Unit Mechanism Learning: From Outcome Prediction to Unit-Specific Generation

Anonymous Author(s)

Working paper, July 16, 2026
Paper version: UML-WP-2026-07-15-V7

Abstract

Supervised learning usually estimates an observational predictive law such as $p(Y \mid X = x)$. We study the finer computation hidden inside that prediction. At world level, U selects a token/unit, one realization $U = u^\star$ is that token together with its embedding, and its realized outcome is $Y(x) = f_\theta(x, E; U)$, or $Y_{u^\star}(x) = f_\theta(x, E; u^\star)$ after the realization is fixed. At learner level, the observed predictors $X = x$ are factual evidence about the unobserved $u^\star$ and, in ordinary factual prediction, the same x is also the mechanism query. Abduction can return a point $\widehat{u} = a_\phi(x)$, a local Gaussian candidate distribution, or a heavy-tailed Cauchy candidate distribution. These are alternative epistemic representations of the same actual token, not redraws of physical identity. Prediction composes an abducted candidate u with the shared generator, $f_\theta(x, E; u)$; equivalently, its density is $p_{\phi,\theta}(y \mid x^F, x^Q) = \int p_\theta(y \mid x^Q, u) q_\phi(u \mid x^F)\, du$. For a same-token counterfactual, the abduction result from factual x^F is held fixed while only the query changes to $x^Q = x'$. We call this problem *Unit Mechanism Learning*. An evidence-only branch uses $f_\theta(E; u, c_0)$, while an explicit-input branch uses $f_\theta(X, E; u, c_0)$; Causal Regression belongs to the former. The current formulation adopts a one-to-one unit–representation code as an identity-encoding modeling assumption, not as a coordinate system identified from one-shot data. Our primary regime contains one factual event per sampled unit. We therefore make the learnability boundary explicit: the induced conditional response law—not a unique physical mechanism—is the direct one-shot learning target and can be learned under predictive conditions, whereas the latent coordinates, the structural factorization, and the split between unit and event uncertainty are generally not identified. We prove a projection result connecting the response law to ordinary prediction, constructive non-identifiability results, and population Fisher consistency under the log score. We then introduce a Cauchy-affine specialization with a bilinear unit–input interaction. Its location and scale are learned, but its candidate distribution has no finite mean or variance; this encodes globally open rather than local Gaussian uncertainty over candidate unit mechanisms. It yields closed-form response distributions and quantiles without a population prior, a KL objective, or latent Monte Carlo sampling. A new evaluation protocol uses one factual event per unit while retaining oracle response curves only for evaluation. In a four-dataset observational stress test, training and validation outcomes are independently deranged within split while test outcomes remain clean. A fresh-seed, eight-contrast extension designed after V2 was unsealed finds that a learned-scale Laplace predictor is the only model to pass a multiplicity-controlled gate that requires both lower clean-test error after training with corrupted train/validation outcomes and less own-clean degradation on every dataset. Direct Cauchy retains a large absolute advantage over matched Gaussian fitting but does not pass that stricter degradation gate; neither current unit factorization nor repaired historical cores establish a robust advantage. Causal and cross-context interpretations require additional assumptions and are not conferred by the framework or these prediction experiments.

1 Introduction

The standard supervised-learning abstraction maps observed predictors W to an outcome Y. A probabilistic predictor estimates $P_{\text{obs}}(Y \mid W = w)$; a point predictor estimates a functional such as a conditional mean or median. This abstraction is often sufficient for forecasting, but it does not require the learner to distinguish why two units with similar observed predictors respond differently, which variables are inputs of the outcome-generating mechanism, or which variation belongs to a new event rather than to the unit.

We study a more structured question. At model level, a unit-selection variable U selects a token/unit and its embedding realization u. Let O be admissible evidence, x the explicit input of an outcome mechanism, and E event-level randomness. In a fixed context c_0, write

$$Y(x) = f_\theta(x, E; U), \qquad Y_u(x) = f_\theta(x, E; u, c_0). \tag{1}$$

For a realized token, we write $U = u^\star$ when it is useful to distinguish its fixed but unobserved embedding from a learner-side candidate u.

The bridge from Equation (1) to ordinary machine learning is *unit abduction*. In the basic supervised specialization, the observed predictors $X = x$ have two operational roles. As factual evidence x^F, they are used to recognize which token embedding could have generated the row; as query x^Q, they specify where to evaluate that token's mechanism. For factual prediction, $x^F = x^Q = x$. The learner may return one of three abduction objects:

$$\begin{aligned} \text{point:} \quad & \widehat{u} = a_\phi(x^F), \\ \text{Gaussian:} \quad & q_\phi(u \mid x^F) = \mathcal{N}\{u; \mu_\phi(x^F), \Sigma_\phi(x^F)\}, \\ \text{Cauchy:} \quad & q_\phi(u \mid x^F) = \prod_j \text{Cauchy}\{u_j; m_{\phi,j}(x^F), \gamma_{\phi,j}(x^F)\}. \end{aligned} \tag{2}$$

The point mode asserts that the evidence can be compressed into one candidate. The Gaussian mode represents local candidate uncertainty around a center. The Cauchy mode has a location and scale but no finite mean or variance; it keeps distant candidate unit mechanisms appreciably possible instead of forcing uncertainty into a finite-variance neighborhood of an average unit. This heavy-tailed openness does not imply that every outcome can be produced by every unit: the support of possible outcomes is still determined by f_θ and the event law.

For either distributional mode, prediction can be written without introducing a separate response-kernel symbol:

$$p_{\phi,\theta}(y \mid x^F, x^Q) = \int p_\theta(y \mid x^Q, u) q_\phi(u \mid x^F)\, du, \qquad p_\theta(y \mid x, u) := \mathcal{L}_E\{f_\theta(x, E; u, c_0)\}. \tag{3}$$

Equivalently, sample a computational candidate $\widetilde{U} \sim q_\phi(\cdot \mid x^F)$ and event noise E, then evaluate $f_\theta(x^Q, E; \widetilde{U}, c_0)$. This sampling notation expresses the learner's uncertainty about the same actual $u^\star$; it does not redraw the physical identity, declare a population prior, or create a posterior over response functions.

We call the resulting learning problem *Unit Mechanism Learning*. Its primary modeling objects are the unit-specific instances $f_{\theta,u,c_0}(\cdot,\cdot) \equiv f_\theta(\cdot,\cdot; u, c_0)$. The central modeling bet is that a response-relevant unit representation can make the unit-specific instances jointly simple within a shared, unit-modulated family. This allocates stable model-side heterogeneity to $u^\star$, uncertainty over candidate representations to q_ϕ, explicit query inputs to X, and event randomness to E. We use *unit-specific mechanism* for the object, *unit-conditioned instance* for its relation to the shared family,

and *shared unit-modulated mechanism family* for f_θ; the instances are not independently fitted per-unit functions.

The current formulation adopts a one-to-one token–embedding identity encoding. This is an *identity-encoding modeling assumption*; one-shot observations do not identify its numerical coordinates, and a future framework extension may relax it. The downstream map $u \mapsto \{x \mapsto p_\theta(\cdot \mid x, u)\}$ need not be injective: distinct candidate representations may induce the same predictive behavior. Ordinary prediction therefore outputs $p_{\phi,\theta}(y \mid x, x)$ in Equation (3), not one selected u or one selected response function.

For a same-token alternative-input query, the separation of roles becomes essential. The learner first abducts from the factual predictors $x^F = x$ and then evaluates

$$p_{\phi,\theta}(y \mid x, x') = \int p_\theta(y \mid x', u) q_\phi(u \mid x)\, du. \tag{4}$$

It must not replace $q_\phi(u \mid x)$ by $q_\phi(u \mid x')$, because that would use the alternative query to recognize a potentially different token. This fixed-abduction, changed-query rule supplies the Layer 3 linkage; causal interpretation still requires the additional assumptions stated below.

The explicit mechanism input is a branch choice, not a property of every abduction-first model. An evidence-only branch may take observed features X_{obs} as O and use them only to infer U, with outcome equation $Y = f_\theta(E; U, c_0)$ with $U \sim q_\phi(\cdot \mid O)$. Causal Regression has this structure: its observed regression features are abduction evidence and do not enter the outcome mechanism directly. The explicit-input model in Equation (1) is a different sibling specialization, not a relabeling of that equation.

