

From Autoregressive Generation to Decision Readouts under Exogenous Chance

Mark Joonho Kim

MindCanvas Co.

July 2026

Abstract

Autoregressive models (ARMs) and energy-based models (EBMs) are equivalent in function space, with the autoregressive factorization obtained by a soft Bellman recursion. We ask what this equivalence means when a sequence alternates between *controllable actions* and *exogenous events*. A teacher-forced model fitted to observational trajectories recovers reachable action and event conditionals, but its ordinary generative readout is not a control solution. The two readouts use different chance-node operators: generation applies an exponential log-moment, whereas control averages under the fixed environment kernel. We derive an exact local-to-global recursion for their value gap, separating locally created entropic optimism from propagated downstream gap. We then give (i) a reference-measure audit showing when event typicality enters a reconstructed energy, (ii) a pathwise decomposition of realized return into soft action advantages and mean-zero event innovations, and (iii) a finite-horizon model-error and support-gated regret bound. All algebraic identities are verified by exhaustive enumeration in a 20,736-trajectory LifeGrid world. A pre-specified 2×2 intervention varies transition entropy independently of continuation-return dispersion. Tabular estimation studies the coverage–error–regret trade-off, and a finite causal Transformer shows that a learned observational model can support a lower-error decision readout without making the model itself a policy. The results delimit a precise bridge from sequence generation to decision support: preserve exogenous chance under its environment law, optimize only controllable actions relative to an explicit reference, and abstain outside supported regions.

Keywords: autoregressive models; energy-based models; soft Bellman recursion; stochastic control; exogenous chance; reference measures; decision support.

1 Introduction

Next-token prediction and sequential decision making share a prefix structure. A current state summarizes a history, and each new symbol extends it. Recent work makes this connection exact: in function space, an ARM can represent an EBM, and the required future correction is a soft Bellman recursion [Blondel et al., 2026]. This is a representational statement, not a license to treat every fitted sequence model as an optimal policy.

The distinction becomes structural when trajectories alternate between choices and events. A person chooses an action, after which market conditions, health shocks, institutional changes, or other circumstances occur without being chosen. A sequence model may accurately fit the joint law of actions and events while answering the wrong question when its generation probabilities are interpreted as action quality. In particular, the standard EBM-to-ARM factorization exponentially tilts every next token. Doing so at an exogenous chance node acts as if favorable random transitions could be selected. Related optimism in probabilistic control is known [Levine, 2018]; what is missing is an exact account of how this mismatch appears and propagates inside the ARM–EBM construction.

We formalize an alternating action–event process and distinguish three objects:

1. the observational law $\rho_\beta = \pi_\beta \times \mu$, learned by teacher forcing on supported prefixes;
2. a reward-tilted *generative* law, whose exact autoregressive factorization uses log-moments at both action and event positions; and
3. a KL-regularized *control* solution, which soft-maximizes controllable actions but averages exogenous events under the environment kernel μ .

The paper makes four principal technical contributions. First, we preserve exact ARM–EBM factorization under a reference measure that contains both an action reference π_{ref} and the event kernel μ . Second, we derive a local-to-global recursion for the generative–control value gap. The recursion isolates an entropic premium created at a current chance node and a term propagated from future nodes. It also shows why transition entropy alone does not determine the gap: dispersion of continuation return is the relevant object. Third, we derive two audits from the same formalism. A relative-energy identity separates event typicality from action typicality, while a pathwise identity decomposes realized return into controllable soft advantages and mean-zero event innovations. Fourth, we give a finite-horizon uniform simulation bound and a support-gated regret corollary that separates estimation error from the opportunity cost of excluding unsupported actions.

The experiments are deliberately aligned with the claims. LifeGrid is a finite, fully enumerable world with $K = 4$ rounds, four actions, three events, and 20,736 complete trajectories. Exhaustive tests verify each equality to floating-point tolerance. A pre-registered 2×2 design changes transition entropy without changing reward dispersion, and changes event-dependent continuation dispersion without changing transition logits. Tabular and Transformer estimators then test the distinction between learning an observational kernel and choosing a readout. Synthetic evidence is not presented as evidence about human lives; it tests implementation, identification under stated assumptions, and qualitative predictions of the mathematics.

Scope of novelty. We do not claim that stochastic control-as-inference optimism, entropic certainty equivalents, general ARM–EBM equivalence, or martingale reward decompositions are new. The narrower contribution is their joint localization in typed action–event sequences: an exact gap-propagation law, a reference-measure interpretation audit, and a support-aware route from a learned sequence law to a control readout.

Road map. [Section 2](#) defines the process and observational semantics. [Section 3](#) develops generative and control recursions. [Section 4](#) gives the audit and attribution identities, and [Section 5](#) gives error and regret bounds. [Section 6](#) reports exact, tabular, and neural experiments.

2 Setup and observational semantics

2.1 Typed alternating trajectories

A trajectory of K rounds is

$$\xi = (a_1, e_1, \dots, a_K, e_K) \in \Xi.$$

At decision state s_t , an agent chooses $a_t \in \mathcal{A}(s_t)$ and reaches chance state $s'_t = s_t \oplus a_t$. Nature then draws $e_t \in \mathcal{E}(s'_t)$ from $\mu(\cdot \mid s'_t)$, producing $s_{t+1} = s'_t \oplus e_t$. Action and event tokens form a tagged disjoint union, so their roles cannot be confused by vocabulary overlap. Rewards may occur at both positions:

$$R(\xi) = \sum_{t=1}^K \{r(s_t, a_t) + r(s'_t, e_t)\}.$$

The state can be a full history, as in our finite experiments, or a learned representation. A terminal state has value zero.

For a finite reference distribution b and score f , define

$$\text{LSE}_{b,\tau}(f) = \tau \log \mathbb{E}_{Z \sim b} \exp\{f(Z)/\tau\}, \quad \tau > 0.$$

The temperature τ controls softness relative to a reference; it is not automatically a user’s risk preference.

2.2 Observational model

Let π_β be the behavior policy that generated data. The observational trajectory law is

$$\rho_\beta(\xi \mid x) = \prod_{t=1}^K \pi_\beta(a_t \mid s_t) \mu(e_t \mid s'_t). \quad (1)$$

Expected teacher-forcing negative log-likelihood is a sum of occupancy-weighted conditional cross-entropies. A function-space optimum therefore matches π_β at reachable action prefixes and μ at reachable event prefixes. This identifies a predictive factorization of (1); it neither changes observed behavior into a normative policy nor supplies causal identification by itself.

Assumption 1 (Policy-invariant event kernel). For every evaluation policy under consideration and every supported state–action pair, the conditional event law is the same kernel $\mu(e \mid s \oplus a)$.

