WHEN SHOULD A WORLD MODEL SWITCH? SURPRISE-GATED STATE-SPACE MODELS FOR HIERARCHICAL EVENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

A predictive world model faces a recurrent dilemma: preserve context while an event is stable, yet rapidly reorganize state when the current dynamics no longer apply. Most video models leave this decision implicit, and most event-boundary detectors mark changes without letting those changes alter the model's internal dynamics. We introduce HEG-VMAMBA, a multimodal latent world model that operationalizes Event Segmentation Theory as explicit predictive computation: **calibrated predictive innovation** becomes an endogenous control signal, and a learned hierarchical **event hazard** jointly (i) expands a selective state-space model's discretization step, attenuating stale history, and (ii) retrieves an input-conditioned **schema** that initializes the next state. Counterfactual utility regularization teaches the gate to switch only when reorganization improves future prediction, and a learned prior supplies event hazards during open-loop rollout; we prove a monotone bound on the direct contribution of pre-boundary state that a fully open schema-mixture removes exactly. HEG-VMAMBA improves controlled open-loop prediction by -11.5% ($5/5$ seeds, p=0.007) (OOD $+6.7$–12.6% in $5/5$ seeds across all four OOD families); on Kinetics-GEBD its internal gate reaches Rel.Dis. mean-F1 0.760 (zero-shot 5-seed ensemble on $17{,}911$ val clips with extracted features; per-seed 0.754 ± 0.032) without boundary-supervised representation learning, and the same gate reduces open-loop prediction error on KTH Actions at recurring action-regime starts while remaining a detector at exogenous cuts. In naturalistic fMRI, endogenous innovation adds unique cross-validated predictivity of $\Delta r_{\mathrm{CV}}{=}{+}0.013$ (p=0.005; $15/16$ subjects) in a pre-specified event network beyond audiovisual features and raw innovation, replicating on an independent film ($\Delta r{=}{+}0.005$, p=0.028), while the gate's discrete boundaries dissociably track sensory-change cortex ($16/16$ subjects). Together, these results establish **event-structured switching** as a principled mechanism for continuous world modeling, and show that its functional decomposition reflects the organization of the human brain.

1 INTRODUCTION

Continuous experience is neither a sequence of independent frames nor a uniformly evolving process. Within an event, the recent past is often the best guide to the future; at an event transition, the same persistence becomes a liability. A capable world model must therefore solve a stability–plasticity problem over time: *when should it keep its current state, and when should it replace the assumptions encoded by that state?*

Current sequence models obscure this decision. Transformers attend over long histories and selective state-space models (SSMs) learn input-dependent memory, but neither is ordinarily trained with an explicit, testable variable that marks a change of the active event model. Generic event-boundary detection does mark changes in video, yet is optimized for localization rather than as a mechanism that reorganizes predictive state. Predictive coding and Event Segmentation Theory provide the relevant account: a prediction failure is informative because it may indicate that the latent dynamics generating the observations have changed (Rao & Ballard, 1999; Zacks et al., 2007; Zacks, 2020; Wang et al., 2024). Prior work treats these elements separately. Sparse-update models decide *when*

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.

state should change but do not couple that decision to both the continuous-time memory kernel and a reusable event schema; surprise-based chunking requires observed error, which is unavailable during open-loop rollout; and naturalistic brain alignment can show correspondence without showing that the same mechanism improves prediction under matched controls.

Event cognition supplies the inductive bias, not a second problem. Event Segmentation Theory casts perception as an active event model replaced when prediction fails (Zacks et al., 2007; Zacks, 2020); predictive processing treats the error itself as the update trigger (Rao & Ballard, 1999; Friston, 2010); and naturalistic neuroimaging reports reusable schemas and a hierarchy of temporal receptive windows (Baldassano et al., 2018; Hasson et al., 2008; Lerner et al., 2011). We introduce HEG-VMAMBA, a hierarchical event-gated video world model that implements each principle as differentiable computation: calibrated innovation as the surprise signal, a learned event hazard as the switch, a schema bank as reusable templates, and nested clocks with Δ control as the timescale hierarchy. The model predicts future multimodal latents; the hazard both contracts stale state through the SSM discretization (**timescale adaptation**) and resets it toward a retrieved schema (**structured switching**). A counterfactual regularizer trains the gate to switch when reorganization improves the future, and a learned boundary prior reproduces the posterior gate during imagined rollout.

The question is therefore when a continuous world model should switch, not whether another detector can localize generic cuts. We first ask whether the same gate improves open-loop prediction under controlled regime shifts and on natural video, and whether it reorganizes state at schema-recoverable regime changes rather than at every high-error frame. Matched ablations then test whether those gains require learned timing, Δ modulation, and schema retrieval rather than generic surprise or perturbation. Making the verbal theory explicit also yields a prediction the theory does not by itself specify: segmentation decomposes into a graded surprise trajectory and a discrete boundary trajectory. We treat cortex as an external test of that decomposition—graded innovation over a higher-order event network, discrete gate boundaries over sensory-change cortex—after the predictive mechanism is in place, not as a parallel target.

Our contributions are:

- **Event-structured switching for continuous world models.** Calibrated multimodal innovation drives a learned event hazard that jointly expands an SSM discretization step, so stale history decays, and retrieves a schema that initializes the next state. Counterfactual utility trains the gate to switch when reorganization improves the future, and a learned prior supplies the same hazard when surprise is unobserved.
- **A quantitative account of state replacement.** We prove a monotone attenuation bound: the direct contribution of pre-boundary state shrinks with gate strength and SSM stability, and a fully open schema mixture removes it exactly.
- **The same variables are mechanistically necessary and neurally specific.** Matched controls show that learned timing, timescale adaptation, and schema retrieval—not generic surprise or perturbation—improve prediction under regime shift. In cortex the two control trajectories **doubly dissociate**: graded innovation tracks a higher-order event network (replicated on an independent film), while the gate's discrete boundaries track sensory-change cortex ($16/16$ subjects).

2 Related Work

Sparse state changes and temporal abstraction. GateL0RD biases recurrent models toward sparsely changing latent states (Gumbsch et al., 2021); THICK builds hierarchical world models from adaptive temporal abstractions (Gumbsch et al., 2024); SUGAR uses surprise-gated recurrence with counterfactual regularization (Humaidan et al., 2021); and SUNTA makes prediction error determine nested video chunks, addressing hierarchical collapse and missing open-loop surprise (Iiyama et al., 2026). Input-dependent SSMs likewise adapt memory to content: HID-SSM stacks hierarchical, input-conditioned timescales for long surgical video (Wu et al., 2025), whereas TIDES argues that making the discretization step Δ input-dependent harms extrapolation across sampling rates and instead routes selectivity through the continuous-time generator with an analytic Δ (Soydan et al., 2026). Our focus is complementary: the gate directly controls the continuous-time discretization of a selective SSM, retrieves an explicit event schema, exposes modality-specific innovations, and is evaluated by a matched mechanistic decomposition across machine tasks and human brain data;

we stress-test the learned Δ under sampling-rate and timing shifts against the TIDES concern in Appendix G.

Video prediction and generic event boundaries. Self-supervised predictive objectives learn semantic dynamics without reconstructing stochastic pixels (van den Oord et al., 2018; Assran et al., 2023), and video backbones such as VideoMAE and VideoMamba give strong dense-temporal representations (Tong et al., 2022; Li et al., 2024). Generic Event Boundary Detection formalizes taxonomy-free boundaries that agree with human segmentation (Shou et al., 2021), and UBoCo learns them by unsupervised contrastive learning (Kang et al., 2022). Recent detectors improve along complementary axes: DyBDet allocates sample-dependent sub-networks for efficient dynamic inference (Zheng et al., 2024), DiffGEBD casts boundaries as denoising diffusion to capture annotator diversity (Hwang et al., 2025), and ESTimator detects online in streaming video by thresholding a causal predictor's error (Jung et al., 2025)—the last, like HEG-VMAMBA, grounded in Event Segmentation Theory. These optimize localization quality or latency; unlike a boundary head trained only to localize, HEG-VMAMBA uses the boundary variable inside the recurrent state update, so we evaluate both boundary quality and whether the same gate improves prediction under distribution shift. Appendix E positions HEG-VMAMBA against these detectors by supervision, causality, and whether the boundary variable updates state.

Predictive processing, event cognition, and naturalistic brain alignment. Predictive processing holds that neural systems maintain hypotheses and update them from error (Friston, 2010; Rao & Ballard, 1999), and Event Segmentation Theory casts experience as a sequence of event models with prediction failures at transitions (Zacks & Tversky, 2001; Zacks et al., 2007; Zacks, 2020). Naturalistic neuroimaging reveals event-structured representations, schemas, and a hierarchy of temporal receptive windows (Baldassano et al., 2017; 2018; Hasson et al., 2008; Lerner et al., 2011), and encoding models and RSA interface learned representations with fMRI (Kriegeskorte et al., 2008; Naselaris et al., 2011). Because correlation with a feature space does not establish a shared computation, the neural analyses are incremental and cross-validated: they test whether the learned variables explain variance beyond pretrained audiovisual features, low-level stimulus change, and the innovation from which the gate is computed. HEG-VMAMBA is treated as a computational hypothesis that yields **dissociable**, a-priori neural predictions.

3 HIERARCHICAL EVENT-GATED VIDEO MAMBA

HEG-VMAMBA implements event-cognition theory as differentiable computation (Fig. 1). Table 3 (appendix) lists the correspondence: each cognitive principle maps to an architectural component, and each component yields a falsifiable machine-learning or neural consequence.

3.1 PROBLEM SETTING

Let $o_t = (v_t, a_t)$ denote a synchronized audiovisual stream and $x_t \in \mathbb{R}^d$ its latent representation. We seek a model that predicts future latents over multiple horizons, represents stable regimes with persistent state, reorganizes that state at learned event boundaries, and continues the same abstraction during open-loop rollout; it emits three interpretable trajectories per layer ℓ: **calibrated innovation** s_t^ℓ, **event hazard** $g_t^\ell \in [0, 1]$, and **event state** h_t^ℓ. Visual and audio encoders f_v, f_a produce synchronized features, and lightweight temporal adapters capture local dynamics before fusion:

$$z_t^v = \phi_v\big(f_v(v_{t-w:t+w})\big), \qquad z_t^a = \phi_a\big(f_a(a_{t-w:t+w})\big), \tag{1}$$

$$x_t = W_f[z_t^v; z_t^a]. \tag{2}$$

The main experiments freeze the large foundation encoders and train the temporal adapters, fusion layer, event-gated SSM, predictors, and schema memory. A smaller end-to-end visual front end is included as a robustness experiment so that conclusions are not restricted to CLIP/Whisper feature geometry (Radford et al., 2021; 2023).

