Learning DiscoSCM under One Factual Event per Token: Stable Identity, Learnability Boundaries, and Semantic Tests

Anonymous Author(s)

Working paper, July 17, 2026
Paper version: LDSCM-WP-2026-07-17-V4

Abstract

How can a Distribution-Consistency Structural Causal Model (DiscoSCM) be learned when each token contributes only one factual outcome? We study the smallest slice that retains the difficulty: a fixed-context, token-modulated outcome mechanism. The key primitive is not an ordinary predictive latent. A Unit Selection Variable names the stable object whose realization must be held fixed across queries. We distinguish the target-population law P_U^{tar}, the actual realization $U = u_i$, and the learner's evidence-indexed belief $Q_i(du) \approx P_{\mathrm{tar}}(U \in du \mid O = o_i)$.

This distinction yields a Perception–Abduction–Reduction contract. A factual record contains admissible token evidence O_i, an observed mechanism query x_i^F, and its outcome y_i^F. Perception maps O_i to an information-preserving or sufficient representation; abduction forms Q_i once; query perception maps a factual or alternative x^Q to the response geometry; and reduction integrates a shared token-modulated mechanism against the stored Q_i. Assignment, sample inclusion, and hypothetical query values are not token evidence by default. Point, Gaussian, and Cauchy abduction are learner-side uncertainty choices, not ontology. A coordinatewise Cauchy belief combined with a mechanism affine in the token code gives an exact Cauchy predictive law without latent Monte Carlo.

We prove Fisher consistency only for the composed outcome law on observed factual support. We also establish latent-reparameterization equivalence, token/event scale non-identifiability, and underdetermination of same-token alternative-query laws. Thus fixed abduction is an identity-preserving interface, not a token-recovery or counterfactual-identification theorem.

The evidence is correspondingly diagnostic. A frozen 1,400-cell E01 confirmatory benchmark does not confirm a Unit-factorization robustness advantage: at the primary endpoint the mean log error ratio is -0.00715 with a 95-percent interval $[-0.0496, 0.0353]$. A selected-token diagnostic then shows that evidence-to-unit abduction can approach its finite-test reference when the true mechanism is frozen, while the jointly learned one-shot factorization fails clean recovery. Masking a randomized factual action from the abductor reduces matched fixed-token curve error by 86.3 percent and makes the cached belief exactly action-invariant, but the control gain is larger and the learned Cauchy belief is strongly underdispersed. These results support the semantic interface and localize its learning bottleneck; they do not establish calibrated Q_i, structural superiority, or Layer 3 identification.

1 Introduction: How to Learn DiscoSCM?

Distribution-Consistency Structural Causal Models (DiscoSCMs) distinguish a stable unit from the event-level randomness that remains after the unit and a query are fixed. This supplies a language for token-level generation and for cross-world questions whose answers need not be deterministic at the unit level [3]. It also creates an immediate learning problem:

If DiscoSCM is accepted as a causal modeling theory, what is the first object that data can actually teach a learner?

A complete answer would have to learn or validate a graph, multiple structural mechanisms, a target-population unit law, event-noise laws, distribution-consistency, and the coupling used for joint cross-world queries. We do not compress these tasks into one theorem. We study the first slice that retains the theory's distinctive pressure: a fixed context, one token-modulated outcome mechanism, and one factual outcome per token. The world-side equation is

$$Y(x) = f_0(x, E; U), \tag{1}$$

where U is the single model-level Unit Selection Variable and E is event-level randomness.

The selector must be kept distinct from both its population law and a learner's uncertainty. A declared target population has composition P_U^{tar}; focal record i has one actual realization $U = u_i$; and admissible factual evidence $O_i = o_i$ induces a learner-side belief

$$Q_i(du) := q_\phi(U \in du \mid O = o_i) \approx P_{\mathrm{tar}}(U \in du \mid O = o_i). \tag{2}$$

Candidate draws from Q_i are hypotheses about the same fixed token, not new physical units. The approximation in Equation (2) is a semantic calibration target, not a consequence of predictive fit.

This separation changes the learning interface. A factual record is written $(O_i, X_i^F, Y_i^F) = (o_i, x_i, y_i)$. The evidence O_i must be available at query time and admissible for learning about token identity or response type; X_i^F is the factual mechanism query at which the outcome was observed. Treatment assignment, sample inclusion, and a hypothetical query value are not token evidence by default. The resulting computation has two branches that meet only at reduction:

$$\begin{aligned} O_i &\xrightarrow{P_F} z_i^F \xrightarrow{\mathcal{A}_\phi} Q_i, \\ x^Q &\xrightarrow{P_Q} z^Q, \\ (Q_i, z^Q) &\xrightarrow{\mathcal{R}_\theta} \mathcal{L}_i(dy; x^Q). \end{aligned} \tag{3}$$

For a distributional learner,

$$p_{\phi,\theta}(y \mid o_i, x^Q) = \int p_\theta\{y \mid P_Q(x^Q), u\}\, Q_i(du). \tag{4}$$

Perception may change evidence or query geometry; abduction forms one belief; reduction holds that belief fixed while the query changes. This Perception–Abduction–Reduction contract is more fundamental than a particular neural architecture or uncertainty family.

One factual event per token exposes an essential limit. A proper factual score can supervise Equation (4) only on the observed support of (O, X^F). It cannot by itself identify the actual u_i, calibrate Q_i, select a unique perception–abduction–mechanism factorization, allocate uncertainty uniquely between token and event, or determine the same token's law at an unobserved query. We make this boundary a central result rather than a qualification appended after the algorithm.

Our theorem-bearing specialization uses coordinatewise Cauchy abduction and a response mechanism affine in the candidate token code. Query perception and the coefficient functions may be nonlinear; the exact bilinear model studied first is the identity-perception special case. Cauchy is motivated by explicit desiderata—global openness, no forced finite-moment commitment, location–scale semantics, and exact affine propagation—but is not claimed to be uniquely correct.

The experiments progressively sharpen the learning claim rather than merely rank architectures. E01 tests an evidence-indexed *composed* response law and does not confirm a robustness advantage

over a capacity-matched direct predictor. Selected-token diagnostics then fix the simulator realization across a hidden query grid: abduction approaches its finite-test reference when the mechanism is frozen, but joint one-shot factorization fails clean recovery. An action-invariant encoder repair greatly improves hidden curves yet fails semantic-specificity and belief-calibration guards. Thus the evidence supports the interface discipline while exposing the unresolved learning bottleneck.

