CaLM: Causal Language Modeling for Counterfactual World Simulations

Anonymous authors
Paper under double-blind review

Abstract

Counterfactual trajectory simulation asks how a sequence would have unfolded had selected actions been different. Autoregressive Transformers provide a powerful framework for sequence modeling by learning the conditional distribution of each token given the preceding tokens. However, fitting these conditional distributions does not ensure that simulations respect causal relationships or correctly account for unobserved confounding when evaluating interventions. We formalize counterfactual trajectory simulation in the language of structural causal models. To evaluate how reliably a simulator class answers counterfactual queries, we introduce *counterfactual coverage*, which holds when the true counterfactual value lies between the lower and upper bounds of the class's estimates. We show theoretically that next-token prediction alone does not guarantee counterfactual coverage for vanilla Transformers. We develop `CaLM` (Causal Language Modeling), a Transformer-based simulator that incorporates constraints specified by an arbitrary causal graph into its attention mechanism. We establish counterfactual coverage for the CaLM simulator class and demonstrate improved counterfactual prediction across chess, buyer-seller negotiation, autonomous driving, clinical trajectories, and poker.

1 Introduction

Autoregressive Transformers provide a general framework for modeling and generating sequences (Vaswani et al., 2017). With suitable sequence representations, this framework has enabled language modeling, text completion (Radford et al., 2019), image generation (Parmar et al., 2018), video generation (Yan et al., 2021), and sequential decision-making (Janner et al., 2021). In the latter setting, recent world simulators use this framework to simulate future interactions between agents and their environments (Micheli et al., 2022; Robine et al., 2023).

World simulators have achieved substantial success by learning statistical dependencies between the trajectory history and the next token, which may represent an action or a state (Janner et al., 2021; Seff et al., 2023; Wu et al., 2024). By fitting the distribution of the next token conditioned on that history, they can accurately answer observational questions such as *"Given the trajectory so far, what actions and outcomes are likely to follow?"* Meanwhile, most recent work on world modeling also emphasizes answering counterfactual (what if) questions as an important capability for planning (Xiang et al., 2025; Parker-Holder et al., 2025; Foss et al., 2025). In other words, these models aim to answer counterfactual questions such as *"How would the subsequent actions and outcomes have changed had an earlier action been different?"*. However, learning observational correlations alone does not guarantee causal predictions. Specifically, we show later that a sequence model trained with next-token prediction and a left-to-right attention masks ("*causal*" masks in literatures) may fit observational conditional distributions well yet generate counterfactual trajectories that contradict the true outcomes. We illustrate this challenge with the following example.

Example 1. Consider a blue ego vehicle (Ego) and an orange trailing vehicle (TV) traveling in the same lane, as shown in Fig. 1(a). At each frame, a driver's action, such as whether to apply the emergency brake, depends on the preceding traffic conditions and interactions with vehicles ahead. In this example, each driver's unobserved persona also influences their actions throughout the trajectory, creating a shared unobserved cause of that driver's actions across frames (Fig. 1(b)). Ego has a normal driving persona, whereas TV's driver is distracted and responds more slowly to changes in traffic.

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

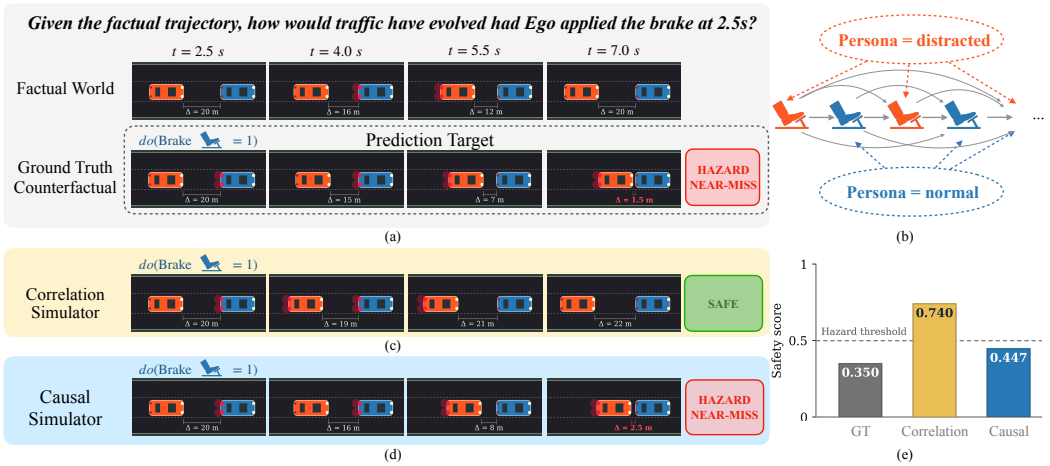

Figure 1: Had Ego (blue) braked at $t = 2.5\,\mathrm{s}$, how would traffic have evolved? (a) Factual and ground-truth counterfactual; the distracted trailing driver reacts late, causing a near miss. (b) Each driver's unobserved persona affects all of their actions. (c, d) Rollouts of a correlation-based and a causal simulator. (e) Outcome scores for correlation-based and a causal simulator.

Now consider the counterfactual question: given the factual trajectory, how would traffic have evolved had Ego applied the emergency brake at $t = 2.5\,\mathrm{s}$, for example, in response to a false collision warning? In the ground-truth counterfactual, TV's distracted driver reacts late to Ego's braking, resulting in a near miss (Fig. 1(a), lower row). In this setting, suppose most trailing drivers in the observational training data are attentive and brake promptly after the vehicle ahead slows, as may occur in heavy traffic. A correlation-based sequence simulator may therefore predict that TV brakes promptly and maintains a safe gap (Fig. 1(c)). This prediction ignores TV's distracted persona, which remains unchanged under the intervention on Ego's braking action. A causal simulator should instead use the factual trajectory to infer the drivers' personas and preserve the underlying characteristic under the intervention. Thus, a causal simulation that retains TV's delayed response can predict a shrinking gap and a near miss, as illustrated in Fig. 1(d). In this example, the causal simulator more closely matches the illustrated ground-truth counterfactual and identifies a hazard that the correlation-based simulator misses. Fig. 1(e) illustrates the decision-making implication: the correlation-based simulator assigns a safety score above the displayed hazard threshold, whereas the causal simulator and ground truth fall below it. □

Motivated by this limitation of correlation-based methods for counterfactual inference, we ask a general question: how can a Transformer provide guarantees for arbitrary counterfactual trajectory simulations? To address this question, we model the problem of counterfactual simulation through the language of causality (Pearl, 2000, Ch. 7). Within this framework, we propose *counterfactual coverage* (Def. 3) to evaluate whether a simulator class is reliable for decision-making: the class satisfies coverage when the interval spanned by its counterfactual estimates contains the true counterfactual quantity. Building on this criterion, we make the following contributions:

1. We develop `CaLM` (Def. 5), a Transformer-based architecture that incorporates causal graphical constraints into its attention mechanism, together with training and inference algorithms for counterfactual trajectory simulation (Algs. 1 and 2).
2. We prove that the `CaLM` simulator class satisfies counterfactual coverage under the stated assumptions (Thm. 6) and show that this property is not guaranteed for vanilla Transformers trained to fit observational conditional distributions (Thm. 4).
3. We evaluate `CaLM` across chess, buyer-seller negotiation, autonomous driving, clinical trajectories, and poker, demonstrating its effectiveness in counterfactual simulations (Sec. 4).

2 Structural Causal Models and Counterfactual Simulations

We begin by describing the causal mechanisms that generate an ordered sequence $\mathbf{Z} = (Z_1, \ldots, Z_L)$. Lowercase letters denote realizations of the corresponding uppercase variables. We leverage structural

causal models as our semantic foundation (Pearl, 2000, Ch. 7) and follow the presentation of Bareinboim et al. (2022).

Definition 2 (Structural Causal Model)**.** A structural causal model (SCM) is a tuple $\mathcal{M} = \langle \mathbf{U}, \mathbf{Z}, \mathcal{F}, P(\mathbf{U}) \rangle$, where $\mathbf{U}$ is a collection of exogenous variables determined by factors outside the model; $\mathbf{Z} = (Z_1, \ldots, Z_L)$ is a sequence of endogenous variables; $\mathcal{F} = \{f_{Z_j}\}_{j=1}^{L}$ is a set of structural functions, each determining Z_j from its endogenous parents $\mathbf{Pa}_{Z_j} \subseteq \{Z_1, \ldots, Z_L\} \setminus \{Z_j\}$ and exogenous inputs $\mathbf{U}_{Z_j} \subseteq \mathbf{U}$; and $P(\mathbf{U})$ is a joint distribution over the exogenous variables. □

To illustrate, each Z_j is determined by $f_{Z_j}(\mathbf{Pa}_{Z_j}, \mathbf{U}_{Z_j})$. Importantly, causal semantics in SCMs differ from the term "*causal*" used in the attention masks in the literature, which refers to restricting attention to preceding positions and the current position (see App. E.1). Here, the mechanism f_{Z_j} depends not only on its parents but also on unobserved exogenous variables $\mathbf{U}_{Z_j}$, and the exogenous inputs to different mechanisms may overlap or be statistically dependent. Consequently, unobserved factors can induce dependencies beyond the causal effects among sequence variables. In this work, we consider the order $\mathbf{Pa}_{Z_j} \subseteq \{Z_1, \ldots, Z_{j-1}\}$. The causal diagram $\mathcal{G}$ induced by an SCM has one vertex per Z_j, with arrow denoting $Z_j \in \mathbf{Pa}_{Z_k}$ and bidirected arrow indicating shared or dependent U. Example 14 in App. F illustrates the corresponding SCM for Example 1 and its causal diagram through drivers' action generation.

Counterfactuals An intervention $do(\mathbf{x})$ replaces the mechanisms for $\mathbf{X} \subseteq \mathbf{Z}$ with constants $\mathbf{x}$, yielding the submodel $\mathcal{M}_{\mathbf{x}}$. For $\mathbf{Y} \subseteq \mathbf{Z}$, the potential outcome $\mathbf{Y}_{[\mathbf{x}]}(\mathbf{u})$ (also written $\mathbf{Y}_{\mathbf{x}}(\mathbf{u})$) is obtained by evaluating $\mathcal{M}_{\mathbf{x}}$ at $\mathbf{U} = \mathbf{u}$. For a joint query Q specifying $\mathbf{Y}_{i[\mathbf{x}_i]} = \mathbf{y}_i$, $i = 1, \ldots, k$, all worlds share the same exogenous assignment, giving (Bareinboim et al., 2022, Def. 7)

$$P^{\mathcal{M}}(Q) = \int_{\mathcal{X}_{\mathbf{U}}} \prod_{i=1}^{k} \mathbf{1}\{\mathbf{Y}_{i[\mathbf{x}_i]}(\mathbf{u}) = \mathbf{y}_i\} \, dP^{\mathcal{M}}(\mathbf{u}). \tag{1}$$

Here, $\mathcal{X}_{\mathbf{U}}$ denotes the joint domain of $\mathbf{U}$, and $\mathbf{1}\{\cdot\}$ is the indicator function. Conditional queries additionally condition on factual evidence of positive probability. For the driving SCM in Example 14, let E_k denote Ego's braking action at $t = 2.5\,\mathrm{s}$ and let $R = 1$ denote a safe continuation. Given the factual trajectory $\mathbf{Z} = \mathbf{z}$, the probability that traffic would have remained safe had Ego applied the emergency brake ($E_k = 1$) is $P^{\mathcal{M}^*}(R_{E_k=1} = 1 \mid \mathbf{Z} = \mathbf{z})$. The observational distribution $P(\mathbf{Z})$ is induced by the SCM when all intervention sets are empty.

2.1 Counterfactual Simulators

The previous subsection formalizes the ground-truth model $\mathcal{M}^*$ that induces the observed distribution $P^{\mathcal{M}^*}(\mathbf{Z})$ and the unknown target counterfactual quantities $P^{\mathcal{M}^*}(Q)$. A *counterfactual simulator* $\mathcal{S}$ is a model fitted to the observational distribution $P^{\mathcal{M}^*}(\mathbf{Z})$ and equipped with a procedure for approximating the target counterfactual distributions.

Ideally, the surrogate counterfactual estimate $P^{\mathcal{S}}(Q)$ should exactly match the ground-truth quantity $P^{\mathcal{M}^*}(Q)$. However, counterfactual quantities may not be uniquely determined by the observational distribution $P(\mathbf{Z})$ and causal structure $\mathcal{G}$ (a limitation known as non-identifiability in causal inference) (Pearl, 2009; Correa et al., 2021; Bareinboim et al., 2022). Thus, to assess whether $P^{\mathcal{S}}(Q)$ provides a suitable estimate of $P^{\mathcal{M}^*}(Q)$ despite this ambiguity, we introduce the following properties.

Definition 3 (Counterfactual coverage)**.** Let $\mathcal{M}^*$ denote the underlying true SCM and $Q \in \mathbb{Q}$ the target query, where $\mathbb{Q}$ is the family of joint and conditional counterfactual queries under consideration. Consider a nonempty simulator class $\Omega_{\mathcal{S}}$ such that $P^{\mathcal{S}}(\mathbf{Z}) = P^*(\mathbf{Z})$ for every $\mathcal{S} \in \Omega_{\mathcal{S}}$. Define the *coverage interval* $\mathcal{I}_Q(\Omega_{\mathcal{S}}) = [\inf_{\mathcal{S} \in \Omega_{\mathcal{S}}} P^{\mathcal{S}}(Q), \sup_{\mathcal{S} \in \Omega_{\mathcal{S}}} P^{\mathcal{S}}(Q)] \subseteq [0, 1]$. The simulator class $\Omega_{\mathcal{S}}$ satisfies *counterfactual coverage* if $P^{\mathcal{M}^*}(Q) \in \mathcal{I}_Q(\Omega_{\mathcal{S}})$. Its *error* is quantified by $\delta_Q(\Omega_{\mathcal{S}}) = \inf_{v \in \mathcal{I}_Q(\Omega_{\mathcal{S}})} \left| v - P^{\mathcal{M}^*}(Q) \right|$, and its *confidence* is quantified by the length of the coverage interval, i.e., $\eta_Q(\Omega_{\mathcal{S}}) = |\mathcal{I}_Q(\Omega_{\mathcal{S}})|$. □

To illustrate, consider the driving query $P^{\mathcal{M}^*}(Q) = P^{\mathcal{M}^*}(R_{E_k=1} = 1 \mid \mathbf{Z} = \mathbf{z})$ from Example 14, which asks whether traffic would have remained safe had Ego applied the emergency brake. For a simulator class fitted to the observational distribution $P(\mathbf{Z})$, coverage means that the true quantity

lies between the bounds of the class's counterfactual estimates. Coverage error measures how far the true quantity lies outside these bounds.

Counterfactual coverage can certify decisions that depend on whether $P^{\mathcal{M}^*}(Q) < \tau$. For example, a safety assessment flags Ego's emergency-braking intervention as hazardous when its counterfactual probability of a safe continuation, $P^{\mathcal{M}^*}(Q) = P^{\mathcal{M}^*}(R_{E_k=1} = 1 \mid \mathbf{Z} = \mathbf{z})$, falls below a threshold τ. Under coverage, $\sup \mathcal{I}_Q(\Omega_{\mathcal{S}}) < \tau$ guarantees $P^{\mathcal{M}^*}(Q) < \tau$, certifying the decision. Without coverage, the true value may exceed this bound. The lower bound works analogously. Interval width $\eta_Q(\Omega_{\mathcal{S}})$ quantifies confidence: narrower intervals provide more informative certificates when coverage holds. Broad intervals may cover the truth but leave decisions unresolved; for example, an upper bound of 1 cannot certify $P^{\mathcal{M}^*}(Q) < \tau$ for $\tau \leq 1$.

We study the counterfactual coverage limitations of three Transformer (TF) baselines $\mathcal{S}_{\mathrm{TF}}^{(b)}$ inducing $P^{\mathcal{S}_{\mathrm{TF}}^{(b)}}(Q)$, $b \in \{0, 1, 2\}$, which differ in training data and inference procedure. For the target query $P(Q) = P(\mathbf{Z}_{Z_j = z'_j} = \mathbf{z}' \mid \mathbf{Z} = \mathbf{z})$, rollout-based procedures initialize the generated trajectory $\mathbf{Z}'$ with the factual prefix $\mathbf{z}'_{<j} = \mathbf{z}_{<j}$ and replacement token z'_j. **B0 (prompt-based simulation)** gives an instruction-following language model, possibly accessed through an API, the factual trajectory and an instruction specifying the intervention (j, z'_j), and requests a counterfactual continuation that retains the factual prefix. **B1 (single-trajectory rollout)** trains a Transformer on individual observed trajectories by next-token prediction and, at inference, continues from the modified prefix $(\mathbf{z}_{<j}, z'_j)$, discarding the factual suffix. **B2 (paired-trajectory rollout)** trains on pairs of observed trajectories and, at inference, continues the modified prefix in the second segment, retaining the complete factual trajectory in the first segment as context. App. F.3 gives formal definitions of $P^{\mathcal{S}_{\mathrm{TF}}^{(b)}}(Q)$ and an example of each procedure.

Population next-token prediction minimizes the cross-entropy with the training distribution, or equivalently its Kullback–Leibler divergence from the model distribution. Thus, in the realizable limit, every population minimizer induces the same observational distribution. For B1 and B2, this also determines their rollout distributions whenever the modified prefixes have positive probability under the respective training distributions. For B2, the relevant distribution is the joint distribution of the two observed trajectories, which can contain more information than a single-trajectory marginal. In a misspecified model class, a unique best in-class distribution requires an additional assumption. For B0, observational fitting alone need not determine the response to a counterfactual instruction outside the training support; the following result therefore explicitly assumes that the fitted class agrees on the query answer. Under this assumption, increasing the amount of observational data can yield zero interval width without guaranteeing coverage. The proof is given in App. A.1.

Theorem 4 (Limitation of the TF baselines)**.** *Fix $b \in \{0, 1, 2\}$ and a counterfactual query Q. Let $\Omega_{\mathrm{TF}}^{(b)}$ be a nonempty class of simulators fitted to baseline Bb's population training distribution, and suppose that all simulators in this class agree on the query answer. Let $\mathcal{S}_{\mathrm{TF}}^{(b)}$ be any member of this class, whose answer is $P^{\mathcal{S}_{\mathrm{TF}}^{(b)}}(Q)$. Suppose there are two SCMs compatible with the graph and this same full observational training distribution that give different values of $P(Q)$. Then*

$$\mathcal{I}_Q(\Omega_{\mathrm{TF}}^{(b)}) = [P^{\mathcal{S}_{\mathrm{TF}}^{(b)}}(Q), P^{\mathcal{S}_{\mathrm{TF}}^{(b)}}(Q)], \qquad \eta_Q(\Omega_{\mathrm{TF}}^{(b)}) = 0, \tag{2}$$

and there exists an SCM $\mathcal{M}^*$ compatible with that graph and observational distribution for which $\delta_Q(\Omega_{\mathrm{TF}}^{(b)}) > 0$. Consequently, the fitted TF class does not guarantee counterfactual coverage. □

3 Causal Modeling with Transformers

Theorem 4 motivates a simulator class that retains the causal mechanisms compatible with the observations. We incorporate the causal diagram $\mathcal{G}$ into the paired-trajectory construction of B2, representing the exogenous variables explicitly and restricting each mechanism to its endogenous and exogenous parents.

Encoding the causal diagram. A mask that permits attention only to parent positions does not, by itself, enforce the required dependencies in a multilayer Transformer. Consider the chain $Z_1 \rightarrow$

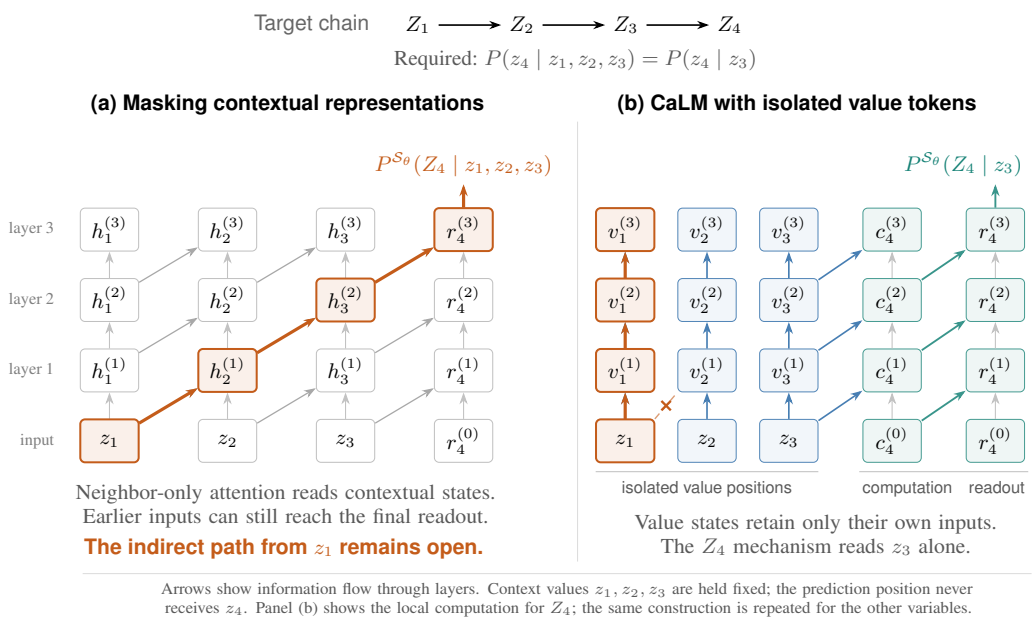

Figure 2: Attention restrictions for the chain $Z_1 \to Z_2 \to Z_3 \to Z_4$. (a) Restricting each position to its immediate predecessor still permits information from z_1 to reach the prediction of Z_4 through successive layers. (b) In the construction of Def. 5, value tokens retain only their supplied values. The computation and readout for Z_4 therefore depend only on its parent value z_3. Here $h_j^{(\ell)}$ is a contextual representation, while $v_j^{(\ell)}$, $c_j^{(\ell)}$, and $r_j^{(\ell)}$ denote value, computation, and readout representations. Ordinary causal influence through generated values is preserved.