The distinction from ordinary prediction is real but must not be overstated. Prediction is an observational projection of the finer model. Conversely, a conditional predictive law does not generally identify the mechanism factorization that generated it. This matters especially in our primary data regime: the training population contains many units but only one factual event per unit. This is a deliberate one-shot design: as in ordinary cross-sectional data and the potential-outcome regime, input dependence must be learned from variation in X across units and shared mechanism structure, while each unit's alternative-input responses remain missing. Under that regime, it is possible to fit and calibrate a conditional response law, but it is generally impossible to verify a unique latent coordinate, a true physical mechanism, or a unit-versus-event uncertainty split. We treat this negative result as part of the framework rather than hiding it in a limitations paragraph.

Our first tractable model uses a factorized Cauchy abduction distribution and a bilinear mechanism. For evidence O, vector input X, and latent dimension d, the model is

$$\widetilde{U}_j \mid O \sim \text{Cauchy}\{m_{U,j}(O), \gamma_j(O)\}, \tag{5}$$

$$\widetilde{Y} = \alpha + \beta^\top X + a^\top \widetilde{U} + X^\top B \widetilde{U} + \sigma E, \qquad E \sim \text{Cauchy}(0, 1). \tag{6}$$

The bilinear term permits unit-specific slopes while preserving affine stability conditional on X. The reduced conditional response is therefore available in closed form and can be trained by factual negative log likelihood. This specialization introduces direct-input and unit–input interaction paths that are absent from Causal Regression; the shared inheritance is the evidence-to-unit abduction pattern and analytic Cauchy propagation, not the outcome mechanism.

Our contributions are:

1. We formulate Unit Mechanism Learning around stable model-side unit representations, learner-side candidate abduction, and the unit-specific instances of a shared, unit-modulated mechanism family, while separating admissible evidence, mechanism input, event randomness, and context.

2. We define the reduced conditional response as the primary one-shot learning target, prove that ordinary prediction is its observational projection, and establish latent and unit/event non-identifiability results.

3. We give a population Fisher-consistency result for response-law learning and derive closed-form Cauchy-affine response parameters, CDFs, and quantiles.

4. We provide a clean-room implementation and a truth-isolated evaluation protocol in which each training unit contributes one factual event while oracle response curves are retained by the evaluator only.

Unit Mechanism Learning is an outcome-generation framework, not an automatic causal identification result. When X is a treatment or action, causal and counterfactual interpretation additionally requires intervention semantics, support, identification, stability, and cross-world assumptions. Modeling variation across contexts is a separate extension that we defer.

2 Relationship to Existing Work

The active HCGM notation is model-level U, token embedding realization u, shared mechanism $Y(x) = f_\theta(x, E; U)$, and learner-side $q_\phi(du \mid O)$. Indexed notation retained from cited literatures is comparison bookkeeping, not a competing unit ontology.

The contribution is not the mere use of a latent variable. Several mature literatures already learn latent structure, individual variation, or conditional response functions.

Causal Regression lineage. Causal Regression uses observed features to form a learner-side abduction law over candidate representations of the latent modeled unit and posits the historical outcome equation $Y = f^{\mathrm{CR}}(U, \varepsilon)$. Its tractable CausalEngine maps X_{obs} through Perception and Abduction before applying an Action map to U and event noise; it has no direct $X_{\mathrm{obs}} \to Y$ mechanism path. The present Cauchy-bilinear model instead introduces an explicit mechanism input and an X–U interaction. The two lines share abduction-first uncertainty over unit representation and an affine Cauchy propagation motif, but impose different outcome mechanisms and should be treated as sibling specializations rather than the same tractable model.

Structural causal models and unit-specific instances. A conventional structural causal model declares population-shared assignments, for example $Y_i = f_Y(X_i, \varepsilon_i)$ [17]. Fixing a unit's exogenous realization already induces the unit-specific response map $f_i^{\mathrm{SCM}}(x) := f_Y(x, \varepsilon_i)$. Our grammar can therefore be embedded in an SCM by decomposing a token's exogenous realization into a stable embedding $U = u^\star$ and event noise E; we do not claim that SCMs lack unit-specific functions or that UML adds expressive power to the SCM formalism. The difference is the modeling and learning target: UML starts from one selected token with a stable realization $U = u^\star$, promotes the unit-conditioned instances $f_{u,c}$ of a shared family to the primary objects, explicitly separates the fixed representation from event noise and learner-side candidate uncertainty, and states what reduced conditional response can be learned from one factual event per unit.

Mixed and varying-coefficient models. Random-effects and mixed-effects models separate population structure from subject- or group-specific deviations. Neural mixed-effects models extend this idea with learned representations and covariance structure [8]. Varying-coefficient models and mixtures of experts allow regression parameters or experts to depend on observed covariates

[5, 7]. Our Cauchy-bilinear model is adjacent to these families. The distinction is primarily the framework-level separation of evidence O from mechanism input X, the directly learned abduction distribution $q_\phi(du \mid O)$ over candidate representations of the same latent modeled unit, and the reduced response-law target with an explicit one-shot identifiability boundary. Here q_ϕ is not a function posterior.

Mechanism learning and nonseparable structural functions. Independent-mechanism learning and transfer work explicitly treats causal mechanisms as reusable objects [15]. Econometric nonseparable and triangular models likewise study structural functions whose responses vary with latent heterogeneity [4, 12]. These literatures rule out a broad "first to learn generating mechanisms rather than predictors" claim. Our narrower claim is the primary-object/one-shot combination above, together with the $O/X/U/E$ role separation and an explicit response-law identifiability boundary.

Amortized latent-variable and function learning. Variational autoencoders learn conditional latent distributions through amortized inference [10]. Neural Processes map a context set to a distribution over functions and support target-conditioned prediction [2, 9]. These are important neighbors, not subsumed methods. Neural Processes typically obtain task- or function-level information from a set of context input–output pairs. Our primary regime instead has one factual event per unit and uses separate admissible evidence to form uncertainty over unit representations. Each candidate u determines the predictive map $x \mapsto p_\theta(\cdot \mid x, u)$ within the shared family, but this map need not be injective in u; a function-space law is only the pushforward of q_ϕ, whereas ordinary prediction composes candidate predictions into $p_{\phi,\theta}(y \mid O, x)$.

Deep causal and counterfactual models. Deep structural causal models use flexible structural equations and abduct exogenous noise for counterfactual inference [16]. DiscoSCM motivates separating stable individual variation from event-level noise [3]. Our base framework is deliberately weaker than a full causal model: it defines an outcome-generating target and states which additional assumptions are required before alternative X values acquire an interventional or counterfactual meaning. General counterfactual non-identifiability from observational data also cautions against interpreting a fitted latent decomposition as the uniquely true mechanism [14].

Individualized treatment-effect learning. Representation-learning approaches estimate heterogeneous treatment effects under assumptions such as strong ignorability [6, 19]. CEVAE combines proxy-variable modeling with a latent confounder model [11]. These methods activate treatment semantics and target causal contrasts. In our hierarchy they are causal specializations rather than definitions of the base outcome-generation problem.

Causal representation identifiability. Causal representation learning studies conditions under which latent causal variables or their graphs can be recovered. Recent guarantees use interventions, multiple environments, temporal structure, grouping, or other restrictions [1, 13, 18, 20]. These results reinforce our boundary: one-shot factual prediction alone should not be presented as identifying the coordinates of the actual $u^\star$ or an identity-encoding map.

The paper's novelty claim is therefore not the invention of mechanism learning or of unit-specific response functions. It is the decision to model each sample as arising from one latent unit with a stable response-relevant representation; infer a learner-side distribution over candidate values of that representation from admissible evidence; let each candidate index a response function in

a shared family; and make their reduced conditional response the direct one-shot target, while preserving the evidence-only versus explicit-input branch distinction and the boundary between predictive adequacy and latent or causal identification. The current formulation explicitly adopts a one-to-one unit–representation code as an identity-encoding modeling assumption rather than a one-shot identification result; a future framework extension may relax it.

3 Problem Setup

3.1 World-side generation and learner-side recognition

We work at a fixed context $C = c_0$ and suppress c_0 when no ambiguity can arise. The model-level unit-selection variable is $U \in \mathcal{U}$. One realization $U = u^\star$ is a particular token/unit together with its embedding. For mechanism input x and realized event noise e, the world computes

$$Y(x) = f_\theta(x, E; U), \qquad U = u^\star,\ E = e \implies Y_{u^\star}(x) = f_\theta(x, e; u^\star, c_0). \tag{7}$$

A row number does not enter this computation. Superscripts such as $(X^{(n)}, Y^{(n)})$ below identify observations only; they do not create row-specific variables U_n or independently fitted functions f_n.

