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

中文问答导读版

Anonymous Author(s)

Working paper, July 16, 2026
Guide version: UML-GUIDE-ZH-2026-07-15-V7

Abstract

摘要导读

问：一个unit 只有一次事实观测，论文究竟声称能学到什么？ 答：world side 固定$U = u^\star$ 后由$f_\theta(x, E; u^\star)$ 生成outcome；learner 只能从evidence 得到关于同一$u^\star$ 的point 或candidate distribution，再与共享generator 组合。它不声称恢复唯一真实的潜坐标或物理机制。

问：看到predictors $X = x$ 后，预测中缺失的步骤是什么？ 答：先把x 当factual evidence 做unit abduction，再把同一个x 当factual query 输入generator。Counterfactual 时固定前一步，只把query 改成x'。

问：abduction 有哪三种形式？ 答：point $\widehat{u} = a_\phi(x)$、局部finite-variance 的Gaussian，以及有location/scale 但无finite mean/variance、保持globally open candidate uncertainty 的Cauchy。它们都不是重新抽physical identity。

问：O, X, U, E 为什么不能混为一组特征？ 答：O 是factual evidence，X^Q 是机制query，world-side $U = u^\star$ 选择token，learner candidate 另写为u 或$\widetilde{U}$，E 是事件随机性。Evidence-only branch 可以没有显式X^Q。

问：闭式Cauchy 模型与因果结论是一回事吗？ 答：不是。Cauchy–bilinear 只提供可解析的响应律；因果解释与跨context 外推都还需要额外设计、识别、稳定性或迁移假设。

问：当前tabular 实验最强的正向结论是什么？ 答：在corrected split-local pairing damage 下，V3 八个比较只有learned-scale Laplace 相对Gaussian 通过完整的absolute-error 与own-clean-degradation gate。

问：这是否验证了unit factorization？ 答：没有。Direct Cauchy 有明显absolute advantage，但factorized model 与repaired core 的比较都没有建立独立的结构优势。

Supervised learning usually estimates an observational predictive law such as $p(Y \mid X = x)$. We study the finer computation hidden inside it. At world level U selects a token/unit, $U = u^\star$ is that token together with its embedding, and $Y(x) = f_\theta(x, E; U)$. At learner level, predictors $X = x$ are first factual evidence about $u^\star$ and then, for factual prediction, the same x is the mechanism query. Abduction may return a point, a Gaussian candidate distribution, or a Cauchy candidate distribution. These are epistemic descriptions of the same actual token, not identity redraws. Distributional prediction composes them as $p_{\phi,\theta}(y \mid x^F, x^Q) = \int p_\theta(y \mid x^Q, u) q_\phi(u \mid x^F)\, du$. A same-token counterfactual holds the abduction result from factual x^F fixed and changes only x^Q to x'. We treat the shared, unit-modulated response family as the primary modeling object and call the 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 explicitly adopts a one-to-one unit–representation code as an identity-encoding modeling assumption, not as a coordinate system identified from one-shot data; a future framework extension may relax this assumption. 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 candidate distribution has learned location and scale but no finite mean or variance, encoding globally open rather than local Gaussian uncertainty. 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

本节导读

问：为什么普通$P(Y \mid W)$ 不足以回答本文的问题？ 答：它是观测投影，可能预测准确，却不必区分机制输入、稳定的unit-specific 差异与新事件噪声。

问：为什么同时需要$u^\star$ 与$q_\phi(u \mid x)$？ 答：$U = u^\star$ 是model-side actual token embedding；q_ϕ 是learner-side candidate uncertainty。后者的随机性不表示physical identity 被反复重抽。

问：为什么主动采用one-shot，而不是为每个unit 增加重复观测？ 答：`events_per_unit=1` 对应常见横截面与potential-outcome 数据结构；信息来自跨unit 的X variation 和共享机制，alternative-input responses 仍是missing，而不是训练样本。

问：Causal Regression 中的observed X 会直接生成Y 吗？ 答：不会。它只作为$O = X_{\text{obs}}$ 进入abduction；用本文统一的argument order，其evidence-only branch 写作$Y = f_\theta(E; U, c_0)$。Cauchy–bilinear model 属于另一条explicit-input branch。