$Z_2 \to Z_3 \to Z_4$ with independent exogenous variables. Its observational distribution satisfies $P(z_4 \mid z_1, z_2, z_3) = P(z_4 \mid z_3)$ whenever the conditioning event has positive probability. However, after one attention layer, the representation at position 2 can contain information about z_1. After two layers, that information can reach position 3, and after three layers it can reach the prediction of Z_4. Holding the supplied parent value z_3 fixed does not hold its contextual representation fixed. Thus, the prediction can depend on z_1 even after conditioning on z_3, as illustrated in Fig. 2(a). Related leakage through contextual keys and values is studied by Kasai et al. (2020, Sec. 2.3); Gloeckler et al. (2024, App. A1.1) characterize how attention paths across layers enlarge the possible dependencies.

To prevent this dependence, we separate a variable's supplied value from the computation of its mechanism. For each Z_j, we use a block of computation tokens C_j, a readout token R_j, and a value token X_j containing z_j. The computation and readout tokens precede the value token, so that the readout predicts its value without receiving the target. Each value token contains only its supplied value and a fixed variable identity. Token positions and block sizes are fixed independently of the supplied values. Exogenous inputs use the same representation. Following canonical causal models (CCMs) and neural causal models (NCMs), we use independent exogenous roots that can be shared among mechanisms (Zhang et al., 2022; Xia et al., 2023). The family includes one exogenous variable for each maximal clique of bidirected edges in $\mathcal{G}$, supplied only to mechanisms in that clique; private exogenous variables are also permitted. An isolated vertex forms a singleton clique, and exogenous domains and marginal distributions can vary.

Definition 5 (`CaLM`)**.** Given a causal diagram $\mathcal{G}$, a `CaLM` simulator $\mathcal{S}_\theta$ represents a sequential SCM with mechanisms

$$Z_j = f_{Z_j,\theta}(\mathbf{Pa}_{Z_j}, \mathbf{U}_{Z_j}), \qquad j = 1, \ldots, L, \tag{3}$$

and the graph-compatible exogenous structure described above. Each mechanism is computed by a Transformer with the following attention restrictions at every layer:

1. A value token attends only to itself.
2. A computation token for Z_j attends only to the value tokens for $\mathbf{Pa}_{Z_j}$ and $\mathbf{U}_{Z_j}$, and to preceding or current computation tokens in its own block.

3. The readout for Z_j attends only to computation tokens in its own block.

Allowed entries of the additive attention mask $\mathbf{M}$ are zero; all other entries are $-\infty$. Residual connections, normalization, and feed-forward layers operate separately at each position. The readout supplies logits, and its largest logit determines the value of Z_j, with a fixed rule for ties. The resulting $f_{Z_j,\theta}$ is deterministic given $(\mathbf{pa}_{Z_j}, \mathbf{u}_{Z_j})$; all randomness in value generation belongs to $\mathbf{U}$. Factual and counterfactual copies use the same mechanisms, including the same variable and position encodings. Counterfactual inference follows Eq. 1, holding the complete exogenous assignment fixed across worlds. Dropout is disabled during this evaluation. □

Compatibility permits a mechanism to ignore an allowed parent, but forbids dependence on inputs excluded by $\mathcal{G}$. After generation, the value token is embedded anew from the generated value, without the hidden state that produced it. Thus, ancestors can affect descendants through generated parent values, without introducing additional dependence when those values are held fixed (Fig. 2(b)).

Proposition 8 in App. A.2 establishes that, at every depth, the readout depends only on the supplied values of its designated endogenous and exogenous parents. Abduction may use all factual evidence to infer $\mathbf{U}$; each mechanism receives that assignment only through its designated exogenous inputs.

Training. Given observational data $\mathcal{D} = \{\mathbf{z}^{(i)}\}_{i=1}^{n}$, we fit $\mathcal{S}_\theta$ by minimizing

$$\mathcal{L}_{\text{obs}}(\theta, \tau) = -\frac{1}{n}\sum_{i=1}^{n} \log \mathbb{E}_{P_\theta(\mathbf{U})}\left[\prod_j p_{\theta,\tau}\left(z_j^{(i)} \mid \mathbf{pa}_{Z_j}^{(i)}, \mathbf{U}_{Z_j}\right)\right], \tag{4}$$

where $p_{\theta,\tau}$ is the softmax of the readout logits divided by $\tau > 0$. Both trajectories in a training pair are observational; their joint exogenous prior shares only the factors specified by the pairing procedure. All unobserved variables, including player ratings in chess, are marginalized. We approximate the expectation by joint exogenous samples and update θ through the logits under teacher forcing, including the prior gradient if it is learned. We anneal τ toward zero during optimization and check that the training probabilities concentrate on greedy outputs. When the winning logit gaps dominate τ, observational fit of the training law transfers to the greedy simulator (Proposition 11). App. B.1 gives the sampled objective and training algorithm; App. B.2 states the concentration condition precisely.

Counterfactual inference. Let $\mathbf{E} = \mathbf{e}$, with $\mathbf{E} \subseteq \mathbf{Z}$, denote the completed factual trajectory, with positive probability under $\mathcal{S}_\theta$. Abduction infers $P^{\mathcal{S}_\theta}(\mathbf{U} \mid \mathbf{E} = \mathbf{e})$ using the entire trajectory, including the suffix after the intervention. The second segment contains a partial counterfactual copy of that same trajectory. For each posterior draw, we replace the intervened mechanisms by the specified values and complete the copy in causal order, using the same mechanisms and complete exogenous assignment. Remaining values are decoded greedily from their counterfactual parent values and designated exogenous inputs. Averaging query indicators over these draws estimates $P^{\mathcal{S}_\theta}(Q)$. App. B.3 gives the algorithm and its correctness argument.

Counterfactual coverage. Coverage follows from soundness and expressiveness: every fitted `CaLM` represents a compatible SCM, and every compatible SCM's counterfactual distribution can be represented by a fitted `CaLM`. CCMs supply the finite representation (Zhang et al., 2022); the NCM class-equivalence argument then applies once the masked Transformer can represent its structural functions (Xia et al., 2023). Theorem 9 in App. A.3 establishes these properties. Let $\Omega(\mathcal{G}, P^*)$ be the SCMs compatible with $\mathcal{G}$ and the observational law $P^*(\mathbf{Z})$, and let Ω_{CaLM} be the `CaLM` simulators fitting that law. For paired training, $\mathbf{Z}$ comprises both observed trajectories and $P^*(\mathbf{Z})$ is their full joint law.

Theorem 6 (Counterfactual coverage of `CaLM`)**.** *Assume finite endogenous domains, the exogenous structure in Def. 5, arbitrary finite exogenous cardinalities and marginal distributions, and sufficient Transformer capacity. Let $\Omega(\mathcal{G}, P^*)$ be nonempty. For every joint counterfactual query Q, or conditional query whose conditioning event has positive probability in every compatible SCM,*

$$\mathcal{I}_Q(\Omega_{\text{CaLM}}) = \left[\inf_{\mathcal{M}\in\Omega(\mathcal{G},P^*)} P^{\mathcal{M}}(Q), \sup_{\mathcal{M}\in\Omega(\mathcal{G},P^*)} P^{\mathcal{M}}(Q)\right]. \tag{5}$$

Consequently, $\delta_Q(\Omega_{\mathrm{CaLM}}) = 0$ for every true SCM $\mathcal{M}^* \in \Omega(\mathcal{G}, P^*)$. This interval has the smallest width, and hence maximal confidence, among closed intervals guaranteeing coverage for every SCM compatible with the same graph and observations. □

The proof is in App. A.4. This is a population guarantee for the full observationally fitted class and exact inference. As in Gumbel-max SCMs, fixing the exogenous assignment fixes the outcomes in every world (Oberst & Sontag, 2019; Lorberbom et al., 2021); coverage additionally requires considering all compatible mechanisms. App. E.3 discusses these connections.

Approximating the coverage interval. In practice, we seek models that fit the observations and attain the smallest and largest query values. For an empirical observational distribution $\widehat{P}_n$, discrepancy D, and tolerance ϵ_n, this gives

$$\inf_{\theta:\, D(P^{\mathcal{S}_\theta}(\mathbf{Z}),\widehat{P}_n)\le\epsilon_n} P^{\mathcal{S}_\theta}(Q), \qquad \sup_{\theta:\, D(P^{\mathcal{S}_\theta}(\mathbf{Z}),\widehat{P}_n)\le\epsilon_n} P^{\mathcal{S}_\theta}(Q). \tag{6}$$

Proposition 7 (Consistent approximation of the coverage interval)**.** *Under the assumptions of Theorem 6, take D to be total variation, D_{TV}, in Eq. 6. Suppose $\epsilon_n \downarrow 0$ and $D_{\mathrm{TV}}(\widehat{P}_n, P^*) \le \epsilon_n$ eventually almost surely. If global optimization and query estimation errors vanish almost surely, the estimated endpoints converge almost surely to those of $\mathcal{I}_Q(\Omega_{\mathrm{CaLM}})$, for joint queries and queries conditioned on factual evidence of positive probability.* □

The proof is in App. A.5; App. B.4 describes the numerical approximation. Finite collections of fitted models may miss the extrema and need not cover the true counterfactual value.

4 Experiments

4.1 Design

In a general sequential environment, the endogenous variables are decoded into a sequence actions from agents, A_t and terminal outcomes Y of the form: $A_1, A_2, \ldots A_T, Y$.

To investigate queries of the form $P(Y_{X=x'} = y \mid \mathbf{Z} = \mathbf{z})$, we use a twin-network construction (Balke & Pearl, 2013). The first segment contains the completed factual trajectory $\mathbf{z}$, and the second contains a partial counterfactual copy of that same trajectory. In the intervened model $\mathcal{M}_{x'}$, the equation for X is replaced by the constant assignment $X = x'$. Completing this copy follows the abduction, intervention, and prediction procedure in Algorithm 2.

The factual trajectory determines the posterior $P^{\mathcal{S}_\theta}(\mathbf{U} \mid \mathbf{Z} = \mathbf{z})$. Each draw includes all exogenous variables and is retained when evaluating the counterfactual copy.

The details of the experimental setup are provided in App. C.

Baselines and metrics. We compare `CaLM` with two Transformers trained on the same paired sequences: Naive, which uses the standard left-to-right mask (App. C.1), and Random-Sparse, which removes attention connections uniformly at random. We report counterfactual MSE, its relative reduction versus Naive, Pearson correlation, directional accuracy of the predicted effect, harm-avoidance rate, and head-to-head win rate against Naive (App. C.3), at three model scales (App. C.2).

4.2 Dataset Construction

We experimented across a diverse array of domains, focusing on multi-player game trajectories like in chess and poker, and real world settings like healthcare, autonomous driving and negotiations, the environments for which are described in detail below. For simulator-based evaluations, the benchmark target is defined by the specified generating simulator and intervention. Counterfactual quantities can be non-identifiable from observational data, so this choice supplies an explicit reference model. The Stockfish score transformation used in chess is a proxy benchmark target and is not by itself the same-game counterfactual win probability in an SCM of the players. No training pair contains

an intervention: both halves of every twin sequence are ordinary observational trajectories, and the intervention appears only at evaluation time. This changes the generation procedure at evaluation time, although the resulting trajectories need not lie outside the support of the observational distribution. Due to space constraints, the details for the domain specific construction are provided in App. C.4.

Driving as a running example. Counterfactual prediction in multi-agent traffic matters for safety assessment, e.g., under ISO 26262 and ISO 21448 (SOTIF). Our driving domain instantiates Example 1: a three-lane highway simulator with seven vehicles, in which each driver's unobserved style (aggressive, normal, cautious, or distracted) sets its reaction time, desired headway, and braking. We intervene on Ego's action at $t = 2.5\,\mathrm{s}$, for example forcing an emergency brake, and query the terminal safety outcome: safe continuation, near miss, or collision (For more details, see App. C.4). Longitudinal motion follows the Intelligent Driver Model, and the four evaluated interventions are an emergency brake check, an aggressive cut-in, a yield-and-merge, and an overtake with full acceleration.

4.3 Results

Here, we provide a summary of results with details in the App. D. Ten results & ablations isolate *where* attention edge severing helps, *why*, and where it does not; space constraints relegate their full treatment to the sections and figures cited here. The aggregated results across all domains and scales are provided in App. D.1. **What the gain is**: a per-intervention breakdown over all 20 canonical maneuvers shows how the advantage varies depending on the intervention and setting. (Fig. 3 below; Tiers S and L in App. D.2), and how the advantage appears only under the counterfactual prediction but `CaLM` tracks the Naive baseline on observational sequences App. D.7; **Why it is causal and not just regularization**: a from-scratch random-sparse control at matched connection density collapses and inverts the sign of physiological effects, while `CaLM` gains at the same budget (App. D.4);. **Where it is largest**: multi-move interventions ($K{=}3$ consecutive interventions widen the advantage (App. D.5) and so do interventions later in the trajectory (App. D.6); **Safety and uncertainty**: `CaLM` enforces strictly monotone hazard under deceleration sweeps whereas Naive inverts (App. D.3); and under weight perturbation `CaLM`'s distance-to-bound distribution stochastically dominates the baseline's in twelve of twelve bucket comparisons (App. D.9). **At the foundation-model scale**: the mechanism applies in the fine-tuning of Llama-3-8B-Instruct (Grattafiori et al., 2024) and Mistral-7B-Instruct (Jiang et al., 2023) on 100k MIMIC-IV (Johnson et al., 2023) inspired trajectories, reducing counterfactual MSE by up to 41% across four clinical protocols (App. D.8), and an 8B `CaLM` model outperforms four frontier (Claude Opus 4.7 (Anthropic, a), Claude Opus 5 (Anthropic, b), GPT-5.5 (OpenAI, b) and GPT-5.6-sol (OpenAI, a)) in-context reasoners on the same benchmark (App. D.10).

4.4 Performance by Intervention Type

Fig. 3 provides a multi-metric evaluation of all four canonical interventions in each domain at model scale *Tier M*; the corresponding evaluations of all 20 scenarios for *Tier S* and *Tier L* are shown in Figs. 5 and 6 in the appendix. Besides counterfactual MSE and Pearson correlation r, we report directional accuracy (DA), the fraction of queries for which the predicted change in outcome has the correct sign; the harm-avoidance rate (HAR), the fraction of harmful interventions that the model also predicts to be harmful; and the head-to-head (H2H) win rate, the fraction of queries on which `CaLM`'s prediction is closer to the ground truth than Naive's. Across the 20 scenarios at Tier M, the standout improvements of `CaLM` are as follows:

- **Chess**: Here, the selected moves are intervened on, while the players' skills remain unobserved causes of subsequent play. `CaLM` consistently reduces MSE by 9.2–16.3%, maintaining at least 96.7% DA and at least a 56.7% H2H win rate, whereas Random-Sparse completely collapses to near-zero or negative correlation.
- **eBay**: On *Rigid Counteroffer*, `CaLM` reduces counterfactual MSE by 66.7% ($0.0035 \to 0.0012$), with $r = 0.997$ vs. 0.991, a 96.7% HAR, and a 73.3% H2H win rate, whereas Random-Sparse increases MSE by 159.5%.
- **Sepsis**: In clinical titration, dense attention and random-sparse controls frequently invert the sign of the treatment effect. Under acute *Vasopressor Escalation*, `CaLM` maintains $r = 0.640$ (vs. 0.602



Figure 3: Performance by intervention type at Tier M for Naive (red), Random-Sparse (orange), and CaLM (blue), across four interventions in each of five domains. Columns: counterfactual MSE, relative MSE reduction versus Naive, Pearson r, directional accuracy, harm-avoidance rate, and head-to-head win rate against Naive.

for Naive and 0.549 for Random-Sparse), with a 73.3% HAR, avoiding excessive vasopressor toxicity, and a 53.3% H2H win rate.

- **Poker**: Confounding between unobserved player aggression and bet size causes dense attention to collapse during aggressive maneuvers. On *Preflop Bluff Squeeze*, CaLM reduces counterfactual MSE by 90.7%, raises the Pearson correlation from $r = 0.143$ to 0.833, improves DA by 46.6%, and wins 86.7% of the H2H comparisons against Naive.
- **Driving**: On *Emergency Brake Check*, CaLM reduces MSE by 59.3%, attains a 96.7% HAR, and wins 76.7% of the H2H comparisons against Naive.

5 Conclusion

We introduced counterfactual coverage to evaluate whether a simulator class is reliable for decision-making, and showed that next-token prediction alone does not guarantee it. We proposed CaLM, which encodes a causal diagram in Transformer attention so that each prediction depends only on its designated endogenous and exogenous parents, and proved that, under the stated assumptions, the CaLM simulator class satisfies coverage with the narrowest interval valid for every compatible SCM. Across chess, negotiation, driving, clinical, and poker benchmarks, CaLM improved counterfactual predictions over dense and randomly sparsified Transformers in most evaluated settings.

Limitations Our guarantees are population results: they assume a correctly specified causal diagram, sufficient model capacity, and exact fitting and inference, and they do not provide finite-sample guarantees for a finite collection of trained models. In our benchmarks, the same simulator generates the training data and defines the evaluation target, so performance under model misspecification and on real-world counterfactual data remains to be studied.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic data sets (e.g. MedGemma rendered the clinical records described in Appendix D.8), to implement methods and evaluation code, and to assist writing. Additionally, we used generative AI tools to create or modify figures, to draft and edit parts of the paper for readability, and to search for related literature. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

App. A contains complete proofs of all theoretical results, with assumptions stated in each theorem. App. B gives the training and inference algorithms, and App. C.4 describes each simulator, tokenization, dataset size, model configuration, and evaluation metric. The code will be released upon acceptance.

REFERENCES

Vassili Alexiadis, James Colyar, John Halkias, Rob Hranac, and Gene McHale. The next generation simulation program. *Institute of Transportation Engineers. ITE Journal*, 74(8):22, 2004.

Anthropic. Introducing Claude Opus 4.7 — anthropic.com. https://www.anthropic.com/news/claude-opus-4-7, a. [Accessed 26-09-2026].

Anthropic. Introducing Claude Opus 5 — anthropic.com. https://www.anthropic.com/news/claude-opus-5, b. [Accessed 26-09-2026].

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Matthew Backus, Thomas Blake, Brad Larsen, and Steven Tadelis. Sequential bargaining in the field: Evidence from millions of online bargaining interactions. *The Quarterly Journal of Economics*, 135(3):1319–1361, 2020.

Alexander Balke and Judea Pearl. Counterfactual probabilities: Computational methods, bounds and applications, 2013. URL https://arxiv.org/abs/1302.6784.

Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 507–556. Association for Computing Machinery, New York, NY, USA, 1st edition, 2022.

Juan Correa, Sanghack Lee, and Elias Bareinboim. Nested counterfactual identification from arbitrary surrogate experiments. *Advances in Neural Information Processing Systems*, 34:6856–6867, 2021.

Laura Evans, Andrew Rhodes, Waleed Alhazzani, Massimo Antonelli, Craig M Coopersmith, Craig French, Flávia R Machado, Lauralyn Mcintyre, Marlies Ostermann, Hallie C Prescott, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Intensive care medicine*, 47(11):1181–1247, 2021.

Aaron Foss, Chloe Evans, Sasha Mitts, Koustuv Sinha, Ammar Rizvi, and Justine T Kao. Causalvqa: A physically grounded causal reasoning benchmark for video models. *arXiv preprint arXiv:2506.09943*, 2025.