The learner faces a different computation because the actual $u^\star$ is hidden. The operational variables have distinct roles:

- $O \in \mathcal{O}$ is admissible factual evidence about the selected token;
- $X^Q \in \mathcal{X}$ is the query supplied to the outcome mechanism;
- $U = u^\star$ is the fixed world-side token realization, while a generic u denotes one learner-side candidate value;
- $E \in \mathcal{E}$ is event-level randomness after token and query are fixed.

In ordinary supervised prediction, the observed predictors $X = x$ are the factual evidence, so $O = X^F = x$, and the same value is also the factual query, $X^Q = x$. The superscripts F and Q name roles rather than new random variables. They become essential for a same-token alternative query: evidence remains $X^F = x$, while the mechanism query becomes $X^Q = x'$.

Definition 1 (Unit abduction)**.** *An abduction operator maps factual evidence to an epistemic description of the same actual token embedding $u^\star$. Three useful modes are*

$$\textit{point:} \quad \widehat{u} = a_\phi(O), \tag{8}$$

$$\textit{Gaussian:} \quad q_\phi(u \mid O) = \mathcal{N}\{u; \mu_\phi(O), \Sigma_\phi(O)\}, \tag{9}$$

$$\textit{Cauchy:} \quad q_\phi(u \mid O) = \prod_{j=1}^{d} \mathrm{Cauchy}\{u_j; m_{\phi,j}(O), \gamma_{\phi,j}(O)\}. \tag{10}$$

The distributional modes are learner uncertainty over candidate embeddings for one fixed token. They are neither population laws nor repeated world-side draws of identity.

Point abduction makes the strongest recognition claim. Gaussian abduction describes local uncertainty with a finite mean and covariance. Coordinatewise Cauchy abduction has learned location and scale but no finite mean or variance; its polynomial tails keep remote candidate unit mechanisms open. This is an uncertainty geometry, not a claim that every outcome can be generated by every unit. Outcome support remains constrained by f_θ and the law of E.

Definition 2 (Unit Mechanism Model)**.** *A Unit Mechanism Model in context c_0 is a pair $(\mathcal{A}_\phi, f_\theta)$, where $\mathcal{A}_\phi(O)$ returns one of the abduction objects above and $f_\theta : \mathcal{X} \times \mathcal{E} \times \mathcal{U} \times \mathcal{C} \to \mathcal{Y}$ is a shared measurable outcome mechanism. For fixed (u, c_0), the unit-conditioned instance is $f_{\theta,u,c_0}(x, e) := f_\theta(x, e; u, c_0)$.*

These unit-conditioned instances are the primary modeling objects, but they are not free functions fitted independently for each row. Their common parameterization is what permits learning across a one-shot population. For readability, we write the outcome distribution induced by event noise as

$$p_\theta(\cdot \mid x, u) := \mathcal{L}_E\{f_\theta(x, E; u, c_0)\}. \tag{11}$$

If a density does not exist, Equation (11) is understood as a probability law; the corresponding event-set kernel is needed only for measure-theoretic proofs.

Point abduction predicts from $f_\theta(x^Q, E; \widehat{u}, c_0)$. A distributional abduction predicts by sampling a computational candidate $\widetilde{U} \sim q_\phi(\cdot \mid O)$ and then evaluating $f_\theta(x^Q, E; \widetilde{U}, c_0)$, or equivalently by

$$p_{\phi,\theta}(y \mid O, x^Q) = \int_{\mathcal{U}} p_\theta(y \mid x^Q, u) q_\phi(u \mid O)\, du. \tag{12}$$

The tilde distinguishes a computational candidate from the actual world-side $U = u^\star$. Sampling $\widetilde{U}$ does not redraw physical identity, and q_ϕ is not automatically a Bayes or function posterior.

For factual supervised prediction, Equation (12) becomes $p_{\phi,\theta}(y \mid X^F = x, X^Q = x)$. For a same-token query at x', it is

$$p_{\phi,\theta}(y \mid X^F = x, X^Q = x') = \int p_\theta(y \mid x', u) q_\phi(u \mid x)\, du. \tag{13}$$

The abduction result remains the one obtained from factual x. Replacing it with $q_\phi(u \mid x')$ would re-recognize the unit from the alternative query and would not preserve same-token linkage.

The token–embedding one-to-one correspondence is a standing *identity-encoding modeling assumption*. It is not a consequence of one-shot data and does not identify numerical coordinates. This assumption is separate from response equivalence: distinct u values may induce the same map $x \mapsto p_\theta(\cdot \mid x, u)$ on the target query domain.

3.2 Observation regime

The training data are

$$\mathcal{D}_n = \{(O^{(i)}, X^{(i)}, Y^{(i)})\}_{i=1}^n, \tag{14}$$

with one factual event per sampled latent modeled unit. Models share $\mathcal{A}_\phi$ and f_θ across units, so learning is possible through population-level regularity. A unit identifier is neither needed nor treated as evidence.

The one-factual-event choice is deliberate. It matches the standard cross-sectional regime and, when X is an action, the potential-outcome observation regime: each sampled unit reveals an outcome only at its realized mechanism input, while responses at alternative inputs remain unobserved. Information about input dependence therefore comes from variation in the observed inputs $X^{(i)}$ across the population, together with shared structure in $(\mathcal{A}_\phi, f_\theta)$, rather than from repeatedly exposing each unit. Under the required support and learnability conditions, repeated observations of the same unit are neither assumed nor required for the primary response-law problem. This one-shot regime does not by itself identify individual alternative-intervention outcomes or their joint cross-world coupling.

Repeated observations would define a different data regime, not complete the primary one-shot design. If available, one could write $Y^{(i,r)} = f_\theta(X^{(i,r)}, E^{(i,r)}; u^{(i)}, c_0)$ and use held-out events to probe within-unit persistence or unit/event separation. Such repeated natural events are not equivalent to observing the same unit under alternative interventions, and they do not by themselves identify cross-world quantities. We treat this as an optional longitudinal extension rather than a prerequisite for Unit Mechanism Learning.

3.3 Causal semantics are optional and additional

If X is merely an observed predictor, $p_{\phi,\theta}(\cdot \mid o, x)$ is a structured predictive object. If X is a treatment or action, evaluating a new x has a causal interpretation only after the study specifies intervention semantics, consistency, support, an identification argument, mechanism stability, and any cross-world coupling needed by a joint counterfactual query. None of these conditions follows from Definition 1.

3.4 Evidence-only and explicit-input branches

Unit abduction supplies the shared primitive $O \mapsto \mathcal{A}_\phi(O)$, which may be a point or a candidate distribution. In the following compact composition notation, $\widetilde{U}$ is a learner-side computational candidate, whereas $U = u^\star$ remains the fixed world-side token realization. What enters the outcome mechanism after abduction is a separate modeling choice:

$$\text{evidence-only branch:} \quad \widetilde{U} \sim \mathcal{A}_\phi(O), \qquad \widetilde{Y} = f_\theta(E; \widetilde{U}, c_0), \tag{15}$$

$$\text{explicit-input branch:} \quad \widetilde{U} \sim \mathcal{A}_\phi(O), \qquad \widetilde{Y} = f_\theta(X^Q, E; \widetilde{U}, c_0). \tag{16}$$

In the first branch, observed features can change the predictive law only through the inferred distribution over U. In the second, X also enters the world-level outcome equation. Causal Regression belongs to the first branch with $O = X_{\text{obs}}$ and historical notation $Y = f^{\text{CR}}(U, \varepsilon)$. The Cauchy-bilinear model developed here belongs to the second. If $O = X = W$ is used in that model, the same observed vector has both an abduction path $W \to U \to Y$ and a direct mechanism path $W \to Y$, plus the W–U interaction. Those restrictions must not be attributed to Causal Regression.

3.5 Framework and specializations

The framework is the conditional-response target, the actual $u^\star$/learner-candidate u distinction, and the separation of O, X^Q, U, E; neither Cauchy noise nor a causal treatment is part of its definition. The models developed around this project fit into the following hierarchy:

Specialization	Mathematical restriction	Additional interpretation
Cauchy-affine unit mechanism	factorized Cauchy $\mathcal{A}_\phi(O)$ and a bilinear $f_\theta(X^Q, E; u, c_0)$	explicit-input predictive propagation
Causal Regression	$O = X_{\text{obs}}$ and $Y = f^{\text{CR}}(U, E)$, with no direct $X_{\text{obs}} \to Y$ path	evidence-only abduction-first prediction
Binary treatment	$X = T \in \{0, 1\}$	arm-specific response laws and, when identified, causal contrasts

Ordinary distributional regression is another observational specialization: it may take $O = X = W$ and learn the reduced conditional response without interpreting the latent factorization. Causal

Regression is not the binary-treatment row and does not assign intervention semantics to its observed regression features. These restrictions instantiate different branches of Unit Mechanism Learning; they do not define its general scope.

