When Are Joint Predictive Models Reusable?

Anonymous authors
Paper under double-blind review

Abstract

Joint predictive models are reused after training under new compositions, observations, and questions. Familiar-task accuracy alone does not establish which later answers remain available. We ask what training records must reveal and what a retained state must preserve for such reuse. For known finite operators, the future query and evidence specify the required readouts. In 32 fresh worlds, records of a predicted missing triple interaction reduce conditional-query excess Brier by 0.0206 relative to equal-budget irrelevant records. At the primary budget, direct fits of the sufficient seven-coordinate target law and the full 64-state law have nearly identical mean risks. In a separate dynamic intervention, removing past–current coupling preserves tested forward predictions but stops revision of the past; a five-coordinate state preserves revision for a fixed past event. Endpoint-learned local mechanisms also predict unseen trees and use their outputs to revise later predictions with fixed parameters. Thus familiar-task accuracy can conceal different reuse abilities, while the information sufficient for a future use can be smaller than a full world law.

1 Introduction

In world-model research, learned models are increasingly used with new conditioning roles, compositions, and feedback after training (Hafner et al., 2025; Alzayer et al., 2026; Zhu et al., 2025; Bakermans et al., 2025). Large language models face a related question: accessible component facts need not yield a correct multi-hop answer (Press et al., 2023; Yu et al., 2026). Consider two mechanisms learned from short chain endpoints. We may wire them into an unseen tree, observe its output, revise a belief about the shared source, and predict another tree. This rewiring is a structural out-of-distribution (OOD) test: the mechanisms and parameters stay fixed on a graph unseen during training. Its observations then change the source belief and the next prediction. How much of this sequence does accuracy on the original chains establish?

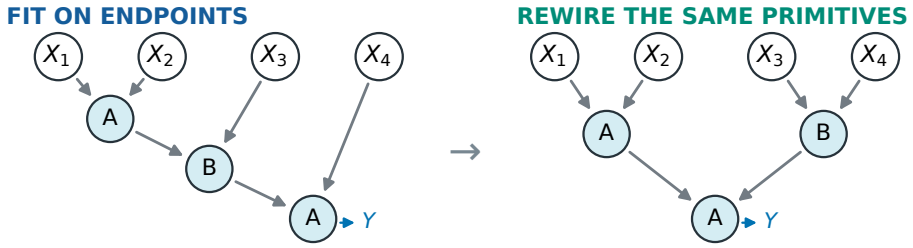


Figure 1: Training sees source coordinates X_i and endpoint Y, with internal outputs hidden. Fixed A, B are rewired into an unseen tree; its output revises the shared-source belief. The checks below organize the experiments.

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.

Across these settings, the common evaluation question is which relationships remain usable when composition, evidence, or the target query changes. We make it concrete for joint predictive models learned from observable records: shared stochastic mechanisms and a retained belief induce a coherent law for each declared experiment.

A tree endpoint changes roles: it is a prediction target before observation and evidence about the shared source afterward. With likelihood L, that observation changes a query f from $\mathbb{E}[f]$ to $\mathbb{E}[fL]/\mathbb{E}[L]$. Knowing the separate predictions need not determine their interaction. A coherent joint law connects these roles, but endpoint records may leave the needed interaction undetermined, or a retained belief may discard it before the next call. New wiring also tests whether the learned local behavior survives a different composition. For a declared future use, we ask which relationships the records reveal and which the retained state carries into the next update.

We derive query-relative information requirements and test three consequences: targeted records recover missing answers, compact beliefs preserve revision, and endpoint-learned local laws support new wiring and feedback.

2 When does a prediction remain useful?

We formalize this sequence by specifying the world, the records available at training time, and the queries asked after the context changes.

Let a world $M \in \mathcal{M}$ specify a source law and stochastic mechanisms. The supplied interface fixes variable identities, valid wiring, and which mechanisms are shared across experiments. A reuse context $c = (G, \iota, L, f)$ specifies an experiment graph, assigned inputs, a nonnegative observation likelihood, and a query $0 \le f \le 1$. Its answer is

$$q_M(c) = \frac{\mathbb{E}_{P_{M,G,\iota}}[fL]}{\mathbb{E}_{P_{M,G,\iota}}[L]}, \qquad 0 < \mathbb{E}_{P_{M,G,\iota}}[L] < \infty. \tag{1}$$

Training contexts $\mathcal{C}_{\text{train}}$ and the declared reuse family $\mathcal{C}_{\text{reuse}}$ can differ in graph, query, evidence, or horizon. Chain-to-tree transfer below changes the graph; a fresh sample of the original chain task does not. We report squared-probability or categorical Brier risk under each experiment's stated context weights, distinguishing average accuracy from exact agreement on a family.

Write $\mathsf{T}(M)$ for the population law of the observable training records. The new answers are identifiable from those records precisely when

$$\mathsf{T}(M_1) = \mathsf{T}(M_2) \implies q_{M_1}(c) = q_{M_2}(c) \quad \forall c \in \mathcal{C}_{\text{reuse}}, \quad M_1, M_2 \in \mathcal{M}. \tag{2}$$

If two worlds produce the same records but different new answers, collecting more records under that design cannot resolve the difference. Conversely, identification alone does not guarantee finite-sample or computational accuracy. The condition is relative to the future use: neither a full world joint nor a particular training objective is automatically required. Correct autoregressive conditionals at every supported prefix recover a finite sequence joint; selected one-step answers need not. Appendix I gives an exact sequence witness, while Section 4 specifies the missing directions for the experiments here.

3 A reusable object: mechanisms and belief

We represent the fitted object as (θ, b_t): learned local mechanisms θ and a belief b_t over uncertainty shared by experiments. The supplied interface names each mechanism's inputs and outputs and valid acyclic wirings. With source $H \sim \mu_{\theta,t}$, a new graph G induces the classical composed law

$$P_{\theta,t,G}(h, v \mid \iota) = \mu_{\theta,t}(h) \prod_{i \in G} K_{\theta,i}(v_i \mid v_{\text{pa}(i)}, h, \iota_i). \tag{3}$$

The factors have fresh noise conditional on their inputs and H; repeated use of a shared variable retains its identity. The graph can change after training, while θ stays fixed (Fritz, 2020; Kozen et al., 2023; Cusumano-Towner et al., 2019). Graph G records which named stochastic laws are called and how their ports connect; fitted models can implement those laws as tables or networks.

In Figure 1, $H = X = (X_1, \ldots, X_4)$: training observes those source coordinates and endpoint Y, while the outputs inside A, B are hidden.

An observation of experiment A has source likelihood $\ell_A(h)$. For a later query in experiment B, let $g_B(h)$ be its conditional answer given $H = h$. With independent local noise conditional on the shared source, one observation updates belief and changes the answer by

$$\mu_{t+1}(h) = \frac{\mu_t(h)\ell_A(h)}{\mathbb{E}_{\mu_t}[\ell_A]}, \qquad q_B(e) = \frac{\mathbb{E}_{\mu_t}[g_B\ell_A]}{\mathbb{E}_{\mu_t}[\ell_A]}, \qquad \Delta_B = \frac{\mathrm{Cov}_{\mu_t}(g_B, \ell_A)}{\mathbb{E}_{\mu_t}[\ell_A]}. \tag{4}$$

The denominator is positive for supported evidence. The product $g_B\ell_A$ links the query to what the observation reveals; separate forward answers do not generally determine it. Here g_B can read an earlier source event or predict a later experiment using the same H. In our chain-to-tree study, H is the source, $K_{\theta,i}$ are endpoint-learned primitives, and G is unseen tree wiring. The observation and retention studies then isolate relationships absent from records or discarded by the state.

For any composed law P, write the old answer, evidence mass, joint term, and revised answer as

$$a = \mathbb{E}_P[f], \quad z = \mathbb{E}_P[L], \quad m = \mathbb{E}_P[fL], \quad q = m/z, \quad \Delta = q - a. \tag{5}$$

The experiments intervene on the availability of m. Exact conditioning is standard; Appendix T gives the query-role and particle implementations. We ask when the training records identify the needed readouts and when the retained belief keeps them through later evidence.

4 Which relationships are needed?

Acquisition asks whether records reveal a required query–evidence relation; retention asks whether a belief summary keeps it available after another update. Known finite operators give exact answers to both questions.

Acquisition. For a finite source law p, known mechanisms turn a query and evidence into $q_c(p) = u_c^T p/(v_c^T p)$, where $u_c(h) = \mathbb{E}[fL \mid h, c]$ and $v_c(h) = \mathbb{E}[L \mid h, c]$. A known column-stochastic training channel reveals Op in the population. The following classical linear-estimability criterion makes Eq. 2 operational.

Proposition 1 (Query identification under a known finite channel)**.** *Let p range over all strictly positive laws on a finite source, let O be known and column-stochastic, and fix $0 \le u \le v$ pointwise with $v \neq 0$. The ratio $u^T p/(v^T p)$ is determined by Op for every such p exactly when $u = kv$ for a constant k, or both $u, v \in \mathrm{row}(O)$. If new evidence acts through a known likelihood $\ell \ge 0$, conditionally independent of the context given the source and history, and $v\ell \neq 0$, apply the same criterion to $u\ell, v\ell$.*

The proof and fixed-fiber variant are in Appendix B (Lenth, 2015; Escanciano, 2022; Basse & Bojinov, 2020). In the six-bit acquisition task, $h_i \in \{-1, +1\}$ pass through a known bit-flip channel with $\gamma = 0.8$. The tested conditional query and evidence likelihood are

$$f(h) = \frac{1 + \gamma h_2}{2}, \qquad L_e(h) = \frac{(1 + \gamma e_0 h_0)(1 + \gamma e_1 h_1)}{4}. \tag{6}$$

Across evidence values, L_e, fL_e span eight functions:

$$1,\ h_0,\ h_1,\ h_2,\ h_0h_1,\ h_0h_2,\ h_1h_2,\ h_0h_1h_2.$$

Pair marginals provide seven directions; the missing one is $\tau = \mathbb{E}[h_0h_1h_2]$. The target block's eight-state law, with seven free coordinates, suffices for the full evaluated query family without fitting the six-bit joint. With $\mu_I = \mathbb{E}[\prod_{i \in I} h_i]$ and $D_e = 1 + \gamma(e_0\mu_0 + e_1\mu_1) + \gamma^2 e_0 e_1 \mu_{01}$,

$$q_e(\tau') - q_e(\tau) = \frac{\gamma^3 e_0 e_1}{2D_e}(\tau' - \tau). \tag{7}$$

This predicts which extra observation changes the answer before fitting. Appendix E gives the full response formula.

Retention. Let T be a past quantity, S the current state, and future evidence act only on S. Suppose the past-query functions including the constant span $\mathcal{U}$ (r dimensions), and permitted evidence likelihoods including an empty history span $\mathcal{V}$ (d dimensions). All declared queries must remain answerable after every permitted evidence history.

Proposition 2 (Cross-time interactions for declared queries)**.** *Over all strictly positive beliefs on the finite Cartesian space (T, S), unnormalized query–evidence readouts span $\mathcal{U} \otimes \mathcal{V}$ (rd dimensions). Separate query and evidence expectations retain only $r+d-1$; the missing $(r-1)(d-1)$ directions are cross interactions. For exact linear summaries returning both conditional answers and evidence probabilities, $rd-1$ free coordinates are necessary and sufficient.*

In the three-state filtering task, $T = S_0$ and $S = S_t$ have three categories. When $\mathrm{rank}(KC) = 3$ for the transition K and sensor C, one further observation generates all nine past–current products. Thus revising the full initial-state posterior needs the cross interactions beyond the separate marginals. All 24 fitted action matrices in the twelve worlds satisfy this condition. For only the fixed event $S_0 = 0$, a recursively closed state keeps its three event–current masses and three current masses: five free coordinates after shared normalization. Section 7 tests that construction. These are universal positive-belief statements, not minimal memory claims about the experiment's reachable trajectories.

Let A_e be the known nonnegative operator for evidence e on unnormalized beliefs, with $U_e(b) = A_e b/(\mathbf{1}^T A_e b)$ for supported evidence. For this update to act on a stored summary S, equal summaries must still agree after that update: $S(b) = S(b')$ implies $S(U_e b) = S(U_e b')$ whenever both updates exist. A summary that decodes today's answers need not support this next step (Proposition 3 in Appendix B). For linear expectation summaries, the required query family $\mathcal{F}$ closes under the update operators:

$$\mathcal{W}_H = \mathrm{span}\{A_w^T f : f \in \mathcal{F} \cup \{1\},\ |w| \le H\}. \tag{8}$$

Here A_w is the chronological product over permitted nonzero-mass evidence histories. If population training records reveal Op and the retained linear state is $B^T p$, with the constant direction included in both spaces, then for all strictly positive initial beliefs, requiring both conditional answers and evidence probabilities gives two tests:

$$\begin{aligned} \mathcal{W}_H \subseteq \mathrm{row}(O) \quad &\Longleftrightarrow \quad \text{records determine the declared readouts,} \\ \mathcal{W}_H \subseteq \mathrm{col}(B) \quad &\Longleftrightarrow \quad \text{the state retains those readouts.} \end{aligned} \tag{9}$$

These are known-operator, linear-summary criteria; a single state updated recursively also needs the well-defined update condition above. Appendix C derives the containments and interaction count; Appendix B gives finite-sample bounds. The same future-use space therefore specifies two different interventions: acquire a missing direction through the records, or retain it in a state that survives the next update.

5 Can endpoint-learned mechanisms be rewired?

The positive reuse test fits two stochastic primitives A, B from chain endpoints, freezes them, builds unseen trees, and uses tree outputs to revise belief about a shared source and predict another tree. The source prior, primitive identities, ports, graph, and sharing semantics are supplied; intermediate states and tree labels are not observed during training. In each of eight fresh binary worlds, all fits see the same $8{,}192$ root/output training tuples and are selected by the same held-out mask score.

The local-kernel fits use either endpoint joint likelihood (K-J) or random-mask conditional likelihood (K-M). Graph-conditioned neural comparators GC-DAG, GC-T, and GC-M receive the same graph interface; each also constructs a coherent joint with the supplied source prior and can perform full Bayesian source-belief updates. The comparison therefore tests learned likelihoods under new wiring, with inference access on both sides. Figure 1 shows the topology change; Figure 2 reports prediction and feedback.

K-J's mean chain/tree excess Brier is $0.000354/0.000320$; GC-DAG's is $0.001992/0.021835$. The prespecified paired tree difference is 0.021515 (95% world-level interval $[0.006638, 0.036392]$); the paired tree-minus-chain degradation difference is 0.019876 ($[0.005408, 0.034345]$). K-M also transfers, with tree risk 0.000275. The ID difference has 90% interval $[0.000444, 0.002833]$ inside the prespecified ± 0.005 practical margin; the achieved ID risks are not equal. Thus the useful object is

a fitted local law that can be called under new wiring, not a particular likelihood objective. The supplied local factorization matches the teacher's mechanism class; representation, parameterization, and optimization still vary together across model families.

The feedback task first observes one source coordinate. Two unseen trees G_1, G_2 then provide successive observations about that same source X, with fresh local noise; the updated belief reanswers the unobserved source coordinates and predicts a third unseen tree G_3. Twelve schedules and four source rotations are averaged within each world. This tests whether the learned tree likelihood remains useful for another question, beyond its forward tree score.

After two unseen-tree observations, K-J gains 0.0311 in old-source Brier and 0.00684 on the next tree; K-M is similar. Product-projecting K-J's updated source belief loses 0.000809 and 0.001643 gain, respectively. GC-DAG's mean gains are negative on both readouts, with intervals spanning zero despite the same full update interface.

The local-row check finds mean teacher TV 0.02267 for K-J and 0.01877 for K-M. In four ternary worlds the tree difference has a descriptive interval crossing zero, and the worst graph/mask error reaches 0.185947; the binary means should not be read as uniform accuracy. In the separate exact-law study, five matched-shallow-risk seed pairs have lower shallow-joint KL for nonlinear fits (0.00045–0.00094 versus 0.02785–0.02800), yet their mean depth-four excess Brier is higher (0.00610 versus 0.000536).

The KL check was post-outcome, and the original four-rule primary match failed (Appendix Figure 14). The full feedback comparison is in Appendix Figure 9; checkpoint summaries are in Appendix E.3, and larger-wiring diagnostics are in Appendix D.

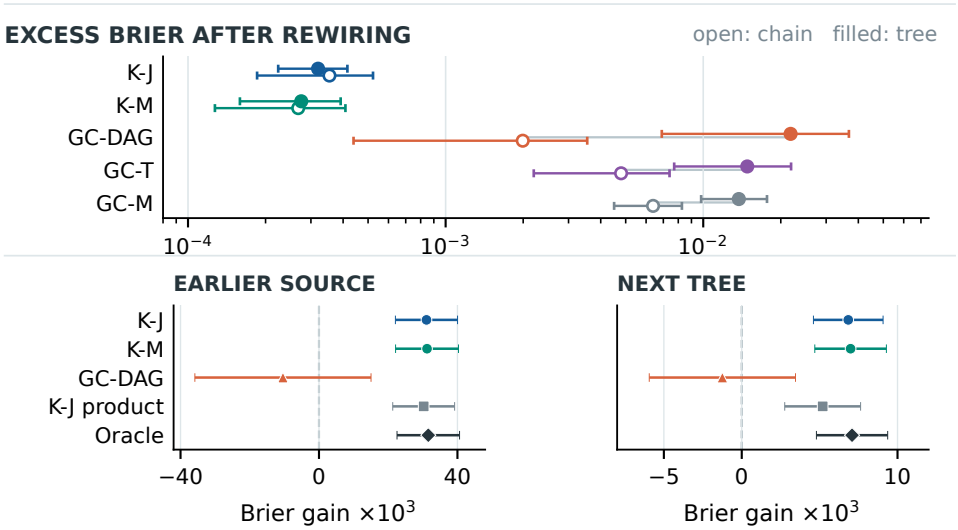


Figure 2: **Endpoint-learned primitives predict and revise under unseen wiring.** The topology is in Figure 1. Top: chain/tree excess Brier on 80 masks and 192 graphs per topology. Bottom: earlier-source and next-tree Brier gain (before minus after two tree observations); positive values mean improvement. Marks and 95% Student-t intervals summarize eight binary worlds. The product ablation projects K-J's source belief; the oracle uses true kernels. Full feedback arms are in Appendix Figure 9.

6 Can records acquire the predicted relation?

The six-bit acquisition experiment tests the missing triple direction from Eq. 7. After four development worlds and a frozen protocol, each of 32 fresh worlds starts with the same $4{,}608$ noisy pair records. At the primary budget, target-triple records (T7) observe (h_0, h_1, h_2), while equal-cost wrong-triple records (W7) observe (h_3, h_4, h_5). Both retain the same lower-order target moments;

only the target records measure the direction the formula selects. Figure 3 shows the resulting laws and risk.

Targeted records reduce mean query excess Brier from 0.022783 (W7) to 0.002182 (T7). The pre-specified paired difference is 0.020601, with 95% interval $[0.018357, 0.022845]$, positive in all 32 worlds. More pairs at the same scalar budget leave error at 0.022037. Direct maximum-likelihood fitting of the eight-state target law (QMLE8, seven free coordinates) reaches 0.000607, almost the same point estimate as a full 64-state fit (0.000606). The predicted information direction matters; direct likelihood estimation uses its records more efficiently than the simple one-moment repair. On the earlier exposed eight-world fitted laws, the seven-coordinate target summary reproduces all 36 evaluated answers to maximum difference 4.44×10^{-16}; an equal-size wrong-direction summary can differ by 0.195. This checks sufficiency within each fitted law, separately from the fresh acquisition test. Appendix E gives the generating laws, decoder, all budgets, and solver checks.

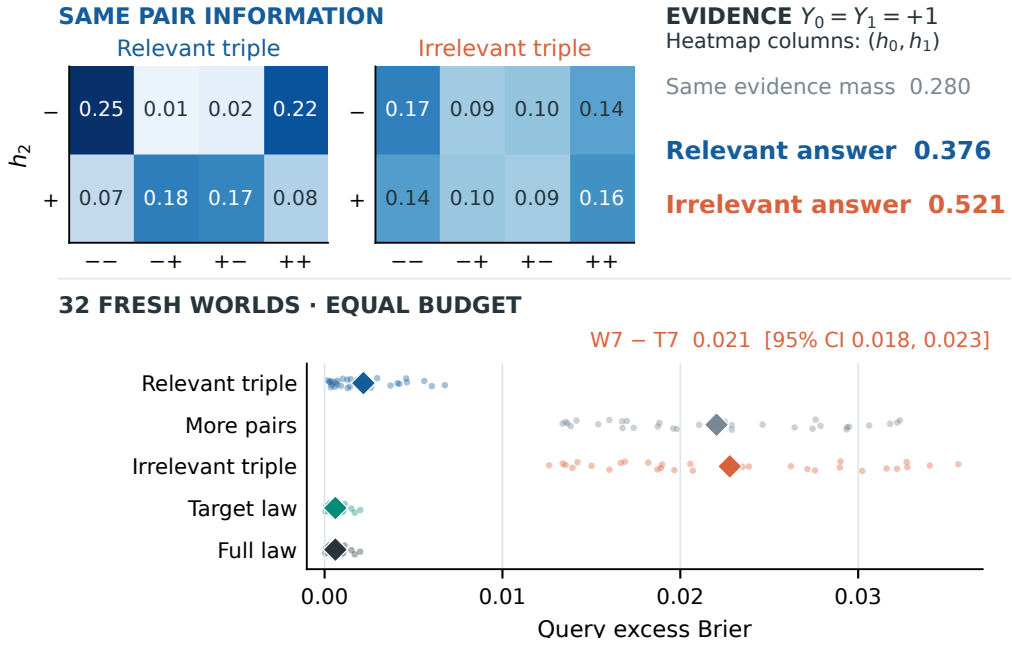


Figure 3: **The query selects an observation direction.** Top: target-block laws with equal pair moments and evidence mass give different answers. Bottom: teacher-weighted excess Brier in 32 fresh worlds at $m = 1{,}024$ triples (MorePairs: equal observed scalars). Dots are worlds, diamonds are means; the W7–T7 interval is paired, other arms descriptive.

In a separate eight-world temporal test, joint maximum-likelihood fitting on linked three-symbol episodes lowers query excess Brier to 0.000066, versus 0.001415 for equal-budget independent short episodes. Earlier objective controls for this task and a frozen-VAE conditioning check retain their own cohorts and metrics in Appendix F. The next test holds the fitted law fixed and changes what its belief retains.

7 Can the state keep that relation through evidence?

The three-state filtering experiment isolates retention under fixed fitted mechanisms. In twelve worlds, the full filter retains a joint belief over initial and current three-state variables. After every observation, its paired intervention replaces that joint by the product of its marginals. The two variants therefore agree on the current-state and tested future-AB predictions, yet differ in whether a later observation can revise an initial-state query. Figure 4 shows one actual belief and the twelve-world effect.

