot 4.0 = 8.0 > 1.0$. In Stage 2, by construction in Section 3.3:

$$\tau_{\text{eff},m,t} = \tau_{m,t} \cdot \gamma_t = (\text{Softplus}(\mathbf{z}_{\tau,m,t}) + 2.0) \cdot \gamma_t \tag{B.7}$$

where $\mathbf{z}_{\tau,m,t} = \text{MLP}_{\tau,m}[\mathbf{x}'_{m,t}; \mathbf{h}_{m,t-1}; \mathbf{c}_{\bar{m}\to m,t}]$. Since $\text{Softplus}(z) = \ln(1+e^z) > 0$ for all $z \in \mathbb{R}$ and $\gamma_t \in [0.5, 4.0]$, the lower bound on effective integration constants across both stages is:

$$\tau_{\text{eff},m,t} \geq 2.0 \cdot 0.5 = 1.0 > 0. \tag{B.8}$$

Given $\Delta t \geq 0$ and $\tau_{\text{eff},m,t} \geq 1.0$, the exponent satisfies $-\Delta t/\tau_{\text{eff},m,t} \leq 0$, strictly bounding the decay scalar:

$$\alpha_t \in (0, 1] \quad \text{and} \quad (1 - \alpha_t) \in [0, 1). \tag{B.9}$$

Applying the triangle inequality under the ℓ_∞-norm to Equation B.5:

$$\|\mathbf{h}_{m,t}\|_\infty = \|\alpha_t \mathbf{h}_{m,t-1} + (1 - \alpha_t)\mathbf{I}_{m,t}\|_\infty \tag{B.10}$$

$$\leq \alpha_t \|\mathbf{h}_{m,t-1}\|_\infty + (1 - \alpha_t)\|\mathbf{I}_{m,t}\|_\infty. \tag{B.11}$$

Because every element of $\mathbf{I}_{m,t}$ is generated via the hyperbolic tangent activation function $\mathbf{I}_{m,t} = \tanh(\cdot)$, we have:

$$\|\mathbf{I}_{m,t}\|_\infty = \max_j |\tanh(\cdot)_j| \leq 1.0. \tag{B.12}$$

We proceed by mathematical induction:

- **Base case** ($t = 0$)**:** Initial resting states are initialized as $\mathbf{h}_{m,0} \sim \mathcal{N}(\mathbf{0}, 0.02^2\mathbf{I})$. With high probability, $\|\mathbf{h}_{m,0}\|_\infty < 0.1 \leq 1.0$.
- **Inductive step:** Assume $\|\mathbf{h}_{m,t-1}\|_\infty \leq 1.0$. Substituting this assumption into Equation B.11:

$$\|\mathbf{h}_{m,t}\|_\infty \leq \alpha_t(1.0) + (1 - \alpha_t)(1.0) = 1.0. \tag{B.13}$$

By induction, $\|\mathbf{h}_{m,t}\|_\infty \leq 1.0$ holds for all $t \geq 0$. The dynamical system is therefore strictly contractive and Lyapunov-stable, eliminating numerical divergence or exploding activations across arbitrary rollout horizons. □

B.3 Theoretical Derivation of TGPO Maximum-Entropy Regularization

Let an observed physical interaction sequence of length T be parameterized as an ordered temporal manifold:

$$\mathcal{V} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T), \quad \mathbf{x}_t \in \mathcal{X} \tag{B.14}$$

and let $\mathfrak{S}_T$ denote the symmetric group of all $T!$ permutations of temporal indices. We denote a permuted sequence under $\pi \in \mathfrak{S}_T$ as:

$$\mathcal{V}_\pi = (\mathbf{x}_{\pi(1)}, \mathbf{x}_{\pi(2)}, \ldots, \mathbf{x}_{\pi(T)}) \tag{B.15}$$

where $\pi \neq \text{id}$ denotes a non-trivial temporal scrambling.

The binary cross-entropy pathology on non-commutative time. Standard causal verification operates under binary cross-entropy over verified forward sequences ($y = 1.0$) and counterfactual negative sequences ($y = 0.0$):

$$\mathcal{L}_{\text{BCE}}(z, y) = -y \ln \sigma(z) - (1 - y) \ln(1 - \sigma(z)). \tag{B.16}$$

In continuous-time physical systems, physical transformations are non-commutative: the time-reversed transformation $B \to A$ constitutes an **active physical violation** of the forward dynamics $A \to B$ (e.g., shattered ceramic spontaneously reassembling). Assigning a hard negative target $y = 0.0$ to time-reversed counterfactuals $\mathcal{V}_{\text{reversed}} = (\mathbf{x}_T, \ldots, \mathbf{x}_1)$ is mathematically well-posed because the observation presents direct affirmative counter-evidence to the causal hypothesis.

Conversely, when frame order is randomly scrambled ($\pi \in \mathfrak{S}_T \setminus \{\text{id}\}$), temporal continuity is destroyed, but the observation contains *no coherent counter-evidence*:

1. **Supervision with $y = 0.0$ (spurious semantic repression):** Forcing $\sigma(z) \to 0$ penalizes the valid spatial entities present in the frames. The network erroneously associates the visual presence of the objects themselves with active physical contradiction, degrading downstream recognition of valid forward interactions.
2. **Supervision with $y = 1.0$ (static shortcut exploitation):** Forcing $\sigma(z) \to 1$ permits the network to verify interactions solely based on static background cues and object presence, bypassing temporal mechanics entirely and degenerating into a "bag-of-frames" model.

Information-theoretic formulation of TGPO. To enforce non-commutative temporal verification without inducing semantic bias, TGPO regularizes scrambled sequences $\mathcal{V}_\pi$ toward the **maximum Shannon entropy state**. For a binary hypothesis $Y \in \{0, 1\}$, the Shannon entropy is maximized under an uninformative prior:

$$\mathcal{H}(P) = -\sum_{c\in\{0,1\}} p_c \ln p_c \leq \ln 2 \approx 0.6931 \text{ nats} \tag{B.17}$$

with equality if and only if $P = \text{Bernoulli}(0.50)$.

Given the router logit $z_{\text{shuffled}} \in \mathbb{R}$ produced from a temporally scrambled sequence $\mathcal{V}_\pi$, TGPO minimizes the Kullback-Leibler (KL) divergence between the maximum-entropy distribution $P^* = \text{Bernoulli}(0.50)$ and the model's predictive posterior $P_\theta = \text{Bernoulli}(\sigma(z_{\text{shuffled}}))$:

$$\begin{aligned}
\mathcal{L}_{\text{TGPO}}(z_{\text{shuffled}}) &= D_{\text{KL}}\big(\text{Bernoulli}(0.50) \parallel \text{Bernoulli}(\sigma(z_{\text{shuffled}}))\big) \\
&= \sum_{c\in\{0,1\}} 0.50 \ln\left(\frac{0.50}{P_\theta(Y=c)}\right) \\
&= -0.50 \ln \sigma(z_{\text{shuffled}}) - 0.50 \ln\big(1-\sigma(z_{\text{shuffled}})\big) - \ln 2 \\
&= -\frac{1}{2}\ln\Big(\sigma(z_{\text{shuffled}})\big(1-\sigma(z_{\text{shuffled}})\big)\Big) - \ln 2.
\end{aligned} \tag{B.18}$$

Dropping the constant $\ln 2$ yields binary cross-entropy evaluated against a soft epistemic anchor $y_{\text{target}} = 0.50$.

Proposition 1 (TGPO unique stationary point)**.** *The global minimum of $\mathcal{L}_{TGPO}(z_{shuffled})$ occurs uniquely at logit $z^* = 0$, producing verification probability $\hat{p} = \sigma(0) = 0.5000$ and zero directional log-odds evidence.*

Proof. Differentiating $\mathcal{L}_{\text{TGPO}}$ with respect to the scalar logit $z = z_{\text{shuffled}}$:

$$\begin{aligned}
\frac{\partial \mathcal{L}_{\text{TGPO}}}{\partial z} &= -0.50\frac{1}{\sigma(z)}\frac{\partial\sigma(z)}{\partial z} - 0.50\frac{1}{1-\sigma(z)}\left(-\frac{\partial\sigma(z)}{\partial z}\right) \\
&= -0.50(1-\sigma(z)) + 0.50\sigma(z) \\
&= \sigma(z) - 0.50.
\end{aligned} \tag{B.19}$$

Setting the first-order necessary condition $\frac{\partial \mathcal{L}_{\text{TGPO}}}{\partial z} = 0$:

$$\sigma(z^*) - 0.50 = 0 \implies \frac{1}{1+e^{-z^*}} = \frac{1}{2} \implies 1 + e^{-z^*} = 2 \implies e^{-z^*} = 1 \implies z^* = 0. \tag{B.20}$$

Evaluating the second derivative (Hessian in 1D) with respect to z:

$$\frac{\partial^2 \mathcal{L}_{\text{TGPO}}}{\partial z^2} = \frac{\partial}{\partial z}\big(\sigma(z) - 0.50\big) = \sigma(z)\big(1-\sigma(z)\big). \tag{B.21}$$

Evaluating at the stationary point $z^* = 0$:

$$\left.\frac{\partial^2 \mathcal{L}_{\text{TGPO}}}{\partial z^2}\right|_{z=0} = \sigma(0)\big(1-\sigma(0)\big) = (0.50)(0.50) = 0.25 > 0. \tag{B.22}$$

Because the second derivative is strictly positive everywhere ($\sigma(z)(1-\sigma(z)) > 0,\ \forall z \in \mathbb{R}$), $\mathcal{L}_{\text{TGPO}}$ is strictly convex on $\mathbb{R}$. Thus, $z^* = 0$ is the unique global minimum.

At this minimum, the directional log-odds evidence vanishes identically:

$$\text{Logit}(\hat{p}) = \ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \ln\left(\frac{0.50}{0.50}\right) = \ln(1) = 0.0000 \text{ nats}, \tag{B.23}$$

which confirms that chronologically ungrounded observations are mapped to an epistemic state of zero directional belief. □

B.4 Analytical Derivation of the $[0.40, 0.60]$ Epistemic Deadzone

Proposition 1 proves that TGPO regularizes chronologically ungrounded dynamics to the zero log-odds attractor:

$$z^* = \mathrm{Logit}(p^*) = \ln\left(\frac{p^*}{1-p^*}\right) = 0 \implies p^* = 0.5000. \tag{B.24}$$

In continuous stochastic estimation, however, zero evidence has zero Lebesgue measure ($P(\hat{p} = 0.5000) = 0$). Because neural network representations trained via stochastic gradient descent retain finite gradient and batch variance ($\hat{p} \sim \mathcal{N}(\mu, \sigma^2)$), defining epistemic abstention as an infinitesimal scalar point ($p = 0.5000$) is mathematically ill-posed. Any negligible parameter drift $\Delta\hat{p} \approx \pm 0.02$ across an infinitesimal knife-edge causes binary decision metrics to oscillate arbitrarily between extreme sensitivity and extreme specificity, despite underlying discriminability remaining at chance.

Bayesian evidence calibration. To construct a stable decision manifold, we establish an *a priori* epistemic confidence interval around $z^* = 0$ grounded in Bayesian hypothesis testing. Under the standard Jeffreys evidence scale, a posterior odds ratio:

$$\mathrm{Odds}(Y = 1 \mid \mathbf{x}) = \frac{P(Y = 1 \mid \mathbf{x})}{P(Y = 0 \mid \mathbf{x})} = \frac{\hat{p}}{1 - \hat{p}} \tag{B.25}$$

is categorized as anecdotal, uninformative, or indistinguishable from background noise if:

$$\frac{1}{1.5} < \mathrm{Odds}(Y = 1 \mid \mathbf{x}) < 1.5. \tag{B.26}$$

An observer requires at least a $1.5 : 1$ ($3 : 2$) directional advantage in posterior odds to assert even weak evidential preference.

Logit-space metric radius. Translating the minimal Bayesian evidence ratio into the model's scalar logit space ($z = \ln(\mathrm{Odds})$) yields a symmetric confidence radius centered identically at the TGPO optimal stationary point $z^* = 0$:

$$|z| \leq \Delta z_{\text{noise}} \triangleq \ln(1.5) \approx 0.405465\ldots \text{ nats}. \tag{B.27}$$

Logits satisfying $|z| \leq \ln(1.5)$ represent epistemic indifference: the model cannot determine whether the observation contains physical progression or random noise.