3.2 CALIBRATED PREDICTIVE INNOVATION

Predictive-processing accounts treat the *precision-weighted* prediction error, rather than the raw error, as the quantity that indicates a change of latent dynamics (Friston, 2010; Rao & Ballard, 1999). A

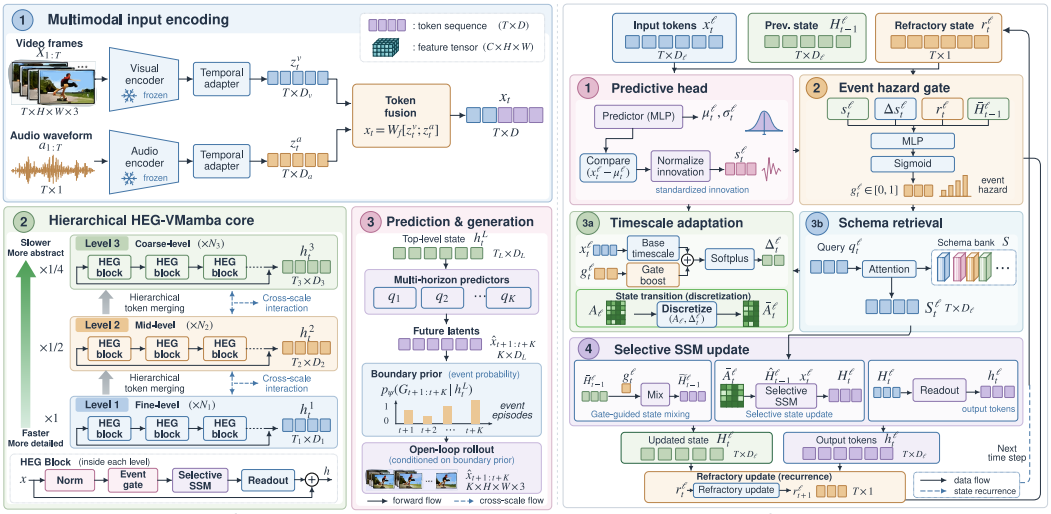


FIGURE 1: **HEG-VMAMBA architecture.**

deterministic Euclidean error conflates a genuinely unexpected transition with observations that are merely difficult or noisy. At layer ℓ, a predictive head therefore parameterizes a diagonal Gaussian over the next layer input,

$$p_\theta(x_t^\ell \mid H_{t-1}^\ell) = \mathcal{N}(\mu_t^\ell, \mathrm{diag}((\sigma_t^\ell)^2)), \tag{3}$$

where H_{t-1}^ℓ is the recurrent SSM state. The **innovation** is the dimension-normalized negative log likelihood,

$$s_t^\ell \;=\; \frac{1}{2d_\ell}\sum_{j=1}^{d_\ell}\Big[\frac{(x_{t,j}^\ell - \mu_{t,j}^\ell)^2}{(\sigma_{t,j}^\ell)^2} + 2\log\sigma_{t,j}^\ell\Big]. \tag{4}$$

We maintain stop-gradient robust exponential moments m_t^ℓ, v_t^ℓ and standardize s_t^ℓ into $\tilde{s}_t^\ell$. For multimodal attribution the gate receives joint, visual, and audio innovations $(\tilde{s}_t^{\ell,j}, \tilde{s}_t^{\ell,v}, \tilde{s}_t^{\ell,a})$ and their temporal differences, so a boundary can be attributed to visual, speech/narrative, or cross-modal change.

3.3 A LEARNED, UTILITY-GROUNDED EVENT HAZARD

Event Segmentation Theory holds that a boundary *replaces* the active event model rather than only marking a change (Zacks et al., 2007; Zacks, 2020). The **hazard** below is the differentiable analogue of that replacement: a posterior boundary probability, relaxed with a Hard-Concrete distribution (Louizos et al., 2018), that gates both forgetting and reset rather than a passive readout.

$$r_t^\ell = \rho_\ell r_{t-1}^\ell + g_{t-1}^\ell, \tag{5}$$

$$q_\phi(g_t^\ell \mid o_{\le t}) = \mathrm{HardConcrete}\Big(f_g^\ell\big([\tilde{s}_t^{\ell,j}, \tilde{s}_t^{\ell,v}, \tilde{s}_t^{\ell,a}, \Delta\tilde{s}_t^\ell, \bar{H}_{t-1}^\ell]\big) - \eta_\ell r_t^\ell\Big). \tag{6}$$

$\bar{H}_{t-1}^\ell$ is the mean of the previous SSM state H_{t-1}^ℓ across its state axis. The refractory variable r_t^ℓ prevents sustained uncertainty from saturating boundaries. Unlike a fixed threshold, the gate can treat some large innovations as nuisance while treating smaller but consequential ones as transitions.

Surprise alone does not guarantee that switching is useful. On a random subset of training steps, we compute short counterfactual branches with the gate clamped to keep ($g = 0$) or switch ($g = 1$), yielding future predictive losses $L_{t:t+K}^{\text{keep}}$ and $L_{t:t+K}^{\text{switch}}$. Their stop-gradient difference defines a **switching advantage**,

$$u_t^\ell = \mathrm{sg}\big(L_{t:t+K}^{\text{keep}} - L_{t:t+K}^{\text{switch}}\big), \qquad \mathcal{L}_{\text{cf}} = \mathrm{BCE}\big(g_t^\ell, \mathrm{sigmoid}(u_t^\ell/T_u)\big). \tag{7}$$

The gate is therefore trained to switch when reorganization is expected to improve future prediction, rather than when the instantaneous error is large.

3.4 TIMESCALE ADAPTATION AND STRUCTURED SCHEMA SWITCHING

For each layer, Mamba (Gu & Dao, 2024) parameterizes a stable diagonal continuous-time transition $A_\ell = -\exp(A_{\ell,\log})$. The gate increases the input-dependent discretization step,

$$\Delta_t^\ell = \mathrm{softplus}\big(\Delta_{\mathrm{base}}^\ell(x_t^\ell) + \gamma_\ell g_t^\ell\big)\,. \tag{8}$$

A larger step accelerates the decay of components encoded by A_ℓ. The new event state is not initialized from zero. A layer-specific **schema bank** contains keys $K_\ell \in \mathbb{R}^{M\times d_\ell}$ and state-valued memories $V_\ell \in \mathbb{R}^{M\times(d_{\mathrm{ssm}}d_\ell)}$. The current input retrieves

$$\pi_t^\ell = \mathrm{softmax}(q_\ell(x_t^\ell)K_\ell^\top/\sqrt{d_\ell}), \tag{9}$$

$$S_t^\ell = \mathrm{reshape}(\pi_t^\ell V_\ell), \tag{10}$$

$$\widetilde{H}_{t-1}^\ell = (1-g_t^\ell)H_{t-1}^\ell + g_t^\ell S_t^\ell. \tag{11}$$

The selective SSM then updates

$$H_t^\ell = \exp(A_\ell\Delta_t^\ell)\odot\widetilde{H}_{t-1}^\ell + \overline{B}_t^\ell x_t^\ell, \qquad h_t^\ell = C_t^\ell H_t^\ell + D_\ell x_t^\ell. \tag{12}$$

This separates two functions of event switching: Δ_t controls how quickly the recurrent dynamics forget (**timescale adaptation**), while the schema mixture controls what replaces the stale event state (**structured switching**).

Proposition 1 (Controlled attenuation of the direct pre-boundary contribution) *Assume $A_\ell = -\mathrm{diag}(a_1,\dots,a_n)$ with $a_i \ge a_{\min} > 0$, and let $\Delta(g) = \mathrm{softplus}(d+\gamma g)$ with $\gamma \ge 0$. In the update of Eq. (12), the contribution of the pre-boundary state satisfies*

$$\big\|\exp(A_\ell\Delta(g))(1-g)H_{t-1}^\ell\big\|_2 \le (1-g)e^{-a_{\min}\Delta(g)}\|H_{t-1}^\ell\|_2. \tag{13}$$

The bound is non-increasing in g; at $g=1$ the old-state contribution is exactly zero. For $g<1$, the operation is a controlled soft replacement, not an instantaneous exact reset caused by the matrix exponential alone.

The proof and a non-diagonal extension based on the logarithmic norm are given in Section B.

3.5 LEARNING NESTED TEMPORAL ABSTRACTIONS

Hierarchy is obtained from learned temporal structure rather than from hand-chosen per-layer thresholds. Deeper boundaries are encouraged to be sparser and to be supported by accumulated lower-layer evidence. Define the probability that at least one lower-level boundary occurred in the recent window,

$$c_t^{\ell-1} = 1 - \prod_{j=0}^{w_\ell-1}(1-g_{t-j}^{\ell-1}). \tag{14}$$

The nesting and rate objectives are

$$\mathcal{L}_{\mathrm{nest}} = \sum_{\ell>1,t}[g_t^\ell - c_t^{\ell-1}]_+, \qquad \mathcal{L}_{\mathrm{rate}} = \sum_\ell\left(\frac{1}{T}\sum_t g_t^\ell - \pi_\ell\right)^2, \tag{15}$$

where $\pi_1 > \pi_2 > \cdots$ is a weak target ordering rather than a fixed boundary schedule. Hierarchy is evaluated through empirical event duration, autocorrelation, cross-layer nesting, and alignment to independently annotated fine and coarse boundaries.

3.6 PREDICTIVE OBJECTIVE AND OPEN-LOOP BOUNDARY PRIOR

From h_t^L, horizon-specific predictors q_k predict target latents x_{t+k}. We use a multi-horizon joint-embedding objective,

$$\mathcal{L}_{\mathrm{pred}} = \sum_{k=1}^{K} w_k\ \mathrm{InfoNCE}\left(q_k(h_t^L), \mathrm{sg}(x_{t+k}), \mathcal{N}_{t,k}\right), \tag{16}$$

plus the Gaussian innovation loss from Eq. (4). Schema collapse is discouraged with an entropy-balanced utilization loss and a key-diversity penalty.

Observed surprise is unavailable during imagined rollout. A **boundary prior** $p_\psi(g_{t+1:t+K} \mid h_t)$ is therefore trained to predict the posterior hazards produced under teacher forcing,

$$\mathcal{L}_{\text{prior}} = \sum_{k=1}^{K} \text{BCE}\left(\hat{g}_{t+k}, \text{sg}(g_{t+k})\right). \tag{17}$$

At observation time the posterior gate in Eq. (6) is used; at rollout time the prior gate controls Δ and schema retrieval. Open-loop prediction therefore does not depend on observed surprise.

The complete objective is

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_{\text{nll}}\mathcal{L}_{\text{nll}} + \lambda_{\text{cf}}\mathcal{L}_{\text{cf}} + \lambda_{\text{prior}}\mathcal{L}_{\text{prior}} \\ + \lambda_{\text{sparse}}\mathbb{E}[g] + \lambda_{\text{nest}}\mathcal{L}_{\text{nest}} + \lambda_{\text{rate}}\mathcal{L}_{\text{rate}} + \lambda_{\text{schema}}\mathcal{L}_{\text{schema}}.\end{aligned} \tag{18}$$

4 Experiments

4.1 Evidence hierarchy and common protocol

The experiments address four questions, and neural tests are considered only after the predictive mechanism is established: **(Q1) predictive utility**—whether event-gated reorganization improves long-horizon and OOD prediction; **(Q2) event validity**—whether gates recover intrinsic and human-perceived boundaries rather than arbitrary high-error frames; **(Q3) mechanism identity**—whether gains arise from learned timing, Δ_t modulation, or schema retrieval rather than generic perturbation; and **(Q4) external alignment**—whether the learned variables explain brain dynamics beyond stimulus features and raw error.

