FORECASTING SKILL ALONE DOES NOT CERTIFY INTERVENTION COMPETENCE IN WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

World models are validated by forecasting held-out trajectories, yet they are also used to predict outcomes after actions that change the system. Held-out forecast error does not directly measure whether a model can distinguish outcomes under candidate actions, although it is often treated as evidence of intervention competence. We introduce forecast-gated intervention evaluation: a world model first demonstrates held-out forecast skill and is then tested on interventions. We ask whether rolling its learned dynamics forward adds intervention value beyond label-supervised predictors that use the same observation and intervention but never roll dynamics forward. The benchmark is two simulated Epileptor cohorts (12 and 16 nodes) whose resection outcome is decided by one latent variable, a deliberately favorable case for certification. An ODE-based Epileptic-brain World Model (EWM) clears the gate but trails the no-rollout baseline in both settings: AUROC is 0.519 vs. 0.743 at 12 nodes and 0.509 vs. 0.701 at 16 nodes. Ten further rollout models pass the gate. At the primary label budget, no evaluated rollout clears the prespecified margin over the label-supervised no-rollout head built on TGR-WM's temporal encoder. Swapping the EWM and TGR-WM encoders reverses their intervention-performance ordering, suggesting that intervention transfer is strongly encoder-dependent in this benchmark. Within each EWM member, the decoded signal tracks the intervention-relevant state, but independently trained members exhibit different orientations in that association, so averaging cancels it. A validation-derived orientation bit per member brings EWM only level with the floor. World models intended for interventions should be evaluated on interventions, against strong no-rollout baselines.

1 INTRODUCTION

Learned world models have become a standard approach to prediction and planning: observations are encoded into a latent state, learned dynamics are rolled forward, and actions are selected from predicted futures (Ha & Schmidhuber, 2018; Hafner et al., 2019). Their evaluation, however, still turns mainly on forecast error over held-out trajectories. Accurate forecasts are often taken to imply suitability for downstream planning, including proposed clinical applications (Qazi et al., 2025). When world models are used to predict outcomes after interventions, low forecast error on held-out pre-intervention trajectories does not by itself establish accurate intervention prediction. Epilepsy surgery is a concrete setting where this matters: about one third of patients are drug-resistant (Perucca et al., 2023), surgery renders only 50–70% seizure-free (Avigdor et al., 2026), and selecting tissue to resect means predicting how seizure dynamics change once the network is perturbed. Our benchmark tests a prerequisite: state recovery before resection, not post-resection propagation.

Predictive fidelity need not imply decision value: lower forward-model error need not improve control (Lambert et al., 2020; Yu et al., 2026), and action-conditioned benchmarks show that plausible futures can fail to respect interventions (Cai et al., 2026; Yang et al., 2026). Our contribution is a direct intervention test. After a held-out forecasting gate, we compare forecast-trained rollouts with direct predictors trained on intervention-outcome labels at a specified budget. Both receive the same observations and proposed interventions; architecture and state-access controls examine transfer across models.

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

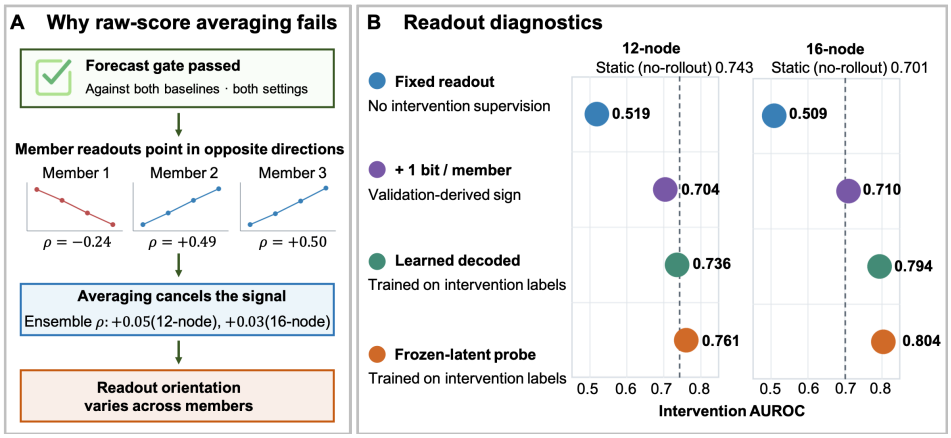


Figure 1: **Why held-out forecasting skill does not certify intervention competence.** **A**, EWM passes the forecasting gate, but opposing member orientations cancel when averaged (12-node correlations shown). **B**, one validation-derived bit per member restores performance to the Static no-rollout floor; learned readouts on frozen decoded signals or latents recover more discrimination. Dots show test AUROC; dashed lines mark Static.

We answer with a forecast-gated intervention evaluation protocol. A world model is admitted to the intervention comparison only after demonstrating held-out forecasting skill; its intervention ranking is then compared against a strong structural predictor with no rollout, using paired patient-clustered bootstrap inference and a prespecified smallest effect size of interest (SESOI) of $+0.05$ AUROC. Architecture and state-access controls then probe which model choices influence the transfer.

We instantiate the protocol in two independently generated cohorts of 1,400 virtual patients each, on 12- and 16-node Epileptor networks, each patient carrying two candidate resections whose outcomes are read from the intervened simulator. The benchmark has a structure we state up front: both resections leave one relay node isolated, and that relay's true excitability alone predicts the outcome at AUROC 0.999, so the intervention task is to recover one intervention-relevant state from the observation window, not to simulate propagation. Forecasting skill and intervention ranking come apart even in this simplified setting. EWM passes the forecasting gate, but its fixed readout ranks interventions near chance because member-specific orientations cancel (Figure 1A). A validation-derived orientation bit restores performance to the no-rollout floor, while learned readouts on the same frozen outputs recover more discrimination (Figure 1B). Across architectures, no rollout model adds the prespecified margin beyond the strongest no-rollout predictor at the primary label budget, and encoder swaps reverse the ordering.

In summary, our main contributions are as follows:

- **A forecast-gated intervention evaluation protocol.** A rollout model enters the intervention comparison only after clearing a held-out forecasting gate; intervention value is then evaluated against strong predictors that use the same observation and intervention without rolling dynamics forward, with paired inference against the SESOI.
- **Forecast-trained EWM members differ in intervention-readout orientation.** EWM clears the gate yet ranks interventions far below the no-rollout floor. Within each ensemble member the decoded signal tracks the intervention-relevant state, but its orientation varies across members and averaging cancels the signal. One validation-derived bit per member brings EWM level with the floor; a 33- to 545-parameter supervised readout on the same frozen latents reaches or exceeds the no-rollout floor.
- **The dissociation extends across architectures, and transfer is strongly encoder-dependent.** A matched GRU-core, a Dreamer-family RSSM among the strongest forecasters in the study, and a transformer world model likewise provide no consistent intervention advantage over the no-rollout floor. TGR-WM clears the floor at 16 nodes, but a label-supervised head on its own temporal encoder beats it in both settings, so no rollout model adds value beyond the strongest no-rollout predictor by the margin at the primary label budget. Matching parameters in both directions

leaves the ordering intact. Encoder swaps reverse the ordering in this benchmark, suggesting that encoder choice strongly influences intervention transfer.

- **A paired-ground-truth intervention benchmark.** Two independently generated Epileptor cohorts provide both candidate-resection outcomes for every virtual patient and expose the intervention-relevant latent state, allowing forecasting and intervention discrimination to be evaluated separately. The generator and a reference implementation of EWM training, forecast gating, intervention scoring, and readout diagnostics accompany the paper.

2 Related Work

World models and learned dynamics. World models share one recipe: encode observations into a latent state, learn a transition, and plan by rolling it forward under actions (Chiappa et al., 2017; Ha & Schmidhuber, 2018; Hafner et al., 2019), from games and continuous control (Schrittwieser et al., 2020; Hafner et al., 2025; Hansen et al., 2024) to clinical treatment planning (Qazi et al., 2025; Yang et al., 2025). Backbones include sequence models (Micheli et al., 2023; Deng et al., 2023), graph networks (Sanchez-Gonzalez et al., 2020), structured world models (Kipf et al., 2020), and neural ODEs (Chen et al., 2018; Rubanova et al., 2019); our architecture controls vary this design choice.

Fidelity versus decision and intervention competence. Predictive fidelity and decision usefulness come apart: value-aware and value-equivalent model learning shows that a model useful for planning need not minimize observation error (Farahmand et al., 2017; Grimm et al., 2020), and objective mismatch shows the converse (Lambert et al., 2020; Janner et al., 2019). Behavioral benchmarks probe this gap through counterfactual physical scenes (Baradel et al., 2020; Thorpe et al., 2026), physical-understanding tests of video models (Motamed et al., 2026), action-conditioned suites (Cai et al., 2026; Yang et al., 2026; Li et al., 2025), and embodied task success (Zhang et al., 2026a), since high-fidelity futures can still violate physics or ignore the action. A decision-centric position sets out the evidence such evaluation should require (Yu et al., 2026). We instead test whether forecast-trained rollout provides intervention-relevant value beyond a predictor that uses the same observation and intervention but never rolls dynamics forward.

Latent-state sufficiency and identifiability. A state sufficient for prediction need not suffice for control or intervention (Kim, 2026), and whether a learned state recovers intervention-relevant factors is often studied as an identifiability question (Zhang et al., 2026b; Lippe et al., 2023); in our experiments, independently trained members exhibit different intervention-readout orientations. We also manipulate state access, replacing the learned state by the true one at training or evaluation.

Epilepsy intervention modelling. Machine learning for epilepsy has concentrated on localization (Li et al., 2021) and self-supervised iEEG/EEG models (Wang et al., 2023; Jiang et al., 2024); neither predicts outcomes under a specified intervention. Virtual-intervention modelling does, simulating resection or stimulation on individualized connectomes (Kini et al., 2019; Sinha et al., 2017; Gleichgerrcht et al., 2018; Sanz Leon et al., 2013; Jirsa et al., 2017; Proix et al., 2017; Hashemi et al., 2023; Wang et al., 2025). Treatment-effect estimators (Bica et al., 2020; Melnychuk et al., 2022) share our target but address confounded observational estimation, not the value of learned rollout (Appendix F).

3 Method

Rollout models must first pass a held-out forecast gate, then be tested against predictors with matched inputs but no rollout. The Epileptic-brain World Model (EWM) encodes the pre-intervention state, applies resection, rolls forward, and scores the trajectory. Figures 2 and 3 summarize the protocol and EWM data flow; notation is in Appendix G.

3.1 Problem formulation

For patient i, the inputs are a connectivity graph with adjacency matrix $A_i \in \mathbb{R}^{N_i \times N_i}$, a pre-intervention observation window $X_i^{\mathrm{pre}} \in \mathbb{R}^{N_i \times T_w}$, and a candidate intervention a; resection re-

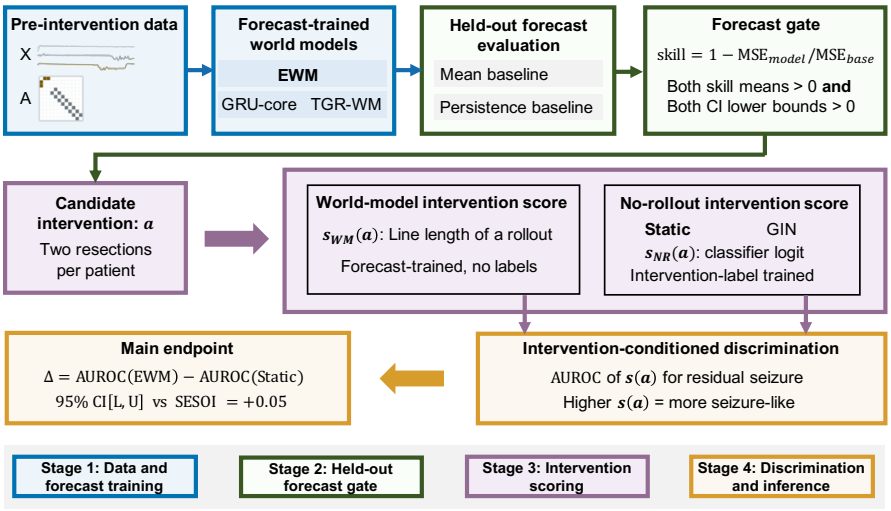


Figure 2: **Four-stage forecast-gated workflow.** Forecast-trained world models must first pass a held-out forecasting gate; rollout and no-rollout models then score the same two resections per patient, followed by intervention discrimination and paired inference. Colors identify the four stages.

moves a node set R from graph coupling. The task is intervention scoring,

$$s_i(a) \;=\; f_\theta\big(X_i^{\mathrm{pre}},\, A_i,\, a\big), \tag{1}$$

a continuous score that is higher the more seizure-like the predicted post-intervention activity remains, so that candidate interventions are ranked by outcome. The model conditions on the pre-intervention window, the original graph, and the proposed resection, without access to post-intervention observations.

3.2 Forecast-gated intervention evaluation

Four-stage evaluation. Stage 1: data and forecast training. World models train on pre-intervention trajectories without intervention labels. **Stage 2: held-out forecast gate.** Each rollout predicts 288 held-out steps from a normalized 96-step context and must clear MSE skill against both baselines. **Stage 3: intervention scoring.** Each model scores the same two resections from matched inputs: rollout models use line length (Eq. 2), while Static and GIN use label-trained logits. The comparison estimates intervention value at explicit supervision budgets, contrasting forecast-trained rollout without intervention labels with label-supervised direct predictors that use the same observation and intervention. **Stage 4: discrimination and inference.** Residual-seizure AUROC is evaluated separately by setting; the confirmatory statistic is the patient-clustered-bootstrap EWM–Static contrast against a prespecified $+0.05$ AUROC SESOI, a pragmatic benchmark threshold that separates a material from a merely positive gain; conclusions follow the CI bounds (Table 1).

Models under comparison. Each model probes a distinct contrast (Table 9, Appendix B.1). Static is the no-rollout floor and GIN adds structural capacity; both use intervention labels unavailable to world models. EWM ODE is the main model and the only confirmatory arm; the remaining rollout models are descriptive controls. GRU-core retains its encoder and decoder but replaces the ODE core; TGR-WM is architecturally distinct but matched in inputs and intervention semantics. A Dreamer-family RSSM, a transformer world model (TWM), parameter-matched variants, and encoder swaps test architecture, capacity, and encoder dependence (Section 4.4).

3.3 EWM architecture

Graph encoder. The encoder extracts six statistics ϕ_n per node from X^{pre} (mean, standard deviation, minimum, maximum, final value, and mean adjacent absolute difference, the readout's line-length feature), a fixed summary that avoids committing to a temporal architecture. A shared MLP embeds ϕ_n into h_n, one degree-normalized graph hop (Kipf & Welling, 2016) adds the network con-

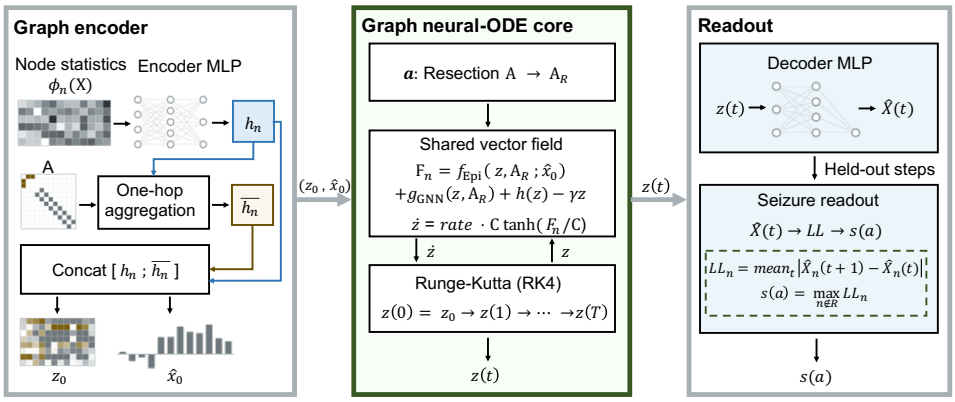


Figure 3: **EWM architecture and intervention readout.** The graph encoder maps (X, A) to $(z_0, \hat{x}_0)$. A candidate resection enters the neural-ODE rollout through A_R; the decoded rollout is scored by the maximum line length over surviving nodes (Eq. 2; full vector field in Appendix A).

Table 1: **Confirmatory protocol**, fixed before evaluation; Figure 2 summarizes the workflow.

Stage	Item	Rule	Protocol value
Setup	Cohorts	Two independently generated Epileptor cohorts; one scalar z-score per setting, fitted on training context only	12- and 16-node; 1,400 each; 980/210/210
	Interventions	Each removes the epileptogenic node and one relay; removed rows and columns of A zeroed; label read from the intervened simulator	2 per patient
1 Training	World models	Forecast objectives on pre-intervention trajectories; no intervention labels	steps 96 : 384
2 Gate	Forecast skill	$1 - \mathrm{MSE}_{\mathrm{model}}/\mathrm{MSE}_{\mathrm{baseline}}$ on held-out steps against the context-mean and persistence baselines; pass iff mean skill > 0 and lower CI bound > 0 for both baselines	steps 96 : 384
3 Scoring	Rollout	Maximum over remaining nodes of the mean decoded line length (Eq. 2); no labels	steps 96 : 384
	No-rollout	Intervention-conditioned logit trained on labels	N/A
	Members	Three members per architecture and setting; test outcomes excluded from selection; scores averaged per patient–intervention pair before AUROC	3
4 Discrimination and inference	Endpoint	$\Delta = \mathrm{AUROC(EWM)} - \mathrm{AUROC(Static)}$ per setting, never pooled; the same paired contrast for every architecture	SESOI $+0.05$
	Uncertainty	Paired patient-clustered bootstrap, 95% CI	10,000 resamples
	Decision	† lower $>$ SESOI; * $0 <$ lower $\leq$ SESOI; $^{\circ}$ lower ≤ 0, upper $\geq$ SESOI; ‡ lower ≤ 0, upper $<$ SESOI	N/A

text $\bar{h}_n$, and two linear heads produce the initial latent state $z_n(0)$ and an unconstrained excitability-like coordinate $\hat{x}_{0,n}$, which no loss calibrates to x_0^{true} (Appendix A).