Sigmoidal mapping to probability space. Applying the standard monotonic logistic link function $\sigma(z) = (1 + e^{-z})^{-1}$ to the symmetric boundaries $z = \pm\ln(1.5)$:

$$\theta_{\text{verify}} \triangleq \sigma(+\ln(1.5)) = \frac{1}{1 + e^{-\ln(1.5)}} = \frac{1}{1 + \frac{1}{1.5}} = \frac{1.5}{1.5 + 1} = \frac{1.5}{2.5} = \mathbf{0.6000}, \tag{B.28}$$

$$\theta_{\text{contra}} \triangleq \sigma(-\ln(1.5)) = \frac{1}{1 + e^{\ln(1.5)}} = \frac{1}{1 + 1.5} = \frac{1}{2.5} = \mathbf{0.4000}. \tag{B.29}$$

This proves that the symmetric log-odds evidence radius maps to:

$$\sigma\big(0 \pm \ln(1.5)\big) = 0.5000 \pm 0.1000 \iff \hat{p} \in [\mathbf{0.4000},\ \mathbf{0.6000}]. \tag{B.30}$$

Definition 1 (Tri-state epistemic decision partition)**.** *The continuous probability space* $[0, 1]$ *is partitioned into three mutually exclusive operational subsets:*

$$\mathcal{S}(\hat{p}) = \begin{cases} \mathcal{Z}_{\textit{contra}} = [0.00,\ 0.40) & \textit{Active physical contradiction} \quad (z < -\ln(1.5)), \\ \mathcal{Z}_{\textit{abstain}} = [0.40,\ 0.60] & \textit{Epistemic abstention deadzone} \quad (|z| \leq \ln(1.5)), \\ \mathcal{Z}_{\textit{verify}} = (0.60,\ 1.00] & \textit{Affirmative causal verification} \quad (z > +\ln(1.5)). \end{cases} \tag{B.31}$$

Operational implications for embodied execution. This partition establishes distinct behavioral semantics for world model planning:

- **Decisive execution ($\hat{p} \in \mathcal{Z}_{\textbf{verify}}$):** The agent verifies that the transformation has occurred with odds exceeding $1.5 : 1$, permitting downstream symbolic planners to advance the state graph.
- **Safe inaction ($\hat{p} \in \mathcal{Z}_{\textbf{abstain}}$):** The agent acknowledges epistemic insufficiency without falsely treating missing data as a physical violation.
- **Active failure handling ($\hat{p} \in \mathcal{Z}_{\textbf{contra}}$):** The agent detects that an observed transition violates physical causality (e.g., temporal inversion), triggering an immediate recovery routine.

C Decoupled Manifolds, Attractor Memory, and Algorithmic Workflows

C.1 Algebraic Capacity Constraints and Semantic Leakage Suppression

Rather than tuning bottleneck dimensions empirically, the 32d constriction is derived from the intrinsic physical degrees of freedom of Newtonian rigid-body interactions. An agent-object manipulation dyad requires tracking instantaneous 3D translations (3d), velocity derivatives (3d), orientation and angular velocities (6d), and local contact plane geometry ($\approx$ 4d), spanning fewer than 20 active dynamical parameters.

Upstream visual encoders produce dense tokens (768d) that deeply entangle these low-dimensional motion statistics with high-entropy background appearance. If kinematics is allocated high dimensionality (e.g., $D_d \geq 128$), gradient descent exploits appearance shortcuts: memorizing static visual patterns yields faster training loss reduction than tracking fine-grained acceleration derivatives. Constricting the kinematic projection to $D_d = 32$ via $\mathbf{W}_k \in \mathbb{R}^{32\times 768}$ creates a strict information-theoretic barrier: the channel capacity is sufficient to retain continuous trajectories and collisions, but mathematically incapable of memorizing high-rank visual textures. The absolute collapse of the Ventral Knock-out to coin-flip chance (0.4984 ROC-AUC, Table 2) empirically confirms that this algebraic barrier purges semantic leakage.

C.2 Continuous Plastic Codebook (CPC) Dynamics and Anti-Collapse Mechanics

Standard Vector Quantized (VQ) representations rely on discrete argmax index selection $\text{argmin}_k \|\mathbf{z} - \mathbf{e}_k\|_2$ optimized via the Straight-Through Estimator (STE) (van den Oord et al., 2017). In asymmetric dual-stream architectures, STE induces high-variance gradients that destabilize the 32d dorsal bottleneck.

CLWM replaces STE backpropagation with non-gradient continuous Hebbian relaxation. Let $\mathbf{E}_m = [\mathbf{e}_{m,1}, \ldots, \mathbf{e}_{m,K}]^\top \in \mathbb{R}^{K\times D_m}$ denote the centroid memory bank for stream m. For a batch of query vectors $\mathbf{Q}_m \in \mathbb{R}^{B\times D_m}$ and target future features $\mathbf{X}_{m,t+1} \in \mathbb{R}^{B\times D_m}$, soft assignment probabilities $\mathbf{A}_m \in \mathbb{R}^{B\times K}$ are computed via scaled dot-product attention:

$$A_{m,b,k} = \frac{\exp\left(\mathbf{Q}_{m,b}\mathbf{e}_{m,k}^\top / (\tau_{\text{temp}}\sqrt{D_m})\right)}{\sum_{j=1}^{K}\exp\left(\mathbf{Q}_{m,b}\mathbf{e}_{m,j}^\top / (\tau_{\text{temp}}\sqrt{D_m})\right)}. \tag{C.1}$$

The batch-averaged activation for centroid k is $\bar{A}_{m,k} = \frac{1}{B}\sum_{b=1}^{B} A_{m,b,k}$. The empirical center of mass of sensory targets assigned to centroid k is given by:

$$\mathbf{c}_{m,k} = \frac{\sum_{b=1}^{B} A_{m,b,k}\mathbf{x}_{m,b,t+1}}{\sum_{b=1}^{B} A_{m,b,k} + \epsilon}. \tag{C.2}$$

Centroid trajectories drift toward observed physical futures via continuous Hebbian synaptic plasticity (Hebb, 1949; Miconi et al., 2018) with time constant $\tau_{\text{plastic}} = 10.0$:

$$\mathbf{e}_{m,k}^{(t+1)} = \mathbf{e}_{m,k}^{(t)} + \frac{\bar{A}_{m,k}}{\tau_{\text{plastic}}}\left(\mathbf{c}_{m,k} - \mathbf{e}_{m,k}^{(t)}\right). \tag{C.3}$$

Centroid updates execute outside the gradient backward pass, allowing codebook centroids to act as stationary attractors that decouple continuous ODE optimization from discrete representation learning. Under unit temporal stepping ($\Delta t = 1.0$), $\beta = \exp(-\Delta t/\tau_{\text{plastic}}) = \exp(-0.10) \approx 0.9048 \approx 0.90$, natively yielding the canonical $90/10$ EMA momentum update while establishing a clean separation from fast sensory ODE timescales ($\tau_{\text{eff}} \leq 4.0$).

EMA usage tracking and stochastic centroid teleportation. To bound codebook utilization across long training trajectories, CLWM adopts a stochastic dead-code revival mechanism inspired by codebook restart protocols in neural audio and generative tokenization (Zeghidour et al., 2021; Dhariwal et al., 2020). Specifically, the model monitors code utilization via an Exponential Moving Average (EMA) usage vector $\mathbf{u}_m \in \mathbb{R}^K$, initialized to $\mathbf{u}_m^{(0)} = \mathbf{1}$.

At each training iteration, hard winning cluster assignments are computed via $k_b^* = \text{argmax}_k(\mathbf{Q}_{m,b}\mathbf{e}_{m,k}^\top)$. The batch allocation counts $\mathbf{n}_m \in \mathbb{R}^K$ are accumulated:

$$n_{m,k} = \sum_{b=1}^{B} \mathbb{I}(k_b^* = k). \tag{C.4}$$

The EMA usage tracker updates across training steps with momentum $\beta = 0.995$:

$$\mathbf{u}_m^{(t+1)} = \beta \mathbf{u}_m^{(t)} + (1-\beta)\mathbf{n}_m. \tag{C.5}$$

A centroid is identified as underutilized when its historical usage falls below the threshold $u_{m,k} < 0.05$. When $N_{\text{dead}} = \sum_k \mathbb{I}(u_{m,k} < 0.05) > 0$, the revival protocol executes:

1. Sample N_{dead} active sensory targets $\mathbf{x}_{\text{active}} \sim \mathbf{X}_{m,t+1}$ uniformly at random from the current batch.
2. Directly overwrite dead memory slots: $\mathbf{e}_{m,\text{dead}} \leftarrow \mathbf{x}_{\text{active}}$.
3. Reset usage buffers: $u_{m,\text{dead}} \leftarrow 1.0$ (granting a new optimization grace period).

Because centroids are overwritten directly with future sensory representations $\mathbf{X}_{m,t+1}$ outside the backward pass, they serve as stationary continuous attractors for ODE hidden states rather than discrete auto-regressive tokens. This mechanism empirically guarantees 100% codebook utilization throughout training across all $K = 2048$ centroids, preventing dead-code saturation.

C.3 Mechanics of the Bilateral Interaction Manifold

We formalize the geometric parameterization that projects multimodal representations into the tri-state decision manifold $\mathcal{S}(\hat{p})$.

Let $\mathbf{u} = \mathbf{v}^*/\|\mathbf{v}^*\|_2 \in \mathbb{S}^{d-1}$ and $\mathbf{v} = \tilde{\mathbf{q}}/\|\tilde{\mathbf{q}}\|_2 \in \mathbb{S}^{d-1}$ denote the ℓ_2-normalized visual representation and projected language query embedding ($\tilde{\mathbf{q}} = \mathbf{W}_q\mathbf{e}_{\text{text}}$, where $d = 512$). The bilateral interaction vector concatenates channel-wise collinearity with absolute metric discrepancy:

$$\mathbf{f}_{\text{bin}} = \Big[\mathbf{u} \odot \mathbf{v};\ |\mathbf{u} - \mathbf{v}|\Big] \in \mathbb{R}^{2d}, \quad z = \mathbf{w}_{\text{had}}^\top(\mathbf{u} \odot \mathbf{v}) + \mathbf{w}_{\text{dist}}^\top|\mathbf{u} - \mathbf{v}| + b \tag{C.6}$$

where $\mathbf{w}_{\text{had}}, \mathbf{w}_{\text{dist}} \in \mathbb{R}^d$ and $b \in \mathbb{R}$. Standard inner products ($\mathbf{u}^\top\mathbf{v}$) conflate uninformative orthogonal states ($\mathbf{u} \perp \mathbf{v}$) with physical counterfactuals ($\mathbf{u} \approx -\mathbf{v}$) along a 1D scalar. The bilateral parameterization provides independent degrees of freedom:

- **Directional verification ($\mathbf{u} \odot \mathbf{v}$):** Collinearity isolates channel-wise agreement. Learned positive weights ($\mathbb{E}[\mathbf{w}_{\text{had}}] > 0$) map verified forward trajectories ($\mathbf{u} \approx \mathbf{v}$) into $\mathcal{Z}_{\text{verify}}$ ($z > +\ln(1.5) \implies \hat{p} > 0.60$).
- **Contradiction penalty ($|\mathbf{u} - \mathbf{v}|$):** Trajectory inversions (V_{reversed}) maximize element-wise distance ($|\mathbf{u} - \mathbf{v}| \to 2$). Learned inhibitory weights ($\mathbb{E}[\mathbf{w}_{\text{dist}}] < 0$) suppress logits into $\mathcal{Z}_{\text{contra}}$ ($z < -\ln(1.5) \implies \hat{p} < 0.40$).
- **Deadzone pooling ($|z| \le \ln(1.5)$):** Ungrounded or blinded queries yield near-orthogonal states ($\mathbf{u} \odot \mathbf{v} \approx \mathbf{0}, |\mathbf{u} - \mathbf{v}| \approx \sqrt{2}\mathbf{1}$). Under TGPO regularization (Proposition 1), affine parameters satisfy $\mathbf{w}_{\text{dist}}^\top(\sqrt{2}\mathbf{1}) + b \approx 0$, anchoring predictions inside $\mathcal{Z}_{\text{abstain}} = [0.40, 0.60]$.