Assumption 1 requires a sufficient state representation, consistency, and overlap. If it fails, μ is only a behavior-conditioned predictive law. Planning outputs must then be labeled scenario projections, not intervention effects.

2.3 Reference policy and objective

Let $\pi_{\text{ref}}(a \mid s) > 0$ be a full-support action reference and define

$$B_{\text{ref}} = \sup_s \log \frac{1}{\min_{a \in \mathcal{A}(s)} \pi_{\text{ref}}(a \mid s)} < \infty.$$

For $M = (\mu, r)$, the finite-horizon KL-regularized control objective is

$$\begin{aligned} J_M^\tau(\pi; x) = & \mathbb{E}_{\pi \times \mu} \left[\sum_{t=1}^K \{r(S_t, A_t) + r(S'_t, E_t)\} \right] \\ & - \tau \mathbb{E}_{\pi \times \mu} \left[\sum_{t=1}^K D_{\text{KL}}(\pi(\cdot \mid S_t) \parallel \pi_{\text{ref}}(\cdot \mid S_t)) \right]. \end{aligned} \quad (2)$$

A uniform π_{ref} recovers entropy regularization up to a fixed additive constant. A non-uniform reference can encode a default or behavior prior, but it must be reported because values and policies are relative to it.

3 Generative and control readouts

3.1 Exact reference-measure factorization

Define the trajectory base measure and reward-tilted density

$$m_{\text{ref}}(\xi \mid x) = \prod_{t=1}^K \pi_{\text{ref}}(a_t \mid s_t) \mu(e_t \mid s'_t), \quad (3)$$

$$p_{\text{gen}}^\tau(\xi \mid x) = \frac{m_{\text{ref}}(\xi \mid x) \exp\{R(\xi)/\tau\}}{Z_\tau(x)}. \quad (4)$$

Proposition 2 (Generative factorization). *With terminal value zero, define*

$$W_{\text{gen}}^\tau(s') = \tau \log \mathbb{E}_{e \sim \mu(\cdot \mid s')} \exp \left\{ \frac{r(s', e) + V_{\text{gen}}^\tau(s' \oplus e)}{\tau} \right\}, \quad (5)$$

$$Q_{\text{gen}}^\tau(s, a) = r(s, a) + W_{\text{gen}}^\tau(s \oplus a), \quad (6)$$

$$V_{\text{gen}}^\tau(s) = \tau \log \mathbb{E}_{a \sim \pi_{\text{ref}}(\cdot \mid s)} \exp\{Q_{\text{gen}}^\tau(s, a)/\tau\}. \quad (7)$$

Then (4) has exact conditionals

$$p_{\text{gen}}^\tau(a \mid s) = \pi_{\text{ref}}(a \mid s) \exp\{(Q_{\text{gen}}^\tau(s, a) - V_{\text{gen}}^\tau(s))/\tau\}, \quad (8)$$

$$p_{\text{gen}}^\tau(e \mid s') = \mu(e \mid s') \exp\{(r(s', e) + V_{\text{gen}}^\tau(s' \oplus e) - W_{\text{gen}}^\tau(s'))/\tau\}. \quad (9)$$

Their chain product equals $p_{\text{gen}}^\tau(\xi \mid x)$.

The proof is a telescoping cancellation of adjacent value terms (Appendix A). The proposition extends reference-measure ARM–EBM factorization to typed action–event sequences. It is a *generative* recursion: both decisions and events are exponentially tilted toward high reward.

3.2 Control preserves exogenous chance

The optimum of (2) instead satisfies

$$W_{\text{ctrl}}(s') = \mathbb{E}_{e \sim \mu(\cdot | s')} [r(s', e) + V_{\text{ctrl}}(s' \oplus e)], \quad (10)$$

$$Q_{\text{ctrl}}(s, a) = r(s, a) + W_{\text{ctrl}}(s \oplus a), \quad (11)$$

$$V_{\text{ctrl}}(s) = \tau \log \mathbb{E}_{a \sim \pi_{\text{ref}}(\cdot | s)} \exp\{Q_{\text{ctrl}}(s, a)/\tau\}, \quad (12)$$

$$\pi_{\text{ctrl}}^*(a | s) = \pi_{\text{ref}}(a | s) \exp\{(Q_{\text{ctrl}}(s, a) - V_{\text{ctrl}}(s))/\tau\}. \quad (13)$$

The action operator is the same as in generation; the chance operator is not. Generation uses an exponential log-moment in (5), whereas control uses the fixed-kernel expectation in (10).

For a continuation function U , define the local chance-readout gap

$$g_\tau(s'; U) = \tau \log \mathbb{E}_\mu \exp\{[r(s', E) + U(s' \oplus E)]/\tau\} - \mathbb{E}_\mu[r(s', E) + U(s' \oplus E)]. \quad (14)$$

Proposition 3 (Local-to-global gap recursion). *Let $\Delta V = V_{\text{gen}}^\tau - V_{\text{ctrl}}$ and $\Delta W = W_{\text{gen}}^\tau - W_{\text{ctrl}}$. Then*

$$\Delta W(s') = g_\tau(s'; V_{\text{gen}}^\tau) + \mathbb{E}_{e \sim \mu(\cdot | s')} \Delta V(s' \oplus e), \quad (15)$$

$$\Delta V(s) = \tau \log \mathbb{E}_{a \sim \pi_{\text{ctrl}}^*(\cdot | s)} \exp\{\Delta W(s \oplus a)/\tau\}. \quad (16)$$

Consequently $\Delta V, \Delta W \geq 0$. If $\bar{g}_j$ is the largest local gap at chance layer j , then $\Delta V(s_t) \leq \sum_{j=t}^K \bar{g}_j$. The two readouts agree exactly when every reachable chance return $r(s', E) + V_{\text{gen}}^\tau(s' \oplus E)$ is μ -almost surely constant.

Equation (15) separates optimism newly created at the current event from gap inherited from future events. By Jensen's inequality $g_\tau \geq 0$. For centered $Z = r(s', E) + U(s' \oplus E)$ with small fluctuation,

$$g_\tau(s'; U) = \frac{\text{Var}_\mu(Z)}{2\tau} + O\left(\frac{\mathbb{E}_\mu|Z - \mathbb{E}Z|^3}{\tau^2}\right). \quad (17)$$

Thus event entropy is not the causal variable. A high-entropy event law gives zero gap if all events have the same continuation value; a low-entropy law can give a large gap when a rare event has an extreme return.

Remark 4 (Known optimism, new localization). Naive inference-based control under stochastic dynamics can act as though favorable outcomes were controllable [Levine, 2018]. Proposition 3 does not rename that phenomenon. It localizes the mismatch in the ARM-EBM readout and gives its exact multi-round propagation.