Our contributions are:

1. We formulate the first learning slice of DiscoSCM and distinguish target-population composition, actual token realization, and focal-token belief.
2. We derive the Perception–Abduction–Reduction contract, separating admissible factual evidence from mechanism queries and requiring one stored Q_i across same-token queries.
3. We give point, Gaussian, and Cauchy beliefs a common learner-side semantics and derive an exact Cauchy law for mechanisms affine in the token code.
4. We prove factual-support Fisher consistency and three negative boundaries: latent reparameterization, token/event scale non-identifiability, and alternative-query underdetermination.
5. We separate composed-law and fixed-selected-token evaluation targets, execute confirmatory and diagnostic protocols, and report their negative and boundary-sharpening results without promoting them to causal identification.

This paper has two active sibling viewpoints. *Unit Mechanism Learning* abstracts the general shared, unit-modulated learning problem without requiring DiscoSCM semantics. *Token Causation* studies the causal ontology and singular-attribution boundary. The present paper is the end-to-end DiscoSCM learning viewpoint; the three papers do not inherit one another's theorems or experimental claims automatically.

2 The First Learnable Slice: One Outcome Mechanism

Let $\mathcal{O}$, $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{E}$ be the admissible-evidence, query, outcome, and event-noise spaces. We fix a context c_0 and suppress it. The world-side outcome mechanism is

$$Y(x) = f_0(x, E; U). \tag{5}$$

For focal record i, the single model-level selector realizes as $U = u_i$. When the factual query is x_i and event noise realizes as e_i, the outcome is

$$y_i = f_0(x_i, e_i; u_i). \tag{6}$$

The subscript zero marks world truth. A learner uses a parameterized family f_θ to approximate the response law induced by f_0. Correct specification may posit $f_0 = f_{\theta_0}$, but the observation regime does not thereby identify θ_0, u_i, or a unique latent factorization.

The sample consists of many focal records

$$D_i = (O_i, X_i^F, Y_i^F) = (o_i, x_i, y_i), \tag{7}$$

each contributed by a different modeled token and containing one factual outcome. O_i is the declared view admissible for learning about the focal token; X_i^F is the mechanism query that actually occurred. The factual outcome Y_i^F supervises population-level parameters during training

but is not an input to the focal-token abduction object in the present slice. Thus our use of *abduction* is response-blind at focal query time; it targets $P(U \mid O)$ rather than an ex-post posterior such as $P(U, E \mid O, X^F, Y^F)$.

A record index is bookkeeping. It belongs to observations, realizations, and beliefs, but it does not create another Unit Selection Variable or enter the mechanism as identity. Every row instantiates the same selector role as $U = u_i$.

This slice intentionally excludes graph discovery, multiple endogenous nodes, context variation, and the general construction of joint counterfactual laws. It nevertheless retains five questions that a learning theory must answer: which token owns the outcome, which factual variables are legal evidence about that token, how a latent token becomes computable, how a new query remains linked to it, and which target is supervised by one-shot data. A theory that cannot answer these questions for one outcome mechanism has not yet learned a complete SCM.

We distinguish four targets throughout:

Factual composed law. The observable conditional law at the realized record, $p(y \mid o_i, x_i)$, after learner-side token uncertainty and event randomness are combined. This is the only target directly supervised by the primary factual score.

Evidence-indexed composed response law. For a changed query x^Q, the learner integrates the response mechanism against the fixed focal belief Q_i. E01 provides evaluator-only synthetic truth for this object.

Fixed-selected-token response law. The simulator conditions on the actual realization, $\mathcal{L}_i^\star(dy; x^Q) = P_\star\{Y(x^Q) \in dy \mid U = u_i\}$. E05, E06, and the perception diagnostic expose this object only inside a synthetic truth vault.

Causal alternative-query law. A marginal or joint interventional law that additionally requires assignment, support, stability, distribution-consistency, and, for joint worlds, an event-noise coupling.

These targets must not be pooled. Success or failure on an evidence-indexed mixture is not recovery of the actual selected-token curve, and neither synthetic target is automatically an identified real-world counterfactual.

3 Unit Selection as an Irreducible Primitive

3.1 What disappears if the selector is deleted?

Suppose Equation (5) were replaced by $Y(x) = f_0(x, E)$. With the same x and event-noise law, the formal language would contain no stable object through which two tokens could have different response mechanisms. Persistent between-token variation would have to be absorbed into E, conflating heterogeneity with repeated-event randomness.

Retaining U permits the token-indexed slice

$$Y_u(x) = f_0(x, E; u), \tag{8}$$

where the family is shared even though its instances differ. The same u can also be held fixed across x and x'. Event-noise coupling cannot replace this role: E records what happened in an event, whereas U records which token is being queried.

This yields a semantic necessity statement, not a naming convention: any theory that represents stable unit heterogeneity, unit-indexed response mechanisms, and same-token cross-query linkage

needs a primitive serving the role of a Unit Selection Variable. It need not be a neural-network input and it is not a renamed row identifier.

The word *selection* has a narrow namespace here: U selects token identity/code. It is not a treatment or action A/T, a sample-inclusion indicator S^{obs}, or an assignment or sampling propensity. When a population law P_U is introduced, it describes token composition in a declared target population. By contrast, Q_i is the evidence-indexed belief about the focal token in record i.

3.2 Population composition, actual realization, and focal belief

The three objects are formally distinct. A declared target population first supplies a reference law P_U^{tar}. Record i then has one unknown actual realization $U = u_i$. After observing admissible evidence $O_i = o_i$, a learner forms the probability measure

$$Q_i(B) := q_\phi(U \in B \mid O = o_i) \approx P_{\text{tar}}(U \in B \mid O = o_i), \qquad B \subseteq \mathcal{U} \text{ measurable.} \tag{9}$$

The right-hand side is the intended calibration target relative to the declared population; it is not identified merely because the composed outcome law fits well. Nor does Equation (9) force an explicit prior-times-likelihood implementation: $\mathcal{A}_\phi$ may be a discriminative abductor.

The direction is reference law plus focal evidence to focal belief. Abduction does not create a new target population and does not resample actual identity. Multiple candidates from Q_i are epistemic hypotheses about the same fixed u_i. If the observed sample is selected, the immediately learnable object may instead be $P_{\text{obs}}(U \in \cdot \mid O = o_i, S^{\text{obs}} = 1)$; transport to P_{tar} then needs an explicit selection, support, and invariance argument.