Manuel Gloeckler, Michael Deistler, Christian Dietrich Weilbach, Frank Wood, and Jakob H. Macke. All-in-one simulation-based inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15735–15766. PMLR, 2024. URL https://proceedings.mlr.press/v235/gloeckler24a.html.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. Physiobank, physiotoolkit, and physionet. *Circulation*, 101(23):e215–e220, 2000. doi: 10.1161/01.CIR.101.23.e215. URL https://www.ahajournals.org/doi/abs/10.1161/01.CIR.101.23.e215.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems*, 34:1273–1286, 2021.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. Cladder: Assessing causal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:31038–31065, 2023.

Alistair Johnson, Tom Pollard, and Roger Mark. Mimic-iii clinical database (version 1.4). *PhysioNet*, 10(C2XW26):2, 2016.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Jungo Kasai, James Cross, Marjan Ghazvininejad, and Jiatao Gu. Non-autoregressive machine translation with disentangled context transformer. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5144–5155. PMLR, 2020. URL https://proceedings.mlr.press/v119/kasai20a.html.

Ilyes Khemakhem, Ricardo Pio Monti, Robert Leech, and Aapo Hyvärinen. Causal autoregressive flows, 2021. URL https://arxiv.org/abs/2011.02268.

Guy Lorberbom, Daniel D. Johnson, Chris J. Maddison, Daniel Tarlow, and Tamir Hazan. Learning generalized Gumbel-max causal mechanisms. In *Advances in Neural Information Processing Systems*, volume 34, pp. 26792–26803, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/e143c01e314f7b950daca31188cb5d0f-Abstract.html.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes, 2022. URL https://arxiv.org/abs/2204.07258.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. *arXiv preprint arXiv:2209.00588*, 2022.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with Gumbel-max structural causal models. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4881–4890. PMLR, 2019. URL https://proceedings.mlr.press/v97/oberst19a.html.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition — openai.com. https://openai.com/index/gpt-5-6/, a. [Accessed 26-09-2026].

OpenAI. Introducing GPT-5.5 — openai.com. https://openai.com/index/introducing-gpt-5-5/, b. [Accessed 26-09-2026].

Jack Parker-Holder, Shlomi Fruchter, et al. Genie 3: A new frontier for world models. *Google DeepMind Blog*, 4, 2025.

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International conference on machine learning*, pp. 4055–4064. PMLR, 2018.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000. 2nd edition, 2009.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009. ISBN 9780521749190.

Karl Pearson and Francis Galton. Vii. note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352):240–242, 1895. doi: 10.1098/rspl.1895.0041. URL https://royalsocietypublishing.org/doi/abs/10.1098/rspl.1895.0041.

PokerTracker Software, LLC. *PokerTracker 4: Poker Tracking & Analysis Software*, 2026. URL https://pokertracker.com. Accessed: September 22, 2026.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. Counterfactual story reasoning and generation, 2019. URL https://arxiv.org/abs/1909.04076.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rajat Rasal, Daniel C. Castro, Nick Pawlowski, and Ben Glocker. Deep structural causal shape models, 2022. URL https://arxiv.org/abs/2208.10950.

Jan Robine, Marc Höftmann, Tobias Uelwer, and Stefan Harmeling. Transformer-based world models are happy with 100k interactions. *arXiv preprint arXiv:2303.07109*, 2023.

Kristina E Rudd, Sarah Charlotte Johnson, Kareha M Agesa, Katya Anne Shackelford, Derrick Tsoi, Daniel Rhodes Kievlan, Danny V Colombara, Kevin S Ikuta, Niranjan Kissoon, Simon Finfer, et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the global burden of disease study. *The Lancet*, 395(10219):200–211, 2020.

Ari Seff, Brian Cera, Dian Chen, Mason Ng, Aurick Zhou, Nigamaa Nayakanti, Khaled S Refaat, Rami Al-Rfou, and Benjamin Sapp. Motionlm: Multi-agent motion forecasting as language modeling. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8545–8556. IEEE, 2023.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Mervyn Singer, Clifford S. Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, Richard S. Hotchkiss, Mitchell M. Levy, John C. Marshall, Greg S. Martin, Steven M. Opal, Gordon D. Rubenfeld, Tom van der Poll, Jean-Louis Vincent, and Derek C. Angus. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):801–810, 02 2016. ISSN 0098-7484. doi: 10.1001/jama.2016.0287. URL https://doi.org/10.1001/jama.2016.0287.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. ISSN 00029556. URL http://www.jstor.org/stable/1412159.

Martin Treiber, Ansgar Hennecke, and Dirk Helbing. Congested traffic states in empirical observations and microscopic simulations. *Physical Review E*, 62(2):1805–1824, August 2000. ISSN 1095-3787. doi: 10.1103/physreve.62.1805. URL http://dx.doi.org/10.1103/PhysRevE.62.1805.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wei Wu, Xiaoxin Feng, Ziyan Gao, and Yuheng Kan. Smart: Scalable multi-agent real-time motion generation via next-token prediction. *Advances in Neural Information Processing Systems*, 37: 114048–114071, 2024.

Kevin Xia, Yushu Pan, and Elias Bareinboim. Neural causal models for counterfactual identification and estimation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vouQcZS8KfW.

Jiannan Xiang, Yi Gu, Zihan Liu, Zeyu Feng, Qiyue Gao, Yiyan Hu, Benhao Huang, Guangyi Liu, Yichi Yang, Kun Zhou, et al. Pan: A world model for general, interactable, and long-horizon world simulation. *arXiv preprint arXiv:2511.09057*, 2025.

Hong Xiong, Feng Wu, Leon Deng, Megan Su, and Li wei H Lehman. G-transformer: Counterfactual outcome prediction under dynamic and time-varying treatment regimes, 2024. URL https://arxiv.org/abs/2406.05504.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform bad for graph representation?, 2021. URL https://arxiv.org/abs/2106.05234.

Junzhe Zhang, Jin Tian, and Elias Bareinboim. Partial counterfactual identification from observational and experimental data. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26548–26558. PMLR, 2022. URL https://proceedings.mlr.press/v162/zhang22ab.html.

Appendix

A Proofs **15**

A.1 Limitation of the TF Baselines 15

A.2 Dependence of the Readout 16

A.3 Soundness and Expressiveness 16

A.4 Counterfactual Coverage 18

A.5 Consistent Approximation of the Coverage Interval 19

B Training and Counterfactual Inference **19**

B.1 Training with Discrete Exogenous Variables 19

B.2 Agreement as the Temperature Vanishes 21

B.3 Counterfactual Inference 22

B.4 Approximating the Coverage Interval 23

C Additional details for experiments **23**

C.1 Baselines 23

C.2 Model Scaling Architecture 24

C.3 Evaluation Metrics 24

C.4 Domain Specifications 25

D Additional Results and Ablations **30**

D.1 Overall Results 30

D.2 Performance by Intervention Type (Tier S and Tier L) 30

D.3 Monotonicity Testing 32

D.4 Sparsity Comparisons 33

D.5 Longer Interventions 33

D.6 Position of Interventions 33

D.7 Observational Parity vs. Counterfactual Divergence 35

D.8 Foundation Model Scaling 35

D.9 Prediction Intervals under Weight Perturbations 36

D.10 Frontier Models 38

E Background and Related Work **38**

E.1 Background 38

E.2 Related Work 39

E.3 Relation to Existing Causal Models 40

F Examples **41**

F.1 Motivation 41

F.2 SCM and Counterfactuals 42

F.3 Transformer Inference Procedures 43

A PROOFS

A.1 LIMITATION OF THE TF BASELINES

Proof of Theorem 4. By assumption, every simulator in $\Omega_{\mathrm{TF}}^{(b)}$ returns $P^{\mathcal{S}_{\mathrm{TF}}^{(b)}}(Q)$ for any representative $\mathcal{S}_{\mathrm{TF}}^{(b)} \in \Omega_{\mathrm{TF}}^{(b)}$. Its coverage interval is therefore $[P^{\mathcal{S}_{\mathrm{TF}}^{(b)}}(Q), P^{\mathcal{S}_{\mathrm{TF}}^{(b)}}(Q)]$, with width zero. Non-identifiability provides two SCMs $\mathcal{M}_0$ and $\mathcal{M}_1$ compatible with the given graph and the full observational training distribution, but with $q_0 = P^{\mathcal{M}_0}(Q) \neq P^{\mathcal{M}_1}(Q) = q_1$. Since the available training distribution is the same, the fitted baseline class and its query value are the same under both SCMs. At least one of q_0, q_1 differs from $P^{\mathcal{S}_{\mathrm{TF}}^{(b)}}(Q)$. Taking the corresponding SCM as $\mathcal{M}^*$ gives $\delta_Q(\Omega_{\mathrm{TF}}^{(b)}) = |P^{\mathcal{M}^*}(Q) - P^{\mathcal{S}_{\mathrm{TF}}^{(b)}}(Q)| > 0$. In fact, the triangle inequality gives

$$\max\{|q_0 - P^{\mathcal{S}_{\mathrm{TF}}^{(b)}}(Q)|, |q_1 - P^{\mathcal{S}_{\mathrm{TF}}^{(b)}}(Q)|\} \geq \frac{|q_0 - q_1|}{2}. \tag{7}$$

Thus the baseline cannot guarantee coverage for every SCM compatible with the available information. □

For completeness, let P_b^* denote the training distribution of the complete observed token sequence for baseline Bb. The population next-token loss equals

$$\mathbb{E}_{P_b^*}[-\log P_\theta(\mathbf{Z})] = H(P_b^*) + D_{\mathrm{KL}}(P_b^* \| P_\theta). \tag{8}$$

If the class is realizable and a global minimum is attained, every minimizer has the same law P_b^*. Every conditional used by the B1 or B2 rollout is then determined on conditioning events with positive probability. For B2, P_b^* includes the joint law of both observed trajectories. This argument does not determine a conditional at a zero-probability prefix, or B0's response to an instruction absent from its fitted distribution. The assumption that fitted simulators agree on the query value is necessary at precisely this step. Similarly, convergence to a best in-class approximation in a misspecified class does not establish uniqueness unless it is assumed separately. Non-identifiability yields the existential failure proved above; an almost-sure claim would additionally require a probability measure over compatible SCMs or query values.

An example with positive-probability prefixes. Let $X \sim \mathrm{Bernoulli}(1/2)$, independently of a response-function index $T = (T_0, T_1) \in \{0,1\}^2$, and set $Y = T_X$. For $0 < \alpha < 1$, give T the probabilities

$$P(T = (0,0)) = P(T = (1,1)) = \frac{1-\alpha}{2}, \qquad P(T = (0,1)) = P(T = (1,0)) = \frac{\alpha}{2}. \tag{9}$$

Every such SCM is compatible with $X \to Y$ and induces $P(X = x, Y = y) = 1/4$ for all x, y. Nevertheless,

$$P(Y_{X=1} = 1 \mid X = 0, Y = 0) = \alpha. \tag{10}$$

B1's population rollout returns $P(Y = 1 \mid X = 1) = 1/2$. The choices $\alpha = 1/4$ and $\alpha = 3/4$ therefore have identical observational laws and coverage error $1/4$ for this baseline. If B2 is trained on independent observational pairs, its paired law is also identical across these SCMs, every paired prefix has positive probability, and its rollout returns the same value $1/2$. This last example uses independent training pairs; a different pairing procedure must be checked against its full joint law, as required by the theorem.

A.1.1 TRAINING DISTRIBUTIONS FOR PAIRED TRAJECTORIES

The paired-trajectory baseline uses two observed trajectories, but their joint training distribution depends on how the pairs are constructed. Write $\mathbf{Z}^{(1)} = \mathbf{Z}_{1:L}$ and $\mathbf{Z}^{(2)} = \mathbf{Z}_{L+1:2L}$ for the two segments, and let P_D denote their joint distribution.

Independent pairs. If the two trajectories are sampled independently from the same observational distribution, then

$$P_D(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) = P(\mathbf{z}^{(1)})P(\mathbf{z}^{(2)}). \tag{11}$$

Under exact population fitting, the first trajectory supplies no information about the second. Consequently, on conditioning events with positive probability, the paired rollout reduces to the single-trajectory rollout in Eq. 58. The example above uses this construction.

Pairs with shared latent factors. Alternatively, two trajectories may share persistent latent factors H, such as player skill, while their trajectory-specific randomness is sampled independently conditional on H. Their joint distribution is

$$P_D(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) = \int P(\mathbf{z}^{(1)} \mid h)P(\mathbf{z}^{(2)} \mid h)P(dh). \tag{12}$$

For example, two games played by the same players can share their underlying skills while differing in their sampled moves. The first trajectory can therefore provide information about the shared factors that informs predictions of the second trajectory. These are paired observational trajectories, not ground-truth factual–counterfactual pairs, because the trajectory-specific background variables are resampled rather than preserved across the two segments. The replacement token is still treated as an observation in the second trajectory rather than explicitly as an intervention. To apply Theorem 4 to this construction, the candidate SCMs must agree on the full paired distribution P_D, not merely on the single-trajectory marginal $P(\mathbf{Z})$. The independent-pair counterexample above does not by itself establish that condition for shared-latent pairs.

A.2 Dependence of the Readout

Proposition 8 (Dependence of the readout)**.** *Under Def. 5, the readout for Z_j is a function only of $(\mathbf{pa}_{Z_j}, \mathbf{u}_{Z_j})$, the model parameters, and fixed variable identities, at every Transformer depth. In particular, changing any other supplied value while holding these inputs fixed cannot change the readout.* □

Proof of Proposition 8. Fix the parameters, variable identities, and supplied endogenous and exogenous values. At layer zero, a value-token representation depends only on its own supplied value, and the computation and readout representations contain only fixed token and position information. Suppose the stated dependence restrictions hold at layer ℓ. A value token attends only to itself, so its representation at layer $\ell + 1$ still depends only on its own value. For variable Z_j, every key and value available to a computation token comes either from a value token for $\mathbf{Pa}_{Z_j}$ or $\mathbf{U}_{Z_j}$, or from a computation token in the same block. By the induction hypothesis, all these inputs, as well as the query representation, depend only on $(\mathbf{pa}_{Z_j}, \mathbf{u}_{Z_j})$ and fixed information. The resulting attention weights and output have the same restriction. The readout receives only these local computation representations, and its own residual also satisfies the restriction. Finally, positionwise normalization, residual additions, and feed-forward maps cannot introduce dependence on another position. The claim follows by induction for every depth. □

The proposition fixes the *supplied* parent values. It does not prevent a generated parent value from changing when one of its own parents changes. This is the distinction between the causal path $Z_1 \to Z_2 \to Z_3 \to Z_4$ and a path through hidden representations that remains available after fixing z_3.

A.3 Soundness and Expressiveness

The graphical restrictions ensure that a `CaLM` simulator represents a plausible SCM. Coverage additionally requires the class to represent every counterfactual distribution compatible with the graph. We establish these properties by adapting the soundness and expressiveness results for NCMs (Xia et al., 2023, Thms. 1–2) and the canonical representation theorem of Zhang et al. (2022, Thm. 2.4). As in those results, endogenous domains are finite, and confounding is represented by independent exogenous roots shared among mechanisms.

Theorem 9 (Soundness and expressiveness of CaLM)**.** *Every simulator satisfying Def. 5 induces an SCM compatible with $\mathcal{G}$ and agrees with that SCM on every counterfactual query. For finite endogenous domains, allow the exogenous variables to have arbitrary finite cardinalities and marginal distributions, and allow sufficient Transformer depth, width, and feed-forward capacity with deterministic categorical decoding. Then, for every SCM $\mathcal{M}$ compatible with $\mathcal{G}$, there exists a CaLM simulator $\mathcal{S}_\theta$ such that*

$$P^{\mathcal{S}_\theta}(Q) = P^{\mathcal{M}}(Q) \qquad \text{for every } Q \in \mathbb{Q}. \tag{13}$$

Conditional queries are restricted to evidence of positive probability in the corresponding SCM. □

We first establish the representation property needed to transfer the NCM expressiveness argument to the masked Transformer.

Lemma 10 (Representation of finite structural functions)**.** *Fix finite endogenous and exogenous domains and the parent sets permitted by Def. 5. With sufficient depth, embedding width, and feed-forward capacity, a CaLM Transformer with deterministic categorical decoding can represent any collection of functions $f_{Z_j} : \mathcal{X}_{\mathbf{Pa}_{Z_j}} \times \mathcal{X}_{\mathbf{U}_{Z_j}} \to \mathcal{X}_{Z_j}$. The construction permits shared Transformer parameters across mechanisms.* □

Proof. Use one computation token for each mechanism. Encode each supplied pair consisting of a variable V and its value v by a basis vector $e_{V,v}$, with disjoint coordinates for distinct pairs. An initial positionwise map can form these codes from distinct token and variable embeddings. Computation tokens have fixed mechanism identities in separate coordinates. Set attention queries and keys to zero, so each computation token takes a uniform average over its permitted inputs, including itself. If there are m_j parent tokens, suitable value projections and the residual produce the code

$$h_j(\mathbf{pa}_{Z_j}, \mathbf{u}_{Z_j}) = b_j + \frac{1}{m_j + 1} \sum_{V \in \mathbf{Pa}_{Z_j} \cup \mathbf{U}_{Z_j}} e_{V,v}, \tag{14}$$

where b_j contains only fixed information identifying the mechanism. The layout fixes m_j independently of the supplied values. This code is injective in the allowed parent assignment: changing any value changes a distinct coordinate. The mechanism identity also distinguishes codes belonging to different functions.

A shared positionwise feed-forward map can assign the required output scores on this finite set of distinct codes. For an explicit construction, choose a linear projection that maps all codes to distinct scalars and interpolate each desired score by a continuous piecewise-linear function. Such a function is a finite linear combination of ReLU functions and an affine term. Choose the correct category's score to be one and every other score to be zero. The next attention layer lets the readout receive this computation token, and categorical decoding returns the required value. Other feed-forward activations that approximate these finite interpolants also suffice: an error smaller than half the score margin preserves every decoded value.

Residuals add only known local information and can be incorporated into the finite interpolation. If layer normalization is used, choose unit gains and include two distinct fixed reference coordinates in each code. Two such codes cannot become identical under normalization: that would require an affine rescaling and shift, which the fixed reference coordinates rule out. The positionwise maps can therefore restore the required codes after normalization. Fixed token roles distinguish value, computation, and readout states, allowing these maps to be implemented with shared parameters. Thus the construction respects all attention restrictions while evaluating every specified finite function. □

Proof of Theorem 9. For soundness, Proposition 8 ensures that each readout is a function only of its specified endogenous and exogenous parents. The deterministic decoding and exogenous prior therefore define

$$\mathcal{M}_\theta = \langle \mathbf{U}, \mathbf{Z}, \{f_{Z_j,\theta}\}_{j=1}^{L}, P_\theta(\mathbf{U}) \rangle. \tag{15}$$

The independent exogenous roots are shared only within bidirected cliques, and the endogenous arguments are exactly the permitted parents, including arguments a function may ignore. Thus $\mathcal{M}_\theta$ is compatible with $\mathcal{G}$. This is the graphical consistency argument of Xia et al. (2023, Thm. 1 and App. A.1), with Proposition 8 establishing the required parent restriction for CaLM. The simulator

evaluates interventions by replacing the same mechanisms as $\mathcal{M}_\theta$, with the same exogenous assignment in every world. It therefore agrees with $\mathcal{M}_\theta$ on all joint counterfactual events and, by taking ratios, on conditional queries with positive-probability evidence.

For expressiveness, fix an SCM $\mathcal{M}$ compatible with $\mathcal{G}$. The canonical representation theorem of Zhang et al. (2022, Thm. 2.4) supplies an SCM $\mathcal{M}^0 = \langle \mathbf{U}^0, \mathbf{Z}, \mathcal{F}^0, P(\mathbf{U}^0)\rangle$ with finite exogenous domains, the same causal structure, and the same joint counterfactual distributions as $\mathcal{M}$. Its exogenous roots remain independent. This is the canonical-model reduction used in Xia et al. (2023, Thm. 2 and App. A.2).

To put $\mathcal{M}^0$ in the exogenous form of Def. 5, follow the grouping construction in that NCM proof. The endogenous children of each $U \in \mathbf{U}^0$ form a bidirected clique, so assign U to one maximal clique containing those children. For each maximal clique C, let $\widetilde{U}_C$ be the tuple of roots assigned to it; an empty tuple is constant. These tuples have finite domains and are mutually independent because they partition the independent roots of $\mathcal{M}^0$. Supplying $\widetilde{U}_C$ to each mechanism in C introduces no forbidden shared input. Each mechanism extracts its original exogenous arguments and ignores any other components. Lemma 10 now supplies a masked Transformer that evaluates all the resulting finite structural functions.