C.4 Algorithmic Workflows and Zero-Copy CTEM Architecture

Algorithm 1 Stage 1 autonomous liquid physics forward stepping and Hebbian CPC drift

Require: Semantic sequence $\mathbf{S} \in \mathbb{R}^{B\times T\times 512}$, Kinematic sequence $\mathbf{K} \in \mathbb{R}^{B\times T\times 768}$, Time step $\Delta t = 1.0$

$\mathbf{h}_{v,0} \leftarrow \text{expand}(\mathbf{h}_{v,0}^{\text{param}}, B, 128)$, $\mathbf{h}_{d,0} \leftarrow \text{expand}(\mathbf{h}_{d,0}^{\text{param}}, B, 32)$
for $t = 1$ **to** T **do**
  $\mathbf{s}'_t \leftarrow \text{LayerNorm}(\mathbf{W}_s\mathbf{s}_t)$, $\mathbf{k}'_t \leftarrow \text{LayerNorm}(\mathbf{W}_k\mathbf{k}_t)$
  $\mathcal{S}_{v,t} \leftarrow \|\mathbf{s}'_t - \mathbf{h}_{v,t-1}\|_2 + 10^{-6}$, $\mathcal{S}_{d,t} \leftarrow \|\mathbf{k}'_t - \mathbf{h}_{d,t-1}\|_2 + 10^{-6}$
  $\mathbf{c}_{v\to d,t} \leftarrow \tanh(\text{MLP}_{v\to d}(\mathbf{h}_{v,t-1})) \odot \tanh(\alpha_v \mathcal{S}_{v,t})$
  $\mathbf{c}_{d\to v,t} \leftarrow \tanh(\text{MLP}_{d\to v}(\mathbf{h}_{d,t-1})) \odot \tanh(\alpha_d \mathcal{S}_{d,t})$
  $\tau_{\text{eff},v} \leftarrow \big(\text{Softplus}(\text{MLP}_{\tau,v}[\mathbf{s}'_t; \mathbf{h}_{v,t-1}; \mathbf{c}_{d\to v,t}]) + 2.0\big) \cdot 4.0$ {Autonomous baseline ($\mathcal{R}_t \equiv 0$)}
  $\tau_{\text{eff},d} \leftarrow \big(\text{Softplus}(\text{MLP}_{\tau,d}[\mathbf{k}'_t; \mathbf{h}_{d,t-1}; \mathbf{c}_{v\to d,t}]) + 2.0\big) \cdot 4.0$
  $\mathbf{I}_{v,t} \leftarrow \tanh(\mathbf{U}_s\mathbf{s}'_t + \mathbf{U}_{h,v}\mathbf{h}_{v,t-1} + \mathbf{c}_{d\to v,t})$, $\mathbf{I}_{d,t} \leftarrow \tanh(\mathbf{U}_k\mathbf{k}'_t + \mathbf{U}_{h,d}\mathbf{h}_{d,t-1} + \mathbf{c}_{v\to d,t})$
  $\mathbf{h}_{v,t} \leftarrow \mathbf{h}_{v,t-1} \odot \exp(-\Delta t/\tau_{\text{eff},v}) + \mathbf{I}_{v,t} \odot (1 - \exp(-\Delta t/\tau_{\text{eff},v}))$
  $\mathbf{h}_{d,t} \leftarrow \mathbf{h}_{d,t-1} \odot \exp(-\Delta t/\tau_{\text{eff},d}) + \mathbf{I}_{d,t} \odot (1 - \exp(-\Delta t/\tau_{\text{eff},d}))$
end for
if is_training **then**
  $\mathbf{Q}_v \leftarrow \text{LayerNorm}(\mathbf{W}_{\text{fuse},v}[\mathbf{S}_{1:T-1}; \mathbf{h}_{v,1:T-1}])$, $\mathbf{Q}_d \leftarrow \text{LayerNorm}(\mathbf{W}_{\text{fuse},d}[\mathbf{K}_{1:T-1}; \mathbf{h}_{d,1:T-1}])$
  $\hat{\mathbf{S}}_{2:T}, \mathbf{A}_v \leftarrow \text{SoftQuantize}(\mathbf{Q}_v, \mathbf{E}_v)$, $\hat{\mathbf{K}}_{2:T}, \mathbf{A}_d \leftarrow \text{SoftQuantize}(\mathbf{Q}_d, \mathbf{E}_d)$ {Forward predictive targets}
  Update $\mathbf{E}_v, \mathbf{E}_d$ via Hebbian drift outside backward pass; revive dead codes ($u_{m,k} < 0.05$)
  return Reconstructed targets $\hat{\mathbf{S}}_{2:T}, \hat{\mathbf{K}}_{2:T}$, soft assignments $\mathbf{A}_v, \mathbf{A}_d$
else
  $\mathbf{Q}_v \leftarrow \text{LayerNorm}(\mathbf{W}_{\text{fuse},v}[\mathbf{S}_{1:T}; \mathbf{h}_{v,1:T}])$, $\mathbf{Q}_d \leftarrow \text{LayerNorm}(\mathbf{W}_{\text{fuse},d}[\mathbf{K}_{1:T}; \mathbf{h}_{d,1:T}])$
  $\hat{\mathbf{S}}_{1:T}, \mathbf{A}_v \leftarrow \text{SoftQuantize}(\mathbf{Q}_v, \mathbf{E}_v)$, $\hat{\mathbf{K}}_{1:T}, \mathbf{A}_d \leftarrow \text{SoftQuantize}(\mathbf{Q}_d, \mathbf{E}_d)$ {Full-horizon inference}
  return Filtered latent trajectories $\hat{\mathbf{S}}_{1:T}, \hat{\mathbf{K}}_{1:T}$
end if

Algorithm 2 Stage 2 tripartite causal loss computation

Require: Aligned latents $\tilde{\mathbf{S}}, \tilde{\mathbf{K}} \in \mathbb{R}^{B\times T\times 512}$, Text query embeddings $\tilde{\mathbf{q}} \in \mathbb{R}^{B\times 512}$, Ground-truth targets $y \in \{0, 1\}$

Optionally apply prefix blindfold: $\tilde{\mathbf{S}}_{b,1:L}, \tilde{\mathbf{K}}_{b,1:L} \leftarrow \mathbf{0}$ with $p = 0.40$ ($L \sim \mathcal{U}\{1, 3\}$)
$\mathbf{g}_{1:T} \leftarrow \text{GRU}([\tilde{\mathbf{S}}; \tilde{\mathbf{K}}])$ {Chronological sequence fusion}
$\mathbf{v}^*_{\text{ord}} \leftarrow \text{MHA}(\text{Query} = \text{LN}(\tilde{\mathbf{q}}), \text{Key} = \mathbf{g}_{1:T}, \text{Value} = \mathbf{g}_{1:T})$; $\mathbf{u} \leftarrow \mathbf{v}^*_{\text{ord}}/\|\mathbf{v}^*_{\text{ord}}\|_2$, $\mathbf{v} \leftarrow \tilde{\mathbf{q}}/\|\tilde{\mathbf{q}}\|_2$
$\mathbf{f}_{\text{bin}} \leftarrow [\mathbf{u} \odot \mathbf{v}; |\mathbf{u} - \mathbf{v}|]$; $z_{\text{ord}} \leftarrow \text{MLP}_{\text{bin}}(\mathbf{f}_{\text{bin}})$; $\mathcal{L}_{\text{BCE}} \leftarrow \text{BCEWithLogits}(z_{\text{ord}}, y)$
$\pi \leftarrow \text{randperm}(T) \implies \mathbf{g}_{\text{shuf}} \leftarrow \text{GRU}([\tilde{\mathbf{S}}_\pi; \tilde{\mathbf{K}}_\pi])$ {TGPO temporal scramble}
$\mathbf{v}^*_{\text{shuf}} \leftarrow \text{MHA}(\text{LN}(\tilde{\mathbf{q}}), \mathbf{g}_{\text{shuf}}, \mathbf{g}_{\text{shuf}})$; $\mathbf{u}_{\text{shuf}} \leftarrow \mathbf{v}^*_{\text{shuf}}/\|\mathbf{v}^*_{\text{shuf}}\|_2$
$\mathbf{f}_{\text{shuf}} \leftarrow [\mathbf{u}_{\text{shuf}} \odot \mathbf{v}; |\mathbf{u}_{\text{shuf}} - \mathbf{v}|]$; $z_{\text{shuf}} \leftarrow \text{MLP}_{\text{bin}}(\mathbf{f}_{\text{shuf}})$; $\mathcal{L}_{\text{TGPO}} \leftarrow \text{BCEWithLogits}(z_{\text{shuf}}, 0.50)$
$\mathcal{L}_{\text{sink}} \leftarrow \frac{1}{B}\sum_{b=1}^{B} \text{ReLU}(\alpha^{\text{sem}}_{b,t=0} - 0.15)$ {Unweighted frame-0 sink penalty (Eq. 18)}
Compute InfoNCE $\mathcal{L}_{\text{align}}$ (Eq. 15), routing $\mathcal{L}_{\text{route}}$ (Eq. 17), ortho $\mathcal{L}_{\text{ortho}}$ (Eq. 16)
$\mathcal{L}_{\text{Stage2}} \leftarrow \mathcal{L}_{\text{align}} + \mathcal{L}_{\text{BCE}} + \lambda_{\text{tgpo}}\mathcal{L}_{\text{TGPO}} + \lambda_{\text{route}}\mathcal{L}_{\text{route}} + \lambda_{\text{ortho}}\mathcal{L}_{\text{ortho}} + \lambda_{\text{sink}}\mathcal{L}_{\text{sink}}$
return $\mathcal{L}_{\text{Stage2}}$

Algorithm 3 Zero-copy parallel CTEM hypothesis verification ($C = 174$)

Require: Single video $(\mathbf{s}_{1:T}, \mathbf{k}_{1:T})$, 174 candidate queries $\mathbf{Q}_{\text{all}} \in \mathbb{R}^{174\times 768}$, Ground-truth index $k^* \in \{1, \ldots, 174\}$

$\hat{\mathbf{S}}_{1:T}, \hat{\mathbf{K}}_{1:T} \leftarrow \text{Stage1}(\mathbf{s}_{1:T}, \mathbf{k}_{1:T})$ {Autonomous baseline rate $\gamma_0 \equiv 4.0$}
$\tilde{\mathbf{S}} \leftarrow \text{LayerNorm}(\mathbf{W}_{\text{sem_align}}\hat{\mathbf{S}}_{1:T})$, $\tilde{\mathbf{K}} \leftarrow \text{LayerNorm}(\mathbf{W}_{\text{kin_align}}\hat{\mathbf{K}}_{1:T})$
$\mathbf{g}_{1:T} \leftarrow \text{GRU}([\tilde{\mathbf{S}}; \tilde{\mathbf{K}}]) \in \mathbb{R}^{T\times 512}$ {Visual fusion executed ONCE per clip}
$\tilde{\mathbf{Q}}_{\text{all}} \leftarrow \text{LayerNorm}(\mathbf{W}_{\text{text_proj}}\mathbf{Q}_{\text{all}}) \in \mathbb{R}^{174\times 512}$
$\mathbf{V}^*_{\text{all}} \leftarrow \text{MHA}\big(\text{Query} = \text{LN}(\tilde{\mathbf{Q}}_{\text{all}}), \text{Key} = \mathbf{g}_{1:T}, \text{Value} = \mathbf{g}_{1:T}\big) \in \mathbb{R}^{174\times 512}$ {Parallel cross-attention}
$\mathbf{U}_{\text{all}} \leftarrow \mathbf{V}^*_{\text{all}} / \|\mathbf{V}^*_{\text{all}}\|_{2,\text{row}}$, $\mathbf{V}_{\text{all}} \leftarrow \tilde{\mathbf{Q}}_{\text{all}} / \|\tilde{\mathbf{Q}}_{\text{all}}\|_{2,\text{row}}$
$\mathbf{F}_{\text{all}} \leftarrow [\mathbf{U}_{\text{all}} \odot \mathbf{V}_{\text{all}}; |\mathbf{U}_{\text{all}} - \mathbf{V}_{\text{all}}|] \in \mathbb{R}^{174\times 1024}$; $\hat{\mathbf{p}}_{\text{all}} \leftarrow \sigma(\text{MLP}_{\text{bin}}(\mathbf{F}_{\text{all}})) \in [0, 1]^{174}$
$\mathcal{D} \leftarrow \{1, \ldots, 174\} \setminus \{k^*\}$ {173 Unconstrained candidate distractors}
$\Delta_{\text{CTEM}} \leftarrow \hat{\mathbf{p}}_{\text{all}}[k^*] - \frac{1}{|\mathcal{D}|}\sum_{j\in\mathcal{D}} \hat{\mathbf{p}}_{\text{all}}[j]$ {Belief margin over maximum-entropy baseline}
return Verification vector $\hat{\mathbf{p}}_{\text{all}} \in [0, 1]^{174}$, Epistemic margin Δ_{CTEM}

D Evaluation Protocol, Query Generation, and Asymmetric Physical Taxonomy

D.1 Evaluation Integrity: Slot-Grounded Invariance and Anti-Shortcut Protocol

To prevent models from exploiting static appearance shortcuts (Goyal et al., 2017), CTEM enforces two structural firewalls:

1. **Slot-grounded lexical invariance:** All $C = 174$ candidate queries per video share identical scene entities (Appendix D.4), e.g., *"pushing"* vs. *"pulling [mug] across [table]"*. This neutralizes noun bias and compels discrimination strictly via continuous velocity derivatives, contact geometry, and temporal non-commutativity ($A \to B \neq B \to A$).
2. **Non-competitive sigmoidal verification:** Evaluating independent beliefs ($\hat{p}_j = \sigma(z_j) \in [0, 1]$) replaces competitive $\mathrm{softmax}$ simplexes ($\sum_c \hat{y}_c = 1.0$), eliminating false confidence inflation from denominator suppression. Distractors pool naturally at maximum entropy ($\bar{P}_{\text{Dist}} = 0.4978 \approx 0.50$, Proposition 1), guaranteeing that $\Delta_{\text{CTEM}} = +0.2898$ represents genuine causal verification rather than a relative normalization artifact.