3.3 Discrete token and computational code

The selector first ranges over a semantic token space. Write the selected token as $t^\star \in \mathcal{U}_{\text{token}}$. A shared numerical mechanism requires a code, so we introduce an identity-to-code map

$$\iota : \mathcal{U}_{\text{token}} \longrightarrow \mathbb{R}^d, \qquad u^\star = \iota(t^\star). \tag{10}$$

We assume ι is injective: a code addresses at most one token, and every modeled token has a code. This is an identity-encoding modeling assumption, not a conclusion from one-shot observations. After making the layers explicit, we use the compressed notation $U = u^\star$ when no confusion can arise.

Injectivity of ι does not imply injectivity of the response map

$$u \longmapsto \{p_0(\cdot \mid x, u) : x \in \mathcal{X}\}. \tag{11}$$

Distinct token codes may induce identical response laws on the query domain. Nor does injectivity give numerical meaning to a particular coordinate system: an invertible recoding can preserve every observable prediction. These facts will reappear as an observational-equivalence theorem.

Thus token identity/code and response-equivalence class are distinct objects. Identifying a response class would not by itself identify the physical token.

The token/embedding analogy to language models is therefore limited but useful. A token first specifies discrete identity; an embedding permits shared computation. Here the token can be a person, device, patient state, or other response-relevant instance rather than a text symbol.

4 Perception, Abduction, and Reduction

4.1 Two information branches meet at response reduction

The world uses the actual u_i without uncertainty, but the learner never receives u_i as a label. It receives the factual record in Equation (7). The first task is therefore to declare which view O_i is legal evidence about the token. A factual treatment, action, or dose may be present in the record while remaining irrelevant or inadmissible for this purpose.

Evidence perception maps that view to a computational representation,

$$Z_i^F = P_F(O_i), \qquad z_i^F = P_F(o_i). \tag{12}$$

If P_F is invertible with measurable inverse on the relevant support, then $\sigma(Z^F) = \sigma(O)$ and no information in O is lost in principle. This is information equivalence, not semantic identification. A non-invertible P_F is also admissible when it is sufficient for the target belief.

Abduction then forms one focal-token belief,

$$z_i^F \xrightarrow{\mathcal{A}_\phi} Q_i, \qquad Q_i(du) = q_\phi(U \in du \mid Z^F = z_i^F). \tag{13}$$

The word *abduction* names learner-side inference about the actual realization. It does not assert a causal arrow $O \to U$ or $U \to O$, and it does not require an explicit Bayesian posterior computation. In this paper Y_i^F trains shared parameters across records but is not fed back into Equation (13) for focal-token recognition.

A separate query-perception map acts on the mechanism query,

$$Z^Q = P_Q(X^Q), \qquad z^Q = P_Q(x^Q). \tag{14}$$

Reduction combines the two branches:

$$\mathcal{L}_i(dy; x^Q) = \mathcal{R}_\theta(Q_i, z^Q)(dy) = \int p_\theta(dy \mid z^Q, u)\, Q_i(du). \tag{15}$$

For factual scoring, $x^Q = x_i^F$. For a same-token alternative query, x^Q changes while O_i, z_i^F, and Q_i remain fixed. Thus the two required contracts are stronger than "cache something once": the cached object must be formed from admissible selection evidence, and the hypothetical query must enter only the query-perception and response branches.

If only marginal response laws are required, Equation (15) integrates over Q_i. If a simulation compares the same candidate token across several queries, it must draw one $\widetilde{u}_i^{(m)} \sim Q_i$ and reuse that realization. Resampling from the same Q_i at every query changes the simulated token even though the marginal belief is unchanged. Whether event noise is also shared is a separate cross-world coupling decision.

4.2 Three focal-token belief parameterizations

The abduction interface can use several epistemic parameterizations:

$$\text{point:} \quad Q_i = \delta_{a_\phi(z_i^F)}, \tag{16}$$

$$\text{Gaussian:} \quad q_\phi(u \mid z_i^F) = \mathcal{N}\{u; \mu_\phi(z_i^F), \Sigma_\phi(z_i^F)\}, \tag{17}$$

$$\text{Cauchy:} \quad q_\phi(u \mid z_i^F) = \prod_{j=1}^{d} \text{Cauchy}\{u_j; m_{\phi,j}(z_i^F), \gamma_{\phi,j}(z_i^F)\}. \tag{18}$$

The point mode makes the strongest concentration commitment: remaining response-relevant uncertainty can be ignored. Gaussian and Cauchy instead encode different tail and moment commitments; neither is uniformly "more uncertain" in every sense. Drawing $\widetilde{U} \sim Q_i$ is an integration device and does not redraw physical identity.

No distributional mode is calibrated merely because it was parameterized. Without oracle token truth, one-shot data compare these choices first through their composed outcome laws. Moreover, "near" and "far" depend on the latent coordinate chart, so uncertainty geometry is a model choice rather than a reparameterization-invariant property of token ontology.

4.3 Gaussian as a pressure test

Gaussian abduction is a natural and important baseline. It encodes local concentration around a finite-mean center, finite covariance, exponentially decaying remote candidates, and Euclidean scale aggregation. None of these is intrinsically wrong. The pressure test is whether one factual event, without labeled latent coordinates, warrants all of them.

We use the following desiderata for the first tractable family:

1. *Global openness*: finite evidence should not exponentially close remote candidate mechanisms.
2. *No forced moment commitment*: finite mean and variance are not treated as part of token ontology.
3. *Location–scale semantics*: the encoder still emits interpretable trainable coordinates.
4. *Affine stability*: the predictive law remains analytic through a shared mechanism affine in the token candidate.
5. *Exact proper likelihood*: training need not approximate a latent integral by Monte Carlo.
6. *Controlled residual influence*: gross residuals do not dominate the location update without bound.

Within coordinatewise symmetric stable location–scale candidates, Gaussian is the $\alpha = 2$ case and Cauchy is the $\alpha = 1$ case. Cauchy has polynomial tails, no finite mean or variance, and exact L_1 scale aggregation under affine sums [7]. These properties make it a natural minimal choice for our desiderata. They do not prove that Cauchy is uniquely correct among all distributions, nor that every dataset benefits from it. Heavy-tailed candidate uncertainty also does not imply that every token can produce every outcome; outcome support remains governed by the response and event-noise laws.

5 A Learnable Conditional-Affine Cauchy Slice

5.1 Model

Figure 1 makes the implemented specialization explicit. Factual evidence is encoded exactly once into a learner-side unit distribution; every factual or alternative query then enters only the shared response mechanism. Location and scale propagate through separate analytic lanes before being recombined as a response law.