4 Reference-measure audit and pathwise attribution

4.1 What a reconstructed energy measures

Let $b(z | s)$ equal π_{ref} at a decision node and μ at a chance node. Suppose a full-support conditional $\hat{p}$ is represented by local score q_b :

$$\hat{p}(z | s) = b(z | s) \exp\{[q_b(s, z) - V_{q_b}(s)]/\tau\}, \quad V_{q_b}(s) = \tau \log \mathbb{E}_{z \sim b} \exp\{q_b(s, z)/\tau\}. \quad (18)$$

Define the inverse edge score $r_b(s, z) = q_b(s, z) - V_{q_b}(s \oplus z)$.

Proposition 5 (Relative-energy audit). *For every complete trajectory,*

$$\sum_{(s, z) \in \xi} r_b(s, z) = \tau \log \frac{\hat{\rho}(\xi | x)}{m_{\text{ref}}(\xi | x)} + V_{q_b}(x). \quad (19)$$

For the observational law $\rho_\beta = \pi_\beta \times \mu$, event terms cancel under the μ -aware base:

$$\tau \log \frac{\rho_\beta(\xi | x)}{m_{\text{ref}}(\xi | x)} = \tau \sum_{t=1}^K \log \frac{\pi_\beta(a_t | s_t)}{\pi_{\text{ref}}(a_t | s_t)}. \quad (20)$$

Under a counting base, the reconstructed energy instead contains both $\sum_t \log \pi_\beta(a_t | s_t)$ and $\sum_t \log \mu(e_t | s'_t)$.

Both bases are mathematically legitimate, but they answer different questions. Counting-base energy measures joint trajectory typicality. Relative energy under (3) removes event typicality when interpreting action deviations from π_{ref} . Conflation arises only when joint likelihood is labeled merit, choice quality, or intentionality. Typicality is also distinct from favorability: $\log \mu(e | s')$ asks whether an event was expected, not whether it helped.

4.2 Choice and event components

Using the control readout, define the soft action advantage and event innovation

$$A_{\text{ctrl}}(s, a) = Q_{\text{ctrl}}(s, a) - V_{\text{ctrl}}(s) = \tau \log \frac{\pi_{\text{ctrl}}^*(a | s)}{\pi_{\text{ref}}(a | s)}, \quad (21)$$

$$I_{\text{event}}(s', e) = r(s', e) + V_{\text{ctrl}}(s' \oplus e) - W_{\text{ctrl}}(s'). \quad (22)$$

By construction, $\mathbb{E}_{e \sim \mu} I_{\text{event}}(s', e) = 0$. Other action-impact baselines may be useful: for any declared $\nu(\cdot | s)$,

$$I'_{\text{action}}(s, a) = Q_{\text{ctrl}}(s, a) - \mathbb{E}_{a' \sim \nu(\cdot | s)} Q_{\text{ctrl}}(s, a'). \quad (23)$$

The baseline must be reported. Using π_β describes deviation from comparable observed behavior; using a user's default policy describes deviation from that user's tendency.

Theorem 6 (Pathwise choice–event decomposition). *For any realized trajectory ξ and terminal $V_{\text{ctrl}}(s_{K+1}) = 0$,*

$$R(\xi) - V_{\text{ctrl}}(s_1) = \sum_{t=1}^K A_{\text{ctrl}}(s_t, a_t) + \sum_{t=1}^K I_{\text{event}}(s'_t, e_t). \quad (24)$$

The identity holds regardless of the policy that produced the realized actions.

The first sum records the realized sequence of controllable deviations from the soft baseline; the second records favorable and unfavorable chance innovations. The theorem is algebraic and predictive. A causal reading of either component still requires Assumption 1 and the usual longitudinal identification conditions.

4.3 A policy-level regret identity

For any full-support Markov policy π , let d_t^π be its state occupancy under μ . The KL-regularized performance difference specializes to

$$J_M^\tau(\pi_{\text{ctrl}}^*; x) - J_M^\tau(\pi; x) = \tau \sum_{t=1}^K \mathbb{E}_{S_t \sim d_t^\pi} D_{\text{KL}}(\pi(\cdot | S_t) \| \pi_{\text{ctrl}}^*(\cdot | S_t)). \quad (25)$$

This identity supplies an exact implementation test and connects local action divergence to global regret.

5 Model error, global regret, and support gating

Let $M = (\mu, r)$ and $\hat{M} = (\hat{\mu}, \hat{r})$ share action sets, π_{ref} , and τ . Rewards at action and event positions are bounded by R_{max} in absolute value. Suppose

$$|\hat{r} - r| \leq \epsilon_r, \quad D_{\text{TV}}(\hat{\mu}(\cdot | s'), \mu(\cdot | s')) \leq \epsilon_\mu$$

uniformly on the relevant region. Let $V_{M,h}^\pi$ be the KL-regularized value of policy π with h rounds remaining.

Theorem 7 (Uniform simulation and policy regret). *For every Markov policy π and $h \leq K$,*

$$\|V_{M,h}^\pi - V_{\hat{M},h}^\pi\|_\infty \leq C_h, \quad (26)$$

$$C_h = 2h\epsilon_r + \epsilon_\mu \sum_{j=1}^h [2R_{\text{max}} + (j-1)(4R_{\text{max}} + \tau B_{\text{ref}})]. \quad (27)$$

The same bound holds between optimal values. If $\hat{\pi}$ is η -optimal in $\hat{M}$, then

$$J_M^\tau(\pi_M^*; x) - J_M^\tau(\hat{\pi}; x) \leq 2C_K + \eta. \quad (28)$$

For a uniform reference over at most A_{max} actions, the order is $O(K\epsilon_r + K^2(R_{\text{max}} + \tau \log A_{\text{max}})\epsilon_\mu + \eta)$.

The theorem controls a complete rollout, not merely one-step Q error. It is intentionally conservative: the uniform maximum error prevents a new policy from exploiting a rare but poorly estimated state. Weighted prediction error may be operationally useful, but it cannot replace ϵ_μ in (27).

5.1 Gating and abstention

Let $G(s) \subseteq \mathcal{A}(s)$ be a support gate and Π_G the policies that put no mass outside $G(s)$. Let π_G^* optimize the true model over Π_G and $\hat{\pi}_G$ be η -optimal for the learned model over Π_G .