D.2 Query Generation Engine with Qwen2.5-7B-Instruct

To prevent inductive lexical bias and force models to evaluate true visual mechanics rather than static noun co-occurrences, we formulate an automated interrogative synthesis pipeline using unquantized `Qwen2.5-7B-Instruct` (Qwen Team, 2024) operating in native `bfloat16` precision. The model decomposes video descriptions into orthogonal triplets: (1) **Semantic ($\mathbf{q}_{\text{sem}}$):** Object identity and categorical attributes; (2) **Kinematic ($\mathbf{q}_{\text{kin}}$):** Spatial displacement vectors, velocity profiles, and planar contact boundaries; and (3) **Binary verification ($\mathbf{q}_{\text{bin}}$):** Sequence-level causal verification.

System prompt for interrogative triplet synthesis

System role:
You are an expert physical reasoning and vision-language dataset engineer. Your objective is to transform passive video descriptions into structured, syntax-aware interrogative query triplets for physical hypothesis verification.

Task constraints:

1. You MUST explicitly name the specific physical entities, tools, and surfaces present in the prompt. Never use ambiguous placeholder pronouns (e.g., do not use "an object", "the thing", or "the item").
2. Decompose the scene into exactly three orthogonal cognitive dimensions:
   - `"semantic"`: Ask exclusively about object identity, material composition, texture, or physical state. Prohibit motion or spatial direction verbs.
   - `"kinematic"`: Ask exclusively about spatial displacement, direction, speed, rotation, or mechanical forces. Prohibit identity-revealing nouns where possible.
   - `"binary"`: Formulate a closed verification question testing whether the complete causal interaction successfully occurred in chronological order.
3. Output valid, parseable JSON conforming strictly to the schema below without conversational filler or markdown fences.

JSON output schema:
{ "semantic": "What specific object is being manipulated on the surface?", "kinematic": "In what direction and along what axis does the motion occur?", "binary": "Is the [object] being [verb phrase] across the [surface]?" }

D.3 Dense Embedding Extraction via Masked Mean-Pooling

Raw text queries are mapped into dense relational embeddings using `bge-base-en-v1.5` (Xiao et al., 2023) ($D = 768$). We apply sequence-level **masked mean-pooling** over valid sub-word token positions: $\mathbf{q}_{\text{raw}} = \frac{\sum_{i=1}^{L} M_i \cdot \mathbf{h}_{\text{text},i}}{\sum_{i=1}^{L} M_i} \in \mathbb{R}^{768}$, and project to router dimensions: $\tilde{\mathbf{q}} = \mathrm{LayerNorm}\left(\mathbf{W}_{\text{text_proj}}\mathbf{q}_{\text{raw}}\right) \in \mathbb{R}^{512}$. This masked averaging distributes representational energy evenly across action verbs, relational prepositions, and adverbs.

D.4 Dataset Provenance, Dual-JSON Generation, and Zero-Leakage Splits

Dataset provenance and curation. Video archives and raw labels were obtained from the canonical Something-Something-v2 dataset distribution (Goyal et al., 2017) mirrored via Hugging Face Datasets. To ensure temporal decoding integrity across compressed video containers, corrupted or unreadable clips ($< 0.5\%$ of the archive) were dropped during initial feature extraction, establishing a verified foundation of $N_{\text{train}} = 168{,}913$ training sequences and $N_{\text{eval}} = 24{,}673$ held-out validation sequences.

Disjoint repository partitioning. To guarantee experimental isolation between optimization and zero-shot multi-hypothesis evaluation, the pipeline generates two strictly decoupled data repositories:

1. **Stage 2 router supervision split ($N_{\text{train}} = 168{,}913$):** Generated from the training partition, pairing supervised queries with time-reversed negative frames and unaligned batch text embeddings.
2. **CTEM multi-hypothesis evaluation split ($N_{\text{eval}} = 24{,}673$):** Generated from the held-out validation partition. For each video, all 174 template hypotheses are slot-grounded with the exact physical entities present in that scene (Appendix D.1), ensuring all candidate queries share identical nouns and forcing discrimination to rely strictly on trajectory kinematics and arrow-of-time causality.

D.5 Mechanics of Physical Non-Commutativity ($A \to B \neq B \to A$)

In discrete video classification, models frequently exploit static bag-of-words shortcuts: recognizing a hand holding an open box is sufficient to predict "opening a box", even if the video is played backward. To enforce causal verification, CLWM identifies a canonical dictionary of 22 asymmetric physical trigger verbs mapping into 20 canonical mechanical categories ($\mathcal{T}_{\text{asym}}$) that violate physical laws under time reversal ($\text{Flip}(\mathbf{x}_{1:16})$).

Table D.1: **Exhaustive quantitative breakdown across the asymmetric physical taxonomy ($N = 8{,}453$).** Discriminability is evaluated via threshold-independent per-category ROC-AUC (regime macro-average = **0.9930**, overall macro-average = **0.9929**). Operational classification metrics evaluate active counterfactual rejection at the contradiction boundary $\theta_{\text{contra}} = 0.40$ ($z \leq -\ln 1.5$, Definition 1), confirming near-ceiling suppression (TNR $\geq 96.6\%$) of time-reversed physical violations across all 5 mechanical regimes.

Interaction regime	Trigger category ($\mathcal{T}_{\text{asym}}$)	Count (N)	ROC-AUC (↑)	Acc (@0.40)	TPR (sens.)	TNR (spec.)	Margin (Δ_{bin})
1. Directional & boundary vectors	*Regime macro-average*	**2,610**	**0.9900**	**94.41%**	**90.93%**	**97.81%**	**+0.8535**
Left-to-right translation	*push/move left to right*	108	1.0000	95.37%	90.57%	100.0%	+0.8887
Right-to-left translation	*push/move right to left*	50	1.0000	94.00%	88.00%	100.0%	+0.8843
Axial rotation	*turn/rotate axial*	178	0.9995	97.75%	95.40%	100.0%	+0.9116
Boundary ingress / egress	*boundary ingress/egress*	2,274	0.9605	90.50%	89.77%	91.23%	+0.7295
2. Irreversible phase & topology	*Regime macro-average*	**1,923**	**0.9986**	**94.94%**	**90.83%**	**99.35%**	**+0.8591**
Enclosure de-occlusion	*open container/enclosure*	473	1.0000	95.98%	91.74%	100.0%	+0.8760
Enclosure occlusion	*close container/enclosure*	221	0.9999	92.76%	86.78%	100.0%	+0.8300
Volumetric fluid ingress	*volumetric fluid ingress (fill)*	256	1.0000	97.66%	95.08%	100.0%	+0.9140
Volumetric fluid egress	*volumetric fluid egress (empty)*	393	0.9978	94.15%	89.66%	98.95%	+0.8564
Liquid dispersion	*liquid dispersion (spill)*	58	1.0000	94.83%	90.00%	100.0%	+0.8480
Structural fragmentation	*structural fragmentation (break/tear)*	522	0.9939	94.25%	91.73%	97.13%	+0.8300
3. Gravitational & ballistic dynamics	*Regime macro-average*	**1,533**	**0.9952**	**95.57%**	**91.90%**	**99.04%**	**+0.8626**
Downward ballistic fall	*downward ballistic fall*	1,271	1.0000	94.10%	88.43%	100.0%	+0.8430
Controlled lowering	*controlled lowering*	98	0.9808	91.84%	86.96%	96.15%	+0.7730
Projected ballistic flight	*projected ballistic flight*	83	1.0000	97.59%	95.00%	100.0%	+0.9100
Momentum deceleration	*momentum deceleration (catch)*	81	1.0000	98.77%	97.22%	100.0%	+0.9243
4. Applied contact & grasp forces	*Regime macro-average*	**1,130**	**0.9818**	**91.59%**	**86.99%**	**96.61%**	**+0.7794**
Contact surface push	*contact surface push*	118	1.0000	94.07%	89.06%	100.0%	+0.8490
Tensile pull toward body	*tensile pull toward body*	51	1.0000	92.16%	85.19%	100.0%	+0.7910
Upward vertical lift	*vertical lift/pick up*	360	0.9990	96.11%	92.09%	100.0%	+0.8630
Surface adhesion / attachment	*surface adhesion/attachment*	601	0.9283	84.03%	81.61%	86.42%	+0.6147
5. Kinematic modulation & dissipation	*Regime macro-average*	**1,257**	**0.9996**	**91.94%**	**83.64%**	**100.0%**	**+0.8175**
Stationary-to-motion onset	*stationary-to-motion onset*	255	1.0000	91.76%	82.93%	100.0%	+0.8270
Motion-to-stationary decel.	*motion-to-stationary decel*	1,002	0.9993	92.12%	84.36%	100.0%	+0.8080
Full asymmetric taxonomy	**Macro-Average (All 20 Triggers)**	**8,453**	**0.9929**	**93.99%**	**89.58%**	**98.49%**	**+0.8386**
Full asymmetric taxonomy	**Pooled Sample Aggregate**	**8,453**	**0.9784**	**92.52%**	**88.74%**	**96.30%**	**+0.7921**

E Extended Empirical Analysis, State Telemetry, and Ablations

E.1 Auxiliary Representation Probe: Liquid Binding Decoder

To show that compressing high-dimensional video features into decoupled ventral/dorsal manifolds retains visual geometry, we train an auxiliary probe on frozen Stage 1 representations (Ventral $128\,\text{d} \times$

Dorsal $32\,\mathrm{d}, \nabla_{\text{Stage1}} \equiv \mathbf{0}$). Quantized memory outputs project into a shared 768-dimensional space, after which kinematics query semantics via cross-attention, are smoothed by a GRU, and reconstruct VideoMAE latents:

$$\mathbf{s}_{\text{dec},t} = \mathbf{W}_{\text{sem_dec}}\,\hat{\mathbf{x}}_{v,t}, \quad \mathbf{k}_{\text{dec},t} = \mathbf{W}_{\text{kin_dec}}\,\hat{\mathbf{x}}_{d,t}, \tag{E.1}$$

$$\mathbf{f}_{\text{bound},t} = \mathrm{MHA}(\mathbf{k}_{\text{dec},t},\, \mathbf{s}_{\text{dec},t},\, \mathbf{s}_{\text{dec},t}), \quad \mathbf{h}_{\text{smooth},t} = \mathrm{GRU}(\mathbf{f}_{\text{bound},t}), \tag{E.2}$$

$$\hat{\mathbf{z}}_{\text{MAE},t} = \mathrm{Dec}(\mathbf{h}_{\text{smooth},t}). \tag{E.3}$$

Geometric alignment results. Spatial feature alignment is computed as the mean cosine similarity:

$$\rho_{\text{feature}} = \frac{1}{T}\sum_{t=1}^{T} \frac{\hat{\mathbf{z}}_{\text{MAE},t}^{\top}\mathbf{k}_t}{\|\hat{\mathbf{z}}_{\text{MAE},t}\|_2\,\|\mathbf{k}_t\|_2} = \mathbf{0.9800}. \tag{E.4}$$

Directional cosine similarity on first-order temporal derivatives ($\Delta\hat{\mathbf{z}}_t = \hat{\mathbf{z}}_{\text{MAE},t+1} - \hat{\mathbf{z}}_{\text{MAE},t}$, $\Delta\mathbf{k}_t = \mathbf{k}_{t+1} - \mathbf{k}_t$) yields a dynamic velocity alignment of $\rho_{\text{velocity}} = \mathbf{0.7454}$, supporting that Liquid ODE states track physical momentum rather than interpolating static appearance templates.