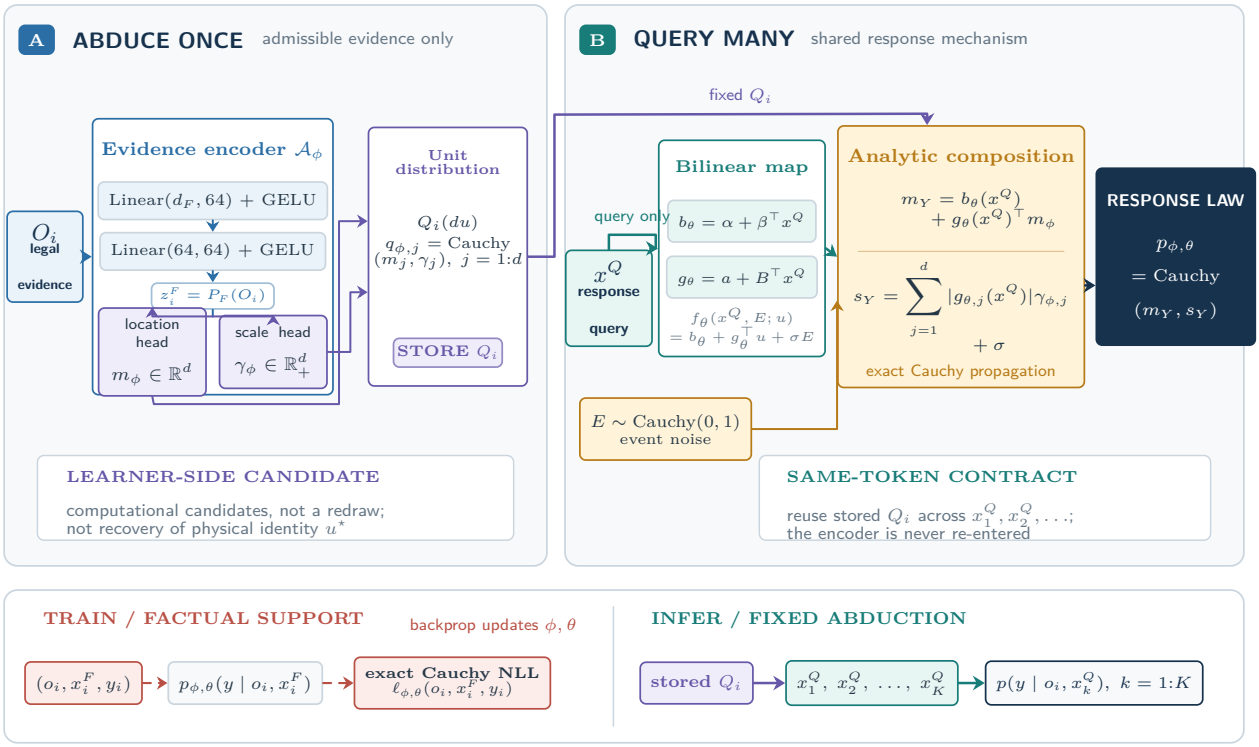

Figure 1: End-to-end computation graph of the Cauchy Unit model. The blue path abducts a token-code candidate from factual evidence; the teal path applies a perceived query to a mechanism affine in the unit candidate; the gold block analytically composes unit uncertainty and event noise. Training specializes to the observed pair (o_i, x_i^F) and minimizes the exact Cauchy negative log likelihood. At same-token evaluation, the stored abduction result is reused for every changed query, so the evidence encoder is never re-entered. The shown layer widths and bilinear query head use the E01 benchmark instantiation, whose latent dimension is $d = 4$; the equations below define the broader conditional-affine family.

The evidence branch emits coordinatewise locations and positive scales,

$$m_\phi(z^F) \in \mathbb{R}^d, \qquad \gamma_{\phi,j}(z^F) = \gamma_{\min} + \text{softplus}\{h_{\phi,j}(z^F)\} > 0, \tag{19}$$

and defines Equation (18). The symbol m denotes a Cauchy location, not an expectation, and γ is a scale, not a standard deviation.

Let $z^Q = P_Q(x^Q)$. The shared mechanism may be nonlinear in z^Q while remaining affine in the unit candidate:

$$f_\theta(z^Q, E; u) = b_\theta(z^Q) + g_\theta(z^Q)^\top u + \sigma_\theta(z^Q)E, \qquad \sigma_\theta(z^Q) > 0. \tag{20}$$

The original bilinear specialization is recovered by setting P_Q to the identity, $b_\theta(x) = \alpha + \beta^\top x$, $g_\theta(x) = a + B^\top x$, and $\sigma_\theta(x) = \sigma$. We assume

$$E \sim \text{Cauchy}(0,1), \qquad E \perp\!\!\!\perp \widetilde{U} \mid z^F, \tag{21}$$

and conditional independence of the candidate coordinates given z^F.

Proposition 1 (Cauchy–affine propagation)**.** *Under Equations* (18)–(21)*, the composed predictive law is*

$$p_{\phi,\theta}(y \mid o, x^Q) = \text{Cauchy}\{y; m_Y(z^F, z^Q), s_Y(z^F, z^Q)\}, \tag{22}$$

where

$$m_Y(z^F, z^Q) = b_\theta(z^Q) + g_\theta(z^Q)^\top m_\phi(z^F), \tag{23}$$

$$s_Y(z^F, z^Q) = \sum_{j=1}^{d} |g_{\theta,j}(z^Q)| \gamma_{\phi,j}(z^F) + \sigma_\theta(z^Q). \tag{24}$$

Proof. Conditional on (z^F, z^Q), Equation (20) is an affine sum of independent Cauchy variables. Locations add with signed coefficients and scales with absolute coefficients, yielding Equations (23) and (24). □

For factual record i, set $z_i^Q = P_Q(x_i)$ and use the exact negative log likelihood

$$\ell_{\phi,\theta}(o_i, x_i, y_i) = \log\{\pi s_Y(z_i^F, z_i^Q)\} + \log\left[1 + \left\{\frac{y_i - m_Y(z_i^F, z_i^Q)}{s_Y(z_i^F, z_i^Q)}\right\}^2\right]. \tag{25}$$

Thus the latent candidate integral is analytic; no latent Monte Carlo is required for training or prediction. This proposition establishes tractability, not identification or universal correctness of the Cauchy family.