Couple this simulator and $\mathcal{M}^0$ through the same assignment of $\mathbf{U}^0$ and its grouped tuples. In any intervened world, replaced variables agree by construction, and all remaining variables agree whenever their parents do. Induction in causal order therefore gives identical trajectories for each exogenous assignment. Using that same assignment in every world establishes equality of all joint counterfactual distributions. Since $\mathcal{M}^0$ and $\mathcal{M}$ already agree on these distributions, Eq. 13 follows. Conditional queries follow by taking ratios with positive denominators. For paired observations, the argument is applied to the SCM for both observed games and therefore preserves their full joint observational law as well. □

A.4 Counterfactual Coverage

Let $P^*(\mathbf{Z})$ denote the observational distribution. Define

$$\Omega(\mathcal{G}, P^*) = \{\mathcal{M} : \mathcal{M} \text{ is compatible with } \mathcal{G},\ P^{\mathcal{M}}(\mathbf{Z}) = P^*(\mathbf{Z})\}, \tag{16}$$

For paired training, apply the construction to the SCM for both observed games, with their shared causes specified by the graph. Here $\mathbf{Z}$ comprises both games, and $P^*(\mathbf{Z})$ is their full joint law. The counterfactual query still concerns one game and its intervened copy. Let Ω_{CaLM} be the class of `CaLM` simulators fitting the same observational law.

Proof of Theorem 6. Let

$$A_Q = \{P^{\mathcal{S}}(Q) : \mathcal{S} \in \Omega_{\mathrm{CaLM}}\}, \qquad B_Q = \{P^{\mathcal{M}}(Q) : \mathcal{M} \in \Omega(\mathcal{G}, P^*)\}. \tag{17}$$

For every $\mathcal{S} \in \Omega_{\mathrm{CaLM}}$, soundness gives a compatible SCM with the same counterfactual distributions. Since observational distributions are a special case, this SCM also induces $P^*(\mathbf{Z})$. For paired training, $\mathbf{Z}$ contains both observed games, so this equality includes their joint law. Hence $A_Q \subseteq B_Q$. Conversely, expressiveness represents every $\mathcal{M} \in \Omega(\mathcal{G}, P^*)$ by a simulator agreeing on all queries, including the observational distribution. That simulator belongs to Ω_{CaLM}, so $B_Q \subseteq A_Q$. Taking infima and suprema proves Eq. 5. These inclusions are the same use of soundness and expressiveness as in the graphical–neural equivalence theorem of Xia et al. (2023, Thm. 3). They also transfer the tight canonical-model bounds of Zhang et al. (2022, Thm. 2.8) to `CaLM`.

For any compatible true SCM, $P^{\mathcal{M}^*}(Q) \in B_Q$, which establishes coverage. If a closed interval J covers every value in B_Q, its lower endpoint is at most $\inf B_Q$ and its upper endpoint is at least $\sup B_Q$. Thus $\mathcal{I}_Q(\Omega_{\mathrm{CaLM}}) \subseteq J$, proving the minimum-width claim even when the extrema are not attained. □

The theorem concerns the population observational law and exact counterfactual inference. For Algorithm 1, Proposition 11 gives a sufficient condition for observational fit of the training law to transfer to the greedy simulator. The following result addresses convergence of the interval endpoints as the observational tolerance vanishes, provided that optimization and query estimation errors also vanish.

A.5 Consistent Approximation of the Coverage Interval

Proof of Proposition 7. For fixed finite endogenous domains, the canonical representation theorem of Zhang et al. (2022, Thm. 2.4) bounds the required exogenous cardinalities in terms of the graph and endogenous domains. Apply it to the fixed exogenous structure of Def. 5, after the grouping in Appendix A.3. The resulting canonical family has finitely many structural-function tables and a product of closed probability simplices for its independent exogenous roots. This family is compact, and observational and joint counterfactual probabilities are continuous on it: for each table, they are polynomials in the exogenous probabilities. By Theorem 9, this family and the expressive `CaLM` class attain the same observational and counterfactual laws. We may therefore analyze both optimization problems on the canonical family, without assuming compactness of the Transformer parameter space.

Fix a sample path on which the assumed error bounds hold. Let F_n contain the canonical models with $D_{\mathrm{TV}}(P^{\mathcal{M}}(\mathbf{Z}), \widehat{P}_n) \le \epsilon_n$, and let F_0 contain those with $P^{\mathcal{M}}(\mathbf{Z}) = P^*(\mathbf{Z})$. These sets are closed subsets of the compact canonical family. Since F_0 is nonempty and $D_{\mathrm{TV}}(\widehat{P}_n, P^*) \le \epsilon_n$ eventually, $F_0 \subseteq F_n$ for all sufficiently large n. Moreover, every $\mathcal{M}_n \in F_n$ satisfies

$$D_{\mathrm{TV}}(P^{\mathcal{M}_n}(\mathbf{Z}), P^*(\mathbf{Z})) \le \epsilon_n + D_{\mathrm{TV}}(\widehat{P}_n, P^*) \le 2\epsilon_n \longrightarrow 0. \tag{18}$$

Thus every accumulation point of such models lies in F_0.

Write $a_n = \min_{\mathcal{M}\in F_n} P^{\mathcal{M}}(Q)$ and $a = \min_{\mathcal{M}\in F_0} P^{\mathcal{M}}(Q)$. For joint queries, continuity ensures that these minima exist, and inclusion gives $a_n \le a$ eventually. If a_n did not converge to a, a subsequence of minimizing models would have query values bounded below a by a fixed positive amount. Compactness supplies a convergent further subsequence, whose limit lies in F_0 and, by continuity, has query value strictly below a, a contradiction. The same argument applies to the maxima. Theorem 6 identifies the limiting extrema with the endpoints of $\mathcal{I}_Q(\Omega_{\mathrm{CaLM}})$.

For a conditional query with factual evidence $\mathbf{E} = \mathbf{e}$ and $P^*(\mathbf{E} = \mathbf{e}) > 0$, every sufficiently large F_n has evidence probability at least $P^*(\mathbf{E} = \mathbf{e}) - 2\epsilon_n$, which is bounded away from zero. The query is a ratio of continuous joint probabilities on these sets, so the same argument applies. Finally, vanishing global optimization and query estimation errors leave both endpoint limits unchanged. The argument holds on an event of probability one, proving the almost-sure statement. □

For independent observations from $P^*(\mathbf{Z})$ on a fixed finite domain, $\epsilon_n = n^{-1/4}$ is one sufficient choice: concentration of the empirical frequencies and the Borel–Cantelli lemma give $D_{\mathrm{TV}}(\widehat{P}_n, P^*) \le \epsilon_n$ eventually almost surely. The tolerance must accommodate sampling error; exact agreement with every empirical distribution need not admit the population-fitted models. The optimization assumption requires feasible candidates whose query values approach the global extrema in Eq. 6, with vanishing query estimation error for the selected candidates. It is not a guarantee that repeated initializations or a penalized objective achieve those extrema.

B Training and Counterfactual Inference

B.1 Training with Discrete Exogenous Variables

Training fits the observational distribution using the supplied causal diagram and data $\mathcal{D} = \{\mathbf{z}^{(i)}\}_{i=1}^{n}$. Both trajectories in a training pair are observational; neither is labeled as a counterfactual. No observed rating or skill label is inserted into the sequence or used as a target. The latent inputs are sampled and marginalized, rather than supervised. For paired training, their joint prior must reflect the pairing procedure: only factors shared by the two observed games are shared in their latent assignments. This training distribution is distinct from the counterfactual construction, where both copies concern the same game and use the same complete exogenous assignment.

We parameterize the simulator as $\mathcal{S}_\theta$ and train its logits using a positive temperature τ. For fixed inputs $(\mathbf{pa}_{Z_j}, \mathbf{u}_{Z_j})$, let $\ell_{j,\theta,a}$ denote the readout logit for category $a \in \mathcal{X}_{Z_j}$. The training probabilities

and structural mechanism are

$$p_{\theta,\tau}(a \mid \mathbf{pa}_{Z_j}, \mathbf{u}_{Z_j}) = \frac{\exp(\ell_{j,\theta,a}/\tau)}{\sum_{b \in \mathcal{X}_{Z_j}} \exp(\ell_{j,\theta,b}/\tau)}, \tag{19}$$

$$f_{Z_j,\theta}(\mathbf{pa}_{Z_j}, \mathbf{u}_{Z_j}) = \arg\max_{a \in \mathcal{X}_{Z_j}} \ell_{j,\theta,a}. \tag{20}$$

For a sampled exogenous assignment, Eq. 19 permits differentiation through the logits even when the exogenous variables are discrete. Training evaluates these probabilities under teacher forcing; it does not differentiate through the argmax in Eq. 20. The structural functions in Eq. 20 are deterministic given their complete inputs. Their observational law is

$$P^{\mathcal{S}_\theta}(\mathbf{Z} = \mathbf{z}) = \int \prod_{j=1}^{L} \mathbf{1}\{f_{Z_j,\theta}(\mathbf{pa}_{Z_j}, \mathbf{u}_{Z_j}) = z_j\}\, dP_\theta(\mathbf{u}). \tag{21}$$

The indicators do not provide useful pathwise gradients through the logits. During training, we replace them by the probabilities in Eq. 19, obtaining

$$P_{\theta,\tau}(\mathbf{Z} = \mathbf{z}) = \int \prod_{j=1}^{L} p_{\theta,\tau}(z_j \mid \mathbf{pa}_{Z_j}, \mathbf{u}_{Z_j})\, dP_\theta(\mathbf{u}). \tag{22}$$

For paired observations, $\mathbf{Z}$ contains both observed games, and the integral marginalizes their shared and private exogenous variables.

The objective in Eq. 4 is the negative log likelihood of Eq. 22. We approximate it by

$$\widehat{\mathcal{L}}_{\mathrm{obs}}(\theta, \tau) = -\frac{1}{n} \sum_{i=1}^{n} \log \left[\frac{1}{K} \sum_{k=1}^{K} \prod_{j} p_{\theta,\tau}\left(z_j^{(i)} \mid \mathbf{pa}_{Z_j}^{(i)}, \mathbf{u}_{Z_j}^{(i,k)}\right) \right], \tag{23}$$

where $\mathbf{u}^{(i,k)} \sim P_\theta(\mathbf{U})$ is a joint draw for the observed sequence or pair. The product includes all observed targets in both games, with the specified shared exogenous inputs, and excludes structural tokens, latent assignments, and padding.

Conditional on each draw, the log probability is differentiable in the logits:

$$\frac{\partial \log p_{\theta,\tau}(z_j \mid \mathbf{pa}_{Z_j}, \mathbf{u}_{Z_j})}{\partial \ell_{j,\theta,a}} = \frac{1}{\tau} \left[\mathbf{1}\{a = z_j\} - p_{\theta,\tau}(a \mid \mathbf{pa}_{Z_j}, \mathbf{u}_{Z_j}) \right]. \tag{24}$$

Thus backpropagation updates the readout, Transformer, and exogenous-value embeddings without differentiating a discrete draw. For a fixed prior and s_{ik} as defined in Algorithm 1, the gradient of the sampled minibatch loss is

$$\nabla_\theta \widehat{\mathcal{L}}_t = -\frac{1}{|\mathcal{B}_t|} \sum_{i} \sum_{k=1}^{K} \frac{e^{s_{ik}}}{\sum_{r=1}^{K} e^{s_{ir}}} \nabla_\theta s_{ik}. \tag{25}$$

The assignments are held fixed in this derivative. If the prior is parameterized by θ, differentiating a sampled objective also requires the derivative of its sampling law, for example through a score-function estimator; exact summation over finite exogenous domains is another option. The sampled probability is unbiased for Eq. 22, but its negative logarithm is generally biased by Jensen's inequality. This sampling approximation is separate from the temperature relaxation.

We anneal τ toward zero while updating θ. Algorithm 1 uses a geometric schedule $\tau_t = \tau_0 \rho^t$, with $\tau_0 > 0$ and $0 < \rho < 1$; every finite optimization step uses a positive temperature. When the winning logit is separated from the alternatives by a gap large relative to τ_t, the training probabilities concentrate on the greedy output. Proposition 11 makes this condition precise and shows that the relaxed observational law then approaches the law of $\mathcal{S}_{\theta_t}$. Thus, in the converged limit, observational fit of the relaxed model transfers to the greedy simulator. Annealing alone does not imply concentration if the logit gaps shrink with τ_t. We therefore check both observational fit under greedy generation and the agreement between the temperature-dependent probabilities and greedy outputs.

Here $\mathrm{LSE}_k(s_k) = \log \sum_k \exp(s_k)$ is evaluated stably. The gradient is taken through the logits at the current positive temperature; annealing is performed after the parameter update. At small

Algorithm 1: `CaLM` training with sampled exogenous variables and temperature annealing.

Input: Observational data $\mathcal{D}$, diagram $\mathcal{G}$, simulator family $\mathcal{S}_\theta$, latent sample count K, steps T, learning rates α_t, initial temperature $\tau_0 > 0$, decay $0 < \rho < 1$.

Output: A simulator $\mathcal{S}_{\widehat{\theta}}$ selected by greedy observational validation fit, with readout concentration diagnostics.

Initialize θ_0 and the optimizer state
for $t = 0, \dots, T-1$ **do**
  Sample an observational minibatch $\mathcal{B}_t$
  foreach $\mathbf{z}^{(i)} \in \mathcal{B}_t$ **do**
    Draw K joint assignments $\mathbf{u}^{(i,k)} \sim P_{\theta_t}(\mathbf{U})$ with the specified sharing between observed games
    Insert observed values and sampled exogenous values; construct the mask in Def. 5
    Compute the logits of $\mathcal{S}_{\theta_t}$ under teacher forcing
    Set $s_{ik} \leftarrow \sum_j \log p_{\theta_t,\tau_t}(z_j^{(i)} \mid \mathbf{pa}_{Z_j}^{(i)}, \mathbf{u}_{Z_j}^{(i,k)})$ using log-softmax of logits divided by τ_t
  Set $\widehat{\mathcal{L}}_t \leftarrow -|\mathcal{B}_t|^{-1} \sum_i [\mathrm{LSE}_k(s_{ik}) - \log K]$
  Estimate $\widehat{\mathbf{g}}_t$ by differentiating $\widehat{\mathcal{L}}_t$ with respect to θ_t, including the prior contribution if learned
  Update $\theta_{t+1} \leftarrow \mathrm{AdamW}(\theta_t, \mathrm{clip}(\widehat{\mathbf{g}}_t), \alpha_t)$
  Anneal $\tau_{t+1} \leftarrow \rho\tau_t$
  Validate $\mathcal{S}_{\theta_{t+1}}$ by sampling exogenous assignments and decoding greedily; monitor readout concentration
Select $\widehat{\theta}$ by greedy observational validation fit and report readout concentration
return $\mathcal{S}_{\widehat{\theta}}$ *with the mechanisms in Eq. 20*

temperatures, sufficient latent samples and stable log probabilities remain necessary to estimate the marginal likelihood. A finite training run approximates the zero-temperature limit; the coverage theorem concerns the observationally fitted greedy SCM class.

For fixed inputs, the same training kernel can be represented as the distribution of $\arg\max_a \{\ell_{j,\theta,a} + \tau G_{j,a}\}$ with independent standard Gumbel variables (Oberst & Sontag, 2019; Lorberbom et al., 2021). These are auxiliary variables for the relaxation; Algorithm 1 evaluates its probabilities directly. The final simulator uses Eq. 20, with all generative randomness supplied by $\mathbf{U}$. The following proposition states when the two observational laws agree in the limit.

B.2 Agreement as the Temperature Vanishes

For each mechanism, define

$$\varepsilon_{j,\theta,\tau}(\mathbf{u}_{Z_j}) = \max_{\mathbf{pa}_{Z_j}} \left[1 - p_{\theta,\tau}\Big(f_{Z_j,\theta}(\mathbf{pa}_{Z_j}, \mathbf{u}_{Z_j}) \mid \mathbf{pa}_{Z_j}, \mathbf{u}_{Z_j}\Big)\right]. \tag{26}$$

The maximum ranges over the finite endogenous parent domain. This quantity measures the probability assigned to values other than the greedy output, with the exogenous inputs fixed.

Proposition 11 (Training and greedy observational distributions)**.** *For finite endogenous domains and any θ and $\tau > 0$,*

$$D_{\mathrm{TV}}\big(P_{\theta,\tau}(\mathbf{Z}), P^{\mathcal{S}_\theta}(\mathbf{Z})\big) \le \sum_{j=1}^{L} \mathbb{E}_{P_\theta(\mathbf{U})}[\varepsilon_{j,\theta,\tau}(\mathbf{U}_{Z_j})]. \tag{27}$$

Consequently, if the right-hand side tends to zero along (θ_t, τ_t) and $P_{\theta_t,\tau_t}(\mathbf{Z})$ converges to $P^*(\mathbf{Z})$ in total variation, then $P^{\mathcal{S}_{\theta_t}}(\mathbf{Z})$ also converges to $P^*(\mathbf{Z})$ in total variation. □

Proof. Draw the same $\mathbf{U}$ for the relaxed and greedy sequences and evaluate both in causal order. Until their first disagreement, they have identical endogenous parent values. At variable Z_j, the conditional probability that the relaxed categorical draw differs from the greedy value is therefore at most $\varepsilon_{j,\theta,\tau}(\mathbf{U}_{Z_j})$. A union bound over the first disagreement and integration over $P_\theta(\mathbf{U})$ bound the probability that the complete sequences differ by the right-hand side of Eq. 27. The coupling inequality for total variation proves the bound. The convergence statement follows by the triangle inequality. □

To relate this bound to annealing, let $\Delta_{j,\theta}(\mathbf{pa}_{Z_j}, \mathbf{u}_{Z_j})$ be the winning logit minus the largest alternative logit, and let $d_{j,\theta}(\mathbf{u}_{Z_j}) = \min_{\mathbf{pa}_{Z_j}} \Delta_{j,\theta}(\mathbf{pa}_{Z_j}, \mathbf{u}_{Z_j})$. Then

$$\varepsilon_{j,\theta,\tau}(\mathbf{u}_{Z_j}) \leq (|\mathcal{X}_{Z_j}| - 1) \exp\{-d_{j,\theta}(\mathbf{u}_{Z_j})/\tau\}. \tag{28}$$

For a singleton output domain the error is zero. For fixed θ, a unique maximizing logit for every parent assignment and almost every exogenous assignment gives $d_{j,\theta} > 0$ almost surely. As $\tau \downarrow 0$, the error therefore converges to zero almost surely and in expectation by dominated convergence. With finite exogenous domains, convergence of the logits to values with unique maxima likewise gives this conclusion during parameter updates. More generally, it suffices for the normalized gaps $d_{j,\theta_t}/\tau_t$ to diverge in probability under $P_{\theta_t}(\mathbf{U})$. The finite-function construction in Lemma 10 has unit score gaps and satisfies the condition.

A temperature schedule alone is insufficient when the parameters change. For binary logits $(0, \tau_t \log 3)$, the training probability of category 1 is always $3/4$, while the greedy mechanism always returns 1, even though $\tau_t \downarrow 0$. This is why Algorithm 1 checks the greedy observational fit and readout concentration. The proposition controls the agreement of the two laws; it does not establish convergence of the optimizer or coverage of a finite collection of fitted models.

B.3 Counterfactual Inference

For counterfactual inference, let $\mathbf{E} = \mathbf{e}$, with $\mathbf{E} \subseteq \mathbf{Z}$, denote the factual evidence specified by the query, with $P^{\mathcal{S}_\theta}(\mathbf{E} = \mathbf{e}) > 0$. For the same-game query, $\mathbf{E}$ contains all variables of the completed factual game, with observed values $\mathbf{e}$. Abduction obtains $P^{\mathcal{S}_\theta}(\mathbf{U} \mid \mathbf{E} = \mathbf{e})$ for the complete exogenous assignment under the greedy structural mechanisms. For each draw from this distribution, the intervention replaces the mechanisms for $\mathbf{X}$ by the constants $\mathbf{x}$, and the remaining mechanisms are evaluated in causal order. A replaced value does not retain the hidden state of its original mechanism. Other mechanisms retain their own endogenous and exogenous parents, including shared causes such as the players' skills. Algorithm 2 implements these steps for a conditional query; an unconditional query uses the exogenous prior.

Algorithm 2: `CaLM` counterfactual inference by abduction, intervention, and prediction.