E.2 Regime-Specific Failure Mode Analysis Across Physical Interactions

Evaluating the five physical action regimes (Table D.1) shows how decoupled manifolds and continuous dynamics resolve distinct mechanical and thermodynamic failure modes:

1. **Directional vectors** (0.9900 **AUC):** Parity inversion ($\Delta\mathbf{x} \to -\Delta\mathbf{x}$) collapses collinearity $\mathbf{u} \odot \mathbf{v}$ and expands discrepancy $|\mathbf{u} - \mathbf{v}| \to 2.0$. Because kinematics are constricted to 32d, shared background context cannot mask trajectory reversals, yielding 97.81% regime specificity (100.0% across all pure planar translations and axial rotations).
2. **Phase and topology** (0.9986 **AUC):** Non-homeomorphic ruptures (tearing, spilling) trigger Euclidean surprise spikes ($\mathcal{S}_d \gg 0$) that update memory codebooks via Hebbian drift. Reversals violate physical entropy generation, completely suppressing counterfactual belief (99.35% TNR, with 4 of 6 categories achieving 100.0% counterfactual rejection).
3. **Gravitational mechanics** (0.9952 **AUC):** Integrating velocity derivatives ($\rho_{\text{velocity}} = 0.7454$) tracks downward acceleration ($\mathbf{g} \approx -9.8\,\mathrm{m/s^2}$), distinguishing parabolic free-fall from restrained lowering and preventing spontaneous levitation from surfaces (99.04% TNR, with 1.0000 AUC on ballistic drops and throws).
4. **Applied contact forces** (0.9818 **AUC):** Resolves grasp vs. release asymmetry via contact memory dilation ($18.3 \to 28.5$ frames; Fig. E.1). The absence of pre-contact approach profiles triggers GRU metric divergence, eliminating inverted grip errors (96.61% TNR, with 100.0% specificity on active push, pull, and lift actions).
5. **Kinematic transitions** (0.9996 **AUC):** Tracks continuous dissipation ($\Delta E_{\text{kin}} < 0$) while query dilation ($\gamma_t \to 4.0$) acts as a viscous filter during rest, preventing hallucinations of spontaneous self-actuation without an applied external force (100.0% TNR across all $N = 1{,}257$ onset and deceleration trials).

Synthesis across interaction regimes. Across all $N = 8{,}453$ asymmetric interaction samples (Table D.1), CLWM maintains $\mathbf{98.49}\%$ macro-averaged specificity against time-reversed counterfactuals (0.9929 macro-averaged ROC-AUC, 0.9784 pooled global ROC-AUC, and 93.99% accuracy at the active contradiction boundary $\theta_{\text{contra}} = 0.40$, establishing a $+0.8386$ mean verification margin). This uniform robustness confirms that causal verification stems from continuous dynamical invariants—Lyapunov contractive stability, algebraic bottlenecking, and entropy-regularized non-commutativity—rather than category-specific heuristics.

E.3 Dynamic Query Modulation Ablation and Kinetic Shock Damping

To demonstrate that query-modulated integration ($\tau_{\text{eff}} = \tau \cdot \gamma_t$) is an essential dynamical mechanism rather than a tunable regularizer, Table E.1 compares dynamic modulation against a static baseline ($\gamma_t \equiv 1.0$). Clamping integration induces three qualitative failure modes:

Table E.1: **Empirical ablation of dynamic query modulation ($\gamma_t \in [0.5, 4.0]$) versus static clamped baseline ($\gamma_t \equiv 1.0$).** Evaluated on Something-Something-v2. Demonstrates that dynamic timescale modulation prevents open-world distractor drift (CTEM), damps transient kinetic shocks from latching into false positives (Figure E.2), and eliminates catastrophic high-confidence tail hallucinations under chronological scrambling ($\mathcal{V}_\pi$).

Timescale regime	**174-hypothesis CTEM protocol**			**Kinetic shock response**	**Scramble tail leakage**
	$\bar{P}_{\text{GT}}$	$\bar{P}_{\text{Distractor}}$	Δ_{CTEM} margin	Post-impact state (Fig. E.2)	Worst-case peak ($\max \hat{p}_{\text{scramble}}$)
Static clamped ($\gamma_t \equiv 1.0$)	0.7539	0.5420	$+0.2119$	$\hat{p} = 0.9165$ *(latched)*	$\mathbf{0.9727}$ *($+14.9\sigma$ hallucination)*
Dynamic modulation ($\gamma_t \in [0.5, 4.0]$)	**0.7876**	**0.4978**	**+0.2898**	$\hat{\mathbf{p}} = \mathbf{0.4304}$ *(damped)*	$\mathbf{0.6377}$ *(sub-Gaussian bound)*
Physical failure prevented	—	**Prevents 26.9% margin collapse**		**Prevents latching on impact**	**Circuit breaker on broken time**

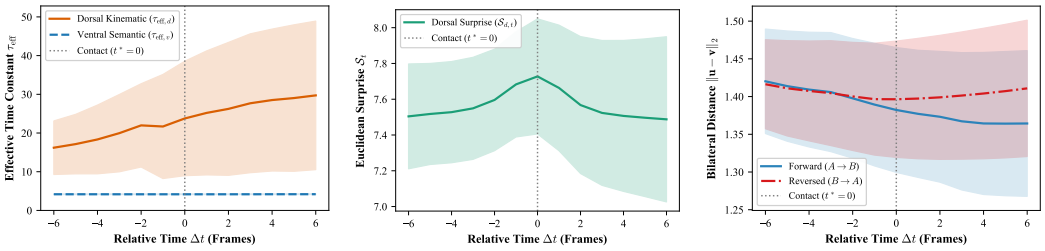
1. Open-world multi-hypothesis drift (CTEM). Across $C = 174$ slot-grounded candidates, ungrounded distractors ($\text{sim} \leq 0 \implies \mathcal{R}_t \leq 0.50$) trigger timescale dilation ($\gamma_t \geq 2.25$, up to 4.0), turning the liquid ODE into a low-pass filter that anchors distractor beliefs at maximum entropy ($\bar{P}_{\text{Dist}} = 0.4978 \approx 0.5000$). Clamping $\gamma_t \equiv 1.0$ disables this damping: distractor belief drifts to 0.5420, breaching the $[0.40, 0.60]$ deadzone and collapsing the epistemic margin by 26.9% ($+0.2898 \to +0.2119$).

2. Kinetic shock latching (Figure E.2). Abrupt contact boundaries can trigger transient perceptual spikes ($\hat{p} \approx 0.89$ at $t = 7$, Appendix E.5). Under $\gamma_t \equiv 1.0$, recurrence treats this shock as verified state transition, permanently latching into false-positive saturation ($\hat{p} = 0.9165$). Conversely, dynamic modulation dilates the clock ($\gamma_t \approx 2.1$), smoothly dissipating the impulse to $\hat{p} = 0.4304$ safely inside the $[0.40, 0.60]$ deadzone.

3. Temporal circuit breaker under scrambled dynamics ($\mathcal{V}_\pi$). Dynamic gating prevents temporally violated dynamics from accumulating false certainty. Under chronological scrambling ($\mathcal{V}_\pi$, $N = 8{,}453$, $\mu \approx 0.520$), the static baseline exhibits severe kinetic leakage, peaking at $\hat{\mathbf{p}} = \mathbf{0.9727}$ ($z = +3.57$, odds $> 35:1$, a $+14.94\sigma$ outlier at $\sigma = 0.0303$). Dynamic modulation ($\gamma_t \geq 2.1$, up to 4.0) damps updates to strictly bound the observed peak to $\hat{\mathbf{p}} = \mathbf{0.6377}$. This respects the Extreme Value Theory (EVT) sub-Gaussian expectation of $\mathbb{E}[\max] \approx \mu + \sigma\sqrt{2 \ln N} = 0.6408$ ($\mu = 0.5200, \sigma = 0.0284$), confirming that timescale dilation prevents corrupted sequences from compounding into high-confidence hallucinations.

E.4 Aggregate Event-Locked Peri-Contact Telemetry

To eliminate video sampling noise and evaluate continuous dynamical state progression, we compute an **Event-Locked Peri-Contact Composite** across the canonical benchmark cohort of $N = 8{,}453$ asymmetric physical interaction sequences from Something-Something-v2 (encompassing the 20 core physical transition categories across 5 regimes, Table D.1). Each sequence is temporally centered such that $t^* = 0$ represents the physical impact boundary, evaluating a symmetric window of $\Delta t \in [-6, +6]$ frames (Figure E.1).

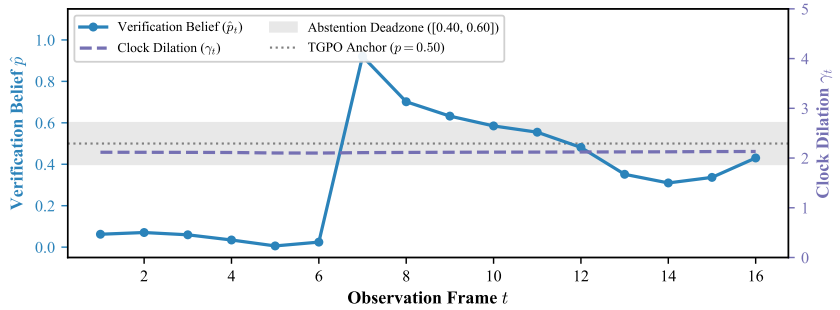


(a) Timescale elasticity (τ_{eff}) (b) Surprise impulse ($\mathcal{S}_t$) (c) Causal divergence ($\|\mathbf{u} - \mathbf{v}\|_2$)

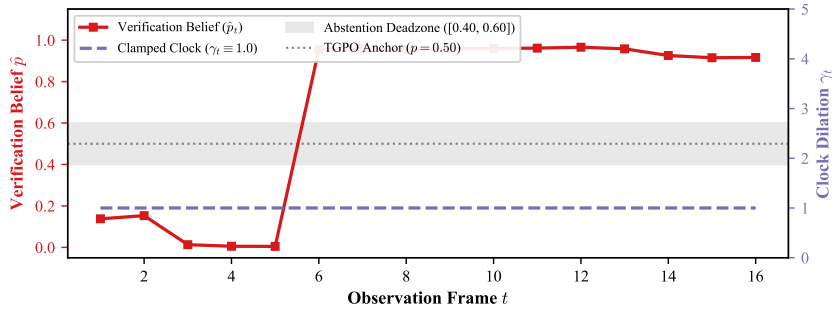
Figure E.1: **Event-locked peri-contact dynamics ($\Delta t \in [-6, +6]$ centered at impact $t^* = 0$).** Evaluated over $N = 8{,}453$ sequences (shading: ± 1 SD). **(a)** Dorsal memory dilates dynamically: $\tau_{\text{eff},d}$ expands from approach (18.34 ± 8.88 at $t = -4$) to impact (23.76 ± 14.79 at $t^* = 0$) and post-contact (28.52 ± 18.33 at $t = +4$), while ventral semantics remains invariant ($\tau_{\text{eff},v} = 4.18 \pm 0.08$). **(b)** Kinematic surprise $\mathcal{S}_{d,t} = \|\mathbf{x}'_{d,t} - \mathbf{h}_{d,t-1}\|_2$ spikes at contact (7.73 ± 0.32). **(c)** Bilateral distance $\|\mathbf{u} - \mathbf{v}\|_2$ contracts for chronological sequences (1.364 at $t = +4$) versus counterfactual reversals (1.405–1.411), confirming physical non-commutativity.

