COUNTERFACTUAL RESPONSE MODELING WITH A BINARY TREATMENT: ABDUCTION-BASED ESTIMATION OF HETEROGENEOUS TREATMENT EFFECTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating a treatment effect conditional on covariates requires more than predicting an observed outcome: the model must represent the same unit under two treatment-indexed worlds. We introduce an abduction-based heterogeneous treatment-effect estimator (AB-HTE) that first infers a distribution over a latent unit representation from pretreatment covariates and then evaluates a shared treatment-indexed structural mechanism. Our primary Cauchy AB-HTE instantiation uses an affine mechanism in the abducted representation and a factorized Cauchy abduction family. This combination yields closed-form model predictive outcome distributions and a differentiable conditional median (location) contrast. Under consistency, exchangeability, positivity, and correct marginal specification, the population factual risk recovers these potential-outcome marginals. It does not reveal the unobserved pair of potential outcomes for any unit. We separate these identified marginal targets from an individual-effect distribution, which additionally depends on a non-identified cross-world coupling. In particular, a non-degenerate Cauchy effect has no ordinary mean, so the primary estimand is a conditional median contrast rather than an informally relabeled CATE. In a ten-seed nonlinear synthetic study that independently corrupts training and validation at the same rate, Cauchy AB-HTE has the smallest degradation under gross-outlier replacement and substantially improves over matched Gaussian AB-HTE, while direct robust learners remain better for several other corruptions. On 100 held-out IHDP simulations, Cauchy AB-HTE improves over its Gaussian twin and direct robust S-learners but not TARNet/CFR-style baselines. On the official 4,218-person WAWS randomized-trial window, Cauchy AB-HTE has the lowest factual MedAE; the aggregate ITT is imprecise, and learned-policy values are reported only as cross-fitted point diagnostics, not formal inference. A hidden-confounding negative control increases effect error by over twenty-fold. The recovery theorem is for an unpenalized population risk; finite-sample consistency and a fixed nonzero regularizer are not covered.

1 INTRODUCTION

For a unit with pretreatment covariates $X = x$, treatment choice concerns two outcomes that are never jointly observed: $Y(1)$ and $Y(0)$. Standard learners estimate conditional response surfaces or their mean difference (Shalit et al., 2017; Künzel et al., 2019; Shi et al., 2019); distributional learners extend this target to potential-outcome laws and risk functionals (Park et al., 2021; Zhou et al., 2022; Kallus & Oprescu, 2023; Ma et al., 2024). Yet the two marginal laws do not determine the distribution of the within-unit difference $Y(1) - Y(0)$, even in a randomized experiment (Fan & Park, 2010; Melnychuk et al., 2024). A point-valued individual-effect law therefore requires an additional cross-world coupling.

We introduce an *abduction-based heterogeneous treatment-effect estimator* (AB-HTE) that keeps marginal identification and cross-world modeling separate. Its primary Cauchy instantiation is an analytically tractable model: an encoder maps $X = x$ to a factorized Cauchy distribution $q_\phi(u \mid x)$

over candidate embeddings for the same actual token, and a binary-treatment mechanism specifies

$$Y(t) = a_0 + a^\top U + t(b_0 + b^\top U) + \sigma_t E_t, \qquad E_t \sim \text{Cauchy}(0, 1). \tag{1}$$

The ontology has two sides. In the data-generating world, $U = u^\star$ selects one actual token/unit embedding and its outcome is generated by $Y_{u^\star}(t) = f_\theta(t, E; u^\star)$. On the learner side, $X = x$ is factual evidence and $q_\phi(u \mid x)$ is epistemic uncertainty about that same unobserved $u^\star$; it is not a population prior and does not redraw the physical unit. Factual prediction first abduces from x and then queries the mechanism at the observed treatment. A counterfactual query keeps that abduction result fixed and changes only t. The interaction $tb^\top U$ carries effect heterogeneity. Cauchy stability makes both arm-specific predictive outcome distributions analytic, so the primary target is the conditional marginal location contrast

$$\tau_{\text{loc}}(x) = \text{med}\{Y(1) \mid x\} - \text{med}\{Y(0) \mid x\} = b_0 + b^\top \mu_\phi(x). \tag{2}$$

A non-degenerate Cauchy effect has no ordinary mean; we therefore do not relabel this quantity as CATE. Under consistency, conditional exchangeability, positivity, correct marginal specification, and an unpenalized population objective, factual log-risk recovers the two predictive distributions and hence this marginal contrast. It does not recover either missing potential outcome for an observed individual or validate the latent coordinates.

Cauchy AB-HTE can additionally impose either independent event resampling or a shared exogenous realization across treatment worlds. Each declared coupling yields an analytic Cauchy law, or its degenerate point-mass limit, for the within-unit contrast, but factual marginals select neither coupling. We report such quantities as model-implied sensitivity analyses and derive sharp Makarov bounds for assumption-lean comparison. Separately, the Cauchy location and log-scale residual scores are bounded; a parameter-gradient bound further requires controlled model Jacobians. Thus the robustness claim concerns tested outcome contamination, not confounding, overlap failure, or unconditional bounded influence.