Physics-informed graph neural-ODE core. Four additive terms compose the vector field (Eq. 4, Appendix A). The Epileptor-structured prior f_{Epi} supplies mechanistic drift as a function of $\hat{x}_0$ (Jirsa et al., 2014), g_{GNN} propagates activity along A, the learned residual absorbs what the prior does not capture, and the leak pulls states toward rest; $C\tanh(\cdot/C)$ bounds the field and fixed-step Runge–Kutta integration keeps rollouts differentiable (Chen et al., 2018). The prior places $\hat{x}_0$ in an excitability-like role without asserting $\hat{x}_0 = x_0^{\mathrm{true}}$ (Appendix A).

Intervention operator, decoder, and readout. Resection zeroes the rows and columns of A for R before rollout and scoring, giving the intervened graph A_R. The window is encoded once, on the original graph, so a rollout on A_R is a prediction about the intervention rather than a fit to it. A decoder returns the observed signal, $\hat{X}_n(t) = \mathrm{MLP}_{\mathrm{dec}}(z_n(t))$, and the fixed readout scores a resection by the most seizure-like remaining node,

$$s_i(a) = \max_{n \notin R} \mathrm{LL}\big(\hat{X}^{(a)}_{i,n}[96{:}384]\big), \quad \mathrm{LL}(y) = \frac{1}{|y| - 1} \sum_t |y_{t+1} - y_t|, \tag{2}$$

where LL is mean line length over the held-out segment, an amplitude-variation measure that rises with seizure-like activity. The maximum encodes the clinical logic that a resection fails if any remaining node still seizes; no threshold is tuned.

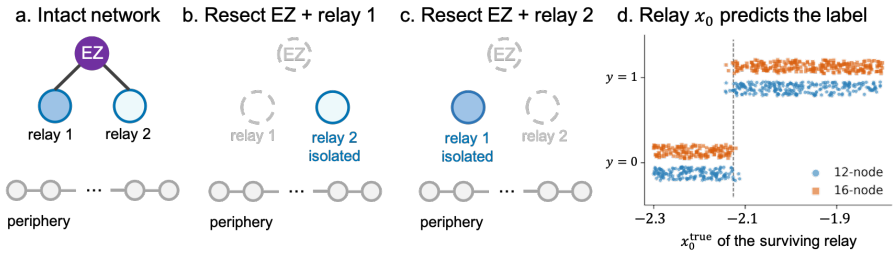


Figure 4: **Benchmark structure.** **a**, EZ couples to two relays; the relays connect neither to each other nor to the weakly coupled periphery, and only relay excitability varies across patients. **b, c**, each resection removes EZ with one relay, leaving the other isolated; the label comes from the simulator rerun on the resected graph. **d**, label against the surviving relay's true excitability, one point per test pair ($n = 420$): AUROC $0.998/0.999$ here and 0.999 over all 2,800 patients; the dashed threshold at -2.13 is $0.976/0.990$ accurate here, 0.981 overall (Appendix C).

Training objective. EWM trains without intervention labels, on the MSE over the held-out steps $[96:384]$, a lighter reconstruction term on the encoding window (weight 0.2), and a physical regularizer that bounds the latent and penalizes the learned residual (weight 0.1); the coefficients were never tuned (Eq. 10, Appendix A). TGR-WM trains on the two reconstruction terms alone, and the state-access variants change only when true x_0 enters (Table 9).

4 Experiments

The protocol in Table 1 governs all results. Within each setting, the confirmatory endpoint is the paired EWM–Static AUROC difference over both intervention rows per patient, with bootstrap resampling clustered by patient. Within-patient selection is descriptive; all other intervals are controls, descriptive contrasts, or post-hoc analyses without multiplicity correction.

Benchmark, and what it measures. Clinical data ordinarily contain the outcome of the resection performed, not both candidate outcomes. The Epileptor simulator supplies both outcomes for each virtual patient, enabling a paired endpoint (Jirsa et al., 2014) (Figure 4). The label comes from rerunning the simulator under each resection, positive when the most active remaining node exceeds 2.2 times the patient's baseline median line length (multipliers 2.0–2.4 change no decision category; Figure 6). The two interventions differ only in which relay survives, and that relay's excitability alone predicts the label at AUROC 0.999 over all 2,800 patients (Appendix C): the benchmark tests whether a model recovers one intervention-relevant latent state, not whether it simulates propagation. This makes the benchmark a deliberately favorable case for certification, requiring recovery of one intervention-relevant state rather than propagation.

4.1 All rollout models pass the forecasting gate

All eleven rollout models, comprising the five in the protocol and six controls (Section 4.4), pass in both settings under training-context-only normalization, every skill mean above 0.5 and every lower bound well above zero (Table 2). The gate holds on the intervened graph as well: scored against the intervened simulator's own trajectory, every rollout on A_R has MSE skill 0.68–0.80 against the context mean and 0.08–0.52 against persistence, all lower bounds above zero (Table 22). The metric does not certify the seizure structure the label depends on: decoded line lengths are about one percent of the simulator's, and the within-patient ordering of nodes is nearly uncorrelated with it (Table 22). These task-relevant quantities serve as post-gate diagnostics of transfer. An event-level alternative is degenerate here: the epileptogenic node and both relays seize in every patient and no peripheral node ever does (Appendix E). Passing the gate makes a general forecasting failure less plausible, but does not show that a model recovered the state needed for intervention ranking. Across ten EWM members at 16 nodes, forecast skill spans 0.651–0.686 while intervention AUROC spans 0.335–0.751; the two are uncorrelated (Spearman $\rho = -0.01$), whereas member-level ρ_{surv} is associated with intervention AUROC (Spearman $\rho = 0.76$; Appendix B).

Table 2: **Every rollout model clears the forecasting gate.** Patient-level skill against the context-mean and persistence baselines (95% patient-bootstrap CI). Pass requires mean skill and the lower CI bound to exceed zero for both baselines (Table 1); all 22 model–setting cells pass, and the smallest lower bound across both baselines is $+0.512$. Lower block: controls and encoder swaps (Appendix E).

		12-node		16-node	
Rollout model	Params	Mean [CI]	Persistence [CI]	Mean [CI]	Persistence [CI]
EWM ODE	4,958	0.601 [0.578, 0.623]	0.612 [0.587, 0.636]	0.679 [0.658, 0.699]	0.642 [0.618, 0.665]
GRU-core	4,506	0.531 [0.512, 0.550]	0.547 [0.526, 0.568]	0.622 [0.602, 0.641]	0.579 [0.557, 0.600]
TGR-WM	12,858	0.718 [0.701, 0.733]	0.720 [0.700, 0.739]	0.749 [0.734, 0.764]	0.720 [0.702, 0.737]
x_0-supervised	4,958	0.586 [0.560, 0.609]	0.600 [0.574, 0.625]	0.680 [0.660, 0.698]	0.643 [0.620, 0.664]
Oracle-conditioned	4,958	0.694 [0.669, 0.717]	0.705 [0.680, 0.728]	0.747 [0.726, 0.765]	0.718 [0.697, 0.738]
RSSM	12,836	0.721 [0.703, 0.737]	0.723 [0.702, 0.743]	0.785 [0.770, 0.799]	0.759 [0.740, 0.776]
TWM	11,074	0.573 [0.553, 0.592]	0.586 [0.564, 0.608]	0.648 [0.629, 0.666]	0.608 [0.585, 0.630]
TGR-WM small	4,953	0.702 [0.685, 0.718]	0.708 [0.688, 0.727]	0.766 [0.752, 0.779]	0.738 [0.720, 0.754]
EWM ODE wide	12,864	0.603 [0.581, 0.623]	0.616 [0.592, 0.638]	0.671 [0.651, 0.690]	0.632 [0.609, 0.653]
EWM ODE + temporal encoder	13,310	0.718 [0.698, 0.737]	0.722 [0.700, 0.743]	0.788 [0.774, 0.800]	0.761 [0.744, 0.777]
TGR-WM + static encoder	4,506	0.578 [0.556, 0.598]	0.592 [0.568, 0.614]	0.642 [0.622, 0.661]	0.601 [0.579, 0.623]

Table 3: **Intervention discrimination and paired difference versus Static.** AUROC and paired Δ (95% CI). *Supervision*: information available beyond the observation window. *Ev.*: C, confirmatory pair; D, descriptive/control; H, post-hoc. Classes: † clears the SESOI; * is positive but does not clear it; $^\circ$ spans zero and the SESOI; ‡ excludes the SESOI without establishing a positive gain (Table 1). Variants and swaps: Table 21; additional contrasts: Appendix B.

				12-node		16-node	
	Model	Ev.	Supervision	AUROC	Δ vs Static [95% CI]	AUROC	Δ vs Static [95% CI]
No rollout	Static (floor)	C	labels	0.743	reference	0.701	reference
	GIN	D	labels	0.810	$+0.067$ $[+0.026, +0.108]^*$	0.823	$+0.123$ $[+0.077, +0.169]^\dagger$
	No-graph MLP	H	labels	0.801	$+0.058$ $[+0.013, +0.104]^*$	0.823	$+0.122$ $[+0.075, +0.172]^\dagger$
	Temporal head	H	labels	0.907	$+0.164$ $[+0.119, +0.213]^\dagger$	0.939	$+0.238$ $[+0.185, +0.290]^\dagger$
Rollout (label-free)	EWM ODE	C	none	0.519	-0.224 $[-0.305, -0.147]^\ddagger$	0.509	-0.192 $[-0.268, -0.117]^\ddagger$
	GRU-core	D	none	0.527	-0.216 $[-0.290, -0.142]^\ddagger$	0.520	-0.181 $[-0.261, -0.102]^\ddagger$
	TGR-WM	D	none	0.793	$+0.049$ $[-0.010, +0.111]^\circ$	0.912	$+0.211$ $[+0.157, +0.266]^\dagger$
	RSSM (Dreamer)	D	none	0.687	-0.057 $[-0.119, +0.005]^\ddagger$	0.692	-0.009 $[-0.074, +0.056]^\circ$
	TWM (transformer)	D	none	0.600	-0.143 $[-0.199, -0.086]^\ddagger$	0.743	$+0.042$ $[-0.008, +0.093]^\circ$
Rollout + sup.	EWM + probe	H	labels	0.761	$+0.018$ $[-0.020, +0.055]^\circ$	0.804	$+0.103$ $[+0.058, +0.149]^\dagger$
	x_0-supervised	D	x_0 (train)	0.725	-0.018 $[-0.064, +0.025]^\ddagger$	0.720	$+0.020$ $[-0.032, +0.072]^\circ$
	Oracle-conditioned	D	x_0 (eval)	0.948	$+0.205$ $[+0.160, +0.251]^\dagger$	0.975	$+0.275$ $[+0.222, +0.327]^\dagger$

4.2 Held-out forecasting skill does not certify intervention value

After passing both forecasting gates, EWM fell below Static: AUROC 0.519 vs. 0.743 at 12 nodes ($\Delta = -0.224$ $[-0.305, -0.147]$) and 0.509 vs. 0.701 at 16 nodes ($\Delta = -0.192$ $[-0.268, -0.117]$). Both paired intervals lie entirely below zero and therefore also exclude the prespecified $+0.05$ margin (Table 3). Across these three arms and two settings, five of six cells clear the margin: TGR-WM at 16 nodes, both TGR-WM-small cells, and both EWM-with-temporal-encoder cells. Every aggregation rule leaves the primary EWM–Static floor comparison intact: a mean of ranks gives 0.677 and 0.549 while the label-supervised arms move by at most 0.006 (Table 16). No EWM member or sub-ensemble clears the SESOI over Static (Tables 19 and 24). The confirmatory interval conditions on three fixed ensemble members and captures patient-level rather than training-run uncertainty. In the ten-member analysis, EWM remains below Static at 16 nodes, at 0.624 versus 0.722 (Table 17).

Patient-level ordering gives the complementary decision view: whether a model ranks the two candidate resections correctly for the same patient. On label-discordant patients, within-patient accuracy is $0.484/0.538$ for EWM at 12/16 nodes, compared with $0.758/0.802$ for Static. TGR-WM reaches $0.842/0.934$, and a label-trained probe on frozen EWM latents reaches $0.789/0.813$ (Table 12).

Static, the higher-capacity GIN, and a no-graph MLP reach the AUROCs of Table 3, the MLP matching GIN: the surviving relay's excitability is largely readable from its own pre-intervention statistics, without rollout or graph. A linear head on TGR-WM's own temporal encoder, trained on labels with no rollout, reaches 0.907 and 0.939 and exceeds TGR-WM in both settings (Tables 3 and 14). Because these predictors use intervention labels whereas the world models do not, this

Table 4: **Every negatively oriented member has a static encoder.** No-rollout floor $0.743/0.701$. Per-member ρ_{surv} and five further models: Table 19.

		12-node			16-node		
Model	Encoder	$\rho_{\mathrm{surv}} < 0$	member AUROC	ensemble	$\rho_{\mathrm{surv}} < 0$	member AUROC	ensemble
EWM ODE	static	1/3	0.380, 0.687, 0.723	0.519	1/3	0.751, 0.498, 0.335	0.509
GRU-core	static	2/3	0.497, 0.500, 0.531	0.527	2/3	0.398, 0.581, 0.507	0.520
TGR-WM, static enc.	static	2/3	0.529, 0.488, 0.474	0.488	2/3	0.680, 0.526, 0.570	0.689
TGR-WM	temporal	0/3	0.795, 0.579, 0.714	0.793	0/3	0.716, 0.937, 0.556	0.912
EWM ODE, temporal enc.	temporal	0/3	0.831, 0.839, 0.879	0.875	0/3	0.729, 0.869, 0.725	0.831

comparison varies both the use of rollout and the availability of intervention supervision. For reference, zero-parameter context scores without intervention-label training reach 0.619 and 0.665, compared with EWM's 0.519 and 0.509 (Appendix E).

The no-rollout floor depends on label budget. With rollout scores frozen, the best supervised 16-node AUROC falls from 0.939 at 980 labels to 0.813 at 200 and 0.622 at 25; Static saturates near 0.70 by 200. Against the supervised temporal head, TGR-WM moves from -0.027 $[-0.052, -0.002]$ at 980 labels to $+0.340$ $[+0.280, +0.400]$ at 25 (Appendix D), making 980 the least favourable evaluated budget for this contrast. EWM and GRU-core remain below every label-supervised arm in both settings, including Static at 25 labels ($0.650/0.535$); six intervals there span zero. A post-transient 16-node context raises EWM to 0.904 but not above its Static floor by the margin (Appendix B).

4.3 Readout orientation contributes to EWM's intervention failure

Frozen latents retain intervention-relevant information. A supervised probe (Alain & Bengio, 2017; Belinkov, 2022) pools temporal summaries of frozen EWM latents and selects its head on validation data: a 545-parameter MLP at 12 nodes and a 33-parameter linear head at 16 nodes (Appendix D). It reaches 0.761 and 0.804 (Table 3): level with Static at 12 nodes, beyond the SESOI above it at 16, and below the no-graph MLP.

A learned readout recovers much of the performance gap. A 7-parameter linear head on the decoded per-node mean, standard deviation, and line length recovers most of the gap and exceeds Static at 16 nodes ($+0.094$ $[+0.044, +0.145]$; Figure 1B; Table 5).

Readout orientation is unstable across independently trained members. Within each EWM member, the surviving relay's decoded line length tracks its excitability ($|\rho_{\mathrm{surv}}|$ 0.12–0.58), but the sign varies across members. The three-member mean is therefore nearly uninformative ($\rho_{\mathrm{surv}} = +0.05$ and $+0.03$; Figure 1A; Table 4). Across the eleven rollout models and both settings, ρ_{surv} is strongly associated with intervention AUROC (Spearman $\rho = 0.95$). One bit per member corrects it: choosing the sign on validation patients and rescoring the frozen rollouts lifts EWM to 0.704 and 0.710, gains of $+0.185$ $[+0.098, +0.272]$ and $+0.200$ $[+0.124, +0.276]$ that bring it level with the floor in both settings (Table 18). It changes sign-consistent models little, with direct x_0-supervision yielding similar performance. The bit is itself a label; without it, EWM falls below the zero-parameter baselines above.

Table 5: **Learned readouts recover information from frozen EWM representations.** Frozen EWM ODE, test AUROC; floor $0.743/0.701$. Upper block: fixed readouts. Lower: learned linear.

Input (per node)	Readout	Head params	12-node	16-node
decoded: mean line length	max over nodes (ordinary EWM)	0	0.519	0.509
latent: coordinate-mean temporal std	max over nodes	0	0.600	0.619
decoded: mean, std, line length	linear on mean⊕max pool	7	0.736	0.794
latent: temporal mean, std	linear on mean⊕max pool	33	0.775	0.804
latent: temporal mean, std	linear on max pool	17	0.752	0.800

4.4 Intervention transfer is strongly encoder-dependent

Four contrasts, all passing the gate, ask what the dissociation depends on (Table 21, Appendix E).

Not on the ODE core. Swapping it for a GRU leaves ranking at chance at 16 nodes; the 12-node value reflects ties, 70 of 95 discordant patients tying under its readout (Table 12).