Corollary 8 (Support-gated regret). *If the assumptions of Theorem 7 hold on all nodes reachable by Π_G , then*

$$\begin{aligned} J_M^\tau(\pi_M^*; x) - J_M^\tau(\hat{\pi}_G; x) &\leq \underbrace{J_M^\tau(\pi_M^*; x) - J_M^\tau(\pi_G^*; x)}_{\text{coverage-restriction cost}} \\ &\quad + \underbrace{2C_K(G) + \eta}_{\text{within-gate estimation/planning cost}}. \end{aligned} \quad (29)$$

Abstention is therefore not free. Increasing a count threshold can lower maximum model error but may exclude valuable actions or leave no supported route to parts of the tree. A deployed system should report coverage and error together, and should switch from ranking to scenario exploration when the gate is not credible.

Proof strategy. A backward induction bounds every fixed policy uniformly. The optimal-value bound follows by taking suprema, and (28) follows from a model-comparison sandwich. Corollary 8 inserts and subtracts the true gated optimum. Complete proofs are in Appendix A.

6 Experiments

6.1 Protocol and evidential scope

Before the full run, we froze the claim hierarchy, factorial interventions, sample sizes, seed counts, support thresholds, architecture, metrics, and rejection rules. The code and protocol are included in the supplementary package. All experiments are synthetic validation of the formal objects; no result is treated as causal evidence about human outcomes.

LifeGrid has $K = 4$ rounds, $|\mathcal{A}| = 4$, $|\mathcal{E}| = 3$, and $(4 \cdot 3)^4 = 20,736$ complete trajectories. The exact state is the full typed history. Rewards and transition logits are generated from an environment seed, π_{ref} is uniform, and $\tau = 1$. The behavior policy has full support. Reusing a seed across factorial cells reuses every base table.

The factorial intervention has two levels of transition entropy and two levels of event-dependent continuation dispersion. Entropy changes only the transition-logit temperature (0.35 versus 2.50); dispersion changes only the scale of event and continuation features (0.15 versus 1.10). Hence the high/low entropy worlds have different μ but the same reward table within a dispersion condition, while the high/low dispersion worlds have identical observational token laws but different returns.

6.2 E1: exhaustive identity checks

We enumerate every trajectory for 20 independent worlds and all four factorial cells. Tests cover normalization and factorization in Proposition 2, both lines of Proposition 3, the two reference bases in Proposition 5, Theorem 6, the conditional mean of event innovations, and the KL-regret identity (25). The pre-specified rejection threshold was 10^{-10} in float64. The largest absolute residual was $1.42\text{e-}14$, so no identity test was rejected (Figure 1). These computations verify the implementation of exact equalities; they are not empirical proofs of the propositions.

6.3 E2: entropy and continuation dispersion

Figure 2 shows the pre-specified 2×2 comparison. The transition intervention changes measured event entropy while preserving it across dispersion levels, confirming that the factors are mechanically separated. With low return dispersion, the mean root gap is only 0.009 and 0.017 in the low- and high-entropy cells. Raising return dispersion increases the gap to 0.440 and 0.896, respectively. The true control regret of deploying the generative action policy is 0.047 and 0.056 in the two high-dispersion cells.

The result supports the proposition’s qualitative prediction: entropy alone is insufficient, while dispersion of chance continuation values creates the mismatch. It also shows interaction rather than pure additivity: a more

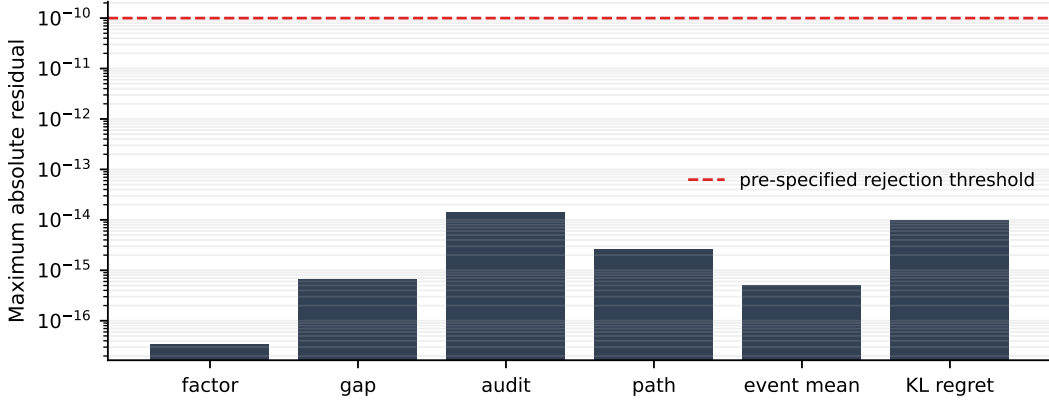


Figure 1: Maximum absolute residual over exhaustive identity tests. The dashed line is the rejection threshold frozen before the run.

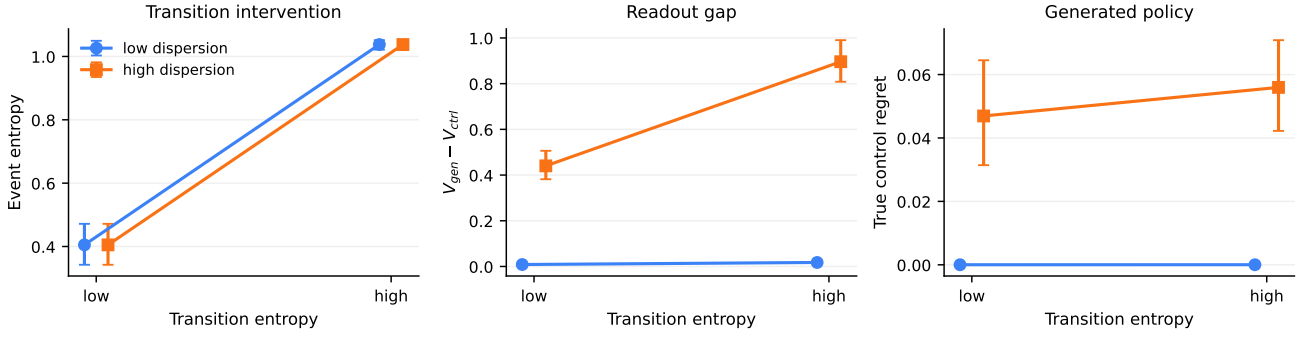


Figure 2: Pre-specified factorial experiment. Points are means and bars are 95% bootstrap intervals over environment seeds. Left: the entropy manipulation is invariant to return dispersion. Middle: the exact root readout gap. Right: true regret of the generative action policy evaluated with the control objective.

diffuse μ exposes more of the return dispersion. This is consistent with the exact object g_τ , which depends on the full conditional distribution of return, not on either scalar factor alone.