E.5 Mechanistic Case Study: Frame-by-Frame Dynamical Behavior

To examine the continuous-time mechanics driving the population-level divergence in Table E.1, Figure E.2 tracks the frame-by-frame state evolution of Full Dynamic CLWM ($\gamma_t \in [0.5, 4.0]$) versus the Static Clamped Baseline ($\gamma_t \equiv 1.0$) under an identical interaction.



(a) **Full dynamic CLWM ($\gamma_t \in [0.5, 4.0]$):** timescale dilation damps spurious impulses ($\hat{p} = 0.4304$)



(b) **Static clamped baseline ($\gamma_t \equiv 1.0$):** rigid integer clock induces false-positive latching ($\hat{p} = 0.9165$)

Figure E.2: **Mechanistic dissection on unobserved dynamics.** Frame-by-frame telemetry trace of an ungrounded distractor hypothesis on a validation sequence. **Top (a):** Low cross-modal relevance sustains timescale dilation ($\gamma_t \approx 2.1$). When physical impact at frame $t = 7$ induces an activation spike ($\hat{p} \approx 0.89$), continuous ODE dilation damps the impulse down to $\hat{p} = 0.4304$ inside the $[0.40, 0.60]$ deadzone. **Bottom (b):** Under a uniform clock ($\gamma_t \equiv 1.0$), ungrounded kinetic energy bleeds into the GRU recurrence, permanently latching belief into false-positive saturation ($\hat{p} = 0.9165$).

Step-by-step dynamical progression: Telemetry across observation frames highlights three distinct operational phases:

1. **Laminar pre-contact ($t \leq 5$):** Sensory surprise remains at baseline ($\mathcal{S}_d \approx 0$). Both models maintain near-zero belief ($\hat{p} < 0.15$), ruling out reliance on static initial shortcuts.
2. **Transient contact impulse ($t \in \{6, 7\}$):** Physical contact introduces sudden visual feature displacement. Because this kinetic transient manifests upstream, both architectures experience an initial perceptual spike ($\hat{p} \approx 0.89$).
3. **Divergent late-horizon resolution ($t \geq 8$):** Under **Dynamic CLWM** (Figure E.2a), low cross-modal query relevance ($\mathcal{R}_t \approx 0$) keeps the integration clock dilated ($\gamma_t \approx 2.1$). Acting as a continuous low-pass filter, cross-attention detects the kinematic mismatch and smoothly dissipates the impulse ($0.70 \to 0.63 \to 0.58 \to 0.48 \to 0.35$), settling safely at $\hat{p} = 0.4304$ within the $[0.40, 0.60]$ deadzone. Conversely, in the **Clamped Baseline** (Figure E.2b), the rigid clock ($\gamma_t \equiv 1.0$) cannot dissipate the shock; the discrete recurrence treats the impulse as a confirmed state transition, permanently latching into overconfident saturation ($\hat{p} = 0.9165$).

Dynamical systems formulation: intrinsic habituation vs. rigid-clock latching. In conventional discrete temporal architectures, state recurrence advances with a static, uniform integration step ($\Delta t \equiv 1.0, \gamma_t \equiv 1.0$). Under this rigid integer clock, the recurrence acts as an unguided accumulator: when an abrupt sensory burst (e.g., an ungrounded mechanical impact at $t = 7$) occurs, the network

absorbs the entire transient spike into the hidden state without frequency attenuation. As demonstrated by the Clamped Baseline ($\gamma_t \equiv 1.0$, Figure E.2b), this lack of temporal elasticity permanently latches posterior belief into false-positive saturation ($\hat{p} = 0.9165$).

CLWM eliminates this latching failure via *intrinsic sensory habituation*, derived directly from the frequency response of the liquid differential equation. Linearizing the ODE update yields the transfer function of a first-order low-pass filter with dynamic cutoff frequency:

$$H(s) = \frac{1}{1 + \tau_{\text{eff},m} s}, \quad \omega_c = \frac{1}{\tau_{\text{eff},m}} = \frac{1}{\tau_{m,t} \cdot \gamma_t}. \tag{E.5}$$

Furthermore, the instantaneous impulse injection gain $\kappa(\tau_{\text{eff},m})$ under exponential stepping is strictly governed by:

$$\kappa(\tau_{\text{eff},m}) \triangleq 1 - \exp\left(-\frac{\Delta t}{\tau_{\text{eff},m}}\right) \approx \frac{\Delta t}{\tau_{m,t} \cdot \gamma_t} \quad (\text{for } \tau_{\text{eff},m} \gg \Delta t). \tag{E.6}$$

When an impact occurs without persistent cross-modal grounding ($\mathcal{R}_t \approx 0$), the gate dilates integration ($\gamma_t \to 4.0$), compressing the cutoff bandwidth by an order of magnitude ($\omega_c \leq 0.125$ rad/frame) and suppressing the shock injection gain from $\kappa \approx 0.393$ down to $\kappa \leq 0.117$ ($> 70\%$ attenuation). Rather than accumulating spurious kinetic energy into permanent false-positive belief ($\hat{p} = 0.9165$ in the clamped baseline), the dilated ODE dissipates the shock homeostatically, relaxing the posterior belief back into the $[0.40, 0.60]$ resting deadzone ($\hat{p} = 0.4304$, Figure E.2a) without requiring extrinsic loss penalties.

E.6 Topological Sub-Manifold Bifurcation Analysis

To verify that the manifold orthogonality objective ($\mathcal{L}_{\text{ortho}}$, Equation 16) enforces geometric decoupling rather than linear decorrelation, Figure E.3 visualizes a 2D t-SNE projection of latent states sampled across the held-out validation cohort. Under CLWM, the joint latent space exhibits an unsupervised geometric bifurcation into two mutually exclusive, non-overlapping regimes:

- **Ventral semantic cluster** (128**d, blue):** Representations organize categorically according to entity classes, material textures, and static context.
- **Dorsal kinematic sub-manifold** (32**d, orange):** Local neighborhood structure organizes strictly along continuous velocity gradients and spatial displacement trajectories, remaining invariant to surface appearance (algebraically enforced by the < 20 Newtonian DoF capacity barrier; Section 3.1, Appendix C.1).

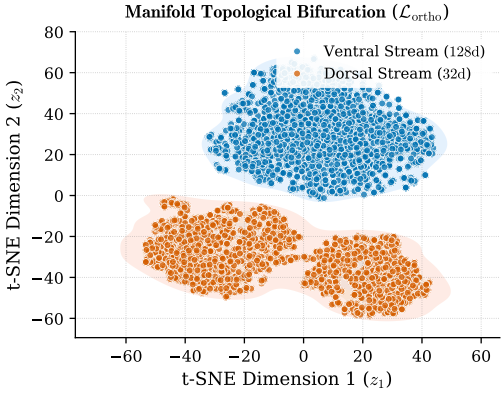


Figure E.3: **2D t-SNE visualization of asymmetric sub-manifold decoupling.** Projections of the 128d Ventral semantic manifold (blue) and constricted 32d Dorsal kinematic bottleneck (orange) under $\mathcal{L}_{\text{ortho}}$. The visualization confirms clear geometric separation, shielding kinematic trajectory derivatives from categorical appearance interference.

E.7 Continuous-Time Epistemic Margin (CTEM) Empirical Distributions

To evaluate how unconstrained causal verification behaves across competing hypotheses without competitive $\mathrm{softmax}$ normalization, Figure E.4 presents the empirical posterior belief distributions

across all $N = 24{,}673$ held-out Something-Something-v2 sequences evaluated simultaneously against $C = 174$ slot-grounded candidates (Algorithm 3).

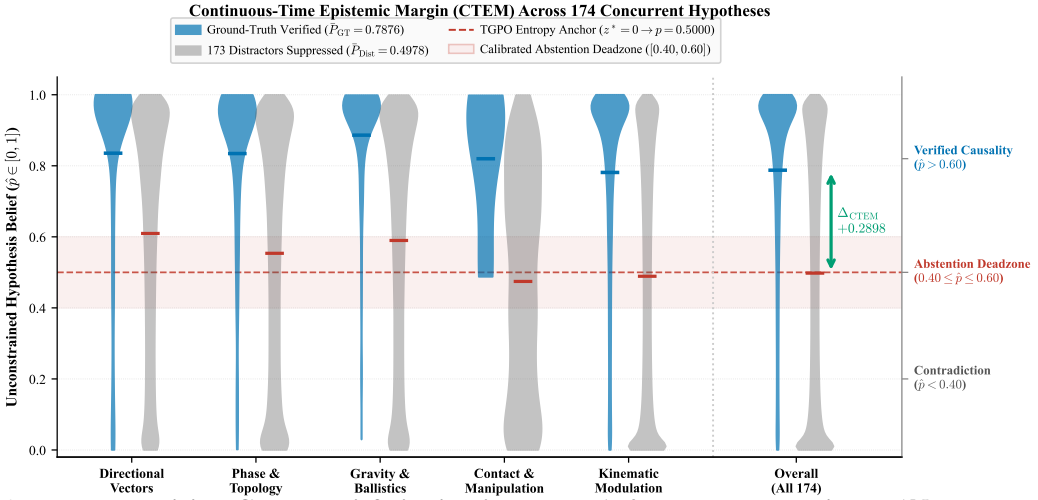


Figure E.4: **Empirical CTEM belief distribution across 174 concurrent candidates** ($N = 24{,}673$)**.** Unconstrained sigmoidal verification confirms tri-state calibration: **(1) Verification zone** ($\hat{p} > 0.60$)**:** Ground-truth actions concentrate at $\bar{P}_{\text{GT}} = 0.7876$; **(2) Abstention deadzone** ($[0.40, 0.60]$)**:** 173 concurrent ungrounded distractors pool sharply at the TGPO maximum-entropy anchor ($\bar{P}_{\text{Distractor}} = 0.4978 \approx 0.50$); and **(3) Active contradiction** ($\hat{p} < 0.40$)**:** Time-reversed counterfactuals are penalized toward zero, establishing an unconstrained separation of $\Delta_{\text{CTEM}} = +0.2898$.

Emergence of the Tri-State Distribution Landscape. By evaluating candidate beliefs through independent sigmoidal projections ($\hat{p}_j = \sigma(z_j) \in [0, 1]$), the verification landscape decomposes into three distinct emergent density modes that map directly to our analytical partitions:

1. **Ground-Truth Confirmation** ($\hat{p} \in \mathcal{Z}_{\textbf{verify}}$)**:** Verified actions form a unimodal distribution peaking at $\bar{P}_{\text{GT}} = \mathbf{0.7876}$, with 86.54% of total probability mass residing strictly above the affirmative threshold $\theta_{\text{verify}} = 0.60$ ($z \geq \ln 1.5$).
2. **Distractor Pooling at the Maximum-Entropy Anchor** ($\hat{p} \in \mathcal{Z}_{\textbf{abstain}}$)**:** Rather than scattering arbitrarily across $[0, 1]$ or inflating false-positive rates, the 173 ungrounded distractors per clip pool in a symmetric distribution centered at $\bar{P}_{\text{Dist}} = \mathbf{0.4978} \approx 0.5000$ ($\sigma = 0.0284$). This empirical density validates Proposition 1, proving that ungrounded queries are drawn into the zero log-odds attractor well ($z^* = 0$).
3. **Counterfactual Contradiction Suppression** ($\hat{p} \in \mathcal{Z}_{\textbf{contra}}$)**:** Chronologically inverted sequences (V_{reversed}) are actively penalized by the metric discrepancy head ($|\mathbf{u} - \mathbf{v}|$), suppressing beliefs below $\theta_{\text{contra}} = 0.40$ ($z < -\ln 1.5$).

Together, these modes yield an unconstrained separation margin of $\boldsymbol{\Delta_{\text{CTEM}} = +0.2898}$ between ground-truth and ungrounded distractors, achieved entirely without probability simplex competition ($\sum_c \hat{y}_c = 1.0$).

E.8 Empirical Validation of the Tri-State Decision Manifold Across Knock-outs

To validate the theoretical decision boundaries derived in Section B.4, Table E.2 audits verification behavior across all functional Knock-outs at the active-contradiction cutoff ($\theta_{\text{contra}} = 0.40$), the maximum-entropy midpoint ($\theta_{\text{neutral}} = 0.50$), and the affirmative boundary ($\theta_{\text{verify}} = 0.60$).