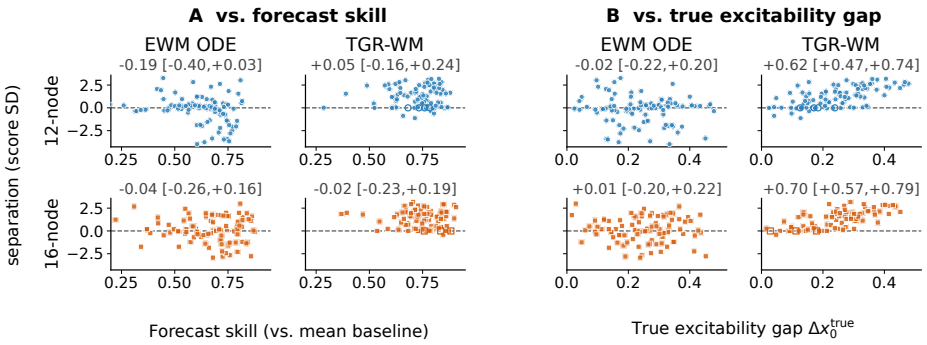


Figure 5: **Within-patient ordering on label-discordant test patients.** Each point gives one patient's seizing-minus-seizure-free score in SD; positive is correct; hollow, ties. **A**, versus forecast skill. **B**, versus the surviving-relay excitability gap. Labels give Spearman ρ with 95% CI; correct fractions and paired contrasts are in Table 12.

Not unique to EWM's design. Neither a Dreamer-family recurrent state-space model (RSSM; Hafner et al., 2019), among the study's strongest forecasters, nor a causal transformer world model (TWM; Micheli et al., 2023; Robine et al., 2023) establishes an advantage.

Not on capacity. Parameter matching both ways preserves the result: TGR-WM shrunk to EWM's count retains its advantage and reaches 0.965 at 16 nodes, the only label-free rollout above the temporal head ($+0.026$ $[+0.010, +0.044]$, short of the SESOI), while widened EWM stays below the floor.

On the encoder. Swapping encoders reverses the ordering in this benchmark, suggesting that encoder choice strongly influences intervention transfer. EWM's ODE core with TGR-WM's temporal encoder clears the SESOI over Static in both settings, whereas TGR-WM's cell with EWM's static encoder falls below the floor in both (Table 4). Because RSSM and TWM also use temporal encoders, the swap points to the specific TGR-WM encoder rather than temporal encoding in general. TGR-WM remains above GIN at 16 nodes ($+0.089$ $[+0.052, +0.128]$, †). At the primary label budget, no evaluated rollout clears the SESOI over the supervised temporal head, and models below Static remain below the stronger baselines (Table 14). On the post-transient window, neither EWM nor TGR-WM clears the SESOI over GIN. The temporal encoder also shows substantially stronger relay–x_0 tracking: $|\rho| = 0.91$–0.94, compared with $|\rho| \leq 0.62$ per member for the six-statistic encoder (Appendix E). Readout orientation then helps explain how much of this encoded signal is exposed by the label-free readout; forecast training did not yield a consistent orientation across members.

Forecast skill does not predict patient-level correctness; TGR-WM separation tracks the true excitability gap (Figure 5).

5 Scope and Conclusion

What the benchmark does and does not test. The confirmatory study uses one synthetic family, two node counts, and a one-state intervention task; it tests intervention-relevant state recovery rather than propagation-dependent intervention (Section 4). Across ten configurations and three outcome definitions, no propagation-dependent setting was oracle-solvable: full excitability reached only 0.71–0.78 AUROC when outcomes depended on propagation (Table 13). The study does not include a real cohort. Single-pulse stimulation recordings could support a real-data test of state recovery, while the full protocol would additionally require intervention outcomes. Why TGR-WM's temporal encoder performs better than the RSSM and TWM encoders remains unexplained.

Conclusion. Held-out forecasting skill did not certify intervention competence even on a one-state task. At the primary label budget, no rollout beat the strongest no-rollout predictor by the SESOI. In EWM, readout orientation contributed to the gap, while frozen latents retained intervention-relevant information; across architectures, transfer was strongly encoder-dependent. Intervention world models should be tested against strong no-rollout baselines at deployment budget.

Reproducibility Statement

Data generation, preprocessing, model configurations, intervention semantics, member aggregation, paired bootstrap inference, and robustness analyses are documented in Appendix B; full EWM component forms appear in Appendix A. For each architecture, the three members' continuous scores are averaged within patient and intervention before AUROC is computed. Appendix B.1 also reports parameter counts, hardware, and per-member training and evaluation time. The supplementary material includes cohort generation and a reference implementation of EWM training, forecast-gate evaluation, intervention scoring, and readout diagnostics.

AI Use Statement

We used generative AI tools to assist with polishing the writing of this paper (grammar, phrasing, and clarity of exposition) and for partial assistance with code and experiment scripting. All AI-assisted text was reviewed and edited by the authors, and all AI-assisted code was reviewed and its outputs verified against logged runs by the authors. We take responsibility for the final content of this work, including its text, claims, and artifacts.

Ethics Statement

This work uses only simulated data generated from a published neural-mass model; no human-subjects data were collected or analyzed. The models are research instruments for studying when learned dynamics aid intervention-conditioned prediction; as the text states, they are not validated for clinical decision support and are not deployment-ready.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations, Workshop Track*, 2017.

Tamir Avigdor, Alyssa Ho, Matthew Moye, William Davalan, Erica Minato, Sana Hannan, Tamzin Holden, Tasha Bouchet, Yingqi Laetitia Wang, Kassem Jaber, Mays Khweileh, Samantha Kaplan, Vojtech Travnicek, David Carlson, and Birgit Frauscher. Epilepsy surgery outcomes and their determinants: a systematic review and individual patient data meta-analysis. *Journal of Neurology, Neurosurgery & Psychiatry*, 97(7):639–648, 2026. doi: 10.1136/jnnp-2025-337158.

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

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Ioana Bica, Ahmed M. Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations*, 2020.

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

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Silvia Chiappa, Sébastien Racanière, Daan Wierstra, and Shakir Mohamed. Recurrent environment simulators. In *International Conference on Learning Representations*, 2017.

Fei Deng, Junyeong Park, and Sungjin Ahn. Facing off world model backbones: RNNs, Transformers, and S4. In *Advances in Neural Information Processing Systems*, 2023.

Amir-massoud Farahmand, André Barreto, and Daniel Nikovski. Value-aware loss function for model-based reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, 2017.

Ezequiel Gleichgerrcht, Brent Munsell, Sonal Bhatia, William A Vandergrift III, Chris Rorden, Carrie McDonald, Jonathan Edwards, Ruben Kuzniecky, and Leonardo Bonilha. Deep learning applied to whole-brain connectome to determine seizure control after epilepsy surgery. *Epilepsia*, 59(9):1643–1654, 2018.

Christopher Grimm, André Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024.

Meysam Hashemi, Anirudh N Vattikonda, Jayant Jha, Viktor Sip, Marmaduke M Woodman, Fabrice Bartolomei, and Viktor K Jirsa. Amortized bayesian inference on generative dynamical network models of epilepsy using deep neural density estimators. *Neural Networks*, 163:178–194, 2023.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, 2019.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations*, 2024.

Viktor K Jirsa, William C Stacey, Pascale P Quilichini, Anton I Ivanov, and Christophe Bernard. On the nature of seizure dynamics. *Brain*, 137(8):2210–2230, 2014.

Viktor K Jirsa, Timothée Proix, Dionysios Perdikis, Michael Marmaduke Woodman, Huifang Wang, Jorge Gonzalez-Martinez, Christophe Bernard, Christian Bénar, Maxime Guye, Patrick Chauvel, et al. The virtual epileptic patient: individualized whole-brain models of epilepsy spread. *Neuroimage*, 145:377–388, 2017.

Keon Woo Kim. Latent state design for world models under sufficiency constraints. *arXiv preprint arXiv:2605.01694*, 2026.

Lohith G. Kini, John M. Bernabei, et al. Virtual resection predicts surgical outcome for drug-resistant epilepsy. *Brain*, 142(12):3892–3905, 2019.

Thomas Kipf, Elise van der Pol, and Max Welling. Contrastive learning of structured world models. In *International Conference on Learning Representations*, 2020.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Nathan Lambert, Brandon Amos, Omry Yadan, and Roberto Calandra. Objective mismatch in model-based reinforcement learning. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pp. 761–770. PMLR, 2020. URL https://proceedings.mlr.press/v120/lambert20a.html.

Adam Li, Chester Huynh, Zachary Fitzgerald, Iahn Cajigas, Damian Brusko, Jonathan Jagid, Angel O Claudio, Andres M Kanner, Jennifer Hopp, Stephanie Chen, et al. Neural fragility as an eeg marker of the seizure onset zone. *Nature neuroscience*, 24(10):1465–1474, 2021.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. WorldModelBench: Judging video generation models as world models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Causal representation learning for instantaneous and temporal effects in interactive systems. In *International Conference on Learning Representations*, 2023.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *International Conference on Machine Learning*, 2022.

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

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

Emilio Perucca, Piero Perucca, H. Steve White, and Elaine C. Wirrell. Drug resistance in epilepsy. *The Lancet Neurology*, 22(8):723–734, 2023. doi: 10.1016/S1474-4422(23)00151-5.

Timothée Proix, Fabrice Bartolomei, Maxime Guye, and Viktor K Jirsa. Individual brain structure and modelling predict seizure propagation. *Brain*, 140(3):641–654, 2017.

Mohammad Areeb Qazi, Maryam Nadeem, and Mohammad Yaqub. Beyond generative ai: World models for clinical prediction, counterfactuals, and planning. *arXiv preprint arXiv:2511.16333*, 2025.

Jan Robine, Marc Höftmann, Tobias Uelwer, and Stefan Harmeling. Transformer-based world models are happy with 100k interactions. In *International Conference on Learning Representations*, 2023.

Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International conference on machine learning*, pp. 8459–8468. PMLR, 2020.

Paula Sanz Leon, Stuart A Knock, M Marmaduke Woodman, Lia Domide, Jochen Mersmann, Anthony R McIntosh, and Viktor Jirsa. The virtual brain: a simulator of primate brain network dynamics. *Frontiers in neuroinformatics*, 7:10, 2013.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588:604–609, 2020.

Nishant Sinha, Justin Dauwels, Marcus Kaiser, Sydney S Cash, M Brandon Westover, Yujiang Wang, and Peter N Taylor. Predicting neurosurgical outcomes in focal epilepsy patients using computational modelling. *Brain*, 140(2):319–332, 2017.

Adam J. Thorpe, Stepan Tretiakov, Cheng-Hsi Hsiao, Su Ann Low, Xingjian Li, Hassan Iqbal, Neel P. Bhatt, Ufuk Topcu, and Krishna Kumar. Physically viable world models: A case for query-conditioned embodied ai. *arXiv preprint arXiv:2605.30542*, 2026.

Dorien Van Blooijs, Max A Van Den Boom, Jaap F Van Der Aar, Geertjan M Huiskamp, Giulio Castegnaro, Matteo Demuru, Willemiek JEM Zweiphenning, Pieter Van Eijsden, Kai J Miller, Frans SS Leijten, et al. Developmental trajectory of transmission speed in the human brain. *Nature Neuroscience*, 26(4):537–541, 2023.

Max A van den Boom, Nicholas M Gregg, Gabriela Ojeda Valencia, Brian N Lundstrom, Kai J Miller, Dorien van Blooijs, Geertjan JM Huiskamp, Frans SS Leijten, Gregory A Worrell, and Dora Hermes. Er-detect: a pipeline for robust detection of early evoked responses in bids-ieeg electrical stimulation data. *Journal of Neuroscience Methods*, 418:110389, 2025.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. BrainBERT: Self-supervised representation learning for intracranial recordings. In *International Conference on Learning Representations*, 2023.

Huifang E Wang, Borana Dollomaja, Paul Triebkorn, Gian Marco Duma, Adam Williamson, Julia Makhalova, Jean-Didier Lemarechal, Fabrice Bartolomei, and Viktor Jirsa. Virtual brain twins for stimulation in epilepsy. *Nature computational science*, 5(9):754–768, 2025.

Tianzhuo Yang, Zihan Shen, Zirui Mi, Zhaoyi Zhang, Jiayi Zhou, Jiaming Ji, Juntao Dai, Jiawei Chen, Boyuan Chen, and Yaodong Yang. Mirabench: Evaluating action-conditioned reliability in robotic world models. *arXiv preprint arXiv:2605.29360*, 2026.

Yijun Yang, Zhao-Yang Wang, Qiuping Liu, Shuwen Sun, Kang Wang, Rama Chellappa, Zongwei Zhou, Alan Yuille, Lei Zhu, Yu-Dong Zhang, and Jieneng Chen. Medical world model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8319–8329, 2025.

Yang Yu, Shiyuan Zhang, Yifei Sheng, Haoxiang Ren, and Haoxin Lin. How should world models be evaluated for embodied decision-making? A decision-making-centric position. *arXiv preprint arXiv:2606.15032*, 2026.

Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal M Patel, Paul Pu Liang, Daniel Khashabi, Cheng Peng, Rama Chellappa, Tianmin Shu, Alan Yuille, Yilun Du, and Jieneng Chen. World-in-world: World models in a closed-loop world. In *International Conference on Learning Representations*, 2026a.

Xiangteng Zhang, Yang Guan, Bo Zhang, Hongyang Li, Ya-Qin Zhang, and Shengbo Eben Li. On the identifiability of controlled world models. *arXiv preprint arXiv:2607.22430*, 2026b.

Appendix Contents

Appendix A. Model details
Explicit EWM encoder, vector field, training objective, and implementation details.

Appendix B. Confirmatory protocol and reproducibility details
Cohort generation, intervention labels, model configurations, paired inference, post-transient robustness, and patient-level evaluation records.

Appendix C. Benchmark structure and the propagation-setting gate
Structure of the intervention task and the attempted propagation-dependent benchmark variants.

Appendix D. Readout and robustness analyses
Supervised probes, readout ablations, stronger no-rollout predictors, label-budget sensitivity, member uncertainty, orientation, and readout compatibility.

Appendix E. Comparison models, encoder controls, and additional robustness
Additional world models, capacity matching, encoder swaps, intervened forecasting, member robustness, additional member sets, and encoder diagnostics.

Appendix F. Extended related work
Intervention prediction, epilepsy modelling, and world-model evaluation literature.

Appendix G. Notation summary
Key symbols and quantities used throughout the manuscript and appendices.

All appendix titles above are clickable in the PDF.

A MODEL DETAILS

This appendix gives the explicit forms deferred from the Method.

Encoder heads. With h_n the shared-MLP embedding of the six per-node statistics and $\bar{h}_n = \frac{1}{\max(\deg_n,1)} \sum_m A_{nm} h_m$ the degree-normalized one-hop aggregate, two linear heads produce

$$z_n(0) = 2\tanh\big(W_z[h_n;\bar{h}_n]\big), \qquad \hat{x}_{0,n} = W_x[h_n;\bar{h}_n], \tag{3}$$

where $z_n(0)$ initializes the rollout and $\hat{x}_{0,n}$ is an unconstrained excitability-like rollout coordinate.

Vector field. Four additive terms compose the field, bounded and rescaled before integration,

$$F_n(z) = f_{\mathrm{Epi}}(z, A; \hat{x}_0)_n + g_{\mathrm{GNN}}(z, A)_n + h(z_n) - \gamma z_n, \qquad \dot{z}_n = \mathrm{rate} \cdot C\, \tanh\big(F_n(z)/C\big), \tag{4}$$

where $C\tanh(\cdot/C)$ bounds the field, rate sets the timescale and γ is the leak. Writing u_n, w_n for the first two coordinates of z_n, the physical prior $f_{\mathrm{Epi}}(z, A; \hat{x}_0)_n$ is a two-dimensional reduced Epileptor,

$$\dot{u}_n = 1 - u_n^3 - 2u_n^2 - w_n + \kappa \sum_m A_{nm}(u_m - u_n), \tag{5}$$

$$\dot{w}_n = \big(4(u_n - \hat{x}_{0,n}) - w_n\big)/\tau, \tag{6}$$

where the learned coordinate $\hat{x}_{0,n}$ occupies the bifurcation-control position and the difference coupling $\kappa \sum_m A_{nm}(u_m - u_n)$ spreads activity along the graph. This notation does not identify $\hat{x}_0$ with simulator truth x_0^{true}. The graph message-passing term couples the full latent states,

$$g_{\mathrm{GNN}}(z, A)_n = \tfrac{1}{\max(\deg_n,1)} \sum_m A_{nm}\, \mathrm{MLP}_g\big([z_n; z_m]\big); \tag{7}$$

the degree clamp keeps the message mean defined on resected graphs, where a removed node and its former neighbours can reach zero degree. The learned residual $h(z_n)$ is an MLP of the latent state, distinct from the encoder embedding h_n.

Training objective. The objective combines a reconstruction loss that weights the held-out window more heavily and a physics-consistency term that keeps the latent bounded and the learned residual small,

$$\mathcal{L}_{\mathrm{recon}} = \mathrm{MSE}\big(\hat{X}[t_{\mathrm{enc}}:],\, X[t_{\mathrm{enc}}:]\big) + 0.2\,\mathrm{MSE}\big(\hat{X}[:t_{\mathrm{enc}}],\, X[:t_{\mathrm{enc}}]\big), \tag{8}$$

$$\mathcal{L}_{\mathrm{phys}} = \mathrm{mean}\big[\mathrm{ReLU}(|z| - z_{\max})^2\big] + 0.1\ \mathrm{mean}\big[h(z)^2\big], \tag{9}$$

giving the objective optimized with Adam,

$$\mathcal{L} = \mathcal{L}_{\mathrm{recon}} + 0.1\,\mathcal{L}_{\mathrm{phys}}. \tag{10}$$

Throughout, MSE denotes the per-element mean over the indicated slice. The residual term is computed as the elementwise mean of $h(z)^2$ over at most 4,096 randomly sampled latent states and all eight latent coordinates.