6.4 E3: finite tabular estimation and support

For ten environment seeds and five data seeds, we sample $N \in \{100, 300, 1000, 3000, 10000, 30000\}$ observational trajectories in the high-entropy/high-dispersion cell. Symmetric Dirichlet smoothing ($\alpha = 0.5$) estimates conditionals, and conditional reward means are learned from observations with Gaussian noise standard deviation 0.10. We report exact trajectory KL, occupancy-weighted TV, true policy regret, and the maximum TV over each declared supported region.

From $N = 100$ to $N = 30,000$, mean exact trajectory KL decreases from 0.790 to 0.179, while occupancy-weighted event TV decreases from 0.147 to 0.060 (Figure 3). Control-value error falls from 0.195 to 0.036. In contrast, generative-value error moves from 0.347 toward 0.893, nearly the exact high/high structural gap 0.896. Better estimation therefore does not make the generative readout a control readout; it makes the model converge more accurately to the different generative objective. These averages assess predictive learning but cannot certify Theorem 7; the theorem requires a maximum error on the entire deployment region.

A visit-count gate exposes the missing trade-off (Figure 4). At $N = 30,000$, the supported chance mass changes from 0.989 at the lowest threshold to 0.704 at the highest. Stronger gating can reduce the maximum error among retained nodes but raises the coverage-restriction cost; in sparse full-history states, that cost can dominate. The computed worst-case bound remains conservative and often vacuous, which is itself an important negative result: a theorem whose premises require full-tree uniform accuracy does not become an operational certificate merely because average prediction is good.

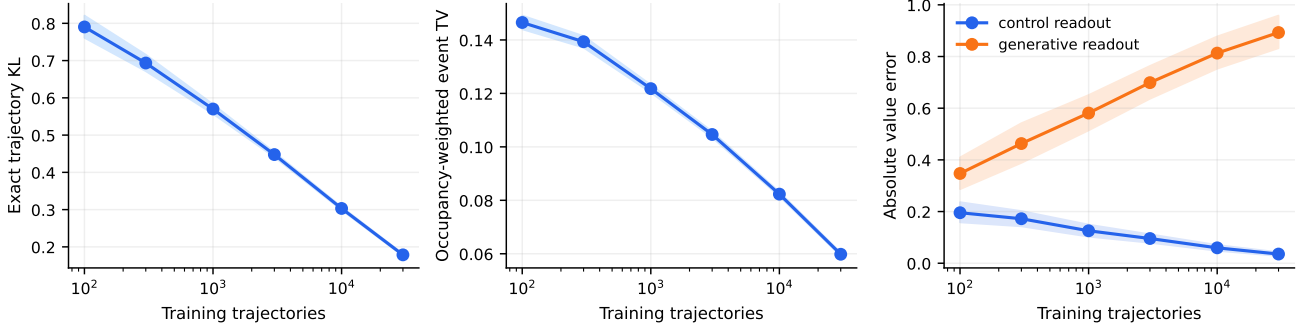


Figure 3: Tabular learning curves. Lines are means and bands are 95% bootstrap intervals over environment/data seeds. KL is evaluated exactly by trajectory enumeration. The right panel shows structural rather than estimation error: as $\hat{\mu}$ improves, the generative readout converges toward the wrong chance operator for the control target.

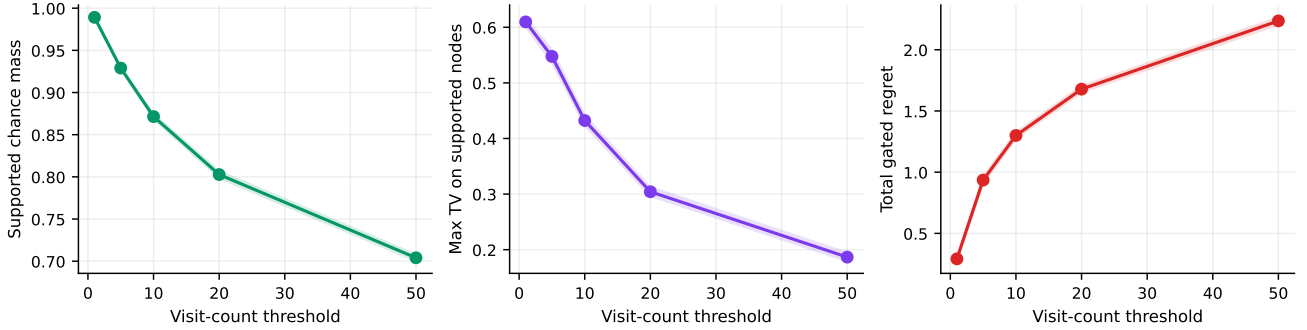


Figure 4: Coverage-error-regret trade-off at $N = 30,000$. A count gate improves the declared error region by excluding data-poor nodes, but exclusion can create substantial opportunity cost.

6.5 E4: finite causal Transformer and two readouts

We train a two-layer causal Transformer (width 32, four heads, feed-forward width 128) by teacher forcing on 30,000 sequences for 1,200 AdamW updates. Typed-token masking permits only actions at action positions and events at event positions. There are five environment seeds and five training seeds per entropy condition. Because dispersion changes only rewards, the same trained observational model is reused for the two dispersion conditions. Evaluation enumerates every valid prefix and every complete trajectory; it is not based on sampled test rollouts.

The learned model reaches mean exact trajectory KL 0.025 and 0.022 in the low- and high-entropy conditions. We then use the same estimated $\hat{\mu}$ in two planners: the chance-corrected control recursion and the naive generative recursion. In the high-entropy/high-dispersion cell, mean control-value error is 0.028 versus 0.968 for the generative readout; true policy regret is 0.003 versus 0.060 (Figure 5). The comparison isolates the readout: both planners receive the same finite model and reward.

The neural experiment does not show that teacher forcing solves control. It shows the more limited and useful statement: a teacher-forced model can estimate observational conditionals, after which a structurally correct readout can materially reduce decision error. Approximation and optimization error remain visible in nonzero KL and maximum prefix TV.