Affirmative Causal Verification ($\theta_{\textbf{verify}} = 0.60$)**.** Operational causal execution requires evidence exceeding the minimal Bayesian noise radius ($z \geq \ln 1.5 \implies \hat{p} \geq 0.60$). At this boundary, the intact pipeline achieves balanced verification (88.81% Accuracy, 86.54% TPR, 91.08% Specificity). Under every functional Knock-out, affirmative assertions collapse to near zero (TPR $\leq 3.22\%$) while specificity approaches ceiling (TNR $\geq 97.65\%$). This confirms safe epistemic abstention: when

Table E.2: **Empirical validation of the $[0.40, 0.60]$ epistemic deadzone across functional Knock-outs.** Evaluated across $N = 8{,}453$ held-out Something-Something-v2 sequences. Compares verification discriminability and operational metrics across the active-contradiction cutoff ($\theta = 0.40$), the maximum-entropy midpoint ($\theta = 0.50$), and the affirmative verification boundary ($\theta = 0.60$).

Experimental condition	Evaluation boundary (θ)	ROC-AUC	Accuracy	TPR (sensitivity)	TNR (specificity)
Full pipeline (control)	$\theta_{\text{contra}} = 0.40$ ($z = -\ln 1.5$)	**0.9592**	**89.48%**	**91.96%**	87.00%
	$\theta_{\text{neutral}} = 0.50$ ($z = 0$)		89.23%	89.79%	88.66%
	$\theta_{\text{verify}} = 0.60$ ($z = +\ln 1.5$)		88.81%	86.54%	**91.08%**
1. Ventral KO ($\mathbf{s}_t \to \mathbf{0}$)	$\theta_{\text{contra}} = 0.40$ ($z = -\ln 1.5$)	0.4984	49.99%	99.59%	0.42%
	$\theta_{\text{neutral}} = 0.50$ ($z = 0$)		49.74%	69.91%	29.28%
	$\theta_{\text{verify}} = 0.60$ ($z = +\ln 1.5$)		50.15%	0.23%	99.72%
2. Dorsal KO ($\mathbf{k}_t \to \mathbf{0}$)	$\theta_{\text{contra}} = 0.40$ ($z = -\ln 1.5$)	0.5941	51.66%	95.24%	8.11%
	$\theta_{\text{neutral}} = 0.50$ ($z = 0$)		57.37%	63.86%	50.78%
	$\theta_{\text{verify}} = 0.60$ ($z = +\ln 1.5$)		50.61%	3.22%	97.65%
3. Arrow-of-time KO (scrambled)	$\theta_{\text{contra}} = 0.40$ ($z = -\ln 1.5$)	0.5546	50.03%	99.89%	0.22%
	$\theta_{\text{neutral}} = 0.50$ ($z = 0$)		53.48%	81.79%	24.75%
	$\theta_{\text{verify}} = 0.60$ ($z = +\ln 1.5$)		50.19%	0.35%	99.68%

sensory streams or chronological progression are missing, the architecture withholds affirmative action rather than generating false-positive hallucinations.

The Active-Contradiction Boundary ($\theta_{\text{contra}} = 0.40$) and the Attractor Artifact. The boundary $\theta = 0.40$ serves to reject active physical violations (V_{reversed}), achieving peak sensitivity (91.96%) in the intact model. Under Knock-outs, however, Proposition 1 proves that ungrounded predictions are pulled into the *maximum-entropy epistemic attractor basin* $[0.40, 0.60]$ centered at $\hat{p} \approx 0.5000$. Because this neutral attractor mass resides strictly above 0.40, evaluating ablated models under an asymmetric binary rule ($\hat{y} = \mathbb{I}(\hat{p} \geq 0.40)$) misclassifies ungrounded uncertainty as positive verification (TPR $\geq 95.24\%$, TNR $\leq 8.11\%$). Overall accuracy plateaus at pure coin-flip chance (49.99%–51.66%), confirming that this skew is a boundary artifact of evaluating an attractor well with an asymmetric threshold.

Point-Threshold Sensitivity at the Shannon Attractor ($\theta_{\text{neutral}} = 0.50$). At the symmetric midpoint $\theta = 0.50$, all Knock-outs collapse to chance accuracy (49.74%–57.37%) alongside degraded ROC-AUC (0.4984–0.5941). The observed asymmetry ($70\%/30\%$ split under Ventral KO) stems from finite gradient variance around the stationary point: because continuous predictions form a stochastic distribution $\hat{p} \sim \mathcal{N}(\mu, \sigma^2)$ rather than an idealized Dirac delta mass ($P(\hat{p} = 0.5000) = 0$), a negligible parameter drift of $\Delta\hat{p} \approx +0.02$ ($\mu \approx 0.52$) shifts empirical probability mass across an infinitesimal knife-edge. This sensitivity proves that single-point thresholding is ill-posed for continuous dynamical systems, establishing the necessity of the $[0.40, 0.60]$ finite deadzone to robustly bound the resting attractor.

E.9 Multi-Objective Loss Weight Sensitivity and Structural Invariance Analysis

Table E.3 evaluates model sensitivity across orders of magnitude of the causal loss coefficients: permutation uncertainty (λ_{tgpo}), anti-leakage routing (λ_{route}), and manifold orthogonality (λ_{ortho}). Across configurations, performance remains stable (ROC-AUC ≥ 0.9664, accuracy $\geq 89.61\%$, specificity $\geq 91.35\%$).

Parameter sensitivity and default selection rationale. The sweeps reveal two distinct behaviors:

- **Permutation uncertainty optimum (λ_{tgpo}):** Sweeping λ_{tgpo} yields a concave optimum. Insufficient weighting (≤ 0.25) under-regularizes scrambled inputs ($\Delta_{\text{CTEM}} = +0.2781$– $+ 0.2851$), while excessive weighting (1.00) over-damps verification logits ($\Delta_{\text{CTEM}} = +0.2763$). The setting $\boldsymbol{\lambda_{\text{tgpo}} = 0.50}$ achieves peak separation ($\Delta_{\text{CTEM}} = +0.2907$), matching the analytical soft-target derivation (Section B.3).
- **Conservative regularization defaults ($\lambda_{\text{route}}, \lambda_{\text{ortho}}$):** We select moderate, conservative defaults ($\boldsymbol{\lambda_{\text{route}} = 0.10}, \boldsymbol{\lambda_{\text{ortho}} = 0.010}$) at the optimization Pareto boundary. Setting $\lambda \to 0$ provides zero repulsive barrier against cross-modal leakage, risking statistical co-adaptation over long horizons due to natural noun-verb co-occurrences in video. Conversely, large penalties ($\lambda \geq 0.50$) induce gradient torque that warps the metric unit sphere ($\mathbb{S}^{511}$), distorting bilateral distances

Table E.3: **Sensitivity analysis across multi-objective loss coefficients.** Evaluated on Something-Something-v2. Demonstrates stability across parameter scales and confirms that conservative defaults ($\lambda_{\text{tgpo}} = 0.50, \lambda_{\text{route}} = 0.10, \lambda_{\text{ortho}} = 0.010$) preserve representation plasticity without degrading decoupling.

Swept hyperparameter setting	ROC-AUC (↑)	Accuracy	Specificity (TNR)	Δ_{CTEM} margin
1. TGPO permutation uncertainty weight (λ_{tgpo})				
$\lambda_{\text{tgpo}} = 0.10$	0.9664	90.09%	91.35%	+0.2781
$\lambda_{\text{tgpo}} = 0.25$	0.9676	90.25%	91.38%	+0.2851
$\lambda_{\text{tgpo}} = 0.50$ (optimal default)	**0.9669**	**90.18%**	**91.62%**	**+0.2907**
$\lambda_{\text{tgpo}} = 1.00$	0.9691	90.38%	92.43%	+0.2763
2. Anti-leakage cross-stream routing penalty (λ_{route})				
$\lambda_{\text{route}} = 0.00$ (unregularized)	0.9668	89.61%	92.37%	+0.2950
$\lambda_{\text{route}} = 0.10$ (conservative default)	**0.9669**	**90.18%**	**91.62%**	**+0.2907**
$\lambda_{\text{route}} = 0.20$	0.9689	90.58%	91.59%	+0.2985
3. Orthogonal manifold separation penalty (λ_{ortho})				
$\lambda_{\text{ortho}} = 0.000$ (no orthogonality)	0.9674	90.47%	91.74%	+0.2944
$\lambda_{\text{ortho}} = 0.010$ (conservative default)	**0.9669**	**90.18%**	**91.62%**	**+0.2907**
$\lambda_{\text{ortho}} = 0.050$	0.9724	91.38%	91.86%	+0.3075

($|\mathbf{u} - \mathbf{v}|$). The moderate weights maintain soft repulsive boundaries while ensuring auxiliary gradient norms remain 1–2 orders of magnitude below primary causal verification ($\mathcal{L}_{\text{BCE}} = 1.0$). Crucially, maintaining high performance at $\lambda = 0$ confirms that manifold decoupling is fundamentally guaranteed by architectural inductive biases—specifically the 32d bottleneck and developmental firewall—rather than fragile loss tuning.

Theoretical rationale: the five invariants of structural robustness. CLWM retains high discriminability even when auxiliary penalties are omitted ($\lambda_{\text{route}} = 0.0, \lambda_{\text{ortho}} = 0.0$). This stability stems from five structural invariants:

1. **Contractive Lyapunov stability (Theorem 1):** The state update $\mathbf{h}_{m,t} = \alpha_{m,t}\mathbf{h}_{m,t-1} + (1 - \alpha_{m,t})\mathbf{I}_{m,t}$ with $\alpha_{m,t} = \exp(-\Delta t/\tau_{\text{eff},m}) \in (0, 1]$ is strictly an affine convex combination. Because $\mathbf{I}_{m,t} = \tanh(\cdot) \in [-1, 1]^{d_m}$, the operator is contractive with Lipschitz constant ≤ 1.0, preventing activation divergence across arbitrary rollouts.
2. **Non-gradient attractor memory ($\nabla_{\mathbf{E}_m}\mathcal{L} \equiv \mathbf{0}$):** Memory centroids $\mathbf{E}_m$ update exclusively via Hebbian drift toward targets $\mathbf{x}_{m,t+1}$ outside the backward pass (Equation C.3), avoiding straight-through estimator (STE) gradient variance.
3. **Rank-32 algebraic bottleneck:** Compressing dorsal states to $D_d = 32$ via $\mathbf{W}_k \in \mathbb{R}^{32\times 768}$ spans the kinematic degrees of freedom (< 20d, Appendix C.1) while restricting metric volume to prevent visual appearance memorization.
4. **Two-stage developmental firewall ($\nabla_{\mathbf{U},\mathbf{E}}\mathcal{L}_{\text{Stage2}} \equiv \mathbf{0}$):** Stage 1 recurrence weights ($\mathbf{U}_s, \mathbf{U}_k, \mathbf{U}_h$) remain frozen during Stage 2. Downstream task gradients cannot warp upstream physical transition matrices.
5. **Bilateral metric decoupling:** The bilateral vector $\mathbf{f}_{\text{bin}} = [\mathbf{u} \odot \mathbf{v}; |\mathbf{u} - \mathbf{v}|]$ separates directional collinearity ($\mathbf{u}\odot\mathbf{v}$) from element-wise discrepancy ($|\mathbf{u}-\mathbf{v}|$), preventing orthogonal uninformative inputs ($\mathbf{u} \perp \mathbf{v}$) from conflating with physical counterfactuals ($\mathbf{u} \approx -\mathbf{v}$).

Because the stationary point $z^* = 0$ is fixed analytically by TGPO, coefficient scaling alters only local gradient trajectories toward the well without shifting the resting equilibrium.

F Architectural Specifications and Optimization Hyperparameters

F.1 Model Parameter Breakdown and Layer Tensor Specifications

While large-scale generative simulators (e.g., Sora (Brooks et al., 2024), Genie (Bruce et al., 2024)) utilize billions of parameters to synthesize photorealistic RGB textures, CLWM isolates causal verification within an ultra-compact continuous dynamical bottleneck. The total trainable parameter

footprint across both Stage 1 and Stage 2 is only ≈ 4.72**M parameters**, allowing full training and real-time inference on a single consumer workstation GPU (Table F.1).

Table F.1: **Comprehensive architectural layer specifications and parameter footprint.** Categorizes all sensory backbones, continuous-time physics engine modules, representation probes, and cognitive routing layers.