On 1,024 episodes per world, the saved current-state and future-AB readouts agree between full and projected filters to numerical precision at all tested horizons. At horizon eight, full versus product initial-state excess Brier is 0.001989 versus 0.023223; the paired increase is 0.021233 (95% interval $[0.016849, 0.025618]$), positive in all twelve worlds. After the first projection, subsequent observations no longer change the old state marginal. The full three-category past query still requires the relevant cross-time coupling. For only the fixed event $S_0 = 0$, the five-coordinate state from Section 4 recursively matches the full event answer to maximum discrepancy 1.22×10^{-15} on the retained episodes. A wrong-direction six-coordinate summary does not. The replay uses existing fits and episodes; the local and recursively projected interventions are distinguished in Appendix F.

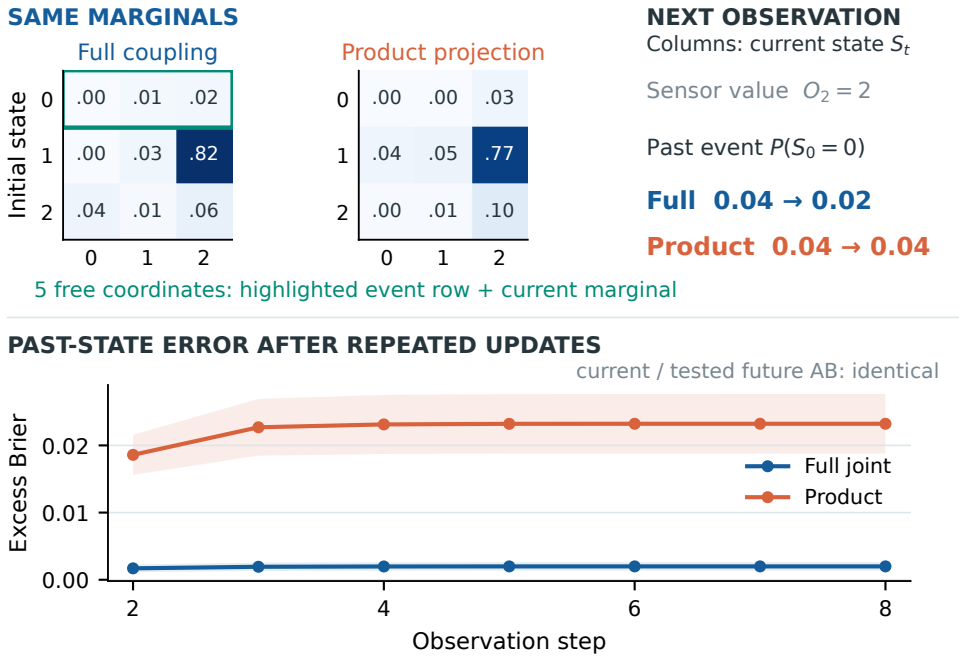


Figure 4: **Same marginals, different ability to revise the past.** Full and product share both marginals; only full coupling revises the past. The green event row and current marginal form the sufficient five-coordinate state. Bottom: twelve-world past-state excess Brier over updates, with descriptive 95% Student-t bands; product projection repeats after each update.

The sufficient state has a direct update rule. Write $r_i = P(S_0 = 0, S_t = i)$ and $m_i = P(S_t = i)$. Both rows pass through the same transition; on likelihood ℓ_i they become $r_i \ell_i / z$ and $m_i \ell_i / z$, where $z = \sum_i m_i \ell_i$. Thus the event answer $\sum_i r_i$ remains available at every step. The following replay checks the predicted direction of compression against the same full-joint answers; its metric is different from the teacher error in Figure 4.

Table 1: Fixed-event replay for the joint-likelihood fit at horizon eight across the exposed twelve worlds, six roles, and 1,024 episodes per world. The state selected by the declared event preserves the full-joint answer; more coordinates in the wrong direction do not.

Retained state for $S_0 = 0$	Mean squared difference to full joint
Correct five-coordinate event state	$< 3 \times 10^{-32}$
Wrong-event six-coordinate state	0.004552
Event–current product	0.006791

A supplied visual measurement. In a separate fresh pixel test, a frozen VAE receives sixteen source pixels. A supplied measurement reveals the *count*, but not the configuration, of four dis-

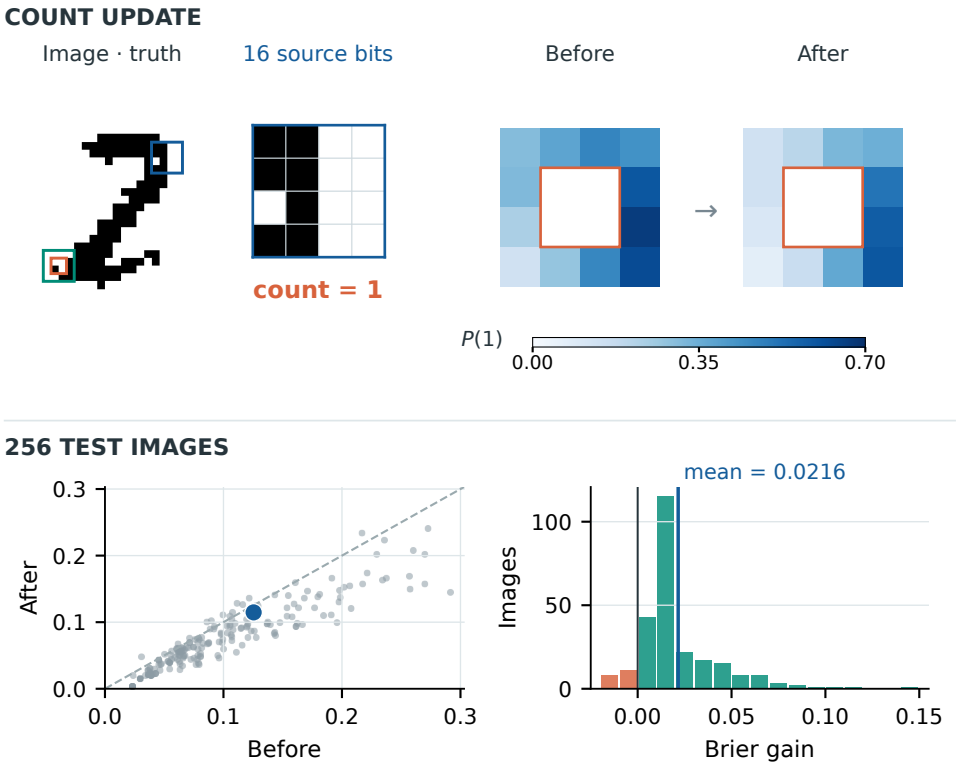


Figure 5: **A frozen generator revises pixel predictions after count evidence.** One label-selected image provides source bits and a hidden patch count of one; the full image is truth for display only. Blue, jade, and coral outlines mark source, twelve scored query pixels, and the counted patch. Before/after maps share a 0–0.7 probability scale. Below: observed-label Brier and its gain across 256 test images, averaged over twelve directions and eight streams within each image. Blue marks the displayed image in the scatter and the cohort mean in the histogram.

joint pixels; particle reweighting then revises twelve query-pixel probabilities. At 16,384 particles, observed-label Brier falls from 0.084778 to 0.063205 across 256 test images (25.4%; mean paired gain 0.0216, image-bootstrap 95% interval $[0.0191, 0.0243]$). Figure 5 shows the operation and all-image comparison; a 16×16 array of 256 complete conditional model draws is in Appendix Figure 8. Numerical reference is stable for only 1,732/3,072 image-direction cases (Appendix F).

8 Related Work

Neural Testbed shows why marginal prediction quality need not imply joint quality or downstream value (Osband et al., 2022). Belief and predictive-state methods ask which history summaries support future decisions (Kaelbling et al., 1998; Littman et al., 2001); invariant subspaces support value-directed compression (Poupart & Boutilier, 2002), and marginal filtering can lose dependencies needed after observation (Pfeffer, 2001; Boyen & Koller, 1998). We turn a declared query and evidence operation into specific acquisition and retention tests.

World models and compositional inference already use learned dynamics, variable conditioning roles, and fixed-parameter feedback (Hafner et al., 2019; 2025; Alzayer et al., 2026; Zhu et al., 2025; Bakermans et al., 2025). Our endpoint experiment measures which fitted local laws survive unseen wiring; the finite analysis specifies the relation that a later answer needs. Flexible VAE conditioning and posterior matching support the pixel instance (Kingma & Welling, 2014; Strauss & Oliva, 2022). The formal criteria use classical linear estimability (Basse & Bojinov, 2020; Lenth, 2015), and the

belief intervention follows the maximum-entropy hierarchy (Schneidman et al., 2003). Further links to composition, filtering, and conditional-query work are indexed in Appendix G.

9 DISCUSSION

Prediction quality measures answers to the questions used to test it; a new graph, query, or evidence history can call on other relationships. For known finite operators, the future query–evidence readouts give a concrete test: do training records determine them, and can the retained state return them after updates? The targeted-record experiment and the five-coordinate event state show that the right information can be acquired and kept without fitting or storing a full world joint. More records in the wrong direction and a state projection that preserves the tested forward answers do not meet those needs.

The endpoint-learning experiment asks whether fitted local laws also remain useful when their composition changes. In the tested binary worlds they predict unseen trees, and observations from those trees revise earlier-source and next-tree predictions with fixed parameters. This suggests evaluating world models against the composition, query, and evidence changes they are meant to support, alongside familiar-task accuracy. The graph interface and mechanism class are supplied in this experiment; the exact information criteria above assume known operators.

For a declared later use, this gives a concrete sequence: specify the graphs, evidence operations, and queries; identify the response functions they require when the operators are known; then test whether the records reveal those functions and whether the retained state can keep updating them. When local mechanisms must be learned, rewiring and feedback tests assess the fitted laws directly. A familiar-task score alone does not answer either question.

REPRODUCIBILITY STATEMENT

The anonymous supplementary materials provide saved-input and saved-law checks for the 32-world directed-acquisition study, saved-prediction checks for event retention, and world-aggregate replay for mechanism recomposition and feedback. The README specifies commands and coverage; retraining, earlier studies, and pixel-generator assets are excluded. Appendix U describes the retained evidence and validation limits.

AI USE STATEMENT

AI tools assisted hypothesis and protocol development, mathematical derivations and proof checks, implementation of synthetic-data generators and training/evaluation code, result interpretation, literature search, manuscript preparation, and artifact-consistency checks.

REFERENCES

Hadi Alzayer, Wenlong Huang, Haonan Chen, Christopher Luey, Lvmin Zhang, Maneesh Agrawala, Gordon Wetzstein, Fei-Fei Li, Yilun Du, Jiajun Wu, and Jia-Bin Huang. Masked visual actions for unified world modeling. *arXiv preprint arXiv:2607.19343v1*, 2026. URL https://arxiv.org/abs/2607.19343v1.

Jacob J. W. Bakermans, Pablo Tano, Reidar Riveland, Charles Findling, and Alexandre Pouget. Compositional meta-learning through probabilistic task inference. *arXiv preprint arXiv:2510.01858*, 2025. URL https://arxiv.org/abs/2510.01858.

Guillaume Basse and Iavor Bojinov. A general theory of identification. *arXiv preprint arXiv:2002.06041v1*, 2020. URL https://arxiv.org/abs/2002.06041v1.

Xavier Boyen and Daphne Koller. Tractable inference for complex stochastic processes. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, 1998. URL https://ai.stanford.edu/~koller/Papers/Boyen+Koller:UAI98.pdf.

Arwen Bradley. Local mechanisms of compositional generalization in conditional diffusion. *arXiv preprint arXiv:2509.16447v3*, 2026. URL https://arxiv.org/abs/2509.16447v3. Version 3, revised March 2026.

Arwen Bradley, Preetum Nakkiran, David Berthelot, James Thornton, and Joshua M. Susskind. Mechanisms of projective composition of diffusion models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 5319–5354, 2025. URL https://proceedings.mlr.press/v267/bradley25a.html.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.neurips.cc/paper_files/paper/2024/hash/2aee1c4159e48407d68fe16ae8e6e49e-Abstract-Conference.html.

Marco F. Cusumano-Towner, Feras A. Saad, Alexander K. Lew, and Vikash K. Mansinghka. Gen: A general-purpose probabilistic programming system with programmable inference. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, pp. 221–236, 2019. doi: 10.1145/3314221.3314642. URL https://www.cs.cmu.edu/~fsaad/assets/papers/2019-CusumanoTownerEtAl-PLDI.pdf.

Cassio Polpo de Campos and Fabio Gagliardi Cozman. Belief updating and learning in semi-qualitative probabilistic networks. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pp. 153–160, 2005. URL https://arxiv.org/abs/1207.1367.

Yilun Du, Conor Durkan, Robin Strudel, Joshua B. Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Sussman Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and MCMC. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 8489–8510, 2023. URL https://proceedings.mlr.press/v202/du23a.html.

Juan Carlos Escanciano. Semiparametric identification and fisher information. *Econometric Theory*, 38(2):301–338, 2022. doi: 10.1017/S0266466621000116. URL https://doi.org/10.1017/S0266466621000116. Published online in 2021.

Tobias Fritz. A synthetic approach to markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, 370:107239, 2020. doi: 10.1016/j.aim.2020.107239. URL https://arxiv.org/abs/1908.07021v8.

Christopher Grimm, André Barreto, Gregory Farquhar, David Silver, and Satinder Singh. Proper value equivalence. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://papers.nips.cc/paper/2021/file/400e5e6a7ce0c754f281525fae75a873-Paper.pdf.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/2de5d16682c3c35007e4e92982f1a2ba-Abstract.html.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2555–2565, 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1lOTC4tDS.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2. URL https://www.nature.com/articles/s41586-025-08744-2.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. In *Proceedings of the Conference on Learning Theory*, 2009. URL https://www.learningtheory.org/colt2009/papers/011.pdf.

Oleg Ivanov, Michael Figurnov, and Dmitry Vetrov. Variational autoencoder with arbitrary conditioning. In *International Conference on Learning Representations (ICLR 2019)*, 2019. URL https://arxiv.org/abs/1806.02382v3.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X. URL https://people.csail.mit.edu/lpk/papers/aij98-pomdp.pdf.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6114.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5637–5664. PMLR, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Klemen Kotar, Wanhee Lee, Rahul Venkatesh, Honglin Chen, Daniel Bear, Jared Watrous, Simon Kim, Khai Loong Aw, Lilian Naing Chen, Stefan Stojanov, Kevin Feigelis, Imran Thobani, Alex Durango, Khaled Jedoui, Atlas Kazemian, and Dan Yamins. World modeling with probabilistic structure integration. *arXiv preprint arXiv:2509.09737*, 2025. URL https://arxiv.org/abs/2509.09737.

Dexter Kozen, Alexandra Silva, and Erik Voogd. Joint distributions in probabilistic semantics. *Electronic Notes in Theoretical Informatics and Computer Science*, 3, 2023. doi: 10.46298/entics.12279. URL https://entics.episciences.org/12279. Proceedings of MFPS XXXIX.

Steven P. Lalley. Markov chains: Basic theory. University of Chicago course notes, n.d. URL https://www.stat.uchicago.edu/~lalley/Courses/312/MarkovChains.pdf.

Russell V. Lenth. Estimability tools for package developers. *The R Journal*, 7(1):195–199, 2015. doi: 10.32614/RJ-2015-016. URL https://journal.r-project.org/articles/RJ-2015-016/.

Yang Li, Shoaib Akbar, and Junier B. Oliva. ACFlow: Flow models for arbitrary conditional likelihoods. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 5831–5841, 2020. URL https://proceedings.mlr.press/v119/li20a.html.

Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. In *Advances in Neural Information Processing Systems*, volume 14, 2001. URL https://proceedings.neurips.cc/paper/2001/file/1e4d36177d71bbb3558e43af9577d70e-Paper.pdf.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision (ECCV 2022)*, 2022. URL https://arxiv.org/abs/2206.01714v6.

Sulin Liu, Peter J. Ramadge, and Ryan P. Adams. Generative marginalization models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 31773–31807, 2024. URL https://proceedings.mlr.press/v235/liu24az.html.

Utkarsh A. Mishra, David He, Yongxin Chen, and Danfei Xu. Compositional diffusion with guided search for long-horizon planning. In *International Conference on Learning Representations (ICLR 2026)*, 2026. URL https://arxiv.org/abs/2601.00126v3.

Utkarsh Aashu Mishra, Yongxin Chen, and Danfei Xu. Generative factor chaining: Coordinated manipulation with diffusion-based factor graph. In *Proceedings of The 8th Conference on Robot Learning*, volume 270, pp. 1461–1472, 2025. URL https://proceedings.mlr.press/v270/mishra25a.html.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Brendan O'Donoghue, Botao Hao, Dieterich Lawson, Morteza Ibrahimi, Xiuyuan Lu, Vikranth Dwaracherla, and Benjamin Van Roy. The neural testbed: Evaluating joint predictions. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://papers.neurips.cc/paper_files/paper/2022/file/5141f6bc105d30edbae48f1d2e0b1e66-Paper-Conference.pdf.

Paolo Perrone. *Categorical Probability and Stochastic Dominance in Metric Spaces*. PhD thesis, University of Leipzig, 2018. URL https://paoloperrone.org/phdthesis.pdf.

Avi Pfeffer. Sufficiency, separability and temporal probabilistic models. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 421–428, 2001. URL https://arxiv.org/abs/1301.2301.

Pascal Poupart and Craig Boutilier. Value-directed compression of POMDPs. In *Advances in Neural Information Processing Systems*, volume 15, 2002. URL https://proceedings.neurips.cc/paper_files/paper/2002/file/14ea0d5b0cf49525d1866cb1e95ada5d-Paper.pdf.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, 2023. doi: 10.18653/v1/2023.findings-emnlp.378. URL https://aclanthology.org/2023.findings-emnlp.378/.

Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. doi: 10.1109/5.18626. URL https://www.cs.cmu.edu/~durand/03-711/Readings/Rabiner89.pdf.

Eric Rossetto and Alessandro Antonucci. Partial identification under causal orders by linear programming. *arXiv preprint arXiv:2608.24427v1*, 2026. URL https://arxiv.org/abs/2608.24427v1.

Elad Schneidman, Susanne Still, II Berry, Michael J., and William Bialek. Network information and connected correlations. *Physical Review Letters*, 91:238701, 2003. doi: 10.1103/PhysRevLett.91.238701. URL https://www.princeton.edu/~wbialek/our_papers/schneidman+al_03b.pdf.

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. doi: 10.1038/s41586-020-03051-4. URL https://www.nature.com/articles/s41586-020-03051-4.

Anders Sjöberg, Jakob Lindqvist, Magnus Önnheim, Mats Jirstrand, and Lennart Svensson. MCMC correction of score-based diffusion models for model composition. *Entropy*, 28(3):351, 2026. doi: 10.3390/e28030351. URL https://doi.org/10.3390/e28030351.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 56242–56280, 2025. URL https://proceedings.mlr.press/v267/song25b.html.

Ryan R. Strauss and Junier B. Oliva. Posterior matching for arbitrary conditioning. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2201.12414v4.

Tao Sun, Utkarsh A. Mishra, Jiaxin Lu, Danfei Xu, and Iro Armeni. Energy-based compositional diffusion planning. *arXiv preprint arXiv:2606.21646v1*, 2026. URL https://arxiv.org/abs/2606.21646v1. Author-designated ICML 2026 paper; citing the inspected preprint.

Michael Thon and Herbert Jaeger. Links between multiplicity automata, observable operator models and predictive state representations—a unified learning framework. *Journal of Machine Learning Research*, 16:103–147, 2015. URL https://jmlr.org/papers/v16/thon15a.html.

Zhe Yu, Wenpeng Xing, Yunzhao Wei, Jie Chen, Hongzhi Wang, Xuyang Teng, and Meng Han. Composition collapse: Stable factual knowledge does not imply compositional reasoning. *arXiv preprint arXiv:2605.26789*, 2026. URL https://arxiv.org/abs/2605.26789.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. In *Proceedings of Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.015. URL https://www.roboticsproceedings.org/rss21/p015.html.

A Mathematical details

Affine characterization. Write $v = \sum_s v_s \delta_s$. If F preserves convex mixtures, then $F(v) = \sum_s v_s F(\delta_s) = vT$, where $T_{s,:} = F(\delta_s)$. Every row is a probability vector, so T is stochastic and unique. Conversely, matrix multiplication preserves every mixture and maps the simplex to itself. Continuity is unnecessary. This is the finite version of the classical correspondence in Perrone (2018).

Total-variation propagation. For a stochastic matrix T, $\|(p-q)T\|_1 \le \sum_s |p_s - q_s| \sum_t T_{st} = \|p-q\|_1$. If $\max_s \mathrm{TV}(T_j(s,\cdot), \hat{T}_j(s,\cdot)) \le \epsilon_j$, convexity gives $\mathrm{TV}(pT_j, p\hat{T}_j) \le \epsilon_j$. Adding and subtracting an intermediate transport yields $\mathrm{TV}(p_j, \hat{p}_j) \le \mathrm{TV}(p_{j-1}, \hat{p}_{j-1}) + \epsilon_j$. Equal initial states and induction give the stated sum bound, capped by one. A vertex attaining the largest row discrepancy also shows $\sup_p \mathrm{TV}(pT, p\hat{T}) = \max_s \mathrm{TV}(T(s,\cdot), \hat{T}(s,\cdot))$. Lalley (n.d., Proposition 5) gives a stronger strict contraction under positive entries; we need only the non-strict calculation above. Inputs are the same exogenous sequence in both recurrences. Adaptive, model-dependent input policies would need additional assumptions.

Unconditional transport is not an evidence update. For likelihood ratio two, the normalized update $B(p) = 2p/(1+p)$ fixes both vertices but maps $1/2$ to $2/3$. It does not preserve original mixture weights. The affine characterization concerns transport before conditioning and does not rule out necessary nonlinear Bayesian evidence updates.

Exact nonlinear witness. For Equation 12, $U([0,1]) = [1/4, 3/4]$, $4p(1-p) \le 1$, and $|(p - 1/4)(p-3/4)| \le 3/16$. Thus $N([0,1]) \subseteq [1/16, 15/16]$. Its perturbation vanishes at $0, 1, 1/4, 3/4$, giving identical first and second iterates from either deterministic initial state. The second iterates are $3/8$ and $5/8$; evaluating the perturbation at either gives $-45/1024$. The third iterates from zero are $7/16$ and $403/1024$; from one they are $9/16$ and $531/1024$. Squared scalar probability error is $2025/1048576$; full two-coordinate squared Euclidean error is twice that value. This is a valid simplex-valued predictor, not a fixed unconditional Markov map. The exact arithmetic and affine/TV arguments were checked independently of the experiment outcomes. This does not identify the trained network with the witness.