Input: Simulator $\mathcal{S}_\theta$; query $P^{\mathcal{S}_\theta}(Q) = P^{\mathcal{S}_\theta}(\mathbf{Y}_{1[\mathbf{x}_1]} = \mathbf{y}_1, \ldots, \mathbf{Y}_{k[\mathbf{x}_k]} = \mathbf{y}_k \mid \mathbf{E} = \mathbf{e})$; number of samples N.
Output: A Monte Carlo estimate of $P^{\mathcal{S}_\theta}(Q)$.

```
Set a ← 0
for s = 1, ..., N do
    Draw u^(s) ~ P^{S_θ}(U | E = e)
    for i = 1, ..., k do
        Initialize a counterfactual copy with the same mechanisms and exogenous assignment u^(s)
        for j = 1, ..., L in causal order do
            if Z_j ∈ X_i then
                Set z_j to its specified value in x_i
            else
                Set z_j ← f_{Z_j,θ}(pa_{Z_j}, u^(s)_{Z_j}) by the greedy readout in Eq. 20
            Insert the value token for z_j with the mask in Def. 5
        Record Y_{i[x_i]}(u^(s)) from this trajectory
    a ← a + ∏_{i=1}^{k} 1{Y_{i[x_i]}(u^(s)) = y_i}
return a/N
```

The factual segment remains available for abduction, including its suffix after the intervention position. The second segment starts with the factual prefix before the earliest intervention and the specified replacement values. Subsequent values are generated with the same mechanisms and exogenous assignment, using parent values from the counterfactual copy. Thus, the segment boundary does not select a different mechanism or a new game. Sharing Transformer weights alone is insufficient if different position encodings change the mechanism between the two segments.

Fix a simulator $\mathcal{S}_\theta$ and evidence of positive probability. For each exact posterior draw $\mathbf{u}$, evaluating the unchanged mechanisms with the specified replacements gives precisely the potential outcomes in

Eq. 1. If $A(\mathbf{u})$ is the product of event indicators evaluated by Algorithm 2, then

$$\mathbb{E}_{\mathbf{U}|\mathbf{E}=\mathbf{e}}[A(\mathbf{U})] = P^{\mathcal{S}_\theta}(\mathbf{Y}_{1[\mathbf{x}_1]} = \mathbf{y}_1, \ldots, \mathbf{Y}_{k[\mathbf{x}_k]} = \mathbf{y}_k \mid \mathbf{E} = \mathbf{e}). \tag{29}$$

With independent posterior draws, the returned average is unbiased for this simulator's query value and converges to it almost surely by the law of large numbers. This establishes correctness for the specified SCM; coverage over compatible SCMs is the separate class-level result of Theorem 6.

For a fully observed factual trajectory, the posterior used by the greedy simulator is

$$dP^{\mathcal{S}_\theta}(\mathbf{u} \mid \mathbf{Z} = \mathbf{z}) = \frac{\prod_j \mathbf{1}\{f_{Z_j,\theta}(\mathbf{pa}_{Z_j}, \mathbf{u}_{Z_j}) = z_j\}}{P^{\mathcal{S}_\theta}(\mathbf{Z} = \mathbf{z})}\, dP_\theta(\mathbf{u}). \tag{30}$$

Here the parent values are taken from the factual trajectory. The complete exogenous assignment is retained for every counterfactual copy. For fixed θ satisfying the concentration condition above, replacing these indicators by the training probabilities gives a posterior that converges to Eq. 30 as $\tau \downarrow 0$, provided the factual probability is positive. This follows by dominated convergence for the unnormalized posterior density and its normalizing constant. At a positive temperature, the two posteriors can differ; Algorithm 2 uses the posterior of the greedy simulator. For discrete evidence with positive probability, rejection sampling from the complete prior gives exact posterior draws in principle: retain precisely those assignments whose factual trajectory agrees with the evidence. An inference network can approximate this posterior. As a direct consequence of using the same mechanisms and exogenous assignment, the empty intervention satisfies

$$P^{\mathcal{S}_\theta}(\mathbf{Z}_\varnothing = \mathbf{z} \mid \mathbf{Z} = \mathbf{z}) = 1. \tag{31}$$

B.4 Approximating the Coverage Interval

Equation 6 optimizes the query over models whose greedy observational law satisfies the chosen fit criterion. A practical approach uses repeated initializations and minimizes $\widehat{\mathcal{L}}_{\mathrm{obs}}(\theta, \tau_t) + \lambda_t s \widehat{P}^{\mathcal{S}_\theta}(Q)$, with $s = 1$ for the lower bound and $s = -1$ for the upper bound. Decreasing λ_t during training prioritizes observational fit. Counterfactual sampling can require a stochastic gradient estimator; replacing the query by a differentiable surrogate does not establish that its extrema match those of $P^{\mathcal{S}_\theta}(Q)$. Each candidate must still be checked against the chosen fit criterion using its greedy observational law; the penalty alone does not enforce the constraint in Eq. 6.

The minimum and maximum over retained models approximate the interval by exploring the fitted class. Even when every retained model fits the observations, a finite collection may miss the extrema and need not cover the true counterfactual value. Accordingly, Theorem 6 does not provide a finite-sample guarantee for the range of these estimates. We distinguish this numerical approximation from the population coverage interval when interpreting the experiments.

Proposition 7 gives consistency when the observational tolerance accommodates sampling error and tends to zero, and the global optimization and query estimation errors vanish. For a lower endpoint, the optimization error is the excess of a retained feasible model's true query value over the feasible infimum; for an upper endpoint, it is the corresponding shortfall from the supremum. The query estimation error must vanish for the selected models, which requires controlling Monte Carlo error as the number of candidates grows. These conditions concern the greedy SCMs used in Eq. 6. Readout concentration relates the training law to those SCMs, but does not itself ensure exploration of the fitted class.

C Additional details for experiments

C.1 Baselines

When trained from scratch, `CaLM` is tested against a dense auto-regressive transformer, referred to as *Naive* (discussed as B2 in Sec.2.1), with identical design, and against a sparse auto-regressive transformer, with the same number of random connections severed as in `CaLM`, referred to as *Random-Sparse*. When pre-trained models are finetuned, we use the exact same training procedure for both `CaLM` and Naive to keep the comparisons fair. What differs is the design discussed in Sec. 3.

C.2 Model Scaling Architecture

To investigate empirical scaling behavior, all five benchmark domains share a standardized 3-tier scaling ladder. What is held fixed across domains is the transformer body (the stack of attention and feed-forward blocks) whose parameter count is identical in every domain, but the vocabulary varies from domain to domain, hence the true model sizes vary across the domains. [1]

- **Small Scale (S)**: $L = 4$ layers, $D = 128$ hidden dim, $H = 4$ attention heads, $d_{\text{ff}} = 512$ feed-forward dimension; body $= 790{,}528$ parameters.
- **Medium Scale (M)**: $L = 8$ layers, $D = 512$ hidden dim, $H = 8$ attention heads, $d_{\text{ff}} = 2048$ feed-forward dimension; body $= 25{,}219{,}072$ parameters.
- **Large Scale (L)**: $L = 16$ layers, $D = 1024$ hidden dim, $H = 16$ attention heads, $d_{\text{ff}} = 4096$ feed-forward dimension; body $= 201{,}539{,}584$ parameters.

All models are optimized using AdamW (Loshchilov & Hutter, 2017) with a cosine learning rate schedule and linear warmup. Weight decay is 0.01 for Driving, Poker, Sepsis and eBay, and 0.05 for Chess, whose larger vocabulary and smaller effective corpus required stronger regularization.

C.3 Evaluation Metrics

For a counterfactual query we write $\hat{y}$ for the model's predicted outcome under $do(A = a')$, y for the simulator ground truth under the same intervention, y_{f} for the *factual* outcome of the untouched trajectory, and n for the number of queries. Let $\tau_i = y_i - y_{\text{f},i}$ and $\hat{\tau}_i = \hat{y}_i - y_{\text{f},i}$. Let $\mathcal{H} = \{i : \tau_i < -0.01\}$, representing all interventions i which decrease the likelihood of preferred outcome (e.g. survival). Let i and j be two scalar interventions from the set of all interventions $\mathcal{Q}$. **Note**: All outcomes are scalars in $[0, 1]$ (e.g. survival probability, win probability, safe driving outcome, or normalized sale price), so every metric below is directly comparable across domains, and is reported for the counterfactual predictions.

Magnitude-Based Metrics: To evaluate general prediction patterns.

- **Mean Squared Error (MSE)** $= \frac{1}{n}\sum_i (\hat{y}_i - y_i)^2$,
- **Absolute Error (AE)** $= |\hat{y}_i - y_i|$.
- **Relative Error Reduction (RER)** $= (\text{MSE}_{\text{Naive}} - \text{MSE}_{\text{CaLM}})/\text{MSE}_{\text{Naive}} \times 100\%$. [2]

Alignment Metric: To evaluate correlation against ground truth evals.

- **Pearson Correlation** (Pearson & Galton, 1895): $r = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$
- **Spearman Correlation** (Spearman, 1904): $\rho = \frac{\sum_{i=1}^{N}(R(\hat{y}_i) - \bar{R})(R(y_i) - \bar{R})}{\sqrt{\sum_{i=1}^{N}(R(\hat{y}_i) - \bar{R})^2}\sqrt{\sum_{i=1}^{N}(R(y_i) - \bar{R})^2}}$ [3]

Decision-Level Metric. To evaluate accuracy in predicting if an intervention improves or worsens a certain outcome.

- **Directional Accuracy (Individual Treatment Effect (ITE) Sign Correctness)** $= \frac{1}{|\mathcal{N}|}\sum_{i\in\mathcal{N}} \mathbb{1}[\text{sign}(\hat{\tau}_i) = \text{sign}(\tau_i)] \times 100\%$, restricted to $\mathcal{N} = \{i : |\tau_i| > 10^{-4}\}$
- **Harm-Avoidance Rate (HAR)** $= \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \mathbb{1}[\hat{\tau}_i < 0] \times 100\%$.
- **Monotonicity-Violation Rate (MVR)**$= \frac{\sum_{i,j\in\mathcal{Q}, y_i - y_j > 0.05} \mathbb{I}(\hat{y}_i < \hat{y}_j)}{\sum_{i,j\in\mathcal{Q}, y_i - y_j > 0.05}}$.

[1] For most of the domains ($|\mathcal{V}| \leq 146$) the embedding contributes under 4% of parameters at the largest tier, giving totals of S=~ 0.8M, M=~ 25M and L=~ 200M respectively. Chess is the exception; its vocabulary of $36{,}126$ tokens makes the embedding dominant at the small tiers, giving totals of S=10.07M, M=62.25M and L=275.56M. Comparisons across domains at a given tier are comparisons at equal transformer-body capacity.

[2] reported w.r.t. the score for the Naive model, unless specified.

[3] where $R(\cdot)$ represents the rank of the prediction across all queries

Paired Comparison Metric: To evaluate model performances against each other per query.

- **Head to Head Win Rate for Model M against Model M' (H2H)** $= \frac{1}{n}\sum_{i\in n} \mathbb{1}[\text{AE}_{\text{M}} - \text{AE}_{\text{M'}}] \times 100\%$ [4]

C.4 Domain Specifications

C.4.1 Chess

Motivation. Chess, represents an ideal domain for causal world modeling: the state space is combinatorially massive ($\approx 10^{40}$ legal positions), the game rules are completely deterministic and Stockfish provides a reproducible engine-based evaluation target. A fundamental counterfactual question is: *"If Bob had played an alternative blunder move at move t, what would Bob's win probability be?"*. In human play, blunders are overwhelmingly correlated with amateur players ($S_W \approx 1200$ Elo), whereas grandmasters rarely hang pieces for no reason. When an exogenous blunder is injected into a game, standard transformers suffer from *correlational bias*; they tend to infer that the player is an amateur, and predict terrible moves on every subsequent turn, instead of taking the true caliber and game state into account.

SCM Formalization. We use the move mechanisms in Eq. 56, with a third terminal outcome for draws in the experiments. The unobserved player variables S_W and S_B represent skill and style, while U_t^W and U_t^B account for additional decision variation. Both types of exogenous inputs are retained when evaluating a counterfactual of the same game.

Under $do(W_t = m')$, the structural equation for that move is replaced by a constant; an intervention on Black's move is defined analogously. Its original endogenous and exogenous parents no longer determine the imposed move. Player skills remain causes of later moves and must be inferred from the factual game; fixing a move does not reveal a player's true skill or guarantee an accurate continuation.

Interventions.

`blunder_tactical_substitution`: Injecting an exogenous alternative blunder move ($do(W_t = m_{\text{blunder}})$) and predicting downstream win probability $P(\text{Black Wins})$, to evaluate whether the model separates localized move error from player strength.

`stockfish_best_countermove`: Substituting a passive or suboptimal move with Stockfish's top-1 engine move ($do(W_t = m^*_{\text{best}})$), simulating grandmaster-level (2800+ Elo) tactical counterplay.

`stockfish_worst_blunder`: Forcing Stockfish's lowest-ranked legal move ($do(W_t = m_{\text{worst}})$), causing acute tactical catastrophe (e.g., hanging a major piece or walking into mate) to test susceptibility to blunder-attribution bias.

`elo_skill_mismatch`: Deliberate skill-action misalignment shocks (e.g., forcing a 2800+ engine refutation on an amateur $S_W < 1800$ Elo player or an unprompted piece hang on a 2500+ Grandmaster $S_W \geq 2400$ Elo), testing whether the model decouples individual moves from player Elo.

Vocabulary. Unlike the other four domains, the chess vocabulary is not learned from the corpus but *combinatorially enumerated* over all syntactically well-formed Standard Algebraic Notation (SAN) strings, giving $|\mathcal{V}| = 36{,}126$ allocated tokens: 5 structural tokens (⟨PAD⟩, ⟨COMP⟩, ⟨READ⟩, ⟨SEP⟩, ⟨UNK⟩); 36,032 SAN move tokens, including check (+), mate (#) and promotion variants; 64 bare destination squares and 2 castling tokens; 3 game result tokens (⟨1-0⟩, ⟨0-1⟩, ⟨1/2-1/2⟩); and 16 tokens for latent skill values (⟨UNOBS_0⟩ to ⟨UNOBS_15⟩). [5]

Data Generation. Reconstructed board positions are evaluated by **Stockfish** at search depth 15 (depth 10 for the worst-move search of `stockfish_worst_blunder`), with scores taken from White's point of view and mates mapped to $\pm 10{,}000$ centipawns. Centipawn scores (cp) are mapped

[4] (ties within 10^{-6})

[5] Only 7,626 of these tokens (21.1%) occur in the Tier L corpus, and a 1,200-token effective vocabulary already covers 97.3% of move occurrences; unused input-embedding rows receive no gradient when embeddings are untied; output rows can still receive gradients through softmax normalization. This is the reason CaLM-Chess departs from the parameter totals of the other domains at every tier (Sec. C.2).

to benchmark scores in $[0, 1]$ using the following logistic transformation:

$$s_W = \frac{1}{1 + 10^{-\mathrm{cp}/400}}, \qquad s_B = 1 - s_W. \tag{32}$$

These transformed engine scores are evaluation proxies; calibration as win probabilities under the players' same-game counterfactual distribution is not established by this transformation. Elo settings control the generating players but are not shown to the models as inputs or targets. The data corpus is generated from controlled Stockfish engine personas.

Dataset Sizes. 11,250, 22,500, 50,000 games for S-,M-,L-sized models respectively with max sequence length being 2000.

C.4.2 eBay Auctions

Motivation. Sequential price negotiations reflect incomplete information bilateral bargaining dynamics, where for instance the observational data conflate multiple latents like the buyer's true price elasticity and seller's rigidity and cost basis.

SCM Formalization. $\mathcal{M}_{ebay} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, where: $\mathbf{U} = \{U_B, U_S\}$ which encode *buyer* persona $U_B = (v_B, K_{\max}, \alpha_B)$ encoding valuation ratio $v_B/L \in [0.50, 1.20]$, impatience threshold $K_{\max} \in [1, 4]$, and concession rate α_B and *seller* persona $U_S = (v_S, \rho_S, \alpha_S)$ encoding reserve cost basis $v_S/L \in [0.60, 0.95]$, rigidity factor $\rho_S \in [0, 1]$, and concession rate α_S respectively; $\mathbf{V} = \{L, B_1, S_1, \ldots, R\}$ where list price L (init. token), alternating buyer offers B_t and seller counteroffers S_t, and bargaining outcome $R \in \{\text{DEAL}, \text{NO_DEAL}\}$ (terminal token) with agreed price P^*;

$\mathcal{F}$:

$$B_t = f_B(B_{<t}, S_{<t}, U_B) \tag{33}$$

$$S_t = f_S(B_{\leq t}, S_{<t}, U_S) \tag{34}$$

$$R = f_R(B_t, S_t, U_B, U_S) \tag{35}$$

Under $do(S_t = s')$, the seller's offer mechanism is replaced by the specified constant. Buyer and seller characteristics remain inputs to the unchanged mechanisms; replacing the offer does not by itself identify buyer price elasticity.

Vocabulary. $|\mathcal{V}_{ebay}| = 146$ discrete tokens including negotiation units like: ($\langle$LIST_PRICE$\rangle$, $\langle$BUYER_OFFER$\rangle$, $\langle$SELLER_COUNTER$\rangle$); outcomes in the form of ($\langle$DEAL$\rangle$, $\langle$NO_DEAL$\rangle$); and 121 discrete offer price ratio bins $\langle$OFFER_000$\rangle$ through $\langle$OFFER_120$\rangle$ discretizing offer ratios Offer/ListPrice from 0% to 120%; and the rest are 16 latent persona bins $\langle$UNOBS_00$\rangle$ through $\langle$UNOBS_15$\rangle$.

Interventions.

`random_offer`: $S_1' \sim \mathcal{U}(0.60, 0.98) \cdot L$ to evaluate global out-of-distribution price elasticity.

`aggressive_counter`: $S_1' = B_1' + 0.05 \cdot L$ to simulate seller offering a generous concession just 5% above buyer's opening bid.

`rigid_counter`: $S_1' = 0.98 \cdot L$ to simulate when seller stubbornly anchors at 98% of list price.

`persona_mismatch`: To simulate deliberate misalignment shocks (e.g., injecting an aggressive concession on an impatient buyer or a rigid counter on a low-valuation buyer).

Further we also sweep counteroffers from from $0.60 \cdot L$ to $0.98 \cdot L$ *to measure monotonic decay of deal probability* $P(\textit{Deal})$.

Data Generation. Behavioral distributions are derived directly from empirical field transactions (Backus et al., 2020), containing millions of real eBay Best Offer transactions; the persona parameter ranges above are hand-specified to span the behavioral regimes that paper documents rather than estimated from it. Ground-truth counterfactual deal probabilities $P^*(\text{Deal}_2 = 1 \mid do(S_1'))$ are obtained by **Monte-Carlo averaging over repeated rollouts** of the alternating-offer simulator under the fixed buyer persona, holding the exogenous draws fixed across the twin worlds. This provides a Monte Carlo estimate under the specified simulator, with sampling error determined by the number of rollouts.

Dataset Size. 10,000 to 50,0000 sessions for S-,M-,L-sized models with max sequence length being 1024.

C.4.3 ICU SEPSIS

Motivation. Septic shock is one of the leading causes of mortality in Internsive Care Units (ICUs) worldwide (Rudd et al., 2020). Clinical management requires continuous observation and interventions in the form of intravenous fluids and vasopressors to restore Mean Arterial Pressure (MAP > 65mmHg), without inducing fatal conditions. However, observational Electronic Health Records (EHR) suffer from extreme confounding by indication, with clinicians prescribing high-dose vasopressors almost exclusively to patients in critical condition. Thus, models trained on merely these logs often spuriously learn that vasopressors predict mortality, and make false inferences about the drug harming the patient. Evaluating counterfactuals accurately requires one to make the mortality prediction based on the patient's underlying physiological frailty and not the external intervention.

SCM Formalization. $\mathcal{M}_{sepsis} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, where: $\mathbf{U} = \{U_F, U_P, \mathbf{N}_V\}$ which encode *patient frailty* $U_F \in [0, 1]$ encoding physiological frailty and organ reserve, *patient presentation* $U_P =$ encoding age, weight and other static conditions, and stochasticity in the vitals measurement at each time step $\mathbf{N}_V \sim \mathcal{N}(0, \mathbf{\Sigma})$; $\mathbf{V} = \{V_t, A_t, Y\}$ where vitals $V_t = (MAP_t, HR_t, SOFA_t)$, clinician treatment actions $A_t = (VASO_t, FLUID_t)$ and survival outcome $Y \in \{\text{SURVIVED}, \text{DECEASED}\}$ (terminal token) at the end of the 72-hour observation window;

$\mathcal{F}$:

$$V_t = f_V(V_{t-1}, A_{t-1}, U_F, \mathbf{N}_{V,t}) \tag{36}$$

$$A_t = f_A(V_t, U_P, U_F, \mathbf{N}_{A,t}) \tag{37}$$

$$Y = f_Y(V_{\leq T}, U_F) \tag{38}$$

In `CaLM`, when the intervention $do(A'_t = a')$ is performed, the incoming edges to A'_t are severed, isolating the true bio-physiological treatment effect.