Implementation details. Each node carries an eight-dimensional latent state whose first two coordinates are the Epileptor variables; all MLPs have width 64 with SiLU activations; the dynamics are integrated with fixed-step RK4 at unit step, the field is bounded at $C{=}5$, and rate, time scale, leak, and coupling strength are learned scalars. Ordinary EWM has 4,958 parameters. Full hyperparameters are in Table 10.

B CONFIRMATORY PROTOCOL AND REPRODUCIBILITY DETAILS

Confirmatory cohort generation. The two node counts were fixed before the confirmatory cohorts were generated. The independently generated confirmatory cohorts contain 1,400 12-node and 1,400 16-node patients, each split $980/210/210$ into train, validation, and test. Nodes 0, 1, and 2 are respectively the epileptogenic zone (EZ) and two relays. The symmetric adjacency has weight 2.0

from the EZ to each relay and weight 0.2 along the undirected peripheral chain $3{-}4{-}\cdots{-}(N-1)$; all other edges and tract lengths are zero, so no transmission delays are used. Simulator excitability is $x_{0,0}^{\text{true}} = -1.6$, $x_{0,n}^{\text{true}} = -2.4$ for peripheral nodes, and independently $x_{0,r}^{\text{true}} = -2.05 + 0.25U_r$, $U_r \sim \text{Uniform}[-1, 1]$, for relays $r \in \{1, 2\}$. Node coordinates are sampled from $\mathcal{N}(0, 30^2)$ but do not affect the zero-delay dynamics.

Each trajectory is generated with TVB's Epileptor ($K_s = -0.2$, $r = 0.00035$), Difference coupling $a = 2.0$, and stochastic Heun integration with step 0.05, using `tvb-library` 2.10. Additive noise uses `nsig` $(0, 0, 0, 10^{-4}, 10^{-4}, 0)$, and a temporal-average monitor samples every 1.0 time unit for a duration of $8{,}000$. Initial conditions are not explicitly supplied, so the simulator's configured default initialization is used; each baseline and intervention has its own independent deterministic random stream. The stored observation is the monitor's first observable, $x_2 - x_1$. No additional observation noise is added. For model input, the observation is decimated by 16 and truncated to 384 steps; steps $0{:}96$ form the context and $96{:}384$ the held-out forecast target.

Normalization is fitted independently for each node-count setting using only the decimated context samples $[0:96]$ from its 980 training patients. One scalar mean and population standard deviation (ddof 0) are pooled over patients, nodes, and context time points, with denominator standard deviation plus 10^{-6}. The fitted values are $(\mu, \sigma) = (0.5402, 0.7522)$ for 12 nodes and $(0.5553, 0.7412)$ for 16 nodes. Each setting's scalar transform is then applied unchanged to the full 384-step trajectories in train, validation, and test. Forecast targets, validation patients, and test patients contribute nothing to the fitted statistics.

Patient parameters, baseline simulations, intervention simulations, and split assignment use disjoint deterministic random streams. Patients are assigned by sorting independently generated split keys rather than labels.

Ground-truth intervention label. The two candidate actions remove the epileptogenic zone together with one relay: nodes $[0, 1]$ and nodes $[0, 2]$ respectively. For ground truth, resection zeros both adjacency rows and columns of the removed nodes, sets their simulator excitability to the healthy value -2.4, and reruns the Epileptor with an intervention-specific deterministic random stream. Let

$$L_n(Y) = \sum_{t=2000}^{7998} |Y_n(t+1) - Y_n(t)|, \qquad \theta_i = 2.2 \ \text{median}_n \, L_n(Y_i^{\text{pre}}). \tag{11}$$

The first 2,000 monitor samples are therefore discarded. The model input, 384 decimated steps, covers monitor samples 0–6,143 of the baseline run, so the 96-step context lies within the interval the label computation discards as transient: the models read excitability from the settling phase of the baseline run, while the label integrates the intervened run after its transient. The patient–intervention label is

$$y_i(a) = \mathbb{1}\left\{ \max_n L_n(Y_i^{(a)}) > \theta_i \right\}, \tag{12}$$

where $y = 1$ means residual network seizure and $y = 0$ means seizure-free (label counts: Table 6). This label comes only from the intervened simulator trajectory. It is distinct from every model's continuous score: rollout models use mean line length over normalized predicted steps $[96:384]$ and maximize over remaining nodes, while Static and GIN use classifier logits.

Table 6: Confirmatory label counts. Entries are positive residual-seizure / negative seizure-free patient–intervention rows. Each intervention contributes one row per patient.

Setting	Split (patients)	EZ+relay 1	EZ+relay 2	Combined
12-node	train (980)	638 / 342	642 / 338	1280 / 680
	validation (210)	134 / 76	149 / 61	283 / 137
	test (210)	133 / 77	122 / 88	255 / 165
16-node	train (980)	640 / 340	634 / 346	1274 / 686
	validation (210)	125 / 85	134 / 76	259 / 161
	test (210)	128 / 82	139 / 71	267 / 153

Post-transient context (post-hoc). The original 96-step context covers decimated steps $[0:96]$ (monitor samples 0–1,535), within the $2{,}000$ samples discarded as transient by the label computation. To test the effect of this window, the three label-supervised predictors were retrained on an

equally long post-transient context, steps $[125:221]$ (monitor samples 2,000–3,535), with the scalar normalizer refitted on the corresponding training context and cohorts, labels, splits, architectures, and optimization otherwise unchanged. All six cells rise (Table 7): Static reaches $0.862/0.898$, GIN $0.869/0.930$, and the supervised temporal head $0.982/0.978$. The original window is therefore less favourable to the no-rollout floor than the tested post-transient window. Refitting the original window with the same code reproduces the frozen normalizers to 2.7×10^{-14} and 1.4×10^{-14}, and the original-window column recomputes the published AUROC in every cell.

Both rollout models were then retrained on the post-transient window at 16 nodes and again passed the forecasting gate. EWM ODE reached 0.904 against the same-window Static value of 0.898, $\Delta = +0.006$ $[-0.012, +0.025]$, still below the $+0.05$ margin; TGR-WM reached 0.991, $\Delta = +0.093$ $[+0.066, +0.123]$. Against GIN (0.930), the corresponding differences were -0.026 $[-0.046, -0.007]$ for EWM and $+0.061$ $[+0.041, +0.084]$ for TGR-WM. Against the temporal head (0.978), they were -0.074 $[-0.100, -0.049]$ and $+0.013$ $[+0.005, +0.024]$, respectively (Table 14). Because an AUROC of 0.978 leaves no feasible $+0.05$ margin, the temporal-head contrast on this window is directional; the GIN contrast remains a feasible margin test. The available post-transient forecast segment contains 279 rather than 288 steps; rescoring over the matched horizon changes every AUROC by at most 0.0008 and no decision class.

What the window changes is the orientation. All three EWM members are positively oriented on the post-transient window ($\rho_{\mathrm{surv}} = +0.79, +0.72, +0.65$; three-member mean $+0.767$ $[+0.728, +0.799]$), whereas one of three is negatively oriented on the original window. The post-transient ensemble therefore reaches 0.904 rather than 0.509, but still does not exceed its same-window floor by the $+0.05$ margin.

Table 7: **Post-transient context.** Test AUROC with an equally long context at decimated steps $[125:221]$; original-window values use steps $[0:96]$. Paired patient-clustered bootstrap, 10,000 resamples. Rollout models were retrained at 16 nodes only; stronger-floor contrasts are in Table 14.

	12-node			16-node		
Model	post-transient	original	Δ [95% CI]	post-transient	original	Δ [95% CI]
Label-supervised predictors (Δ: this model on the two windows)						
Static	0.862	0.743	$+0.118$ $[+0.061, +0.177]$	0.898	0.701	$+0.197$ $[+0.131, +0.262]$
GIN	0.869	0.810	$+0.059$ $[+0.011, +0.108]$	0.930	0.823	$+0.107$ $[+0.061, +0.154]$
Temporal head	0.982	0.907	$+0.075$ $[+0.051, +0.101]$	0.978	0.939	$+0.039$ $[+0.018, +0.063]$
Rollout models, 16 nodes only (Δ: this model against Static on the post-transient window)						
EWM ODE	N/A	N/A	N/A	0.904	0.509	$+0.006$ $[-0.012, +0.025]^{\ddagger}$
TGR-WM	N/A	N/A	N/A	0.991	0.912	$+0.093$ $[+0.066, +0.123]^{\dagger}$

Post-hoc ground-truth threshold sensitivity. The primary label applies a multiplier of 2.2 to each patient's median baseline node line length. Test-set labels were recomputed offline at multipliers 2.0, 2.1, 2.3, and 2.4 from the stored line-length arrays (the 2.2 derivation reproduces the frozen labels exactly), with all scores and member sets fixed. At most 9 of 210 test patients change label in any cell, architecture-level AUROCs move little (Table 8), and all 16 alternative-threshold intervals for EWM–Static and TGR-WM–Static keep their primary decision class (Figure 6).

Table 8: **Post-hoc label-threshold sensitivity: architecture-level test AUROC.** Ground-truth labels are re-derived at each multiplier of the per-patient baseline line-length median; 2.2 is the pre-specified value and reproduces the frozen labels exactly (zero mismatches in 840 rows). Canonical member sets; no simulation or model inference was rerun.

Multiplier	Static	GIN	EWM ODE	GRU-core	TGR-WM	x_0-sup.	Oracle
12-node							
2.0	0.733	0.812	0.523	0.530	0.786	0.712	0.946
2.1	0.742	0.810	0.516	0.530	0.794	0.716	0.945
2.2	0.743	0.810	0.519	0.527	0.793	0.725	0.948
2.3	0.747	0.818	0.505	0.531	0.793	0.717	0.945
2.4	0.749	0.822	0.487	0.531	0.789	0.718	0.942
16-node							
2.0	0.700	0.822	0.508	0.520	0.899	0.718	0.972
2.1	0.703	0.826	0.513	0.512	0.908	0.722	0.977
2.2	0.701	0.823	0.509	0.520	0.912	0.720	0.975
2.3	0.699	0.814	0.512	0.525	0.913	0.718	0.974
2.4	0.702	0.815	0.522	0.530	0.922	0.717	0.974

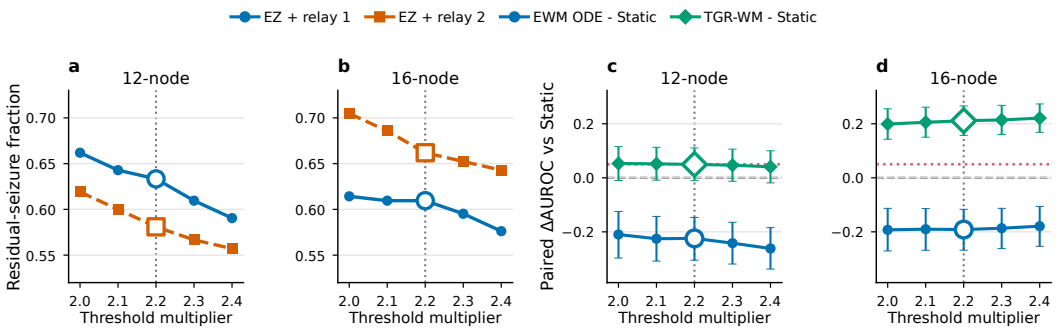


Figure 6: **Label-threshold sensitivity. (a)**, **(b)**, residual-seizure label fraction across multipliers 2.0–2.4; **(c)**, **(d)**, paired ΔAUROC versus Static (whiskers, 95% CIs; dotted, +0.05 SESOI; open markers, the primary value 2.2); no SESOI conclusion changes.

B.1 Model and training configurations

Table 9: **Models under comparison, information access, and the contrast each probes.** No forecast-trained world model receives intervention labels; label-supervised controls are explicitly marked in the training-objective column. For EWM + probe, only the probe head is trained on intervention labels.

Model	Dynamics core	Training objective	True x_0 access	Probes
EWM ODE	Epileptor-prior ODE	forecast + physics reg.	none	main world model
GRU-core	graph message, GRU cell	forecast + latent-bound reg.	none	same-class core swap
TGR-WM	temporal GRU, GRU-cell rollout	forecast	none	second rollout architecture
x_0-supervised	Epileptor-prior ODE	EWM obj. + MSE on $\hat{x}_0$	training only	train-time supervision
Oracle-conditioned	Epileptor-prior ODE	forecast + physics reg.	train and eval	privileged ceiling
No-graph MLP	none (node features + mask)	intervention labels	none	graph-free floor
EWM + probe	frozen EWM latents, small head	head: intervention labels	none	readout diagnostic
Temporal enc. + head	none (TGR-WM encoder)	intervention labels	none	encoder ceiling, no rollout
RSSM	recurrent state-space	forecast + KL	none	comparison world model
TWM	causal transformer	forecast	none	comparison world model
TGR-WM small	TGR-WM, EWM-size widths	forecast	none	capacity, downward
EWM ODE wide	EWM, TGR-WM-size widths	forecast + physics reg.	none	capacity, upward
EWM ODE, temporal enc.	ODE core, TGR-WM encoder	forecast + physics reg.	none	encoder swap
TGR-WM, static enc.	GRU cell, EWM encoder	forecast	none	encoder swap

The seven protocol models are Static, GIN, EWM ODE, GRU-core, TGR-WM, x_0-supervised, and Oracle-conditioned. Table 9 summarizes the rollout, state-access, and additional control models; Static and GIN are specified in Table 11. All models receive only the normalized 96-step pre-intervention window and baseline graph unless explicitly identified as a true-state control. Static and GIN compute six per-node context summaries (mean, standard deviation, minimum, maximum, final value, and mean adjacent absolute difference). For each candidate they use the post-resection adjacency and a keep mask: removed nodes are disconnected before message passing and excluded from mean pooling. Thus the no-rollout models use the same source observation, graph, and intervention definition as the rollout models, but their continuous outputs are intervention-conditioned logits rather than the rollout line-length score. Table 10 lists EWM's confirmatory hyperparameters and Table 11 the no-rollout comparator configurations.

Table 10: Hyperparameters of EWM in the confirmatory experiments. The table describes EWM ODE; N/A denotes a parameter not used by the fixed-grid implementation.

Group	Hyperparameter	Confirmatory value
Data	Node-count handling	Size-agnostic dense graph operations; separate models are trained for the 12- and 16-node settings.
	Trainable parameters	4,958; identical in both settings and across the three members.
	Input/trajectory length	Normalized $[N, 384]$ trajectory; steps 0:96 are the encoder context and 96:384 are the 288-step held-out forecast segment.
	Decimation	16 from the stored simulator observation.
	Per-node input statistics	Six: mean, standard deviation, minimum, maximum, final value, and mean adjacent absolute difference over the 96-step context.
	Normalization	One scalar z-score per node-count setting, fitted only on all training-patient context values [0:96] (ddof 0; denominator standard deviation $+10^{-6}$), then applied unchanged to full train/validation/test trajectories.
Encoder	Node path	Linear $6 \to 64$, followed by SiLU.
	Graph aggregation	One weighted, degree-normalized one-hop aggregation, $Ah/\max(\deg, 1)$.
	Local/graph fusion	Concatenate the 64-dimensional local and aggregated vectors (128 dimensions).
	z_0 head	Linear $128 \to 8$, followed by $2\tanh(\cdot)$; each coordinate is bounded to $[-2, 2]$.
	$\hat{x}_0$ head	Linear $128 \to 1$ per node; no explicit output bound or normalization.
Dynamics	Latent dimension	8 per node.
	Structured coordinates	Coordinates 1–2 are reduced-Epileptor (u, w); coordinates 3–8 are learned latent coordinates.
	Graph message MLP	Concatenated sender/receiver state $16 \to 64 \to 8$, with SiLU; weighted degree-normalized message mean.
	Message-passing depth	One graph-message operation per ODE right-hand-side evaluation.
	Residual field	$8 \to 64 \to 8$ MLP with SiLU.
	Difference coupling	Learned scalar κ, initialized to 1.0.
	Slow time scale	$\tau = \mathrm{softplus}(\ell_\tau) + 1$, with ℓ_τ initialized to 2.0.
	Leak γ	$\gamma = \mathrm{softplus}(\ell_{\mathrm{decay}})$, with ℓ_{decay} initialized to -1.0; applied as $-\gamma z$.
	Integration rate	$r = \mathrm{softplus}(\ell_{\mathrm{rate}}) + 0.01$, with ℓ_{rate} initialized to -2.0; multiplies the full field.
	Field saturation	$C = 5.0$ through $C\tanh(f/C)$.
ODE solver	Solver	`torchdiffeq` fixed-grid fourth-order Runge–Kutta (RK4); adjoint disabled.
	Step size	1.0 normalized model-time unit.
	Rollout points	384 output points ($t = 0, \ldots, 383$).
	Fixed/adaptive	Fixed step.
	Error tolerances	N/A; no `rtol`/`atol` is supplied to the fixed-grid RK4 call.
Decoder	Dimensions	Linear $8 \to 64 \to 1$ per node and time point.
	Activation	SiLU after the first linear layer.
	Output	One normalized trajectory value per node per rollout point.
Training	Optimizer	Adam.
	Learning rate	3×10^{-3}.
	Batch size	8 patients.
	Epochs	100.
	Scheduler	None.
	Weight decay	0.
	Gradient clipping	Global maximum norm 1.0.
	Mixed precision	Not used; no autocast or gradient scaler (standard PyTorch FP32).
	Determinism	Python, NumPy, CPU/GPU tensor-library, and loader-shuffle randomness are initialized independently for each ensemble member, with deterministic algorithm selection.
	Ensemble policy	Three independently trained members per model and setting, fixed before evaluation; members share identical scientific and optimization configuration.
	Accessible splits	Training batches alone contribute optimization gradients. Validation and test data select no checkpoint and no hyperparameter, and test data are excluded from training and model selection.
	Checkpoint rule	Final epoch (100) only; no validation- or test-metric selection.
Loss	Total objective	$\mathcal{L}_{\mathrm{recon}} + 0.1\mathcal{L}_{\mathrm{phys}}$.
	Held-out forecast term	$\mathrm{MSE}_{96:384}$, coefficient 1.0.
	Encoding-window term	$\mathrm{MSE}_{0:96}$, coefficient 0.2.
	Physical regularization	Outer coefficient 0.1.
	Bound term	$\mathrm{mean}\,\mathrm{ReLU}(\lvert z\rvert - 4)^2$.
	Residual-magnitude term	0.1 mean squared residual per latent coordinate inside $\mathcal{L}_{\mathrm{phys}}$.
	Other non-zero terms	None.
Evaluation	Forecast context/horizon	96 context steps / 288 held-out steps.
	Intervention	Zero both rows and columns of removed nodes before rollout.
	Intervention readout	Maximum across remaining nodes of mean adjacent absolute difference over predicted steps 96:384.
	Score direction	Higher is more seizure-like.