The paper contributes: (i) a treatment-indexed Cauchy–affine model with closed-form predictive distributions and location/quantile contrasts; (ii) an explicit identification chain from causal assumptions through Fisher-consistent factual risk, together with non-identification of the latent factorization and cross-world effect law; and (iii) a layer-valid evaluation that separates synthetic coupling truth, semi-synthetic response-surface truth, and factual-only real data. Every reported cell is generated from a completeness-checked, machine-readable artifact under the truth boundary appropriate to that layer.

2 Related Work

Heterogeneous-effect learning. Classical and modern approaches include causal forests, meta- and R-learners, and representation learners such as TARNet/CFRNet and Dragonnet (Wager & Athey, 2018; Künzel et al., 2019; Nie & Wager, 2021; Shalit et al., 2017; Shi et al., 2019). Cauchy AB-HTE also shares information across arms, but uses a distribution-valued unit representation, an affine treatment–unit interaction, and exact Cauchy propagation. Its decisive control is a similar-capacity direct Cauchy S-learner; the exact parameter counts are reported rather than asserted equal.

Distributional and generative counterfactual models. GANITE, latent-variable and deep structural causal models learn counterfactual generators or implement abduction–action–prediction (Yoon et al., 2018; Louizos et al., 2017; Pawlowski et al., 2020). Distributional learners estimate potential-outcome laws, quantiles, or risk functionals with kernels, collaborating networks, orthogonal losses, or diffusion models (Park et al., 2021; Zhou et al., 2022; Kallus & Oprescu, 2023; Ma et al., 2024). Cauchy AB-HTE sacrifices flexible density approximation for closed-form Cauchy kernels and coupling-explicit sensitivity analysis.

Cross-world dependence and robustness. Potential-outcome marginals only partially identify the effect distribution; Makarov bounds and recent orthogonal learners expose this boundary (Fan & Park, 2010; Melnychuk et al., 2024). Latent or rank restrictions can identify additional functionals only under extra assumptions (Yin et al., 2018; Shin, 2024). Robust and quantile regression, including robust HTE methods, already limit outcome sensitivity (Huber, 1964; Koenker & Bassett, 1978; Li

et al., 2023; Uehara, 2026). Cauchy AB-HTE's narrower contribution is the combination of shared latent abduction, treatment-indexed affine mechanisms, analytic effect propagation, and bounded residual scores. Extended comparisons appear in Appendix D.

3 Problem Setup and Estimands

We observe i.i.d. data $\mathcal{D}_n = \{(X_i, T_i, Y_i)\}_{i=1}^n$, where X contains only pretreatment covariates, $T \in \{0, 1\}$, and $Y = Y(T)$. The subscript i is only a record index; it does not define an indexed family of unit variables or response functions. At the model level U is the unit selection variable, and one realization $U = u^\star$ denotes the actual token and its embedding for the record.

Assumption 1 (Identification of potential-outcome marginals)**.** *We assume consistency, no interference, conditional exchangeability* $(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$*, and positivity* $0 < e(x) := \mathbb{P}(T = 1 \mid X = x) < 1$ *on the target population.*

Assumption 1 is not supplied by abduction. If an unobserved variable affects treatment and outcome after conditioning on X, exchangeability may fail; without additional assumptions, factual regression then cannot guarantee identification of a causal effect.

Let $F_t(y \mid x) = \mathbb{P}\{Y(t) \le y \mid X = x\}$ and $Q_t(p \mid x) = \inf\{y : F_t(y \mid x) \ge p\}$ for $p \in (0, 1)$. We study the identified marginal quantile contrast

$$\tau_p(x) = Q_1(p \mid x) - Q_0(p \mid x), \qquad \tau_{\mathrm{loc}}(x) = \tau_{1/2}(x). \tag{3}$$

This is not generally the pth quantile of $\Delta = Y(1) - Y(0)$: the former uses two marginals, whereas the latter requires their joint cross-world law. A continuous counterexample and the full statement are given in Appendix B.5.

The conventional CATE $\mathbb{E}\{Y(1) - Y(0) \mid X = x\}$ need not exist in the primary Cauchy model. All effects are therefore described as marginal location/quantile contrasts unless a finite-mean model or a declared coupling justifies another object. Pointwise positivity is sufficient for identification; stable weighting requires strong overlap $\epsilon \le e(x) \le 1 - \epsilon$. Implemented overlap diagnostics appear in Appendix C; no overlap-restricted target is reported.

4 Abduction-Based HTE Estimation

AB-HTE composes latent abduction, treatment-indexed action, and analytic reduction. We call the factorized-Cauchy instantiation below *Cauchy AB-HTE*; the architecture-matched Gaussian ablation is *Gaussian AB-HTE*. In the general unit-abduction grammar, evidence may yield (i) a point estimate $\widehat{u}_\phi(x)$, (ii) a Gaussian distribution expressing local uncertainty, or (iii) a Cauchy distribution expressing heavy-tailed uncertainty over candidate embeddings for the same actual token. This paper implements the latter two and uses Cauchy as its primary model; the point case remains a deterministic limiting alternative rather than a reported estimator.