6.6 Summary of claim support

7 Related work

ARM-EBM equivalence. Blondel et al. [2026] establish an explicit function-space bijection, connect it to the soft Bellman equation, and analyze EBM-to-ARM distillation. We retain their reference-measure logic but type the sequence into actions and events. This reveals that the factorization that is correct for generation is generally

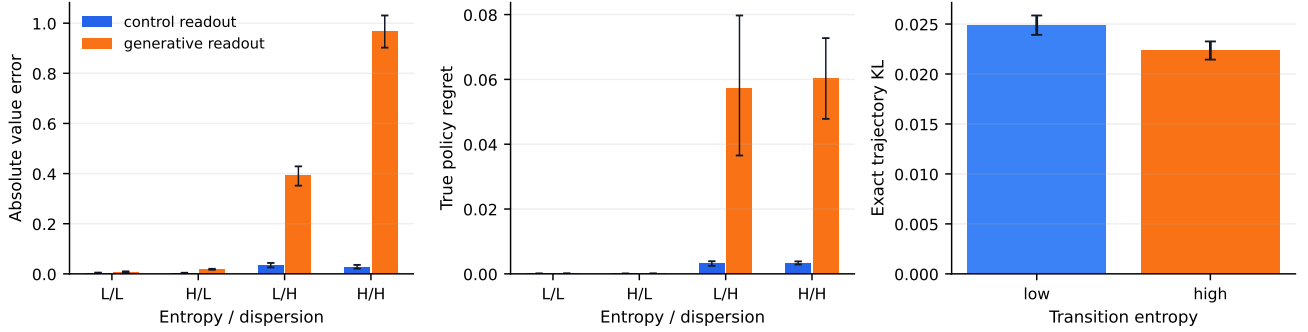


Figure 5: Finite Transformer results. L/L, H/L, L/H, and H/H denote transition entropy / return dispersion. Bars are means and error bars are 95% bootstrap intervals. The right panel reports exact trajectory KL of the observational ARM.

Table 1: What each experiment can and cannot establish.

Experiment	Supported conclusion	Not established
Exact enumeration	Code satisfies stated equalities to tolerance	Generality of the proofs (established analytically)
Factorial intervention	Gap tracks chance-return dispersion; entropy modulates exposure	A universal scalar law in entropy or variance
Tabular estimation	Prediction improves; coverage and uniform error conflict	A non-vacuous certificate for sparse full histories
Transformer	Same learned model benefits from a control readout	Causal identification, human utility, real-world safety

not the operator required for control.

Regularized control and control as inference. Maximum-entropy inverse RL [Ziebart et al., 2008], energy-based policies [Haarnoja et al., 2017], and the general theory of regularized MDPs [Geist et al., 2019] connect exponential tilting, convex regularization, and Bellman operators. Levine [2018] explains why naive inference can be optimistic under stochastic dynamics, and Arriegas et al. [2023] gives a Bayesian-inference construction consistent with the fixed-dynamics constraint. Our gap recursion complements these results by separating local entropic optimism and its downstream propagation inside an alternating sequence factorization. Stochastic GFlowNets likewise isolate environmental stochasticity from agent choices [Pan et al., 2023].

Sequence models for decisions. Decision Transformer conditions sequence generation on return [Chen et al., 2021], while Trajectory Transformer performs model-based sequence search [Janner et al., 2021]; world-model approaches learn dynamics before controller optimization [Ha and Schmidhuber, 2018]. We focus on a different boundary: teacher forcing remains an estimator of the observational law, and the normative change is made explicitly in the readout operator.

Risk sensitivity and stochastic attribution. Entropic utility and risk-sensitive MDPs are classical [Howard and Matheson, 1972, Bäuerle and Rieder, 2011]. Modern analyses quantify the statistical cost of exponential-utility objectives [Fei et al., 2021]. Martingale-based decompositions have also been used to separate reward uncertainty [Vadori et al., 2020]. Our g_τ is a diagnostic of generative/control mismatch, not a proposal that users should be risk-seeking. The pathwise innovation identity is specialized to typed action–event sequences and the same reference-policy control value.

Offline support and causal interpretation. Off-policy evaluation quantifies policy-shift error [Precup et al., 2000, Jiang and Li, 2016, Thomas et al., 2015]; support-constrained offline RL trades extrapolation control against conservatism [Mao et al., 2023, Kim and Oh, 2023]. Corollary 8 makes this trade-off explicit for the present

finite-horizon model comparison. Moving from conditional $P(e \mid s, a)$ to intervention claims requires the assumptions of longitudinal causal inference [Pearl, 2009, Hernán and Robins, 2020].

8 Limitations and responsible scope

Predictive is not causal. The learned kernel is observational. A $\text{do}(a)$ interpretation requires sufficient state, consistency, positivity, and control of time-varying confounding. If hidden variables jointly affect action and event, the planning semantics fail even when prediction is calibrated.

Finite full-history worlds are both useful and pathological. Exhaustive enumeration gives unusually strong implementation checks. It also creates a rapidly expanding state space in which count gates lose coverage. Learned state abstractions are necessary for scale, but abstraction error then becomes a new assumption rather than disappearing.

Function space versus finite models. Proposition 2 is exact in function space. The Transformer experiment adds representation, approximation, optimization, and calibration error. Our results show a readout effect in one controlled architecture, not universal dominance across model classes.

Reward specification. The scalar reward is declared, not discovered as a universal good. A practical system needs versioned user preferences, multiple value axes, sensitivity analysis, and the ability to reject a scalarization. The present paper analyzes planning conditional on r .

Bounds are conservative. Theorem 7 is a uniform worst-case guarantee. It can be vacuous in sparse spaces; the experiment reports that failure rather than replacing maximum error with a weighted average. Tighter occupancy-aware guarantees would require restrictions on policy shift or concentrability.

High-stakes deployment. Medical, legal, financial, employment, and family decisions require domain expertise and human approval. The formalism can support structured scenario comparison; it does not justify autonomous prescription. Longitudinal personal data also require minimization, granular consent, export, deletion, and security controls outside the scope of this paper.

9 Conclusion

An autoregressive model can represent a globally reward-tilted trajectory distribution without thereby solving a stochastic control problem. In typed action–event sequences, the difference is one operator at chance nodes: exponential tilting for generation versus expectation under the environment law for control. That small formal change has three consequences. It yields an exact recursion for how chance optimism is created and propagated; it clarifies when event typicality contaminates an interpretation of sequence energy; and it supports a pathwise separation of controllable advantages from exogenous innovations.

The experiments follow the logical order of the paper. Exhaustive code tests validate the equalities, an independent factorial design identifies the relevant return-dispersion mechanism, finite estimators reveal coverage limits, and a causal Transformer demonstrates that an observational model and a decision readout can be trained and evaluated as separate components. The operational rule is therefore precise: learn what actions and events occurred, preserve exogenous chance under its supported conditional law, optimize only controllable choices relative to an explicit reference, and abstain when the model cannot support the comparison.

References

Argenis Arriojas, Jacob Adamczyk, Stas Tiomkin, and Rahul V. Kulkarni. Bayesian inference approach for entropy regularized reinforcement learning with stochastic dynamics. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 99–109, 2023.