Table 11: Reproducible confirmatory no-rollout comparators. Both models use the common training settings shown below.

Component	Static	GIN
Input features	Six per-node 96-step summaries: mean, standard deviation, minimum, maximum, final value, and mean adjacent absolute difference	Same six features
Intervention/adjacency	Weighted post-intervention adjacency with removed rows/columns zero; keep mask excludes removed nodes	Same post-intervention adjacency and keep mask
Graph architecture	$6 \to 64$ SiLU encoder; three residual layers $h \leftarrow h + \mathrm{ReLU}(W(Ah/\max(\deg, 1)))$	Linear $6 \to 64$ encoder; three residual GIN blocks $h \leftarrow h + \mathrm{ReLU}(\mathrm{MLP}((1+\epsilon)h + Ah))$; learned ϵ initialized at 0; each MLP $64 \to 64 \to 64$ with SiLU
Pooling	Arithmetic mean over kept nodes	Arithmetic mean over kept nodes
Classifier head	$64 \to 64$ SiLU $\to 1$	$64 \to 64$ SiLU $\to 1$
Output/loss	Higher-is-more-seizure continuous logit; binary cross-entropy with logits	Same output orientation and loss
Parameters	17,153	29,636

Common optimization. Adam, learning rate 3×10^{-3}, batch size 8, 100 epochs, no scheduler, zero weight decay, gradient-norm clip 1.0, deterministic execution, and final-epoch-only checkpointing. Each model and setting uses three independently trained members with identical configuration. Only training intervention labels contribute to binary cross-entropy optimization. Validation labels do not affect gradients, hyperparameters, weighting, or checkpoint selection, and test labels are excluded from training and model selection.

TGR-WM optimizes the same reconstruction loss as EWM, including the 0.2 encoding-window term, and carries no physics term; GRU-core keeps EWM's objective without the residual-magnitude term, which its core lacks. Every learned member uses the common optimization settings of Table 10; for each member, Python, NumPy, PyTorch, and loader randomness are seeded independently under deterministic algorithms, and the sole checkpoint rule is the final epoch.

The Oracle-conditioned model receives simulator x_0^{true} at the rollout boundary during both training and evaluation and serves as a privileged ceiling.

Confirmatory intervention scores and paired inference. Both adjacency rows and columns are zeroed before rollout or static scoring. Within patient, intervention, model, and setting, continuous scores are averaged over the three-member set before architecture-level AUROC. EWM ODE is compared with Static separately in each setting using 10,000 paired patient-clustered resamples, seeded deterministically from a protocol constant and the node count; each resample retains both interventions and all members of every sampled patient and recomputes the member mean, both AUROCs, and their difference. For an interval $[L, U]$ and SESOI $+0.05$: $L > 0.05$ is a meaningful positive advantage (†); $0 < L \leq 0.05$ a positive advantage short of the threshold ($*$); $L \leq 0$ and $U < 0.05$ excludes a $+0.05$ advantage (‡); $L \leq 0$ and $U \geq 0.05$ detects no advantage while $+0.05$ remains compatible ($\circ$).

Model capacity. Trainable parameter counts are identical in the 12- and 16-node settings and across members, because every module acts on per-node features with degree-normalized dense aggregation and no layer carries a node-count dimension: $17{,}153$ (Static), $29{,}636$ (GIN), $4{,}958$ (EWM ODE, x_0-supervised, and Oracle-conditioned alike), $4{,}506$ (GRU-core), and $12{,}858$ (TGR-WM). Parameter count is therefore unchanged across node-count settings, and the state-access contrasts are capacity-matched. TGR-WM encodes the 96-step context with a temporal GRU of hidden width 32, 64- and 32-dimensional graph projections with layer normalization, a $16 \to 64 \to 8$ edge message MLP, an eight-dimensional rollout GRU cell with 17 inputs, and the same $8 \to 64 \to 1$ decoder as EWM. Its $2.6\times$ larger parameter count is addressed by the matched variants of Appendix E.

Compute. Each member was trained on one GPU (four CPU cores, 32 GB memory) in full FP32 with deterministic algorithms. Per-member wall-clock was 0.19–0.30 hours for the two static comparators, 1.9–3.5 hours for GRU-core and TGR-WM, and 12.2–23.7 hours for the three ODE-core variants; the 42 confirmatory streams total 327.6 single-GPU hours, and test-set inference took 50.1 minutes over all 42. Environment: Python 3.11, PyTorch 2.5.1 (CUDA 12.4), `torchdiffeq` 0.2.5, PyTorch Geometric 2.8.0.

Member sets. Each architecture uses three independently trained members. Test outcomes are excluded from training and model selection, and the same member sets are used for all reported inference; within patient, intervention, model, and setting, member scores are averaged before the architecture-level AUROC is computed, and the patient-bootstrap intervals describe patient-level uncertainty with the member set held fixed.

Within-patient intervention ordering (exploratory). A patient is label-discordant when the two resections have different outcomes (95/210 at 12 nodes; 91/210 at 16). Using the three-member-mean scores, EWM correctly orders $0.484/0.538$ of these patients, against Static's $0.758/0.802$; the paired differences are -0.274 $[-0.411, -0.126]$ and -0.264 $[-0.385, -0.143]$. TGR-WM is above Static at 16 nodes and uncertain at 12, while the EWM probe returns to the Static level without establishing an advantage (Table 12). Ties are reported separately because the maximum-line-length readout can assign both interventions the same unaffected peripheral node. For GRU-core at 12 nodes, the maximum falls on a periphery node in 94.5% of member-level rows, against 30.2% for EWM ODE and 50.7% for TGR-WM.

Patient-level forecast-skill and excitability-gap analysis (exploratory). Figure 5 uses the same discordant test patients and three-member-mean scores. The separation is $s(\text{seizing}) - s(\text{seizure-free})$ divided by the standard deviation of the model's pooled test scores, so that line-length scores and classifier logits are comparable; forecast skill is the three-member-mean skill against the context-mean baseline. Because both resections remove the epileptogenic node, the label is decided by the surviving relay, and in all 186 discordant patients the seizing resection leaves

the more excitable relay, so the gap $\Delta x_0^{\mathrm{true}}$ is positive throughout. Spearman correlations (95% percentile intervals, 3,000 patient resamples) of separation with the gap at 12/16 nodes: EWM ODE -0.02 $[-0.22, +0.20]$ / $+0.01$ $[-0.20, +0.22]$; GRU-core $+0.17$ / -0.01; TGR-WM $+0.62$ $[+0.47, +0.74]$ / $+0.70$ $[+0.57, +0.79]$; Static $+0.40$ / $+0.44$; GIN $+0.53$ / $+0.32$; x_0-supervised $+0.32$ / $+0.40$; Oracle-conditioned $+0.72$ / $+0.85$. Separation with forecast skill excludes zero on the positive side for no architecture in either setting. Median separations in score SD are EWM ODE $-0.04/+0.07$ against TGR-WM $+1.20/+1.52$.

Post-hoc state-access paired contrasts. Computed from the frozen patient-level exports; no model was rerun. The contrasts $\Delta_{x_0} = \mathrm{AUROC}(x_0\text{-supervised}) - \mathrm{AUROC}(\mathrm{EWM})$ and $\Delta_{\mathrm{Oracle}} = \mathrm{AUROC}(\mathrm{Oracle\text{-}conditioned}) - \mathrm{AUROC}(\mathrm{EWM})$ follow the confirmatory rules exactly: identical member aggregation and score orientation, 10,000 paired patient-clustered resamples per setting, and the same deterministic seeding; the only change is the comparator, and all architectures use their fixed triplets. A gate first reproduced the frozen marginal AUROCs and the EWM-minus-Static intervals at full precision. The results are $\Delta_{x_0} = +0.206$ (95% CI $[+0.137, +0.278]$) at 12 nodes and $+0.211$ ($[+0.138, +0.286]$) at 16, and $\Delta_{\mathrm{Oracle}} = +0.429$ ($[+0.372, +0.489]$) and $+0.466$ ($[+0.407, +0.527]$). These are supporting contrasts; no SESOI decision rule is applied.

Table 12: **Within-patient intervention ordering on label-discordant patients.** Accuracy is the fraction for which the residual-seizure resection receives the higher score; ties count as incorrect, and the tie-excluded fraction is descriptive. Intervals are 95% percentile intervals from 10,000 patient-bootstrap resamples; paired differences against Static reuse the same sampled patients. Exploratory analysis; no SESOI classification is applied.

Model	correct	ties	accuracy [95% CI]	excl. ties	paired Δ vs. Static [95% CI]
12-node ($n = 95$ label-discordant patients)					
Static	72/95	0	0.758 [0.663, 0.842]	0.758	reference
GIN	77/95	0	0.811 [0.726, 0.884]	0.811	+0.053 [−0.032, +0.137]
EWM ODE	46/95	0	0.484 [0.379, 0.579]	0.484	−0.274 [−0.411, −0.126]
GRU-core	21/95	70	0.221 [0.137, 0.305]	0.840	−0.537 [−0.653, −0.421]
TGR-WM	80/95	4	0.842 [0.768, 0.916]	0.879	+0.084 [−0.011, +0.189]
x_0-supervised	74/95	2	0.779 [0.695, 0.863]	0.796	+0.021 [−0.063, +0.105]
Oracle-conditioned	90/95	0	0.947 [0.895, 0.989]	0.947	+0.189 [+0.105, +0.274]
No-graph MLP	75/95	0	0.789 [0.705, 0.863]	0.789	+0.032 [−0.053, +0.116]
EWM + probe	75/95	0	0.789 [0.705, 0.863]	0.789	+0.032 [−0.053, +0.116]
16-node ($n = 91$ label-discordant patients)					
Static	73/91	0	0.802 [0.714, 0.879]	0.802	reference
GIN	80/91	0	0.879 [0.802, 0.945]	0.879	+0.077 [+0.022, +0.143]
EWM ODE	49/91	0	0.538 [0.440, 0.637]	0.538	−0.264 [−0.385, −0.143]
GRU-core	49/91	1	0.538 [0.440, 0.637]	0.544	−0.264 [−0.407, −0.121]
TGR-WM	85/91	3	0.934 [0.879, 0.978]	0.966	+0.132 [+0.055, +0.220]
x_0-supervised	57/91	15	0.626 [0.527, 0.725]	0.750	−0.176 [−0.286, −0.066]
Oracle-conditioned	88/91	0	0.967 [0.923, 1.000]	0.967	+0.165 [+0.088, +0.242]
No-graph MLP	80/91	0	0.879 [0.813, 0.945]	0.879	+0.077 [+0.022, +0.143]
EWM + probe	74/91	0	0.813 [0.725, 0.890]	0.813	+0.011 [−0.055, +0.077]

Evaluation records. Patient-level exports retain the setting, model, member, patient, intervention, continuous score, and label; world-model records also retain forecast metrics.

C BENCHMARK STRUCTURE AND THE PROPAGATION-SETTING GATE

Structure of the intervention task. The cohort generator builds two kinds of edges only: weight 2.0 between the epileptogenic node 0 and each relay $(1, 2)$, and weight 0.2 along the peripheral chain $3-4-\cdots-(N-1)$. There is no edge between the two relays and none between $\{0, 1, 2\}$ and the periphery; this holds for all 2,800 patients. Both candidate resections remove node 0, so after either one the surviving relay is isolated and the peripheral chain is unchanged. Consequently a rollout under the two resections differs only at the surviving relay: on the common surviving nodes, the per-node latent statistics of the frozen 12-node EWM members are identical to machine precision across the two resections (maximum absolute difference 0 over 30 test patients), and any scorer, learned or fixed, separates the two resections only through which relay enters the surviving set. The residual-seizure label is decided almost entirely by that relay's true excitability: over all 2,800 patients, x_0^{true} of the surviving relay alone predicts the label at AUROC 0.9986 in both settings,

with a single threshold (-2.126 and -2.127) reaching accuracy 0.981. This is by construction: relay excitability is sampled as $-2.05 + 0.25\,U[-1, 1]$, a band that straddles the seizure threshold. Consistently, the graph-free MLP of Table 3 nearly matches GIN without access to the adjacency.

Attempt to construct a propagation-dependent setting. We evaluated three screens within the same simulator family, fixing each screen and its gates before that screen was run. The first searched for a generator configuration in which the intervention effect reaches the periphery. Each candidate was simulated for 200 patients and required to pass four gates: (G1) the surviving relay's excitability alone predicts the label at AUROC ≤ 0.85; (G2) propagation occurs, with peripheral seizure in more than half of the positive labels; (G3) the label remains predictable from the full simulator state at AUROC ≥ 0.9; and (G4) label balance lies in $[0.40, 0.60]$ with at least 30% label-discordant patients. Candidates varied relay–periphery coupling $W_{RP} \in \{0.6, 1.0\}$ and peripheral excitability centre $c \in \{-2.30, -2.20, -2.13\}$ (baseline -2.4). The uncoupled baseline validated the screen (G2 $= 0$, G3 $= 0.9985$). No configuration passes all four gates (Table 13). As peripheral excitability approaches threshold, intervention-dependent recruitment appears and then gives way to autonomous peripheral seizures; G1 falls only when G4 collapses, while G3 declines from 0.999 to 0.782. Within this generator family, outcomes are therefore either dominated by single-node excitability or fail the full-state solvability gate once propagation becomes substantial.

A second screen changed the outcome rather than the connectome. On the same 200-patient records at $c = -2.20$, the label was redefined as seizure in any peripheral node, with G1 ≤ 0.85, G3 ≥ 0.85, positive rate in $[0.30, 0.70]$, and discordance ≥ 0.25. The task now depends on propagation (G1 $0.712/0.649$; discordance $0.39/0.29$), but the full-state oracle reaches only $0.764/0.714$ (a held-out MLP on all x_0, $0.739/0.680$), so neither configuration passes the solvability gate and no full cohort was generated.

A third screen tested whether this ceiling was caused by stochastic realization noise. On 200 newly simulated patients at $c = -2.20$, each resection was rerun nine times under independent noise streams and the peripheral-seizure label was aggregated over $k \in \{1, 5, 9\}$ realizations. The single-realization result reproduces the second screen to within 0.015; at $k = 9$, G3 remains $0.776/0.726$, with G1, positive rate, and discordance all within 0.03 of their single-realization values. Across realizations, 89% of patient–resection outcomes are already near-certain, only 2–3$\%$ fall in the undecided band, and the modal label differs from the single-realization label for 4.0–4.3% of pairs. Reducing the integration noise tenfold raises G3 only to $0.798/0.745$. The low ceiling is therefore not explained by a single noise realization.

Across the propagation-dependent screens, the full excitability state improves AUROC over the surviving relay alone by only 0.047–0.065, approximately one SESOI. A stronger benchmark for rollout value would therefore require an outcome whose propagation-dependent determinants are recoverable from the observation window; we did not obtain such a regime within this generator family.

D Readout and robustness analyses

Probe definition. For each canonical EWM ODE member, the encoder, dynamics core, and decoder are frozen, and only the probe head is trained. Each test and training patient is rolled out under each resection exactly as in the main analysis, with the unintervened graph in the encoder and A_R in the vector field. The per-node latent trajectory $z_n(t) \in \mathbb{R}^8$ over the held-out steps $[96 : 384]$ is summarized by its temporal mean and standard deviation (16 features per node); these are pooled over the surviving nodes by mean and max (removed nodes excluded, the same keep-mask convention as Static and GIN), giving a 32-dimensional input to the head. For each member, pooled features are standardized using feature-wise training-split means and standard deviations (standard deviation $+10^{-8}$); the same transform is applied to validation and test features. Because the two resections leave identical latents on the common surviving nodes (Appendix C), the probe, like every other scorer on this benchmark, separates them only through the surviving relay's contribution to the pool. The head outputs one logit, oriented higher-is-more-seizure, and is trained with binary cross-entropy on the training-split intervention labels with the main experiments' optimizer settings (Adam, learning rate 3×10^{-3}, batch 8, 100 epochs, gradient-norm clip 1.0, final epoch,

Table 13: **Simulation-only gate for a propagation-dependent setting** (200 patients per configuration; 12 nodes). For the network-wide outcome, passing requires G1 ≤ 0.85, G2 > 0.5, G3 ≥ 0.9, positive rate in $[0.40, 0.60]$, and discordance ≥ 0.3; the peripheral-outcome screens use G1 ≤ 0.85, G3 ≥ 0.85, positive rate in $[0.30, 0.70]$, and discordance ≥ 0.25, with G2 satisfied by definition. The baseline row is an instrument check. The ten screen candidates comprise six in the first screen, two in the second, and two in the third; multiple k values in the third screen are repeated evaluations of the same two configurations. The third screen evaluated $k \in \{1, 5, 9\}$; the table shows $k = 1$ and $k = 9$. Bold indicates passed gates; no candidate passes all required gates.