问：本文最容易被误读成什么？ 答：把结构化生成模型直接当成因果识别或跨环境机制发现；二者都需要正文模型之外的额外假设。

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, write $U = u^\star$ to distinguish its fixed but unobserved embedding from a learner-side candidate u. The bridge to ordinary machine learning is unit abduction. Predictors $X = x$ first act as factual evidence x^F for recognizing the token and, in factual prediction, also act as the query $x^Q = x$. Three abduction modes make different commitments:

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

The Gaussian expresses local finite-variance candidate uncertainty. The Cauchy has location and scale but no finite mean or variance, so remote candidate mechanisms remain appreciably possible. This does not mean every unit can generate every outcome; f_θ and E still constrain outcomes.

Distributional prediction is

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

Equivalently, $\widetilde{U} \sim q_\phi(\cdot \mid x^F)$ is a computational candidate supplied to $f_\theta(x^Q, E; \widetilde{U}, c_0)$. It does not redraw the actual token, declare a population prior, or create a function posterior.

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 predictive map $u \mapsto \{x \mapsto p_\theta(\cdot \mid x, u)\}$ need not be injective. Ordinary prediction therefore outputs $p_{\phi,\theta}(y \mid x, x)$, not one selected u or one selected response function. For a same-token query at x', the learner must keep $q_\phi(u \mid x)$ fixed and change only the query:

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

Re-abducting from x' could select a different token and break Layer 3 linkage.

An explicit mechanism input is a branch choice rather than a universal requirement. In an evidence-only branch, the observed covariates enter only through abduction,

$$U \sim q_\phi(\cdot \mid O), \qquad Y = f_\theta(E; U, c_0). \tag{5}$$

Causal Regression belongs to this branch with $O = X_{\text{obs}}$: its observed features provide evidence about U but do not form a direct $X_{\text{obs}} \to Y$ mechanism path. The explicit-input grammar in Equation (1) is a sibling specialization, not a relabeling of that model.

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

$$\widetilde{U}_j \mid O \sim \text{Cauchy}\{m_{U,j}(O), \gamma_j(O)\}, \tag{6}$$

$$\widetilde{Y} = \alpha + \beta^\top X + a^\top \widetilde{U} + X^\top B \widetilde{U} + \sigma E, \qquad E \sim \text{Cauchy}(0,1). \tag{7}$$

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. Unlike the evidence-only Causal Regression branch, this specialization contains both a direct $X \to Y$ path and an X–U interaction. Their shared inheritance is candidate-representation abduction and analytic Cauchy propagation, not an identical 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.

本节导读

问：贡献是否只是"再加一个latent variable"？ 答：不是；重点是分离$O/X/U/E$、以响应律为目标，并正面限定one-shot 的不可识别边界。

问：它与Neural Processes 或混合效应模型的关键差别？ 答：这里每个unit 仅有一条事实事件，另用可采纳证据调制共享机制；并不声称取代这些成熟方法。

问：本文的q_ϕ 是Neural Process 式function posterior 吗？ 答：不是。它定义在同一actual token 的candidate embeddings 上；function-space law 只是派生pushforward，ordinary prediction 直接组合为$p_{\phi,\theta}(y \mid O, x)$。

问：SCM 是否只有shared f，而本文才有unit-specific f_u？ 答：不是。固定SCM 的exogenous realization 本来也诱导unit-specific response map；本文的新意在于把shared family 的unit-conditioned instances 设为primary objects，并把one-shot ordinary prediction 限定为reduced conditional response。

问：与深度因果模型的关系？ 答：基础框架可嵌入SCM，但刻意弱于完整因果模型；处理变量的介入或反事实含义必须另行激活。

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

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 $x \mapsto p_\theta(\cdot \mid x, u)$, but this map need not be injective in u; a function-space law is only the pushforward of q_ϕ, whereas ordinary prediction composes candidates 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 Regression lineage. Causal Regression provides the nearest internal starting point: observed features form a learner-side abduction law over candidate representations of the latent modeled unit, and its historical response notation is $Y = f^{\mathrm{CR}}(U, E)$ without a direct observed-feature path. The present Cauchy-bilinear model instead uses $Y = f_\theta(X, E; U, c_0)$ and adds both a direct input effect and an input–unit interaction. The two models share an abduction-first motif and analytic Cauchy propagation, but they are sibling specializations with different outcome-mechanism restrictions rather than successive versions of one tractable equation.