Conditional KL certificate. Let $\delta = \max_{a,s} D_{\mathrm{KL}}(T^*_a(s,\cdot) \| T_{\theta,a}(s,\cdot)) < \infty$ in nats. The trajectory chain rule gives $D_{\mathrm{KL}}(P(S_{0:L}) \| Q(S_{0:L})) \le L\delta$. Marginalizing to final outputs cannot increase KL, and

$$D_{\mathrm{KL}}(P(c, y \mid a) \| Q(c, y \mid a)) = D_{\mathrm{KL}}(P(c \mid a) \| Q(c \mid a)) + \mathbb{E}_{c \sim P} D_{\mathrm{KL}}(P(y \mid c, a) \| Q(y \mid c, a)).$$

The first term is nonnegative. Bernoulli Pinsker bounds squared probability error by half the conditional KL, giving $e_L \le L\delta/2$. Permutation averaging preserves this bound by convexity. Mean row divergence cannot replace the maximum without an occupancy argument. The earlier matched-objective diagnostics have larger maximum row KL but smaller measured OOD error after conditional training; the valid loose certificate does not predict that ranking.

Joint and conditional objectives. Factorization gives $\mathcal{L}_J = \mathcal{L}_C + \mathbb{E}_a H(P(c \mid a), Q_\theta(c \mid a))$. For a freely parameterized local row, rescaling each context pair jointly while preserving its ratio leaves conditional predictions unchanged. This local ambiguity need not survive recurrence sharing and the full depth-two constraints. The identity neither guarantees joint-training superiority nor makes joint loss necessary for a joint-state model.

B IDENTIFICATION OF FUTURE POSTERIOR QUERIES

This appendix imports the scoped finite-channel derivation used to design the information experiments. It specializes linear estimability (Lenth, 2015; Escanciano, 2022) and identification as constancy on observation fibers (Basse & Bojinov, 2020). The ratio statement below is our elementary derivation from those principles; it is not asserted to occur verbatim in those references.

One interaction, two readouts. For a fixed query and incremental likelihood on the common retained world, write $a = \mathbb{E}_P[f]$, $m = \mathbb{E}_P[fL]$, $z = \mathbb{E}_P[L] > 0$, $q = m/z$, and $\Delta = q - a$. If a is constant on the admissible family of laws, q is constant on that family if and only if Δ is constant. This proves the identification equivalence used in the main text. It does not equate identifying an exact revision with certifying its sign, and the equivalence need not hold if the earlier answer varies.

For estimates $\hat{a}, \hat{m}, \hat{z}$ with $\hat{z} > 0$, direct subtraction gives

$$\hat{q} - q = \frac{(\hat{m} - m) - q(\hat{z} - z)}{\hat{z}}, \qquad \widehat{\Delta} - \Delta = (\hat{q} - q) - (\hat{a} - a). \tag{10}$$

With the same exact earlier answer, the updated-answer error and revision error are identical, pointwise. More generally, their difference is precisely the earlier-answer error. The first identity retains cancellation between numerator and denominator errors: preserving a normalized readout need not preserve both unnormalized quantities separately. If $|\hat{m} - m| \le \epsilon_m$, $|\hat{z} - z| \le \epsilon_z < z$, and $|\hat{a} - a| \le \epsilon_a$, then $|\widehat{\Delta} - \Delta| \le (\epsilon_m + q\epsilon_z)/(z - \epsilon_z) + \epsilon_a$. These are elementary probability and ratio identities, not a new learning guarantee.

Universal identification (Proposition 1). Let O be column stochastic, p range over the interior of the probability simplex, and $0 \le u \le v$ with $v \neq 0$. The ratio $q(p) = u^T p/(v^T p)$ is a function of Op for every such p if and only if either $u = kv$ for a constant k, or both u and v belong to $\mathrm{row}(O)$. Membership is sufficient because numerator and denominator are then known linear functionals of Op; the constant case is immediate. For necessity, take $z \in \ker(O)$. Column stochasticity gives $\mathbf{1}^T z = 0$, so $p + tz$ stays in the simplex for sufficiently small positive and negative t. Constancy along this fiber implies

$$(u^T z)(v^T p) - (v^T z)(u^T p) = 0.$$

This linear identity holds on the open simplex, hence $(u^T z)v - (v^T z)u = 0$ as a vector. Unless u, v are proportional, both coefficients vanish for every $z \in \ker(O)$, giving the claimed row-space condition. At a particular interior fiber with value q_0, the weaker necessary and sufficient condition is $u - q_0 v \in \mathrm{row}(O)$. Boundary supports instead require feasible directions. Neither statement identifies hidden causal mechanisms.

Identification after feedback. Suppose a fresh evidence event has a known nonnegative likelihood ℓ_e on the retained source, and is conditionally independent of the integrated context randomness given that source and existing history. Then replace u, v by $u\ell_e, v\ell_e$ in the preceding argument. Provided $v\ell_e \neq 0$, the updated nonconstant ratio is universally identified from the earlier Op exactly when both products belong to $\mathrm{row}(O)$. A constant ratio is again an exception. This is a direct corollary, not a new identification theorem. The likelihood must represent genuinely new evidence conditional on the existing history. If new records also enlarge the training-observation experiment, its information contract must be enlarged accordingly; this result does not prohibit learning from new data.

Role compatibility and evidence bookkeeping. Appendix T derives the common-joint identities for query roles and evidence assimilation. Moving an evidence event from query context into history preserves the answer; it does not supply a second likelihood factor. Distinct events with identical recorded values are not duplicate events. These identities are necessary checks, not a finite certificate for every possible learned conditional.

Current answers versus future revisions. Even all queries over one experiment's observables, with shared uncertainty marginalized, may omit the coupling between experiments. For Bernoulli trials with a shared rate R, the priors $R = 1/2$ surely and $P(R = 0) = P(R = 1) = 1/2$ both predict a fair first trial. After a success they predict next-trial success probabilities $1/2$ and 1, respectively. The observation model and shared-rate interpretation are supplied. This classical Bayesian example separates one-trial prediction from revision; it is not an empirical result.

Proposition 3 (State sufficiency through successive updates)**.** *Suppose equal represented states give equal required current-query answers. For each permitted update, require*

$$S(\mu) = S(\mu') \quad \Longrightarrow \quad S(U_e\mu) = S(U_e\mu')$$

on the domain where the update is defined. When update availability is determined by the represented state, this condition is necessary and sufficient for the full belief updates to induce well-defined updates on S. Together with the readouts, these induced updates reproduce the required answers along finite admissible histories.

If $S(\mu)$ is a proposed sufficient state, equal states must determine equal required queries and, for each permitted update U_e, obey

$$S(\mu) = S(\mu') \quad \Longrightarrow \quad S(U_e\mu) = S(U_e\mu')$$

where both updates exist. Update availability must also be determined by the represented state. The implication makes $\overline{U}_e(S(\mu)) = S(U_e\mu)$ independent of the representative μ; conversely, any such induced update requires the implication. With well-defined updates and readouts, induction proves Proposition 3 and covers finite admissible histories. Below, invariant expectation coordinates give a sufficient construction; no small or efficiently learnable realization follows automatically.

Indistinguishable training laws. If $Op_+ = Op_-$ but their future-query probabilities differ by Δ, independent training records from this channel have the same law at every sample size. Any estimator based on those records has worst-source expected absolute error at least $\Delta/2$ and expected squared error at least $\Delta^2/4$. The estimator has the same output law in both worlds. For every output, the two absolute errors sum to at least Δ, and the two squared errors sum to at least $\Delta^2/2$; taking expectations proves the bounds. Restrictions on the source family can change this conclusion by supplying additional information. Larger models alone cannot distinguish identical observation laws.

Estimation after identification. Fix M required functions and, before sampling, representations $w = O^T\lambda_w + r_w$ with $\|\lambda_w\|_\infty \le B_w$ and $\|r_w\|_\infty \le \beta_w$. For N independent training records Z_j, estimate $\langle p, w\rangle$ by $N^{-1}\sum_j \lambda_w(Z_j)$. Hoeffding's inequality and a union bound give simultaneous errors at most

$$\epsilon_w = \beta_w + B_w\sqrt{2\log(2M/\delta)/N}$$

with probability at least $1-\delta$. If $\langle p, v\rangle \ge \zeta > \epsilon_v$, then the ratio estimate satisfies

$$|\hat{q} - q| \le \frac{\epsilon_u + \epsilon_v}{\zeta - \epsilon_v}.$$

Indeed, writing numerator and denominator as N_q, Z_q gives $|\hat{N}_q - q\hat{Z}_q| \le \epsilon_u + q\epsilon_v$ and $\hat{Z}_q \ge \zeta - \epsilon_v$. Clipping to $[0,1]$ cannot increase scalar error. This separates approximation, channel conditioning, sample size, and evidence sensitivity. Multiplying u, v and their bounds by the same positive constant leaves the ratio bound invariant. No bound is claimed for an unresolved denominator or an unbounded, adaptively enlarged query family. Independently accurate scalar estimates also need not form a compatible joint law.

Representing the required operations. For a kernel, $(K^*g)(x) = \sum_y K(y \mid x)g(y)$; evidence acts as $(M_Lg)(x) = L(x)g(x)$. The span of required numerator, denominator, and intermediate functions provides sufficient linear expectation coordinates when it is invariant under the permitted operations. A bounded horizon requires only the corresponding truncated closure. The dimension describes this class of linear representations, not a lower bound for arbitrary nonlinear encodings. Positivity, realizability, and numerical conditioning remain separate requirements. For the sequential instance, this is the predictive-state/observable-operator construction (Littman et al., 2001, Eq. (3),

Lemma 1) (Thon & Jaeger, 2015). To see the correspondence, let $A_{a,o}$ be the unnormalized belief update on an augmented finite state containing shared latent information. If the columns of B span the required function space, $\mathbf{1} = Bc_0$, and $A_{a,o}^T B = BC_{a,o}$, then $r = B^T b$ updates as

$$r' = \frac{C_{a,o}^T r}{c_0^T C_{a,o}^T r}.$$

The denominator must be positive. For $u = B\alpha$ and $v = B\beta$, the readout is $\alpha^T r / \beta^T r$. General expectation coordinates need not themselves be test probabilities or define a realizable state for arbitrary r. This correspondence explains the inherited construction; it does not identify a new graph's unknown mechanisms from observations of the old graph.

C TASK-RELATIVE DEPENDENCY SPACES AND CONCRETE BOUNDARIES

This appendix connects the information criterion to the exact requirements of E1 and RF-2. It uses linear estimability, invariant readout spaces, and tensor products. The construction inherits observable-state and belief-compression theory (Poupart & Boutilier, 2002; Littman et al., 2001; Thon & Jaeger, 2015). All necessity statements below quantify over strictly positive distributions on the declared finite state space. They do not establish minimum memory for arbitrary encodings or a single reachable trajectory.

Forward composition and backward evidence. Expanding the finite sums gives

$$((K_2 \circ K_1)^* g)(x) = \sum_z \sum_y K_1(y \mid x) K_2(z \mid y) g(z) = \sum_y K_1(y \mid x)(K_2^* g)(y).$$

For experiments conditionally independent given H, D_t, their common law factors as $\mu_t(h) K_A(y_A \mid h) K_B(y_B \mid h)$. Conditioning on $Y_A = e$ and summing Y_B gives Eq. 4. Subtracting $\mathbb{E}_{\mu_t} g_B$ yields the covariance expression. Dependence beyond H must instead be included in the common law. No conditional-independence conclusion is inferred from the existence of two separately trained predictors.

Linear readout sufficiency. Let A_e be known nonnegative, column-substochastic operators on unnormalized beliefs. For a word $w = e_1 \cdots e_k$, put $A_w = A_{e_k} \cdots A_{e_1}$. Its evidence mass is $v_w^T p$, $v_w = A_w^T \mathbf{1}$, and its query numerator is $u_{f,w}^T p$, $u_{f,w} = A_w^T f$. Allow all words over this fixed operator alphabet, excluding those with identically zero evidence mass. Protocol-restricted histories instead require history-dependent spaces or an augmented protocol state; unrestricted closure is then only a sufficient enlargement. For any linear observation Mp whose row space contains $\mathbf{1}$, a functional $g^T p$ is determined for every positive p precisely when $g \in \mathrm{row}(M)$. Sufficiency is linear reconstruction. Otherwise choose $z \in \ker M$ with $g^T z \neq 0$; $\mathbf{1}^T z = 0$, and small perturbations $p \pm \epsilon z$ stay positive and have the same observations but different readouts.

If query answers and evidence masses are both requested, they determine every $u_{f,w}^T p, v_w^T p$ and are determined by them. Include the known constant function in B, or equivalently augment its columns by $\mathbf{1}$ when testing containment. Applying the argument to all words of length at most H proves the two containments following Eq. 8. Taking a basis of $\mathcal{W}_H$ gives the minimum within linear expectation summaries: $\dim \mathcal{W}_H - 1$ independent coordinates. For an existing information space $\mathcal{V}_0$ containing the constant, the minimum number of additional linear directions is

$$\dim(\mathcal{V}_0 + \mathcal{W}_H) - \dim \mathcal{V}_0.$$

This number specifies information directions, not a sample count or computational cost.

Recursive realization and its scope. The spaces obey $\mathcal{W}_0 = \mathrm{span}(\mathcal{F} \cup \{1\})$ and $\mathcal{W}_{h+1} = \mathcal{W}_h + \sum_e A_e^T \mathcal{W}_h$. Finite state size makes this sequence stabilize at a common invariant space. If the columns of B form a basis there, write $A_e^T B = BC_e$ and $\mathbf{1} = Bc_0$. Then

$$r = B^T p, \qquad r' = \frac{C_e^T r}{c_0^T C_e^T r}.$$

The denominator is the incremental evidence probability and must be positive. To retain the probability of the entire observed history from the original prior, store an additional scalar Z and multiply it by this denominator at each step, or use unnormalized coordinates. A normalized state alone need not retain that mass. For bounded horizon, the appropriate spaces decrease with the remaining horizon; $\mathcal{W}_H$ need not be invariant. Coordinates can be signed, and arbitrary coordinate vectors need not correspond to valid probability laws.

Conditional answers alone. If evidence masses are not outputs, call (f, w) active when $u_{f,w}$ and v_w are not proportional. Proposition 1 implies that the minimum initial linear summary spans the constant and the numerator/denominator functions of active pairs only. The constant is included for span tests but is not counted as a free stored coordinate. Constant ratios need no such directions. For example, $f = (0, 0, 1)$ and $A = (1, 0, 0)^T(1/5, 1/2, 4/5)$ give an initial answer p_3 and answer zero after every nonempty repeated update. One free coordinate suffices for the answers, whereas including evidence probability requires two.

An initial summary that decodes every future answer given the word need not be recursively updateable in the same coordinates. For a four-state example, let $f = (0, 0, 1, 1)$, $D = \mathrm{diag}(1/4, 1/2, 3/4, 3/4)$, and $d = D\mathbf{1}$. Let the stochastic operator E send states $1, 2, 3$ to state 1 and state 4 to state 2. Pure D words depend only on $p_1, p_2, p_3 + p_4$; words containing E have answer zero. The initial active space is $\mathrm{span}\{1, f, d\}$. However, $d^T Ep = 1/4 + p_4/4$: the two priors $(1/4, 1/4, 1/4, 1/4)$ and $(1/4, 1/4, 1/8, 3/8)$ have equal initial coordinates but updated d moments $5/16$ and $11/32$. This distinguishes static decoding from recursive realization without asserting impossibility for a different nonlinear state.

Cross-time tensor space. Under Proposition 2, all declared query/history pairings are required. A future query numerator is the expectation of $g(T)v_w(S)$. Since the queried functions span $\mathcal{U}$ and the future likelihoods span $\mathcal{V}$, these products span $\mathcal{U} \otimes \mathcal{V}$. Products of bases are linearly independent on the Cartesian space, giving dimension rd. The spaces of separate expectations are $\mathcal{U} \otimes 1$ and $1 \otimes \mathcal{V}$, intersecting only in the constant. Their sum has dimension $r + d - 1$, leaving $(r - 1)(d - 1)$ cross directions. Linear estimability proves necessity and sufficiency of $rd - 1$ free coordinates when evidence masses are also outputs. Additional current queries outside $\mathcal{V}$, or operations that modify T, require their own enlarged space.

E1: exact missing direction. For binary variables, the products $\prod_{i \in S} h_i$ form a basis for functions of h. The known nonzero bit-flip signal preserves each observed block's span. Complete pair information on six variables spans the constant, six singleton functions, and fifteen pair functions, for rank 22. Expanding Eq. 6 over all four outcomes gives eight functions on the first three variables. Their intersection with the pair space has dimension seven. The missing direction is $h_0h_1h_2$; adding the target triple raises the observation rank to 23 and covers the query space. All 36 evaluated queries depend on those same three variables. Complete pairs and triples have rank 42, also covering this space without identifying the full 64-state source.

For pair-matching laws, the mixed-expectation difference is exactly

$$\delta m = \frac{\gamma^3}{8} e_0 e_1 \, \delta\mathbb{E}[h_0 h_1 h_2] = 0.064 e_0 e_1 \, \delta\mathbb{E}[h_0 h_1 h_2].$$

Because a and z agree, $\delta q = \delta\Delta = \delta m / z$. The saved projection preserves pairs only to its reported numerical tolerance, so the empirical comparison uses the full ratio decomposition in Eq. 10. A maximum-entropy pair model can itself have nonzero higher-order moments; projection does not simply set the triple moment to zero. These are population and fixed-law calculations. Targeted versus random triple performance at finite budget depends on how observations are allocated.

RF-2: exact additional directions. Let $i = S_0$, $x = S_t$, and $r_o(x) = \sum_{x'} K(x, x')C(x', o)$. A next-observation posterior of initial category j has numerator $\mathbf{1}\{i = j\}r_o(x)$ and denominator $r_o(x)$. If KC has rank three, the r_o span all current-state functions, so the nine numerator functions span the full 3×3 joint space. Every such posterior ratio is nonconstant; the ratio criterion therefore gives necessity even without separately requiring evidence probabilities. Current readouts span three directions, both marginals five, and these posterior readouts nine: respectively $2, 4, 8$ independent coordinates. The four cross directions are precisely $(3 - 1)(3 - 1)$.

More generally, a past-query space of dimension r and an m-state current variable with a full-rank response give $\mathcal{U}\otimes\mathbb{R}^m$ of dimension rm. Current-only operators preserve this space. For conditional-only necessity, the past-query space must contain a nonconstant function. A single nonconstant past event gives $r = 2$, hence dimension six in the three-state experiment. All 24 saved fitted action matrices over the 12 worlds pass the rank-three response check; the smallest response singular value across them exceeds 0.328. This verifies the operator premise, not reachability of every positive belief.

If $b(i,x) = u(i)v(x)$, prediction and a new observation give $b'(i,x') \propto u(i)(vK)(x')C(x',o)$. Summing over x' leaves the marginal u. Thus recursive product projection fixes the past-state answer after its first application. The current marginal follows the same filter as before projection. The experiment separately checks this identity on its saved current and future-AB readouts and measures the accumulated backward-query error.

D FINITE-DATA MECHANISM RECOMPOSITION: IMPLEMENTATION AND FULL RESULTS

This experiment complements the exact-label, sixteen-state Boolean study with finite endpoint records, heterogeneous mechanism worlds and graph-conditioned comparators. The full retained cohort has eight binary and four ternary worlds, with respectively five and three fits per world, totaling 52 fitted models. Training never receives the teacher's numerical kernel or a tree endpoint label.

Mechanism distribution and graph interface. For each binary primitive, draw a truth table uniformly from the ten Boolean functions that depend on both inputs. Independently for each input row, flip its preferred output with probability $\epsilon \sim \mathrm{Uniform}(0.05, 0.20)$. For a ternary primitive, draw $a, b \in \{1,2\}$ and $c \in \{0,1,2\}$; its preferred output is $(au + bv + c) \bmod 3$, with probability $1-\epsilon$, and each alternative has probability $\epsilon/2$. Fitted local tables are unrestricted within their categorical rows; the preferred-output rule is not supplied. The two local tables contain 16 binary or 54 ternary logits, including softmax redundancy.

Sources are independent uniform variables with a known prior. Mechanism calls have fresh independent local noise conditional on the shared source. A fan-out references the same intermediate random variable, so exact latent summation retains its dependence. Learning fixes a point estimate of the mechanisms; test-time updates concern source uncertainty, not a posterior over learned parameters.

For a three-letter primitive word (t_1, t_2, t_3) and root permutation π, the training chain has $h_1 = t_1(X_{\pi_1}, X_{\pi_2})$, $h_2 = t_2(h_1, X_{\pi_3})$, $Y = t_3(h_2, X_{\pi_4})$. The tree instead uses $h_2 = t_2(X_{\pi_3}, X_{\pi_4})$ and $Y = t_3(h_1, h_2)$. All eight words and all 24 root permutations give 192 graphs of each kind. The diamond uses $h_2 = t_2(h_1, X_{\pi_3})$, $Y = t_3(h_1, h_2)$; its matched chain uses $Y = t_3(h_2, X_{\pi_3})$. Both use three roots and three calls. The seven-call diagnostic repeats the four-root permutation twice among eight leaves, sampling 32 common word/permutation pairs. Chain calls follow execution order; the actual tree implementation assigns the word in bottom-up level order. Larger graphs thus also reuse roots, so this is a bounded size-and-sharing diagnostic.

Models and training actually executed. K-J maximizes endpoint conditional likelihood through exact latent elimination. K-M samples one of 80 masks uniformly for each tuple. GC-DAG has width 64, root/value embeddings, two primitive-specific two-layer MLPs with hidden width 128, residual parent additions, layer normalization and a sink softmax readout. GC-T/GC-M have four Transformer layers, four heads, width 64 and feed-forward width 128, with ordered-edge attention biases and no positional embeddings. GC-T predicts the sink given all roots; GC-M has a target token and masks unobserved values. GC-DAG has $66{,}690$ binary parameters, GC-T/GC-M $134{,}946$; these are different representation and optimization recipes, not parameter-matched networks.

Each training set has $4{,}096$ tuples from each of chain lengths two and three; the validation set has $1{,}024$ of each. In the executed optimizer, each step samples uniformly from the distinct graphs present, then samples 256 records from that graph with replacement. With all 96 two-call and 192 three-call graphs present, the effective graph-length frequencies are $1/3$ and $2/3$, although

the dataset contains equal counts. All arms use this graph sampling scheme. Adam runs at most $1{,}500$ steps, with gradient clipping at 5; validation every 100 steps selects the best checkpoint, with early stopping after five non-improving validations once 500 steps have elapsed. A 10^{-6} uniform probability mixture ensures positive support. Fits use float32; teacher references and probability summaries use float64.

Two binary development worlds and two learning-rate candidates per arm yield 20 development fits. The selected rates are 0.03 for K-J/K-M, 0.003 for GC-DAG and 0.001 for GC-T/GC-M. GC-DAG is the fixed primary comparator selected by mean development validation loss. The score equally weights all 80 masks on the first 512 validation records for every arm; GC-M uses its native mask head, the others their coherent joint. Fresh-world recipe choices are fixed before those worlds are generated. The population unit for intervals is the independently generated mechanism world, including its finite records and fitted model.