5.2 Algorithmic contract

The ontology implies a two-stage interface:

$$\texttt{abduce(factual\_evidence)} \longrightarrow \texttt{UnitDistribution},$$
$$\texttt{response\_law\_from\_unit(unit\_result, query)} \longrightarrow \texttt{ResponseLaw}.$$

A convenience operation `response_law(evidence, query)` composes them for factual use. An alternative query must reuse the first return value and call only the second operation; it must not re-enter the encoder.

Training proceeds as follows: (i) perceive admissible evidence o_i and form a point, Gaussian, or Cauchy Q_i; (ii) perceive the factual mechanism query x_i^F; (iii) evaluate the response law and optimize the proper factual score; and (iv) at evaluation time, reuse the stored Q_i for every alternative query belonging to the same focal token. Point, Gaussian, and Cauchy models share this contract, which makes their differences attributable to uncertainty geometry rather than different same-token semantics.

5.3 A bounded-influence consequence

For fixed scale s, let $r = y - m$. Ignoring constants, the Cauchy loss is $\ell(r; s) = \log s + \log(1 + r^2/s^2)$ and

$$\frac{\partial \ell}{\partial m} = -\frac{2r}{s^2 + r^2}. \tag{26}$$

Its magnitude is bounded by $1/s$ and tends to zero as $|r| \to \infty$. This is a precise optimization-level robustness consequence. It does not imply robustness to arbitrary covariate shift, query extrapolation, or mechanism misspecification; those remain empirical pressure tests.

6 What One-Shot Data Can and Cannot Learn

6.1 Factual predictive-law consistency

Let P_{train} be the distribution of factual records (O, X^F, Y^F) and let $r_0(y \mid o, x)$ be its conditional outcome density. For a composed model, define its factual density

$$r_{\phi,\theta}(y \mid o, x) := p_{\phi,\theta}(y \mid o, x^Q = x). \tag{27}$$

Theorem 1 (Fisher consistency on factual support)**.** *Suppose the candidate density class is dominated, contains r_0, and all relevant log risks are finite. Every population minimizer of*

$$\mathcal{Q}(r) = \mathbb{E}_{P_{\text{train}}}[-\log r(Y^F \mid O, X^F)] \tag{28}$$

satisfies $r(\cdot \mid O, X^F) = r_0(\cdot \mid O, X^F)$ P_{train}-almost surely.

Proof. For any candidate r,

$$\mathcal{Q}(r) - \mathcal{Q}(r_0) = \mathbb{E}\Big[\mathrm{KL}\{r_0(\cdot \mid O, X^F) \| r(\cdot \mid O, X^F)\}\Big] \geq 0. \tag{29}$$

Equality holds only when the conditional densities agree almost surely. □

This is the paper's positive learnability result. It concerns the composed factual law in Equation (27); it does not recover a unique u_i, calibrated Q_i, encoder, mechanism, or same-token response at an unobserved query. Strictly proper scores justify analogous distributional targets under their own integrability conditions [2].

6.2 Latent observational equivalence

Theorem 2 (Latent reparameterization)**.** *Let $h : \mathbb{R}^d \to \mathbb{R}^d$ be a measurable bijection with measurable inverse. Push the candidate distribution forward by h and define*

$$\widetilde{q}_\phi(\cdot \mid z^F) = h_{\#} q_\phi(\cdot \mid z^F), \qquad \widetilde{f}_\theta(z^Q, e; v) = f_\theta\{z^Q, e; h^{-1}(v)\}. \tag{30}$$

Then the original and transformed models induce exactly the same composed predictive law for every (o, x^Q).

Proof. If $\widetilde{U} \sim q_\phi(\cdot \mid z^F)$, then $V = h(\widetilde{U}) \sim h_{\#} q_\phi(\cdot \mid z^F)$ and $\widetilde{f}_\theta(z^Q, E; V) = f_\theta(z^Q, E; \widetilde{U})$ almost surely. □

Consequently, one-shot likelihood cannot assign unique semantic meaning to latent axes. Injective token coding is compatible with this result: an invertible recoding remains injective while changing every numerical coordinate.

Proposition 2 (Token/event scale split is not identified)**.** *Consider the scalar predictive composition*

$$\widetilde{U} \mid z^F \sim \mathrm{Cauchy}\{m(z^F), \gamma_U(z^F)\}, \quad E \sim \mathrm{Cauchy}(0, \gamma_E), \quad \widetilde{Y} = \widetilde{U} + E. \tag{31}$$

For any constant $0 < \delta < \gamma_E$, replacing $\gamma_U(z^F)$ by $\gamma_U(z^F) + \delta$ and γ_E by $\gamma_E - \delta$ leaves the composed predictive law unchanged.

Proof. The composed scale is $\gamma_U(z^F) + \gamma_E$, which is invariant under the stated transformation. □

The architecture may label these components epistemic and aleatoric, but one-shot outcome likelihood identifies only their composed effect without additional information.

6.3 Factual support does not determine alternative queries

Proposition 3 (Alternative-query underdetermination)**.** *Suppose each focal token contributes one observed pair (O_i, X_i^F) to the factual risk. Any two conditional-law functions that agree P_{O,X^F}-almost surely at these factual pairs have identical factual population risk, even if they disagree for the same evidence o_i at an unobserved query $x' \neq x_i$.*

Proof. The factual risk integrates the score only over the observed distribution of (O, X^F). Values at same-evidence alternative queries outside that integration support do not change the risk. □

Shared structural restrictions can extrapolate an alternative-query law, and synthetic oracle queries can test that extrapolation. Neither operation turns the factual-support theorem into calibration of Q_i, recovery of u_i, or nonparametric identification of individual counterfactuals. The selected-token diagnostics later show an empirical manifestation of this boundary: similar factual likelihoods can coexist with sharply different hidden response surfaces.

7 Same-Token Queries and the Causal Boundary

7.1 Fixed abduction, changed query

Let $Q_i = \mathcal{A}_\phi\{P_F(o_i)\}$ denote the point or distributional belief formed once from admissible factual evidence. The same-token alternative-query interface evaluates

$$\texttt{response\_law\_from\_unit}(Q_i, x'), \tag{32}$$

or, in density form,

$$p_{\phi,\theta}(y \mid o_i, x') = \int p_\theta\{y \mid P_Q(x'), u\}\, Q_i(du). \tag{33}$$

Recomputing Q_i from x' would use the alternative query to recognize a potentially different token. Point abduction follows the same rule by retaining its single point mass. Caching is necessary but not sufficient: the cached belief must itself exclude treatment, action, or query coordinates that are not admissible evidence about U.