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

本节导读

问：model-side unit 与$O/X/U/E$ 各自承担什么角色？ 答：world-side $U = u^\star$ 选择token；O 是factual evidence，X^Q 指定机制query，E 是事件随机性，learner candidate 另写为u 或$\widetilde{U}$。

问：predictors $X = x$ 在ordinary prediction 中做了什么？ 答：它先以x^F 身份提供token-recognition evidence，再以$x^Q = x$ 身份输入generator。Counterfactual 固定从x^F 得到的abduction result，只改变query。

问：unit 与representation 必须一一对应吗？ 答：当前formulation 把injective unit-to-representation map 作为identity-encoding assumption，而不是one-shot identification。不同embeddings 仍可诱导相同predictive behavior。

问：为什么每个unit 默认只有一个factual event？ 答：`events_per_unit=1` 是本文主动研究的one-shot regime，而不是待修补的数据缺陷；它对应常见横截面数据，并在X 是action 时对应potential-outcome setting。未观测alternative inputs 只作为synthetic evaluator truth，不能进入训练。

问：evidence-only 与explicit-input 两条branch 的结构差别是什么？ 答：前者是$O \to U \to Y$，即$Y = f_\theta(E; U, c_0)$；后者还允许$X \to Y$ 与$X \times U \to Y$。Causal Regression 属于前者。

问：把X 换成另一个值就得到因果效应吗？ 答：不会；仍需介入语义、一致性、支持、识别、机制稳定性及必要的跨世界耦合。

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; one realization $U = u^\star$ denotes a token/unit together with its embedding. The shared mechanism is

$$Y(x) = f_\theta(x, E; U), \qquad Y_{u^\star}(x) := f_\theta(x, E; u^\star). \tag{8}$$

A row index is observation bookkeeping and does not enter this computation. The operational variables have distinct roles:

- $O \in \mathcal{O}$ is admissible factual evidence about the token;
- $X^Q \in \mathcal{X}$ is the explicit input of the mechanism call;
- $U = u^\star$ is the fixed actual token realization, while u denotes a learner-side candidate value;
- $E \in \mathcal{E}$ is event-level randomness after query and token are fixed.

Ordinary tabular regression uses the observed $X = x$ in both roles: $O = X^F = x$ for abduction and $X^Q = x$ for factual mechanism evaluation. A same-token query keeps $X^F = x$ fixed and changes only X^Q to x'.

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; invertible reparameterizations of u are observationally equivalent.

Definition 1 (Unit abduction)**.** *The learner may return a point $\widehat{u} = a_\phi(O)$, a Gaussian $q_\phi(u \mid O) = \mathcal{N}\{u; \mu_\phi(O), \Sigma_\phi(O)\}$, or a coordinatewise Cauchy candidate distribution. These describe uncertainty about the same actual $u^\star$ and do not redraw the token.*

For fixed u, $f_{\theta,u,c_0}(x,e) := f_\theta(x,e;u,c_0)$ is a unit-conditioned instance of the shared generator, not a free function fitted to one row. Define its event-noise-induced prediction by

$$p_\theta(\cdot \mid x,u) := \mathcal{L}_E\{f_\theta(x,E;u,c_0)\}. \tag{9}$$

Distributional abduction produces

$$p_{\phi,\theta}(y \mid O,x^Q) = \int_{\mathcal{U}} p_\theta(y \mid x^Q,u) q_\phi(u \mid O)\, du. \tag{10}$$

Equivalently, a computational candidate $\widetilde{U} \sim q_\phi(\cdot \mid O)$ is inserted into f_θ. Cauchy has location and scale but no finite mean or variance, so it represents globally open heavy-tailed candidate uncertainty; it does not imply that every outcome can be generated by every unit.