4 What One-Shot Data Can Learn

4.1 Prediction as projection

Let W denote the covariates visible to an ordinary predictor. They may be a coarsening of (O, X) and may include variables that are not mechanism inputs.

Proposition 1 (Observational projection)**.** *Suppose the data are generated in the fixed context by response law $\mathcal{R}_0(\cdot \mid O, X)$. For every measurable outcome event A and every regular conditional distribution,*

$$P_{\text{obs}}(Y \in A \mid W = w) = \int \mathcal{R}_0(A \mid o, x)\, P_{\text{obs}}(\,\mathrm{d}o,\, \mathrm{d}x \mid W = w). \tag{17}$$

If $W = (O, X)$, the integral reduces to $\mathcal{R}_0(A \mid o, x)$ almost surely.

Proof. Condition first on (O, X) and apply the tower property. The fixed-context data-generating assumption identifies the inner conditional probability with $\mathcal{R}_0(A \mid O, X)$. □

Thus ordinary prediction is not wrong; it is the projection retained by the chosen observation and sampling process. The reverse map is generally not unique.

4.2 Observational equivalence

Definition 3 (One-shot observational equivalence)**.** *Two Unit Mechanism Models $(\mathcal{A}, f)$ and $(\widetilde{\mathcal{A}}, \widetilde{f})$ are observationally equivalent on a domain if their response laws agree for almost every observed (o, x).*

Theorem 1 (Latent reparameterization)**.** *Let $h : \mathcal{U} \to \widetilde{\mathcal{U}}$ be a measurable bijection with measurable inverse. Define the pushforward $\widetilde{\mathcal{A}}(o) = h_{\#}\mathcal{A}(o)$ and*

$$\widetilde{f}(x, e, \widetilde{u}) = f\{x, e, h^{-1}(\widetilde{u})\}. \tag{18}$$

Then $(\mathcal{A}, f)$ and $(\widetilde{\mathcal{A}}, \widetilde{f})$ induce the same response law for every (o, x).

Proof. If the learner-side candidate $V \sim \mathcal{A}(o)$, then $\widetilde{V} = h(V)$ has law $h_{\#}\mathcal{A}(o)$. Substitution gives $\widetilde{f}(x, E, \widetilde{V}) = f(x, E, V)$ almost surely. □

The theorem rules out coordinate-level recovery claims without additional normalizations or semantics.

4.3 Observable coefficient geometry

The bilinear specialization has an observable reduction that clarifies both its relationship to varying-coefficient regression and the conditions under which its structural restriction can matter.

Proposition 2 (Varying-coefficient reduction)**.** *For the Cauchy-bilinear response law, define*

$$b_0(o) = \alpha + a^\top m_U(o), \tag{19}$$
$$b(o) = \beta + B m_U(o). \tag{20}$$

Then the response location is

$$m(o, x) = b_0(o) + b(o)^\top x, \tag{21}$$

and the evidence-indexed slope vectors satisfy

$$b(o) \in \beta + \mathrm{col}(B). \tag{22}$$

Consequently, response-location observations can identify at most the functions b_0 and b, not their particular factorization through m_U, a, and B.

Proof. Expand Equation (36):

$$m(o, x) = \alpha + \beta^\top x + \{a + B^\top x\}^\top m_U(o) \tag{23}$$
$$= \alpha + a^\top m_U(o) + x^\top \{\beta + B m_U(o)\}. \tag{24}$$

The slope displacement $b(o) - \beta = B m_U(o)$ belongs to $\mathrm{col}(B)$. The latent reparameterization result gives the non-uniqueness of the factorization. □

Corollary 1 (Identifiable slope subspace at minimal rank)**.** *Suppose the conditional response law is identified on a domain whose x-support contains an open set, so that the affine slope $b(o)$ in Equation (21) is identified for almost every o. Let*

$$\mathcal{S} = \mathrm{span}\{b(o) - b(o') : o, o' \in \mathcal{O}\} \tag{25}$$

have dimension r. Then $\mathcal{S}$ is an identifiable response-law functional. Every bilinear factorization contains $\mathcal{S} \subseteq \mathrm{col}(B)$; for a minimal-rank factorization with $\mathrm{rank}(B) = r$, the two subspaces are equal. The coordinates $m_U(o)$ and the basis B remain non-unique.

Proof. The slope is the gradient of the identified affine response location with respect to x, hence its affine hull and direction subspace $\mathcal{S}$ are identified. Equation (22) implies $\mathcal{S} \subseteq \mathrm{col}(B)$. If both spaces have dimension r, containment implies equality. Any invertible change of basis within this subspace changes the latent coordinates but not $\mathcal{S}$. □

This result isolates a positive target that survives latent non-identifiability: the response-relevant slope subspace, not its chosen coordinate system. It also explains why a scalar-input experiment with latent dimension at least one cannot by itself demonstrate a low-rank slope advantage; the subspace restriction becomes informative when the input dimension exceeds the response-relevant rank.

Proposition 3 (Unit/event scale is not identified)**.** *Consider the scalar specialization*

$$\widetilde{U} \mid O = o \sim \mathrm{Cauchy}\{m_U(o), \gamma_U(o)\}, \qquad E \sim \mathrm{Cauchy}(0, \gamma_E), \qquad \widetilde{Y} = \widetilde{U} + E, \tag{26}$$

with $\gamma_E > 0$. For any constant $0 < \delta < \gamma_E$, the alternative scales

$$\widetilde{\gamma}_U(o) = \gamma_U(o) + \delta, \qquad \widetilde{\gamma}_E = \gamma_E - \delta \tag{27}$$

induce exactly the same conditional response law.

Proof. Cauchy stability gives $\widetilde{Y} \mid O = o \sim \text{Cauchy}\{m_U(o), \gamma_U(o) + \gamma_E\}$. The alternative scales have the same positive sum. □

Only the total predictive scale is identified in this example. Architectural labels such as "candidate-unit uncertainty" and "event uncertainty" remain modeling commitments unless additional data or restrictions validate the split.

4.4 Response-law Fisher consistency

Theorem 2 (Log-score Fisher consistency)**.** *Let $r_0(y \mid o, x)$ be the true conditional response density and let $r(y \mid o, x)$ range over a dominated model class containing r_0. Assume the relevant log risks are finite. Then every minimizer of*

$$\mathcal{Q}(r) = \mathbb{E}[-\log r(Y \mid O, X)] \tag{28}$$

satisfies $r(\cdot \mid O, X) = r_0(\cdot \mid O, X)$ almost surely. The latent factorization producing r need not be unique.

Proof. Subtract the risk at r_0 and condition on (O, X):

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

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

This is a population statement about the response law. Finite-sample rates, misspecification, optimization, and latent recovery are separate questions.

5 A Cauchy-Bilinear Unit Mechanism

This section develops the paper's explicit-input tractable branch. It is not the Causal Regression outcome equation. Causal Regression uses observed features only as abduction evidence and then generates Y from U and event noise. Here X enters the outcome mechanism directly and interacts with U. The mathematical inheritance is limited to the abduction-first pattern and affine Cauchy propagation.

5.1 Cauchy mode of unit abduction

The evidence encoder outputs locations and positive scales:

$$m_U(O) = h_{\text{loc}}(O), \tag{30}$$

$$\gamma(O) = \gamma_{\min} + \text{softplus}\{h_{\text{scale}}(O)\}. \tag{31}$$

It directly defines the learner-side candidate distribution

$$\mathcal{A}_\phi(O) = \prod_{j=1}^{d} \text{Cauchy}\{m_{U,j}(O), \gamma_j(O)\}. \tag{32}$$

There is no population prior, likelihood for O, or variational KL penalty. The subscript ϕ appears only to distinguish learned parameters; the framework definition does not depend on this parameterization. The vector $m_U(O)$ is a Cauchy location, not an expectation, and $\gamma(O)$ is a scale, not a standard deviation. Neither the mean nor the variance exists. Relative to Gaussian abduction, the polynomial tails encode globally open candidate-unit uncertainty: factual evidence need not confine the actual token to a finite-variance neighborhood of a typical unit. This choice does not enlarge the outcome support beyond what the generator and event noise permit.