Configuration (c, W_{RP})	G1 relay-x_0 oracle	G2 peripheral seizure	G3 full-state ceiling	Positive rate	Discordant fraction
baseline $(-2.40, 0)$	0.999	0.000	**0.999**	0.618	**0.495**
$-2.30, 0.6$	0.999	0.009	**0.956**	**0.578**	**0.525**
$-2.30, 1.0$	0.997	0.198	**0.980**	**0.555**	**0.530**
$-2.20, 0.6$	0.923	**0.511**	0.867	**0.583**	**0.455**
$-2.20, 1.0$	0.915	**0.590**	0.844	**0.542**	**0.455**
$-2.13, 0.6$	**0.569**	**0.884**	0.782	**0.432**	0.135
$-2.13, 1.0$	**0.568**	**0.886**	0.786	0.350	0.150
Peripheral-seizure outcome (second screen; G2 satisfied by definition)					
$-2.20, 1.0$	**0.712**	N/A	0.764	**0.320**	**0.390**
$-2.20, 0.6$	**0.649**	N/A	0.714	0.297	**0.285**
Modal peripheral-seizure outcome over k realizations (third screen)					
$-2.20, 1.0, k{=}1$	**0.730**	N/A	0.779	**0.330**	**0.390**
$-2.20, 1.0, k{=}9$	**0.728**	N/A	0.776	**0.323**	**0.365**
$-2.20, 0.6, k{=}1$	**0.658**	N/A	0.723	**0.308**	**0.275**
$-2.20, 0.6, k{=}9$	**0.664**	N/A	0.726	**0.312**	0.245

no weight decay unless stated). One head is trained per member; member logits are averaged per patient–intervention pair before AUROC, and the paired bootstrap follows the confirmatory rules.

Head selection. Four heads were compared on the validation split only: linear (33 parameters), MLP $32 \to 16 \to 1$ (545), the same MLP with weight decay 10^{-4}, and MLP $32 \to 32 \to 1$. The head with the highest validation AUROC was selected separately in each setting before test evaluation: the 545-parameter MLP at 12 nodes (validation 0.824, test 0.761) and the linear head at 16 nodes (validation 0.805, test 0.804). Parameters at inference, head plus frozen EWM, are 5,503 and 4,991, against 17,153 for Static. Paired contrasts: versus Static $+0.018$ $[-0.020, +0.055]$ and $+0.103$ $[+0.058, +0.149]$; versus ordinary EWM $+0.242$ $[+0.167, +0.317]$ and $+0.295$ $[+0.227, +0.365]$. The probe differs from the confirmatory readout in three respects at once, latent rather than decoded input, mean and max rather than max pooling, and learned rather than fixed weights; the ablation below separates them.

Readout ablation. Five readouts were evaluated on the same frozen canonical EWM members and rollouts; learned variants use linear heads with no validation-based model selection. Inputs are either the decoded signal $\hat{X}_n(t)$ over steps $96:384$, summarized per node by mean, standard deviation, and mean line length (3 features), or the latent $z_n(t)$ summarized by temporal mean and standard deviation (16 features). Readouts are either fixed, the maximum over surviving nodes of one scalar per node (mean line length for the decoded signal, which is the confirmatory readout; the coordinate-averaged temporal standard deviation for the latent), or a learned linear head on mean $\oplus$ max pooling over surviving nodes, trained with the probe's settings. Table 5 gives the results. The decoded-signal fixed readout reproduces ordinary EWM exactly (0.519/0.509; paired Δ 0.000 in every resample). Replacing the fixed reduction by a learned linear head on the same decoded statistics gains $+0.218$ and $+0.285$; replacing decoded by latent input under a learned head gains $+0.038$ and $+0.010$; dropping the mean branch of the pooling costs 0.023 and 0.004; and a fixed maximum over nodes applied to latents reaches only 0.600/0.619, above ordinary EWM but below every learned head and below Static. The gain is therefore attributable almost entirely to replacing the fixed per-node scalar-and-maximum reduction by a learned readout, not to the representation or to the pooling. The linear-head latent cell reaches 0.775 at 12 nodes; the main text reports the validation-selected probe (0.761).

Graph-free MLP. The no-graph MLP receives the same six per-node summary statistics as Static and the keep mask, applies a per-node MLP and masked mean pooling, and never sees the adjacency; its parameter budget (17,153) matches Static. It is trained on intervention labels with the common

optimization settings, three members, and reaches 0.801 and 0.823 ($+0.058$ $[+0.013, +0.104]$ and $+0.122$ $[+0.075, +0.172]$ versus Static; $+0.000$ $[-0.019, +0.019]$ versus GIN at 16 nodes). It is a post-hoc control for interpreting the probe and provides a stronger raw-feature comparator than Static.

Stronger no-rollout floors (post-hoc). Static is the protocol floor, but GIN (0.810/0.823) and the graph-free MLP (0.801/0.823) are the strongest no-rollout predictors in the study. Table 14 recomputes every paired contrast against GIN, and against the supervised temporal predictor of Appendix E, with the confirmatory bootstrap (10,000 patient-clustered resamples, the same seeding). Against GIN, TGR-WM's 16-node advantage remains above the SESOI ($+0.089$ $[+0.052, +0.128]$), as does TGR-WM small's; TGR-WM at 12 nodes falls to class ‡, the encoder-swapped ODE model to $*$ at 12 nodes and $\circ$ at 16, and the probe below GIN in both settings. Every class ‡ against Static is ‡ against GIN as well, and the MLP column differs from GIN by at most 0.01 with identical classes.

Table 14: **Paired contrasts against stronger no-rollout predictors.** Δ AUROC versus GIN and the supervised temporal head, with 95% patient-clustered bootstrap CIs and decision classes as in Table 1. The graph-free MLP gives nearly identical contrasts to GIN. The final block uses predictors retrained on the post-transient window; for a head at AUROC 0.978, a $+0.05$ margin is infeasible.

	12-node		16-node	
Model	vs GIN	vs temporal pred.	vs GIN	vs temporal pred.
EWM ODE	-0.291 $[-0.366, -0.218]^{\ddagger}$	-0.389 $[-0.455, -0.323]^{\ddagger}$	-0.314 $[-0.380, -0.248]^{\ddagger}$	-0.429 $[-0.493, -0.366]^{\ddagger}$
GRU-core	-0.283 $[-0.353, -0.216]^{\ddagger}$	-0.381 $[-0.445, -0.317]^{\ddagger}$	-0.303 $[-0.375, -0.229]^{\ddagger}$	-0.419 $[-0.477, -0.361]^{\ddagger}$
TGR-WM	-0.017 $[-0.069, +0.036]^{\ddagger}$	-0.115 $[-0.155, -0.074]^{\ddagger}$	$+0.089$ $[+0.052, +0.128]^{\dagger}$	-0.027 $[-0.052, -0.002]^{\ddagger}$
x_0-supervised	-0.085 $[-0.128, -0.043]^{\ddagger}$	-0.183 $[-0.231, -0.135]^{\ddagger}$	-0.103 $[-0.152, -0.054]^{\ddagger}$	-0.218 $[-0.265, -0.173]^{\ddagger}$
Oracle-conditioned	$+0.138$ $[+0.105, +0.173]^{\dagger}$	$+0.041$ $[+0.014, +0.068]^{*}$	$+0.152$ $[+0.117, +0.189]^{\dagger}$	$+0.037$ $[+0.019, +0.056]^{*}$
RSSM	-0.123 $[-0.182, -0.066]^{\ddagger}$	-0.221 $[-0.274, -0.169]^{\ddagger}$	-0.131 $[-0.183, -0.081]^{\ddagger}$	-0.247 $[-0.299, -0.195]^{\ddagger}$
TWM	-0.209 $[-0.263, -0.156]^{\ddagger}$	-0.307 $[-0.364, -0.251]^{\ddagger}$	-0.080 $[-0.122, -0.039]^{\ddagger}$	-0.196 $[-0.241, -0.152]^{\ddagger}$
TGR-WM small	$+0.063$ $[+0.026, +0.102]^{*}$	-0.035 $[-0.070, +0.001]^{\ddagger}$	$+0.141$ $[+0.105, +0.180]^{\dagger}$	$+0.026$ $[+0.010, +0.044]^{*}$
EWM ODE wide	-0.129 $[-0.174, -0.082]^{\ddagger}$	-0.226 $[-0.274, -0.179]^{\ddagger}$	-0.368 $[-0.440, -0.294]^{\ddagger}$	-0.483 $[-0.544, -0.423]^{\ddagger}$
EWM ODE, temporal enc.	$+0.066$ $[+0.030, +0.102]^{*}$	-0.032 $[-0.063, -0.001]^{\ddagger}$	$+0.008$ $[-0.039, +0.055]^{\circ}$	-0.108 $[-0.145, -0.073]^{\ddagger}$
TGR-WM, static enc.	-0.322 $[-0.389, -0.254]^{\ddagger}$	-0.420 $[-0.484, -0.357]^{\ddagger}$	-0.134 $[-0.186, -0.081]^{\ddagger}$	-0.249 $[-0.301, -0.199]^{\ddagger}$
EWM probe	-0.049 $[-0.084, -0.013]^{\ddagger}$	-0.147 $[-0.190, -0.104]^{\ddagger}$	-0.019 $[-0.050, +0.011]^{\ddagger}$	-0.135 $[-0.174, -0.096]^{\ddagger}$
Post-transient window, 16 nodes (Appendix B)				
EWM ODE	N/A	N/A	-0.026 $[-0.046, -0.007]^{\ddagger}$	-0.074 $[-0.100, -0.049]^{\ddagger}$
TGR-WM	N/A	N/A	$+0.061$ $[+0.041, +0.084]^{*}$	$+0.013$ $[+0.005, +0.024]^{*}$

The no-rollout floor as a function of the label budget (post-hoc, exploratory). Every no-rollout predictor trains on the 980 training patients' intervention labels; no rollout model sees a label at all. To measure what that asymmetry is worth, each label-supervised predictor was retrained from scratch on nested subsets of $n \in \{25, 50, 100, 200, 490, 980\}$ training patients, drawn by one permutation taken independently of the labels, with cohorts, splits, labels, normalizer, architectures and optimization otherwise unchanged; the $n = 980$ column is the published result, not retrained. The rollout side is frozen throughout: per-patient scores are read from the confirmatory exports, which reproduce the published AUROCs to within 4.9×10^{-4}. A further control, *frozen encoder + head*, copies TGR-WM's forecast-trained encoder weights, freezes them, and trains only the 33-parameter node head, so that it differs from the supervised temporal head in nothing but whether the encoder saw labels.

Figure 7 and Table 15 give the result. At 16 nodes, the strongest label-supervised predictor falls from 0.939 at $n = 980$ to 0.813 at $n = 200$ and 0.622 at $n = 25$, while Static is already near 0.70 by $n = 200$. The TGR-WM contrast against the temporal head changes from -0.027 $[-0.052, -0.002]$ at $n = 980$ to $+0.340$ $[+0.280, +0.400]$ at $n = 25$. EWM ODE and GRU-core remain below every label-supervised arm at every evaluated budget in both settings. The frozen-encoder control remains below the end-to-end supervised head at every budget in both settings.

These exploratory contrasts are uncorrected for multiplicity.

Member aggregation and member-level uncertainty (post-hoc). The protocol averages the three members' raw scores before computing AUROC. Two checks assess the effects of the aggregation rule and the member set. First, the aggregation rule: Table 16 recomputes every model

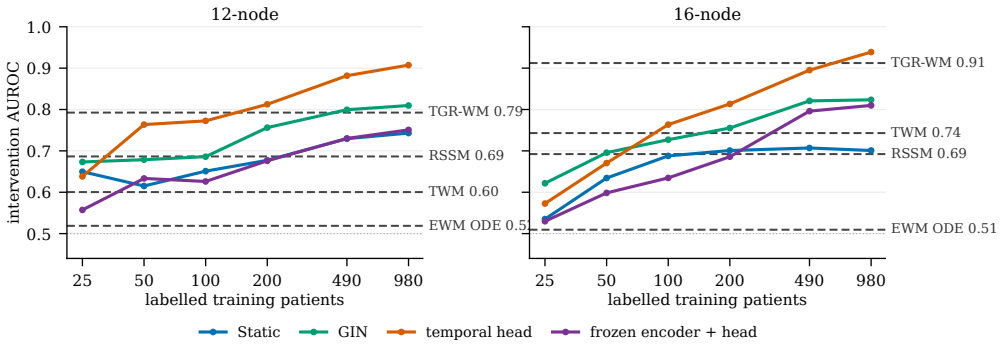


Figure 7: **The no-rollout floor as a function of the label budget.** Curves: label-supervised predictors retrained on n labelled training patients (nested subsets; $n = 980$ is the published column). Dashed: label-free rollout models, whose scores are frozen and do not depend on n. Dotted, chance. The no-graph MLP tracks GIN to within 0.01 at every budget and is omitted; the remaining rollout models are in Table 15.

Table 15: **Test AUROC by label budget** (three members per arm). Upper block: label-supervised predictors, retrained at each budget. Lower block: label-free rollout models, frozen, independent of n. *Frozen enc. + head* is TGR-WM's forecast-trained encoder, frozen, with a 33-parameter head.

	12-node						16-node					
Arm / labelled patients	25	50	100	200	490	980	25	50	100	200	490	980
Static	0.650	0.615	0.651	0.677	0.729	0.743	0.535	0.634	0.688	0.701	0.707	0.701
GIN	0.673	0.679	0.686	0.756	0.799	0.810	0.622	0.696	0.727	0.755	0.821	0.823
No-graph MLP	0.653	0.647	0.684	0.737	0.772	0.801	0.590	0.666	0.744	0.774	0.799	0.823
Temporal head	0.638	0.764	0.773	0.813	0.882	0.907	0.573	0.671	0.763	0.813	0.895	0.939
Frozen enc. + head	0.557	0.633	0.626	0.676	0.730	0.751	0.530	0.598	0.635	0.686	0.796	0.810
TGR-WM small				0.873						0.965		
TGR-WM				0.793						0.912		
EWM ODE + temporal enc.				0.875						0.831		
TWM				0.600						0.743		
RSSM				0.687						0.692		
TGR-WM + static enc.				0.488						0.689		
EWM ODE wide				0.681						0.455		
GRU-core				0.527						0.520		
EWM ODE				0.519						0.509		

under a mean of raw scores, a mean of ranks, and a mean of per-member standardized scores. The label-supervised predictors are insensitive to the choice, moving by at most 0.006, because their members output comparably scaled logits; the label-free rollout models are not, because a line-length score has no fixed scale across members. EWM ODE moves from 0.519 to 0.677 at 12 nodes under rank averaging and from 0.509 to 0.549 at 16, remaining below the floor in both; TGR-WM moves from 0.793 to 0.746 and from 0.912 to 0.811, remaining at or above it. Recomputing the paired intervals under the two alternative rules shows that the primary EWM–Static decision class is also unchanged: EWM ODE against Static is ‡ under rank averaging (-0.066 $[-0.122, -0.011]$ and -0.154 $[-0.218, -0.091]$) and under per-member standardization (-0.118 $[-0.184, -0.052]$ and -0.192 $[-0.265, -0.121]$), as it is under the protocol rule; GRU-core, EWM ODE wide and TGR-WM with the static encoder are ‡ under every rule in both settings; TGR-WM small stays † under every rule. Two cells soften without changing side: EWM ODE with the temporal encoder falls from † to $*$ at 16 nodes under the two alternative rules, and TGR-WM is ○ against Static at 12 nodes under all three. Second, the member set: the confirmatory intervals hold it fixed, so they describe patient-level uncertainty only. Table 17 adds a member bootstrap, resampling the three members with replacement alongside the patients. The intervals widen substantially. At 16 nodes, EWM against Static softens from ‡ to ○, TGR-WM against Static from † to ○, and EWM ODE with the temporal encoder from † to $*$. TGR-WM remains class ‡ against the supervised temporal head in both settings; Table 17 reports the remaining changes. The confirmatory analysis conditions on the fixed member set.

Member depth at 16 nodes (declared extension). Seven additional members each of EWM ODE, Static, and TGR-WM were trained at 16 nodes under the same settings, giving ten per arm; the original three-member ensemble remains the confirmatory endpoint. All twenty rollout members clear the forecasting gate, the smallest lower bound being $+0.582$; the gate is undefined for Static, which performs no rollout. The ten-member ensembles reach 0.624 for EWM ODE, 0.722 for Static, and 0.909 for TGR-WM. EWM remains below Static: $\Delta = -0.098$ $[-0.158, -0.038]$ with patient resampling and $[-0.259, +0.040]$ when members are resampled as well, both class ‡ (Table 17). Eight of the ten members are positively oriented and two are not (ρ_{surv} from -0.32 to $+0.58$), and the ensemble rises with the net orientation, from 0.509 on the original three, whose ρ_{surv} average $+0.13$, to 0.624 on the ten, which average $+0.28$. The gate separates none of this. Across the ten members held-out skill against the context mean spans 0.651 to 0.686 while intervention AUROC spans 0.335 to 0.751, a range twelve times wider, and the two are uncorrelated (Spearman -0.01); ρ_{surv} orders the same ten members at Spearman 0.76. Architecture, data, optimizer and window are identical across these members, so the only source of the spread is the random seed.

Table 16: **Sensitivity to the aggregation rule** (test AUROC of the original three-member ensembles). Raw: the protocol's mean of member scores. Rank: mean of within-model ranks. Std: mean of per-member standardized scores. The label-supervised predictors are insensitive to the rule; the label-free rollout models are not.