3.2 Evidence-only and explicit-input branches

The common primitive is evidence-to-unit abduction. In the following compact composition notation, $\widetilde{U}$ is a learner-side computational candidate rather than a newly sampled unit identity. Two distinct outcome-mechanism branches can follow it:

$$\text{evidence-only:} \quad \widetilde{U} \sim \mathcal{A}_\phi(O), \qquad \widetilde{Y} = f_\theta(E;\widetilde{U},c_0), \tag{11}$$

$$\text{explicit-input:} \quad \widetilde{U} \sim \mathcal{A}_\phi(O), \qquad \widetilde{Y} = f_\theta(X^Q,E;\widetilde{U},c_0). \tag{12}$$

Causal Regression belongs to the evidence-only branch with $O = X_{\text{obs}}$. Its observed covariates change the response law by changing the abducted distribution of U; they do not directly enter the world-level outcome mechanism. The Cauchy-bilinear model developed below is an explicit-input sibling. When its observational benchmark sets $O = X = W$, the same row covariates travel through $W \to \mathcal{A}_\phi(W) \to U \to Y$ and also through $W \to Y$, with an additional W–U interaction. Equality of data columns therefore does not imply equality of structural roles.

3.3 Observation regime

The training data are

$$\mathcal{D}_n = \{(O_i,X_i,Y_i)\}_{i=1}^n, \tag{13}$$

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 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 $Y^{(n,r)} = f_\theta(X^{(n,r)},E^{(n,r)};u^{(n)},c_0)$, with superscripts serving only as observation bookkeeping, 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.

3.4 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.

3.5 Framework and specializations

The framework is the conditional-response target, the actual $U = u^\star$/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. These restrictions instantiate Unit Mechanism Learning; they do not define its general scope. Causal Regression is not the binary-treatment row and does not acquire intervention semantics merely because its evidence is denoted by X.

4 What One-Shot Data Can Learn

本节导读

问：one-shot 数据能识别响应律，就能识别内部机制吗？ 答：不能；相同响应律可由不同潜在分解产生，观测预测并不存在唯一逆映射。

问：latent reparameterization 反例打掉了什么主张？ 答：任意可逆重参数化并同步改写机制都保持输出律不变，因此潜坐标本身没有无条件的唯一语义。

问：U/E 的scale split 为什么也不可识别？ 答：Cauchy 稳定性使unit-level scale 与event-level scale 只通过总尺度进入响应分布，正的尺度搬移可保持同一似然。

问：Fisher consistency 的边界在哪里？ 答：严格真响应密度在候选类内且可积性成立时，log score 在分布层面一致；它不保证参数、潜变量、因果机制或有限样本优化被识别。

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,*

$$P_{\text{obs}}(Y \in A \mid W = w) = \int \mathcal{R}_0(A \mid o, x)\, P_{\text{obs}}(\,\mathrm{d}o,\, \mathrm{d}x \mid W = w). \tag{14}$$

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 2 (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*

$$\widetilde{f}(x, e, \widetilde{u}) = f\{x, e, h^{-1}(\widetilde{u})\}. \tag{15}$$

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*

$$b_0(o) = \alpha + a^\top m_U(o), \tag{16}$$

$$b(o) = \beta + B m_U(o). \tag{17}$$

Then the response location is

$$m(o, x) = b_0(o) + b(o)^\top x, \tag{18}$$

and the evidence-indexed slope vectors satisfy

$$b(o) \in \beta + \mathrm{col}(B). \tag{19}$$

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 (33):

$$m(o,x) = \alpha + \beta^\top x + \{a + B^\top x\}^\top m_U(o) \tag{20}$$

$$= \alpha + a^\top m_U(o) + x^\top\{\beta + B m_U(o)\}. \tag{21}$$

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 (18) is identified for almost every o. Let*

$$\mathcal{S} = \mathrm{span}\{b(o) - b(o') : o, o' \in \mathcal{O}\} \tag{22}$$

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 (19) 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*

$$\widetilde{U} \mid O = o \sim \mathrm{Cauchy}\{m_U(o), \gamma_U(o)\}, \qquad E \sim \mathrm{Cauchy}(0, \gamma_E), \qquad \widetilde{Y} = \widetilde{U} + E, \tag{23}$$

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

$$\widetilde{\gamma}_U(o) = \gamma_U(o) + \delta, \qquad \widetilde{\gamma}_E = \gamma_E - \delta \tag{24}$$

induce exactly the same conditional response law.