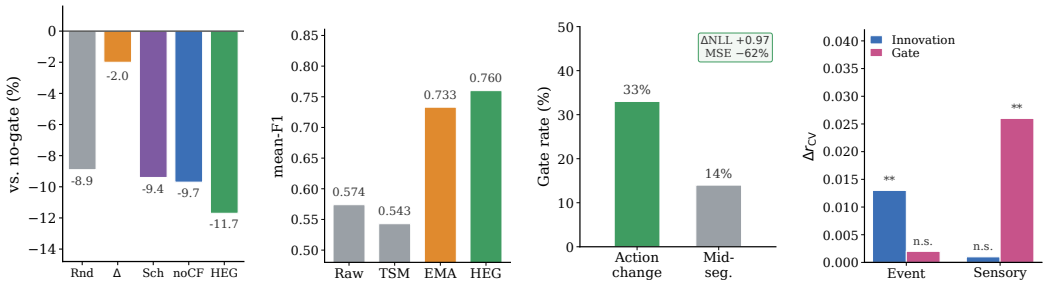
All learned models are matched in hidden size, parameter count, pretraining examples, optimizer steps, and search budget; each primary result spans $\geq$5 seeds, with per-seed values or standard deviations and paired tests, and all thresholds frozen on validation. Reported controlled baselines are a no-gate Mamba, GateL0RD, THICK, and a raw-surprise threshold; video comparisons use same-feature UBoCo/TSM and EMA readouts, with published detectors treated as positioning rather than a same-pipeline leaderboard. Mechanistic controls add a trained no-gate model, rate/amplitude/autocorrelation-matched random and shifted gates, oracle keep/switch policies, and Δ-only / schema-only / no-counterfactual ablations. Figure 2 summarizes the main results across all four questions.

4.2 Controlled regime-shift world modeling

We construct a controlled continuous-dynamics benchmark in which a latent context selects one of several transition laws; boundaries change the active dynamics, while independent nuisance interventions alter appearance, camera, or audio without changing the regime. Training uses bounded event durations; test sets separately vary duration, transition frequency, noise, missing modalities, and unseen regime compositions. Models predict latent targets and observations for up to 256 open-loop steps (multi-step NLL, calibrated uncertainty, rollout survival, boundary AP/F1, detection delay, regime decoding, OOD). The primary test is whether the learned switch improves subsequent prediction after controlling for the same prediction error, not merely whether surprise increases at a transition.

Table 1: **Controlled RegimeWorld** under partial observability (matched $\sim$134k params, 5 seeds; mean $\pm$ s.d.). Open-loop NLL@16, MSE@64, and NLL@16 on a modality-conflict OOD split. HEG-VMamba (full) attains the lowest in-distribution and OOD NLL; MSE@64 saturates. Within-model ablations: Table 2.

Model	16-step NLL ↓	64-step MSE ↓	OOD NLL ↓
Mamba predictor (no gate)	16.4 ± 6.4	1.20 ± 0.30	17.3 ± 6.7
GateL0RD (Gumbsch et al., 2021)	14.8 ± 7.0	0.89 ± 0.21	15.7 ± 7.6
THICK (Gumbsch et al., 2024)	14.5 ± 9.1	0.90 ± 0.23	15.5 ± 8.8
Raw-surprise threshold	11.5 ± 3.5	1.15 ± 0.33	14.3 ± 4.9
HEG-VMamba (full)	$\mathbf{7.3 \pm 1.8}$	1.00 ± 0.28	$\mathbf{8.4 \pm 2.3}$



(a) Controlled utility (b) Zero-shot GEBD (c) KTH prediction (d) fMRI dissociation

FIGURE 2: **Four-question evidence.** Details in the appendix (Tables 1, 2 and 4 and Fig. 15).

Across held-out regimes HEG-VMAMBA improves open-loop prediction by -11.5% ($5/5$ seeds, p=0.007) over a retrained no-gate control, reaches boundary quality $\mathrm{AP}_{\mathrm{gate}}$=0.52, $\mathrm{AP}_{\mathrm{innov}}$=0.58 vs. raw 0.40, null 0.13 ($5/5$), and retains an OOD advantage of $+6.7$–12.6% in $5/5$ seeds across all four OOD families under unseen durations and compositions. At matched capacity ($\sim$134k params, identical splits/seeds) it attains lower open-loop NLL than GateL0RD, THICK, a no-gate Mamba, and a raw-surprise gate, both in distribution and on the modality-conflict OOD split ($5/5$ seeds, $p < 0.05$; Table 1). Nuisance interventions increase raw feature change without producing comparable persistent gates.

4.3 GENERIC AND ONLINE EVENT-BOUNDARY DISCOVERY

We then evaluate transfer to human-perceived changes in natural video. On Kinetics-GEBD (Shou et al., 2021) no boundary labels are used during representation learning; a single validation-set calibration maps hazards to boundaries and is then fixed. The online protocol uses only past and current input; we report detection latency and compute.

The internal HEG-VMAMBA hazard attains Rel.Dis. **mean-F1** 0.760 (zero-shot 5-seed ensemble on 17,911 val clips with extracted features; per-seed 0.754 ± 0.032) ($+0.216$ over an unsupervised UBoCo/TSM readout and $+0.186$ over raw feature-change (identical features/split/protocol)), above the same-feature unsupervised readouts and at the published supervised online ESTimator average (0.748; Jung et al., 2025) without boundary supervision, at online latency $\approx$0.01 s median causal delay (≤ 1 window). A dedicated in-domain unsupervised detector (UBoCo) reports mean-F1 0.867 (F1@0.05 0.703; Kang et al., 2022); we do not match that number under zero-shot transfer. A tuned EMA-surprise detector is competitive at the strict tolerance (full leaderboard in Table 4, appendix); the learned gate is better calibrated and has near-zero latency. On held-out Sherlock annotations it exceeds raw innovation by AP 0.18–0.26 vs. raw 0.11, perm. z up to $+4.1$ (innovation)/$+6.0$ (gate) under one-to-one matching, tolerance curves, and count-matched nulls (thresholds frozen from validation). The same untrained gate transfers to TAPOS Olympic sub-action boundaries (Shao et al., 2020; Shou et al., 2021) at zero-shot mean-F1 0.491 (flow-only zero-shot 5-seed ensemble, N=771; above the supervised TransParser 0.474, below unsupervised PA 0.543), above supervised TransParser, TCN, CTM, and ISBA and the unsupervised SceneDetect baseline, and below unsupervised PA. The model is pretrained on FineVideo (Farré et al., 2024), so its audio stream is out of domain on silent or crowd-noise Olympic footage; removing that stream increases mean-F1 by 0.02–0.04. This is a transfer result, not a TAPOS state-of-the-art comparison.

The same gate is then tested for open-loop *prediction*, which distinguishes an intervention on state from a boundary readout. On KTH Actions (Schuldt et al., 2004) (official person split; identical flow+CLIP features (Radford et al., 2021); 5 seeds; no test tuning) the learned switch lowers boundary-conditioned open-loop error at discrete recurring action-regime starts by open-loop NLL $+0.97\pm0.13$ and latent MSE $-62\pm16\%$ at recurring action-regime starts ($5/5$ seeds, $5/5$ above a circular-shift null). The gate activates on 33% of action changes versus 14% of mid-segment frames and outperforms a matched no-gate control, while remaining null at intra-action periodic phase. The prediction improvement therefore appears at recurring, schema-recoverable regime changes rather than at generic periodicity or exogenous cuts.

4.4 MECHANISTIC DECOMPOSITION

Table 2 separates the two gate actions and their timing; the key comparison is full HEG-VMAMBA versus a from-scratch no-gate model, with inference-time removal reported only as an intervention on a trained system. Matched-null gates retain amplitude, sparsity, and autocorrelation while shifting their temporal locations.

TABLE 2: **Mechanism-identification matrix** on REGIMEWORLD under partial observability (obs. fraction 0.35; 5 seeds). Rows above the rule are trained independently; † is an inference-time timing intervention on the trained full model. Predictive utility is the paired change in aggregate open-loop (K=10) rollout MSE vs. the trained no-gate control (negative = better; $^{*}p<0.05$, $^{**}p<0.01$). Full HEG-VMAMBA is the strongest independently-trained configuration.

Variant	Timing	Δ	Schema	Utility	Role
Trained no-gate	×	×	×	0 (ref.)	baseline
Random matched gate	×	✓	✓	$-8.9\%^{*}$	matched perturbation
Δ-only	✓	✓	×	-2.0%	timescale only
Schema-only	✓	×	✓	-9.4%	schema only
No counterfactual loss	✓	✓	✓	$-9.7\%^{*}$	no utility loss
HEG-VMamba full	✓	✓	✓	$\mathbf{-11.7\%}^{**}$	full model
Raw-surprise override†	×	✓	✓	$-11.8\%^{**}$	inference-time intervention

Full HEG-VMAMBA improves over a from-scratch no-gate model by -22% (oracle -33%) at event onsets ($5/5$ seeds, p=0.001), and learned timing improves by 2.9–3.2% ($5/5$, $p<0.003$) over rate-, amplitude-, and autocorrelation-matched gates. Timescale control carries the boundary-detection signal (Δ-only AP 0.57 ($\approx$ full 0.52, no-gate 0.29)) and contributes in combination: the full model exceeds schema-only in aggregate rollout under partial observability. Schema retrieval provides a structured, content-bearing reset (-58% vs. current-input, -1.7% vs. zero-erasure, -1.5% vs. wrong-regime ($5/5$, $p\leq 0.009$)), outperforming a current-input reset and, under partial observability, both a wrong-regime schema and a zero-state reset, because the post-boundary latent cannot be recovered from the current observation. Under full observability, schema content matches a zero reset; hiding the latent in both modalities widens the schema-versus-zero gap to $+2.3\%$ ($5/5$, p=0.026). We also report schema entropy, effective slots ($10.3/16$ effective slots, regime–slot MI 0.41 (23% of H), wrong-schema swap $+1.5\%$ (p=0.009)), gate calibration (ECE 0.031, Brier 0.076, NLL 0.275 vs. Platt-matched surprise 0.074/0.079/0.333 ($5/5$, p=0.003)), and cross-layer nesting. A factorial long-form and modality analysis (Section F) yields a long-horizon gain of -41% open-loop error versus a copy-last baseline across horizons 1–16 ($5/5$ seeds; the gate-specific contribution is $\approx$0). On natural video, vision carries next-state prediction and audio is redundant without degrading the visual model; in a controlled complementary-modality setting where the streams observe disjoint latent halves, audiovisual fusion improves single-step NLL by 0.26 nats over vision-only ($5/5$, $p=9\times10^{-5}$).

4.5 NATURALISTIC FMRI AS AN EXTERNAL MECHANISTIC TEST

We test a frozen world model on Sherlock (16 participants from the public release; $\sim$50 min of continuous audiovisual narrative; Chen et al., 2017) and replicate with the same frozen checkpoint and pipeline on an independent movie-fMRI dataset (the "(500) Days of Summer" subset of the Naturalistic Neuroimaging Database; Aliko et al., 2020; OpenNeuro `ds002837`; distinct film, scanner, TR, and preprocessing). No fMRI signal trains the model. Encoding uses contiguous nested folds with train-only scaling/reduction and an HRF buffer; joint models add low-level stimulus and audiovisual-encoder features, then calibrated innovation, then the residualized gate and state, with autocorrelation-preserving nulls (circular shift, block permutation, phase randomization; full protocol in Section C).