5.2 Simple shared parameterization of unit-specific mechanisms

For $X \in \mathbb{R}^p$, $U \in \mathbb{R}^d$, and $B \in \mathbb{R}^{p \times d}$, the world-side mechanism for the selected token is

$$Y(x) = \alpha + \beta^\top x + a^\top U + x^\top BU + \sigma E, \qquad E \sim \text{Cauchy}(0,1), \quad \sigma > 0. \tag{33}$$

After $U = u^\star$ is fixed, its realized outcome is obtained by inserting that same $u^\star$. The learner does not know it; for prediction the learner uses a computational candidate $\widetilde{U} \sim \mathcal{A}_\phi(O)$ in the same shared equation. The tilde marks epistemic computation, not a second world-side unit-selection variable. The coefficient on U at target input x is

$$g(x) = a + B^\top x. \tag{34}$$

Hence $a^\top U$ gives a unit-modulated intercept and $x^\top BU$ gives unit-modulated slopes. The mechanism remains a shared low-complexity family; the encoder, rather than a separate arbitrary function for every unit, carries the flexible evidence map. Removing the direct-input structure by setting $\beta = 0$ and $B = 0$ gives a no-input response law, but does not make the current architecture a faithful CausalEngine rewrite: event-noise placement, Action parameterization, and the learned system remain different.

5.3 Closed-form factual prediction

For factual supervised prediction, $O = X^F = x$ and the mechanism query is also $X^Q = x$. More generally, for evidence $O = o$ and query $x^Q = x$, conditional independence and affine stability yield

$$Y \mid O = o, X = x \sim \text{Cauchy}\{m(o,x), s(o,x)\}, \tag{35}$$

where

$$m(o,x) = \alpha + \beta^\top x + g(x)^\top m_U(o), \tag{36}$$

$$s(o,x) = \sum_{j=1}^{d} |g_j(x)| \gamma_j(o) + \sigma. \tag{37}$$

The factual objective is the Cauchy negative log likelihood

$$\ell(y; m, s) = \log(\pi s) + \log\left[1 + \left\{\frac{y-m}{s}\right\}^2\right]. \tag{38}$$

No candidate samples are required in training or prediction because the integral over $q_\phi(u \mid o)$ is analytic. For a same-token query x', the model reuses the parameters $m_U(o), \gamma(o)$ abducted from factual evidence and substitutes x' only into $g(x')$, $m(o,x')$, and $s(o,x')$. Recomputing m_U or γ from x' would violate the same-token query contract.

5.4 Software contract

The implementation exposes three operations:

$$\begin{aligned} \texttt{abduce(evidence)} &\longrightarrow (\mu_U, \gamma_U), \\ \texttt{response\_law(evidence,input)} &\longrightarrow (m_Y, s_Y), \\ \texttt{loss(evidence,input,outcome)} &\longrightarrow \text{factual NLL}. \end{aligned}$$

This contract keeps evidence and target input separate even when an observational application chooses $O = X = W$. The software field `mu_U` is a legacy API name for the Cauchy location m_U; it is not interpreted as a finite mean.

6 Properties of the Tractable Model

Proposition 4 (Cauchy-bilinear propagation)**.** *Under Equation (32), independent latent coordinates, and Equation (33), the response law is exactly the Cauchy distribution in Equations (35)– (37).*

Proof. For fixed x, Equation (33) is

$$\widetilde{Y} = \alpha + \beta^\top x + g(x)^\top \widetilde{U} + \sigma E, \qquad \widetilde{U} \sim \mathcal{A}_\phi(O). \tag{39}$$

Here $\widetilde{U}$ is the learner-side candidate integrated for prediction, not a redraw of the actual token. An affine combination of independent Cauchy variables is Cauchy. Locations add with signed coefficients, while scales add with absolute coefficients. This gives Equations (36) and (37). □

Proposition 5 (Binary-input specialization)**.** *For scalar $X = T \in \{0, 1\}$, the model has arm-specific latent weights $g(0) = a$ and $g(1) = a + B^\top$. It therefore recovers the two-arm Cauchy-affine response family used by abduction-based heterogeneous-treatment models, without making treatment semantics part of the base definition.*

Proof. Substitute $T = 0$ and $T = 1$ into $g(T) = a + B^\top T$. □

Corollary 2 (CDF, quantile, and coverage convergence)**.** *Fix (o, x). Suppose estimated response parameters $(\widehat{m}_n, \widehat{s}_n)$ converge in probability to (m_0, s_0) with $s_0 > 0$. Then the estimated Cauchy CDF converges pointwise at every finite threshold, every fixed quantile in $(0, 1)$ converges, and the coverage of a learned central $(1 - \alpha)$ quantile interval under the target response law converges to $1 - \alpha$.*

Proof. The Cauchy CDF

$$F_{m,s}(y) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{y - m}{s}\right) \tag{40}$$

and quantile

$$Q_q(m, s) = m + s \tan\{\pi(q - 1/2)\} \tag{41}$$

are continuous in (m, s) on $s > 0$. Apply the continuous mapping theorem. The target distribution is continuous, so convergence of both interval endpoints gives the coverage result. □

These results concern the induced response distribution. They do not imply that $m_U(O)$, $\gamma(O)$, a, B, or σ are separately identified.

7 Evaluation Design

The experiments are organized by what their data can validate. Every runner separates training data from evaluator-only truth and writes a frozen configuration and checksummed artifacts.

7.1 Single-row synthetic response laws

The primary benchmark creates a finite population of 6,000 units. Evidence is eight-dimensional, the latent response state is four-dimensional, and the mechanism input is scalar. Each training unit contributes exactly one factual triple (O_i, X_i, Y_i). The evaluator retains the generating response law on a 21-point target-input grid; neither the grid nor the oracle response parameters are passed to training or early stopping.

Accordingly, the benchmark fixes `events_per_unit` $= 1$ by design. The simulator can retain a response-law grid at alternative inputs because it has access to the declared data-generating process, but this grid is evaluator-only truth: it never enters fitting or model selection and does not count as additional factual events. Response-curve error therefore measures recovery of a simulator-defined target under the declared DGP; it is not evidence that alternative interventions were observed for real units or that individual counterfactuals are nonparametrically identified.

The matched scenario follows the Cauchy-bilinear family. Six stress scenarios remove heterogeneity, weaken evidence, narrow target support, introduce a nonlinear mechanism, replace Cauchy event noise by Gaussian noise, or contaminate 10% of outcomes. Configuration selection uses seeds 0–4; confirmatory evaluation uses seeds 100–109 after freezing the model configuration.

7.2 Baselines and ablations

The primary matrix contains the proposed Cauchy model, a structurally matched Gaussian model, a deterministic-unit ablation, direct Cauchy and Gaussian distributional predictors on (O, X), and a four-expert Cauchy mixture. The encoders use matched hidden widths, and parameter counts are reported. This matrix distinguishes the framework factorization from the heavy-tailed loss and from extra capacity.

7.3 Binary-treatment specialization

On IHDP, pretreatment covariates play the role of O and the binary treatment plays the role of X. Hyperparameters are selected on replications 0–9 without outcome-surface truth and frozen before evaluation on replications 100–199. The oracle fields are used only for final response-location contrast error. They are simulator truth, not observed effects of real individuals.

7.4 Observational regression sanity checks

California Housing, Concrete, Kin8nm, and Bodyfat use the observational special case $O = X = W$. Clean data and 20%/40% label-permutation and outlier conditions are evaluated over seeds 11, 22, 33, 44, and 55. These experiments can stress-test factual one-shot outcome-generation prediction, calibration, and sensitivity. By themselves they cannot identify the latent factorization, a repeated-unit mechanism, or intervention-level causal structure.

We additionally isolate robustness under marginal-preserving outcome shuffle. Each tabular row is one factual unit–event, so damaging its context–outcome pairing is a deliberate stress test of one-shot conditional outcome-generation prediction. The current protocol first creates disjoint train, validation, and test splits. It then selects a proportion ρ within train and, independently, the same proportion within validation, and applies a single-cycle derangement to the selected outcomes. Each split-local outcome multiset is preserved, selected links have no fixed points, and test outcomes remain clean. Validation is corrupted because it selects the training checkpoint; leaving it clean would change the training procedure, create a selection-regime mismatch, and make checkpoint selection optimistically cleaner. Feature transforms use train features only, and the outcome transform uses corrupted train outcomes only. All models share split and corruption fingerprints within a dataset, seed, and ratio.