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

Only the total response scale is identified in this example. Architectural labels such as "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*

$$\mathcal{Q}(r) = \mathbb{E}[-\log r(Y \mid O, X)] \tag{25}$$

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):

$$\mathcal{Q}(r) - \mathcal{Q}(r_0) = \mathbb{E}[\mathrm{KL}\{r_0(\cdot \mid O, X) \,\|\, r(\cdot \mid O, X)\}] \geq 0. \tag{26}$$

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

本节导读

问：为什么在Gaussian 之外选择Cauchy abduction? 答：Gaussian 表达围绕中心的局部finite-variance uncertainty；Cauchy 有location/scale 但无finite mean/variance，用heavy tails 保留远端candidate mechanisms。它不保证任何outcome 都能由任何unit 生成。

问：闭式响应律具体避免了什么? 答：训练与预测无需latent Monte Carlo；可直接计算Cauchy NLL、CDF 与quantile。

问：解析方便是否意味着内部参数可解释? 答：不意味着；闭式可计算性与$m_U, \gamma, a, B, \sigma$ 的分别可识别性是两件事。

问：这一节是在重写Causal Regression 吗? 答：不是。这里主动选择explicit-input branch；$\beta^\top X$ 与$X^\top BU$ 在Causal Regression 的evidence-only outcome equation 中都不存在。

This section instantiates the explicit-input branch. It is not the Causal Regression outcome equation: here X acts directly on Y and interacts with U, whereas Causal Regression uses observed X only as abduction evidence.

5.1 Cauchy mode of unit abduction

The evidence encoder outputs locations and positive scales:

$$m_U(O) = h_{\mathrm{loc}}(O), \tag{27}$$

$$\gamma(O) = \gamma_{\min} + \mathrm{softplus}\{h_{\mathrm{scale}}(O)\}. \tag{28}$$

It directly defines

$$\mathcal{A}_\phi(O) = \prod_{j=1}^{d} \mathrm{Cauchy}\{m_{U,j}(O), \gamma_j(O)\}. \tag{29}$$

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. Here $m_U(O)$ is a Cauchy location rather than an expectation, and $\gamma(O)$ is a scale rather than a standard deviation.

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 is

$$Y(x) = \alpha + \beta^\top x + a^\top U + x^\top BU + \sigma E, \qquad E \sim \mathrm{Cauchy}(0, 1), \quad \sigma > 0. \tag{30}$$

Fixing $U = u^\star$ computes the actual token outcome. Prediction instead inserts a computational candidate $\widetilde{U} \sim \mathcal{A}_\phi(O)$ into the same shared equation; this does not redraw the world-side token. The coefficient on U at target input x is

$$g(x) = a + B^\top x. \tag{31}$$

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. Setting $\beta = 0$ and $B = 0$ removes the explicit-input terms and yields a no-input response law. This algebraic degeneration is not a faithful rewrite of the repaired CausalEngine system, whose Action map, event-noise placement, latent width, and training contract are also different.

5.3 Closed-form factual prediction

Conditional independence and affine stability yield

$$Y \mid O = o, X = x \sim \text{Cauchy}\{m(o,x), s(o,x)\}, \tag{32}$$

where

$$m(o,x) = \alpha + \beta^\top x + g(x)^\top m_U(o), \tag{33}$$

$$s(o,x) = \sum_{j=1}^{d} |g_j(x)| \gamma_j(o) + \sigma. \tag{34}$$

The factual objective is the Cauchy negative log likelihood

$$\ell(y; m, s) = \log(\pi s) + \log\left[1 + \left\{\frac{y-m}{s}\right\}^2\right]. \tag{35}$$

No candidate samples are required in training or prediction. For a same-token query x', reuse $m_U(o), \gamma(o)$ abducted from factual evidence and replace only x by x' in g, m, s; re-running abduction on x' would break the same-token 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 Cauchy location m_U, not a finite mean.