- Nicole Bäuerle and Ulrich Rieder. *Markov Decision Processes with Applications to Finance*. Springer, 2011.
- Mathieu Blondel, Michaël E. Sander, Germain Vivier-Ardisson, Tianlin Liu, and Vincent Roulet. Autoregressive language models are secretly energy-based models: Insights into the lookahead capabilities of next-token prediction. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. arXiv:2512.15605.
- Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In *Advances in Neural Information Processing Systems*, volume 34, 2021.
- Yingjie Fei, Zhuoran Yang, Yudong Chen, Zhaoran Wang, and Qiaomin Xie. Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. In *Advances in Neural Information Processing Systems*, volume 34, 2021.
- Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2160–2169, 2019.
- David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31, 2018.
- Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1352–1361, 2017.
- Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.
- Ronald A. Howard and James E. Matheson. Risk-sensitive markov decision processes. *Management Science*, 18(7): 356–369, 1972.
- Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. In *Advances in Neural Information Processing Systems*, volume 34, 2021.
- Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.
- Byeongchan Kim and Min-Hwan Oh. Model-based offline reinforcement learning with count-based conservatism. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 16728–16746, 2023.
- Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.
- Yixiu Mao, Hongchang Zhang, Chen Chen, Yi Xu, and Xiangyang Ji. Supported trust region optimization for offline reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 23829–23851, 2023.
- Ling Pan, Dinghuai Zhang, Moksh Jain, Longbo Huang, and Yoshua Bengio. Stochastic generative flow networks. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 1628–1638, 2023.
- Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.
- Doina Precup, Richard S. Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, 2000.
- Philip S. Thomas, Georgios Theodorou, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015.
- Nelson Vadori, Sumitra Ganesh, Prashant Reddy, and Manuela Veloso. Risk-sensitive reinforcement learning: A martingale approach to reward uncertainty. *arXiv preprint arXiv:2006.12686*, 2020.
- Brian D. Ziebart, Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2008.

A Proofs

A.1 Proof of Proposition 2

At a decision state, (8) is normalized because its denominator is $\exp\{V_{\text{gen}}^\tau(s)/\tau\}$; at a chance state, (9) is normalized because its denominator is $\exp\{W_{\text{gen}}^\tau(s')/\tau\}$. Multiplying the two conditionals for round t gives

$$\pi_{\text{ref}}(a_t | s_t) \mu(e_t | s'_t) \exp \left\{ \frac{r(s_t, a_t) + r(s'_t, e_t) + V_{\text{gen}}^\tau(s_{t+1}) - V_{\text{gen}}^\tau(s_t)}{\tau} \right\}.$$

Over K rounds the value terms telescope. With terminal value zero, the chain product is

$$m_{\text{ref}}(\xi | x) \exp\{[R(\xi) - V_{\text{gen}}^\tau(x)]/\tau\}.$$

Summing over ξ shows that $Z_\tau(x) = \exp\{V_{\text{gen}}^\tau(x)/\tau\}$, which proves the factorization.

For the teacher-forcing statement, expand expected NLL by position and condition on the reached prefix. Every term is a nonnegative occupancy weight times a conditional cross-entropy. On a positive-occupancy prefix, the cross-entropy is minimized exactly by the data conditional. Applying this separately to the tagged action and event positions of (1) yields π_β and μ on reachable prefixes. $\square$

A.2 Proof of Proposition 3

At a chance state set $Z_{\text{gen}}(e) = r(s', e) + V_{\text{gen}}^\tau(s' \oplus e)$. Since $V_{\text{ctrl}} = V_{\text{gen}}^\tau - \Delta V$,

$$W_{\text{ctrl}}(s') = \mathbb{E}_\mu[Z_{\text{gen}}(E) - \Delta V(s' \oplus E)].$$

Subtracting this expression from (5) gives (15). At a decision state, $Q_{\text{gen}}^\tau(s, a) = Q_{\text{ctrl}}(s, a) + \Delta W(s \oplus a)$. Substitute (13) into the ratio of the two action partitions to obtain (16).

Jensen's inequality gives $g_\tau \geq 0$. Starting from zero terminal gap, (15) and (16) imply nonnegativity by backward induction. Also,

$$\mathbb{E}X \leq \max X, \quad \tau \log \mathbb{E} \exp(X/\tau) \leq \max X,$$

so the same induction gives the sum of layerwise maxima. Strict convexity of the exponential gives $g_\tau = 0$ exactly when its argument is constant μ -almost surely. Finally, (17) is the cumulant expansion of $\tau \log \mathbb{E} \exp\{(Z - \mathbb{E}Z)/\tau\}$. $\square$

A.3 Proof of Proposition 5

Rearranging (18),

$$q_b(s, z) = \tau \log \frac{\hat{p}(z | s)}{b(z | s)} + V_{q_b}(s).$$

Therefore

$$r_b(s, z) = \tau \log \frac{\hat{p}(z | s)}{b(z | s)} + V_{q_b}(s) - V_{q_b}(s \oplus z).$$

Summation along a complete trajectory telescopes the values and multiplies the conditional ratios, giving (19). Setting $\hat{\rho} = \rho_\beta$ and $b = \pi_{\text{ref}}$ at actions and $b = \mu$ at events cancels every event ratio and yields (20). Under a uniform counting base, neither factor cancels. $\square$

A.4 Proof of Theorem 6

At each round, (21) and (22) give

$$\begin{aligned} A_{\text{ctrl}}(s_t, a_t) + I_{\text{event}}(s'_t, e_t) &= Q_{\text{ctrl}}(s_t, a_t) - V_{\text{ctrl}}(s_t) \\ &\quad + r(s'_t, e_t) + V_{\text{ctrl}}(s_{t+1}) - W_{\text{ctrl}}(s'_t) \\ &= r(s_t, a_t) + r(s'_t, e_t) + V_{\text{ctrl}}(s_{t+1}) - V_{\text{ctrl}}(s_t). \end{aligned}$$

Summing over t telescopes the value terms and proves (24). Centering in (22) follows directly from (10). $\square$

A.5 Proof of the KL-regret identity (25)

The Gibbs variational identity applied at any decision state gives, for every distribution π_s ,

$$\begin{aligned} V_{\text{ctrl}}(s) &- \{\mathbb{E}_{a \sim \pi_s} Q_{\text{ctrl}}(s, a) - \tau D_{\text{KL}}(\pi_s \| \pi_{\text{ref}, s})\} \\ &= \tau D_{\text{KL}}(\pi_s \| \pi_{\text{ctrl}, s}^*). \end{aligned}$$

Take expectation under the state occupancy generated by π and μ , sum over time, and expand $Q_{\text{ctrl}} = r + \mathbb{E}_\mu[r + V_{\text{ctrl}}]$. Adjacent expected values telescope, leaving the difference between $V_{\text{ctrl}}(x) = J_M^\tau(\pi_{\text{ctrl}}^*; x)$ and $J_M^\tau(\pi; x)$. $\square$