The initial development matrix used $\rho \in \{0, .1, .2, .3, .4\}$ and ten models. A first frozen run retained a pooled development-pool corruption object; its labels could cross the later train/validation boundary and its scaling contract did not match the intended split-local estimand. We preserve that run as negative design history rather than current confirmation. Corrected fast V2 splits first and independently corrupts train and validation at ratios $\{0, .2, .4\}$ on fresh seeds 301–310. Extension V3

repeats ratios $\{0, .4\}$ on fresh seeds 401–410 for ten models: learned-scale direct Cauchy, Gaussian, and Laplace predictors; the current Cauchy Unit model; matched point Cauchy, Huber, median pinball, and MSE predictors; and hash-pinned repaired CausalEngine and MLP-Cauchy cores under the V3 training contract. V1, V2, and V3 rows are never pooled.

For candidate c, baseline b, dataset d, and seed s, V3 defines

$$D_{d,s}(c, b) = \log \frac{\mathrm{MAE}_{d,s,.4}(b)}{\mathrm{MAE}_{d,s,.4}(c)}, \tag{42}$$

$$G_{d,s}(c, b) = \log \frac{\mathrm{MAE}_{d,s,.4}(b)}{\mathrm{MAE}_{d,s,0}(b)} - \log \frac{\mathrm{MAE}_{d,s,.4}(c)}{\mathrm{MAE}_{d,s,0}(c)}. \tag{43}$$

Thus $D > 0$ means lower absolute clean-test MAE after damaged training, whereas $G > 0$ means less deterioration from the model's own clean condition. Each seed block averages the four datasets equally. Exact one-sided sign tests are computed separately for the ten D and G blocks, their intersection–union value is $\max(p_D, p_G)$, and Holm adjustment controls the frozen family of eight contrasts. A robust pass additionally requires positive paired Student-t interval lower bounds and positive mean D and G on every dataset. An advantage in D alone is labeled *absolute only*. These are prediction endpoints; the protocol does not test counterfactual validity, latent-state recovery, repeated-unit dynamics, or identification of the unit factorization.

7.5 Repaired-stack implementation audit

The clean-room models above are new designs rather than line-by-line rewrites of the historical CausalEngine or MLP-Cauchy. In particular, current Unit has a direct $X \to Y$ path and an X–U interaction, whereas CausalEngine uses X_{obs} only to abduct U before applying its Action map. Their comparison is therefore a full-system audit, not a matched test of unit factorization or a single Action component. Before freezing a confirmatory shuffle claim, we therefore run a five-arm development audit: current Unit v1, current learned-scale direct Cauchy, the exact repaired CausalEngine core under the current AdamW contract, and repaired CausalEngine and fixed-scale MLP-Cauchy estimators under the historical robust-shift hyperparameters. The last two arms use 128/64/32 ReLU networks, Adam, 500 epochs, 15% internal validation, patience 20, and batch size 256. They use the repaired source and current paired four-dataset protocol, not the frozen historical executable.

The repaired source is loaded read-only and fixed by source-tree SHA-256 plus five core-file hashes; it is not redistributed because no license could be confirmed for the snapshot. Every arm shares the outer split, shuffle mask, and clean test within a dataset–seed–ratio cell. The artifact separately records the actual fit split, scaler-fit indices, standardized data, model seed, runtime, and implementation hashes. Current Unit is rerun inside the audit and must replay the prior development artifact before aggregation is accepted. The repaired MLP-Cauchy optimizes $\log(1+r^2)$ at fixed standardized scale and returns a point prediction; its implied fixed-scale score is labeled separately rather than reported as learned distributional calibration.

Extension V3 reuses only two repaired cores and places them under the corrected V3 split, scaling, external validation, AdamW, and stopping contract. These are matched-protocol full-system comparisons, not exact executions of the historical training bundle. Architecture, latent width, Action law, direct-input path, event-noise semantics, activation, scale semantics, and parameter count can all differ, so a result against a repaired arm cannot be attributed to one component.

7.6 Metrics

We report negative log likelihood, mean absolute error of the predictive median, pinball loss at quantiles $0.1, 0.5, 0.9$, central 80%/90% interval coverage and width, probability-integral-transform calibration, CDF error, and response-location curve error when oracle curves exist. A non-degenerate Cauchy law has no finite mean or variance; CRPS also requires a finite first moment. We therefore avoid moment recovery and CRPS as primary metrics.

8 Current Executable and Empirical Evidence

The evidence record separates executable mechanics, development diagnostics, protocol provenance, corrected fresh-seed replay, and the final full-model extension. These stages answer different questions and their rows or effect estimates are never pooled. The closed-form response calculations, gradient path, evidence/input interface, and oracle-data firewall are executable and covered by tests. End-to-end smoke runs exercise the synthetic, IHDP, and real-regression pipelines; their row counts establish execution only.

Table 1: Executable pipeline status. Row counts are smoke checks, not confirmatory evidence.

Pipeline	Completed rows
Single-row synthetic	8
IHDP binary-input	4
California regression	8

Table 2: Synthetic smoke checks. These runs validate the executable pipeline and are not confirmatory performance claims.

Scenario	Model	NLL	80% cov.	Curve error
matched	cauchy_unit	1.616	0.833	0.169
matched	direct_cauchy	1.690	0.833	0.242
matched	deterministic_unit	2.037	0.500	0.233
matched	gaussian_unit	3.684	0.842	0.311
no_heterogeneity	cauchy_unit	0.829	0.817	0.038
no_heterogeneity	direct_cauchy	0.885	0.767	0.040
no_heterogeneity	deterministic_unit	0.910	0.767	0.058
no_heterogeneity	gaussian_unit	4.000	0.875	0.095

8.1 Development and repaired-stack diagnostics

The marginal-preserving shuffle development suite contains 1,000 paper-local rows across four datasets, five seeds, five ratios, and ten models. Direct Cauchy has lower severe-shuffle clean-test MAE than matched direct Gaussian in all eight 30%/40% dataset–ratio cells. This signal does not establish a unit-factorization advantage: only Kin8nm at 40% has a positive paired interval for the current Cauchy Unit model versus direct Cauchy, while the other seven severe cells are inconclusive.

The repaired-stack development audit adds 500 rows from a hash-pinned, read-only historical source under current paired data. Same-run current Unit replays the preceding development artifact

exactly in all 100 cells. Under the current training contract, repaired CausalEngine has 29.7% higher equal-dataset geometric MAE at 40% shuffle than current Unit. The repaired fixed-scale MLP-Cauchy is 28.1% worse than current learned-scale direct Cauchy at 40%. The effects are heterogeneous, and the historical-hyperparameter arms change training as well as architecture. They support an implementation audit conclusion, not universal inferiority or an Action-component attribution.

Table 3: Exact repaired CausalEngine core versus current Unit v1 under the current training contract. Values are five-seed mean clean-test MAE and are development evidence only.

Dataset	Unit clean	Repaired core clean	Unit 40%	Repaired core 40%
California	0.3466	0.3388	0.3800	0.3761
Concrete	4.1905	3.9809	5.2364	6.0602
Kin8nm	0.0553	0.0624	0.0637	0.0767
Bodyfat	0.6211	1.0794	2.2048	4.2163

Table 4: Repaired point MLP trained with fixed-scale Cauchy loss versus the current learned-scale direct Cauchy. Values are five-seed mean clean-test MAE and are development evidence only.

Dataset	Direct clean	Repaired MLP clean	Direct 40%	Repaired MLP 40%
California	0.3436	0.3233	0.3777	0.4138
Concrete	4.3839	3.9195	5.6005	6.3269
Kin8nm	0.0556	0.0554	0.0672	0.0862
Bodyfat	0.7254	1.1328	1.8788	3.0802

8.2 Protocol provenance and corrected V2 replay

The first frozen shuffle run, V1, completed its scheduled matrix and passed its then-declared direct-Cauchy versus direct-Gaussian endpoint. A subsequent audit showed that V1 deranged outcomes over a combined development pool before the train and validation rows were used. Outcome donors could therefore cross the intended train–validation boundary, and outcome normalization followed the clean pooled contract rather than corrupted training outcomes. V1 is retained with its checkpoint and outputs as negative design history, but we do not interpret its numerical estimate as evidence for the corrected split-local estimand.

Fast V2 was frozen after this defect and partial V1 outcomes were visible. It uses disjoint splits, independently deranges train and validation at the same ratio, fits transforms from training data only, leaves test clean, and uses fresh seeds 301–310. All 480 scheduled fits completed. At 40% corruption, direct Cauchy passed the narrow frozen clean-test MAE gate after training with corrupted train/validation outcomes against matched direct Gaussian: the equal-dataset seed-block log MAE ratio was 0.4721, with a 95% interval $[0.3903, 0.5540]$, 10/10 positive seed blocks, and exact one-sided sign-test $p = 0.0009766$. This corresponds to a 37.63% geometric reduction in clean-test MAE. The own-clean degradation log advantage was 0.1502 $[0.0019, 0.2984]$, but this descriptive component had 8/10 positive seed blocks and exact sign-test $p = 0.0547$. In contrast, current Cauchy Unit versus direct Cauchy was inconclusive at 40%: 0.0305 $[-0.0392, 0.1002]$, with 5/10 positive seed blocks. V2 is corrected post-unsealing prediction evidence, not a second pre-unsealing confirmation.