6 Properties of the Tractable Model

本节导读

问：闭式传播依赖哪些脆弱条件？ 答：固定x 后机制须对独立Cauchy 坐标仿射，尺度按系数绝对值相加；换分布或依赖结构会改变代数。

问：CDF、quantile 与coverage 收敛结论说了多少？ 答：它们由响应位置/尺度的一致性和连续映射得到，只约束输出分布，不反推潜在分解。

问：二元输入特化自动成为treatment-effect 模型吗？ 答：不会；代数上可形成两臂响应族，但treatment semantics 与识别仍来自额外研究设计假设。

Proposition 4 (Cauchy-bilinear propagation)**.** *Under Equation* (29)*, independent latent coordinates, and Equation* (30)*, the response law is exactly the Cauchy distribution in Equations* (32)*–* (34)*.*

Proof. For fixed x, Equation (30) is

$$\widetilde{Y} = \alpha + \beta^\top x + g(x)^\top \widetilde{U} + \sigma E, \qquad \widetilde{U} \sim \mathcal{A}_\phi(O). \tag{36}$$

Here $\widetilde{U}$ is a computational candidate, 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 (33) and (34). □

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

$$F_{m,s}(y) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{y - m}{s}\right) \tag{37}$$

and quantile

$$Q_q(m, s) = m + s \tan\{\pi(q - 1/2)\} \tag{38}$$

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

本节导读

问：truth-isolated evaluation 防止了哪种泄漏？ 答：训练和早停对每个unit 只见一条事实记录；oracle 响应曲线由评估器隔离保存，只在冻结后评分。

问：oracle 有多个target inputs，为什么仍然是one-shot？ 答：每个unit 的训练数据仍只有一个factual event；alternative-input grid 是simulator 私藏的评分真值，不是额外观测，也不能用来选模型。

问：为什么开发种子与确认种子必须分开？ 答：否则反复调参会把evaluator truth 变相带回模型选择，削弱所谓确认性结果。

问：观测回归数据能检验Unit Mechanism 的哪一层？ 答：它能stress-test one-shot conditional outcome-generation prediction、校准与鲁棒性；但当$O = X = W$ 且每个unit 只有一次事实事件时，不能单独识别latent factorization、repeated-unit mechanism 或intervention-level causal structure。

问：tabular shuffle 是实验错误，还是有意设计？ 答：它是有意的split-local pairing damage。Train 和validation 各自在split 内独立打乱部分outcome，同时保持各自outcome multiset 不变；它检验的是prediction-level outcome-generation robustness，不是intervention。

问：`events_per_unit=1` 与$O = X = W$ 在这里意味着什么？ 答：每行是一名unit 的一次factual event，unit evidence 与mechanism input 共用行索引；可学习的是跨unit 的one-shot conditional outcome-generation law，不是repeated-event dynamics。

问：为什么validation 也必须加noise？ 答：validation 用于checkpoint selection，本质上属于训练流程。协议先split，再在train 与validation 内按同一比例、不同seed 独立derange outcome；test 始终clean。

问：V1 的设计问题是什么？ 答：V1 在合并development pool 上先shuffle，label donor 可以跨越后来的train/validation 边界，scaler contract 也不是corrupted-train-only。它不是“validation 没加噪”，而是没有split-local 独立加噪。

问：这次是否真的复用了过去的核心实现？ 答：是。实验以source-tree 与五个core-file hashes 固定repaired CausalEngine，并直接调用旧forward/loss；同时加入旧MLP-Cauchy。源码只读加载，不在论文包中分发。

问：为什么还要在同一runner 重跑当前模型？ 答：这样split、mask、scaler、seed 和runtime 都成为同一artifact 的可核验证据；current Unit 还必须逐cell 复现先前development rows。

问：V3 的完整robustness gate 为什么同时需要D 与G？ 答：D 检查train/validation outcomes 均被derange 的训练流程之后的absolute clean-test MAE，G 检查相对自身clean condition 的退化；只过D 的结果只能标为`absolute_only`。

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

$$D_{d,s}(c,b) = \log \frac{\mathrm{MAE}_{d,s,.4}(b)}{\mathrm{MAE}_{d,s,.4}(c)}, \tag{39}$$

$$G_{d,s}(c,b) = \log \frac{\mathrm{MAE}_{d,s,.4}(b)}{\mathrm{MAE}_{d,s,0}(b)} - \log \frac{\mathrm{MAE}_{d,s,.4}(c)}{\mathrm{MAE}_{d,s,0}(c)}. \tag{40}$$

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. 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. In particular, the current Unit model has a direct $X \to Y$ term and an X–U interaction, while Causal Regression uses the tabular features only as evidence for abduction before an evidence-only response mechanism. The audit is therefore a matched-protocol full-system comparison, not a matched factorization or Action-law comparison.