Tokenization. $|\mathcal{V}_{sepsis}| = 43$ representing units like: structural markers (⟨PATIENT_CONTEXT⟩, ⟨VITALS⟩, ⟨ACTION⟩); MAP bins (⟨MAP_HYPO⟩ for < 65 mmHg, ⟨MAP_NORM⟩, ⟨MAP_ELEV⟩); Heart Rate bins (⟨HR_BRADY⟩, ⟨HR_NORM⟩, ⟨HR_TACHY⟩); SOFA organ failure bins (⟨SOFA_0_2⟩ to ⟨SOFA_9_PLUS⟩); vasopressor dosage levels ⟨VASO_0⟩ through ⟨VASO_3⟩; fluid boluses ⟨FLUID_0⟩ through ⟨FLUID_3⟩; terminal outcomes ⟨SURVIVED⟩, ⟨DECEASED⟩; and 16 latent frailty tokens ⟨FRAILTY_00⟩ through ⟨FRAILTY_15⟩.

Interventional Queries.

`vaso_escalation`: Performing acute vasopressor escalation ($do(\text{VASO} = 3)$) for septic shock reversal.

`vaso_withholding`: Performing conservative vasopressor withholding ($do(\text{VASO} = 0)$) to avoid ischemic necrosis and arrhythmia.

`fluid_bolus`: Performing high-volume crystalloid fluid resuscitation ($do(\text{FLUID} = 3)$) to simulate aggressive clinician intervention.

`balanced_shock_protocol`: Performing joint balanced protocol ($do(\text{VASO} = 2, \text{FLUID} = 2)$) to simulate a middle ground approach.

Further, we also sweep vasopressor doses $\{0, 1, 2, 3\}$ to reconstruct non-linear biophysical perfusion curves to understand how the models obey monotonicity laws.

Dataset Generation. Ground-truth trajectories are generated by a discrete-time hemodynamic simulator that advances MAP, HR and SOFA on an hourly grid via additive difference equations in treatment dose, latent frailty and per-patient drug sensitivity. Parameter ranges and the clinician behavior policy follow the ICU sepsis literature (Goldberger et al., 2000) in MIMIC-III and MIMIC-IV (Johnson et al., 2016) (Johnson et al., 2023), the Sepsis-3 consensus definitions (Singer et al., 2016) and are aligned with clinical protocols from the Surviving Sepsis Campaign (SSC 2021) (Evans et al., 2021). Thus, the data samples generated mimic a patient's electronic health record.

Dataset Size. 2000, 15,000, 42,000 records for S-,M-,L-sized models with max sequence length being 1024.

C.4.4 6-MAX POKER

Motivation. Multi-agent games with imperfect information are a great investigative tool to understand modern artificial agents' strategic reasoning and belief attribution skills. Particularly, in 6-Max No-Limit Texas Hold'em, players can only observe public betting sequences and community cards, whilst the opponents' private set and playing archetypes remain hidden. Asking counterfactual questions like *If Hero had executed an aggresive 3-bet squeeze bluff instead of folding, what is the likelihood that opponents would fold* is natural and often an essential skill for top poker players. Observed data often has extreme confounding between players' holdings and bets, thus making this a challenging environment to counterfactually reason on.

SCM Formalization. $\mathcal{M}_{poker} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, where: $\mathbf{U} = \{H_0, H_1, \ldots, H_5, U_0, U_1, \ldots U_5, C\}$ which encode *private cards (hole cards)* $H_i \in P(Deals)$, *strategic personas* $UiP =$ governing VPIP (voluntary put in pot), PFR (pre-flop raise) and 3-bet frequency, and aggression factor across sample Tight-Aggressive (TAG), Loose-Aggressive (LAG), Tight-Passive (Nit), and Calling Station (Fish) archetypes and community card deck runouts $C = (C_{\text{flop}}, C_{\text{turn}}, C_{\text{river}})$; $\mathbf{V} = \{\mathbf{A}, \mathbf{P}, Y\}$ where public betting actions $\mathbf{A} = (A_1, A_2, \ldots A_K)$ across streets $s \in \{\text{PREFLOP}, \text{FLOP}, \text{TURN}, \text{RIVER}\}$, pot balances $\mathbf{P}$, and showdown outcome $Y \in \{\text{HERO_WIN}, \text{OPP_WIN}, \text{POT_SPLIT}, \text{HERO_FOLD}, \text{ALL_FOLD_HERO_WINS}\}$;

$\mathcal{F}$:

$$A_{i,s} = f_A(\mathbf{A}_{<i,s}, H_i, U_i, P_{<i,s}) \tag{39}$$

$$Y = f_Y(H_0, \ldots, H_5, C, \mathbf{A}) \tag{40}$$

In `CaLM`, when the intervention $do(A'_{hero} = a')$ is performed, the incoming edges communicating the *hole cards* and *strategic persona* to A'_{hero} are severed, while all of the information of the opponents is preserved.

Tokenization. $|\mathcal{V}| = 51$ tokens: 4 special tokens ($\langle$PAD$\rangle$, $\langle$BOS$\rangle$, $\langle$EOS$\rangle$, $\langle$SEP$\rangle$); 6 table positions ($\langle$POS_UTG$\rangle$ to $\langle$POS_BB$\rangle$); 6 street markers ($\langle$PREFLOP$\rangle$ to $\langle$SHOWDOWN$\rangle$); 7 betting actions ($\langle$FOLD$\rangle$, $\langle$CHECK$\rangle$, $\langle$CALL$\rangle$, $\langle$BET_HALF$\rangle$, $\langle$BET_POT$\rangle$, $\langle$RAISE_3BET$\rangle$, $\langle$ALL_IN$\rangle$); 7 hand equity bins ($\langle$HAND_AIR$\rangle$ to $\langle$HAND_NUTS$\rangle$); 5 showdown outcomes; and 16 style tokens per seat.

Interventions.

`preflop_bluff_squeeze`: Hero squeeze-raises preflop ($do(\text{Hero} = \langle\text{RAISE_3BET}\rangle)$) holding weak equity ($< 35\%$, e.g., $\langle$HAND_AIR$\rangle$) facing an open-raise and flat-call.

`half_pot_cbet`: Hero fires a half-pot continuation bet on the flop ($do(\text{Hero} = \langle\text{BET_HALF}\rangle)$) as preflop aggressor.

`pot_control_check`: Hero checks behind or flat-calls ($do(\text{Hero} = \langle\text{CALL}\rangle)$) with medium showdown value.

`value_shove`: Hero commits remaining stack on polarized nut holdings ($do(\text{Hero} = \langle\text{ALL_IN}\rangle)$).

Dataset Generation. The generated trajectories are synthesized by a simulator for 6-Max No-Limit Texas Hold'em with archetype frequencies set to published aggregate statistics from PokerTracker multi-million online hand histories (PokerTracker Software, LLC, 2026). Multi-way action trees are produced by a behavioral policy combining equity thresholds, pot odds and the persona parameters, providing an account of a poker game.

Dataset Size. 2,100, 8,750, 24,500 games for S-,M-,L-sized models with max. sequence length being 1024.

C.4.5 DRIVING

Example 12. Accurate counterfactual prediction in multi-agent traffic is paramount for autonomous fleet safety certification under *ISO 26262* and *ISO 21448 (SOTIF)*. While training autonomous

driving agents and testing them for counterfactual scenarios it is imperative that they understand the environment, the surrounding driver personas, and driving conditions and ascertain safety hazards to implement correct decision making protocols. Naive transformers are often catastrophically off as depicted in Fig. 1, and predict *safe* outcomes when perhaps emergency hazard collision protocols or human intervention should be invoked. `CaLM` tends to be more faithful at reproducing the ground truth counterfactual trajectories and thus is a more reliable simulator. □

The trajectories and gap values Δ in Fig. 1 are manually constructed illustrations, not recorded model rollouts; red gap labels mark distances below $4\,\mathrm{m}$. Separately, panel (e) reports archived emergency-braking evaluation scores for Scenario 18: 0.350 for ground truth, 0.740 for the naive baseline, and 0.447 for `CaLM`. These safety scores are not probabilities of the illustrated trajectories.

Motivation. Autonomous vehicles operate in dense multi-agent traffic, making split-second, safety-critical decisions while interacting with heterogeneous human drivers. Extreme maneuvers like sudden braking or aggressive lane cuts are statistically rare. Furthermore, while evaluating hypothetical counterfactuals like *"What would the collision risk have been if the ego vehicle (the one being tested) had performed a brake check instead of cruising?*, an observational predictor conflates the action with the latent driver style, hallucinating reckless behavior or a phantom road hazard. Preserving physical laws and unobserved characteristics of the surroundings is necessary to make a justified prediction.

SCM Formalization. $\mathcal{M}_{driving} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, where: $\mathbf{U} = \{U_{EGO}, U_{TV}, \mathbf{N}\}$ which encode *driver behavioral styles* (Aggressive, Normal, Cautious, Distracted) $U_{TV} \sim P(U)$ parameterizing human reaction latency $\tau \in [0.3, 1.4]\,\mathrm{s}$, desired headway $T \in [0.8, 2.2]\,\mathrm{s}$, desired speed $v_0 \in [24, 34]\,\mathrm{m/s}$, maximum acceleration $a_{\max} \in [1.0, 3.0]\,\mathrm{m/s^2}$, and comfortable braking $b \in [1.5, 3.0]\,\mathrm{m/s^2}$. $\mathbf{N} = \{N_S, N_A, N_R\}$ denotes independent physical disturbances.; $\mathbf{V} = \{S_t, A_t, R_t, Y\}$ denoting *spatial traffic states* $S_t = (\mathrm{Lane}_t, \mathrm{Headway}_t, \mathrm{Speed}_t)$, *ego action choices* A_t, trailing vehicle kinematic *responses* R_t, and terminal *safety outcome* $Y \in \{\mathrm{SAFE}, \mathrm{NEAR_MISS}, \mathrm{REAR_COLLISION}, \mathrm{SIDESWIPE}\}$

$\mathcal{F}$:

$$S_t = f_S(S_{t-1}, A_{t-1}, R_{t-1}, N_{S,t}) \tag{41}$$

$$A_t = f_A(S_t, U_{\mathrm{EGO}}, N_{A,t}) \tag{42}$$

$$R_t = f_R(S_t, A_t, U_{\mathrm{TV}}, N_{R,t}) \tag{43}$$

$$Y = f_Y(S_{\leq T}, A_{\leq T}, R_{\leq T}) \tag{44}$$

In `CaLM`, when the intervention $do(A'_t = a')$ excises the incoming edges, the trailing vehicles, as they should, respond strictly according to kinematics without confounding the driver styles.

Tokenization. $|\mathcal{V}| = 49$ discrete tokens encompassing: 4 special tokens ($\langle$PAD$\rangle$, $\langle$BOS$\rangle$, $\langle$EOS$\rangle$, $\langle$SEP$\rangle$); 3 spatial lanes ($\langle$LANE_LEFT$\rangle$, $\langle$LANE_CENTER$\rangle$, $\langle$LANE_RIGHT$\rangle$); 5 longitudinal headway bins ($<5\,\mathrm{m}$ to $>45\,\mathrm{m}$); 5 speed bins ($<2\,\mathrm{m/s}$ to $>35\,\mathrm{m/s}$); 7 ego actions ($\langle$MAINTAIN$\rangle$, $\langle$ACCEL_MILD$\rangle$, $\langle$ACCEL_FULL$\rangle$, $\langle$DECEL_MILD$\rangle$, $\langle$EMERGENCY_BRAKE$\rangle$, $\langle$LANE_CHANGE_LEFT$\rangle$, $\langle$LANE_CHANGE_RIGHT$\rangle$); 5 trailing vehicle responses ($\langle$RESPONSE_MAINTAIN$\rangle$ to $\langle$RESPONSE_COLLISION$\rangle$); 4 outcomes ($\langle$SAFE_CONTINUATION$\rangle$, $\langle$NEAR_MISS$\rangle$, $\langle$COLLISION_REAR$\rangle$, $\langle$COLLISION_SIDESWIPE$\rangle$); and 16 driver styles ($\langle$STYLE_00$\rangle$ to $\langle$STYLE_15$\rangle$).

Interventions. At step $t = 5$ (2.5 s into Game 2), an exogenous maneuver $do(A'_{\mathrm{EGO}} = a')$ is injected across four safety-critical tactical regimes:

`emergency_brake_check`: $do(A = \langle\mathrm{EMERGENCY_BRAKE}\rangle)$ $(-6.5\,\mathrm{m/s^2})$, isolating trailing vehicle headway collapse and reaction delays.

`aggressive_cut_in`: $do(A = \langle\mathrm{LANE_CHANGE_LEFT}\rangle)$ into fast traffic flow ($v \approx 32\,\mathrm{m/s}$) with tight headway gap.

`yield_and_merge`: $do(A = \langle\mathrm{LANE_CHANGE_RIGHT}\rangle)$ into slow-lane flow ($v \approx 22\,\mathrm{m/s}$).

`overtake_accelerate`: $do(A = \langle\mathrm{ACCEL_FULL}\rangle)$ $(+2.5\,\mathrm{m/s^2})$ rapidly closing forward distance.

We also sweep braking deceleration in the set $\{1.0, 0, -2, -6.5\}\,\mathrm{m/s^2}$ to evaluate continuous Time-to-Collision (TTC) hazard monotonicity.

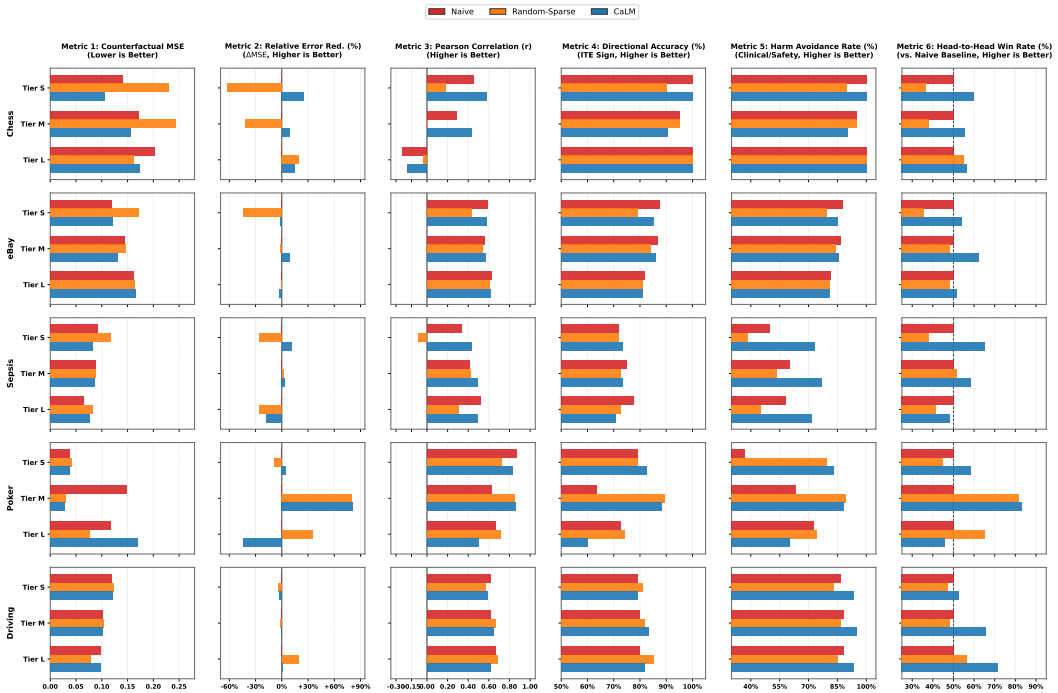

Figure 4: Overall Results across all scales and domains

Dataset Generation. Data is generated using a 3-lane highway simulator governing 7 interacting vehicles. Longitudinal kinematics follow the continuous *Intelligent Driver Model (IDM)* Treiber et al. (2000):

$$\dot{v} = a_{\max}\left[1 - \left(\frac{v}{v_0}\right)^4 - \left(\frac{s^*(v,\Delta v)}{s}\right)^2\right], \quad s^*(v,\Delta v) = s_0 + vT + \frac{v\Delta v}{2\sqrt{a_{\max}b}} \tag{45}$$

IDM parameter archetypes are chosen to span the aggressive/defensive range inspired from the (Alexiadis et al., 2004). Thus, each trajectory depicts a scenario of vehicles driving on a highway.

Dataset Size. 20,000, 70,000, 100,000 scenarios for S-,M-,L-sized models with max sequence length being 256.

D Additional Results and Ablations

D.1 Overall Results

Fig. 4 presents the comprehensive cross-scale benchmark suite across all five environments and model scales (Tier S, Tier M, and Tier L). Across most comparisons, `CaLM` exhibits robust advantages. In the duels, `CaLM` wins against the dense Naive baseline in 13 of the 15 domain $\times$ scale configurations.

D.2 Performance by Intervention Type (Tier S and Tier L)

Similar to Fig. 3, we present the other two scales here: **Tier S** in Fig. 5 and **Tier L** in Fig. 6.



Figure 5: Even at compact Tier S capacity, `CaLM` performs strongly in comparison to Naive and Random-Sparse Models.



Figure 6: Even at the largest scale we tested, CaLM performs reliably in most scenarios

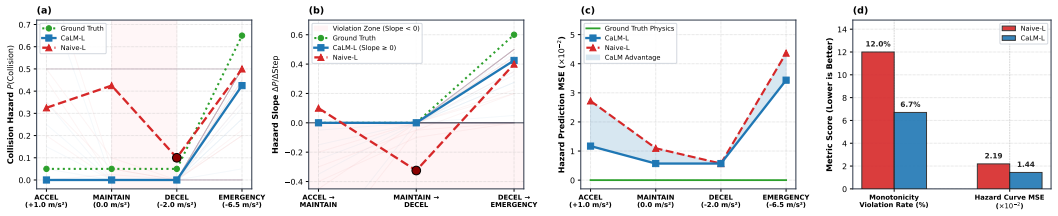

Figure 7: Case Study: Monotonicity Preservation in Driving (a) Collision hazard response across increasing braking severity; (b) Collision Hazard Slope across transitions; (c) Hazard prediction error; (d) Overall statistics

D.3 Monotonicity Testing

Here, we test monotonicity in driving by intervening with acceleration taking values from from $a = \{1, 0, -2, -6.5\}$ m/s^2. Predicted rear-end collision risk should increase as the distance between the vehicles decreases monotonically with increasing deceleration. The exact driving dynamics are defined by variables reaction times and distance between the drivers. In Fig. 7(a) Naive baseline exhibits an erratic non-monotonic dip (a 33% hazard drop) in this example, whereas CaLM gives a monotone response in the displayed case (faint background traces show all other scenarios). Fig.7(b) highlights Naive's steep plunge into the negative violation zone ($\Delta P = -0.325 < 0$, shaded pink), whereas the displayed CaLM response has non-negative slope $\Delta P \geq 0$, more closely mimicking the ground truth. Fig.7(c) and Fig.7(d) highlight overall statistics with CaLM demonstrating lower MSE across all the *acceleration*-related interventions. Naive baseline exhibits a **12**% violation, perhaps due to inaccurate assumptions about the driving style (e.g. say an aggressive driver instead of a distracted one) correlated with the external intervention. CaLM cuts the monotonicity violations in almost half.

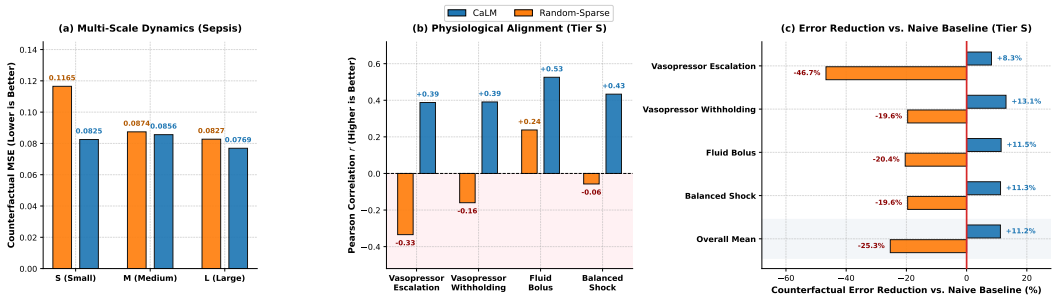

Figure 8: Comparing a Random-Sparse Model against CaLM.

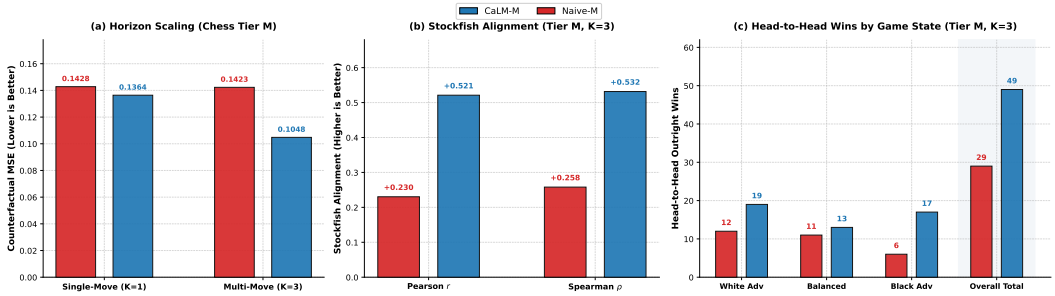

Figure 9: Impact of multi-move interactions on the performance of CaLM and Naive.