An encoder outputs $\mu_\phi(x) \in \mathbb{R}^d$ and $\gamma_\phi(x) \in \mathbb{R}_+^d$ and defines

$$q_\phi(u \mid X = x) = \prod_{j=1}^{d} \mathrm{Cauchy}(u_j; \mu_{\phi,j}(x), \gamma_{\phi,j}(x)), \tag{4}$$

with conditionally independent coordinates and scale floor $\gamma_{\phi,j}(x) \ge \gamma_{\min} > 0$. This is an amortized conditional abduction result: it represents learner-side uncertainty about the fixed actual realization $u^\star$ selected by U. It is not a Bayes-exact posterior, a population distribution that resamples identity, or an identified physical coordinate system. We retain the symbol μ_ϕ only because it is the legacy implementation field name; mathematically it is a Cauchy location (and coordinatewise median), not a mean, while $\gamma_\phi(x)$ is a scale, not a standard deviation; non-degenerate Cauchy coordinates have neither finite means nor finite variances. The heavy tails keep remote candidate unit embeddings from being exponentially suppressed by the family alone. They do not imply that every outcome is attainable: outcome support remains determined by f_θ and the event-noise support. Write $\theta = (a_0, a, b_0, b, \sigma_0, \sigma_1)$ and $\vartheta = (\phi, \theta)$ for the full fitted parameter.

For unit-Cauchy event noise $E_t \perp\!\!\!\perp U \mid X$ and arm scale $\sigma_t > 0$,

$$Y(t) = c_t + w_t^\top U + \sigma_t E_t, \tag{5}$$

$$c_t = a_0 + tb_0, \qquad w_t = a + tb, \tag{6}$$

or equivalently

$$Y(t) = a_0 + a^\top U + t(b_0 + b^\top U) + \sigma_t E_t. \tag{7}$$

At world level, fixing $U = u$ gives the token-level generator $Y_u(t) = f_\theta(t, E; u)$. At prediction time, drawing a candidate $\widetilde{U} \sim q_\phi(\cdot \mid x)$ and event noise, then evaluating the same shared generator, is a computational representation of epistemic and event uncertainty; it is not a claim that the actual token changes. The $tb^\top U$ interaction is required for heterogeneous location effects. Fixing E_t to unit scale removes a trivial event-scale gauge; latent and event dispersion remain structurally, rather than observationally, decomposed.

Affine Cauchy closure gives the predictive outcome distribution

$$m_t(x) = a_0 + tb_0 + (a + tb)^\top \mu_\phi(x), \tag{8}$$

$$s_t(x) = \sum_{j=1}^{d} |a_j + tb_j| \gamma_{\phi,j}(x) + \sigma_t, \tag{9}$$

$$Y(t) \mid X = x \sim \mathrm{Cauchy}\big(m_t(x), s_t(x)\big). \tag{10}$$

Consequently,

$$\tau_p(x) = m_1(x) - m_0(x) + \{s_1(x) - s_0(x)\} \tan\big\{\pi(p - \tfrac{1}{2})\big\}, \tag{11}$$

$$\tau_{\mathrm{loc}}(x) = m_1(x) - m_0(x) = b_0 + b^\top \mu_\phi(x). \tag{12}$$

Here and throughout, marginal quantile levels satisfy $p \in (0, 1)$.

For a factual observation (x_i, t_i, y_i), the reduced negative log-likelihood is

$$\ell_i = \log\{\pi s_{t_i}(x_i)\} + \log\left[1 + \left\{\frac{y_i - m_{t_i}(x_i)}{s_{t_i}(x_i)}\right\}^2\right], \tag{13}$$

and the training objective is

$$\mathcal{L}_{\mathrm{fact}} = \frac{\sum_i \omega_i \ell_i}{\sum_i \omega_i}. \tag{14}$$

With $\omega_i = 1$, this is the ordinary conditional maximum-likelihood objective; other fixed nonnegative weights define a weighted empirical risk, not the ordinary likelihood. The experiments use $\omega_i = 1$, AdamW decay 10^{-5} on all parameters, and gradient clipping at norm 10. The Fisher-consistency statement in Section 5 concerns the correctly specified, unregularized population risk, not that finite-sample optimizer. Closed-form marginalization, gradients, regularization qualifications, and nondifferentiable scale points are detailed in Appendix B.2.

The identified marginal contrast does not require a joint potential-outcome law. For model-implied within-unit distributions, Cauchy AB-HTE reuses one draw from the factual abduction result for the same token and compares two event-noise couplings:

Assumption 2 (Shared-unit independent-resampling coupling)**.** *Conditional on X, $U_1, \ldots, U_d, E_0, E_1$ are mutually independent, $E_0, E_1 \sim \mathrm{Cauchy}(0, 1)$, and the same latent unit U is used in both worlds.*

Thus both treatment queries use the same $q_\phi(\cdot \mid x)$ obtained from factual pretreatment evidence. The alternative treatment is never fed back into the encoder to abduct a new unit. Under this assumption,

$$\Delta_{\mathrm{ind}} = b_0 + b^\top U + \sigma_1 E_1 - \sigma_0 E_0, \tag{15}$$

$$\Delta_{\mathrm{ind}} \mid x \sim \mathrm{Cauchy}\big(\tau_{\mathrm{loc}}(x), \kappa_{\mathrm{ind}}(x)\big), \tag{16}$$

$$\kappa_{\mathrm{ind}}(x) = \sum_j |b_j| \gamma_{\phi,j}(x) + \sigma_1 + \sigma_0. \tag{17}$$

A shared-exogenous realization instead takes $E \sim \mathrm{Cauchy}(0,1)$ with $E \perp\!\!\!\perp U \mid X$ and gives, when $\kappa_{\mathrm{shared}}(x) > 0$,

$$\Delta_{\mathrm{shared}} \mid x \sim \mathrm{Cauchy}\left(\tau_{\mathrm{loc}}(x), \underbrace{\sum_j |b_j|\gamma_{\phi,j}(x) + |\sigma_1 - \sigma_0|}_{\kappa_{\mathrm{shared}}(x)}\right). \tag{18}$$