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

本节导读

问：V1、V2、V3 为什么不能合并？ 答：V1 是pooled-corruption negative design history；V2 在缺陷与部分结果可见后采用corrected split-local protocol；V3 又在V2 后注册八个新比较。它们的estimand、gate 与RNG streams 不同。

问：V2 修正后给出什么结果？ 答：Direct Cauchy 相对direct Gaussian 在40% pairing damage 下把clean-test MAE 几何降低37.63%，D 为10/10 wins；但V2 是post-unsealing correction。

问：V2 的G 是否也确认通过？ 答：没有。V2 的G 均值与区间为正，但只有8/10 wins，exact sign $p = 0.0547$；它是descriptive component，不是V2 primary gate。

问：V3 最强的完整结论是什么？ 答：800 个fits 全部完成，八个注册比较中只有learned-scale direct Laplace 相对matched Gaussian 通过完整gate：MAE 降低31.37%，own-clean degradation 相对降低22.88%，Holm-adjusted $p = 0.0078125$。

问：Direct Cauchy 是否也完整通过？ 答：没有。它相对Gaussian 的absolute MAE 降低41.40%，但G 只有8/10 wins，Bodyfat 的平均G 为负，因此是`absolute_only`。

问：Cauchy 与Laplace 如何比较？ 答：Cauchy 在40% 下的absolute MAE 更低约14.62%，但相对自身clean 的退化反而更差；这正说明absolute performance 与degradation robustness 不能混为一个词。

问：普通robust loss 是否通过？ 答：Point Cauchy 与Huber 对MSE 的方向一致，但全局Holm 后都是$p = 0.0752$；pinball 有absolute advantage，却在Bodyfat 的G 反向。它们都不是V3 robust pass。

问：这些结果验证了unit factorization 吗？ 答：没有。V2 factorized-versus-direct 不确定；V3 current Unit 相对repaired CausalEngine 也未通过。正确结论是“尚未展示factorization advantage”，不是“已证明factorization 无用”。

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.

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.

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.

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

本节导读

问：latent 不可识别是否让框架失去价值？ 答：不会；可校准的$p_{\phi,\theta}(y \mid O, x)$、unit-conditioned generator 的共享参数化与清晰软件接口仍有价值，但解释应落在可验证响应行为上。

问：从$q_\phi(u \mid O)$ sample candidate 是否表示identity 被重新抽取？ 答：不是。world-side $U = u^\star$ 固定；$\widetilde{U}$ 只服务learner-side computation。Same-token query 固定factual evidence 给出的abduction result。

问：不同unit 是否必须有不同response function？ 答：不必须。Unit–representation injectivity 是identity-encoding assumption，但不同embeddings 仍可以产生相同predictive behavior。

问：Cauchy 上的经验收益能代表整个框架吗？ 答：不能；Cauchy 只是便于闭式传播的specialization，分布选择收益不能偷换成框架验证。

问：为什么pairing damage 不是causal mechanism identification？ 答：它只破坏unit context 与factual outcome 的观测配对，并在clean test 上测预测；它没有给同一unit 产生alternative outcomes，也不是treatment randomization。

问：这轮empirical headline 应如何表述？ 答：它是objective-family robustness：Laplace/Gaussian 完整通过，Cauchy/Gaussian 只有absolute advantage，unit factorization 尚未建立优势。

问：下一轮什么实验才可能验证unit mechanism 的额外价值？ 答：需要一个direct predictor 无法轻易吸收的observable discriminator，例如high-dimensional low-rank slope geometry、held-out within-unit events，或明确的multi-query support；目标应是可验证response functional，而不是给latent 坐标命名。

问：Causal Regression 与本文模型共享什么、不共享什么？ 答：共享abduction-first 表示与Cauchy 仿射传播；不共享outcome mechanism。前者是evidence-only，后者有direct-input 与input–unit interaction。

问：跨context 只需把C 加进网络吗？ 答：不够；跨环境机制变化需要跨context 数据以及不变性、稳定性或transport 假设。

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 selection $U = u^\star$, stable within the declared context and time scale. Point, Gaussian, and Cauchy abduction results are alternative epistemic descriptions of that same fixed token. Sampling $\widetilde{U}$ is a computational device, not a physical identity redraw. For a new same-token query, factual evidence and its abduction result remain fixed while only the mechanism input changes.