Recorded design/execution differences. The original plan specified uniform record sampling, the full validation split and six ternary development fits. Actual execution uses graph-grouped sampling, the 512-record validation census and binary-selected recipes directly on the four ternary worlds; those six ternary development fits were not run. The preliminary development competence check used the first 24 chain-3 graphs; reported fresh metrics use all 192. Networks receive a fixed canonical graph serialization, without positional embeddings; the planned explicit reindexing augmentation was not run. The seven-call tree word uses level order rather than the proposed postorder. These details define the retained experiment; no fits are changed for this import. The fixed feedback sequence has no general event-ID deduplication API, so the new MR semantic checks do not certify that interface. The existing RF bookkeeping tests belong to their separate implementation.

Unseen-tree feedback task. Each of twelve fixed schedules selects three distinct, previously unseen trees G_1, G_2, G_3 over the same four-coordinate source X; four rotations choose which source coordinate X_i is first observed. The method predicts the other source coordinates and G_3's output, then observes the outputs of G_1 and G_2 in order. Each tree has fresh local noise conditional on the same X. After each observation, each fitted model has the full source-belief update interface and reanswers the source and G_3 queries with fixed parameters; the no-update and product-projection controls alter only that belief step. The source rotations and schedules are integrated exactly within each world, including teacher outcomes, before the world-level feedback gains are compared.

Table 2: All retained topology results: mean categorical excess Brier over eight binary or four ternary worlds. GC-M uses its coherent forward-slice Bayesian adapter. All models are frozen during evaluation; seven-call and diamond diagnostics reuse the binary fits. Dashes denote arms outside the ternary matrix.

Setting	K-J	K-M	GC-DAG	GC-T	GC-M
Binary chain-3	0.000354	0.000268	0.001992	0.004799	0.006386
Binary tree-3	0.000320	0.000275	0.021835	0.014820	0.013747
Binary diamond	0.000319	0.000288	0.030704	0.018975	0.016640
Binary matched chain	0.000401	0.000275	0.002805	0.008024	0.008514
Binary chain-7	0.000364	0.000260	0.002450	0.004987	0.007018
Binary tree-7	0.000330	0.000252	0.018567	0.012216	0.008938
Ternary chain-3	0.003182	–	0.013197	–	0.011927
Ternary tree-3	0.003078	–	0.013130	–	0.011862

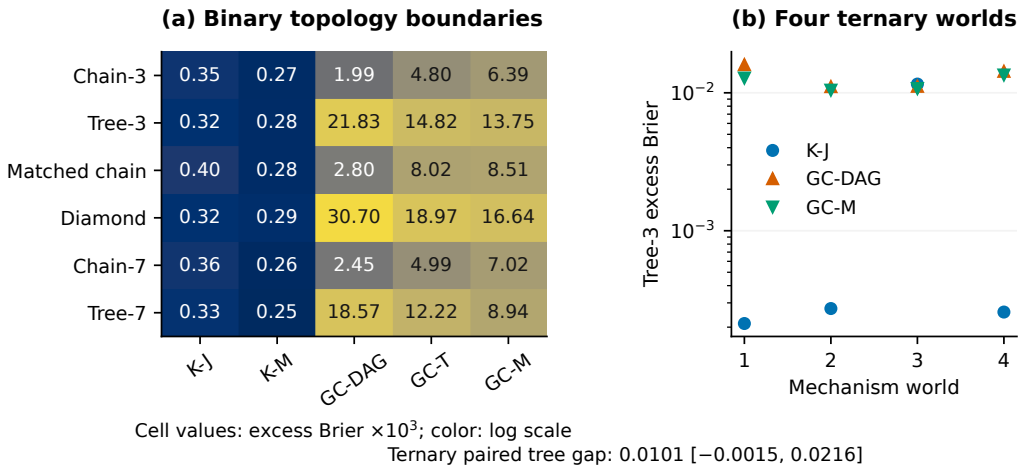


Figure 6: **Effects across wiring patterns and alphabets.** Left: mean binary excess Brier, with cell values multiplied by 10^3 and logarithmic color intensity. The diamond retains a shared intermediate variable; seven-call graphs also repeat source variables. Right: every ternary world's tree error, with log vertical scale; the paired GC-DAG-minus-K-J mean difference has a 95% interval crossing zero. The ternary recipes come directly from binary development. The plots summarize these finite graph families and do not establish asymptotic scalability.

Native masks, dependence retention and uncertainty. GC-M's native binary chain/tree errors are $0.008492/0.014826$, compared with $0.006386/0.013747$ from its coherent adapter. Both views retain the topology pattern. The projected K-J control replaces each source posterior by the product of its marginals after every observation. Its singleton marginals are unchanged at projection, but its joint predictions and later updates can change. The true-kernel oracle uses the same schedules and exact expectation over every source and feedback outcome. It has no finite-data estimation error. Strong coherent baselines use the same full source posterior interface as K-J; the projection is a compression ablation only.

The predeclared primary binary test uses a paired-world Student-t interval at 95%, with a 0.002 practical point-difference threshold. The ID tolerance check uses a 90% interval and 0.005 margin. Secondary topology, feedback and ternary intervals are descriptive, unadjusted for multiple comparisons. The primary half-width is 0.014877, wider than the planned precision target 0.0025; all eight worlds remain included. No positive secondary result substitutes for the primary comparison.

Reconstruction and cost. The import reconstructs all family/arm means from 56 retained raw result files, recomputes paired-world intervals from the complete cohort, and binds the 52 checkpoints and source records by hash. The earlier execution audit compares the finite hidden-state summation to a separately written enumeration and obtains zero maximum discrepancy on its fixtures. A scalar versus batched GC-M query differs by 7.56×10^{-10}. These are arithmetic and implementation checks, not independent training replication. All new paper figures are generated from the retained records. The experiment ran on four CPU threads to avoid an occupied GPU. Development and confirmation run bodies took approximately 33.2 minutes together, excluding the failed first checkpoint-save attempt, implementation, audit and figure preparation. No GPU-active time or large-model training was used in this batch.

E Directed repair and mechanism diagnostics

E.1 E1 query space, retention, and fresh acquisition

The E1 observation design and primary query are defined in Section 4. The complete 36-query family has function-space rank eight, including the constant, over the 64 source states. Its seven free

coordinates are the six nonconstant moments through order two on (h_0, h_1, h_2) and $\tau = \mathbb{E}[h_0 h_1 h_2]$. Equation 7 shows that changing τ while holding the six lower moments fixed changes the conditional answer in a prescribed direction and leaves the evidence probability fixed. This is a query-relative representation result; the full six-variable law has 63 free coordinates.

For the exposed eight-world T condition, four saved estimator laws per world were decoded in four ways. The exact target-block marginal and a summary of its six lower moments plus its own τ retain all seven free coordinates. A pair-only summary keeps just the six lower moments; a wrong-direction summary instead adds $\mathbb{E}[h_3 h_4 h_5]$ at the same coordinate budget. The two sufficient states match every answer from their respective full law to maximum absolute difference 4.44×10^{-16}; the pair-only and wrong-direction states can differ by 0.195002. The reported retention comparison uses each estimator's own fitted law. An initial audit summary had reused the historical J-MLE pair projection as the reference for C-PL, AR, and MD; the corrected replay recomputes each projection, with pair-moment residual at most 6.58×10^{-9}. It changes neither this four-state comparison nor the fresh confirmation below. Both initial and corrected records remain indexed in `research/2026-09-25-directed-repair-design/retention/`.

For the primary evidence tuple $e = (e_0, e_1)$, the full conditional response whose triple-direction difference appears in Eq. 7 is

$$q_e = \frac{1}{2} + \frac{\gamma\mu_2 + \gamma^2(e_0\mu_{02} + e_1\mu_{12}) + \gamma^3 e_0 e_1 \tau}{2D_e}, \qquad D_e = 1 + \gamma(e_0\mu_0 + e_1\mu_1) + \gamma^2 e_0 e_1 \mu_{01}.$$

The confirmation froze code and configuration after four development worlds, then generated 32 independent new worlds. Every arm shares 4,608 noisy pair records (9,216 observed scalars) through the known bit-flip channel. T7 and W7 each add m triple records (three observed scalars each); MorePairs adds $3m/2$ pairs at the same scalar budget. T7 and W7 preserve the same six lower-order target moments and differ in the triple direction used for reconstruction. QMLE8 fits the seven-coordinate target marginal directly from T7's records; F64 fits the full source law with the same data. The learner receives neither the teacher law nor its restricted generative parameterization. Risk is the teacher-evidence-weighted sum of squared errors for the binary query probability over four evidence values, with no extra division by four.

For each world, draw independent signs $s_0, \ldots, s_8 \in \{-1, +1\}$ and ρ with a uniform random sign and $|\rho| \sim \mathrm{Unif}(0.4, 0.7)$. The teacher source law is

$$p(h) = \frac{1 + \rho h_0 h_1 h_2 + 0.015 \sum_{i=0}^{5} s_i h_i + 0.02 \sum_{j=0}^{2} s_{6+j} h_j h_{j+3}}{64}, \qquad h \in \{-1, +1\}^6.$$

Each prefix record draws $h \sim p$, chooses one of the 15 pairs uniformly, and observes its two bits after independent flip probability 0.1. The T7 and W7 suffixes use the same fresh source draws and bit-flip noise but reveal respectively bits $(0, 1, 2)$ or $(3, 4, 5)$; MorePairs uses independent fresh pair records.

The shared prefix first fits an eight-state law $\widehat{p}_{\mathrm{pre}}$ for (h_0, h_1, h_2) by the known partial-observation likelihood, then extracts its six single and pair moments $\widehat{\mu}_i, \widehat{\mu}_{ij}$. For an observed pair $J \subset \{0, \ldots, 5\}$, this eight-state fit uses only $I = J \cap \{0, 1, 2\}$: a target state x assigns the record likelihood $\prod_{i \in I} P(Y_i = y_i \mid h_i = x_i)$ under the known flip channel. A mixed pair contributes a singleton likelihood; a pair entirely outside the target block contributes a constant and is omitted from the eight-state objective. The 64-state fit instead uses both observed bits of every pair. Let $c(x) = 1 + \sum_i \widehat{\mu}_i x_i + \sum_{i<j} \widehat{\mu}_{ij} x_i x_j$ for $x \in \{-1, +1\}^3$. The pair-only decoder is the maximum-entropy eight-state law matching these six moments. T7 holds them fixed, estimates $\widehat{\tau}_{\mathrm{raw}} = m^{-1}\gamma^{-3} \sum_{r=1}^{m} Y_{0r} Y_{1r} Y_{2r}$, and clips it to $[\max_{x_0 x_1 x_2 = 1} -c(x), \min_{x_0 x_1 x_2 = -1} c(x)]$. It reads queries from $\widehat{p}_{\mathrm{T7}}(x) = [c(x) + \widehat{\tau} x_0 x_1 x_2]/8$. At $m = 0$, T7 and W7 both use the pair-only decoder. W7 estimates the corresponding moment on (h_3, h_4, h_5), but its target-block decoder remains the shared pair-only law; its target-query predictions are therefore unchanged across extra budgets. MorePairs refits the six target-block moments using the extra pairs, then applies the same maximum-entropy decoder. QMLE8 and F64 maximize the known partial-observation likelihood on respectively eight and 64 states using T7's records. Likelihood fits constrain each state probability to at least 10^{-8} and select a maximum-entropy fit within 10^{-7} of the best normalized negative log-likelihood; the frozen solver settings and receipts are indexed in the experiment record.

Table 3: Fresh E1 confirmation: mean primary query excess Brier over 32 worlds. All arms share the pair prefix; T7/W7 and MorePairs have equal additional scalar budgets. QMLE8/F64 use T7's records, so their difference reflects fitting and representation rather than additional observations.

Extra triples m	T7	W7	MorePairs	QMLE8	F64
0	.022783	.022783	.022783	.022783	.022808
128	.003371	.022783	.022635	.003005	.003002
512	.002414	.022783	.022267	.001070	.001069
1024	.002182	.022783	.022037	.000607	.000606

The sole prespecified primary test compares W7 and T7 at $m = 1024$ by paired world. The mean difference is 0.0206010718, Student-t 95% interval $[0.0183566786, 0.0228454650]$, positive in 32/32 worlds. The interval half-width 0.0022444 satisfies the frozen 0.003 precision target, and the mean exceeds the frozen 0.005 material-effect target. Other arms and budgets are descriptive. The T7 raw third-moment estimate has RMSE 0.05908 at $m = 1024$; feasibility clipping occurs in two worlds. A secondary equal-family average over all 36 queries at that budget is 0.001015 for T7, 0.007884 for W7, 0.007540 for MorePairs, and 0.000302 for QMLE8/F64. This secondary score has a different weighting from the primary query. Frozen inputs, per-world predictions, budget accounting, and independent arithmetic checks are in `research/2026-09-25-directed-repair-design/`.

E.2 RF-2: A FIXED PAST-EVENT STATE

Fix $E = \mathbf{1}\{S_0 = 0\}$. A recursive state can retain the three masses $b(E = 1, S_t = s)$ and the three masses $b(S_t = s)$, subject to one normalization: five free coordinates. Transition and observation on S_t update each vector linearly before common renormalization. This state returns the event posterior, current-state marginal, future-AB endpoint law, and evidence probability, but it does not return all three initial-state category posteriors.

We replayed the saved oracle, J-L, C-L, and AR laws across the original twelve worlds, six roles, 1,024 episodes per world, and eight update steps. Relative to the full 3×3 belief, maximum event-answer error for the correct state is 1.22×10^{-15}; current, future-AB, and evidence-probability errors are below 10^{-10}. For J-L at horizon eight, mean squared event-answer differences from the full belief are 2.05×10^{-32} for the correct five-coordinate state, 2.05×10^{-32} for a correct six-coordinate grouping, 0.004552 for a wrong six-coordinate grouping, and 0.006791 for product retention. The unit remains the world; roles, episodes, and steps are repeated measurements. The wrong grouping stores the current-state masses and the row $P(S_0 = 1, S_t = s)$, plus the initially fitted split $\alpha_0 = P(S_0 = 0 \mid S_0 \neq 1)$. On later updates it reconstructs the event answer as $\alpha_0[1 - P(S_0 = 1)]$, holding that split fixed. Its extra coordinate does not restore the missing event–current row.

E.3 LOCAL-KERNEL AND CHECKPOINT DIAGNOSTICS

For the eight saved binary confirmation worlds, teacher-versus-fitted mean row TV is 0.022672 for K-J and 0.018770 for K-M; weighting rows by teacher reachability gives 0.02257 and 0.01899. Exact endpoint-law replay matches the original chain/tree risk means within 3.54×10^{-10}. The following hybrids use the same 192 chain and 192 tree graphs and 80 masks per graph as the original analysis.

Table 4: Teacher replacement of one fitted primitive at a time; mean categorical endpoint-query excess Brier. Hybrid risks are nonadditive and do not isolate one causal source of structural transfer error.

Fit	Graph	Both learned	Only A learned	Only B learned
K-J	chain	.000354	.000131	.000213
K-J	tree	.000320	.000133	.000195
K-M	chain	.000268	.000146	.000132
K-M	tree	.000275	.000156	.000130

Four separate development worlds were used for 12 new K-J/K-M/GC-DAG fits. Across 143 saved checkpoints, chain validation selected fits before testing a fixed subset of 24 paired chain/tree graphs per world (all eight operation words, three root permutations per word) and 80 masks per graph. Mean chain/tree risks at the first and selected checkpoints were, respectively, K-J $0.001729/0.002535$ and $0.000213/0.000252$; K-M $0.004508/0.007198$ and $0.000119/0.000144$; GC-DAG $0.007793/0.035951$ and $0.001918/0.032141$. The selected GC-DAG minus K-J tree difference is 0.031889, with descriptive four-world interval $[0.022965, 0.040814]$. Checkpoint points are not independent repetitions. These checkpoint summaries diagnose optimization and transfer in these recipes without adding a new composition confirmation or isolating the cause of the GC-DAG gap.

F SECONDARY STUDIES AND FULL RESULT DETAILS

The following results retain their original cohorts and metrics. They support the main acquisition and retention interpretation without being pooled with the 32-world directed observation test or the eight-world composition test.

Earlier observation controls (E1 and E2). In the separate eight-world E1 comparison, pair-only P, targeted-triple T, and random-triple R each use $12{,}288$ noisy observed scalars. Joint maximum likelihood improves from 0.021928 excess Brier under P to 0.000711 under T, a paired reduction of 0.021217 with 95% interval $[0.016893, 0.025542]$. Conditional pseudo-likelihood, autoregressive, and minimum-distance fits also have low T error. Pair projection of T fits leaves 0.020978 error. It preserves old answer a and evidence probability z to numerical precision but changes m and q (Figure 10a). The exact target-block marginal and a six-moment-plus-correct-triple summary reproduce all 36 answers to maximum difference 4.44×10^{-16}; pair-only and wrong-direction summaries can differ by 0.195.

E2 estimates mixture weights for six supplied three-state stochastic mechanisms. Each world uses $4{,}096$ budget units: either $12{,}288$ independent one-symbol episodes (Short), or $4{,}096$ three-symbol episodes (Linked or Weak). Short and Linked train at signal $\alpha = .8$, Weak at $\alpha = .25$; all evaluate at $.8$ on 512 common episodes across two policies and histories of lengths $0, 1, 2, 4, 8$. Joint maximum-likelihood multiclass excess Brier improves from 0.001415 to 0.000066 for Short versus Linked, a paired reduction of 0.001349 with interval $[0.000354, 0.002344]$. Conditional pseudo-likelihood and minimum distance also benefit. The two original primary p-values after Holm correction are 0.0000159 and 0.01496; intervals are unadjusted. Full objective comparisons are below.

Table 5: Separate eight-world E1 and E2 confirmations. Mean future-query excess Brier; lower is better. E1 columns are pair, targeted-triple, and random-triple observations; E2 columns are Short, Linked, and Weak records. Metrics and cohorts are distinct.

Experiment	Method	P / Short	T / Linked	R / Weak
E1	J-MLE	0.021928	0.000711	0.005006
	C-PL	0.020974	0.000646	0.009481
	AR	0.021468	0.000831	0.010049
	MD	0.021959	0.000708	0.006921
	Pair projection	0.021928	0.020978	0.021652
E2	J-MLE	0.001415	0.000066	0.000713
	C-PL	0.001295	0.000147	0.000555
	MD	0.001415	0.000070	0.000730
	Projection	0.001415	0.001417	0.001818

RF-1 conjunction and RF-2 replay. The twelve-world RF-1 joint learner improves the old answer after feedback relative to keeping it, by 0.07285 against the post-feedback oracle (95% interval $[0.06305, 0.08265]$); conditional and autoregressive learners give similar gains. Post-feedback and subsequent-query excess Brier are 0.001567 and 0.001810. The prespecified new-composition accuracy conjunction fails: its upper confidence limit 0.003116 exceeds the 0.0025 target. The effect interval half-width is 0.00980, versus the planned 0.001 precision.

RF-2 evaluates six role rotations and horizons $0, 1, 2, 4, 8$. Future-AB is the endpoint distribution after two transitions without intervening future observations. A float64 replay finds maximum full/product current and future readout discrepancy 5.56×10^{-16}; that agreement is relative to the fitted model, not zero teacher error. Projecting the same full predictive belief locally keeps a, z but sets $m' = az$ and $\Delta' = 0$. At step two, full revision RMS is 0.130470, versus below 10^{-15} locally. Recursive projection begins to diverge at the second observation and accumulates past error: at step eight its error is 0.023223, versus 0.001990 for the local one-step projection. The fixed-event five-coordinate state has event-answer mean squared difference below 3×10^{-32} from the full joint; a wrong-event six-coordinate grouping gives 0.004552, and an event–current product gives 0.006791. All are descriptive replays of exposed episodes. The separate uncertain-mechanism numerical diagnostic leaves $1{,}337/1{,}105{,}920$ cases unresolved; the table below uses the finite, fixed-mechanism reference.

Table 6: RF-2 horizon-eight mean excess Brier across twelve worlds. J-L and C-L are joint- and conditional-likelihood fits; AR is autoregressive. Full and projected J-L have identical tested forward readouts.

Learned law / retained belief	Initial state	Current state	Next AB
J-L	0.001989	0.001186	0.002826
J-L-projected	0.023223	0.001186	0.002826
C-L	0.001994	0.001186	0.002826
AR	0.001989	0.001186	0.002826

Fresh frozen-pixel update (RF-3). The frozen two-dimensional-latent VAE supplies a law over binarized MNIST. Sixteen observed pixels in one quadrant condition twelve query pixels around a central 2×2 region elsewhere. Only the four-pixel count is revealed; their configuration is hidden, and source, counted, and query pixels are disjoint. The twelve source/target rotations use 256 fresh images, eight streams, and nested $4{,}096/16{,}384$-particle budgets. Paired image intervals use $2{,}000$ bootstrap replicates. The earlier exposed RF-3 run queried distant pixels on 128 images; its prior-sampling gain was 0.000374 with interval $[-0.000238, 0.000995]$. The fresh test changes cohort, binarization, and target geometry.

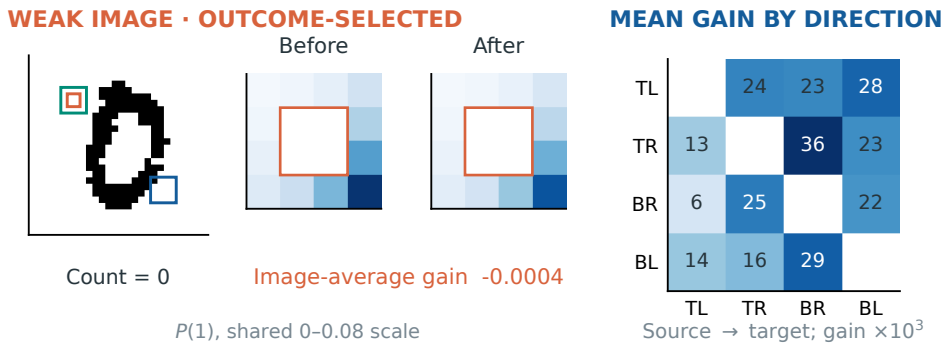


Figure 7: Fresh RF-3 weak-case and direction checks. The example is selected after observing its nonpositive image-average gain; the before/after maps show one direction on a common probability scale. Blue, coral, and jade outlines mark source, counted, and query pixels. The matrix reports mean observed-label Brier gain over all 256 images for each source–target direction (in 10^{-3}); diagonal directions are not tested. The label-selected example and all-image scores appear in Figure 5.