For marginal laws, Equation (33) integrates all candidates. To compare the same simulated token across several queries, draw one $\widetilde{u}_i^{(m)} \sim Q_i$ and reuse it for every query. Independently drawing a candidate from the same Q_i for each query preserves the marginal belief but changes the simulated token. This candidate-reuse decision is separate from the coupling chosen for event noise.

Equation (32) is a computational linkage contract, not an identification theorem. A causal reading additionally requires:

1. a declared intervention meaning for the changed coordinates of x^Q;
2. token sufficiency and a justified observation/assignment process;
3. positivity or support for the target query;
4. control of confounding or another identification argument;
5. stability of the outcome mechanism across factual and target regimes;
6. the relevant DiscoSCM distribution-consistency condition; and
7. an event-noise coupling when a joint cross-world law is requested.

Full support of $Q_i(du)$, including a Cauchy candidate law, concerns latent-code support. It is not treatment positivity or overlap in the target population. Unit abduction also does not repair sample-selection bias or treatment confounding without a separate sampling/assignment and identification argument. For a marginal alternative-world law, E can be integrated under the declared target event law. A joint sample of $Y(x)$ and $Y(x')$ additionally requires a choice for $(E_x, E_{x'})$; shared, independent, and other couplings generally produce different joint laws.

7.2 Contrast with Deep Structural Causal Models

Deep Structural Causal Models (DSCMs) combine flexible conditional mechanisms with abduction–action–prediction in a declared causal graph [8]. Their central abducted quantities are SCM exogenous variables conditioned on observed endogenous variables. Our current slice begins with epistemic uncertainty about the code of the actual token and separates that stable code from event noise. The comparison is:

Question	Deep SCM	Current Learning DiscoSCM slice
Primary abducted object	Exogenous noise in a declared graph	Focal belief about the actual token code
Evidence	Observed endogenous variables	Declared admissible evidence O_i exposed by the focal token
Stable object	Set by the SCM counterfactual contract	Unit Selection Variable U
Event noise	Represented among SCM exogenous variables	Explicitly separated from stable U as E
Current scope	Counterfactual inference in multi-node deep SCMs	Learning one token-modulated outcome mechanism

This is not a dominance claim. A fuller DiscoSCM may need both token abduction and exogenous-event abduction. The present comparison isolates which uncertainty answers "which token?" and which answers "what happened in this event?"

8 Evidence Ladder: Distinct Targets, Distinct Verdicts

The experiments do not form one pooled leaderboard. E01 evaluates an evidence-indexed composed response law, whereas E05–E07 use a simulator vault to fix the selected realization and evaluate its response surface. Neither is an identified real-world counterfactual. This section keeps each target, protocol, and verdict separate.

8.1 E01: evidence-indexed composed-law robustness

The canonical 1,400-cell confirmatory matrix completed without failed fits; its frozen-gate status is `not_confirmed`.

8.1.1 Primary benchmark and truth firewall

The primary simulator draws a hidden actual code to generate each token's one factual record (w, x, y). The bounded E01 scoring oracle is narrower: for each clean test evidence row w, it stores the location and scale of the composed conditional response law over a target-query grid. It does not treat the hidden code as a recovery target or provide 21 realized potential outcomes for that

code. Oracle response-law parameters are unavailable to fitting, validation, early stopping, and hyperparameter selection. A row identifier is allowed only for bookkeeping and artifact alignment.

The bounded experiment uses four worlds: matched, no heterogeneity, weak token evidence, and narrow factual query support. The no-heterogeneity world tests whether token factorization invents structure; weak evidence tests whether a distribution honestly widens; and narrow support separates interpolation from unsupported extrapolation. Nonlinear misspecification and outcome contamination remain separate extensions rather than being silently pooled into this experiment.

Each world contains 6,000 tokens and uses a 60/20/20 train/validation/test split. The oracle contains 21 equally spaced target queries on $[-1, 1]$. For each world, seed, and corruption ratio, every model receives the same split and corruption design. The corruption acts only on the combined train–validation outcome pool: a SHA-pinned Sattolo subset derangement breaks a fraction $\rho \in \{0, .1, .2, .3, .4\}$ of factual input–outcome links while preserving the outcome multiset and moving every selected link. Test rows and oracle grids remain clean and are fingerprinted before and after corruption.

For a token model the evaluator calls `abduce(evidence)` once, stores the resulting `UnitDistribution`, and reuses it for all 21 target queries. Thus an alternative query cannot trigger token re-recognition. Direct predictors use their direct-query path and never receive the oracle curve. In the current implementation, test-token abduction reads w, not that token's clean factual outcome y; the latter is retained only for factual evaluation metrics.

8.1.2 Comparators and metrics

The comparison matrix includes deterministic, Gaussian, and Cauchy Unit models; matched direct Gaussian and Cauchy predictors; a varying-coefficient Cauchy predictor; and a four-expert Cauchy mixture. The primary contrast uses 5,267 parameters for the Cauchy Unit model and 5,216 parameters for a direct Cauchy network with hidden dimensions $(66, 66)$, a gap below one percent. All inferential fits use AdamW with learning rate 10^{-3}, weight decay 10^{-5}, batch size 256, at most 300 epochs, and patience 30.

The primary endpoint is evaluator-hidden curve-location error. Secondary endpoints are the 30-percent-shuffle error, degradation relative to clean training, factual NLL and median error, pinball losses, PIT-ECE, interval coverage and width, hidden-grid CDF error, and—for narrow support—separate inside- and outside-support curve errors. Cauchy means, variances, and CRPS are not reported because a non-degenerate Cauchy lacks the required finite moments.

The mechanics smoke contains eight rows. Development uses seeds $11, 22, 33, 44, 55$ for a 700-row diagnostic matrix. Only after that matrix and its analysis are complete may the implementation freeze source, dependency, configuration, analysis code, gates, and output namespace. Confirmation then uses seeds 101 through 110 for 1,400 rows. A failed or completed confirmatory namespace cannot be overwritten or reopened under another official checkpoint. All 1,400 planned confirmatory cells completed, with zero failed fits.

8.1.3 Promotion gates

At matched $\rho = .4$, let $E_{D,s}$ and $E_{U,s}$ be the direct-Cauchy and Cauchy-Unit hidden curve errors for confirmatory seed s, and define

$$\Delta_s = \log \frac{\max(E_{D,s}, 10^{-12})}{\max(E_{U,s}, 10^{-12})}.$$

The primary robustness gate requires a positive mean Δ, a positive lower endpoint of the two-sided 95-percent Student-t interval, and a one-sided exact sign-test $p \leq .05$; zero effects are discarded by the predeclared tie rule.