Predictors are evidence and, factually, the query. Observing $X = x$ first supplies x^F for token recognition and then $x^Q = x$ for factual mechanism evaluation. A counterfactual query keeps $x^F = x$ fixed and changes only x^Q to x'. Abducting again from x' could switch tokens.

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 observation regime. It does not imply the same outcome mechanism: Causal Regression is evidence-only, whereas the current bilinear specialization is explicit-input. The one-row 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.

Cauchy is a specialization, not the definition. Cauchy stability makes the first distributional-abduction model exactly tractable. It has location and scale but no finite mean or variance, expressing globally open rather than local Gaussian candidate uncertainty. It does not make every outcome compatible with every unit; the generator and event law still determine outcome support. Point, Gaussian, mixture, flow-based, or implicit abduction laws remain legitimate alternatives.

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 fixed factual abduction supplies the learner-side approximation. 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.

Abduction does not decide the outcome-mechanism branch. The shared primitive $O \mapsto \mathcal{A}_\phi(O)$ is learner-side uncertainty over candidate representations of a fixed modeled token. A computational candidate $\widetilde{U}$ supports both an evidence-only prediction $\widetilde{Y} = f_\theta(E; \widetilde{U}, c_0)$ and an explicit-input prediction $\widetilde{Y} = f_\theta(X^Q, E; \widetilde{U}, c_0)$. Causal Regression instantiates the former; the Cauchy-bilinear model studied here instantiates the latter. Analytic Cauchy propagation survives in both, but direct input and input–unit interaction claims do not transfer across the branch boundary.

10 Conclusion

结论导读

问：全文最应保留的一句话是什么？ 答：world side 由固定$U = u^\star$ 与f_θ 计算unit outcome；learner side 先从factual predictors abduct 同一token，再在query 上组合prediction。

问：理论、实现与证据应如何分层？ 答：Cauchy–bilinear 给闭式机制，truth-isolated protocol 给评估纪律；V2/V3 支持bounded objective-family robustness，但仍不支持factorization、latent recovery 或因果结论。

问：下一步何时才能谈因果或跨context？ 答：只有加入相应介入/识别或迁移/不变性假设并提供匹配数据后，才能升级这些主张。

Unit Mechanism Learning treats each sample as arising from one token selection $U = u^\star$. Predictors $X = x$ first serve as factual evidence for recognizing that token and, in factual prediction, as the generator query. Point, Gaussian, Cauchy, or other abduction objects are then composed with f_θ; distributional modes yield $p_{\phi,\theta}(y \mid O, x)$ by integrating candidate predictions. 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 predictive behavior. It contains distinct evidence-only and explicit-input branches: Causal Regression belongs to the former, while the Cauchy-bilinear model provides a first tractable implementation of 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 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

附录合并导读

问：Additional Proof Details 补强了哪条警告？ 答：即便推广到$Y = b + wU + \sigma E$，只要总Cauchy 尺度不变，U/E 尺度标签仍无自动语义；输入变化至多施加限制，不能免除识别条件。

问：Reproducibility Checklist 的核心防线是什么？ 答：论文表格只读paper-local outputs，oracle 与learner 隔离，V1/V2/V3 分开保存，并从prediction arrays 重算V3 指标与八个比较。

问：清单通过后是否等于科学主张成立？ 答：不等于；它保证可追踪与防泄漏，确认性结论仍需预定矩阵、足够运行与合适统计比较。

A.1 Projection under coarsened predictors

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

$$P_{\text{obs}}(Y \in A \mid W) = \mathbb{E}[I_A \mid W] \tag{41}$$
$$= \mathbb{E}\{\mathbb{E}[I_A \mid O, X, W] \mid W\} \tag{42}$$
$$= \mathbb{E}\{\mathcal{R}_0(A \mid O, X) \mid W\}, \tag{43}$$

where the last equality uses the fixed-context generation assumption. This is Equation (14) 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 (35).

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.