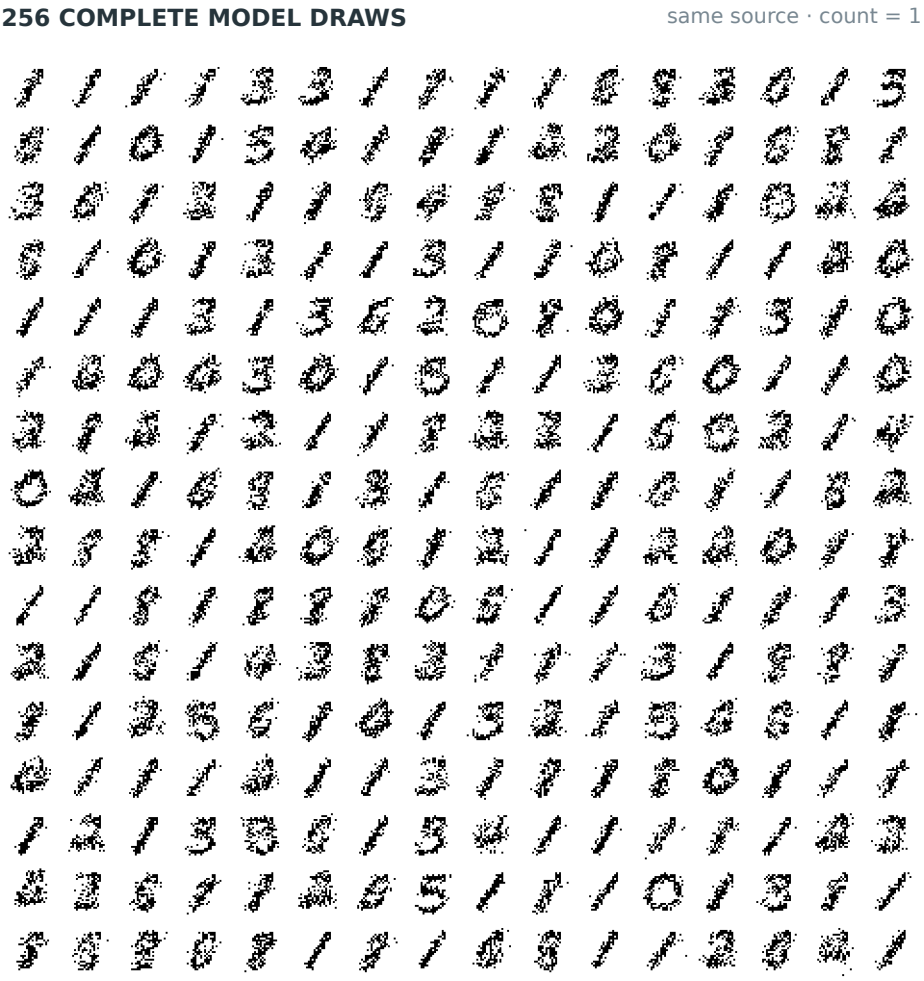


Figure 8: Full-image visual draws from the frozen RF-3 generator. All 256 complete 28×28 images condition on the same label-selected test case: its sixteen source bits and a count of one in the hidden target patch. Source bits are fixed in every draw; the latent variable and unseen pixels are sampled. This grid is separate from the 256 test images scored in Figure 5.

G Additional related work

World Models, PlaNet, and Dreamer learn latent dynamics for planning; PlaNet also filters a belief state (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020; 2025). Predictive-state methods study future-test readouts (Littman et al., 2001; Thon & Jaeger, 2015), and MuZero and value equivalence connect representations to planning values (Schrittwieser et al., 2020; Grimm et al., 2021). Linked observations enter spectral HMM learning under rank assumptions (Hsu et al., 2009). Distribution-shift benchmarks such as WILDS examine other ways a trained predictor can face new contexts (Koh et al., 2021).

Diffusion Forcing and DFoT vary conditioning histories (Chen et al., 2024; Song et al., 2025); MVA and UWM vary condition and target roles (Alzayer et al., 2026; Zhu et al., 2025), while PSI adds extracted structure to queries (Kotar et al., 2025). Generative Factor Chaining recombines skill joints (Mishra et al., 2025); composable diffusion and corrected or multimodal samplers assemble learned densities (Liu et al., 2022; Du et al., 2023; Sjöberg et al., 2026; Mishra et al., 2026). Compatibility and projective analyses formalize conditions for composition (Sun et al., 2026; Bradley et al., 2025; Bradley, 2026).

VAEAC, ACFlow, and Posterior Matching extend generative conditioning (Ivanov et al., 2019; Li et al., 2020; Strauss & Oliva, 2022); MaM learns consistent marginal functions (Liu et al., 2024). Kernel and joint-measure semantics formalize composition and conditioning (Fritz, 2020; Kozen et al., 2023), and Gen implements reusable generative functions (Cusumano-Towner et al., 2019). Linear estimability and conditional-query bounds inform the finite specifications here (Escanciano, 2022; Rossetto & Antonucci, 2026; de Campos & Cozman, 2005).

H Full figure comparisons and numerical diagnostics

These panels retain full comparisons and numerical diagnostics. Cohorts and scores follow their respective result sections.

Mechanism recomposition

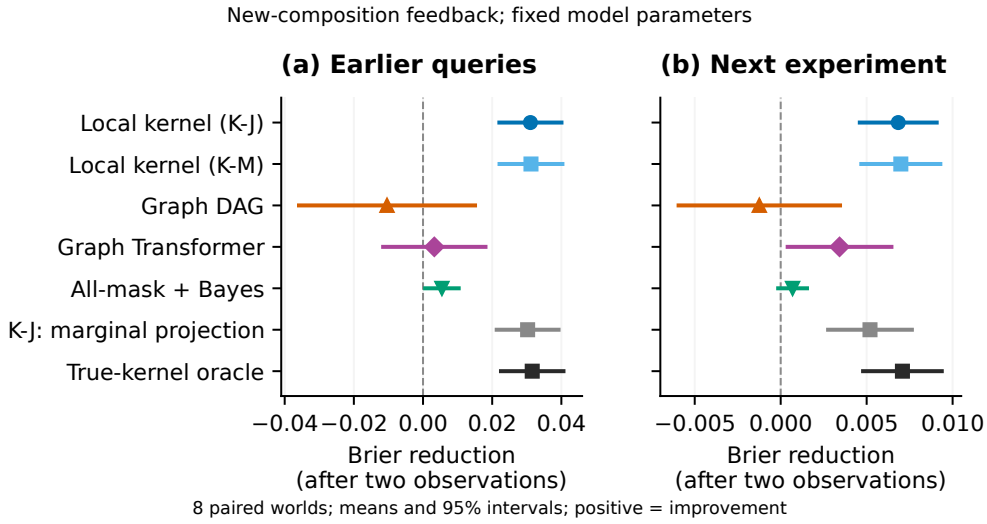


Figure 9: Feedback after two unseen-tree observations in eight binary worlds: old-source and next-tree Brier reductions. Dots and bars show paired-world means and 95% Student-t intervals; positive values mean improvement. Horizontal scales differ.

OBSERVATION INFORMATION AND RETAINED DEPENDENCE

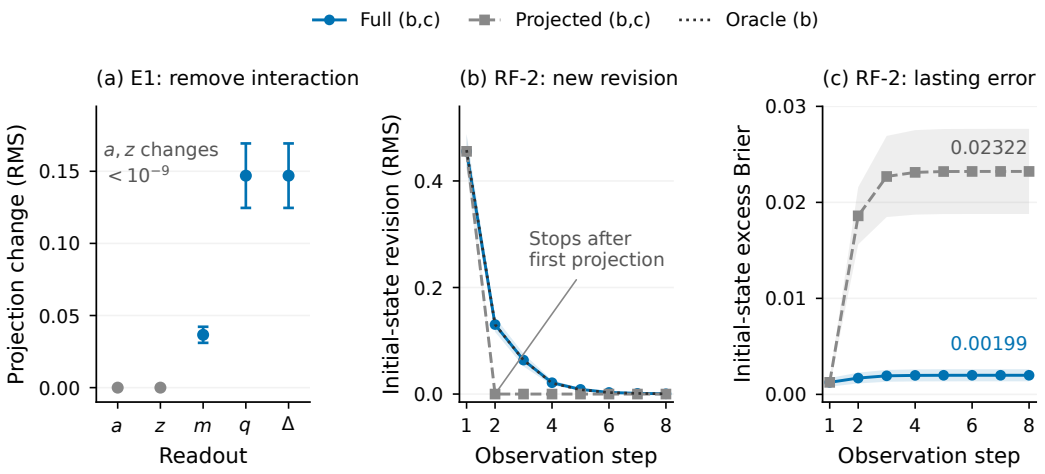


Figure 10: Measured interaction removal and successive revision. (a) In eight E1 worlds, pair projection barely changes the old answer a or evidence probability z, but changes the joint term m, new answer q, and revision Δ. (b,c) Across twelve RF-2 worlds, recursive product projection stops subsequent initial-state revision and accumulates error. Lines/bands are descriptive world-level 95% Student-t intervals. The local one-step counterfactual is distinct from the recursively projected path.

Frozen VAE inference and numerical reference (E3). The E3 VAE was trained on $50{,}000$ binarized MNIST images and then frozen. Posterior-matching (PM) proposals and an all-mask conditional comparator used validation-selected learning rates; five training runs completed 20 epochs. At test time, $1{,}000$ images supplied eight corner-patch or adjacent-patch-union masks, with 16 central query pixels. Particle interfaces used $128/512/2048$ particles and four seeds. At 2048 particles, observed-label Brier was 0.235952 for prior importance sampling, 0.237393 for direct PM, 0.235990 for corrected PM, and 0.241555 for the all-mask comparator. Corrected minus direct PM was -0.001403 with image-bootstrap interval $[-0.002453, -0.000381]$; the interval against prior sampling crossed zero. The original Gauss–Hermite reference passed $0/8000$ cases under both successive 0.002 maximum-query-change thresholds. Bounded outcome-aware Gauss–Legendre checks at $128^2/256^2/512^2$ nodes passed $6666/8000$ cases; on this stable subset, 2048-particle model-reference MAE was 0.006624 for prior sampling, 0.029716 for direct PM, and 0.005323 for corrected PM. The reference subset does not measure generator accuracy, whereas the observed-label scores include every case. This uses the established arbitrary-conditioning route of Strauss & Oliva (2022).

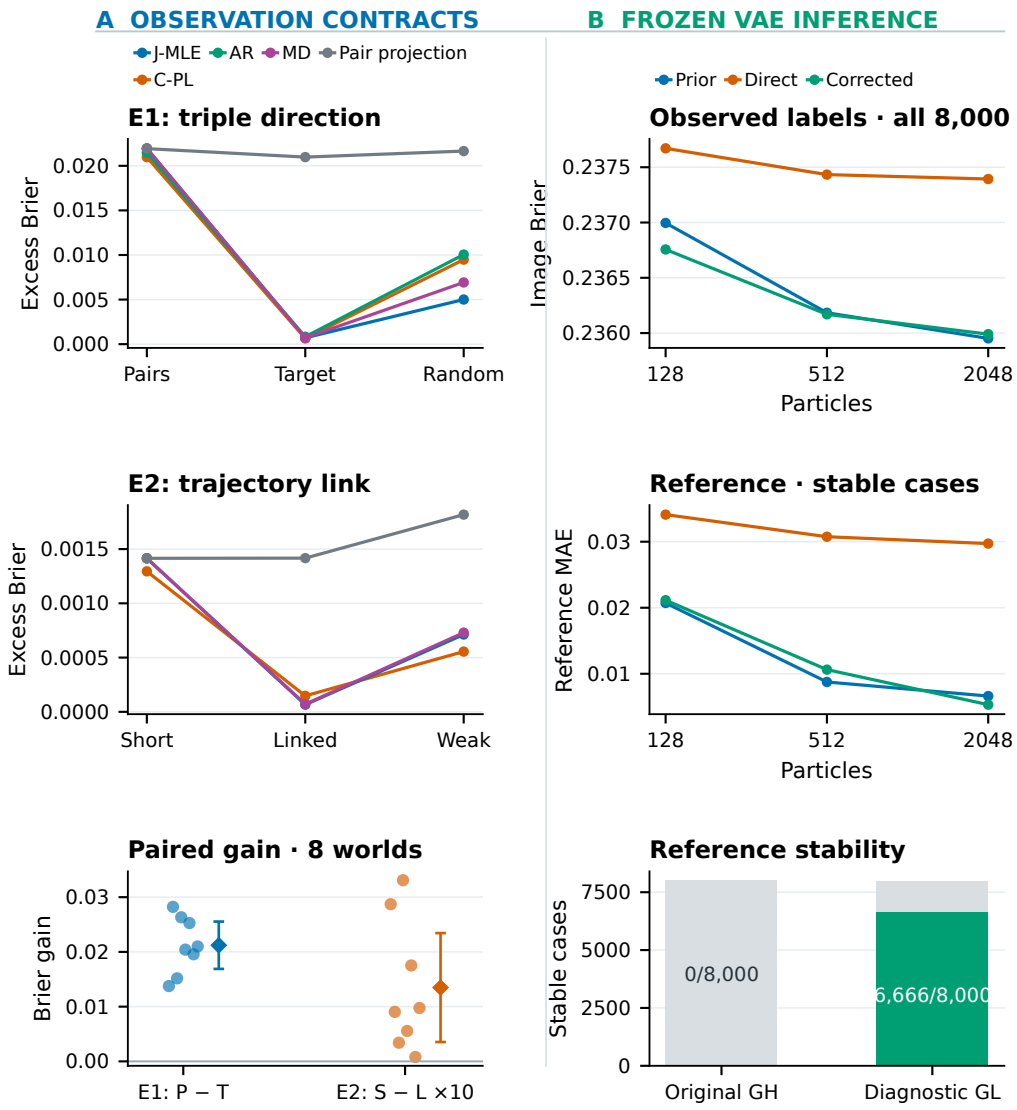


Figure 11: Separate diagnostics in one indexed view. Left: E1 and E2 objectives in distinct eight-world cohorts and metrics; paired points are worlds and diamonds show means with 95% Student-t intervals. E2 paired gains are multiplied by ten for display. Right: frozen-VAE observed-label Brier includes all 8,000 image–mask cases, model-reference MAE uses the 6,666 stable cases, and the bottom panel shows reference stability. The original numerical reference passed zero cases. Scales and cohorts are not pooled.

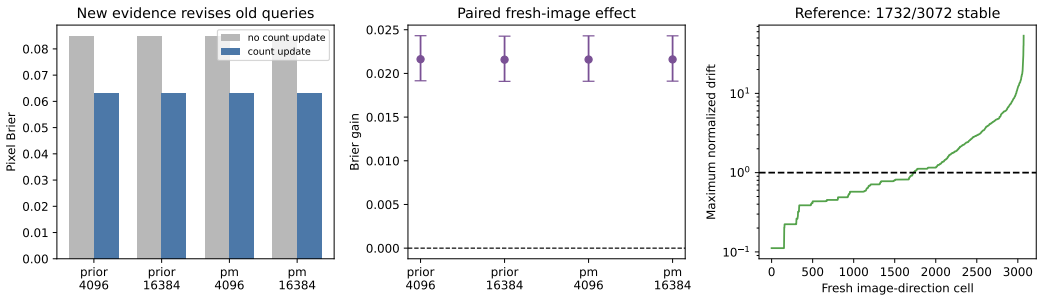


Figure 12: Fresh RF-3 numerical view retained from the original analysis. Observed-label Brier gain includes every image-direction case; the numerical-reference panel counts stable and unresolved cases separately. The label-selected actual image and all-image scores are in Figure 5; a weak actual image and direction gains are in Figure 7.

Table 7: Fresh RF-3: old pixel queries before and after count evidence. Lower Brier is better; gain is before minus after. PM denotes a corrected posterior-matching proposal for importance sampling. Intervals are paired by image.

Proposal / particles	Before	After	Gain	Gain 95% interval
Prior / 4096	0.084835	0.063208	0.021627	[0.01916, 0.02432]
Prior / 16384	0.084778	0.063205	0.021573	[0.01909, 0.02426]
PM / 4096	0.084800	0.063209	0.021591	[0.01910, 0.02430]
PM / 16384	0.084802	0.063199	0.021603	[0.01912, 0.02430]

I Error propagation under supplied composition

Section 4 separates observation information from retained state sufficiency. This section complements those boundaries with error guarantees under accurate local mechanisms. The first results propagate local probability error through supplied composition; the final witness shows why agreement on shallow trajectories does not generally identify deeper nonlinear behavior.

The general factorization in Eq. 3 gives a classical local-to-global explanation for topology and query reuse. Include any learned source distribution as a root factor. Let P_i and Q_i be true and learned local stochastic laws on this augmented finite DAG, and suppose $D_{\mathrm{KL}}(P_i(\cdot \mid u)\|Q_i(\cdot \mid u)) \le \delta_i$ for every true-support parent assignment u. The KL chain rule gives

$$D_{\mathrm{KL}}(P_G\|Q_G) = \sum_i \mathbb{E}_{P_G} D_{\mathrm{KL}}(P_i(\cdot \mid V_{\mathrm{pa}(i)})\|Q_i(\cdot \mid V_{\mathrm{pa}(i)})) \le \sum_i \delta_i. \tag{11}$$

Marginalization and the conditional KL chain rule then bound the evidence-averaged conditional KL for every fixed represented evidence/query partition by the same sum. For a binary query, Pinsker gives the corresponding squared-probability bound. These are standard probability identities, not a new theorem, and the experiments do not claim that small shallow scalar risk certifies the required local KL premise. In the chain, the stronger rowwise premise specializes to the existing $e_L \le L\delta/2$ certificate in Appendix A.

This also identifies why the local stochastic kernel is the reusable object in the paper. Under the stated rowwise premise, the same conditional tables can be placed at new locations in a supplied factorization, while the resulting joint and its conditionals inherit a local-to-global error bound. An endpoint predictor for one fixed shallow query does not by itself provide that factor-level contract.

A map $F : \Delta_n \to \Delta_n$ preserves every convex mixture if and only if $F(v) = vT$ for a unique row-stochastic matrix, whose rows are $F(\delta_s)$. This is the classical affine-extension property of a simplex (Perrone, 2018, Proposition 1.1.8). It distinguishes probability transport from merely normalizing

a vector. The statement concerns unconditional propagation through a fixed mechanism. Bayesian evidence updates generally do not preserve mixtures with their original weights.

Equation 18 makes the corresponding experimental intervention concrete. Chain execution feeds a point mass into one parent slot, whereas tree contraction averages the same conditional table over uncertainty in both parents. The held-out topology test therefore probes the stochastic-kernel semantics under a new contraction pattern, rather than only another input length.

Stochastic maps are nonexpansive in total variation. If the maximum row error is ϵ_j at step j, telescoping gives $\mathrm{TV}(v_L, \hat{v}_L) \le \min\{1, \sum_j \epsilon_j\}$ from identical initial states. Exact row identification implies exact finite-depth transfer by induction. These are classical Markov calculations (Lalley, n.d.; Rabiner, 1989), not new inference algorithms. The displayed TV bound controls the terminal-state distribution; it does not uniformly control rare conditional queries.

Exact shallow fits do not force deeper fits for unrestricted nonlinear simplex maps. Consider the illustrative Bernoulli-state recurrences

$$U(p) = \tfrac{1}{4} + \tfrac{1}{2}p, \qquad N(p) = U(p) + 4p(1-p)(p-\tfrac{1}{4})(p-\tfrac{3}{4}). \tag{12}$$

$N([0,1]) \subseteq [1/16, 15/16]$. Starting at either 0 or 1, both recurrences agree at depths one and two. At depth three, $N^{\circ 3} - U^{\circ 3} = -45/1024$, with squared probability error $2025/1048576$. This independently checked author witness clarifies non-identification beyond supplied shallow trajectories. It neither claims new general theory nor shows that the trained nonlinear network implements this map. The experiments do not directly measure mixture violations or identify them as the sole cause of the observed gap.

J Exact-law task, models, and supervision

To ask whether shallow supervision recovers something reusable, we must separate the numerical object learned from the structure supplied around it. This section therefore answers two questions in order: what local stochastic law is learned from endpoint supervision, and how can the frozen law be inserted into a different supplied composition? We define the joint-state process and learned kernel first, then the chain-to-tree wiring, the nonlinear comparison class, and finally the exact shallow supervision available to each arm.

For depth L, let four input traces be $x^{(r)} = (x_0^{(r)}, \dots, x_L^{(r)})$ for $r \in \{1, \dots, 4\}$. At each position j, one gate g_j is sampled and shared by all four traces, while gates are sampled independently across positions:

$$s_0^{(r)} = x_0^{(r)}, \qquad s_j^{(r)} = g_j(s_{j-1}^{(r)}, x_j^{(r)}), \qquad g_j \sim \pi. \tag{13}$$

Write $V_j = (s_j^{(1)}, \dots, s_j^{(4)}) \in \{0,1\}^4$ for the joint state and $a_j = (x_j^{(1)}, \dots, x_j^{(4)}) \in \{0,1\}^4$ for the observed input column. The final state V_L therefore has a 16-way joint distribution. The original readout observes three output coordinates $c = (c_1, c_2, c_3)$ and predicts the fourth, with target $p^*(y = 1 \mid c, a)$. The later query-reuse evaluation instead chooses any one output as the query and any nonempty subset of the other three as evidence. Inputs are externally supplied and do not update the gate prior. We separately train with a uniform prior over AND, OR, XOR, XNOR and a uniform prior over all sixteen Boolean functions of two bits. The second prior defines a separately trained family, not an unseen-rule evaluation of the first family's model.

The joint state is essential to the reuse question. Because the same hidden gate is applied to all four traces at a position, their outputs are generally correlated. Four separate Bernoulli marginals would not determine the conditional distribution of one output given the others. The 16-state representation retains exactly this joint information, so endpoint supervision can train a local transition through global joint targets and the frozen model can later answer changed evidence/query partitions from the same composed joint.

The simultaneous bits of four traces form sixteen joint states. The exact transition applies one shared gate componentwise across the four traces:

$$T_a^*(s, s') = \sum_g \pi(g) \mathbf{1}\{s' = g(s, a)\}. \tag{14}$$

A row-softmax tensor with $4{,}096$ learned logits parameterizes T_θ:

$$v_0 = \delta_{a_0}, \quad v_j = v_{j-1} T_{\theta,a_j}, \quad \hat{p}(y = 1 \mid c, a) = \frac{v_L(c,1)}{v_L(c,0) + v_L(c,1)}. \tag{15}$$

Here $T_{\theta,a}(s, y)$ is the learned probability of next joint state y from current joint state s when the observed input column is a. In depth-two-only optimization, the learner receives the observed columns $a_{0:2}$ and the exact final 16-way target distribution of V_2. It does not receive the sampled gate identities, intermediate states V_1, the teacher transition T^*, or direct depth-one targets. The 16-state representation, weight sharing, matrix multiplication, and conditioning operation are supplied; recomposition experiments additionally supply the evaluation graph. Only the numerical logits of T_θ are learned. A float64 dynamic program supplies exact endpoint-joint labels, with direct gate enumeration checking reduced cases. Exact inference is linear in depth for fixed context size; no speedup over this reference is claimed.