If $\kappa_{\mathrm{shared}}(x) = 0$, the contrast is the point mass $\Delta_{\mathrm{shared}} = \tau_{\mathrm{loc}}(x)$ almost surely. For either coupling c with $\kappa_c(x) > 0$,

$$\mathbb{P}_\vartheta(\Delta_c > 0 \mid x) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{\tau_{\mathrm{loc}}(x)}{\kappa_c(x)}\right). \tag{19}$$

For the degenerate case, this probability is $\mathbb{I}\{\tau_{\mathrm{loc}}(x) > 0\}$. These laws are sensitivity choices, not factual identification formulas.

5 Identification and Properties

The following chain states what follows from causal assumptions, what the factual objective recovers under model specification, and what remains a cross-world commitment. Full statements are collected in Appendix A; proofs are in Appendix B.

Proposition 1 (Marginal identification)**.** *Under Assumption 1, $F_t(y \mid x) = \mathbb{P}(Y \le y \mid T = t, X = x)$; hence each $\tau_p(x)$ is identified P_X-almost surely wherever positivity holds.*

Proposition 2 (Structural sufficient condition)**.** *If $T \perp\!\!\!\perp (U, E_0, E_1) \mid X$ in Equation* (7)*, then $(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$. Dependence on residual latent information instead leaves the factual likelihood causal only under an additional assumption.*

Proposition 3 (Marginal contrast is not an effect quantile)**.** *In general, $Q_1(p \mid x) - Q_0(p \mid x) \neq Q_{Y(1)-Y(0)}(p \mid x)$; the right-hand side depends on the cross-world coupling.*

Proposition 4 (Affine Cauchy closure)**.** *For independent $Z_j \sim \mathrm{Cauchy}(\mu_j, \gamma_j)$,*

$$r + \sum_j v_j Z_j \sim \mathrm{Cauchy}(r + \sum_j v_j \mu_j, \sum_j |v_j|\gamma_j). \tag{20}$$

This statement assumes the displayed scale is positive; if every stochastic coefficient is zero, the sum is a point mass at its displayed location.

Proposition 5 (Analytic response kernels)**.** *Under the coordinate independence in Equation* (4) *and $E_t \perp\!\!\!\perp U \mid X$, Equation* (7) *yields the Cauchy kernel in Equations* (8)*–*(10) *and all marginal quantile contrasts in Equation* (11)*.*

Corollary 1 (Closed-form location contrast)**.** $\tau_{\mathrm{loc}}(x) = b_0 + b^\top \mu_\phi(x)$.

Theorem 1 (Fisher consistency of factual Cauchy risk)**.** *Under Assumption 1, suppose one shared Cauchy AB-HTE parameter $\vartheta^\star = (\phi^\star, \theta^\star)$ realizes both true conditional Cauchy kernels and has finite weighted cross-entropy. For unpenalized population log-risk over finite-risk candidates and any measurable weight $0 < c \le \omega(X,T) \le C < \infty$,*

$$R_\omega(\vartheta) - R_\omega(\vartheta^\star) = \mathbb{E}_{X,T}[\omega(X,T) D_{\mathrm{KL}}(p^\star(\cdot \mid X,T) \| p_\vartheta(\cdot \mid X,T))] \ge 0. \tag{21}$$

Every finite-risk global minimizer therefore recovers $m_t^\star, s_t^\star$ and $\tau_{\mathrm{loc}}^\star$ almost surely. This does not identify a unique latent factorization; fixed nonzero penalties and finite-sample consistency require separate analysis.

Proposition 6 (Latent gauge non-identification)**.** *For nonsingular diagonal D and $h \in \mathbb{R}^d$ whose transformed encoder remains in the declared model class, including its scale floor, the transformation*

$$U' = DU + h, \quad \mu'_\phi = D\mu_\phi + h, \quad \gamma'_\phi = |D|\gamma_\phi, \quad w'_t = D^{-\top} w_t, \tag{22}$$

with the corresponding intercept shift, leaves both response kernels unchanged. Thus latent coordinates and mechanism coefficients are not identified. A separate constructive ambiguity for the latent/event scale allocation is given in Appendix A.

Table 1: Claim boundary.

Quantity	Status
F_t, τ_p	Identified under Assumption 1
$m_t, s_t, \tau_{\mathrm{loc}}$	Population-risk recovery under correct unpenalized specification
F_Δ without coupling	Partially identified by Makarov bounds
$p_\vartheta(\Delta_c \mid x)$, latent coordinates	Model-implied; not identified by factual marginals

Table 2: Evidence is reported only at the layer where its target is observed or fixed by the generator. Full data-generating processes and protocols are in Appendix C.