After these controls, endogenous innovation adds **unique** cross-validated predictivity of $\Delta r_{\mathrm{CV}}=+0.013$ (p=0.005; $15/16$ subjects) in a pre-specified posterior-medial event network (a-priori ROIs), absent in a V1/A1 sensory control and exceeding a human-boundary regressor; it reproduces across held-out narrative halves and **replicates** on the independent film in the event network's inter-subject shared response ($\Delta r=+0.005$, p=0.028, event-specific), although the frozen learned-calibration increment does not transfer to this out-of-distribution film. The two control trajec-

tories **dissociate**: the gate's discrete boundaries—the same relative-peak signal used for GEBD and human-boundary detection—add unique predictivity in the sensory-change network ($\Delta r_{\mathrm{CV}}{=}{+}0.021$–$0.031$, perm. $p{<}0.003$, $z{=}{+}5.9$–7.5, $16/16$ subjects (robust to a 60-component audiovisual control)) and none in the event network, complementary to graded innovation. Higher-order surprise updating and bottom-up transition detection therefore align with separable, system-specific signals, as predicted by the model's decomposition. The model's event-state geometry shows no incremental partial RSA (near zero, opposite in sign to the human ceiling; Section C), so the correspondence is in transient boundary and innovation dynamics rather than a persistent schema geometry. These findings support incremental computational correspondence, not representational identity or a causal biological mechanism.

5 Conclusion and Discussion

Innovation as a control signal. HEG-VMamba uses prediction failure to change computation rather than only to increase a loss: the learned hazard shortens the memory kernel and selects a structured replacement state, and counterfactual utility trains the gate to switch when reorganization improves the future. The controlled decomposition indicates that the dominant effect is a well-timed state reset, detected by timescale-sensitive innovation and executed by the gate. Schema content is redundant when observations reveal the post-boundary state and necessary under partial observability. The fMRI analysis shows a **double dissociation**: graded innovation tracks a higher-order event network, whereas the gate's discrete boundaries track sensory-change cortex, so the surprise signal and the boundary detector correspond to separable computations. On natural video the effect depends on regime type. The gate is an effective detector at generic Kinetics-GEBD cuts (exogenous novelty) without lowering open-loop error; it improves open-loop prediction at discrete, schema-recoverable regimes (KTH action starts, open-loop NLL $+0.97{\pm}0.13$ and latent MSE $-62{\pm}16\%$ at recurring action-regime starts ($5/5$ seeds, $5/5$ above a circular-shift null)); and it is null on slow naturalistic scene recurrence, for which stream-specific adaptation remains open. Matched no-gate, timing, rate, amplitude, and autocorrelation controls rule out generic perturbation and simple change detection.

Interpretation and limitations. Event-cognition theory is used as architectural motivation; neural evaluation is used only after the mechanism shows predictive value, valid timing, and specificity. Incremental fMRI tests correspondence, not that cortex implements the same equations. The model inherits encoder biases, a finite schema dictionary, a rollout prior that stays calibrated but does not anticipate the timing of distant exogenous boundaries, and an observational rather than action-conditioned setting. We probe these directly: Appendix G stress-tests the learned Δ under irregular sampling and rate shifts; Appendix H tracks prior-versus-posterior calibration over open-loop rollouts to $K{=}120$; Appendix I quantifies the compute and latency overhead of the counterfactual branch and gate; and Appendix J sweeps schema count, counterfactual horizon, and the rate target. These limits bound the claims without removing the question of when predictive state should persist or reorganize.

Outlook. HEG-VMamba converts calibrated multimodal innovation into hierarchical event switching—adapting SSM timescales and retrieving reusable schemas—and yields a testable principle for continuous world modeling together with a dissociable neural prediction derived from that principle.

REPRODUCIBILITY STATEMENT

This statement points to the artefacts rather than restating them. Anonymous source code, configuration files, environment locks, data manifests, checkpoint hashes, and seed- or participant-level result records are submitted as supplementary material; every numerical table is generated from those records. The model, gate, and training objective are specified in Section 3, with a one-step update in Algorithm 1 and the theory-to-component map in Section A. Assumptions and a complete proof of the attenuation bound are in Section B. Dataset construction, feature extraction, official and held-out splits, boundary matching, and fMRI preprocessing—including nested cross-validation and autocorrelation-preserving nulls—are in Section C; public-benchmark coverage is listed in Sections D and L. Controlled follow-ups (ablations, Δ stress tests, prior calibration, compute, and sensitivity) are in Sections F to J, and the neural protocol is expanded in Section K.

AI USE STATEMENT

In this work, we used generative AI tools to edit the manuscript for readability (wording, grammar, and local clarity). We have not used generative AI tools to generate synthetic data, develop theoretical models or conceptual frameworks, formulate or prove mathematical claims, propose or refine hypotheses, design or give feedback on methodology or experiments, implement methods, clean or reformat datasets, support qualitative analysis, or interpret results. We reviewed every AI-suggested edit and accepted only those that preserved the authors' intended claims, citations, and numbers. We take responsibility for the final content of this work, including text, citations, and artifacts.

REFERENCES

Sarah Aliko, Jiawen Huang, Florin Gheorghiu, Stefanie Meliß, and Jeremy I. Skipper. A naturalistic neuroimaging database for understanding the brain using ecological stimuli. *Scientific Data*, 7:347, 2020. doi: 10.1038/s41597-020-00680-2.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Christopher Baldassano, Janice Chen, Asieh Zadbood, Jonathan W. Pillow, Uri Hasson, and Kenneth A. Norman. Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3):709–721.e5, 2017. doi: 10.1016/j.neuron.2017.06.041.

Christopher Baldassano, Uri Hasson, and Kenneth A. Norman. Representation of real-world event schemas during narrative perception. *Journal of Neuroscience*, 38(45):9689–9699, 2018. doi: 10.1523/JNEUROSCI.0251-18.2018.

Janice Chen, Yuan Chang Leong, Christopher J. Honey, Chung H. Yong, Kenneth A. Norman, and Uri Hasson. Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, 20(1):115–125, 2017. doi: 10.1038/nn.4450.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095889.

Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Finevideo. Hugging Face dataset, 2024. URL https://huggingface.co/datasets/HuggingFaceFV/finevideo. CC-BY.

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010. doi: 10.1038/nrn2787.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024. Presented at the First Conference on Language Modeling (COLM), 2024.

Christian Gumbsch, Martin V. Butz, and Georg Martius. Sparsely changing latent states for prediction and planning in partially observable domains. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Christian Gumbsch, Noor Sajid, Georg Martius, and Martin V. Butz. Learning hierarchical world models with adaptive temporal abstractions from discrete latent dynamics. In *International Conference on Learning Representations*, 2024.

Uri Hasson, Eunice Yang, Ignacio Vallines, David J. Heeger, and Nava Rubin. A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*, 28(10):2539–2550, 2008. doi: 10.1523/JNEUROSCI.5487-07.2008.

Dania Humaidan, Sebastian Otte, Christian Gumbsch, Charley Wu, and Martin V. Butz. Latent event-predictive encodings through counterfactual regularization, 2021. arXiv:2105.05894.

Jaejun Hwang, Dayoung Gong, Manjin Kim, and Minsu Cho. Generic event boundary detection via denoising diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14084–14094, 2025.

Tomoshi Iiyama, Masahiro Suzuki, and Yutaka Matsuo. SUNTA: Hierarchical video prediction with surprise-based chunking, 2026. arXiv:2607.02087.

Hyungrok Jung, Daneul Kim, Seunggyun Lim, Jeany Son, and Jonghyun Choi. Online generic event boundary detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13741–13750, 2025.

Hyolim Kang, Jinwoo Kim, Taehyun Kim, and Seon Joo Kim. UBoCo: Unsupervised boundary contrastive learning for generic event boundary detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20073–20082, 2022. doi: 10.1109/CVPR52688.2022.01944.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Yulia Lerner, Christopher J. Honey, Lauren J. Silbert, and Uri Hasson. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31(8): 2906–2915, 2011. doi: 10.1523/JNEUROSCI.3684-10.2011.

Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. VideoMamba: State space model for efficient video understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 15084 of *Lecture Notes in Computer Science*, pp. 237–255, 2024. doi: 10.1007/978-3-031-73347-5_14.

Christos Louizos, Max Welling, and Durk P. Kingma. Learning sparse neural networks through L_0 regularization. In *International Conference on Learning Representations*, 2018.

Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fmri. *NeuroImage*, 56(2):400–410, 2011. doi: 10.1016/j.neuroimage.2010.07.073.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023.

Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999. doi: 10.1038/4580.

Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local SVM approach. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, pp. 32–36, 2004. doi: 10.1109/ICPR.2004.1334462.

Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Intra- and inter-action understanding via temporal action parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 730–739, 2020. doi: 10.1109/CVPR42600.2020.00081.

Mike Zheng Shou, Stan Weixian Lei, Weiyao Wang, Deepti Ghadiyaram, and Matt Feiszli. Generic event boundary detection: A benchmark for event segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8075–8084, 2021. doi: 10.1109/ICCV48922.2021.00797.

Taylan Soydan, Miguel A. Bessa, Dirk Mohr, and Rui Barreira. TIDES: Implicit time-awareness in selective state space models, 2026. arXiv:2605.09742.

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*, 2022. arXiv:2203.12602.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018.

Yuxi Candice Wang, R. Alison Adcock, and Tobias Egner. Toward an integrative account of internal and external determinants of event segmentation. *Psychonomic Bulletin & Review*, 31(2):484–506, 2024. doi: 10.3758/s13423-023-02375-2.

Haoyang Wu, Tsun-Hsuan Wang, Mathias Lechner, Ramin Hasani, Jennifer A. Eckhoff, Paul Pak, Ozanan R. Meireles, Guy Rosman, Yutong Ban, and Daniela Rus. Holistic surgical phase recognition with hierarchical input dependent state space models, 2025. arXiv:2506.21330.

Jeffrey M. Zacks. Event perception and memory. *Annual Review of Psychology*, 71:165–191, 2020. doi: 10.1146/annurev-psych-010419-051101.

Jeffrey M. Zacks and Barbara Tversky. Event structure in perception and conception. *Psychological Bulletin*, 127(1):3–21, 2001. doi: 10.1037/0033-2909.127.1.3.

Jeffrey M. Zacks, Nicole K. Speer, Khena M. Swallow, Todd S. Braver, and Jeremy R. Reynolds. Event perception: A mind-brain perspective. *Psychological Bulletin*, 133(2):273–293, 2007. doi: 10.1037/0033-2909.133.2.273.

Ziwei Zheng, Lijun He, Le Yang, and Fan Li. Fine-grained dynamic network for generic event boundary detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

This appendix follows the main-text argument. Section A is the theory–component map and layer update; Section B proves the attenuation bound; Section C states the protocols; Section D gives public-benchmark tables and curves; Section E positions recent detectors and selective SSMs; Section F collects long-form, modality, and qualitative ablations; Section G stress-tests the learned Δ; Section H tracks prior-versus-posterior calibration; Section I reports compute and latency; Section J sweeps schema count, horizon, and rate; Section K expands the neural analyses; Section L lists the reporting facts.

A Theory-to-component map and layer update

Each row of Table 3 records one correspondence from Section 3: a cognitive principle becomes a differentiable component, which yields a consequence we test. Algorithm 1 is the corresponding one-step update.

TABLE 3: **Theory → component → prediction.** Each row maps an event-cognition principle to a differentiable component and a consequence that can be confirmed or refuted.