Table 5: Fresh-RNG robustness extension at 40% split-local outcome derangement. D is the clean-test MAE log advantage after training with corrupted train/validation outcomes and G is the own-clean degradation log advantage. Intervals are paired 95% Student-t intervals over ten equal-dataset seed blocks; p_{H} is the global Holm-adjusted intersection–union sign-test value over the eight frozen contrasts.

Candidate / baseline	D [95% CI]	G [95% CI]	Wins D/G	Data$^+$ D/G	p_{IUT}	p_{H}	Status
Direct Cauchy / Gaussian	0.534 [0.458, 0.611]	0.105 [0.022, 0.188]	10/8	4/3	0.0547	0.2188	Absolute only
Direct Laplace / Gaussian	0.376 [0.310, 0.443]	0.260 [0.155, 0.364]	10/10	4/4	0.0010	0.0078	Robust pass
Direct Cauchy / Laplace	0.158 [0.104, 0.212]	−0.155 [−0.279, −0.031]	10/2	4/3	0.9893	0.9893	Absolute only
Point Cauchy / MSE	0.319 [0.272, 0.366]	0.252 [0.139, 0.366]	10/9	4/4	0.0107	0.0752	Multiplicity
Point Huber / MSE	0.156 [0.127, 0.185]	0.160 [0.052, 0.268]	10/9	4/4	0.0107	0.0752	Multiplicity
Median pinball / MSE	0.420 [0.361, 0.480]	0.208 [0.073, 0.342]	10/9	4/3	0.0107	0.0752	Absolute only
Current Unit / repaired Engine	0.292 [0.219, 0.366]	0.076 [−0.017, 0.169]	10/7	3/3	0.1719	0.3438	Not passed
Direct Cauchy / repaired MLP-Cauchy	0.209 [0.150, 0.268]	0.085 [0.011, 0.158]	10/8	4/3	0.0547	0.2188	Absolute only

Data$^+$ counts datasets with positive mean effect. "Multiplicity" means both components pass their unadjusted gates but the frozen Holm family does not reject; "absolute only" means only the D component is claimable.

8.3 Fresh-RNG full-model extension V3

V3 was designed after V2 was unsealed. It reruns ten models at ratios zero and 40% on the four datasets with fresh seeds 401–410. All 800 scheduled fits completed without an observed execution failure. Prediction arrays, split and corruption fingerprints, model and runtime hashes, and the two repaired source adapters were verified before the aggregate was admitted. The eight registered contrasts are evaluated jointly with the D/G, interval, per-dataset, and Holm–IUT gates defined in the preceding section.

Only one contrast passes the full robustness gate. Learned-scale direct Laplace versus matched direct Gaussian has $D = 0.3764$ $[0.3100, 0.4429]$ and $G = 0.2598$ $[0.1554, 0.3642]$. Both components are positive in all ten seed blocks and on all four datasets. The exact intersection–union p-value is 0.0009766 and the Holm-adjusted value is 0.0078125. These effects correspond to a 31.37% geometric reduction in clean-test MAE after training with corrupted train/validation outcomes and a 22.88% relative reduction in own-clean degradation.

Direct Cauchy retains a strong absolute advantage over direct Gaussian: $D = 0.5345$, corresponding to a 41.40% geometric MAE reduction at 40%. It is nevertheless *absolute only*. The G component is positive in 8/10 seed blocks, its sign-test value is 0.0547, and Bodyfat has negative mean G. Direct Cauchy is also absolute-only against direct Laplace: its 40% MAE is lower, but its own-clean degradation advantage is negative ($G = -0.1549$). Thus the present evidence distinguishes lower clean-test error after training with corrupted train/validation outcomes from slower deterioration rather than using either quantity alone as "robustness."

Among matched point objectives, Cauchy and Huber each have positive component intervals and positive dataset means against MSE, but neither survives the frozen eight-comparison Holm correction (adjusted $p = 0.0752$). Median pinball has a strong absolute advantage over MSE, but Bodyfat has negative mean G. These are directional or absolute results, not registered robustness passes.

The observational results do not validate the unit factorization. Current Cauchy Unit does not pass against repaired CausalEngine: California has negative mean D and G, the seed-block G interval crosses zero, and $p_{\mathrm{IUT}} = 0.1719$. Current direct Cauchy has an 18.85% absolute clean-test MAE advantage after training with corrupted train/validation outcomes over repaired MLP-Cauchy but fails the G and Bodyfat direction gates. Because both repaired comparisons change multiple architecture and scale-semantic components, they cannot isolate factorization. The supported conclusion is absence of a demonstrated factorization advantage, not proof that factorization can never help.

Taken together, the strongest empirical statement concerns objective-family robustness under one specified pairing-damage process. Corrected V2 shows a large absolute clean-test MAE advantage after training with corrupted train/validation outcomes for direct Cauchy over Gaussian, while the stricter V3 family admits direct Laplace versus Gaussian as its sole joint absolute-and-degradation pass. None of these results establishes latent unit recovery, causal or counterfactual validity, or a general advantage for Unit Mechanism Learning over direct conditional prediction.

9 Discussion and Boundaries

Notation contract. The active ontology is model-level: U selects a token/unit and $U = u$ is its embedding realization, with shared mechanism $Y(x) = f_\theta(x, E; U)$. Observation indices are bookkeeping only. When the actual embedding must be distinguished from learner candidates we write $U = u^\star$ and reserve u inside $q_\phi(u \mid O)$ for a candidate value. Learner-side uncertainty concerns that same actual token; it is not a redraw of physical identity.

A framework can be useful without latent identification. The composed distribution $p_{\phi,\theta}(y \mid O, x) = \int p_\theta(y \mid x, u) q_\phi(u \mid O)\, du$ is sufficient for calibrated query-specific prediction. A simple shared parameterization of unit-specific mechanisms is also a useful inductive bias and software boundary. Neither benefit requires claiming that the learned coordinates are the unique true unit state. Interpretability claims must concern stable response behavior or validated summaries, not merely the existence of a latent vector.

A fixed modeled unit is not a fresh latent draw. Each sample is generated after one model-level selection $U = u^\star$, stable within the declared context and time scale. A point, Gaussian, or Cauchy abduction result is the learner's epistemic account of that fixed realization. Sampling $\widetilde{U}$ from a distributional result is a computational device; it does not redraw physical identity at every prediction, and q_ϕ is not a function posterior. For a new query about the same token, the factual evidence and its abduction result remain fixed while only the mechanism input changes.

Predictors are evidence and, factually, the query. In the basic supervised specialization, observing $X = x$ first supplies factual evidence x^F for recognizing the token. The same value is then used as $x^Q = x$ to compute its factual outcome distribution. For a same-token query, $x^F = x$ remains fixed and x^Q changes to x'. Running the abduction encoder on x' would answer a different question because the alternative input would be allowed to select a different candidate-token distribution.

Identity encoding and response equivalence are different. The current formulation requires different modeled units to have different representations through an identity-encoding modeling assumption. Its coordinates are not identified by one-shot data, and a future framework extension may relax this assumption. Within the current formulation, the map $u \mapsto \{x \mapsto p_\theta(\cdot \mid x, u)\}$ need

not be injective: distinct representations may induce the same response behavior over the target query domain. Prediction can therefore be adequate without recovering identity or selecting a unique response function.

One row per unit is a deliberate hard case. The regime matches many cross-sectional learning problems and the historical Causal Regression setting. That shared observation regime does not make the outcome mechanisms equal: Causal Regression is evidence-only, whereas the current bilinear model has direct and interactive mechanism-input paths. The one-shot regime also makes the limitations sharp. Repeated events, randomized target inputs, multiple proxies, known measurement models, or structural restrictions could identify more of the factorization. Those conditions should be studied explicitly rather than attributed to the base model.

The observable geometry can be weaker than the latent story. In the bilinear specialization, response locations reduce exactly to an evidence-varying intercept and slope. The latent factorization is therefore not empirically distinguished by location fit alone. Its observable restriction is that slope vectors occupy a low-dimensional affine subspace. This restriction is vacuous for a scalar input when the latent dimension is at least one, but can become testable for high-dimensional inputs with a smaller response-relevant rank. The identifiable object is then the slope direction subspace under the conditions of the preceding corollary, not a preferred latent basis.