D.4 Sparsity Comparisons

A natural question to ask is if the gains for CaLM are arising from the causal masking or merely from the attention sparsity acting as a regularizer. To address this question definitively, we employ a trained-from-scratch random sparse model with identical training parameters. The number of attention connections severed in the random sparse model is exactly the same as in the CaLM model, but instead of implementing the graphical constraints, the masking is chosen uniformly at random. Fig.8(a) shows that CaLM consistently outperforms generic random sparsity across all scales, most prominently at the smallest scale. 8(b) highlights that random sparsity is at times harmful, and induces negative correlations with the ground truth. 8(c) re-iterates this point, showcasing the negative consequences of randomly-masked connections when compared to a dense transformer. Figs. 6, 5, 3, and 4 also plot Random-Sparse for comparisons in other domains, scales and interventions.

D.5 Longer Interventions

We also wanted to test how does the performance of CaLM and Naive models get affected when the interventions aren't an isolated single action but a sequence of actions. For example, in chess, intervening on k moves, say $t, t+1, \ldots, t+k-1$ compounds the tension with the unobserved playing style. To investigate horizontal scaling, we evaluated both of them on multi-move interventions ($k = 3$). The observations summarized in Fig. 9(a),(b) show a larger measured advantage for the three-move setting than for the single-move setting for CaLM. Further, Fig. 9(c) highlights how this advantage is more pronounced in game states in favor of a particular player. Note that, two evaluated horizons do not establish a general scaling law.

D.6 Position of Interventions

We thought another curious dimension to test is the temporal positioning of interventions. In chess, trajectories usually partition into three distinct strategic phases, we map roughly to the same buckets: (i) the *Opening* (≤ 20 half-moves), (ii) the *Middlegame* ($21 \rightarrow 50$ half-moves), and (iii) the *Endgame* (> 50 half-moves). We split the interventions into these three categories to understand the performance split by phase. Fig. 10 reveals that CaLM is slightly worse when predicting counterfactual

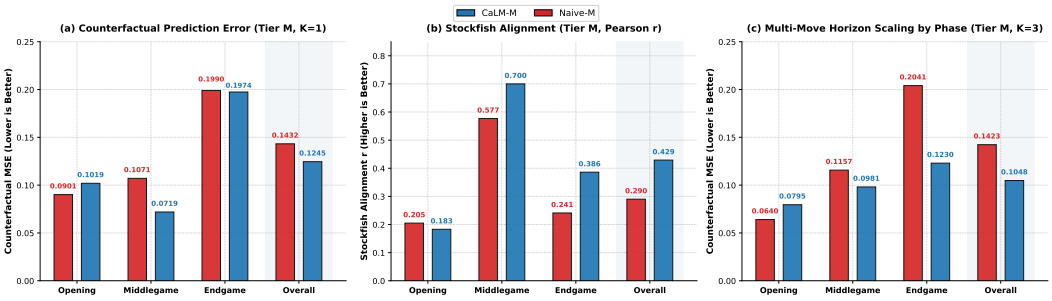

Figure 10: Effect of position of intervention on the performance.

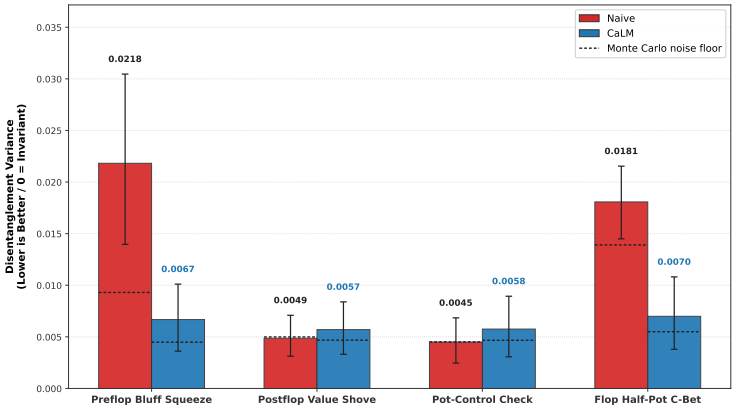

Figure 11: Prediction variance across factual contexts under a fixed intervention. Effectively, these are comparisons of sensitivity to the selected contexts. Lower variance indicates improved causal accuracy only if the target counterfactual probability is invariant across those contexts.

trajectories for interventions during the opening *Opening* but its advantage is clear in the *Middlegame* and *Endgame* scenarios.

D.6.1 Variation Across Factual Contexts

An intervention replaces the mechanism of the targeted variable and leaves the other mechanisms unchanged. It does not generally make the outcome independent of the exogenous variables that still affect it. For example, in the SCM $A = U, Y = A \oplus U$, intervening on A leaves Y dependent on U. Likewise, changing the factual evidence can change the posterior over the causes of a counterfactual outcome. Consequently, invariance across player skills, driver behavior, or patient frailty is not implied by the causal diagram alone.

For a fixed intervention $do(A = a')$ and outcome $Y = y$, let us define the predictions across m factual contexts by

$$\hat{p}_i = P^{\mathcal{S}_\theta}(Y_{A=a'} = y \mid \mathbf{E} = \mathbf{e}^{(i)}), \qquad \bar{p} = \frac{1}{m}\sum_{i=1}^{m}\hat{p}_i, \qquad \sigma^2 = \frac{1}{m}\sum_{i=1}^{m}(\hat{p}_i - \bar{p})^2. \tag{46}$$

Fig. 11 reports this variance across $m = 15$ rollouts for different interventions for poker. The relative reductions achieved by `CaLM` are 82% and 64% (w.r.t. the Monte Carlo noise floor) from Naive's non-floor variances. Interpreting that reduction as removal of spurious dependence requires verifying that the true counterfactual quantity is unchanged across the compared contexts; otherwise, some variation may reflect real differences in causal mechanisms or in the posterior over exogenous variables. The measurements alone do not establish counterfactual coverage or a guarantee of causal invariance.

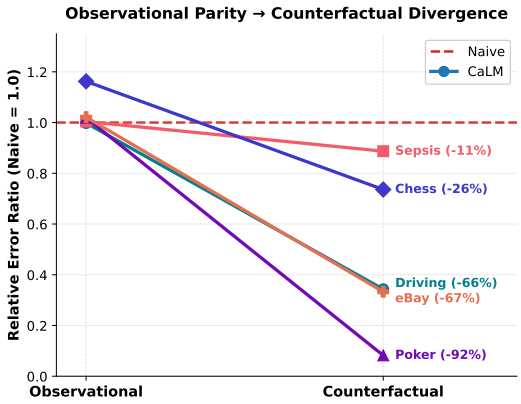

Figure 12: Observational Parity vs. Counterfactual Divergence. On passive observational trajectories (left), CaLM maintains near parity with Naive models. Targeted counterfactual intervention (right) causes catastrophic causal confusion in Naive models, while CaLM's causal graph surgery consistently unlocks dramatic error reductions across all evaluated domains and model scales and varied interventions: *Driving (M)*: Emergency Brake; *eBay (M)*: Rigid Counteroffer; *Poker (M)*: Preflop Bluff Squeeze; *Chess (M)*: Tactical Blunder Substitution; and *Sepsis (Lye)*: Vasopressor Escalation.

D.7 Observational Parity vs. Counterfactual Divergence

A foundational principle of Pearl's Causal Hierarchy is the asymmetry between Layer 1 (Observational Association, $P(Y = y' \mid A = a')$) and Layer 3 (Counterfactual, $P(Y = y'_{A=a'} \mid Y = y)$) Bareinboim et al. (2022). An essential desideratum for any counterfactual simulator is that enforcing graph constraints must not impair passive observational modeling; rather, the architecture should maintain parity on unconfounded sequences while decisively eliminating confounding under intervention.

To evaluate this trade-off, we compare CaLM and Naive autoregressive transformers across both passive observational tasks and targeted interventional queries and report the results in Fig. 12. Unconfounded passive pretraining validation loss is similar between CaLM and Naive, with CaLM being slightly worse in Chess, while in the confounded counterfactual setting CaLM decisively outperforms Naive (26.4% in Chess, 65.7% in Driving, 91.8% in Poker) which reveals that CaLM unlocks robust counterfactual reasoning, without a significant penalty on observational prediction ability.

D.8 Foundation Model Scaling

A central question for practical causal modeling is whether the benefits of CaLM demonstrated on models trained from scratch at various scales could also extend to instruction-tuned foundation models with rich linguistic properties. Pretrained LLMs often internalize strong observational correlations, and when tasked with evaluating counterfactual interventions fail to de-confound these biases.

To empirically investigate foundation model transfer, we scaled CaLM to two open-weight models, Llama-3-8B-Instruct (Grattafiori et al., 2024) and Mistral-7B-Instruct (Jiang et al., 2023). We generated a 100k sized health-record patient trajectory format dataset based on MIMIC-IV, generated using MedGemma (Sellergren et al., 2025), that the models were fine-tuned on. A sample generation is provided below:

Sample from Dataset

```
[PATIENT\_PROFILE]
70-year-old female admitted to the ICU with Gram-negative
bacteremia of unknown primary
source and septic shock. Initial presentation notable for
hypotension and signs of end-organ
hypoperfusion. Admission hemodynamics: MAP 52.3 mmHg,
HR 103.4 bpm. Initial SOFA score 8.
Admission lactate 4.18 mmol/L.
[/PATIENT\_PROFILE]

[HOUR\_1\_EXAM]
Patient remains hypotensive with persistent tachycardia.
MAP 57.8 mmHg, HR 107.9 bpm.
Lactate elevated at 4.61 mmol/L. Peripheral perfusion
poor with delayed capillary refill (>3s).
[/HOUR\_1\_EXAM]

[HOUR\_1\_ORDER]
Initiate Norepinephrine infusion at 0.08 mcg/kg/min; infuse
500 mL Lactated Ringer's bolus.
[/HOUR\_1\_ORDER]

...

[OUTCOME]
Patient stabilized by Hour 20 with resolution of
lactic acidosis and weaning of vasopressors.
[/OUTCOME]
```

Interventions tested were the same as in the trained-from-scratch models. Each patient trajectory couples granular natural-language clinical admission history, acute physiological vital signs governed, clinical interventions and the final outcomes. All models were trained with early stopping for 3 epochs on dedicated NVIDIA H100 80GB GPUs. Results are presented in Fig. 13, revealing that CaLM achieves notable error reductions across all protocols. CaLM consistently outperforms Naive in biophysical alignment across all four interventions on both model families. CaLM wins outright in the vast majority of patient cases. CaLM consistently exceeds Naive in predicting the true sign and direction of survival response. For example, on Llama-3-8B-Instruct, CaLM achieves a 41.0% MSE reduction margin is most dramatic in the *Fluid Bolus* interventions with 41.0% MSE reduction, and On Mistral-7B, CaLM cuts error on *Vasopressor Withholding* interventions by 22.7%. The H2H win-ratio on Llama-3-8B is $1.88\times$ and on Mistral-7B it is $1.66\times$. The false prediction of benefit when an intervention is physiologically detrimental CaLM demonstrates more than $2\times$ higher Harm Avoidance Rate than Naive which is a massive amount for any model deployed in high-stakes scenarios. These foundation-scale findings establish that CaLM's causal graph surgery principles are not limited to lightweight models.

D.9 Prediction Intervals under Weight Perturbations

We introduced varying scales of Gaussian noise ($\epsilon \in \{10^{-2}, 10^{-1}, 0.5\}$) to all billions of weights in the fine-tuned Llama-3-8B and Mistral-7B models. For each counterfactual query across the 400 MIMIC-IV evaluation trajectories, we sampled 10 random perturbations, pruning to the 5 with predictions closest to the ground truth probability to form a robustness bound $[\min, \max]$, and if they contain the ground truth probability.

These intervals use the known evaluation target to select models. The procedure does not require the perturbed models to satisfy the observational-fit constraint in Eq. 6, so it does not implement that optimization or inherit Thm. 6's guarantee. The following comparisons report inclusion, interval width, and distance to the interval for this perturbation procedure.

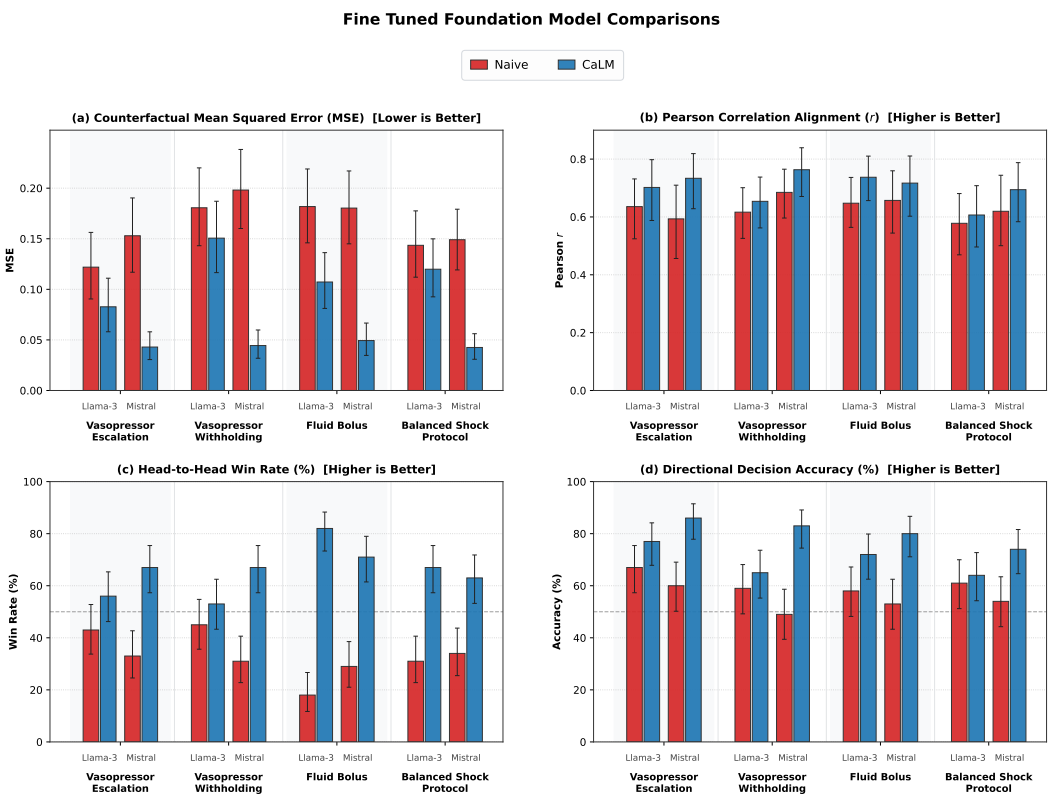

Figure 13: Comparing Fine-Tuned Models.

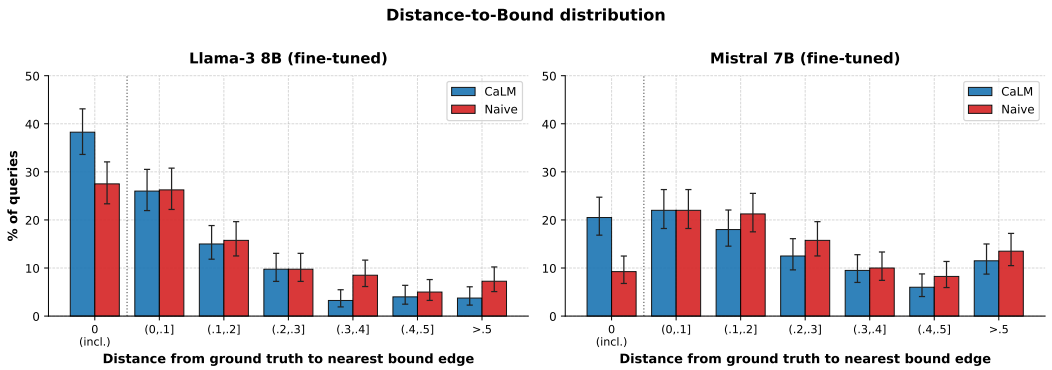

Figure 14: Distance-to-bound distribution under weight perturbation ($\epsilon = 10^{-1}$), fine-tuned foundation models. Each of the 400 seed-matched MIMIC-IV queries per arm is bucketed by the distance from the ground-truth counterfactual probability to the nearest edge of its perturbation interval; the leftmost bucket, left of the dotted divider, is exact inclusion. `CaLM` (blue) carries more mass at inclusion on both backbones, and Naive (red) carries at least as much mass in every bucket to the right of the divider; the distributions are ordered, not merely different in mean. Bounds are formed by sampling 10 weight perturbations and pruning to the 5 predictions closest to ground truth, so they are oracle-assisted reachability intervals rather than calibrated confidence intervals; Note: `CaLM`'s intervals are also the wider ones here (0.195 vs. 0.179 on Llama-3, 0.076 vs. 0.052 on Mistral), so inclusion rate and distance must be read together.

Fig. 14 shows that at low noise ($\epsilon = 10^{-2}$), `CaLM`'s perturbed bounds are systematically closer to the true counterfactual probability, exhibiting an 18.9% relative improvement over Naive on Llama-3 and a 7.7% improvement on Mistral. Further, at $\epsilon = 10^{-1}$, `CaLM` achieves a 2.22× inclusion rate ratio

Table 1: Perturbation Bounds (CaLM vs Naive on Llama-3-8B)

ϵ (Noise)	Inclusion Rate (%)		Avg Bound Width		Distance-to-Bound	
	CaLM	Naive	CaLM	Naive	CaLM	Naive
10^{-2}	**4.75**	3.00	0.020	0.016	**0.247**	0.304
10^{-1}	**38.25**	27.50	0.195	0.179	**0.111**	0.158
0.5	**84.50**	83.00	0.482	0.481	**0.012**	0.014

Table 2: Perturbation Bounds (CaLM vs Naive on Mistral-7B)

ϵ (Noise)	Inclusion Rate (%)		Avg Bound Width		Distance-to-Bound	
	CaLM	Naive	CaLM	Naive	CaLM	Naive
10^{-2}	0.75	**1.25**	0.008	0.005	**0.267**	0.290
10^{-1}	**20.50**	9.25	0.076	0.052	**0.200**	0.243
0.5	**74.75**	73.50	0.315	0.316	**0.021**	0.024

over Naive on Mistral-7B. CaLM's distance-to-bound advantage is 17.5% on Mistral, demonstrating that the causal attention structure provides even stronger *relative* robustness on a less well-calibrated backbone architecture.

D.10 Frontier Models

A natural question is whether the causal-graph surgery in CaLM meaningfully outperforms a closed-weight frontier model with only in-context calibration on the same benchmark. To evaluate this, we compared four frontier systems spanning two labs and two model generations: Claude Opus 4.7 (Anthropic, a), Claude Opus 5 (Anthropic, b), GPT-5.5 (OpenAI, b) and GPT-5.6-sol (OpenAI, a) against the fine-tuned CaLM-Llama-3-8B models on the identical stratified evaluation set. All models see the same twin trajectory text at query time, and none is shown the ground-truth terminal outcome for the query sample.

The comparison is only meaningful if the two systems are asked the same question in the same way, so every frontier model is evaluated by a literal replication of CaLM's own evaluation hook, in which the trajectory is appended with `\n\n[OUTCOME]\nClinical status:` and the model is forced to emit exactly one word (`Survived` or `Deceased`), bucketed to $\{1.0, 0.0\}$. The five few-shot exemplars are raw trajectories carrying no supervision, matching CaLM's pure language-modeling training signal and each prediction is a single forward pass.

GPT-5.5 is statistically indistinguishable from Opus 5 on every metric and the two models agree on 88.7% of the individual predictions and produce near-identical outcome distributions. That two independently developed frontier systems land on the same operating point suggests this is a property of in-context counterfactual inference on this benchmark rather than an idiosyncrasy of any one model family.

Fine-tuning with a causal-graph attention mask remains competitive with, and even superior to, four frontier closed-weight models, and it wins the per-sample head-to-head by a significant.

This suggests that frontier models are not necessarily reliable counterfactual simulators, and they can't be always be expected to out of the box accurately predict counterfactual outcomes despite the massive parameter size and training data they have been exposed to.