Cognitive / neural principle	**HEG component**	**Falsifiable consequence**
Prediction error signals a change of latent dynamics (Rao & Ballard, 1999; Zacks et al., 2007)	Calibrated innovation s_t (Section 3.2)	Uniquely predicts event-network BOLD, event-specific and replicated (Section 4.5)
The active event model is *replaced* at boundaries (Zacks, 2020)	Learned event hazard g_t (Section 3.3)	Detects human/GEBD boundaries (mean-F1 0.760); onsets track sensory-change cortex
Update only when it improves the future (Humaidan et al., 2021)	Counterfactual utility regularizer (Section 3.3)	Switching lowers open-loop error under structured regime shifts (-11.7%, Table 2)
Reusable event schemas (Baldassano et al., 2018)	Schema bank / structured reset (Section 3.4)	Schema content is necessary under partial observability (vs. current-input / zero / wrong-regime)
Cortical hierarchy of temporal receptive windows (Hasson et al., 2008; Lerner et al., 2011)	Nested clocks + Δ timescale control (Sections 3.4 and 3.5)	Timescale control carries the boundary signal (Δ-only AP 0.57 vs. full 0.52)
Observed surprise is unavailable in imagination	Open-loop boundary prior (Section 3.6)	A learned prior sustains gating during rollout (Section 3.6)

Algorithm 1 One observed-sequence update of a HEG-VMamba layer

Require: input x_t, previous SSM state H_{t-1}, refractory state r_{t-1}
1: predict (μ_t, σ_t) from H_{t-1} and compute calibrated innovation s_t
2: sample/relax posterior hazard g_t from innovation, state, and refractory features
3: retrieve event schema S_t and form $\widetilde{H}_{t-1} = (1 - g_t)H_{t-1} + g_t S_t$
4: compute $\Delta_t = \mathrm{softplus}(\Delta_{\mathrm{base}}(x_t) + \gamma g_t)$
5: update $H_t = \exp(A\Delta_t) \odot \widetilde{H}_{t-1} + \overline{B}_t x_t$
6: update refractory state; return output h_t, s_t, g_t, schema assignment, and H_t

B Proof of the attenuation proposition

For diagonal $A = -\operatorname{diag}(a_1, \ldots, a_n)$ with $a_i \geq a_{\min} > 0$,

$$\|\exp(A\Delta)\|_2 = \max_i e^{-a_i\Delta} \leq e^{-a_{\min}\Delta}. \tag{19}$$

The old-state term in Eqs. (11) and (12) is $\exp(A\Delta(g))(1-g)H_{t-1}$, hence submultiplicativity gives

$$\|\exp(A\Delta(g))(1-g)H_{t-1}\|_2 \leq (1-g)e^{-a_{\min}\Delta(g)}\|H_{t-1}\|_2. \tag{20}$$

Because $\Delta(g) = \mathrm{softplus}(d+\gamma g)$ has derivative $\gamma\,\mathrm{sigmoid}(d+\gamma g) \geq 0$, both $(1-g)$ and $e^{-a_{\min}\Delta(g)}$ are non-increasing on $[0,1]$; their nonnegative product is non-increasing. At $g = 1$, the multiplicative factor $(1-g)$ is zero, so this direct pre-boundary contribution is removed exactly; indirect dependence through the inferred gate is analyzed separately. For $g < 1$, the bound specifies a soft attenuation.

For a general stable A, an analogous result holds when the Euclidean logarithmic norm satisfies $\mu_2(A) \leq -a_{\min}$, since $\|e^{A\Delta}\|_2 \leq e^{\mu_2(A)\Delta}$.

C EXPERIMENTAL PROTOCOLS

All thresholds, feature pipelines, and hyperparameters below were frozen on validation data (or on an a-priori split) before the corresponding test evaluation. Boundary labels never enter the self-supervised world-model loss.

C.1 CONTROLLED REGIMEWORLD

A latent context selects one of several linear-oscillatory transition laws. Visual observations see a fraction 0.35 of the latent dimensions (partial observability); an audio identity cue marks the active regime. Independent nuisance interventions change appearance, camera, or audio without changing the dynamics. Training uses bounded event durations; held-out tests separately vary duration, transition frequency, noise, missing modalities, and unseen regime compositions, including a modality-conflict OOD split that scrambles the audio identity cue. Models are matched at $\sim$134k parameters, trained for 45 epochs on identical splits, and evaluated over 5 seeds. Open-loop targets are multi-step NLL and MSE for up to 256 steps; boundary labels used only at evaluation are the latent-regime switches. No test combination of regime, duration, and nuisance seed is used for hyperparameter selection.

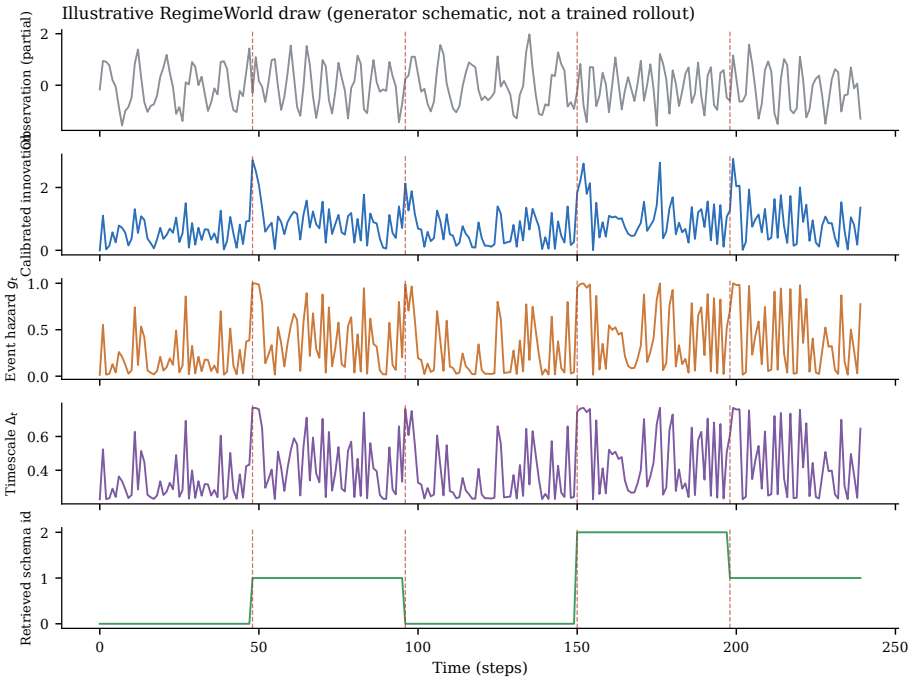


FIGURE 3: Illustrative REGIMEWORLD draw (generator schematic, not a trained rollout): partial observation, calibrated innovation, event hazard, timescale Δ_t, and retrieved schema identity across a latent-regime switch.

C.2 BOUNDARY MATCHING

Predicted and reference boundaries are matched one-to-one. For each relative-distance tolerance, a bipartite graph connects predictions and references within tolerance; maximum-cardinality matching defines true positives, and unmatched predictions and references define false positives and false negatives. We report the full Rel.Dis. tolerance curve, its mean (mean-F1), AP from continuous hazards, and video-level bootstrap intervals. Peak-pick thresholds are selected once on a held-out development half or the official validation split and then frozen.

C.3 PUBLIC VIDEO BENCHMARKS

Kinetics-GEBD. The gate, pretrained on FineVideo (Farré et al., 2024) with flow+CLAP features (Elizalde et al., 2023), is evaluated zero-shot on the K400-GEBD validation clips for which we extracted features (17,911 of the official val set). Features are extracted at a 0.15 s stride. A single threshold is chosen on a held-out development half and reported on the disjoint test half (N=8997). Online latency is the On-GEBD causal median delay, using only past and current windows. No GEBD boundary labels are used in representation learning.

TAPOS. The same untrained gate is evaluated on TAPOS Olympic sub-action boundaries (Shao et al., 2020) under the official Rel.Dis. protocol of Shou et al. (2021). The threshold is selected on TAPOS-train and reported on official TAPOS-val clips with available YouTube features (N=771 of 1,790 official val instances). Because the FineVideo-pretrained audio stream is out of domain on silent or crowd-noise Olympic footage, we also report a flow-only diagnostic that zeros the audio stream at inference (still zero-shot; no TAPOS training).

KTH Actions. We train the full HEG-VMAMBA model and a matched no-gate control on the official person split of KTH (Schuldt et al., 2004) (train persons {11–18}, dev $\{19, 20, 21, 23, 24, 25, 1, 4\}$, test $\{22, 2, 3, 5\text{–}10\}$) with Farneback flow-magnitude (256-d) + CLIP-B/32 (512-d) features at 12.5 Hz, 5 seeds, and no boundary labels. Open-loop error is compared under learned, forced-keep, and oracle gate policies at (i) discrete action-regime starts on concatenated official subsequences and (ii) intra-action cycle-phase peaks from a detector locked on train structure only (0.4–4 Hz). Significance uses a circular shift of onset labels. Coverage is $403/599$ official videos; the remaining locomotion files were unavailable at retrieval time and were not replaced by a change of split or hyperparameters.

C.4 FMRI INFERENCE

A frozen world model (never trained on BOLD) is applied to Sherlock (16 participants from the public release; ∼50 min; Chen et al., 2017) and, with the same checkpoint and pipeline, to the "(500) Days of Summer" subset of the Naturalistic Neuroimaging Database (Aliko et al., 2020) (OpenNeuro `ds002837`; 20 subjects, TR 1.0 s, 5470 volumes). Movie time is PAL-corrected onto the BOLD grid before feature extraction. Encoding uses contiguous nested folds, an HRF buffer around held-out blocks, and train-only scaling and dimensionality reduction. Joint models add low-level stimulus variables and audiovisual-encoder principal components (default 30; a 60-component control is reported for gate onsets), then calibrated innovation, then the residualized gate and state. Group inference is a circular-shift permutation that preserves autocorrelation; block-permutation and phase-randomized surrogates are used as robustness checks. ROIs are pre-specified (posterior-medial event network vs. V1/A1 sensory control) with no peak selection (Fig. 15). Whole-brain maximum-voxel maps, when shown, are exploratory and use corrected extreme-value nulls. Selection and evaluation never share a fold.

D PUBLIC-BENCHMARK RESULTS

Tables 4 to 6 expand the results of Section 4.3. Kinetics-GEBD is the detection comparison; TAPOS is a transfer evaluation that does not surpass the strongest published TAPOS baselines; KTH is the open-loop prediction evaluation, restricted to schema-recoverable action-regime starts.

TABLE 4: **Zero-shot Kinetics-GEBD** (K400 val clips with extracted features, N=17,911; Rel.Dis. F1; threshold selected on a held-out development half and reported on the complementary half, N=8997). Shared flow+CLAP features, no boundary labels. Delay is On-GEBD causal median latency (s). In-domain UBoCo (0.867 / 0.703) is not a same-pipeline row.