The no-heterogeneity guardrail defines geometric Unit advantage as $1 - \text{gmean}(E_U/E_D)$. Its upper 95-percent bound, $1 - \exp(-\text{CI}_{.95}^{\text{high}})$, may not exceed 10 percent. Primary robustness and permission to attribute it to token heterogeneity are reported separately; overall acceptance requires both. Geometry, loss-family, and capacity secondary hypotheses receive separate within-family Holm correction. No secondary result changes either gate.

A structural advantage is therefore promotable only if the Unit model improves the hidden matched-query target, survives fresh-seed confirmation, and does not manufacture a material gain in the no-heterogeneity world. Weak-evidence and unsupported-query failures remain visible rather than being hidden by a pooled average.

8.1.4 Frozen confirmatory verdict

The confirmatory run did not establish the primary robustness claim. At matched $\rho = .4$, the ten-seed statistic was

$$\overline{\Delta} = -0.00714785, \qquad \text{CI}_{.95} = [-0.0495877,\ 0.0352920].$$

The Cauchy Unit model had lower hidden curve error in four of ten seeds, giving a one-sided exact sign-test p-value of 0.828125. Thus every component of the primary gate failed: the mean was not positive, the interval lower endpoint was not above zero, and the sign-test p-value was not at most .05. The point estimate is a 0.72-percent geometric disadvantage for the Unit model, not evidence of a robustness advantage.

The no-heterogeneity guardrail passed. Its 95-percent upper bound on apparent Unit advantage was 6.01 percent, below the predeclared 10-percent ceiling. This permits only the narrow statement that the experiment did not detect a material manufactured advantage after heterogeneity was removed. It cannot rescue the failed primary endpoint, and it does not show that token heterogeneity caused an advantage. Overall acceptance is therefore false and the confirmatory status is `not_confirmed`.

Supportive secondary results are kept subordinate to that verdict. In the Cauchy-matched world at $\rho = .3$, Cauchy Unit was favored over Gaussian Unit on the predeclared absolute factual, curve, and calibration endpoints after within-geometry-family Holm adjustment; for hidden curve-location error, $\overline{\Delta} = 1.21548$ and the adjusted p-value is 0.02539. However, the relative clean-to-40-percent-shuffle degradation contrast pointed against Cauchy ($\overline{\Delta} = -0.56276$; candidate-direction adjusted $p = 1$). Because these analyses are secondary and the matched DGP itself uses Cauchy uncertainty, they do not confirm Cauchy epistemic geometry, and they cannot change the failed primary gate.

Existing tabular shuffle artifacts are therefore classified as external predictive-robustness context and are not pooled into this experiment. The bounded score in Equation (26) supplies a testable downstream prediction; it is not the ontological reason for introducing U or abduction.

The observed result is not positive. Even if it had been, this experiment could support only synthetic fixed-abduction composed-law robustness. It does not establish recovery of the actual U, identification of individual counterfactuals, distribution-consistency, or a complete Layer 3 joint law.

8.2 E05: selected-token target correction

E05 changes the evaluator target rather than reinterpreting E01. The simulator draws one actual code u_i, generates noisy token evidence and one factual outcome, and then holds that same u_i

fixed over a 21-query response grid. The learner never receives the selected code. A finite-test evidence-conditioned Bayes reference quantifies the irreducible error from observing noisy evidence rather than u_i itself.

The semantic matrix completed all 72 formal cells and 18 oracle-isolated diagnostics. With the true mechanism frozen, the learned evidence-only abductor approached the finite-test reference: its maximum normalized excess was 0.00729, below the predeclared 0.05 ceiling. This is a bounded result for the abduction subproblem under a frozen matched mechanism, not general recovery of Q_i.

When abductor and mechanism were learned jointly from the one-shot factual likelihood, the clean semantic gates failed. Matched normalized excess was 0.5307 and no-heterogeneity normalized error was 0.6808, both above the 0.20 ceiling. Different initialization basins had similar factual likelihoods but sharply different hidden token paths. The formal status was therefore `development_ready=false`; no development or confirmatory shuffle claim was run. This result empirically localizes the theoretical factorization and alternative-query underdetermination described earlier.

8.3 E06: admissible-evidence interface repair

E06 tests a narrower failure. Its factual input is $[z, a]$, where z is a token cue and a is independently randomized factual action satisfying $P(U \mid z, a) = P(U \mid z)$. Both arms cache one belief across queries, but the `full_x` abductor reads $[z, a]$ whereas the matched-capacity `z_only` abductor masks the action. The mechanism receives the full query in both arms.

In matched clean data, masking action reduced fixed-realization curve MAE from 0.223607 to 0.030654, an 86.3-percent improvement, and made abduction location and scale exactly invariant to alternative actions. This establishes interface utility in the declared DGP. It does not establish token-semantic recovery: the no-heterogeneity control improved even more (0.234974 versus a matched gain of 0.192953), so the semantic-specificity guard failed. With the true mechanism frozen, the learned Cauchy belief was also strongly underdispersed: nominal 80-percent interval coverage was 0.3822, compared with 0.7917 for the exact truncated-Uniform reference.

The authorized 40-percent stress diagnostic retained an absolute advantage for the masked interface but degraded much more from clean than the full-input arm. E06 therefore supports the rule that cached abduction must be constructed from admissible evidence; it does not support calibrated Q_i, general action masking, corruption robustness, or confirmatory superiority.

8.4 E07: query perception and conditional response diagnostics

The preconfirmatory E07 diagnostic asks whether the response branch should perceive raw queries and whether its coefficient functions should be nonlinear in perceived query coordinates while remaining affine in the token candidate. Six globally parameter-matched arms completed 108 formal fits, with 48 nonranking oracle diagnostics. Explicit perception and conditional response forms improved hidden surfaces relative to the raw bilinear model in the tested warped worlds, and a capacity-matched factorized candidate beat the direct predictor in one predeclared contrast.

The diagnostic nevertheless returned `semantic_not_ready`. The invertible-flow candidate did not beat a residual-MLP perception map, several predeclared improvement gates were missed, isolated learned response paths remained above their ceilings, initialization basins persisted, and the no-unit-heterogeneity control still produced spurious evidence-dependent unit effects. E07 therefore motivates the conditional-affine model class in Section 5 and identifies perception, response learning, and unit-effect regularization as separate bottlenecks. It does not validate a P–A–R architecture, establish that invertibility is necessary, or promote a factorized-model advantage.