Layer	Available truth	Valid primary evidence
Fully synthetic	Both response surfaces and a declared coupling	Location-effect error, marginal calibration, coupling sensitivity
Semi-synthetic	Known response surfaces, coupling only if specified	Location-effect and marginal-distribution error
Real RCT/observational	Factual outcomes; assignment design or assumptions	Factual score/calibration, design-based contrasts, cross-fitted policy value

Proposition 7 (Coupling-specific effect laws)**.** *The declared independent-resampling and shared-exogenous couplings yield the Cauchy laws in Equations* (16) *and* (18) *whenever the corresponding scale is positive; a zero shared scale yields a point mass. Their common median is $\tau_{\mathrm{loc}}(x)$, while their scales and benefit probabilities differ.*

Proposition 8 (No standard CATE under a non-degenerate effect law)**.** *If either coupling-specific Cauchy effect has positive scale, $\mathbb{E}(\Delta_c \mid X = x)$ does not exist.*

Proposition 9 (Cross-world non-identification)**.** *Two shared-unit affine Cauchy models can induce identical marginals for both arms but different distributions of $Y(1) - Y(0)$. Factual marginals therefore do not identify either coupling-specific effect law.*

Theorem 2 (Makarov bounds)**.** *For continuous identified marginals and each fixed δ, every compatible effect CDF satisfies pointwise sharp Makarov bounds over unrestricted couplings; in particular*

$$1 - \overline{F}_\Delta(0 \mid x) \le \mathbb{P}(\Delta > 0 \mid x) \le 1 - \underline{F}_\Delta(0 \mid x), \tag{23}$$

where the full sup/inf definitions are given in Appendix A (Fan & Park, 2010).

Proposition 10 (Bounded residual scores)**.** *For the Cauchy NLL and $s \ge s_{\min} > 0$,*

$$\left|\frac{\partial \ell}{\partial m}\right| \le \frac{1}{s}, \qquad \left|\frac{\partial \ell}{\partial \log s}\right| \le 1. \tag{24}$$

Corollary 2 (Parameter-gradient condition)**.** *If $\|\nabla_\vartheta m\| \le G_m$ and $\|\nabla_\vartheta \log s\| \le G_s$, then*

$$\|\nabla_\vartheta \ell\| \le G_m / s_{\min} + G_s. \tag{25}$$

This is not an unconditional bounded-influence theorem.

6 Empirical Evaluation

The evaluation asks whether Cauchy AB-HTE improves location-effect estimation and marginal calibration under clean and contaminated outcomes, whether any gain is more than a Cauchy likelihood effect, and how coupling-dependent summaries vary under sensitivity analysis. We separate three evidence layers because they expose different ground truth.

Design. The realizable Cauchy–affine generator is an engineering check; the primary benchmark is instead a nonlinear Gaussian-noise HTE SCM outside the Cauchy AB-HTE model family. For operator C_ρ and fitting-and-selection algorithm $\mathcal{A}$, the primary protocol is

$$\widehat{\tau}_\rho = \mathcal{A}\big(C_\rho(D_{\mathrm{tr}}; \xi_{\mathrm{tr}}), \; C_\rho(D_{\mathrm{val}}; \xi_{\mathrm{val}})\big). \tag{26}$$

Table 3: Clean-test location-effect MedAE when training and validation receive independent corruption from the same operator at $\rho = 0.4$ (mean $\pm$ SE over ten frozen seeds; lower is better).

Method	Clean	Global	Within-arm	Outlier	Systematic
Cauchy AB-HTE (ours)	0.071 ± 0.003	0.103 ± 0.010	0.086 ± 0.005	0.086 ± 0.005	0.125 ± 0.009
Gaussian AB-HTE	0.065 ± 0.003	0.186 ± 0.010	0.178 ± 0.008	0.206 ± 0.005	0.115 ± 0.011
Cauchy S	0.065 ± 0.003	0.099 ± 0.010	0.077 ± 0.005	0.084 ± 0.005	0.119 ± 0.009
Huber S	0.057 ± 0.003	0.131 ± 0.013	0.102 ± 0.006	0.128 ± 0.007	0.089 ± 0.005
Pinball S	0.059 ± 0.004	0.107 ± 0.012	0.082 ± 0.005	0.092 ± 0.004	0.115 ± 0.004
Cauchy T	0.099 ± 0.003	0.125 ± 0.007	0.114 ± 0.006	0.119 ± 0.007	0.176 ± 0.011

Here $\xi_{\mathrm{tr}} \perp\!\!\!\perp \xi_{\mathrm{val}}$, and $\widehat{\tau}_\rho$ is scored on D^0_{te}. Thus early stopping and checkpoint selection see independently corrupted validation records from the same operator and rate; only the final causal evaluation population is clean. The weaker clean-validation oracle is an appendix ablation. Rates are $\rho \in \{0, .1, .2, .3, .4, .5\}$; global and within-arm shuffles, gross outliers, asymmetric and systematic shifts are primary, while treatment-record flips and arm-differential shifts are reported separately. Splits, exact DGP and operator equations, validation semantics, and eligibility rules for semi-synthetic and real data are fixed in Appendix C; counterfactual oracle values are never used for model selection. The real layer uses the official 1986-W40–1987-W18 WAWS randomized-trial window with documented propensity $3/8$ (Lachowska et al., 2015).

Baselines and decisive ablations. Implemented comparisons include matched Gaussian AB-HTE, Huber and pinball S-learners, direct Cauchy S- and T-learners, and, on IHDP, TARNet and a CFR-style linear-MMD representation learner. Broader forest, R-learner, Dragonnet, and conditional-distributional comparisons remain future work. Three contrasts are decisive: Cauchy AB-HTE versus Gaussian AB-HTE isolates the likelihood family; Cauchy AB-HTE versus a similar-capacity direct Cauchy S-learner tests whether shared abduction adds beyond robust outcome regression; and matched-corrupted versus clean validation separates pipeline robustness from a trusted-selection-oracle setting. All comparisons share partition-level corruption draws, selection rules, and training budgets. Cauchy AB-HTE and direct Cauchy S have 1,556 and 1,378 trainable parameters, respectively; they are not called exactly capacity matched.