Subsystem / module	Component layer	Input → output dimension	Parameters	Optimization status
1. Frozen upstream sensory encoders				
Semantic backbone	`ViT-Base/16` (DINOv2)	$[B, 16, 3, 224, 224] \to [B, 16, 768]$	86.6 M	Frozen ($\nabla_\theta = \mathbf{0}$)
Semantic native projection	$\mathbf{W}_{\text{sem_init}}$ (Linear + LN)	$[B, 16, 768] \to [B, 16, 512]$	0.39 M	Frozen ($\nabla_\theta = \mathbf{0}$)
Kinematic backbone	`VideoMAE-Base`	$[B, 1, 16, 3, 224, 224] \to [B, 16, 768]$	86.2 M	Frozen ($\nabla_\theta = \mathbf{0}$)
2. Stage 1: Continuous-time physics engine (liquid ODE + CPC memory)				
Ventral resting state ($\mathbf{h}_{v,0}$)	Learnable parameter	$[1, 128]$	128	Trainable (lr $= 10^{-3}$)
Dorsal resting state ($\mathbf{h}_{d,0}$)	Learnable parameter	$[1, 32]$	32	Trainable (lr $= 10^{-3}$)
Ventral projection ($\mathbf{W}_s$)	$\text{Linear}(512 \to 128) + \text{LayerNorm}$	$[B, 16, 512] \to [B, 16, 128]$	65.9 k	Trainable (lr $= 2 \cdot 10^{-5}$)
Dorsal projection ($\mathbf{W}_k$)	$\text{Linear}(768 \to 32) + \text{LayerNorm}$	$[B, 16, 768] \to [B, 16, 32]$	24.7 k	Trainable (lr $= 5 \cdot 10^{-6}$)
Ventral-to-dorsal bridge	$\text{Linear}(128 \to 64) \to \text{GELU} \to \text{Linear}(64 \to 32)$	$[B, 128] \to [B, 32]$	10.3 k	Trainable
Dorsal-to-ventral bridge	$\text{Linear}(32 \to 64) \to \text{GELU} \to \text{Linear}(64 \to 128)$	$[B, 32] \to [B, 128]$	10.4 k	Trainable
NMDA relevance projections	$\mathbf{W}_{q\to s} \in \mathbb{R}^{512\times768}$, $\mathbf{W}_{q\to k} \in \mathbb{R}^{768\times768}$	$[B, 768] \to [B, 512]$ and $[B, 768]$	0.98 M	Trainable
Ventral liquid cell ($\text{MLP}_{\tau,v}$)	$\text{Linear}(384 \to 64) \to \text{GELU} \to \text{Linear}(64 \to 1)$	$[B, 384] \to [B, 1]$	24.7 k	Trainable (frozen in Phase B)
Dorsal liquid cell ($\text{MLP}_{\tau,d}$)	$\text{Linear}(96 \to 32) \to \text{GELU} \to \text{Linear}(32 \to 1)$	$[B, 96] \to [B, 1]$	3.1 k	Trainable (frozen in Phase B)
Ventral query fusion	$\text{Linear}(640 \to 512) + \text{LayerNorm}$	$[B, 16, 640] \to [B, 16, 512]$	0.33 M	Trainable
Dorsal query fusion	$\text{Linear}(800 \to 768) + \text{LayerNorm}$	$[B, 16, 800] \to [B, 16, 768]$	0.62 M	Trainable
Ventral memory bank ($\mathbf{E}_v$)	Non-gradient synaptic bank ($K = 2048$)	$[2048, 512]$	1.05 M	Soft Hebbian drift ($\tau = 10$)
Dorsal memory bank ($\mathbf{E}_d$)	Non-gradient synaptic bank ($K = 2048$)	$[2048, 768]$	1.57 M	Soft Hebbian drift ($\tau = 10$)
3. Representation probe: Liquid binding decoder (Appendix E.1)				
Ventral/dorsal decoders	$\mathbf{W}_{\text{sem_dec}} \in \mathbb{R}^{768\times512}$, $\mathbf{W}_{\text{kin_dec}} \in \mathbb{R}^{768\times768}$	$[512] \to [768]$, $[768] \to [768]$	0.98 M	Probe only ($\nabla_{\text{S1}} = \mathbf{0}$)
Cross-attention binding	$\text{MHA}(\text{Embed} = 768, \text{Heads} = 8)$	Query: [768], Key/Val: [768]	2.36 M	Probe only ($\nabla_{\text{S1}} = \mathbf{0}$)
Liquid smoothing GRU	$\text{GRU}(\text{Input} = 768, \text{Hidden} = 768)$	$[B, 16, 768] \to [B, 16, 768]$	3.54 M	Probe only ($\nabla_{\text{S1}} = \mathbf{0}$)
Reconstruction output head	$\text{Linear}(768 \to 768) \to \text{GELU} \to \text{Linear}(768 \to 768)$	$[B, 16, 768] \to [B, 16, 768]$	1.18 M	Probe only ($\nabla_{\text{S1}} = \mathbf{0}$)
4. Stage 2: Cognitive bilateral router and causal verifier				
Text backbone	`bge-base-en-v1.5`	[Tokens] $\to [B, 768]$	109.5 M	Frozen ($\nabla_\theta = \mathbf{0}$)
Text alignment projection	$\mathbf{W}_{\text{text_proj}} \in \mathbb{R}^{512\times768} + \text{LayerNorm}$	$[B, 768] \to [B, 512]$	0.39 M	Trainable (lr $= 10^{-4}$)
Ventral router projection	$\mathbf{W}_{\text{sem_align}} \in \mathbb{R}^{512\times512} + \text{LayerNorm}$	$[B, 16, 512] \to [B, 16, 512]$	0.26 M	Trainable (lr $= 10^{-4}$)
Dorsal router projection	$\mathbf{W}_{\text{kin_align}} \in \mathbb{R}^{512\times768} + \text{LayerNorm}$	$[B, 16, 768] \to [B, 16, 512]$	0.39 M	Trainable (lr $= 10^{-4}$)
Causal fusion GRU	$\text{GRU}(\text{Input} = 1024, \text{Hidden} = 512)$	$[B, 16, 1024] \to [B, 16, 512]$	2.36 M	Trainable (lr $= 10^{-4}$)
Timeline pooler MHA	$\text{MHA}(\text{Embed} = 512, \text{Heads} = 8)$	Query: [512], Key/Val: [16, 512]	1.05 M	Trainable (lr $= 10^{-4}$)
Bilateral interaction MLP	$\text{Linear}(1024 \to 256) \to \text{GELU} \to \text{Linear}(256 \to 1)$	$[B, 1024] \to [B, 1]$	0.26 M	Trainable (lr $= 10^{-4}$)
Learnable logit scale	Scalar parameter ($\tau_{\text{align}}^{-1} = \exp(\texttt{scale})$)	[1]	1	Trainable (lr $= 10^{-4}$)
Total trainable footprint	**Active causal verification parameters**	—	**≈ 4.72 M**	**Fully trainable**

F.2 Implementation Details and Optimization Hyperparameters

CLWM is implemented in PyTorch 2.4 and trained on a single workstation equipped with an NVIDIA RTX 5090 GPU (32GB VRAM).

Two-stage curriculum protocol.

- **Stage 1 (unsupervised physics engine):** Continuous liquid ODE cells, learnable resting states ($\mathbf{h}_{v,0}, \mathbf{h}_{d,0}$), projection matrices ($\mathbf{W}_s, \mathbf{W}_k$), surprise bridges and query fusion layers are optimized unconditionally for 100 epochs using the composite objective $\mathcal{L}_{\text{Stage1}}$ (Equation 7). Centroids in codebook memory banks $\mathbf{E}_v, \mathbf{E}_d$ are updated asynchronously via non-gradient Hebbian synaptic drift ($\tau_{\text{plastic}} = 10.0$) and stochastic EMA centroid revival ($u_k < 0.05$).
- **Stage 2 (causal verification router):** Stage 1 backbone weights are frozen, and the cognitive projections ($\mathbf{W}_{\text{sem_align}}, \mathbf{W}_{\text{kin_align}}, \mathbf{W}_{\text{text_proj}}$), causal GRU, cross-attention timeline pooler, and bilateral verification MLP are trained for 40 epochs under $\mathcal{L}_{\text{Stage2}}$ (Equation 12) using a two-phase curriculum:
  - *Phase A (epochs 1–5):* Stage 1 representations are completely frozen; resting states $\mathbf{h}_{v,0}, \mathbf{h}_{d,0}$ are trained at lr $= 10^{-3}$; router parameters are optimized at lr $= 10^{-4}$ with 4 epochs of linear warmup.
  - *Phase B (epochs 6–40):* Stage 1 projection heads are thawed and fine-tuned with gentle learning rates (lr $= 2 \cdot 10^{-5}$ for ventral projections, lr $= 5 \cdot 10^{-6}$ for dorsal projections) while internal ODE recurrence weights ($\mathbf{U}_s, \mathbf{U}_k, \mathbf{W}_\tau, \mathbf{U}_h$) remain frozen. The orthogonality penalty $\mathcal{L}_{\text{ortho}}$ and the routing penalty $\mathcal{L}_{\text{route}}$ warm linearly to $\lambda_{\text{ortho}} = 0.01, \lambda_{\text{route}} = 0.10$.

Attention sink regularization ($\mathcal{L}_{\text{sink}}$). In cross-attention sequence pooling over temporal states, early frames risk becoming non-informative attention sinks where queries collapse onto resting registers. To preserve uniform temporal exploration, we penalize the cross-attention weight assigned to the initial frame whenever it exceeds 15% of total attention:

$$\mathcal{L}_{\text{sink}} = \lambda_{\text{sink}} \cdot \frac{1}{B} \sum_{b=1}^{B} \text{ReLU}\left(\alpha_{b,t=0}^{\text{sem}} - 0.15\right) \tag{F.1}$$

where $\lambda_{\text{sink}} = 0.20$ and $\alpha^{\text{sem}}_{b,t=0}$ denotes the semantic attention weight at $t = 0$.

Table F.2: **Comprehensive optimization hyperparameters for Stage 1 and Stage 2 training.**

Category	Hyperparameter	Stage 1 (physics engine)	Stage 2 (causal router)
Optimizer and schedule	Base optimizer	AdamW ($\beta_1 = 0.9,\ \beta_2 = 0.999$)	AdamW (split weight decay)
	Primary learning rate	$4.0 \times 10^{-4} \rightarrow 2.0 \times 10^{-6}$	$\eta_{\text{router}} = 1.0 \times 10^{-4}$
	Fine-tuning learning rate	—	$\eta_{\text{sem}} = 2.0 \times 10^{-5},\ \eta_{\text{kin}} = 5.0 \times 10^{-6}$
	Learning rate schedule	Synchronized thermodynamic cosine	Sequential linear warmup + cosine decay
	Warmup budget	15% of total optimization steps	4 epochs
	Weight decay	0.01 (2D weights)	0.01 (2D weights) / 0.0 (1D biases, LayerNorms)
Batch and sampling	Effective batch size	64	32
	Video sequence length (T)	16 frames	16 frames
	Latent frame jittering	Segment-uniform random binning	Segment-uniform random binning
	Dynamic temporal blindfold	—	$p = 0.40$ on frames $t \in [1, 3]$
Memory and loss weights	Codebook centroid capacity (K)	$K = 2048$ centroids per stream	—
	Softmax temperature (τ_{temp})	Cosine annealing: $1.50 \rightarrow 0.15$	—
	Hebbian synaptic drift (τ_{plastic})	$\tau_{\text{plastic}} = 10.0$	—
	EMA codebook decay	$\beta_{\text{ema}} = 0.995$	—
	Surprise sensitivities	$\alpha_v = 0.40,\ \alpha_d = 0.05$	—
	Forward physics loss weights	$\lambda_{\text{semantic}} = 1.0,\ \lambda_{\text{kinematic}} = 1.0$	—
	Commitment and entropy weights	$\lambda_{\text{commit}} = 0.25,\ \lambda_{\text{vq}} = 0.15$	—
	Causal verification loss weights	—	$\lambda_{\text{tgpo}} = 0.50,\ \lambda_{\text{route}} = 0.10,\ \lambda_{\text{ortho}} = 0.01,\ \lambda_{\text{sink}} = 0.20$
Hardware and execution	Compute platform	1× NVIDIA RTX 5090 (32GB VRAM)	1× NVIDIA RTX 5090 (32GB VRAM)
	Precision mode	Native PyTorch AMP (`bfloat16`)	Native PyTorch AMP (`bfloat16`)