	12-node			16-node		
Model	raw	rank	std	raw	rank	std
Static	0.743	0.743	0.743	0.701	0.703	0.701
GIN	0.810	0.811	0.810	0.823	0.829	0.827
No-graph MLP	0.801	0.800	0.802	0.823	0.823	0.823
Temporal head	0.907	0.906	0.908	0.939	0.934	0.937
EWM ODE	0.519	0.677	0.625	0.509	0.549	0.509
GRU-core	0.527	0.509	0.518	0.520	0.492	0.493
TGR-WM	0.793	0.746	0.770	0.912	0.811	0.859
EWM ODE, temporal enc.	0.875	0.896	0.883	0.831	0.806	0.781
TGR-WM, static enc.	0.488	0.492	0.503	0.689	0.632	0.611
EWM ODE wide	0.681	0.609	0.585	0.455	0.543	0.442
TGR-WM small	0.873	0.854	0.857	0.965	0.961	0.965
x_0-supervised	0.725	0.698	0.722	0.720	0.651	0.644
Oracle-conditioned	0.948	0.909	0.946	0.975	0.983	0.975

Table 17: **Adding members as a second random factor.** Paired Δ AUROC with the confirmatory patient-clustered bootstrap (member set fixed) and with members resampled with replacement alongside patients. With only three members, member resampling is a coarse sensitivity analysis. Indented rows repeat the two Static contrasts at 16 nodes with ten independently trained members per arm; the original three-member ensemble remains the confirmatory endpoint. TS: supervised temporal head.

	12-node		16-node	
Contrast	patients only	+ members	patients only	+ members
EWM ODE − Static	$[-0.305, -0.147]^{\ddagger}$	$[-0.368, +0.023]^{\ddagger}$	$[-0.268, -0.117]^{\ddagger}$	$[-0.407, +0.060]^{\circ}$
ten members per arm	N/A	N/A	$[-0.158, -0.038]^{\ddagger}$	$[-0.259, +0.040]^{\ddagger}$
TGR-WM − Static	$[-0.010, +0.111]^{\circ}$	$[-0.139, +0.112]^{\circ}$	$[+0.157, +0.266]^{\dagger}$	$[-0.119, +0.276]^{\circ}$
ten members per arm	N/A	N/A	$[+0.137, +0.236]^{\dagger}$	$[+0.108, +0.244]^{\dagger}$
EWM ODE, temporal enc. − Static	$[+0.084, +0.182]^{\dagger}$	$[+0.069, +0.188]^{\dagger}$	$[+0.073, +0.188]^{\dagger}$	$[+0.002, +0.205]^{*}$
TGR-WM − TS	$[-0.155, -0.074]^{\ddagger}$	$[-0.311, -0.060]^{\ddagger}$	$[-0.052, -0.002]^{\ddagger}$	$[-0.364, +0.027]^{\ddagger}$
TGR-WM small − TS	$[-0.070, +0.001]^{\ddagger}$	$[-0.264, +0.079]^{\circ}$	$[+0.010, +0.044]^{*}$	$[+0.002, +0.064]^{*}$

Orientation on held-out patients (post-hoc). The per-member orientations of Table 19 are read from the test split. To ask what a single bit of orientation is worth without using test labels, each original member was rescored on the validation patients under the confirmatory rollout and readout and given the sign $s_m = +1$ if its validation AUROC is at least 0.5 and -1 otherwise. Test scores are then standardized per member, multiplied by s_m, and averaged; the protocol's own raw-score mean is reported alongside and reproduces the published test AUROCs to within 10^{-9}. Table 18 gives the result. For EWM ODE the bit is worth $+0.185$ $[+0.098, +0.272]$ and $+0.200$ $[+0.124, +0.276]$, which places it at -0.039 $[-0.083, +0.007]$ and $+0.009$ $[-0.046, +0.063]$ against Static: level with the floor in both settings, above it in neither. GRU-core at 16 nodes behaves the same way ($+0.126$ $[+0.057, +0.193]$), and EWM ODE wide gains $+0.152$ $[+0.078, +0.229]$ without reaching the floor. Where members already agree in orientation, the correction changes little; for TGR-WM

it changes AUROC by -0.023 and -0.053. The orientation bit uses validation labels and therefore quantifies the cost of inconsistent label-free readout orientations across members.

Table 18: **One bit of orientation per member.** Test AUROC of the original three-member ensembles scored by the protocol and after giving each member the sign of its validation AUROC (scores standardized per member before averaging). Paired patient-clustered bootstrap, 95% CI, confirmatory seeding. Static 0.743/0.701.

	12-node			16-node		
Model	protocol	oriented	vs protocol	protocol	oriented	vs protocol
EWM ODE	0.519	0.704	$+0.185$ $[+0.098, +0.272]$	0.509	0.710	$+0.200$ $[+0.124, +0.276]$
GRU-core	0.527	0.530	$+0.003$ $[-0.089, +0.096]$	0.520	0.646	$+0.126$ $[+0.057, +0.193]$
EWM ODE wide	0.681	0.585	-0.096 $[-0.140, -0.054]$	0.455	0.608	$+0.152$ $[+0.078, +0.229]$
TGR-WM, static enc.	0.488	0.497	$+0.010$ $[-0.106, +0.122]$	0.689	0.614	-0.076 $[-0.138, -0.015]$
TGR-WM	0.793	0.770	-0.023 $[-0.045, -0.001]$	0.912	0.859	-0.053 $[-0.070, -0.038]$
EWM ODE, temporal enc.	0.875	0.883	$+0.007$ $[+0.004, +0.011]$	0.831	0.781	-0.050 $[-0.064, -0.037]$
x_0-supervised	0.725	0.722	-0.002 $[-0.017, +0.012]$	0.720	0.644	-0.076 $[-0.113, -0.040]$

Readout compatibility (post-hoc diagnostic). One inference pass over every frozen checkpoint exports, per test patient, resection, and member, the decoded per-node mean line length over steps $96:384$, the quantity the confirmatory readout maximizes. Table 20 reports ρ_{surv}, the Spearman correlation between the three-member-mean line length of the surviving relay and that relay's x_0^{true} over the 420 test pairs (95% CI, 3,000 patient-bootstrap resamples), and f, the fraction of member-level rows whose readout maximum falls on the surviving relay. EWM ODE's member-mean ρ_{surv} does not differ from zero in either setting, but this is not because the decoded line length carries no information: Table 19 shows that within each member it tracks the relay's excitability at $|\rho|$ 0.12–0.58, with an orientation that differs between members (-0.24, $+0.49$, $+0.50$; $+0.58$, $+0.12$, -0.32), so that averaging member scores cancels it. The same holds for GRU-core, EWM ODE wide, TGR-WM + static encoder, and x_0-supervised at 16 nodes; every TGR-WM, TGR-WM small, RSSM, TWM, and temporal-encoder member is positively oriented. Per-member intervention AUROC follows the orientation: EWM members reach 0.380, 0.687, 0.723 and 0.751, 0.498, 0.335, so no member exceeds Static at 12 nodes and one does at 16, and the mean of member AUROCs is 0.597 and 0.528, below the floor under either estimator. For the static-encoder models, the orientation of the decoded excitability signal varies across independently trained members, leaving those orientations unaligned under the label-free readout; temporal-encoder members are consistently positive in this analysis. GRU-core at 12 nodes fails differently: a weakly informative value but a maximum on a periphery node in 94.5% of rows, so both candidates receive the same score. f is not itself a success criterion (TGR-WM has $f = 0.37$ at 16 nodes and AUROC 0.912: when the surviving relay is quiescent the maximum should fall elsewhere); what matters is that the relay's line length rises with its excitability, and across the eleven rollout models and both settings ρ_{surv} orders intervention AUROC at Spearman 0.95 (0.94 without the Oracle cells). The encoder swap moves this coefficient with the encoder: EWM's ODE core reaches $+0.83$ and $+0.77$ once given the temporal encoder, and TGR-WM's cell falls to -0.24 and $+0.35$ once it is removed, its maximum landing on the surviving relay in only 2% of 12-node rows. Figure 8 shows one patient: the simulator's dangerous relay seizes (line length 0.239 against 0.059 for the safe one), TGR-WM reproduces seizure-like excursions for it (0.035 against 0.003), and EWM decodes a smooth oscillation for the safe relay and a single bump for the dangerous one, reversing the order (0.0019 against 0.0025). Both rollouts decode amplitudes one to two orders of magnitude below the simulator's; the endpoint is rank-based, and the figure shows which model preserves the order.

E Comparison models, encoder controls, and additional robustness

Design. Four post-confirmatory comparison models were evaluated with the same intervention semantics and evaluation rules: two capacity-matched variants and two additional world models. *TGR-WM small* reduces TGR-WM's widths until its trainable parameters (4,953) lie within 0.1% of EWM ODE's 4,958; *EWM ODE wide* increases EWM's MLP widths from 64 to 182, giving 12,864 parameters, within 0.05% of TGR-WM's 12,858. *RSSM* (Hafner et al., 2019) and *TWM* (Micheli

Table 19: **Every negatively oriented member has a static encoder.** For each of the three original members: signed Spearman ρ_{surv} between the member's decoded line length of the surviving relay and that relay's x_0^{true}, and the member's own test AUROC; *neg.* counts members with $\rho_{\mathrm{surv}} < 0$; *ens.* is the protocol's three-member-mean-score AUROC (Tables 3 and 21). All 13 negatively oriented member–setting cells lie in the upper block; none of the 30 temporal-encoder cells is negative. Orientation is not sufficient for intervention value: every RSSM and TWM member is positively oriented and both ensembles stay below the floor. Descriptive; no confidence intervals.

	12-node				16-node			
Model	member ρ_{surv}	neg.	member AUROC	ens.	member ρ_{surv}	neg.	member AUROC	ens.
Static encoder (EWM's six fixed per-node statistics)								
EWM ODE	$-0.24, +0.49, +0.50$	1/3	0.380, 0.687, 0.723	0.519	$+0.58, +0.12, -0.32$	1/3	0.751, 0.498, 0.335	0.509
GRU-core	$-0.08, -0.01, +0.36$	2/3	0.497, 0.500, 0.531	0.527	$-0.19, +0.19, -0.19$	2/3	0.398, 0.581, 0.507	0.520
x_0-supervised	$+0.46, +0.26, +0.36$	0/3	0.715, 0.602, 0.655	0.725	$+0.02, +0.53, -0.12$	1/3	0.565, 0.720, 0.560	0.720
EWM ODE wide	$-0.11, +0.44, +0.35$	1/3	0.468, 0.681, 0.586	0.681	$+0.30, -0.22, +0.02$	1/3	0.592, 0.399, 0.600	0.455
TGR-WM, static enc.	$-0.34, +0.04, -0.36$	2/3	0.529, 0.488, 0.474	0.488	$+0.39, -0.11, -0.43$	2/3	0.680, 0.526, 0.570	0.689
Temporal encoder								
TGR-WM	$+0.61, +0.41, +0.55$	0/3	0.795, 0.579, 0.714	0.793	$+0.45, +0.89, +0.04$	0/3	0.716, 0.937, 0.556	0.912
TGR-WM small	$+0.92, +0.14, +0.45$	0/3	0.962, 0.618, 0.738	0.873	$+0.85, +0.90, +0.78$	0/3	0.948, 0.939, 0.944	0.965
EWM ODE, temporal enc.	$+0.74, +0.73, +0.84$	0/3	0.831, 0.839, 0.879	0.875	$+0.57, +0.84, +0.60$	0/3	0.729, 0.869, 0.725	0.831
RSSM	$+0.22, +0.44, +0.41$	0/3	0.616, 0.721, 0.646	0.687	$+0.46, +0.33, +0.37$	0/3	0.709, 0.647, 0.673	0.692
TWM	$+0.54, +0.17, +0.21$	0/3	0.757, 0.535, 0.432	0.600	$+0.56, +0.24, +0.11$	0/3	0.766, 0.597, 0.554	0.743

Table 20: **Readout compatibility**: Spearman ρ_{surv} between the decoded line length of the surviving relay and its true excitability (95% patient-bootstrap CI; test patients, three-member mean), the fraction of member-level rows in which the readout maximum falls on the surviving relay (f), and the intervention AUROCs of Tables 3 and 21.

	12-node			16-node		
Model	ρ_{surv} [95% CI]	f	AUROC	ρ_{surv} [95% CI]	f	AUROC
EWM ODE	$+0.050$ $[-0.053, +0.148]$	0.698	0.519	$+0.033$ $[-0.077, +0.137]$	0.493	0.509
GRU-core	$+0.277$ $[+0.187, +0.366]$	0.055	0.527	$+0.043$ $[-0.058, +0.143]$	0.514	0.520
TGR-WM	$+0.626$ $[+0.563, +0.679]$	0.493	0.793	$+0.844$ $[+0.816, +0.866]$	0.365	0.912
x_0-supervised	$+0.487$ $[+0.415, +0.552]$	0.515	0.725	$+0.429$ $[+0.350, +0.499]$	0.255	0.720
Oracle-conditioned	$+0.902$ $[+0.880, +0.917]$	0.641	0.948	$+0.960$ $[+0.946, +0.968]$	0.734	0.975
RSSM	$+0.366$ $[+0.275, +0.449]$	0.625	0.687	$+0.383$ $[+0.296, +0.465]$	0.587	0.692
TWM	$+0.506$ $[+0.431, +0.573]$	0.618	0.600	$+0.505$ $[+0.432, +0.575]$	0.989	0.743
TGR-WM small	$+0.766$ $[+0.726, +0.798]$	0.625	0.873	$+0.902$ $[+0.884, +0.916]$	0.765	0.965
EWM ODE wide	$+0.486$ $[+0.416, +0.552]$	0.419	0.681	-0.070 $[-0.173, +0.027]$	0.405	0.455
EWM ODE, temporal enc.	$+0.829$ $[+0.789, +0.860]$	0.628	0.875	$+0.774$ $[+0.723, +0.815]$	0.444	0.831
TGR-WM, static enc.	-0.240 $[-0.331, -0.144]$	0.021	0.488	$+0.352$ $[+0.262, +0.435]$	0.096	0.689

et al., 2023; Robine et al., 2023) are comparison world models at parameter counts comparable to TGR-WM's (Table 2).

Encoder swap. EWM ODE and TGR-WM differ in encoder and rollout core. Two further models cross these components and otherwise retain their parent architecture: *EWM ODE + temporal encoder* keeps EWM's ODE core, decoder, readout, and loss but replaces the six fixed per-node statistics with TGR-WM's temporal GRU encoder (hidden width 32 over the 96-step context, with 64- and 32-dimensional graph projections and layer normalization) followed by EWM's heads for $z_n(0)$ and $\hat{x}_{0,n}$ (13,310 parameters); *TGR-WM + static encoder* keeps TGR-WM's rollout GRU cell, decoder, readout, and loss but replaces its temporal encoder with EWM's static encoder, whose $\hat{x}_0$ head supplies the cell's excitability input (4,506 parameters). The latter is shape-identical to GRU-core, module for module. Its third input is the encoder's $\hat{x}_0$, whereas GRU-core uses a constant channel and adds a latent-bound regularizer. Together, the swaps test whether intervention transfer follows the encoder or the rollout core. Paired contrasts against the parent architectures use the same patient-clustered bootstrap.

Results. All six additional rollout models pass the forecasting gate in both settings (Table 2). TGR-WM small has higher AUROC than TGR-WM in both settings and clears the SESOI versus Static in both; EWM ODE wide improves at 12 nodes but remains below the floor and falls below chance at 16; RSSM and TWM establish no advantage in either setting (Table 3). The encoder swap follows the encoder: EWM ODE + temporal encoder reaches 0.875 and 0.831 ($+0.357$ $[+0.287, +0.427]$ and $+0.322$ $[+0.252, +0.392]$ over ordinary EWM; $+0.083$ $[+0.041, +0.125]$ and -0.081 $[-0.110, -0.054]$ against TGR-WM), while TGR-WM + static encoder falls to 0.488

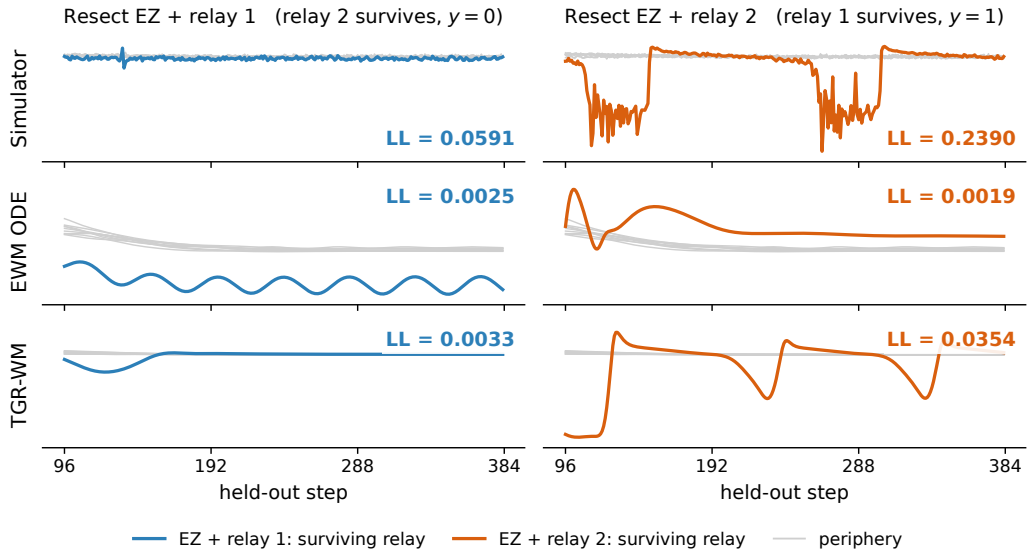


Figure 8: **What the readout sees**: one 12-node test patient (median $|\Delta x_0^{\text{true}}|$ among label-discordant patients; relay excitabilities -2.00 and -2.24), held-out steps under both resections. Rows: simulator truth, EWM ODE (member 0), TGR-WM (member 0); blue, the surviving relay under EZ+relay 1 ($y = 0$); orange, under EZ+relay 2 ($y = 1$), the colours of Figure 6; grey, periphery; removed nodes omitted. LL is the mean line length of the surviving relay, the quantity the readout maximizes; vertical scales are per row and omitted. The simulator's dangerous relay (right, $y = 1$) seizes; TGR-WM reproduces seizure-like excursions and orders the two resections correctly, EWM decodes a smooth oscillation for the safe relay and reverses the order. Both rollouts decode amplitudes far below the simulator's.