Model	Mean-F1 ↑	F1@0.05 ↑	Precision ↑	Delay (s) ↓
Raw feature difference	0.574	0.445	0.560	0.002
UBoCo/TSM novelty (Kang et al., 2022)	0.543	0.408	0.411	0.001
EMA-surprise threshold	0.733	**0.582**	0.690	0.011
HEG-VMAMBA gate (per-seed)	0.754 ± 0.032	0.573 ± 0.030	–	–
HEG-VMamba gate (5-seed ensemble)	**0.760**	0.565	**0.759**	0.010

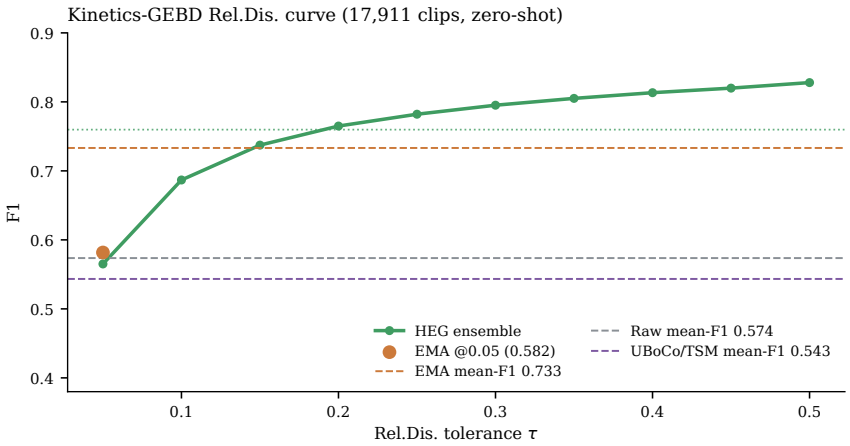


FIGURE 4: **Kinetics-GEBD Rel.Dis. curve.** F1 versus tolerance τ (zero-shot ensemble) with same-feature mean-F1 references. EMA is competitive at τ=0.05; HEG-VMAMBA leads the shared-feature curve. Published in-domain UBoCo (0.867) is not drawn.

TABLE 5: **Zero-shot TAPOS** sub-action boundaries (Rel.Dis. F1; threshold on train, report on val with features, N=771). Published rows are Table 4 of Shou et al. (2021). Flow-only zeros out-of-domain audio at inference. Transfer result, not a TAPOS leaderboard.

Method	Supervision	Mean-F1 ↑	F1@0.05 ↑
SceneDetect	unsup.	0.051	0.035
ISBA	sup.	0.302	0.106
TCN	sup.	0.330	0.237
CTM	sup.	0.350	0.244
TransParser	sup.	0.474	0.289
PA	unsup.	0.543	0.360
PC	sup.	0.642	0.522
HEG-VMAMBA gate (flow+CLAP ens.)	zero-shot	0.468	0.272
HEG-VMAMBA gate (flow-only ens.)	zero-shot	**0.491**	0.281

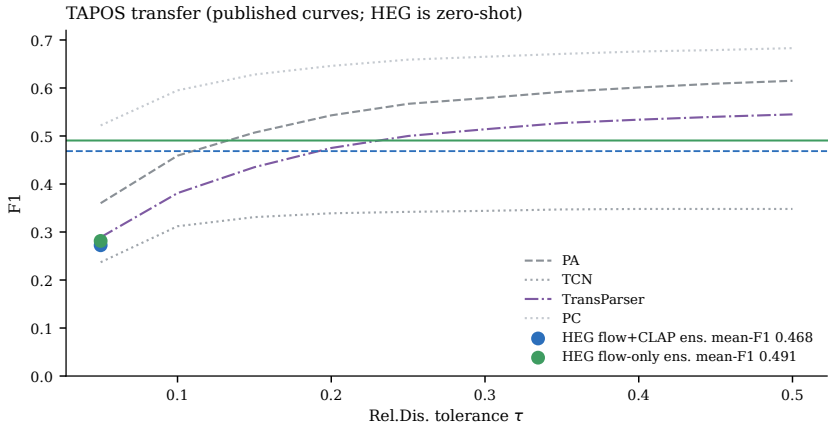


FIGURE 5: TAPOS Rel.Dis. curves from Shou et al. (2021) with zero-shot HEG-VMAMBA operating points. Horizontal lines mark ensemble mean-F1 (flow+CLAP and flow-only).

TABLE 6: KTH open-loop prediction (official person split; flow+CLIP; 5 seeds). Learned-vs-keep k=1 NLL and latent MSE at two a-priori onset classes.

Onset class	Learned vs. keep	vs. shift null	Gate rate
Action-regime starts	NLL $+0.97\pm0.13$; MSE $-62\pm16\%$ (5/5)	5/5 above 95% null	33% vs. 14% mid
Intra-action cycle peaks	null (learned $\approx$ keep)	does not exceed null	flat vs. mid-segment

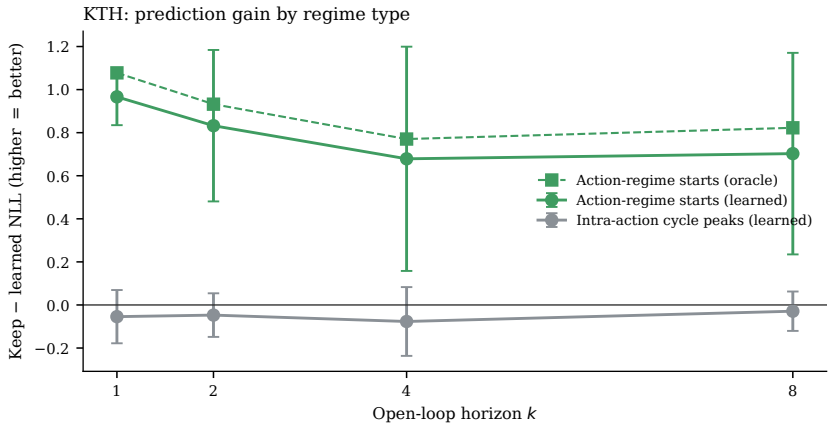


FIGURE 6: KTH open-loop prediction by horizon. Positive values are keep-minus-learned NLL (higher is better). Error bars: s.d. over 5 seeds.

E POSITIONING AGAINST RECENT DETECTORS AND SELECTIVE SSMS

Recent detectors and selective SSMs target objectives that differ from ours. The fair numeric comparison is the official Rel.Dis. *mean*-F1 (average over tolerances 0.05–0.5), not the single strictest operating point F1@0.05. Table 7 groups methods by what they are given.

Supervised *offline* detectors train on Kinetics-GEBD boundary labels and use bidirectional context. With a ResNet-50 backbone they report mean-F1 0.890 (DyBDet; 0.910 with CSN) and 0.875 (DiffGEBD). These are dedicated detection systems, not world models, and are not a like-for-like score for an unlabeled online gate.

The closest published detector is ESTimator, the On-GEBD state of the art: causal and EST-motivated, but a dedicated supervised threshold on predictor error, with no recurrent state update, timescale control, or schema. It reports online mean-F1 0.748 (F1@0.05 0.620).

HEG-VMAMBA uses *no* boundary supervision, runs online, and its boundary variable drives a predictive state update. Its zero-shot mean-F1 of 0.760 exceeds the same-feature unsupervised readouts (UBoCo/TSM 0.543, EMA-surprise 0.733) and sits at the published supervised online ESTimator average (0.748). It does *not* match the published in-domain UBoCo (mean-F1 0.867, F1@0.05 0.703): that system trains a TSM encoder on Kinetics-GEBD videos without boundary labels, whereas HEG-VMAMBA is FineVideo-pretrained and evaluated zero-shot. This is a positioning comparison, not a same-pipeline leaderboard: ESTimator uses a ResNet-50 backbone trained on GEBD labels and reports on the official validation set; HEG-VMAMBA uses flow+CLAP features and a held-out development/test split of 17,911 clips with extracted features. At the strictest tolerance, labels still help (F1@0.05 0.565 vs. 0.620). TAPOS remains a transfer-breadth result (Table 5): published supervised TAPOS mean-F1 is 0.547 (ESTimator), 0.747 (DyBDet) and 0.752 (DiffGEBD); we do not claim TAPOS detection state of the art.

TABLE 7: **Positioning** on Kinetics-GEBD Rel.Dis. mean-F1 (average over 0.05–0.5). Published detector rows use their official protocol (ResNet-50 or in-domain TSM). The same-feature block is the comparable set: FineVideo flow+CLAP, no boundary labels (Table 4). HID-SSM and TIDES are mechanistic, not scored.

Method	Bound. sup.	Online	Mean-F1	F1@0.05	Mechanism / role
Published detectors (not same pipeline)					
DyBDet (Zheng et al., 2024)	✓	×	0.890	0.796	offline; multi-exit
DiffGEBD (Hwang et al., 2025)	✓	×	0.875	0.784	offline; diffusion
ESTimator (Jung et al., 2025)	✓	✓	0.748	0.620	online; predictor-error (EST)
UBoCo (Kang et al., 2022)	×	×	0.867‡	0.703‡	in-domain TSM+RTP; not zero-shot
Same-feature comparison (ours; no boundary labels)					
UBoCo/TSM readout	×	×	0.543†	0.408†	novelty on shared flow+CLAP
EMA-surprise	×	✓	0.733	0.582	causal error EMA threshold
HEG-VMamba	×	✓	**0.760**	0.565	**gate updates predictive state**
Selective SSMs in other domains					
HID-SSM (Wu et al., 2025)	—	✓	—	—	surgery; Cholec80 causal Acc 94.5%
TIDES (Soydan et al., 2026)	—	—	—	—	analytic Δ (time series)

†Same-feature UBoCo/TSM readout (Table 4), not the published UBoCo pipeline. ‡Table 1 of Kang et al. (2022): unsupervised UBoCo-Res50 Average F1 86.7 and F1@0.05 70.3 (in-domain; trained on Kinetics-GEBD videos without boundary labels).

For the two selective-SSM entries, the comparison is mechanistic rather than numerical. HID-SSM (Wu et al., 2025) stacks input-conditioned local and global SSM blocks for full-length surgical video; it shares our use of input-dependent timescales and hierarchy but has no explicit surprise variable, schema retrieval, or counterfactual grounding, and is supervised by phase labels. TIDES (Soydan et al., 2026) is the sharpest challenge to our design: it argues that making Δ a learned function of the input removes its physical-time meaning and harms extrapolation to unseen sampling rates, and proposes to keep Δ analytic while moving selectivity onto the continuous-time generator. Section G stress-tests that claim directly.

F Additional analyses

F.1 Long-form and modality factors

To locate the source of long-form and multimodal gains we cross dataset type, hours, contiguous context length, and modality. Within each video the SSM state is carried across consecutive windows. Shuffled short windows are the control, and each video is trained at several context lengths, so a long-form prior cannot be inferred from the corpus alone. A factorial mixed-effects model yields a long-horizon gain of -41% open-loop error versus a copy-last baseline across horizons 1–16 ($5/5$ seeds; the gate-specific contribution is ≈ 0). Open-loop error is flat across contiguous contexts of 32–96 frames (≈ 0.38; slope ≈ 0, $p{=}0.44$): contiguous context length alone does not improve prediction on this corpus. Testing whether that flatness depends on dataset type requires a second long-form corpus with the same features. Modality factorization is $V \approx AV > A$ on FineVideo ($p{=}0.76$); complementary-control synergy $+0.26$ nats ($5/5$, $p{=}9\times10^{-5}$). On FineVideo a matched-capacity retrain isolates vision as the predictive carrier: audio is redundant and does not degrade vision-only prediction. The complementary-modality control in the main text shows that this is a property of natural video rather than of the mechanism (Fig. 7).