Metrics and reporting. On oracle test populations the primary metric is median absolute location-effect error,

$$\mathrm{MedAE}_\tau = \mathrm{median}_{i \in \mathcal{I}_{\mathrm{te}}} \left| \widehat{\tau}_{\mathrm{loc}}(x_i) - \tau^\star_{\mathrm{loc}}(x_i) \right|. \tag{27}$$

We additionally report clean-to-corrupted degradation, marginal log score and calibration, sign/rank accuracy, and location-policy regret. Effect-law quantiles and probability of benefit are scored only when the generator fixes the coupling. On real trials, aggregate ITT and fixed marginal contrasts receive design-based uncertainty. Learned score strata and policies use cross-fitted scores; their point estimates are reported, while stored first-order standard errors condition on those scores and are not formal training-algorithm inference. Synthetic results use at least ten data/corruption replications with paired interval estimates; the single real trial uses five-fold out-of-fold estimates with design-based aggregate contrasts and conditional learned-score diagnostics.

Section 7 reports the checked evidence for the primary degradation curves, matched ablations, semi-synthetic transfer, coupling sensitivity, and factual-only real-data checks. Global, within-arm, and treatment-record shuffles remain separate. Exact DGPs, corruption operators, leakage controls, overlap diagnostics, all metric formulas (including the trial policy estimator), replication rules, and implementation status appear in Appendix C.

7 Results

All numbers below are generated from machine-readable artifacts after a completeness check. Training and early stopping use factual records only. Synthetic oracle effects are exposed only after model selection, and IHDP hyperparameters are selected on replications 0–9 without loading test response surfaces before evaluation on held-out replications 100–199.

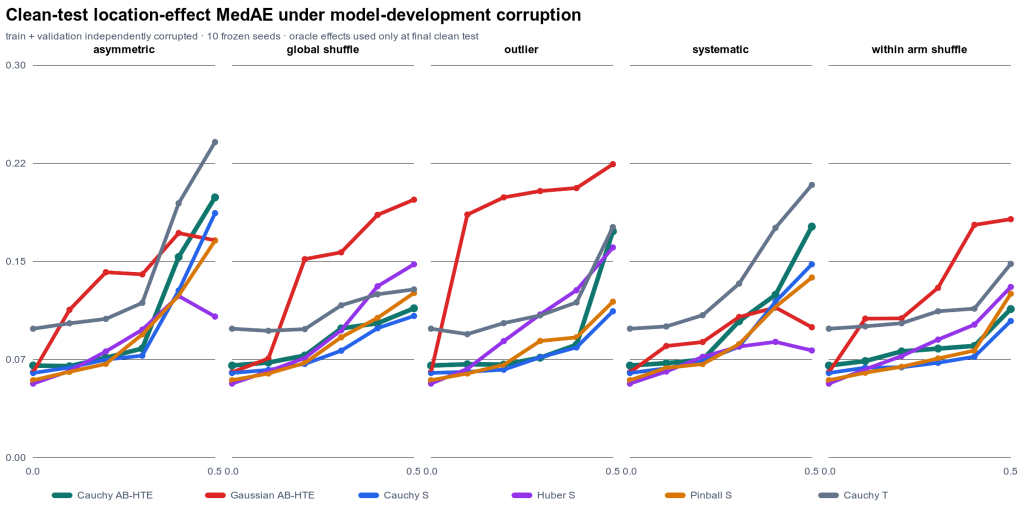

Figure 1: Complete ten-seed robustness curves on the nonlinear randomized SCM. Training and validation receive independently sampled corruption from the same operator and rate; the oracle test population remains clean. Curves are diagnostic rather than a claim of one universally robust method.

Table 4: Identification boundary (effect MedAE, mean ± SE). Hidden confounding is a negative control, not a leaderboard condition.

Method	Randomized	Observed confounding	Hidden confounding
Cauchy AB-HTE (ours)	0.071 ± 0.003	0.072 ± 0.005	1.445 ± 0.053
Gaussian AB-HTE	0.065 ± 0.003	0.067 ± 0.004	1.528 ± 0.025
Cauchy S	0.065 ± 0.003	0.070 ± 0.004	1.564 ± 0.031
Huber S	0.057 ± 0.003	0.057 ± 0.002	1.570 ± 0.024
Pinball S	0.059 ± 0.004	0.064 ± 0.003	1.576 ± 0.028
Cauchy T	0.099 ± 0.003	0.112 ± 0.006	1.530 ± 0.027

Robustness is operator specific. On clean finite-noise data, Huber S is best among the implemented methods (.057 versus Cauchy AB-HTE's .071). At 40% gross-outlier replacement, Cauchy AB-HTE reaches .086 and increases .016 from its own clean error; its error is numerically 58% below matched Gaussian AB-HTE (.206), while direct Cauchy S is slightly better (.084). This supports a heavy-tail advantage over the Gaussian structural twin, but not an unconditional advantage for latent abduction. Direct Cauchy S is also best under global and within-arm shuffle, and Huber S is best under asymmetric and systematic shifts. Figure 1 shows that validation contamination changes checkpoint selection but does not create one universally robust method; the matched-versus-clean validation comparison is isolated in Appendix Table 9. Across the five operators, Cauchy AB-HTE's macro MedAE is .111 under pipeline corruption versus .105 with trusted validation (paired gap $.006 \pm .002$); the sign is not uniform across every estimator/operator, so the gap is diagnostic rather than additive.