9 Related Work

Structural causal models separate causal mechanisms and support formal intervention and counterfactual queries [9]. Deep SCMs make these mechanisms expressive and retain the classical abduction–action–prediction workflow [8]. DiscoSCM adds an explicit distribution-consistency perspective and motivates the separation between stable token selection and event-level randomness [3]. Our contribution is not a replacement for these theories, but a bounded learning formulation for one DiscoSCM outcome mechanism.

Latent-variable learning and variational autoencoders provide powerful tools for inference in flexible generative models [5]. Causal representation learning asks when latent causal variables can be recovered and emphasizes that identifiability needs interventions, environments, grouping, or other inductive structure [1, 10]. Our reparameterization result enforces the same discipline: predictive success does not automatically give semantic latent coordinates.

Representation learning for treatment effects shares information across observed units to estimate missing potential outcomes [6, 11]. Our factual-support/alternative-query distinction is compatible with that literature but targets a different object: a general token-modulated outcome law with an explicit abduction result and event noise, not a claim that one-shot outcomes nonparametrically identify individual counterfactuals.

Distributional regression, proper scoring rules, and mixture models offer strong observational comparators [2, 4]. They can match the factual conditional law without endorsing our latent ontology. This is why direct distributional predictors belong in every benchmark. Stable distributions provide the analytic tool used by our first specialization [7]; Cauchy stability is a tractability property, not evidence that nature has a Cauchy token distribution.

10 Discussion

The first question in learning DiscoSCM is not which neural architecture to use. It is how stable token identity becomes a learning object without being confused with population composition, treatment assignment, sample inclusion, or event noise. The single selector U supplies the ontological role; the actual realization $U = u_i$ fixes who is being queried; and Q_i records what a learner can infer about that fixed realization from admissible evidence. This is the conceptual contribution to which the model families are subordinate.

The Perception–Abduction–Reduction contract makes the corresponding software and statistical obligations explicit. Evidence perception and query perception are separate branches. Abduction acts once on legal evidence; reduction holds the resulting Q_i fixed while the query changes. Merely caching an encoder output is insufficient if a randomized action or a hypothetical query has already contaminated that output. Likewise, reusing the same marginal Q_i is insufficient for a simulated paired comparison if a new candidate token is drawn for each query.

The positive mathematical result remains deliberately narrow: a correctly specified composed factual law is Fisher-consistent on observed support, and a Cauchy belief propagated through a conditional response affine in the token candidate is exactly tractable. The negative map is at least as important. Latent coordinates, belief–response factorization, token/event scale allocation, and same-token alternative-query laws are not recovered by that theorem. These limits prevent "end-to-end" from silently changing meaning from differentiable factual training to focal-belief calibration, selected- token recovery, or causal identification.

The empirical sequence now separates target definition, interface correctness, and semantic recovery. E01 found no robustness advantage for the evidence-indexed composed law. E05 isolated

a near-reference abduction subproblem only after freezing the true mechanism, while joint one-shot factorization failed its clean gates. E06 showed that fixed caching does not prevent contamination by selection-irrelevant query coordinates; its interface repair was large but neither token-specific nor calibrated. E07 found useful signals for query perception and conditional response functions, but failed isolated-path, initialization-stability, and no-spurious-heterogeneity gates. These are not four attempts at the same leaderboard. They progressively reveal which object was being scored and where the current learner remains underdetermined.

The next obligation is therefore not more shuffle seeds. The unresolved work is to stabilize the isolated response path, constrain or gauge the belief–mechanism factorization, suppress evidence-dependent unit effects when heterogeneity is absent, and obtain additional information that can calibrate the focal belief. Candidate routes include shared-token multi-view or multi-mechanism evidence, explicit heterogeneity shrinkage, and designs whose assignment and query support permit a named causal estimand. Each route needs its own theorem and fresh evidence gate.

Learning one outcome mechanism is not the completion of DiscoSCM. It is the first place where ontology, observation process, learnability boundary, and falsifiable evidence can be forced to agree. *Unit Mechanism Learning* and *Token Causation* remain useful sibling viewpoints, but no theorem or experimental verdict transfers to them automatically.

References

[1] Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 372–407, 2023.

[2] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[3] Heyang Gong, Chaochao Lu, and Yu Zhang. Distribution-consistency structural causal models. *arXiv preprint arXiv:2401.15911*, 2024. URL https://arxiv.org/abs/2401.15911.

[4] Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994.

[5] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

[6] Christos Louizos, Uri Shalit, Joris M. Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[7] John P. Nolan. *Univariate Stable Distributions: Models for Heavy Tailed Data*. Springer, 2020.

[8] Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In *Advances in Neural Information Processing Systems*, volume 33, pages 857–869, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/0987b8b338d6c90bbedd8631bc499221-Abstract.html.

[9] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

[10] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[11] Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3076–3085, 2017.

A Notation and Claim Boundary

Symbol	Meaning
$t^\star$	actual discrete focal token
$u^\star = \iota(t^\star)$	world-side computational code of that token
$\widetilde{U}$	learner-side candidate used for integration; not an identity redraw
$O_i = o_i$	admissible factual evidence for focal token i
$z_i^F = P_F(o_i)$	perceived evidence supplied to abduction
Q_i	stored focal-token belief formed once from z_i^F
x_i^F	factual query at which y_i^F was observed
x^Q	query supplied to the outcome mechanism
$z^Q = P_Q(x^Q)$	perceived query supplied to the response mechanism
E	event-level randomness after token and query are fixed
f_0 / f_θ	world-side mechanism / learner family
$p_{\phi,\theta}$	composed learner predictive law

B Continuity of Cauchy Predictive Parameters

The following lemma is useful for implementation checks but is not the main learnability theorem.

Proposition 4 (Predictive-law continuity)**.** *Fix (z^F, z^Q) and an oracle-correct conditional-affine response head. Suppose estimated Cauchy locations and positive scales converge coordinatewise to target values, and the target predictive scale is positive. Then the predictive location and scale in Equations (23)– (24) converge. The corresponding CDF converges at every finite threshold, and every fixed quantile in $(0, 1)$ converges.*

Proof. The predictive parameters are finite continuous affine functions of the component locations and scales. The Cauchy CDF and quantile are

$$F_{m,s}(y) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{y-m}{s}\right), \tag{34}$$

$$Q_p(m, s) = m + s \tan\{\pi(p - 1/2)\}, \tag{35}$$

which are continuous on $s > 0$. □

This result assumes convergence of the learned component parameters and an oracle response head. It therefore checks propagation after learning; it does not prove that those parameters are estimable from one-shot factual data.