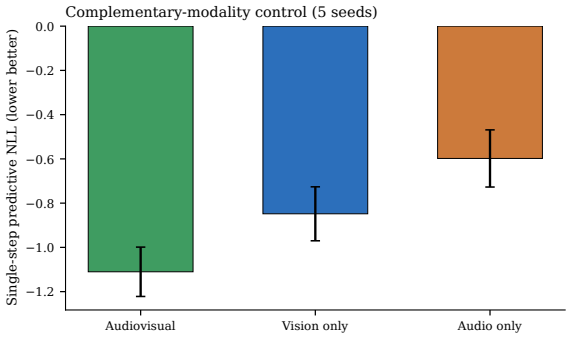


Figure 7: Complementary-modality control: streams observe disjoint latent halves. Audiovisual fusion has lower single-step NLL than either stream (5 seeds).

F.2 Qualitative mechanism inventory

Table 8 extends Table 2 with a learned open-loop prior and two extra rows (a raw-innovation threshold and a posterior-only gate). It records which components each variant includes; the quantitative comparison remains Table 2.

Table 8: Mechanism-identification inventory (qualitative extension of Table 2). *Prior* marks whether a learned open-loop boundary prior is trained.

Variant	Timing	Δ	Schema	Prior	Role
Trained no-gate	×	×	×	×	baseline
Raw-innovation threshold	×	✓	✓	×	change-point
Matched random/shifted gate	×	✓	✓	×	matched perturbation
Δ-only	✓	✓	×	✓	timescale only
Schema-only	✓	×	✓	✓	schema only
No counterfactual utility	✓	✓	✓	✓	no utility loss
Posterior-only gate	✓	✓	✓	×	teacher forcing
HEG-VMamba	✓	✓	✓	✓	full model

Figures 8 to 11 expand the quantitative decomposition of Table 2.

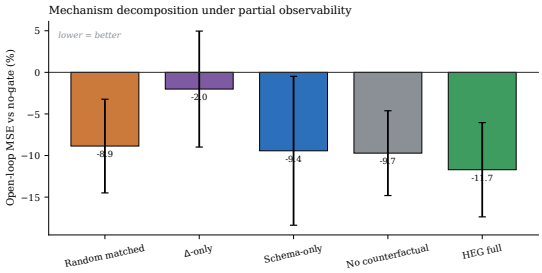


FIGURE 8: REGIMEWORLD mechanism decomposition vs. no-gate MSE (5 seeds; negative better).

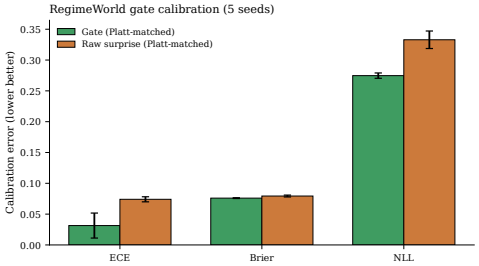


FIGURE 9: Gate vs. raw-surprise calibration after Platt scaling (5 seeds).

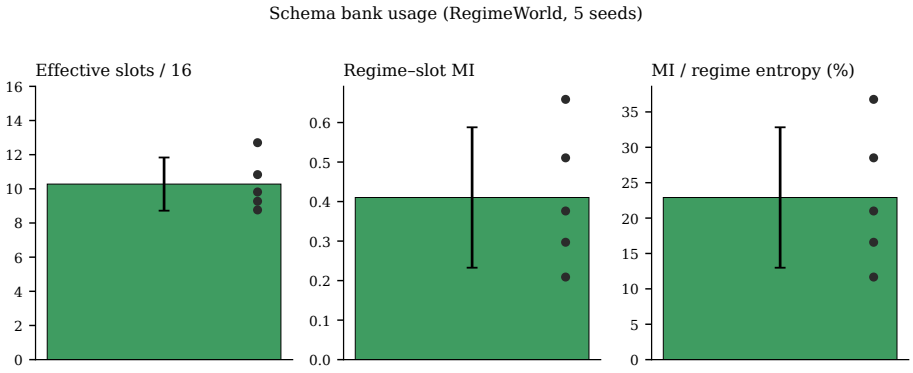


FIGURE 10: Schema-bank usage on the partially observed REGIMEWORLD task: effective slots out of 16, regime–slot MI, and MI as a fraction of regime entropy (5 seeds).

F.3 SCHEMA GENERALIZATION UNDER DISTRIBUTION SHIFT

To test whether schemas encode reusable regime structure rather than memorized training schedules, we re-evaluate frozen partially observed REGIMEWORLD checkpoints (5 seeds) on OOD generative shifts that share the regime bank but change the schedule or dynamics (longer/shorter durations, faster oscillations). Table 9 reports regime–slot MI, the modal-slot *consistency* (fraction of a regime's boundaries that retrieve that regime's most-frequent slot, across different sequences), and the causal wrong-schema swap cost. The regime→slot mapping is preserved or strengthened under shift and the same regime reuses the same slot across unseen schedules (consistency ≈ 0.49–0.54 over 10–16 effective slots), indicating reusable structure. The wrong-schema swap cost is modest at this observation fraction (0.35), consistent with the main-text finding that schema *content* becomes strictly necessary only under heavier occlusion; here it degrades gracefully rather than collapsing.

TABLE 9: Schema generalization on OOD shifts sharing the regime bank (5 seeds). MI and modal-slot consistency transfer; the wrong-schema swap remains non-beneficial.

Condition	Regime–slot MI frac	Slot consistency	Wrong-swap cost	Eff. slots
In-distribution	0.229	0.488	1.5%	10.28
Long durations	0.299	0.514	1.8%	10.49
Short durations	0.172	0.445	−0.3%	10.32
Faster dynamics	0.331	0.543	3.0%	10.43

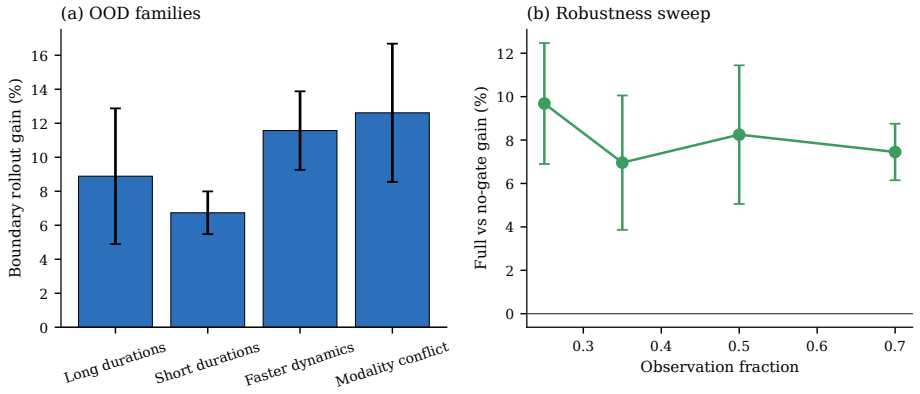


FIGURE 11: (a) Boundary-conditioned rollout gain on held-out OOD families. (b) Full vs. no-gate gain across observation fractions. Error bars: s.d. over seeds.

G LEARNED-Δ ROBUSTNESS UNDER TIMING SHIFT

TIDES (Soydan et al., 2026) argues that a learned, input-dependent discretization step Δ loses its physical-time meaning and can fail to extrapolate across sampling regimes. We test this directly on the frozen partially observed REGIMEWORLD model (5 seeds) by *resampling* the evaluation sequences onto new time grids and remapping the boundary labels: zero-order-hold *missingness* at randomly dropped frames (20% and 40%); genuine *rate downshift* that subsamples to a shorter sequence ($\times\frac{1}{2}$, $\times\frac{1}{3}$); *rate upshift* that interpolates to a longer sequence ($\times 2$); and piecewise *time-warping* that varies the local speed while spanning the full range (no static padding). For each transform we report boundary-detection AP and open-loop rollout gain (keep$-$learned) for the learned Δ-gate, the Δ-gate disabled at eval, and—for the rate shifts, where the physical inter-frame ratio is known—a *timestamp-aware* variant that rescales Δ by that ratio (the analytic-Δ prescription of TIDES; Table 10). This also tests a hybrid ablation that disables learned Δ at evaluation.

TABLE 10: Learned-Δ stress test on frozen partially observed REGIMEWORLD checkpoints (5 seeds). Detection and gain survive missingness and time-warping; predictive gain drops under a global frame-rate change. Rescaling Δ by the known sampling ratio partially recovers the largest shifts.

Transform	Detect. AP ↑	Gain (learned Δ)	Gain (Δ off)	Gain (rescaled Δ)
Clean	0.521	17.8%	17.4%	—
Missing 20%	0.380	15.5%	14.8%	—
Missing 40%	0.339	12.2%	11.2%	—
Rate $\times\frac{1}{2}$	0.292	4.8%	3.2%	3.9%
Rate $\times\frac{1}{3}$	0.429	-0.7%	-1.9%	0.3%
Rate $\times 2$	0.158	1.8%	1.6%	3.8%
Time-warp	0.360	10.1%	8.8%	—

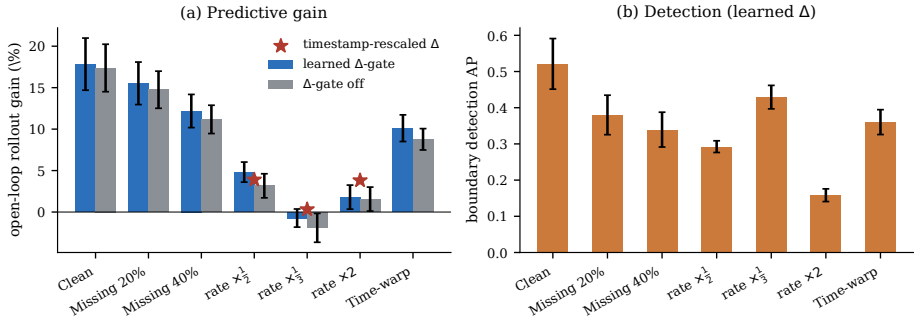


FIGURE 12: Learned-Δ stress test (5 seeds). (a) Open-loop gain by transform; red stars mark timestamp-aware Δ rescaling. (b) Boundary-detection AP.

The gate as a *detector* and its predictive gain both survive frame jitter, missingness, and local time-warping—the perturbations that preserve relative event timing. The predictive gain degrades only under a genuine change of the global frame rate, when the model's implicit per-step Δ no longer matches elapsed time: exactly the extrapolation gap TIDES predicts for learned Δ. Crucially, our diagnostic both localizes the failure to the rollout objective (not detection) and shows that the TIDES remedy works: rescaling Δ by the known sampling ratio partially recovers the gain, and the learned Δ-gate is beneficial rather than the culprit. Training-time timestamp conditioning is the natural route to full recovery.

H BOUNDARY-PRIOR CALIBRATION OVER LONG ROLLOUTS

The posterior gate is computed from observed innovation; in open-loop rollout no observations exist. We compare, over rollouts to $K{=}120$, a state-only boundary *prior* (a small head $\sigma(\mathrm{MLP}(\bar{s}_t))$ trained with plain BCE on the frozen backbone and temperature-scaled on validation) against the *posterior* gate forced to run on the model's own hallucinated frame. Rollouts are primed on 8 real frames; at each horizon we score the driving probability against the true boundary (Table 11).

TABLE 11: Prior vs. posterior boundary calibration in open-loop rollout (5 seeds, by horizon k). The posterior cannot fire without observed surprise. The temperature-scaled state-only prior is calibrated and flat to $K{=}120$; ranking AP sits at the base rate (0.091).