Estimation does not repair identification. Moving from random assignment to observed nonlinear confounding changes Cauchy AB-HTE's effect MedAE only from .071 to .072 under the declared overlap. Adding an unobserved common cause and sorting on gains raises it to 1.445, more than twenty times the randomized error. In that negative control, sign accuracy collapses to .613, the prevalence of positive oracle effects and hence the performance of an almost treat-all rule. The failure is expected: end-to-end optimization cannot recover an estimand absent exchangeability. Propensity, ESS, and weighted-balance diagnostics are reported in Appendix C.

Table 5: Declared-coupling sensitivity on the realizable Cauchy SCM (ten seeds). The factual fit is identical across rows.

Coupling	Pred. scale	Declared scale	Scale MAE	Benefit-prob. MAE
Independent resampling	0.372	0.353	0.030	0.025
Shared realization	0.067	0.153	0.086	0.054

Table 6: IHDP response-surface recovery on 100 held-out benchmark replications. Hyperparameters were frozen on replications 0–9 using factual validation only.

Method	Effect MedAE	Spearman ↑	Policy regret ↓
Cauchy AB-HTE (ours)	0.787 ± 0.022	0.820	0.022
Gaussian AB-HTE	0.912 ± 0.036	0.721	0.028
Cauchy S	3.036 ± 0.187	0.112	0.324
Huber S	1.058 ± 0.044	0.691	0.055
Pinball S	1.099 ± 0.038	0.653	0.054
Cauchy T	0.893 ± 0.024	0.793	0.040
TARNet	0.699 ± 0.021	0.852	0.024
CFR-style MMD	0.700 ± 0.022	0.855	0.023

Causal record corruption exposes different weaknesses. Treatment-label flips and arm-differential shifts are not pooled with outcome shuffle. At $\rho = .4$, Cauchy AB-HTE obtains MedAE .198 under treatment flips but .646 under opposite arm shifts. Direct Cauchy S is better for flips (.179), whereas Cauchy AB-HTE is best for the manufactured contrast; nevertheless, its large absolute error shows that a bounded residual score does not remove systematic arm bias. Full results appear in Table 10.

The factual likelihood cannot choose a coupling. The same fitted model and factual test NLL produce the two rows of Table 5. Their effect locations are identical, but the mean predicted scale changes from .372 to .067, and the predicted probability of benefit changes accordingly. The shared-realization scale is also less accurately recovered, consistent with the separately constructed latent/event allocation ambiguity: factual marginals do not identify how uncertainty should cancel across worlds.

Semi-synthetic transfer is competitive, not state of the art. IHDP supplies simulated response surfaces, so Table 6 scores $\mu_1(x) - \mu_0(x)$, not a real person's effect. Cauchy AB-HTE improves over its Gaussian twin (.787 versus .912 MedAE), direct Cauchy S (3.036), and robust point S-learners, but TARNet/CFR-style MMD remain better ($.699/.700$): shared affine abduction does not dominate a flexible balanced representation.

A real RCT permits causal evaluation without ITE labels. WAWS lacks individual counterfactual truth. On its official window ($n = 4{,}218$), mean ITT is 118 ± 366 dollars; Cauchy AB-HTE has the lowest out-of-fold factual MedAE ($5{,}918$), while Gaussian AB-HTE has the best NLL/PIT. Policy point values span $.472$–$.494$, with differences from assigning ER to all units ranging from $-.010$ to $+.013$. Cauchy AB-HTE's top-minus-bottom score-stratum contrast is 374 dollars with a first-order SE of $1{,}235$. Because these diagnostics condition on learned cross-fit scores, they are not formal learned-policy inference. They are factual/randomization-based evidence, not ITE accuracy (Table 11).

8 Discussion and Conclusion

Cauchy AB-HTE trades distributional flexibility for exact Cauchy propagation. Its affine symmetric kernels may fail for multimodal, bounded, or asymmetric outcomes; experiments must decide when analytic reduction merits that restriction.

Its causal interpretation is conditional, not automatic. Consistency, exchangeability, positivity, and correct marginal specification are required; $q_\phi(u \mid X)$ neither observes hidden confounders nor

identifies latent coordinates. Fisher consistency concerns the unpenalized population log-risk, not finite-sample optimization, neural generalization, or a fixed nonzero regularizer. Bounded residual scores likewise do not establish unconditional bounded influence.

The unit ontology is equally conditional but precise: the world contains one actual realization $U = u^\star$, whereas $q_\phi(u \mid x)$ records the learner's uncertainty about that same token after seeing factual pretreatment evidence. Counterfactual treatment prediction keeps this abduction result fixed and changes the treatment query. Re-encoding an alternative treatment would silently change the unit and is outside the model.

Most importantly, observational data identify potential-outcome marginals under the causal assumptions, not an individual's missing counterfactual or the cross-world copula. Independent-resampling and shared-exogenous effect laws are declared sensitivity models. Policy claims using their dispersion or benefit probability must be accompanied by coupling sensitivity and assumption-lean Makarov bounds.