Table 21: **Capacity-matched variants and encoder swaps.** AUROC and paired Δ versus Static (95% CI; decision classes as in Table 1). Paired contrasts against the parent architectures are reported in the text.

	12-node		16-node	
Model	AUROC	Δ vs Static [95% CI]	AUROC	Δ vs Static [95% CI]
TGR-WM small	0.873	$+0.130$ $[+0.081, +0.178]^{\dagger}$	0.965	$+0.264$ $[+0.211, +0.317]^{\dagger}$
EWM ODE wide	0.681	-0.062 $[-0.112, -0.011]^{\ddagger}$	0.455	-0.246 $[-0.327, -0.163]^{\ddagger}$
EWM ODE, temporal enc.	0.875	$+0.132$ $[+0.084, +0.182]^{\dagger}$	0.831	$+0.130$ $[+0.073, +0.188]^{\dagger}$
TGR-WM, static enc.	0.488	-0.255 $[-0.329, -0.182]^{\ddagger}$	0.689	-0.012 $[-0.069, +0.047]^{\ddagger}$

and 0.689 (-0.305 $[-0.372, -0.237]$ and -0.223 $[-0.270, -0.177]$ against TGR-WM). Its 16-node interval against Static, $[-0.069, +0.047]$, excludes $+0.05$ by a narrow margin. These controls are descriptive and lie outside the confirmatory endpoint.

Forecasting under intervention (post-hoc). The gate scores forecasts of the intact system. To test the same rollouts on the intervened system, each model's rollout on A_R was scored against the intervened simulator trajectory that produced the label (decimated and normalized with the frozen scalar; steps $96 : 384$; surviving nodes), with context-mean and persistence baselines constructed from the pre-intervention context. All 22 model–setting cells have positive skill against both baselines; the context-mean lower confidence bound is above zero in every cell (smallest $+0.650$; Table 22). Intervened-system skill can exceed intact-system skill because the baselines are constructed from the pre-intervention context and therefore describe the post-resection system poorly. Despite the positive MSE skill, decoded line-length amplitudes are only 0.006–0.013 of the simulator values for all models except TWM (0.07), and median within-patient Spearman correlations between decoded and simulator per-node line lengths remain between -0.15 and $+0.20$. Thus the rollouts can forecast the intervened trajectories in MSE while poorly preserving the amplitude and spatial ordering of the seizure activity used by the intervention readout. An event-level gate on the intact held-out window is degenerate in this family: the epileptogenic node and both relays exceed $2.2\times$

the patient's median node line length in every test patient, no peripheral node does, and every model ranks those three nodes above the periphery (node-level AUROC 1.0 in all 22 cells).

Table 22: **Forecasting the intervened system.** Rollouts on A_R scored against the intervened simulator trajectory (steps $96:384$, surviving nodes): member-averaged MSE skill against the context-mean baseline (95% patient-bootstrap CI), skill against persistence, median within-patient Spearman ρ between decoded and simulator per-node line lengths, and median ratio of decoded to simulator line length. Intact-graph skill from Table 2 for comparison.

	12-node					16-node				
Model	intact	intervened skill [CI]	persist.	ρ_{LL}	ratio	intact	intervened skill [CI]	persist.	ρ_{LL}	ratio
EWM ODE	0.601	0.727 [0.699, 0.754]	0.377	$+0.18$	0.007	0.679	0.778 [0.757, 0.800]	0.439	$+0.11$	0.008
GRU-core	0.531	0.722 [0.694, 0.750]	0.356	-0.13	0.007	0.622	0.775 [0.754, 0.797]	0.431	$+0.13$	0.008
TGR-WM	0.718	0.713 [0.686, 0.739]	0.302	$+0.10$	0.008	0.749	0.769 [0.747, 0.792]	0.423	$+0.04$	0.008
x_0-supervised	0.586	0.716 [0.688, 0.744]	0.336	$+0.13$	0.008	0.680	0.766 [0.744, 0.789]	0.388	$+0.09$	0.009
Oracle-conditioned	0.694	0.757 [0.732, 0.782]	0.463	$+0.18$	0.007	0.747	0.805 [0.785, 0.825]	0.523	$+0.15$	0.008
RSSM	0.721	0.685 [0.656, 0.713]	0.119	$+0.16$	0.012	0.785	0.763 [0.741, 0.785]	0.358	$+0.13$	0.013
TWM	0.573	0.734 [0.706, 0.761]	0.414	$+0.13$	0.068	0.648	0.781 [0.760, 0.802]	0.455	$+0.20$	0.074
TGR-WM small	0.702	0.680 [0.650, 0.708]	0.081	$+0.13$	0.007	0.766	0.766 [0.744, 0.788]	0.326	$+0.16$	0.007
EWM ODE wide	0.603	0.722 [0.694, 0.749]	0.375	$+0.05$	0.007	0.671	0.778 [0.757, 0.799]	0.426	$+0.06$	0.008
EWM ODE, temporal enc.	0.718	0.749 [0.724, 0.774]	0.445	$+0.16$	0.008	0.788	0.777 [0.756, 0.799]	0.441	$+0.09$	0.007
TGR-WM, static enc.	0.578	0.735 [0.707, 0.762]	0.426	-0.15	0.007	0.642	0.778 [0.756, 0.799]	0.433	-0.11	0.006

Control configurations. Table 23 lists the architecture and loss of the capacity-matched variants, encoder swaps, and no-rollout controls.

Table 23: **Configurations of the control models.** Decoder for every rollout model: Linear $8 \to 64$, SiLU, Linear $64 \to 1$; readout: maximum over surviving nodes of mean line length on steps $96:384$. Forecast loss: $\mathrm{MSE}_{96:384} + 0.2\,\mathrm{MSE}_{0:96}$.

Model	Encoder	Core	Loss	Params
TGR-WM	per-node GRU (input 1, hidden 32) over the 96-step context; Linear $64 \to 64$, SiLU, Linear $64 \to 32$, residual, LayerNorm; heads $32 \to 8$ ($z_0 = 2\tanh$), $32 \to 1$ ($\hat{x}_0$)	edge MLP $16 \to 64 \to 8$ with degree-normalized weighted sum; GRUCell(17, 8) on $[z,\ \text{message},\ \hat{x}_0]$	forecast	12,858
TGR-WM small	as TGR-WM with widths reduced to EWM's parameter count	as TGR-WM, latent 8	forecast	4,953
EWM ODE, temporal enc.	TGR-WM's encoder and projections; heads $32 \to 8$, $32 \to 1$	EWM's graph neural-ODE, fixed-grid RK4	forecast + latent physics (0.1)	13,310
TGR-WM, static enc.	EWM's encoder: six statistics, Linear $6 \to 64$, SiLU, one hop, concat $[h; \bar{h}]$; heads $128 \to 8$, $128 \to 1$	TGR-WM's edge MLP and GRUCell(17, 8)	forecast	4,506
No-graph MLP	six statistics, Linear $6 \to 64$, SiLU; three residual Linear $64 \to 64$ with ReLU; no adjacency	none; masked mean pool, Linear $64 \to 64$, SiLU, Linear $64 \to 1$	BCE on labels	17,153
Temporal head	TGR-WM's encoder and projections; one degree-normalized hop on the post-resection adjacency	none; per-node Linear $32 \to 1$, masked mean pool	BCE on labels	9,697

Member jackknife (post-hoc). With three members per model, the patient bootstrap does not describe training-run uncertainty. Each key contrast was therefore recomputed for every pair of two-member sub-ensembles (one member dropped from each side; nine pairs per contrast) with the same patient-clustered bootstrap. At 12 nodes, EWM–Static remains ‡, EWM ODE with the temporal encoder remains † against both Static and ordinary EWM, TGR-WM with the static encoder remains ‡ against TGR-WM, and TGR-WM remains ‡ against the temporal predictor; TGR-WM–Static spans ‡, ○, and ∗, while TGR-WM small versus the temporal predictor spans ‡ and ∗. At 16 nodes, EWM–Static can soften from ‡ to ○, and the two † contrasts versus Static (TGR-WM and EWM ODE with the temporal encoder) can soften to ○; the negative contrasts against the temporal predictor and the downward encoder swap remain stable. No EWM member or sub-ensemble clears the SESOI over Static (Table 24).

Additional member set at 16 nodes. The jackknife leaves three 16-node classes dependent on the original member set, so a second member set was trained for that setting: three new members each of Static, EWM ODE, TGR-WM, EWM ODE + temporal encoder, and the supervised temporal head, with the same architectures and optimization as the original members. The two sets are analyzed separately. All three rollout models pass the forecasting gate in the additional set (skill against the context mean 0.676, 0.763, and 0.775, with lower bounds above zero). The three contested classes are unchanged: EWM ODE remains below Static (‡), while TGR-WM and the swapped ODE core remain above the SESOI (†), whether the additional models are compared with the additional or

Table 24: **Member jackknife.** Each contrast recomputed for the nine pairs of two-member sub-ensembles (one member dropped on each side), with the confirmatory patient-clustered bootstrap: original three-member Δ with class, the range of the nine sub-ensemble Δ estimates, and the classes they take. TS: the supervised temporal head.

Contrast	12-node original Δ	12-node jackknife range	12-node classes	16-node original Δ	16-node jackknife range	16-node classes
EWM ODE − Static	$-0.224^{\ddagger}$	$[-0.277, -0.007]$	‡	$-0.192^{\ddagger}$	$[-0.354, +0.010]$	‡, ∘
TGR-WM − Static	$+0.049^{\circ}$	$[-0.043, +0.060]$	‡, ∘, ∗	$+0.211^{\dagger}$	$[+0.016, +0.227]$	∘, †
EWM ODE, temporal enc. − Static	$+0.132^{\dagger}$	$[+0.118, +0.143]$	†	$+0.130^{\dagger}$	$[+0.049, +0.162]$	∘, †
EWM ODE, temporal enc. − EWM ODE	$+0.357^{\dagger}$	$[+0.127, +0.418]$	†	$+0.322^{\dagger}$	$[+0.046, +0.509]$	∘, †
TGR-WM, static enc. − TGR-WM	$-0.305^{\ddagger}$	$[-0.323, -0.193]$	‡	$-0.223^{\ddagger}$	$[-0.357, -0.033]$	‡
TGR-WM − TS	$-0.115^{\ddagger}$	$[-0.205, -0.097]$	‡	$-0.027^{\ddagger}$	$[-0.216, -0.003]$	‡
TGR-WM small − TS	$-0.035^{\ddagger}$	$[-0.168, +0.056]$	‡, ∗	$+0.026^{*}$	$[+0.023, +0.039]$	∗
EWM ODE, temporal enc. − TS	$-0.032^{\ddagger}$	$[-0.045, -0.014]$	‡	$-0.108^{\ddagger}$	$[-0.183, -0.068]$	‡

original Static (Table 25). TGR-WM and the swapped ODE core again do not clear the SESOI over the supervised temporal head (-0.040 $[-0.071, -0.010]$ and -0.022 $[-0.049, +0.006]$). Member dispersion is similar: the within-model AUROC range is 0.33–0.37 wide for the three rollout models (0.386–0.719 for EWM ODE, 0.545–0.893 for TGR-WM, and 0.545–0.914 for the swapped ODE core), compared with widths of 0.016 and 0.028 for Static and the temporal head. Finer ordering differs across member sets: the swapped ODE core is 0.081 below TGR-WM in the original set and 0.019 above it in the additional set.

Table 25: **Additional member set at 16 nodes.** The original confirmatory member set and a second set trained under the same protocol. Each Δ compares a set against its own Static; against the original Static the additional-set contrasts are -0.135 $[-0.208, -0.061]$, $+0.193$ $[+0.138, +0.250]$ and $+0.212$ $[+0.157, +0.270]$, with the same classes. Classes as in Table 1.

Model	AUROC original	AUROC additional	Δ vs Static [95% CI] original	Δ vs Static [95% CI] additional
Static	0.701	0.717	reference	reference
Temporal head (no rollout)	0.939	0.934	$+0.238$ $[+0.185, +0.290]^{\dagger}$	$+0.217$ $[+0.170, +0.265]^{\dagger}$
EWM ODE	0.509	0.566	-0.192 $[-0.268, -0.117]^{\ddagger}$	-0.152 $[-0.222, -0.082]^{\ddagger}$
TGR-WM	0.912	0.894	$+0.211$ $[+0.157, +0.266]^{\dagger}$	$+0.177$ $[+0.126, +0.229]^{\dagger}$
EWM ODE, temporal enc.	0.831	0.913	$+0.130$ $[+0.073, +0.188]^{\dagger}$	$+0.196$ $[+0.145, +0.249]^{\dagger}$

Supervised temporal predictor (post-hoc). To separate what TGR-WM's encoder contributes from what its rollout contributes, a no-rollout predictor was built on the encoder alone: TGR-WM's temporal GRU encoder as reused in the swap, a per-node linear layer, masked mean pooling over surviving nodes under Static's intervention conditioning, and a logit; binary cross-entropy on training intervention labels, no decoder, no physics term, 9,697 parameters, three members per setting. The predictor reaches 0.907 and 0.939. TGR-WM minus this predictor is -0.115 $[-0.155, -0.074]$ and -0.027 $[-0.052, -0.002]$, respectively. Thus the supervised head exceeds TGR-WM in both settings, although the 16-node difference is smaller than the SESOI. No rollout model clears the SESOI against this predictor; TGR-WM small at 16 nodes is the only rollout model above it, at $+0.026$ $[+0.010, +0.044]$, while the encoder-swapped ODE model lies below it in both settings (-0.032 $[-0.063, -0.001]$ and -0.108 $[-0.145, -0.073]$). Table 14 lists every contrast.

Excitability recovery by encoder (post-hoc). On the pre-resection graph, each encoder's excitability output $\hat{x}_{0,n}$ was compared with $x_{0,n}^{\mathrm{true}}$ over the 420 relay nodes per setting. For rollout models the orientation of $\hat{x}_0$ varies across independently trained members, so per-member $|\rho|$ is reported. With the static encoder, EWM ODE reaches 0.50–0.59 and 0.17–0.60, x_0-supervised 0.50–0.58 and 0.56–0.62, and TGR-WM + static encoder 0.55–0.59 and 0.56–0.60. EWM ODE with the temporal encoder reaches 0.91–0.94 in both settings. The six-statistic encoder therefore recovers relay excitability only moderately ($|\rho| \leq 0.62$), whereas the temporal encoder recovers it much more strongly. This difference is consistent with the stronger intervention performance of the temporal-encoder swap and the supervised temporal head.

Zero-parameter context baselines (post-hoc). Three scores read the normalized 96-step context alone, with no training: the surviving relay's context line length, its context variance, and the max-

imum context line length over surviving nodes. They reach AUROC $0.619/0.665$, $0.620/0.661$, and $0.488/0.530$, all class ‡ against Static (the relay scores correlate with the relay's x_0^{true} at Spearman 0.31–0.35). These zero-parameter context scores remain below Static, indicating that the stronger no-rollout performance is not explained by a single raw context amplitude statistic.

F Extended related work

Predicting outcomes under interventions in healthcare. Healthcare machine learning estimates patient outcomes under time-varying treatment sequences from observational data, using balanced recurrent representations (Bica et al., 2020) and transformer estimators (Melnychuk et al., 2022). These estimators condition on interventions and address causal estimation under observational confounding rather than evaluating simulator-style rollout models. Our question is complementary: whether explicit rollout dynamics add intervention-discriminative value beyond a no-rollout predictor using the same observation and intervention.

Epilepsy localization and intervention modelling. Localization methods score tissue for seizure involvement using onset patterns, high-frequency oscillations, or dynamical markers such as neural fragility (Li et al., 2021); a localization score alone does not model intervention-conditioned outcomes, whereas a no-rollout discriminative predictor can condition directly on a candidate intervention, as our structural baselines do. The Virtual Brain, Epileptor, and Virtual Epileptic Patient lines simulate seizure dynamics on individualized connectomes and infer patient-specific excitability for virtual intervention (Sanz Leon et al., 2013; Jirsa et al., 2014; 2017; Proix et al., 2017); recent work amortizes this inversion and extends virtual brain twins to stimulation targeting (Hashemi et al., 2023; Wang et al., 2025). Measured single-pulse stimulation responses, where perturbation and response are both observed, provide a natural real-data setting for this protocol (Van Blooijs et al., 2023; van den Boom et al., 2025).

G Notation summary

Table 26: Key notation used throughout.

Symbol	Meaning
X_i^{pre}	pre-intervention observation window of patient i ($T_w = 96$ context steps)
A_i	structural connectivity graph (adjacency matrix)
a	candidate intervention
R	resected node set
A_R	adjacency with the rows and columns of R zeroed
$z_n(t)$	latent state of node n
$\hat{x}_{0,n}$	learned excitability-like rollout coordinate
$x_{0,n}^{\mathrm{true}}$	simulator excitability (privileged state)
$s_i(a)$	continuous intervention score; higher is more seizure-like
$\mathrm{LL}(\cdot)$	mean line length (mean adjacent absolute difference)
T	rollout length (384 output time points)
κ	Epileptor difference-coupling strength
Δ	paired AUROC difference versus Static
ρ_{surv}	Spearman correlation of surviving-relay decoded line length with x_0^{true}
f	fraction of member-level rows whose readout maximum falls on the surviving relay