Abduction does not decide the outcome-mechanism branch. The shared primitive is an evidence-indexed learner distribution over candidate representations of a fixed modeled unit. Causal Regression historically writes this composition as $Y = f^{\mathrm{CR}}(U, E)$; the present tractable model uses $Y = f_\theta(X, E; U, c_0)$. Observed features affect prediction in both models, but only the latter grants them a direct structural path after abduction. Calling both models abduction-first does not license collapsing their equations, causal graphs, or empirical comparisons.

Cauchy is a specialization, not the definition. Cauchy stability makes the first distributional-abduction model exactly tractable and produces polynomial tails. Unlike a Gaussian candidate law, a Cauchy law has a location and scale but no finite mean or variance. We use that geometry to encode globally open uncertainty over which individual mechanism is compatible with factual evidence, rather than local uncertainty around an average unit. It does not assert that every outcome is compatible with every unit; f_θ and the event law still determine outcome support. Point, Gaussian, mixture, flow-based, or implicit abduction laws remain legitimate alternatives. Empirical gains from a Cauchy likelihood must not be mislabeled as evidence for the entire Unit Mechanism Learning framework.

Outcome shuffle tests a predictive mapping, not identification. With one factual event per unit and $O = X = W$, split-local outcome derangement deliberately damages the unit-context–outcome pairing while preserving each training or validation outcome multiset. It asks whether a predictor learned from damaged pairings still predicts clean factual outcomes. It does not emulate treatment randomization, create alternative outcomes for the same unit, or reveal a separately identified latent unit state. Validation is corrupted because it selects the checkpoint and is therefore part of the training procedure; test remains the clean evaluation object.

The empirical headline is about objective families. Under this particular damage process, learned-scale Laplace provides the only multiplicity-controlled joint advantage in clean-test MAE

after training with corrupted train/validation outcomes and own-clean degradation over Gaussian in V3. Direct Cauchy repeatedly achieves lower absolute clean-test MAE under that training procedure, including in corrected V2, but its relative degradation is less stable across seeds and datasets in the full extension. The factorized model does not obtain a corresponding advantage. These results do not support a blanket claim that heavier tails, Cauchy likelihoods, or the framework are uniformly more robust.

Versioned evidence should not be collapsed. V1 records a historical pooled-corruption error; V2 is a corrected design chosen after that error and partial outcomes were visible; V3 is a fresh-seed post-V2 extension whose eight-comparison family is multiplicity-controlled but not pre-unsealing with respect to the earlier evidence. Keeping the versions separate is part of the scientific result. Their conclusions are conditional on four fixed datasets, frozen RNG streams, split-local outcome derangement, clean-test MAE, and the shared training contracts. They do not establish robustness to feature corruption, additive noise, missingness, selection bias, adversarial contamination, or a population of datasets.

A stronger mechanism claim needs a discriminating observation. The next experiment should not merely add more tabular benchmarks. It must introduce information or a restriction that makes a unit-mechanism prediction different from flexible direct conditional prediction—for example, high-dimensional inputs with a low-rank slope geometry, repeated factual events with held-out within-unit prediction, or explicitly designed multi-query support. The target should be an observable invariant or response functional, not an unvalidated latent coordinate.

Context is a separate axis. The present model fixes $C = c_0$. Context Mechanism Learning would ask how the same unit's mechanism changes across environments, institutions, domains, or regimes. It requires cross-context data and invariance or transport assumptions; adding a context variable to the notation is not enough.

Causal questions require extra assumptions. The stable realization $U = u^\star$ provides a place to express same-token linkage across alternative queries, while the fixed factual abduction result provides the learner-side approximation to that linkage. It does not identify that linkage from factual data. Binary and continuous treatment models are specializations whose causal interpretation depends on their study designs and cross-world semantics.

10 Conclusion

Unit Mechanism Learning treats each sample as arising from one selection $U = u^\star$ of a token with a stable response-relevant embedding. Predictors $X = x$ serve as factual evidence for recognizing that token and, in factual prediction, as the query at which its shared generator is evaluated. The learner returns a point, Gaussian, Cauchy, or other candidate-abduction object; distributional modes compose candidates with f_θ through $p_{\phi,\theta}(y \mid O, x) = \int p_\theta(y \mid x, u) q_\phi(u \mid O)\, du$. The framework separates evidence, explicit mechanism input when present, stable unit heterogeneity, event randomness, and context; it does not require different embeddings to induce different response behavior. It contains an evidence-only branch and an explicit-input branch; Causal Regression belongs to the former, while the Cauchy-bilinear model studied here belongs to the latter. It also makes its limits explicit: prediction is a projection of the finer model, but the latent factorization cannot generally be recovered from prediction alone. The Cauchy-bilinear model provides a first

tractable implementation. Current tabular evidence supports bounded objective-family robustness under damaged outcome pairings—most cleanly for learned-scale Laplace versus Gaussian—but does not validate the unit factorization. Stronger mechanism, causal, and cross-context claims require observably discriminating data and assumptions.

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] Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J. Rezende, S. M. Ali Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018.

[3] Mingming Gong. Distribution-consistency structural causal models. *arXiv preprint*, 2024. DiscoSCM manuscript.

[4] Guido W. Imbens and Whitney K. Newey. Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512, 2009. doi: 10.3982/ECTA7108.

[5] Wenxin Jiang and Martin A. Tanner. Hierarchical mixtures-of-experts for generalized linear models: Some results on denseness and consistency. In *Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics*, 1999.

[6] Fredrik D. Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 3020–3029, 2016.

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

[8] Seyed Mostafa Kia and Andre F. Marquand. Neural processes mixed-effect models for deep normative modeling of clinical neuroimaging data. In *Proceedings of the 2nd International Conference on Medical Imaging with Deep Learning*, pages 297–314, 2019.

[9] Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, S. M. Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. In *International Conference on Learning Representations*, 2019.

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

[11] 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.

[12] Rosa L. Matzkin. Nonparametric estimation of nonadditive random functions. *Econometrica*, 71(5):1339–1375, 2003.

[13] Hiroshi Morioka and Aapo Hyvärinen. Causal representation learning made identifiable by grouping of observational variables. In *Proceedings of the 41st International Conference on Machine Learning*, pages 36249–36293, 2024.

[14] Arash Nasr-Esfahany and Emre Kiciman. Counterfactual (non-)identification in structural causal models. *arXiv preprint arXiv:2301.09031*, 2023.

[15] Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Sch"olkopf. Learning independent causal mechanisms. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 4036–4044, 2018.

[16] 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.

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

[18] 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.

[19] 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.

[20] Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. General identifiability and achievability for causal representation learning. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, pages 2314–2322, 2024.

A Additional Proof Details

A.1 Projection under coarsened predictors

Let $I_A = \mathbf{1}\{Y \in A\}$. Under the response-law model,

$$P_{\text{obs}}(Y \in A \mid W) = \mathbb{E}[I_A \mid W] \tag{44}$$
$$= \mathbb{E}\{\mathbb{E}[I_A \mid O, X, W] \mid W\} \tag{45}$$
$$= \mathbb{E}\{\mathcal{R}_0(A \mid O, X) \mid W\}, \tag{46}$$

where the last equality uses the fixed-context generation assumption. This is Equation (17) written as a conditional expectation.

A.2 Why scale labels are not semantics

The constructive scale example extends beyond the identity mechanism. If $Y = b + wU + \sigma E$, the conditional scale is $|w|\gamma_U(o) + |\sigma|\gamma_E$. Any positive reallocation that preserves this total produces the same one-shot response density. Variation of w with an observed input can restrict the admissible reallocations, but separate identification still requires explicit conditions; it is not implied by fitting Equation (38).

B Reproducibility Checklist

- All manuscript tables are generated from paper-local outputs.
- Synthetic oracle arrays are stored outside learner-facing batches.
- Development and confirmatory seed sets are disjoint.
- V1, corrected V2, and extension V3 retain separate checkpoints, manifests, RNG streams, and effect estimates; no row is pooled across protocols.
- Corrected tabular robustness runs split first, independently derange train and validation outcomes within split, fit transforms from train only, and keep test outcomes clean.
- Extension V3 stores prediction-level arrays and recomputes aggregate metrics before evaluating the frozen eight-contrast family.
- Cauchy checks use CDFs and quantiles rather than empirical moments.
- Data fetchers record URLs and SHA-256 values when upstream artifacts provide stable hashes.
- The arXiv package excludes virtual environments, caches, raw downloads, and sibling-paper outputs.