Throughout, *local stochastic law* denotes the conditional probability rule at one reusable factor, while *stochastic kernel* denotes its normalized matrix/tensor representation such as T_θ or K_θ. We use *Boolean gate* for the teacher-side primitive and *joint distribution* for the object obtained after composing the local laws.

Balanced-tree recomposition. The chain tensor is indexed as $T_{\theta,a}(s, y)$. We reuse exactly this tensor as the two-parent local factor

$$K_\theta(y \mid u, v) = T_{\theta,v}(u, y), \tag{16}$$

where $u, v, y \in \{0, \ldots, 15\}$ encode four-trace joint states. Chain training therefore exercises K_θ with u distributed as the recurrent state and $v = a_j$ fixed by an observed input column. The held-out topology uses the same entries while allowing both u and v to be uncertain. If p_L and p_R are the two child-subtree state distributions, an internal tree node is composed as

$$p_{\mathrm{par}}(y) = \sum_{u,v} p_L(u) p_R(v) K_\theta(y \mid u, v), \tag{17}$$

and Eq. 17 is applied bottom-up to a balanced binary tree. The teacher uses the same operation with the exact local factor and independently sampled node gates. Thus the topology test changes how a frozen local stochastic kernel is wired; it does not add topology-specific parameters or training.

Equivalently, the chain is the special case in which the second parent distribution is a point mass:

$$p'_{\mathrm{chain}}(y) = \sum_{u,v} p(u) \delta_{a_j}(v) K_\theta(y \mid u, v), \qquad p'_{\mathrm{tree}}(y) = \sum_{u,v} p_L(u) p_R(v) K_\theta(y \mid u, v). \tag{18}$$

The tree test therefore changes the kind of uncertainty presented to the same local factor: both parent slots can now carry composed distributions. This is why the chain-to-tree intervention tests more than additional applications of the original recurrence.

Nonlinear simplex comparator. The comparator uses the same initial delta, sixteen-dimensional simplex state, input columns, and conditional readout. It instead applies

$$N_\phi(v, a) = \mathrm{softmax}(W_2 \,\mathrm{ReLU}\,(W_1[v; \mathrm{onehot}(a)] + b_1) + b_2)\,. \tag{19}$$

Its hidden width is 80, with $3{,}936$ parameters. This yields a valid probability vector without enforcing mixture preservation. Architectures and optimization recipes differ; the experiment does not isolate every consequence of linearity while holding all other properties fixed.

Complete shallow joint supervision. All structural arms minimize final-output joint cross-entropy. Depth-two-only training uses the complete population of ordered four-trace inputs. A depth-two trace contains three bits and thus has $2^3 = 8$ possible values, giving $8^4 = 4{,}096$ ordered four-trace groups, including repeats. Each training target is the exact final 16-way endpoint joint; there are no intermediate-state, gate-identity, or depth-one labels. Full coverage adds a unit-weight depth-one joint loss: a depth-one trace has $2^2 = 4$ values, giving $4^4 = 256$ ordered groups. These one-step labels directly cover every transition row. Within an architecture, coverage arms share initialization, depth-two minibatches, and depth-two update counts. Full coverage adds supervision and computation; total compute is not matched. Depth-two labels may still indirectly identify local transitions, so removing one-step labels does not establish primitive discovery.

K EXPERIMENTAL CONTRACTS

The earlier exact-law studies answer four progressively narrower evidentiary questions. Their primary confirmation asks whether a law learned from shallow chains survives a post-training change of graph. The development study supplies the within-topology length evidence and the matched-risk architecture contrast. The outcome-aware extension asks what new conditional queries can be answered from an already learned joint. Earlier objective studies are retained separately because they address the training loss rather than the paper's main recomposition claim. We first define metrics shared by these contracts.

We average predictions over the six demonstration permutations with consistent output-context reindexing. Contexts are weighted by exact teacher mass; impossible contexts have zero weight. Using the single Bernoulli-probability convention,

$$e_L = \mathbb{E}[(\hat{p} - p^*)^2], \quad H_L = \mathbb{E}[(p^* - p_{\text{prior}})^2], \quad C_L = 1 - e_L/H_L. \tag{20}$$

The expectation averages the retained input distribution and native context mass. e_L is excess expected squared error over Bayes; H_L is the available improvement from demonstrations. An informative stratum excludes universal query/demo equality or complement and requires $|p^* - p_{\text{prior}}| \geq 0.1$.

For the topology and query-reuse experiments, let

$$\mathcal{Q}_{\text{mask}} = \{(q, E) : q \in \{1, \ldots, 4\}, \emptyset \neq E \subseteq \{1, \ldots, 4\} \setminus \{q\}\}, \qquad |\mathcal{Q}_{\text{mask}}| = 28. \tag{21}$$

For evaluated input group i and mask $m = (q, E)$, let e range over evidence assignments, define $w_{i,m}(e) = P_i^*(V_E = e)$, $p^*_{i,m}(e) = P_i^*(V_q = 1 \mid V_E = e)$, let $\hat{p}_{i,m}(e)$ be the corresponding conditional from the learned joint, and let $\rho_{i,q} = P_i^*(V_q = 1)$. For n evaluated input groups,

$$\begin{aligned} e_m &= \frac{1}{n} \sum_{i=1}^{n} \sum_{e} w_{i,m}(e) \big(\hat{p}_{i,m}(e) - p^*_{i,m}(e)\big)^2, \\ H_m &= \frac{1}{n} \sum_{i=1}^{n} \sum_{e} w_{i,m}(e) \big(\rho_{i,q} - p^*_{i,m}(e)\big)^2, \qquad C_m = 1 - e_m/H_m. \end{aligned} \tag{22}$$

Teacher-zero evidence cells have zero weight. If a cell has positive teacher evidence mass but zero learned evidence mass, the learned conditional is undefined and the corresponding gate fails. We use C_m only when $H_m > 0$. Metrics are computed separately for each seed. Reported "mean mask error" averages e_m over masks and seeds, while "worst seed/mask" is the maximum e_m over seeds and masks. For predicted and teacher joints $\hat{P}_i$ and P_i^*, mean joint TV is $n^{-1} \sum_i \frac{1}{2} \|\hat{P}_i - P_i^*\|_1$, then averaged over seeds when a table aggregates multiple runs. These conventions are shared by the balanced-tree, matched-chain, and query-reuse evaluations.

Joint TV and the mask errors probe different levels of the same reuse mechanism. Joint TV checks the composed generative object before a particular downstream query is chosen; the mask errors then check the conditionals extracted from that object. Using both therefore tests whether recomposition preserves the joint and whether the same joint remains useful after the evidence/query role changes.

Prospective held-out-topology confirmation. After a separate development probe succeeded, we froze a claim-specific confirmation before fresh training and before constructing its confirmation inputs. Five new seeds $111, 222, 333, 444, 555$ train only on the complete depth-two chain endpoint joint law, with no direct depth-one labels. Checkpoint selection uses only the fixed $1{,}680$-row distinct-input depth-two census. Once all selections are sealed, the learned local law is composed on supplied balanced binary trees of depths three, four, and five; the depth-four tree with 16 leaves and 15 mechanisms is primary. Every confirmation trace group is disjoint from the corresponding development topology group. Each tree is evaluated on all 28 represented evidence/query masks, and a 15-mechanism chain using the same 16 exogenous leaves is the matched topology control. The primary gate requires every seed to satisfy mask error $\leq 10^{-3}$, at least 99% headroom capture on meaningful masks, mean joint TV ≤ 0.01, and no undefined learned conditional for both tree and matched chain. No rescue seeds or secondary-depth rescue are allowed. This is fresh initialization-seed confirmation of a supplied balanced-tree recomposition claim, not independent implementation, learned topology, or arbitrary graph-family transfer.

Structural development study. The factorial design contains two priors, two coverages, two models, and seeds $77, 88, 99$. Linear models receive $6{,}000$ Adam updates at learning rate 0.03; nonlinear models receive $25{,}000$ at 0.003 with cosine decay to one percent. All use batch size 512, gradient norm cap 5, float32, and disabled TF32. One recipe per model is declared, without post-outcome repair or tuning sweeps.

Matching risk is computed over all $1{,}680$ depth-two ordered quadruples with four distinct inputs. These are training-population cases, not held-out validation. This contract deliberately fits the finite shallow law before evaluating greater length. Checkpoints minimize $|\log(e_{\mathrm{ID}}/\tau)|$ on a declared monitoring schedule, with earliest-checkpoint tie breaking. At $\tau = 10^{-4}$ both risks must lie within 10% of target and their ratio must be at most 1.1. Every seed and full run completion are required for recipe eligibility.

The primary comparison is four-rule, depth-two-only training at OOD depth four. Other prior and coverage conditions are predeclared secondary comparisons. A practical difference requires at least 10^{-3} excess Brier in every eligible seed. Best-ID checkpoints separately test accurate transfer: $e_L \leq 10^{-3}$ and $C_L \geq 0.99$ in every seed. Meaningful headroom requires $H_L \geq 0.01$ and informative-stratum mass at least 0.1.

All checkpoint selections seal before OOD construction. Each family uses $2{,}048$ input-only sampled unique unordered groups at depths $4, 5, 8, 16$, with four distinct traces per group. The four-rule family excludes previous OOD raw groups. Semantic-orbit disjointness is not asserted. The question follows earlier outcomes; fresh seeds and groups strengthen development evidence without making this a blind independently designed study.

Outcome-aware query-reuse extension. After the structural and confirmation outcomes were known, a separate analysis reuses the five all-rule/full-coverage best-ID stochastic checkpoints with no additional training. At depths two, four, and sixteen it evaluates all 28 choices of one query output and a nonempty evidence subset among the other three. The criterion requires excess Brier error at most 10^{-3} for every seed/mask and at least 99% capture of prior-to-posterior headroom whenever headroom is at least 0.01; all masks have meaningful headroom. For each evaluated input group, the analysis computes one raw joint and reuses it across all 28 masks before conditioning; it fails on undefined learned conditionals with positive teacher support. Because the checkpoints and source groups had already been exposed, this is a validated outcome-aware extension, not a fresh training replication or confirmation of a new architecture comparison.

Earlier objective studies. The earlier models share the same linear joint-state architecture and differ only in joint versus exact teacher-context-weighted conditional loss. They train at depths one and two with the four-rule prior. A corrected raw-group holdout removes query-role leakage but retains known semantic equivalents. A fixed-budget study and an ID-matched continuation use different seed cohorts and OOD raw groups. Their cross-study reversal cannot be attributed solely to checkpoint matching. The latter selects at held-out ID target 10^{-5}, with the same relative matching tolerances and practical floor 10^{-4}. This held-out matching metric is distinct from the structural study's complete shallow-population risk. Appendix P preserves the earlier results and their limits.

L RECOMPOSING A LEARNED LOCAL STOCHASTIC LAW

The compositional-generalization question is now concrete: after seeing only shallow chains, does the frozen learned factor still behave correctly when its surrounding composition changes? We report the graph change first, then use longer chains and changed queries as supporting evidence. The final subsection asks whether the original shallow predictive score could have certified this capability by itself.

L.1 PRIMARY TEST: RECOMPOSITION ON A NEW GRAPH

Before this test, low depth-two endpoint error leaves open two operational interpretations: the learner may have fit the observed chain composition, or it may have recovered the numerical local law needed beyond that composition. The tree test separates them. Training exposes only depth-two chains, where the second parent is an observed point mass; evaluation replaces the chain with a

supplied balanced binary tree, where both parents can be uncertain, while keeping the learned local factor frozen. All five fresh depth-two-only learners pass the primary tree gate, and all five matched 15-mechanism chain controls pass as well. At tree depth four, mean excess Brier across the 28 masks is 1.804×10^{-9}, the worst seed/mask value is 2.686×10^{-9}, and mean joint TV is 7.224×10^{-5} (Table 8). Predeclared depth-three and depth-five trees also pass, including the depth-five composition with 31 mechanisms. The result is therefore a topology-change result rather than only greater-depth execution of the training recurrence: a numerical local law learned through shallow chain endpoints remains accurate when wired into a different supplied topology on development-disjoint inputs. Because the graph and factor-sharing interface are given, this tests recomposition of a learned local stochastic law, not graph discovery.

Table 8: Fresh-seed held-out-topology confirmation. Five depth-two-chain-trained stochastic-kernel models are frozen and recomposed on supplied balanced binary trees. Mean error and joint TV average five seeds; worst seed/mask is the maximum over the five seeds and all 28 represented conditional-query masks. The matched chain uses the same 16 exogenous leaves as the primary 15-mechanism tree.

Topology	Mechanisms	Mean mask error	Worst seed/mask	Mean joint TV
Tree d=3	7	1.547×10^{-9}	3.281×10^{-9}	5.913×10^{-5}
Tree d=4	15	1.804×10^{-9}	2.686×10^{-9}	7.224×10^{-5}
Tree d=5	31	2.282×10^{-9}	3.794×10^{-9}	9.016×10^{-5}
Matched chain	15	1.798×10^{-9}	6.681×10^{-9}	5.580×10^{-5}

L.2 Supporting reuse under repeated composition and changed queries

Longer chains. Changing topology is the main intervention; longer chains ask the simpler question of whether the same numerical law also survives more repeated uses of the supplied recurrence. Table 9 reports every linear arm at its best-ID checkpoint. Both depth-two-only models transfer through all tested lengths. Every linear seed/depth combination passes the accuracy and meaningful-headroom criteria. At depth sixteen, prior headroom is 0.0600 for four rules and 0.1279 for all rules, with informative mass 0.707 and 0.863. The demonstration signal has not disappeared. This is supporting evidence for repeated composition without direct one-step supervision; it does not by itself test a changed graph.

Table 9: Linear stochastic-kernel models: mean OOD excess Brier at best-ID checkpoints, seeds $77, 88, 99$. The two priors are separately trained families.

Prior	Training labels	$L = 4$	$L = 5$	$L = 8$	$L = 16$
Four rules	Depth 2 only	8.943×10^{-9}	7.501×10^{-9}	4.909×10^{-9}	4.092×10^{-9}
Four rules	Depths 1 + 2	3.639×10^{-9}	3.221×10^{-9}	2.246×10^{-9}	1.893×10^{-9}
All 16 rules	Depth 2 only	2.004×10^{-9}	2.906×10^{-9}	4.504×10^{-9}	6.091×10^{-9}
All 16 rules	Depths 1 + 2	5.906×10^{-9}	8.354×10^{-9}	1.266×10^{-8}	1.548×10^{-8}

Changed query roles. The query-reuse extension asks whether the retained model remains accurate when the evidence/query partition changes. Every one of the 28 masks passes in every retained stochastic seed at depths four and sixteen. At depth sixteen, the mean excess Brier across masks is 5.52×10^{-9} and the largest seed/mask risk is 2.14×10^{-8}; mean joint total variation is 1.06×10^{-4} (Table 10). The corresponding depth-four values are similarly small. This adds query-role reuse to the length and topology changes above: for each evaluated input group, the same predicted joint distribution can be conditioned in all represented ways after training. The mask information was already present in the joint training labels; this experiment tests reuse of the learned joint distribution, not discovery of previously unrepresented variables.

Table 10: Outcome-aware query reuse for five retained best-ID stochastic checkpoints. Mean error averages the per-mask errors over the 28 masks and five seeds; worst seed/mask is the maximum per-mask error over those seed/mask units.

Depth	Mean mask error	Worst seed/mask	Mean joint TV
2	7.961×10^{-10}	2.113×10^{-9}	4.524×10^{-5}
4	3.036×10^{-9}	7.735×10^{-9}	9.091×10^{-5}
16	5.521×10^{-9}	2.141×10^{-8}	1.060×10^{-4}

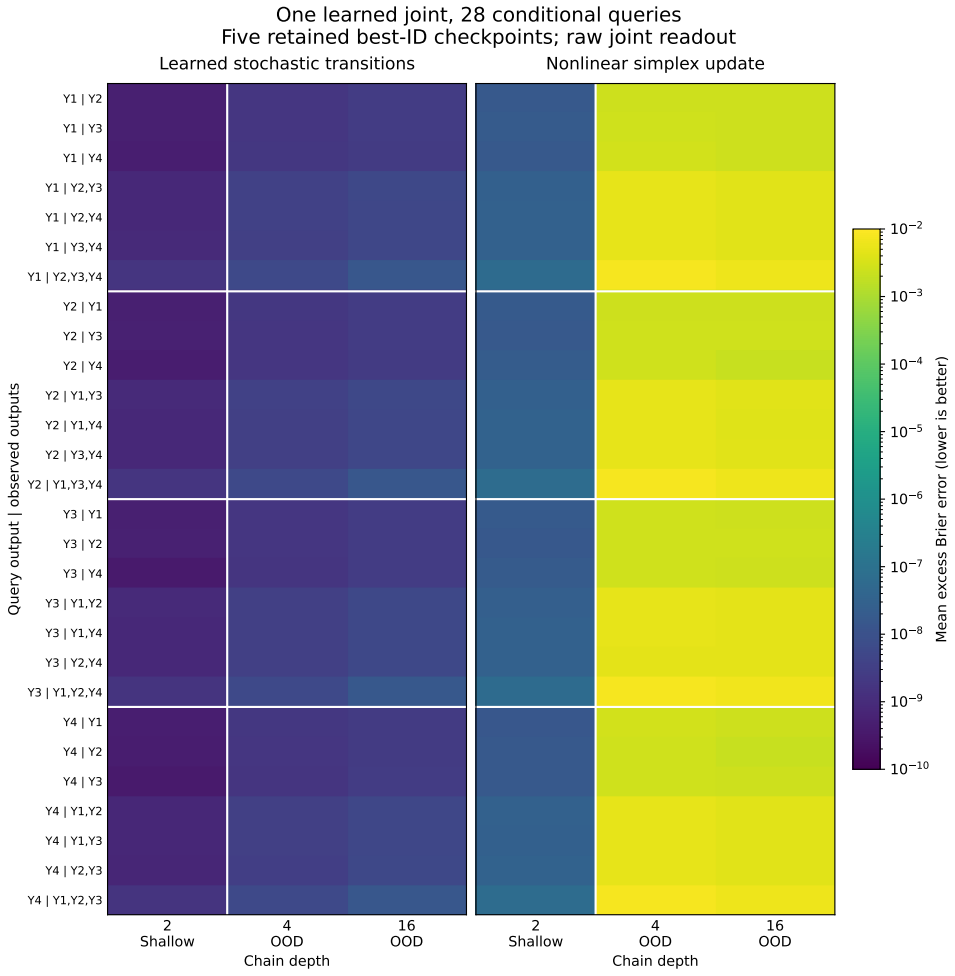


Figure 13: All represented evidence/query masks for the retained all-rule checkpoints. The stochastic model preserves low error at greater depth; nonlinear rows are descriptive because this extension does not create a new matched comparison.

Together, the positive results show three consequences of retaining a probabilistically composable local object: it can be rewired into the tested tree family, iterated farther within the training topology, and queried in multiple represented ways. The following comparison measures how the retained representation affects these uses at matched shallow predictive quality.

L.3 Preserving probabilistic structure through composition

The positive recomposition result shows that shallow supervision can recover a reusable law in this setting. We compare update classes at matched shallow conditional risk to measure their accuracy

Table 11: All structural matching outcomes at target 10^{-4}. Four-rule/depth-two- only is primary; the other rows are predeclared secondary conditions. Selected ID errors are means, while eligibility is tested per seed.

Prior	Depth labels	Linear ID	Nonlinear ID	Eligible pairs
Four	2 only	9.707×10^{-5}	2.389×10^{-3}	0 / 3
Four	1 + 2	1.021×10^{-4}	3.290×10^{-3}	0 / 3
All 16	2 only	1.010×10^{-4}	1.154×10^{-4}	0 / 3
All 16	1 + 2	9.539×10^{-5}	9.781×10^{-5}	3 / 3

Table 12: Eligible secondary contrast: all-rule prior, full shallow labels, depth-four OOD error. Difference is nonlinear minus linear. All three pairs are shown; they are not uncertainty estimates over task families.

Seed	Linear ID	Nonlinear ID	Linear OOD	Nonlinear OOD	Difference
77	9.526×10^{-5}	9.855×10^{-5}	5.514×10^{-4}	6.645×10^{-3}	6.094×10^{-3}
88	9.519×10^{-5}	9.875×10^{-5}	5.551×10^{-4}	5.970×10^{-3}	5.415×10^{-3}
99	9.574×10^{-5}	9.611×10^{-5}	5.543×10^{-4}	6.138×10^{-3}	5.584×10^{-3}

after deeper composition. The primary four-rule/depth-two contrast fails the matching gate; a predeclared secondary all-rule/full-coverage condition is eligible and is then prospectively replicated.

The primary comparison fails its shallow-fit match. The four-rule, depth-two-only primary contrast is ineligible in every seed: nonlinear shallow errors remain approximately 0.0024, above the 10^{-4} target. Table 11 retains all conditions. The four-rule full-coverage comparison is also ineligible. Their large OOD errors do not license equal-fit architecture conclusions.

An eligible secondary condition shows a material transfer difference. With the all-sixteen-rule prior and full shallow labels, all three pairs meet both matching tolerances. Nonlinear-minus-linear depth-four excess Brier differences are 0.006094, 0.005415, and 0.005584, each exceeding the declared 10^{-3} floor (Table 12). This supports a structural-recipe-associated difference in that secondary development condition; it does not pass the primary gate. The all-rule depth-two-only condition misses the tight target/ratio tolerances in all seeds. Those tolerances are not relaxed after observing its OOD gaps.

Prospective five-seed replication confirms the selected contrast. A separately frozen test fixes the all-rule prior, full shallow coverage, target 10^{-4}, depth-four primary metric, and the same fitting recipes. It uses seeds $101, 202, 303, 404, 505$ and $2{,}048$ new input groups per depth, excluding every development group. Finer ID monitoring changes checkpoint resolution without altering learning updates; all selections still precede OOD construction.

All five pairs meet the original tolerances and exceed the 10^{-3} effect floor. Mean depth-four errors are 0.0005362 for linear and 0.006099 for nonlinear updates, an 11.37 ratio. These correspond to 99.53% and 94.68% of available prior-to-posterior headroom captured, respectively. The paired mean difference is 0.005562, with the frozen 95% Student-t interval $[0.005275, 0.005850]$ over initialization seeds, conditional on the common test-input set. It does not estimate uncertainty across new input populations or task families. The smallest seed effect is 0.005227. Table 13 reports every pair. This confirms the selected all-rule/full-coverage contrast under its own protocol; it does not pass the original four-rule primary or confirm depth-two-only architecture attribution.

Predeclared secondary depths five, eight, and sixteen retain positive mean gaps of 0.004695, 0.004304, and 0.004347. These extensions do not expand the primary confirmation claim beyond the fixed family and coverage condition.

A post-outcome check measures the full shallow joint law. At these matched checkpoints, nonlinear updates also have lower full-depth-two joint KL and total variation in all five seeds: KL ranges from 0.000449 to 0.000941, versus 0.02785–0.02800 for linear updates; TV ranges from 0.00743 to 0.00995, versus 0.03224–0.03237. Thus the nonlinear comparator's larger OOD residual is not

Table 13: Prospective replication on five fresh seeds and new input groups. All pairs match shallow risk; OOD columns are depth-four excess Brier. Difference is nonlinear minus linear. The confidence interval in the text is conditional on this common test set.