A.6 Proof of Theorem 7

Fix a policy π and let $\delta_h = \sup_s |V_{M,h}^\pi(s) - V_{\hat{M},h}^\pi(s)|$. The statewise KL penalty is identical in the two models and cancels. At a chance node, add and subtract the μ -expectation of $f(e) = \hat{r}(s', e) + V_{\hat{M},h-1}^\pi(s' \oplus e)$. Using

$$|\mathbb{E}_P f - \mathbb{E}_Q f| \leq D_{\text{TV}}(P, Q) \text{span}(f)$$

and one action-position plus one event-position reward error gives

$$\delta_h \leq \delta_{h-1} + 2\epsilon_r + \epsilon_\mu \left[2R_{\max} + \text{span}(V_{\hat{M},h-1}^\pi) \right]. \quad (30)$$

For any policy, a one-round expected reward plus KL penalty lies in $[-2R_{\max} - \tau B_{\text{ref}}, 2R_{\max}]$. Expectations do not expand span, hence

$$\text{span}(V_{M,h}^\pi) \leq h(4R_{\max} + \tau B_{\text{ref}}).$$

Substitution in (30) and summation from $\delta_0 = 0$ give (27).

The bound is uniform in π , so $|J_M^\tau(\pi) - J_{\hat{M}}^\tau(\pi)| \leq C_K$. With π_M^* optimal in M and $\hat{\pi}$ η -optimal in $\hat{M}$,

$$\begin{aligned} J_M^\tau(\pi_M^*) - J_M^\tau(\hat{\pi}) &\leq [J_M^\tau(\pi_M^*) - J_{\hat{M}}^\tau(\pi_M^*)] + \eta \\ &\quad + [J_{\hat{M}}^\tau(\hat{\pi}) - J_M^\tau(\hat{\pi})] \\ &\leq 2C_K + \eta. \end{aligned}$$

This proves the theorem. $\square$

A.7 Proof of Corollary 8

Apply Theorem 7 after restricting the policy class to Π_G , where its assumptions hold. Add and subtract $J_M^\tau(\pi_G^*)$:

$$\begin{aligned} J_M^\tau(\pi_M^*) - J_M^\tau(\hat{\pi}_G) &= [J_M^\tau(\pi_M^*) - J_M^\tau(\pi_G^*)] \\ &\quad + [J_M^\tau(\pi_G^*) - J_M^\tau(\hat{\pi}_G)]. \end{aligned}$$

The first bracket is the coverage-restriction cost; the second is at most $2C_K(G) + \eta$. $\square$

B Experimental and implementation details

B.1 LifeGrid construction

For each environment seed, we draw fixed base tables for action reward, event reward, transition logits, and behavior logits. The previous event and current depth index the tables; the represented state remains the complete history. Event effects are centered across events before scaling. The four cells are then formed only by

$$T_\mu \in \{0.35, 2.50\}, \quad d_r \in \{0.15, 1.10\}.$$

The event law is $\text{softmax}(\ell/T_\mu)$, and d_r scales only event-dependent reward and continuation features. This construction guarantees that dispersion does not change the observational sequence distribution, while entropy does not change reward tables.

Chance-corrected readout algorithm. Given $\hat{\mu}, \hat{r}, \pi_{\text{ref}}, \tau$, set $\hat{V} = 0$ at terminal states and sweep backward from K to 1. At each s, a , compute

$$\hat{W}(s \oplus a) = \mathbb{E}_{e \sim \hat{\mu}}[\hat{r}(s \oplus a, e) + \hat{V}(s \oplus a \oplus e)], \quad \hat{Q}(s, a) = \hat{r}(s, a) + \hat{W}(s \oplus a).$$

Then set $\hat{V}(s) = \tau \log \mathbb{E}_{a \sim \pi_{\text{ref}}} \exp\{\hat{Q}(s, a)/\tau\}$ and $\hat{\pi}(a | s) = \pi_{\text{ref}}(a | s) \exp\{[\hat{Q}(s, a) - \hat{V}(s)]/\tau\}$. The naive generative readout changes only the first update, replacing the expectation with $\tau \log \mathbb{E}_{\hat{\mu}} \exp\{(\hat{r} + \hat{V})/\tau\}$.

B.2 Tabular estimator

The behavior and event conditionals use symmetric Dirichlet smoothing $\alpha = 0.5$. Reward estimates are conditional sample means with a zero prior. Independent zero-mean Gaussian noise with standard deviation 0.10 is added to each observed reward. A chance node is declared supported at threshold c if its state-action visit count is at least c , and the corresponding action is admitted at that state. If no action passes, the implementation retains the most visited action to keep the recursion defined. The reported raw supported mass counts only nodes that actually pass the threshold; the true value loss of the resulting fallback-completed policy class is reported as the coverage-restriction cost.

B.3 Transformer estimator

The vocabulary contains BOS, four typed action tokens, three typed event tokens, and PAD. Sequences contain eight predicted tokens. A parity mask prevents event tokens at action positions and action tokens at event positions. The network has two pre-norm Transformer encoder layers, width 32, four attention heads, GELU feed-forward width 128, learned positional embeddings, and no dropout. AdamW uses learning rate 3×10^{-3} , weight decay 10^{-4} , gradient clipping at 1, batch size 512, and 1,200 updates. All prefix conditionals are extracted exactly after training.

B.4 Metrics and uncertainty

Trajectory KL is computed by summing over all 20,736 paths. Conditional TV is reported both as observational-occupancy weighted error and as a maximum on declared supported nodes. Policy regret always evaluates the extracted policy in the true environment under (2). Figure intervals use 4,000 nonparametric bootstrap resamples with a fixed reporting seed. Reproducibility seeds are functions only of the environment, data, training replicate, and pre-declared sample size.

B.5 Software checks

The exact and tabular suites use NumPy float64. Neural training uses PyTorch on CPU; exact post-training evaluation returns to NumPy float64. Unit tests independently check each identity, the support-gated bound, and estimator normalization. The complete command is `bash scripts/run_full.sh`.

C Additional numerical summary

Table 2: Mean exact root gap and true regret of the generative policy across factorial cells.

Transition entropy	Return dispersion	$V_{\text{gen}} - V_{\text{ctrl}}$	Generative-policy regret
Low	Low	0.009	—
High	Low	0.017	—
Low	High	0.440	0.047
High	High	0.896	0.056

The omitted low-dispersion regret entries are visually near zero and are available in the machine-readable results. Reporting exact gaps and policy regret separately is important: a positive normalization/value gap need not materially change the action policy.