E Background and Related Work

E.1 Background

Autoregressive Transformers. Transformers combine multi-head self-attention and position-wise feed-forward layers with positional information, residual connections, and normalization (Vaswani

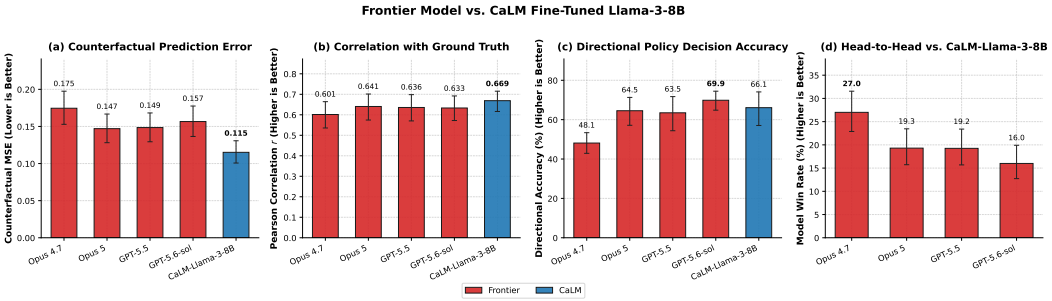

Figure 15: Frontier models vs. fine-tuned CaLM-Llama-3-8B under MIMIC-IV, 400 stratified samples. `CaLM` leads on both magnitude metrics, (a) MSE and (b) Pearson r. In (c), frontier directional accuracy rises with model generation until GPT-5.6-sol becomes comparable to `CaLM`. Panel (d) reports each frontier model's own per-sample win rate against `CaLM`: it falls monotonically across generations, reaching its lowest value.

et al., 2017). Let $\mathbf{Z} = (Z_1, \ldots, Z_L)$ denote an ordered sequence of L token variables, with realization $\mathbf{z} = (z_1, \ldots, z_L)$. At a given layer, the input representations form a matrix $\mathbf{H} \in \mathbb{R}^{L \times d}$, with one d-dimensional row per token position. A self-attention head computes queries $\mathbf{Q} = \mathbf{H}\mathbf{W}_Q$, keys $\mathbf{K} = \mathbf{H}\mathbf{W}_K$, and values $\mathbf{V} = \mathbf{H}\mathbf{W}_V$. Here, $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{d \times d_k}$ and $\mathbf{W}_V \in \mathbb{R}^{d \times d_v}$ are learned projection matrices, so $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{L \times d_k}$ and $\mathbf{V} \in \mathbb{R}^{L \times d_v}$. With an additive mask $\mathbf{M} \in (\mathbb{R} \cup \{-\infty\})^{L \times L}$, the head produces

$$\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}; \mathbf{M}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} + \mathbf{M}\right)\mathbf{V} \in \mathbb{R}^{L \times d_v}, \tag{47}$$

where softmax is applied row-wise. An entry $M_{ij} = 0$ allows position i to attend to position j, whereas $M_{ij} = -\infty$ gives that position zero attention weight. *Lower-triangular-masked self-attention* uses the mask

$$M_{ij} = \begin{cases} 0, & j \le i, \\ -\infty, & j > i. \end{cases} \tag{48}$$

Thus, each position can attend to itself and all preceding positions, but not to any subsequent position. With shifted prediction targets, the final representation at position i predicts Z_{i+1} from the observed prefix $z_{1:i}$, without access to z_{i+1} or later tokens. This mask enforces sequence order; it does not by itself specify the structural causal relationships among the token variables. The autoregressive model factorizes the sequence probability as

$$p_\theta(z_{1:L}) = \prod_{i=1}^{L} p_\theta(z_i \mid z_{<i}), \tag{49}$$

where the first factor is conditioned on a beginning-of-sequence token. Training minimizes the negative log-likelihood $-\sum_{i=1}^{L} \log p_\theta(z_i \mid z_{<i})$ of the observed sequence.

E.2 Related Work

Counterfactual simulation with world and language models. Autoregressive world models learn observational trajectory distributions and simulate future states by rolling forward under candidate action sequences (Janner et al., 2021; Micheli et al., 2022; Robine et al., 2023; Hu et al., 2023). More recent world models similarly emphasize hypothetical or "counterfactual" prediction as a capability for planning and interaction (Xiang et al., 2025; Assran et al., 2025). In language generation, counterfactual story rewriting conditions on an original narrative and a modified event and generates a revised continuation, often using supervised counterfactual rewrites (Qin et al., 2019); benchmarks such as CLADDER instead prompt language models with verbalized causal questions (Jin et al., 2023). These interfaces motivate the three baselines in Sec. 2: B0 prompts an instruction-following model with the factual trajectory and a counterfactual question, B1 replaces selected tokens and rolls forward from the modified prefix, and B2 retains the completed factual trajectory alongside a partial modified

copy. They therefore differ in how much factual evidence is available, but that distinction alone does not determine the semantics of the generated alternative. In particular, B1 discards a factual suffix that may reveal persistent latent factors, whereas B0 and B2 retain this evidence without requiring the model to interpret it as evidence about latent causes or to hold those causes fixed during generation. This is consequential under unobserved confounding: a player's skill, a patient's frailty, or an agent's intent may affect several decisions, so the complete factual trajectory changes what should be inferred about that factor, while the same inferred factor should be reused after the intervention.

Structured and graph-constrained Transformers. Transformer masks and structural encodings have been used to impose dependency patterns beyond ordinary left-to-right generation. For example, the Disentangled Context Transformer predicts tokens from varying subsets of context (Kasai et al., 2020), Graphormer incorporates graph topology through structural encodings (Ying et al., 2021), and the Simformer represents flexible conditionals for simulation-based inference with a Transformer diffusion model (Gloeckler et al., 2024). Closer to causal estimation, the Causal Transformer combines separate treatment, covariate, and outcome streams with balanced representations to estimate longitudinal potential outcomes (Melnychuk et al., 2022), while G-Transformer uses a Transformer implementation of longitudinal g-computation to simulate outcomes under dynamic treatment regimes (Xiong et al., 2024). These methods address time-varying treatment effects under the corresponding longitudinal identification assumptions; in particular, sequential ignorability rules out hidden treatment outcome confounding after conditioning on the observed history. Their target is therefore complementary to arbitrary joint and conditional Layer-3 queries in a supplied causal diagram with shared unobserved causes.

A hidden Transformer representation is not, by itself, an SCM exogenous variable: the latter has a specified role in the data-generating mechanisms, may be shared across several variables, is inferred from factual evidence, and is held fixed when the intervened world is evaluated. `CaLM` encodes these requirements explicitly. Moreover, a parent-only attention mask is insufficient in a multilayer Transformer because a contextualized parent representation can already contain information from nonparents. `CaLM` therefore isolates value tokens from contextual updates and restricts each mechanism's computation and readout at every layer, so that its output depends only on the designated endogenous parents and exogenous inputs (Sec. 3). The distinction from graph Transformers is thus not that they are incapable of causal reasoning, but that their structural encodings were designed for different dependency and prediction targets and do not by themselves define this SCM counterfactual semantics.

Counterfactual inference with causal models. Structural causal models distinguish an interventional distribution, which averages outcomes after replacing a mechanism, from a conditional or joint counterfactual, which additionally conditions on factual evidence and evaluates multiple worlds using the same exogenous assignment (Pearl, 2009; Bareinboim et al., 2022). Deep structural causal models and causal autoregressive flows make abduction and counterfactual prediction tractable using variational or invertible neural mechanisms (Rasal et al., 2022; Khemakhem et al., 2021); these formulations use specified, predominantly Markovian exogenous structures, and the former explicitly excludes unobserved confounding. This differs from semi-Markovian settings, where a shared exogenous cause can induce bidirected edges and persistent confounding across several observed variables.

E.3 Relation to Existing Causal Models

Canonical causal models (CCMs) provide a finite representation of the structural functions and exogenous distributions needed to preserve all counterfactual distributions in a given graph (Zhang et al., 2022). They express counterfactual bounds as constrained optimization over these finite mechanisms and exogenous probabilities. NCMs represent the mechanisms by neural networks, retaining the graph's parent restrictions and shared exogenous inputs, and use neural optimization to estimate counterfactual quantities and test identification (Xia et al., 2023). `CaLM` represents the mechanisms with a shared Transformer. Its attention restrictions and finite-function representation permit the same soundness and expressiveness arguments to apply to sequence generation.

Gumbel-max SCMs make categorical sampling explicit as a deterministic function of its parents and exogenous Gumbel variables (Oberst & Sontag, 2019). For a categorical kernel $p(\cdot \mid \mathbf{pa}_{Z_j})$, the

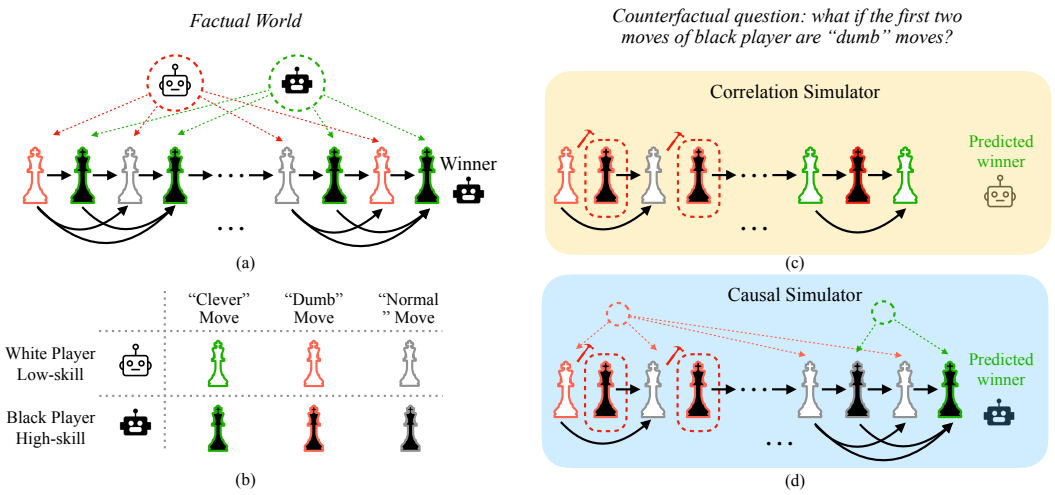

Figure 16: A chess example contrasting correlation-based and causal counterfactual rollouts when Black's first two moves are replaced by "dumb" moves.

mechanism takes the form

$$Z_j = \arg\max_{a \in \mathcal{X}_{Z_j}} \{\log p(a \mid \mathbf{pa}_{Z_j}) + G_{j,a}\}, \qquad G_{j,a} \overset{\text{iid}}{\sim} \text{Gumbel}(0,1). \tag{50}$$

The Gumbel variables are part of $\mathbf{U}$: abduction conditions on the observed outcome, and every intervened world reuses the inferred variables. Lorberbom et al. (2021) study mechanisms as implicit couplings between outcomes under different interventions and learn generalized Gumbel-max mechanisms while preserving prescribed interventional marginals. They replace argmax operations by softmax relaxations for gradient-based training (Lorberbom et al., 2021, Sec. 5.3). This illustrates why a differentiable training procedure and deterministic structural evaluation must be specified separately.

In `CaLM`, temperature annealing connects the training probabilities to a greedy mechanism whose exogenous inputs already account for its randomness. A fixed exogenous assignment therefore determines both the factual trajectory and every counterfactual trajectory, enabling abduction without additional output sampling. Selecting one mechanism by a criterion such as treatment-effect variance, as in Lorberbom et al. (2021), does not establish coverage over all SCMs compatible with the observations. Our coverage result uses the expressive fitted class supplied by the CCM and NCM constructions; temperature annealing addresses the agreement between training and structural evaluation within that class.

F Examples

F.1 Motivation

In addition to the driving setting in Example 1, we provide a chess example to further motivate counterfactual trajectory simulation.

Example 13. Consider two agents, White and Black, playing chess, as shown in Fig. 16(a). At each turn, a player's current decision depends on both players' previous moves. Each player's skill (often unobserved) also influences their decisions throughout the game, creating a shared unobserved cause of that player's moves across time. In this example, Black is the more skilled player and tends to choose more "clever" moves (shown in green in Fig. 16), while White tends to choose more "dumb" moves (shown in red). These correlations create a pattern in the recorded games: a player's "dumb" move and the opponent's "clever" response are likely to be followed by another "dumb" move from the same player. The causal relationships are illustrated in Fig. 16(a), and the corresponding symbols are shown in Fig. 16(b).

Now consider the counterfactual question: given the factual trajectory, how would the game have continued had Black's first two moves been replaced by "dumb" moves? A correlation-based sequence

simulator, such as one trained with naive next-token prediction, may predict further "dumb" moves by Black and more "clever" moves by White because of correlations in the observational training data (Fig. 16(c)). This prediction incorrectly relies on those correlations without accounting for the players' skills as unobserved confounders when generating the counterfactual trajectory. A causal simulator should instead use the factual trajectory to infer the confounder (Black's skill) and preserve that inference when predicting subsequent moves. Thus, even when Black's first two moves are replaced by "dumb" moves, the inference that Black is highly skilled leads to predictions of subsequent "clever" moves (Fig. 16(d)). In this example, the causal simulator more closely matches the ground truth, producing a different counterfactual rollout and predicting a different winner. □

F.2 SCM AND COUNTERFACTUALS

Example 14 (SCM for Example 1)**.** Consider an SCM $\mathcal{M}^* = \langle \mathbf{U}^*, \mathbf{Z}, \mathcal{F}^*, P^*(\mathbf{U}^*) \rangle$ for the driving illustration, with ordered endogenous sequence $\mathbf{Z} = (E_1, T_1, \ldots, E_m, T_m, R)$. Here, E_j and T_j denote the braking actions of Ego and TV at frame j, respectively, and R is a binary terminal safety outcome ($R = 1$ denotes a safe continuation and $R = 0$ denotes a hazard, including a near miss or collision). For this illustration, $E_j = 1$ denotes emergency braking by Ego, while $T_j = 1$ denotes braking by TV; a value of 0 denotes the absence of the corresponding braking action. Let U_E and U_T denote the drivers' unobserved personas, and let U_j^E and U_j^T capture additional independent exogenous variation in their actions. For fixed initial traffic conditions and deterministic vehicle dynamics, the traffic states can be recovered from the action history, giving the structural mechanisms

$$E_j = f_j^E(E_{<j}, T_{<j}, U_E, U_j^E), \tag{51}$$

$$T_j = f_j^T(E_{\leq j}, T_{<j}, U_T, U_j^T), \tag{52}$$

$$R = g(E_{1:m}, T_{1:m}). \tag{53}$$

Each persona is shared across the corresponding driver's action mechanisms, inducing paths such as $E_1 \leftarrow U_E \rightarrow E_3$ and $T_1 \leftarrow U_T \rightarrow T_3$, as illustrated in Fig. 1(b). In the setting of Example 1, U_E corresponds to a normal persona and U_T to a distracted persona whose response mechanism delays braking after Ego decelerates.

For a counterfactual example, let k denote the frame corresponding to $t = 2.5\,\mathrm{s}$ and consider the intervention $do(E_k = 1)$. This intervention replaces the mechanism for E_k with the constant 1, leaving the other mechanisms and the exogenous assignment unchanged. Given a factual trajectory $\mathbf{Z} = \mathbf{z}$ of positive probability, the corresponding counterfactual trajectory distribution is $P^{\mathcal{M}^*}(\mathbf{Z}_{E_k=1} \mid \mathbf{Z} = \mathbf{z})$. The question "Given the factual trajectory, what is the probability that traffic would have remained safe had Ego applied the emergency brake at frame k?" is expressed by $P^{\mathcal{M}^*}(Q) = P^{\mathcal{M}^*}(R_{E_k=1} = 1 \mid \mathbf{Z} = \mathbf{z})$. □

Example 15 (Sequential SCM for Example 13)**.** Consider the sequential SCM $\mathcal{M}^* = \langle \mathbf{U}^*, \mathbf{Z}, \mathcal{F}^*, P^*(\mathbf{U}^*) \rangle$ inducing the ordered endogenous sequence $\mathbf{Z} = (W_1, B_1, \ldots, W_m, B_m, R)$, where R is a binary variable denoting the terminal game outcome ($R = 1$ implies Black wins and $R = 0$ implies White wins). Let S_W and S_B denote the unobserved skills of White and Black, and let U_t^W and U_t^B capture additional independent exogenous variation in their decisions at turn t. For a fixed initial board, the structural mechanisms in $\mathcal{F}^*$ are

$$W_t = f_t^W(W_{<t}, B_{<t}, S_W, U_t^W), \tag{54}$$

$$B_t = f_t^B(W_{\leq t}, B_{<t}, S_B, U_t^B), \tag{55}$$

$$R = g(W_{1:m}, B_{1:m}). \tag{56}$$

The induced causal diagram is shown in Fig. 16. To illustrate, White's move W_t is determined by the history $(W_{<t}, B_{<t})$, White's skill S_W, and the turn-specific exogenous variable U_t^W. Black's move B_t similarly depends on $(W_{\leq t}, B_{<t})$, Black's skill S_B, and U_t^B. Importantly, Black's skill is a shared cause of Black's actions across time, as represented by paths such as $B_1 \leftarrow S_B \rightarrow B_3$; the same applies to White. Consequently, Black's moves at different turns may remain dependent even after conditioning on the preceding move history, because Black's skill is a common cause influencing decisions across turns.

For a counterfactual example, consider intervening on Black's first two moves $\mathbf{X} = \{B_1, B_2\}$, with replacement values $\mathbf{x}' = (b_1', b_2')$ chosen to be legal in the intervened trajectory. The question "Given the first j observed sequence elements, what moves would Black and White make had Black's first two moves been replaced by $\mathbf{x}'$?" asks for the counterfactual trajectory distribution $P^{\mathcal{M}^*}(\mathbf{Z}_{\mathbf{x}'} \mid \mathbf{Z}_{1:j} = \mathbf{z}_{1:j})$. One may also ask about the outcome: "Given the completed game, what is the probability that Black would have won had Black's first two moves been replaced by $\mathbf{x}'$?" Under the outcome convention above, this question is expressed by the conditional counterfactual probability $P^*(Q) = P^{\mathcal{M}^*}(R_{\mathbf{x}'} = 1 \mid \mathbf{Z} = \mathbf{z})$. □

F.3 Transformer Inference Procedures

Consider the query $P(Q) = P(\mathbf{Z}_{Z_j = z_j'} = \mathbf{z}' \mid \mathbf{Z} = \mathbf{z})$, where $\mathbf{z}$ is the factual trajectory and z_j' is the replacement value at position j. The generated trajectory $\mathbf{Z}'$ retains the factual prefix $\mathbf{z}'_{<j} = \mathbf{z}_{<j}$ and the replacement token z_j'. The following procedures define candidate counterfactual estimates, not necessarily the ground-truth counterfactual probabilities.

B0: Prompt-based simulation. The model receives the factual trajectory $\mathbf{z}$ and $[\mathrm{Instruction}]$, which specifies j, z_j', and a request to generate a counterfactual continuation while retaining the factual prefix. Its responses induce

$$P^{\mathcal{S}_{\mathrm{TF}}^{(0)}}(Q) = P^{\mathcal{S}_{\mathrm{TF}}^{(0)}}(\mathbf{Z}'_{>j} = \mathbf{z}'_{>j} \mid [\mathrm{Instruction}], \mathbf{Z} = \mathbf{z}). \tag{57}$$

This procedure applies to instruction-following language models, including large models accessed through an API. For example, the prompt supplies a completed chess game and asks: "Had Black played b_1' instead of b_1, how would the game have continued? Keep White's first move unchanged." Here, $j = 2$ in the ordering $(W_1, B_1, \ldots, W_m, B_m, R)$, and b_1' is a legal alternative move.

B1: Single-trajectory replacement and rollout. A Transformer trained on individual observed trajectories receives only the modified prefix $(\mathbf{z}_{<j}, z_j')$ and samples the continuation autoregressively:

$$P^{\mathcal{S}_{\mathrm{TF}}^{(1)}}(Q) = P^{\mathcal{S}_{\mathrm{TF}}^{(1)}}(\mathbf{Z}_{>j} = \mathbf{z}'_{>j} \mid \mathbf{Z}_{<j} = \mathbf{z}_{<j}, Z_j = z_j'). \tag{58}$$

For the same chess intervention, the input is (w_1, b_1'), and the model generates the remaining moves and outcome without access to the original factual suffix.

B2: Paired-trajectory replacement and rollout. A Transformer is trained on pairs of observed trajectories separated by $\langle \mathrm{SEP} \rangle$; both segments are observational during training. At inference, the first segment contains the completed factual trajectory, and the second contains its modified prefix, giving

$$\begin{aligned} P^{\mathcal{S}_{\mathrm{TF}}^{(2)}}(Q) = P^{\mathcal{S}_{\mathrm{TF}}^{(2)}}\Big(&\mathbf{Z}_{L+j+1:2L} = \mathbf{z}'_{>j} \mid \mathbf{Z}_{1:L} = \mathbf{z}, \langle \mathrm{SEP} \rangle, \\ &\mathbf{Z}_{L+1:L+j-1} = \mathbf{z}_{<j}, Z_{L+j} = z_j'\Big). \end{aligned} \tag{59}$$

For example, the input is $[\mathbf{z}; \langle \mathrm{SEP} \rangle; w_1; b_1']$, and the model completes the second game while retaining the original completed game as context. Possible constructions of the observational training pairs are discussed in Appendix A.1.1.