Seed	Linear ID	Nonlinear ID	Linear OOD	Nonlinear OOD	Difference
101	9.379×10^{-5}	1.003×10^{-4}	5.345×10^{-4}	5.761×10^{-3}	5.227×10^{-3}
202	9.779×10^{-5}	9.971×10^{-5}	5.408×10^{-4}	6.047×10^{-3}	5.506×10^{-3}
303	9.550×10^{-5}	1.004×10^{-4}	5.296×10^{-4}	6.255×10^{-3}	5.725×10^{-3}
404	9.809×10^{-5}	9.933×10^{-5}	5.387×10^{-4}	6.064×10^{-3}	5.525×10^{-3}
505	9.500×10^{-5}	1.016×10^{-4}	5.375×10^{-4}	6.367×10^{-3}	5.829×10^{-3}

accompanied by worse shallow fit on either additional metric. This diagnostic was chosen after outcomes; it is not a formal joint-risk-matched comparison and does not identify a causal mechanism. The nonlinear model still retains substantial predictive headroom; the result concerns residual accuracy, not absence of transfer.

Training objective is a separate axis. Both earlier joint and conditional kernels retain accurate transfer through depth five. At the eligible held-out ID target 10^{-5}, mean depth-four errors are 1.599×10^{-5} and 7.149×10^{-6}, respectively, but the absolute difference is below the declared 10^{-4} practical floor. A same-seed/input replay reverses the residual ranking under a fixed-budget checkpoint policy (Appendix P). Both losses can therefore learn a reusable joint-state model in this setting; the paper's main claim concerns the retained probabilistic object and its uses rather than necessity of a loss labeled generative.

M SECONDARY REPLICATION DEPTHS

Table 14: Prospective replication, matched checkpoints. Depth four is primary; depths five, eight, and sixteen are predeclared secondary extensions.

Depth	Linear mean	Nonlinear mean	Mean difference
4 (primary)	5.362×10^{-4}	6.099×10^{-3}	5.562×10^{-3}
5 (secondary)	6.820×10^{-4}	5.377×10^{-3}	4.695×10^{-3}
8 (secondary)	7.931×10^{-4}	5.097×10^{-3}	4.304×10^{-3}
16 (secondary)	8.114×10^{-4}	5.159×10^{-3}	4.347×10^{-3}

N POST-OUTCOME SHALLOW JOINT-LAW DIAGNOSTICS

Table 15: Full depth-two joint-law diagnostics at the replication's matched scalar conditional-risk checkpoints. These are post-outcome analyses, not an additional joint-risk matching protocol. KL is teacher to model; TV is total variation.

Seed	Linear KL	Nonlinear KL	Linear TV	Nonlinear TV
101	2.793×10^{-2}	9.412×10^{-4}	3.224×10^{-2}	9.953×10^{-3}
202	2.785×10^{-2}	6.965×10^{-4}	3.229×10^{-2}	8.353×10^{-3}
303	2.793×10^{-2}	5.867×10^{-4}	3.225×10^{-2}	8.582×10^{-3}
404	2.795×10^{-2}	6.500×10^{-4}	3.237×10^{-2}	8.896×10^{-3}
505	2.800×10^{-2}	4.494×10^{-4}	3.231×10^{-2}	7.426×10^{-3}

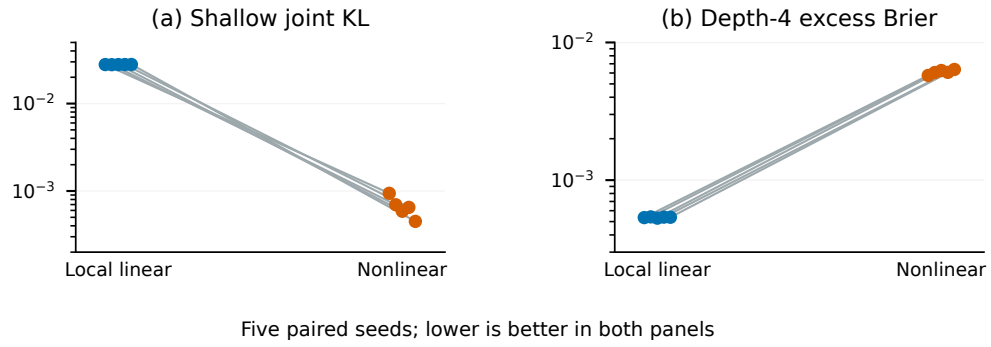


Figure 14: Five paired seeds in the separate exact-law study, selected at matched scalar shallow risk. Each line joins local linear and nonlinear fits from one seed. The nonlinear fit has lower depth-two joint KL but higher depth-four excess Brier in every seed. KL was examined after the outcome, not used for matching; the original four-rule match failed, and representation and optimization differ.

O DESCRIPTIVE STRUCTURAL CURVES

Table 16: All-rule family at each run's best-ID checkpoint, averaged over three seeds. This descriptive table has unequal achieved ID fit.

Depth labels	Update	Best ID	OOD $L = 4$	OOD $L = 16$
2 only	Linear	4.563×10^{-10}	2.004×10^{-9}	6.091×10^{-9}
2 only	Nonlinear	3.285×10^{-8}	1.011×10^{-2}	9.054×10^{-3}
1 + 2	Linear	1.057×10^{-9}	5.906×10^{-9}	1.548×10^{-8}
1 + 2	Nonlinear	2.621×10^{-8}	7.268×10^{-3}	6.186×10^{-3}

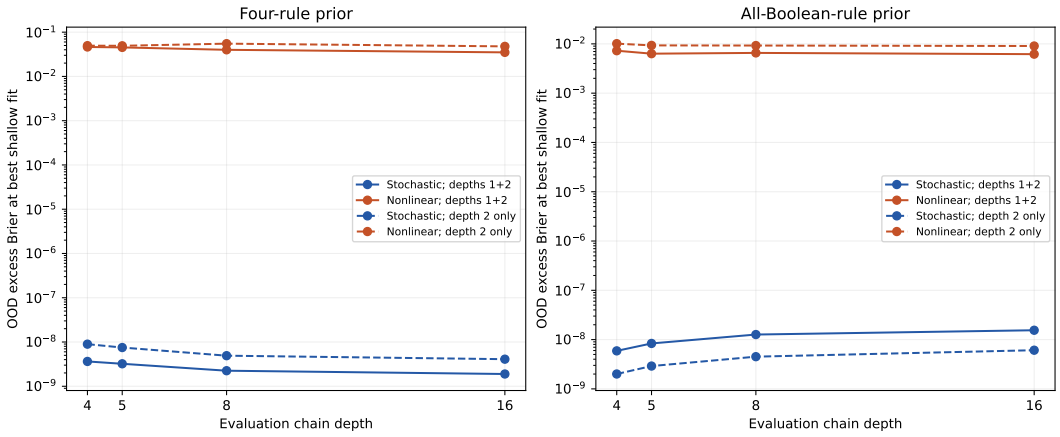


Figure 15: Unmodified structural-study figure: mean best-ID error versus depth for both priors and label coverages. These curves describe retained capability; matched-risk comparisons have separate eligibility tests.

P EARLIER OBJECTIVE COMPARISON

This appendix retains the earlier four-rule objective studies. Their corrected raw-group split trains on 256 depth-one and 3,298 depth-two tuples, with 312 validation rows representing 13 groups. Known semantic symmetries connect 288 rows in 12 groups to training equivalents. The fixed-budget and ID-matched studies differ in seed cohorts and OOD raw groups as well as checkpoint selection; the cross-study reversal alone does not isolate matching as its cause. The fixed-budget study uses seeds 11,22,33 and 2,500 updates. The continuation uses seeds 44,55,66, 6,000 updates per kernel and 25,000 per GRU; targets are selected using only the exposed ID validation set. Its primary target is 10^{-5}, with 10% tolerance, paired risk ratio at most 1.1, and practical floor 10^{-4}. New OOD raw groups number 189 at depth three and 2,048 each at depths four and five. The restricted depth-three support differs from the earlier input distribution.

Both learned kernels transfer beyond training depth. For each recipe's best-ID checkpoint, every seed at every evaluated depth meets the declared accurate-transfer criterion: $e_L \leq 10^{-3}$ and $C_L \geq 0.99$. Table 17 reports mean errors. The minimum capture across both kernels, seeds, and depths is 99.998858%. Errors also remain small in the informative stratum beyond universal copying. These checkpoints have different ID errors; the table establishes each recipe's capability, without isolating an objective effect.

Table 17: Frozen-weight transfer at each recipe's best-ID checkpoint, averaged over seeds $44, 55, 66$. Training uses depths one and two. Lower excess Brier is better; ID fit is not matched in this table.

Objective	OOD $L = 3$	OOD $L = 4$	OOD $L = 5$
Joint kernel	6.564×10^{-9}	6.917×10^{-9}	6.091×10^{-9}
Conditional kernel	6.360×10^{-7}	5.716×10^{-7}	5.324×10^{-7}

Fixed budget and matched fit give different residual rankings. Table 18 reproduces the predecessor's fixed-budget comparison. Joint training achieves lower ID error and lower OOD error. At depth four, both objectives recover more than 99.98% of available headroom. Its residual ranking thus comes with substantially different achieved ID fit and a small absolute effect.

Table 18: Predecessor with $2{,}500$ updates per recipe, seeds $11, 22, 33$. The mean excess Brier values have unequal ID fit.

Objective	ID $L = 2$	OOD $L = 3$	OOD $L = 4$
Joint kernel	2.579×10^{-7}	4.265×10^{-7}	4.340×10^{-7}
Conditional kernel	1.213×10^{-5}	7.937×10^{-6}	8.695×10^{-6}

In the continuation, all three seed pairs meet the primary ID-matching tolerances. Table 19 shows a mean OOD error of 1.599×10^{-5} for joint and 7.149×10^{-6} for conditional training. Conditional training has approximately 2.24 times smaller residual error, with a mean absolute difference of 8.84×10^{-6}. Each paired difference is negative (Table 20), and both recipes capture more than 99.97% of headroom. Neither direction meets the predeclared practical floor of 10^{-4}.

Table 19: Primary ID-matched comparison at target 10^{-5} and OOD depth four. Selected update counts differ; all declared training runs finish.

Objective	Selected update	ID excess Brier	OOD $L = 4$
Joint	900	9.904×10^{-6}	1.599×10^{-5}
Conditional	2,633	9.982×10^{-6}	7.149×10^{-6}

Table 20: Every primary seed pair. Difference is conditional minus joint depth-four excess Brier. Three initializations do not define uncertainty over task families.

Seed	Joint ID error	Conditional ID error	OOD difference $C - J$
44	9.730×10^{-6}	9.938×10^{-6}	-8.798×10^{-6}
55	1.008×10^{-5}	1.000×10^{-5}	-8.990×10^{-6}
66	9.899×10^{-6}	1.001×10^{-5}	-8.719×10^{-6}

Joint training reaches the selected ID accuracy in fewer updates. Conditional training has the smaller OOD residual at that level of aggregate ID error. Matching a scalar validation metric does not equate local error distributions, optimization paths, or all properties of fit. The result therefore supports an objective-associated ranking at the stated matching rule, not an isolated causal effect of the objective on a universal notion of extrapolation.

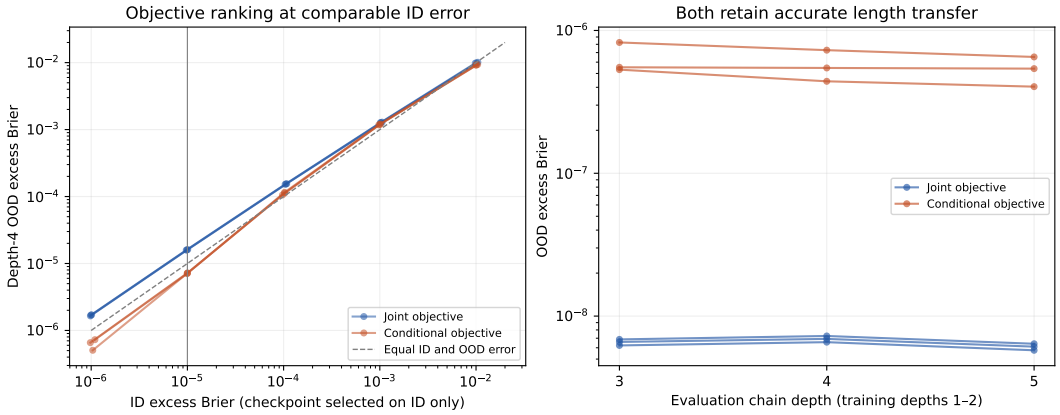


Figure 16: Retained continuation figure. Left: fixed ID targets and all seed outcomes; the vertical line marks the eligible primary objective target. Other targets are descriptive and may fail matching. Right: best-ID checkpoint transfer, with individual seed lines. The panels answer different questions, and lines are not confidence intervals.

A paired checkpoint replay removes the cohort and input confound locally. An outcome-aware replay uses the continuation's same three seeds, initializations, first 2,500 minibatches, evaluator, and 2,048 depth-four groups. At update 2,500, mean joint and conditional errors are 4.438×10^{-7} and 8.521×10^{-6}. At the original ID-matched checkpoints, the corresponding means are 1.599×10^{-5} and 7.149×10^{-6}. Every seed reverses ranking (Table 21). The reversal therefore occurs within these training trajectories under the two checkpoint policies. Every absolute gap remains below 10^{-4}; this does not establish practical superiority or a unique mechanism.

All six replayed original selected checkpoints are tensor-exact at retained anchors; 15,400 loss values and 814 ID-monitor values agree exactly. The original optimizer states and original update-2,500 tensors were not retained, so those cannot be compared directly. Separate arithmetic recomputation checks all 24 ID/OOD prediction metrics with maximum difference 9.15×10^{-20}. This is an outcome-aware development repair, not a fresh scientific confirmation.

Table 21: Same-seed, same-input replay. OOD is depth-four excess Brier. Difference is conditional minus joint. All paired signs reverse with checkpoint policy.

Seed	Selection	Joint OOD	Conditional OOD	Difference
44	2,500	4.393×10^{-7}	8.369×10^{-6}	7.930×10^{-6}
44	ID matched	1.586×10^{-5}	7.062×10^{-6}	-8.798×10^{-6}
55	2,500	4.482×10^{-7}	8.723×10^{-6}	8.275×10^{-6}
55	ID matched	1.614×10^{-5}	7.155×10^{-6}	-8.990×10^{-6}
66	2,500	4.439×10^{-7}	8.470×10^{-6}	8.026×10^{-6}
66	ID matched	1.595×10^{-5}	7.231×10^{-6}	-8.719×10^{-6}

The generic recurrent comparison remains ineligible. Both declared GRU fitting recipes fail the 10^{-4} ID target in every seed (Table 22). The best validation errors range from 0.00974 to 0.01305. A post-outcome diagnostic finds depth-two training error 3.47×10^{-5} but validation error 0.01133 for width 128, seed 66. Thus the failure already includes generalization to held-out depth-two inputs. It is not localized to greater chain length. These experiments cannot establish that joint-state representation causes the observed OOD advantage over the GRU.

Table 22: Best-ID GRU errors after the two declared fitting recipes. No seed reaches the required 10^{-4} target; an architecture comparison is unavailable.

Recipe	Seed 44	Seed 55	Seed 66	Matches at 10^{-4}
GRU, width 64	0.01276	0.01277	0.01123	0 / 3
GRU, width 128	0.00974	0.01305	0.01133	0 / 3

Q CONTROLS AND RETAINED PREDECESSOR FIGURE

Table 23: Predecessor controls, computed under the exact task distribution on the retained input sample. Oracle representatives and products of exact gate marginals are diagnostic controls with oracle information, not learned baselines. One consistent MAP sequence discards posterior uncertainty.

Control	ID $L=2$	OOD $L=3$	OOD $L=4$
Exact dynamic program	0	0	0
Eight oracle representatives	4.444×10^{-34}	5.820×10^{-3}	1.579×10^{-2}
Product of gate marginals	5.901×10^{-2}	4.846×10^{-2}	3.454×10^{-2}
Prior + universal copy/complement	6.077×10^{-2}	5.846×10^{-2}	5.717×10^{-2}
Input-dependent prior	1.149×10^{-1}	8.207×10^{-2}	6.821×10^{-2}
One consistent MAP device	1.314×10^{-1}	1.566×10^{-1}	1.481×10^{-1}

The exact dynamic program and full enumeration agree. The product-of-marginals control destroys posterior dependence among gates even though each marginal is computed exactly. Its error shows that these dependencies can matter in this fixture; it does not attribute the difference between learned architectures. The oracle representative control is a deterministic posterior approximation using eight midpoint-stratified representatives, not a trained particle method.

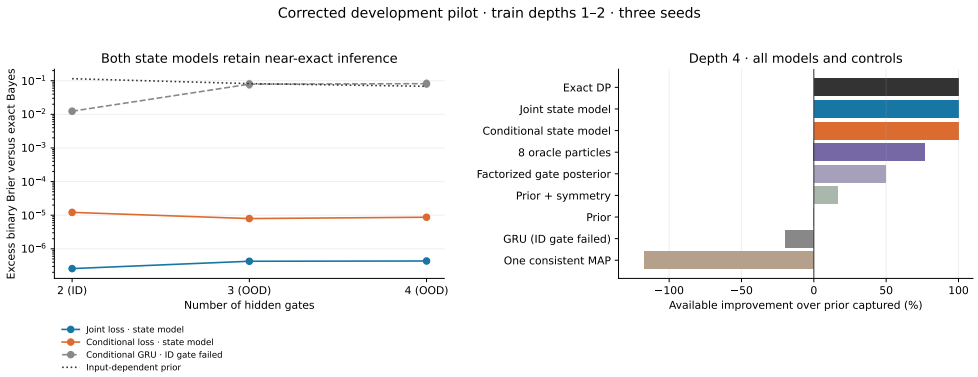


Figure 17: Unmodified predecessor figure. Seed means are descriptive development outcomes. Its fixed-budget ranking and the continuation's ID-matched ranking use different selection rules, seeds, and OOD input samples.

R Execution accounting and evidence chronology

The structural study completes all 24 runs within a 600-second training cap; 398.37 seconds elapse from start to result seal. Its original primary remains ineligible. The earlier continuation retains twelve complete runs, 186,000 updates, and 133.27 seconds to result seal on an RTX 5090 D, within its 450-second cap. These are execution-accounting facts, not speedup claims.

The first working draft remains preserved in a separate snapshot. The September 8 revision added completed structural development evidence and classical explanatory algebra. The present revision additionally imports the validated September 14 query-reuse extension and reorganizes the paper around compositional and conditional reuse; it does not relabel that outcome-aware extension as fresh confirmation. The September 17 held-out-topology study then freezes five new depth-two-only training runs and opens development-disjoint balanced-tree inputs only after checkpoint selection. All five tree units and all five matched-chain controls pass its primary gate, licensing the supplied balanced-tree recomposition claim. A prospectively frozen ten-run replication of the selected all-rule condition completes in 185.46 seconds. Its five fresh initialization units and new input groups support only the locked claim-specific contrast; the development results and original failed primary retain their earlier labels. The inherited driver uses a generic development-status string in its raw JSON, preserved unchanged; the frozen protocol and source/selection chronology define the prospective replication contract.

R.1 Finite observations: can a frozen law use new evidence?

The confirmed recomposition study uses exact endpoint probabilities. We additionally test whether finite observations can teach a law that supports later belief updates. These are adaptive, exposed development experiments, not additional confirmations. They change both the observation contract and the task; they do not enlarge the scope of the balanced-tree confirmation.

In a synthetic shared-world workflow (D8), each training cluster contains three episodes coupled by a hidden mode. Models learn from 8,192 clusters per world without true mode or coefficient labels. At evaluation, weights are frozen and new episodes provide evidence about the shared mode. J-J and J-C have the same normalized local-law representation, initialization and batches; they differ in joint versus conditional training loss. Structure and the mechanism family are supplied. For a fixed target and history length k, let p_k^* be its oracle event probability and $\hat{p}_k$ the learned prediction. We report

$$R_k = \mathbb{E}[(\hat{p}_k - p_k^*)^2], \qquad \Delta_8 = \mathbb{E}[(\hat{p}_0 - p_8^*)^2] - R_8.$$

Thus Δ_8 compares predictions before and after evidence against the same target oracle; it isolates update benefit from a change in evaluation target.

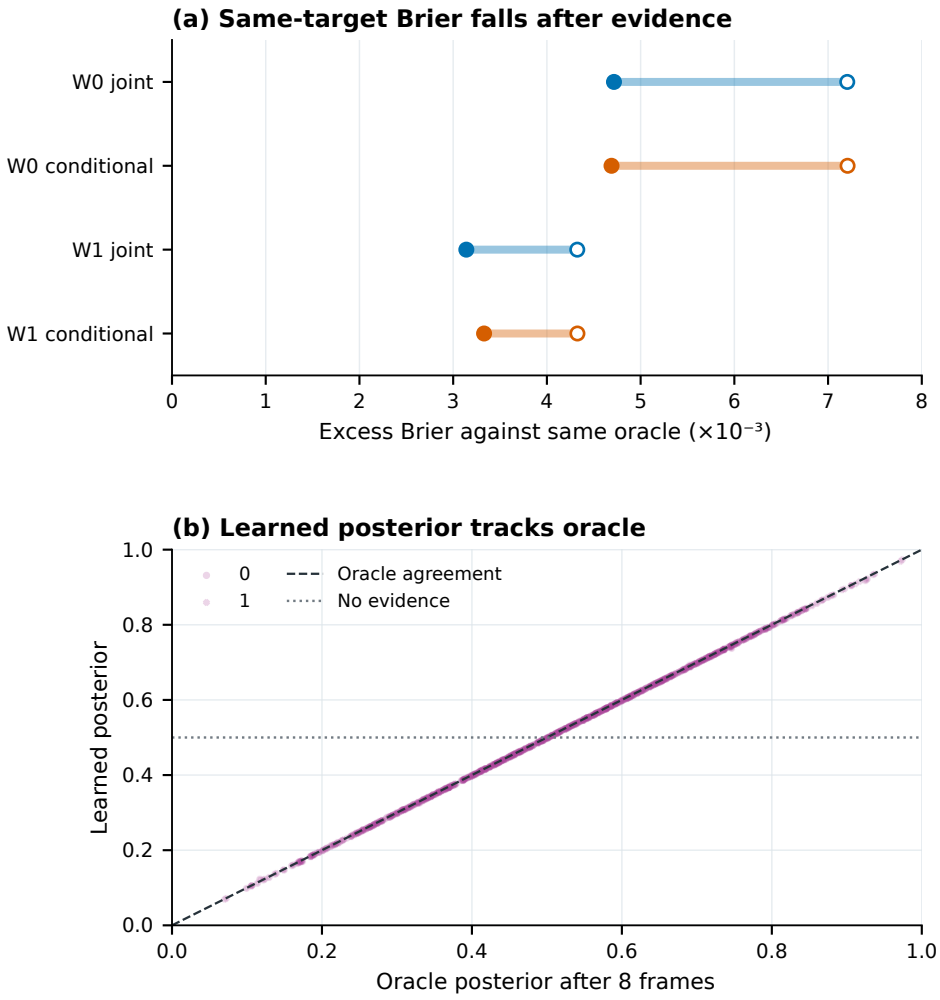


Figure 18: **What reuse changes in two controlled development settings.** **(a)** Finite-observation workflow: new evidence reduces the risk of both frozen local-law models in both exposed worlds (512 targets per world). Open circles are each model's own prediction without evidence; filled circles use eight observations. **(b)** Four-coordinate Gaussian imaging simulator: all 1,024 diagnostic targets from two worlds, with no point selection. A learned direct numeric law tracks the oracle posterior, whereas the no-evidence prediction is 0.5. The prior is supplied and clean/noisy pairs are available during training. The two panels have different information contracts and are not a pooled benchmark.