	State-only prior (temp-scaled)			Posterior
Horizon k	ECE ↓	NLL ↓	AP	ECE ↓
1–20	0.008	0.31	0.094	0.256
21–50	0.014	0.31	0.093	0.284
51–80	0.014	0.31	0.090	0.283
81–120	0.012	0.31	0.093	0.282
Overall	0.012	0.31	0.092	0.279

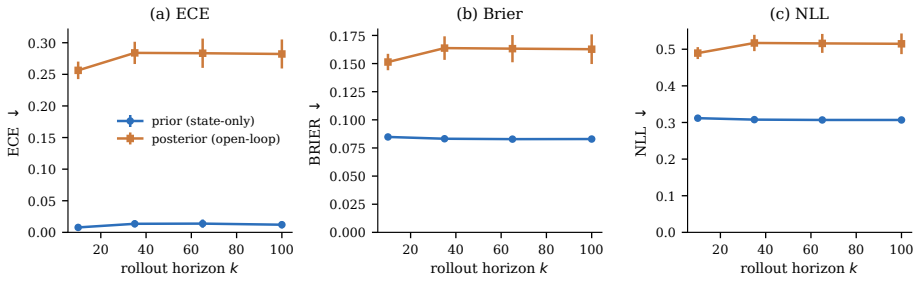


FIGURE 13: Prior vs. posterior calibration over open-loop rollout (5 seeds): ECE (a), Brier (b), NLL (c) versus horizon k. The state-only prior (blue) is flat to $K{=}120$; the hallucinated posterior (orange) is worse.

The diagnostic supports the design and sharpens the limitation. It confirms that a posterior gate cannot function open-loop—justifying the posterior→prior split—and it shows that a lightweight, temperature-scaled prior is well calibrated and *stable* out to $K{=}120$ rather than drifting: its ECE stays near 0.01 and its NLL is flat across horizons. The remaining limitation is discrimination rather than calibration: because event boundaries here are exogenous, the prior's long-horizon boundary-ranking AP sits at the base rate, so the prior maintains a sensible, well-calibrated boundary rate but does not anticipate the exact timing of distant boundaries—appropriate, calibrated behavior for open-loop rollout.

I Compute and Latency

Table 12 quantifies the overhead of the gate, schema, and counterfactual branch relative to a plain SSM core and to simpler EMA/raw-threshold gates, on the controlled model ($131{,}953$ parameters; batch 128, $T{=}128$; single RTX 6000-Ada-class GPU). The counterfactual branch and schema-future auxiliary are *training-only* and absent at inference.

Table 12: Compute/latency on the controlled model. Inference cost over a plain SSM is one gate MLP plus schema retrieval ($\sim$1450 seq/s). Relative overhead is inflated by the tiny backbone ($d{=}64$). The counterfactual branch is training-only ($+12.7\%$ wall-time).

Configuration	Params	Latency (ms/seq) ↓	Throughput (seq/s) ↑
Plain SSM core	72.7k (core)	0.191	5236
EMA / raw-threshold gate	+predictor	0.338	2954
HEG-VMamba (full inference)	132.0k	0.688	1454
Training-only branch (per batch of 128):			
No counterfactual branch	—	331.6 ms	peak 446 MiB
+ counterfactual + schema-future	—	373.5 ms (+12.7%)	peak 2990 MiB

The gate and schema together are 26% of parameters. At inference the deployed cost over a bare SSM is a single gate MLP and one schema read; the counterfactual branch is paid only during training, where it costs 13% wall-time and additional activation memory from the K-step unroll.

J Sensitivity to Schema Count, Horizon, and Rate Target

We sweep schema count, counterfactual horizon K, and the rate target π, one at a time around the default configuration (16 schemas, $K{=}10$, $\pi{=}0.10$), retraining the full model for 3 seeds each (Table 13). Results are stable and the trends are interpretable rather than knife-edged.

TABLE 13: Sensitivity sweeps on REGIMEWORLD (full model, 3 seeds; default settings in bold). Detection AP and rollout gain are stable across schema count and a 20× range of K; π trades recall against calibration.

Value	Detect. AP	Rollout gain	Gate ECE	Regime–slot MI frac	Eff. slots
Schema count (default **16**):					
4	0.469±0.035	15.1%	0.004	0.154	3.41
8	0.529±0.012	17.4%	0.023	0.190	5.64
16	0.517±0.036	17.4%	0.020	0.234	9.38
32	0.538±0.030	18.4%	0.031	0.410	16.32
64	0.515±0.026	16.1%	0.015	0.386	33.45
Counterfactual horizon K (default **10**):					
2	0.646±0.023	19.0%	0.074	0.225	10.10
5	0.569±0.026	20.8%	0.052	0.273	10.72
10	0.517±0.036	17.4%	0.020	0.234	9.38
20	0.497±0.033	13.5%	0.010	0.192	8.35
40	0.479±0.037	11.9%	0.007	0.214	8.39
Rate target π (default **0.10**):					
0.05	0.499±0.045	15.7%	0.011	0.223	8.97
0.08	0.510±0.036	16.6%	0.016	0.229	9.09
0.10	0.517±0.036	17.4%	0.020	0.234	9.38
0.15	0.530±0.036	18.3%	0.027	0.236	9.60
0.20	0.549±0.029	19.3%	0.036	0.241	9.86

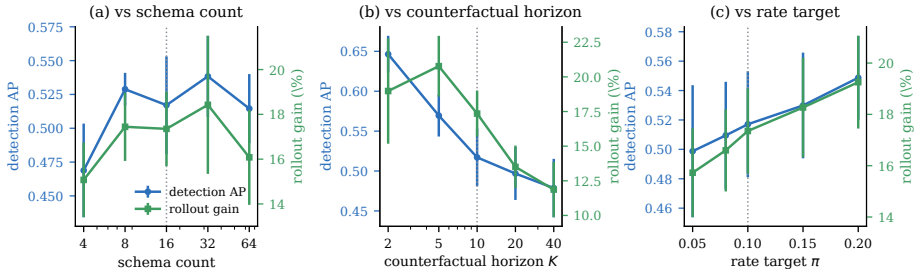


FIGURE 14: Sensitivity sweeps (3 seeds; dotted line = default setting). Detection AP (left) and open-loop gain (right) vs. (a) schema count, (b) horizon K, (c) rate target π.

The mechanism is robust: as few as 4 schemas or a horizon as short as K=2 already recover most of the gain, while very long horizons ($K \geq 20$) dilute the counterfactual signal and reduce it. Larger banks are used (effective slots grow with capacity) without destabilizing detection. The rate target behaves as a single recall/calibration dial—higher π raises detection AP and rollout gain but increases gate ECE—so the default π=0.10 is a deliberate midpoint rather than a tuned optimum.

K NEURAL ANALYSES

Figure 15 is the group-level double dissociation of Section 4.5. No voxelwise statistical map is reported: group inference used region-wise summary statistics, not volume maps. Event and sensory ROIs are pre-specified (PMC/AG/mPFC/TPJ vs. V1/A1; no peak selection). Figures 16 to 18 report the independent-film replication, the recurring-scene null, and the subject-level increment.

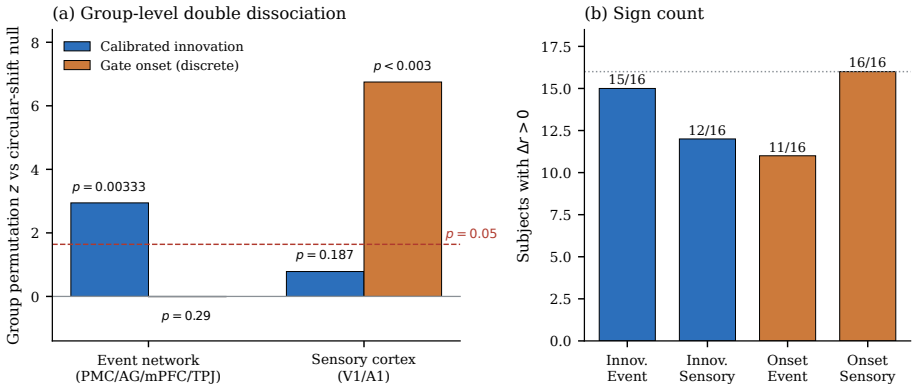


FIGURE 15: **Double dissociation** on Sherlock (16 subjects; circular-shift group null, α=100). (a) Permutation z for calibrated innovation and discrete gate onsets. (b) Subjects with $\Delta r > 0$. Innovation is event-network specific; gate onsets are sensory-specific.

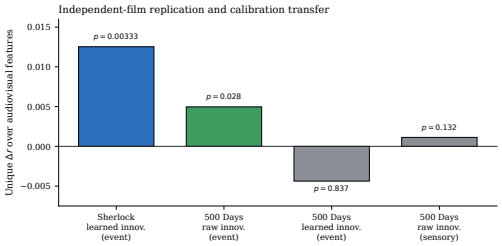


FIGURE 16: **Independent-film replication** (n=20). Frozen learned calibration does not transfer.

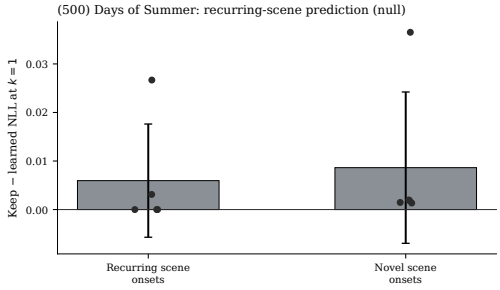


FIGURE 17: Recurring-scene null: no keep-vs-learned NLL gain at k=1.

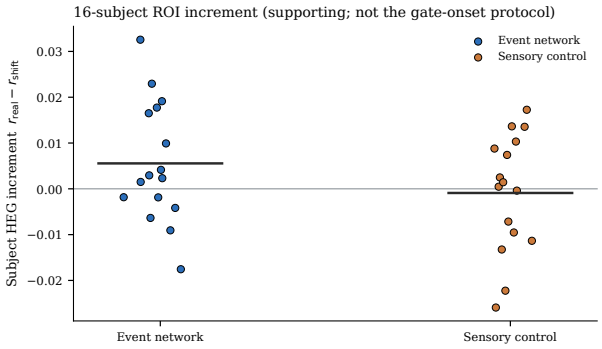


FIGURE 18: Subject-level increment ($r_{\text{real}} - r_{\text{shift}}$) on Sherlock event vs. sensory ROIs (supporting; not Fig. 15).

L REPORTING DETAILS

TABLE 14: Reporting facts for the results in this paper. Negative and non-significant tests are retained in the main text.

Item	Setting
Controlled capacity / seeds	$\sim$134k parameters; 5 seeds $\{13, 17, 21, 29, 37\}$; 45 epochs; obs. fraction 0.35
Video seeds	$\geq$5 on Kinetics-GEBD, TAPOS, and KTH; KTH uses the official person split; thresholds frozen on validation
GEBD coverage	17,911 K400-val clips with extracted features; held-out test half N=8997; flow+CLAP @ 0.15 s
TAPOS coverage	N=771 official val clips with features (of 1,790 official val instances); train-only threshold
KTH coverage	403/599 videos; official person split; flow+CLIP @ 12.5 Hz
fMRI	Sherlock n=16; ds002837 n=20; frozen checkpoint; circular-shift group null; a-priori ROIs
Encoders	Frozen CLIP / Whisper or flow+CLAP / flow+CLIP as named per experiment; adapters trained
Nulls	Circular shift (primary); block permutation and phase randomization as robustness