Both local-law objectives yield positive update benefit in both workflow worlds (Figure 18a). Relative to the oracle's no-update headroom $H_8 = \mathbb{E}[(p_0^* - p_8^*)^2]$, J-J recovers 33% and 18%; J-C recovers 34% and 13%. These are useful but incomplete updates, below the development target of 75%. The correct-family direct numeric fit wins in one world and loses in the other; no uniform advantage is established.

A separate noisy-imaging pilot (D7) uses four real cosine coordinates of a synthetic Gaussian scene, a known correlated prior, and an unknown four-mode observation law. Training sees 2,048 clean/noisy clusters with three frames; evaluation uses eight noisy frames without the clean scene. The direct joint and conditional fits both achieve $R_8 < 4.5 \times 10^{-7}$ in both worlds, against oracle headroom exceeding 0.032. This demonstrates learned update feasibility under a supplied classical family, not natural-image reconstruction or superiority of a neural joint model.

The extensions support the distinction between retaining a reusable probability law and choosing a particular training objective. Their failures matter too: longer-history errors grow, Qwen 0.5B reuse varies sharply by world, and some alternative updates increase risk. Appendix S reports these controls and the limits of dependency, decision and cost diagnostics.

S FINITE-OBSERVATION DEVELOPMENT: CONTROLS AND BOUNDARIES

All results in this section use exposed development worlds. The workflow and imaging pilots each contain two worlds; Qwen contains three. Each reported prediction risk uses 512 diagnostic targets per world. These targets are not independent training replications, and we do not construct confidence intervals over them as if they were. No additional fresh-world confirmation is claimed.

Workflow controls. J-J and J-C learn the same four-mode normalized local law. C-K directly fits the supplied numeric family. C-H is a full-history Transformer; A-R-episode normalizes across episodes and is not a complete within-endpoint autoregressive baseline. J-S is the fixed-source conditional control. The all-mask conditional fit remains underfit; it cannot establish that conditional learning lacks the representational capacity for reuse. Training recipes and achieved risks differ, so this is a capability diagnostic, not an architecture-isolated causal comparison.

Table 24: Workflow risks and own-model update benefits, all multiplied by 10^3. Lower R_8 is better; positive Δ_8 means updating helps. Both worlds and all retained controls are shown. The all-mask and J-S recipes are subsequent development fits, not members of a fresh matched confirmation.

Model	R_8, world 0	R_8, world 1	Δ_8, world 0	Δ_8, world 1
J-J	4.716	3.140	2.489	1.184
J-C	4.689	3.330	2.519	0.996
C-K	6.409	2.770	0.822	1.079
C-H	11.116	7.116	-2.600	-2.174
A-R-episode	9.654	6.724	-0.617	-0.465
J-S	6.123	7.691	1.123	-3.590
all-mask	13.230	6.425	-3.981	-1.113

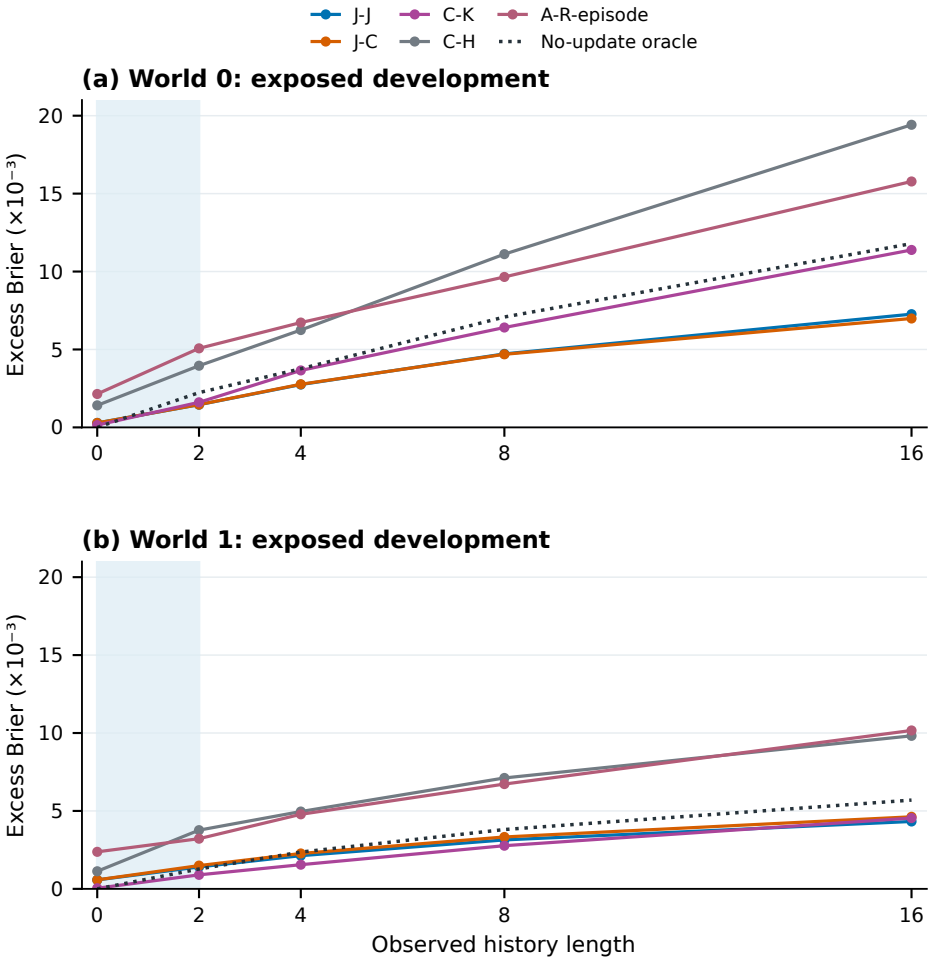


Figure 19: **Beyond training history, update accuracy degrades.** Training clusters contain three episodes, allowing histories up to two (shaded). Each point is evaluated against the oracle at that history length. The dotted line holds the oracle's no-evidence prediction fixed, so it measures available update headroom rather than a learned baseline. Increasing risk across histories alone does not imply that evidence harms a model: the oracle target also changes. Table 24 makes the same-target before/after comparison.

Small pretrained model. Qwen 0.5B uses rank-four LoRA on query/value projections. Nine adapted fits cover joint, conditional and history objectives across three worlds; a paired frozen control uses the same diagnostic inputs. Joint and conditional arms retain the same normalized-law representation. Figure 20 shows $100(1 - R_8/H_8)$, including negative values: negative recovery means worse risk than the oracle's fixed no-evidence predictor. Joint recovery is approximately $-121\%, 17\%, 26\%$; conditional recovery is $-79\%, -90\%, 25\%$. These are heterogeneous feasibility results, not a scaling trend. No comparison across model sizes or universal improvement from adaptation is established.

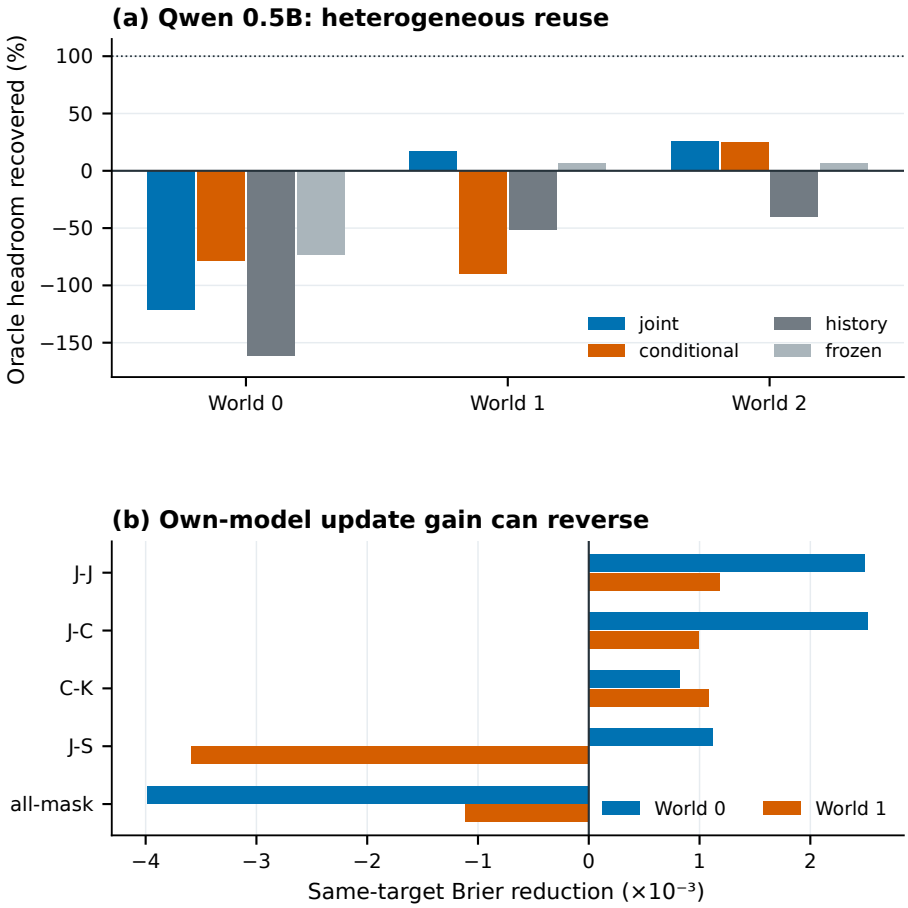


Figure 20: **The boundary of useful updates. (a)** All three Qwen worlds, including paired frozen controls and negative recovery; 100% means oracle agreement. **(b)** Same-target workflow risk reduction can reverse sign for fixed-source and all-mask fits. Panels use different denominators and should not be numerically pooled. These results constrain the scope of the positive demonstrations rather than invalidate the earlier tree confirmation.

Dependency removal is a single-endpoint test. The corrected, converged maximum-entropy projection diagnostic tests $P(\text{shipped} \mid \text{stock}, \text{health})$ on the first 128 exposed endpoints per world. Projecting the oracle to singleton marginals produces risks 1.88×10^{-3} and 5.46×10^{-3}; retaining pair marginals reduces them to 9.96×10^{-6} and 2.77×10^{-6}. This supports a role for dependencies in this query. It does not demonstrate that all higher-order structure is necessary, that pair marginals suffice for cross-episode updates, or that a complete joint is the minimal reusable representation. Endpoint degeneracy also makes this an insufficient test of general high-order necessity.

Decisions and cost. In the retained 64-cluster decision diagnostic per world, the oracle never buys an observation at cost 0.01. The declared setting therefore provides little support for a useful active-observation advantage. It does not upper-bound the regret of weak learned policies, which can still choose poor actions or waste observations. The current repeated-query timing implementation recomputes histories per query; it does not measure amortization from compiling one posterior and reusing it. Together with unmatched achieved quality, this prevents a claim of general quality–cost superiority or its impossibility.

Evidence status and reproducibility. The paper generator recomputes prediction risks from retained arrays, verifies the followthrough manifest, and emits vector figures, plot data and a quantitative ledger. Reported aggregate checks are same-author arithmetic validation, not an independent scientific replication. Original failed development runs remain in the research record; only the corrected converged projection is interpreted here. None of these extensions proves a new general learning theorem or that a joint objective is necessary. Full-history conditional models and other sufficient predictive representations remain legitimate alternatives.

T BAYESIAN INTERPRETATION AND EQUIVALENCE UNDER A SHARED JOINT LAW

This appendix makes explicit the classical probability semantics of the method. The shared joint law supports composition, changed query roles, and evidence revision; Bayesian inference provides their common interpretation. The identities below are algebraic consequences of conditioning. They do not introduce a new inference rule. Generative-program inference, filtering, and inference with a frozen generator supply established implementations (Cusumano-Towner et al., 2019; Rabiner, 1989; Strauss & Oliva, 2022).

One law, two ways to express revision. Fix the learned parameters, supplied interfaces, and variable-sharing contract. For a finite collection of permitted experiments, instantiate a common joint law over their shared uncertainty and local variables. Let P_t denote this law conditioned on the history D_t, including retained past variables when they are queried. For a fresh evidence event e, let L_e be its nonnegative incremental likelihood on this common space, with $0 < \mathbb{E}_{P_t}[L_e] < \infty$. The revised law is

$$P_{t+1}(d\omega) = \frac{L_e(\omega)P_t(d\omega)}{\mathbb{E}_{P_t}[L_e]}. \tag{23}$$

For any bounded query f, substituting this expression gives

$$\mathbb{E}_{P_{t+1}}[f] = \frac{\mathbb{E}_{P_t}[fL_e]}{\mathbb{E}_{P_t}[L_e]}. \tag{24}$$

Thus storing an updated posterior and querying it gives the same answer as querying the original joint with the new evidence. If the evidence is an event among the instantiated variables, L_e is its indicator. If it is a noisy or composite measurement, its observation channel defines L_e. The equality requires the same joint, observation model, and evidence history, with exact inference on both sides. Different fitted laws or numerical approximations need not give identical answers.

In coordinate notation, for disjoint T, C, B and supported events,

$$q_{T|C}(y \mid x, D_t) = \sum_b q_{T|C,B}(y \mid x, b, D_t) q_{B|C}(b \mid x, D_t), \tag{25}$$

$$q_{T|C}(y \mid x, D_t \cup \{X_B = b\}) = q_{T|C,B}(y \mid x, b, D_t). \tag{26}$$

The first identity marginalizes an unobserved quantity; the second assimilates its observed value. Moving the same event into history changes its bookkeeping, not its information content. For a sequence of events, the likelihood factors are incremental ones conditioned on the preceding history and experiment selection. Multiplying these factors gives the batch posterior by the chain rule, without assuming unconditional independence of the observations. Each event is included once.

For a composed law, a query target T and evidence coordinates C are roles:

$$q^t_{T|C}(x_T \mid x_C; G, \iota) = \frac{P_{\theta,t,G}(X_T = x_T, X_C = x_C \mid \iota)}{P_{\theta,t,G}(X_C = x_C \mid \iota)}.$$

Unmentioned variables are marginalized, and the denominator must be positive. The same variable can be a target in one question and evidence in another (Li et al., 2020; Strauss & Oliva, 2022). For disjoint supported I, J, C, the common law also gives

$$q_{I,J|C} = q_{I|C} q_{J|C,I} = q_{J|C} q_{I|C,J}.$$

Independently normalized prediction heads need not obey this identity.

A shared belief implements the same calculation. When H retains all uncertainty needed to connect experiments, integrating out a new experiment's local variables gives $\ell_e(h) = p_\theta(e \mid h, D_t, G_e, \iota_e)$. Applying Eq. 23 to H gives

$$\mu_{\theta,t+1}(h) = \frac{\mu_{\theta,t}(h)\ell_e(h)}{\mathbb{E}_{\mu_{\theta,t}}[\ell_e]}, \qquad q_c^{t+1} = \frac{\mathbb{E}_{\mu_{\theta,t}}[u_c\ell_e]}{\mathbb{E}_{\mu_{\theta,t}}[v_c\ell_e]}, \tag{27}$$

where $u_c(h) = \mathbb{E}[fL \mid h, G, \iota, D_t]$ and $v_c(h) = \mathbb{E}[L \mid h, G, \iota, D_t]$. If the query context's local randomness is independent of e conditional on H, D_t, its functions u_c, v_c remain unchanged and its revised answer is

$$\frac{\mathbb{E}_{\mu_{\theta,t+1}}[u_c]}{\mathbb{E}_{\mu_{\theta,t+1}}[v_c]} = \frac{\mathbb{E}_{\mu_{\theta,t}}[u_c\ell_e]}{\mathbb{E}_{\mu_{\theta,t}}[v_c\ell_e]}.$$

If the context reuses a realized local variable, that variable must be retained or the answer computed from the full common joint. Replacing it by a fresh draw changes the law. A new graph similarly constructs a new experiment law from reusable mechanisms; it is not generally a marginal of the old graph's observable joint. The cross-experiment joint used here is supplied by the sharing contract.

Why dependence enables revision. For the same bounded query and likelihood,

$$\mathbb{E}_{P_{t+1}}[f] - \mathbb{E}_{P_t}[f] = \frac{\mathrm{Cov}_{P_t}(f, L_e)}{\mathbb{E}_{P_t}[L_e]}. \tag{28}$$

This follows by subtracting $\mathbb{E}_{P_t}[f]$ from Eq. 24. For a fixed query and event, revision therefore requires nonzero covariance with that event's likelihood. Independence implies no change, while dependence alone need not change every query for every event. The joint stores the cross-moment that separate marginals can lose. RF-2 measures this consequence by preserving current-state and tested future-AB predictions while removing the dependence needed for earlier-state queries. The information analysis in Appendix B gives the corresponding finite-channel conditions for retaining query expectations and their likelihood-weighted revisions.

Dynamic and learned-generator realizations. In a dynamic model, marginalizing the preceding state gives the prediction law $\sum_s K_\theta(s' \mid s, h, a_t)b_t(h, s)$; multiplying by the observation likelihood and normalizing gives

$$b_{t+1}(h, s') \propto L(o_{t+1} \mid h, s', a_t) \sum_s K_\theta(s' \mid s, h, a_t)b_t(h, s). \tag{29}$$

Retaining past-state dependence allows later observations to revise earlier states, as in classical smoothing (Rabiner, 1989). For a frozen generator, a learned prior and decoder define $p_\theta(z, x)$; a supplied measurement channel extends this joint. Particle reweighting approximates its posterior queries. With particles $z_j \sim \nu(z \mid D_t)$ of adequate support,

$$w_j \propto \frac{p_\theta(z_j, D_t)}{\nu(z_j \mid D_t)}, \qquad w'_j \propto w_j p_\theta(e \mid z_j, D_t), \qquad \widehat{q}^{t+1} = \frac{\sum_j w'_j \mathbb{E}[f \mid z_j, D_t, e]}{\sum_j w'_j}. \tag{30}$$

Our pixel experiment freezes a VAE and composes its decoder with a four-pixel count operator. The count likelihood is a coefficient of $\prod_i[(1 - p_i(z)) + p_i(z)s]$; disjoint query pixels use their decoder probabilities. Prior and corrected posterior-matching proposals implement established inference routes (Kingma & Welling, 2014; Strauss & Oliva, 2022). This realizes the same probability semantics with learned components. It does not require a Bayesian treatment of the fitted parameters: the experiments keep θ fixed and revise uncertainty under the fitted law.

The equivalence concerns exact probability operations under a specified model. It does not establish that an arbitrary trained generator recovers the true joint, that separate conditional heads are compatible, or that approximate inference is exact. Nor does it guarantee that every realized observation improves every prediction. The experiments measure predictive quality and revision effects under their respective fitted laws, observation distributions, and inference budgets. Flexible world models can share these semantics; the empirical questions here concern learning reusable mechanisms, retaining the dependencies needed by later queries, and computing their consequences accurately.

U REPRODUCIBILITY AND EVIDENCE PROVENANCE

The anonymous supplementary materials contain the 32-world directed-acquisition observations, fitted laws, query operators, and prediction checks; the twelve-world fixed-event retention predictions and error recomputation; and the eight-binary-world and four-ternary-world mechanism and feedback aggregates with interval replay. The included README gives CPU commands and a file-hash manifest. These checks reuse stored artifacts and do not repeat training or recursive state-operator evaluation. The earlier audits described below refer to retained research records; their complete protocols, raw archives, checkpoints, and pixel-generator assets are outside this bundle.

Early objective and structure studies. The objective audit independently recomputed 216 prediction arrays and checked 72 checkpoint choices. Structural analysis recomputed 384 arrays with zero aggregate discrepancy; an independent structural audit also recomputed all retained strata and checked complete runs and selections. CPU/CUDA probability replay is not bitwise identical; reported risk differences leave gate classifications unchanged. The independent confirmation audit checks all 80 prediction files across five strata, every selected pair, fresh-input exclusions, and the prescribed interval; all 137 checks pass. The query-reuse extension additionally preserves 60 raw artifacts, 1,680 mask records, and 12 aggregate group decisions; a separately implemented arithmetic and forward replay passed its recorded validation. These audits are not full independent training reruns. The held-out-topology confirmation separately uses five fresh training seeds and development-disjoint inputs; an alternative flattened tree composition plus scalar conditional implementation reproduces all checked forward and metric values to at most 1.64×10^{-15} absolute discrepancy, while reusing the same trained checkpoints and frozen inputs.

Observation and retention studies. The two finite-data context confirmations preserve eight fresh worlds each, raw records, checkpoints, primary paired-world analyses, Holm adjustment over their two primary tests, and separate arithmetic and LP replays. Their conditional comparisons remain descriptive. The VAE reference failure and outcome-aware numerical diagnosis are kept distinct; neither audit constitutes independent training reproduction. The role-flexible import binds its protocols, analyses, raw predictions, and retained figure by hash. It checks the saved full/projected prediction identity across all twelve RF-2 worlds. Fresh RF-3 retains its pre-evaluation protocol, dtype repair record, raw particle predictions, paired image analysis, and arithmetic replay; its numerical reference and the earlier exposed run remain separately reported.

Mechanism recomposition and derived analyses. The finite-data mechanism import verifies all retained graph-family means against raw query records and reconstructs the paired-world statistics and new figures. Actual graph sampling, validation census and unrun ternary development fits are disclosed in Appendix D; this import performs no new model training or independent training reproduction. The interaction analysis reuses the full E1 and RF-2 cohorts and saved fits. It computes query, evidence, and mixed expectations directly, reproduces the original scores, and keeps local counterfactual projections separate from recursive trajectories. Its new moment statistics and intervals are descriptive analyses of exposed data. Finite-channel ranks, all retained RF-2 fitted response matrices, and stated algebraic examples have separate checks. These checks do not constitute new training or an independent confirmation cohort. Protocols, curves, targets, predictions, and checkpoints remain retained locally; no public release is claimed.