Cauchy AB-HTE gives an auditable route from latent abduction to a conditional location contrast while separating identified marginals from model-implied effect laws. Its gross-outlier advantage over Gaussian AB-HTE is narrow: robust learners win other corruptions and TARNet/CFR-style baselines lead on IHDP. WAWS's null aggregate randomized contrast and unstable policy-point ordering show that factual fit is not ITE validation; broader replication is needed before stronger practical claims.

REFERENCES

Alicia Curth and Mihaela van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1810–1818. PMLR, 2021. URL https://proceedings.mlr.press/v130/curth21a.html.

Yanqin Fan and Sang Soo Park. Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory*, 26(3):931–951, 2010. doi: 10.1017/S0266466609990168.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35 (1):73–101, 1964. doi: 10.1214/aoms/1177703732.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

Nathan Kallus and Miruna Oprescu. Robust and agnostic learning of conditional distributional treatment effects. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 6037–6060. PMLR, 2023. URL https://proceedings.mlr.press/v206/kallus23a.html.

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. doi: 10.2307/1913643.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019. doi: 10.1073/pnas.1804597116.

Marta Lachowska, Merve Meral, and Stephen A. Woodbury. The effects of eliminating the work search requirement on job match quality and other long-term employment outcomes. Technical Report 2015-14, U.S. Department of Labor, 2015. URL https://www.dol.gov/sites/dolgov/files/OASP/legacy/files/2013-2014_DOL_Scholars_Paper_Series_Lachowska_Chapter.pdf.

Ruohong Li, Honglang Wang, Yi Zhao, Jing Su, and Wanzhu Tu. Robust estimation of heterogeneous treatment effects: An algorithm-based approach. *Communications in Statistics—Simulation and Computation*, 52(10):4981–4998, 2023. doi: 10.1080/03610918.2021.1974883.

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.

Yuchen Ma, Valentyn Melnychuk, Jonas Schweisthal, and Stefan Feuerriegel. DiffPO: A causal diffusion model for learning distributions of potential outcomes. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1384. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4d254c96c5ff500dda2ac0a58987aeba-Abstract-Conference.html.

Valentyn Melnychuk, Stefan Feuerriegel, and Mihaela van der Schaar. Quantifying aleatoric uncertainty of the treatment effect: A novel orthogonal learner. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3336. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bdabb5d4262bcfb6a1d529d690a6c82b-Abstract-Conference.html.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021. doi: 10.1093/biomet/asaa076.

Junhyung Park, Uri Shalit, Bernhard Schölkopf, and Krikamol Muandet. Conditional distributional treatment effect with kernel conditional mean embeddings and U-statistic regression. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8401–8412. PMLR, 2021. URL https://proceedings.mlr.press/v139/park21c.html.

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

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. doi: 10.1093/biomet/70.1.41.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350.

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*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3076–3085. PMLR, 2017. URL https://proceedings.mlr.press/v70/shalit17a.html.

Claudia Shi, David M. Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8fb5f8be2aa9d6c64a04e3ab9f63feee-Abstract.html.

Myungkou Shin. Distributional treatment effect with latent rank invariance. *arXiv preprint arXiv:2403.18503*, 2024. URL https://arxiv.org/abs/2403.18503.

Eichi Uehara. Bayesian X-learner: Calibrated posterior inference for heterogeneous treatment effects under heavy-tailed outcomes. *arXiv preprint arXiv:2604.27394*, 2026. URL https://arxiv.org/abs/2604.27394. Preprint.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839.

Yanyan Yin, Lan Liu, and Zhi Geng. Assessing the treatment effect heterogeneity with a latent variable. *Statistica Sinica*, 28:115–135, 2018. doi: 10.5705/ss.202016.0150.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ByKWUeWA-.

Fan Zhou, Ramsey D. Carson, and David Carlson. Estimating potential outcome distributions with collaborating causal networks. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=q1Fey9feu7.

A Full Theoretical Statements

This section separates algebraic closure, causal identification, statistical recovery of observable response kernels, and cross-world modeling commitments. Proofs are given in Appendix B.

A.1 Identified marginal targets

Proposition 11 (Identification of potential-outcome marginals)**.** *Under Assumption 1, for* $t \in \{0, 1\}$,

$$F_t(y \mid x) = \mathbb{P}(Y \le y \mid T = t, X = x). \tag{28}$$

Consequently, conditional quantile contrasts $\tau_p(x)$*, including* $\tau_{\mathrm{loc}}(x)$*, are identified* P_X*-almost surely wherever positivity holds.*

The proposition concerns observable response kernels. It does not identify a particular latent coordinate system or a joint law for $(Y(0), Y(1))$.

Proposition 12 (A structural sufficient condition for exchangeability)**.** *Suppose the potential outcomes are generated by Equation* (7) *and*

$$T \perp\!\!\!\perp (U, E_0, E_1) \mid X. \tag{29}$$

Then $(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$*. Conversely, if treatment depends on residual information in* (U, E_0, E_1) *after conditioning on* X*, the factual likelihood alone generally identifies* $p(Y \mid X, T)$ *rather than* $p(Y(t) \mid X)$*.*

Pointwise positivity identifies each conditional arm where it holds. Stable inverse-propensity estimation requires the stronger overlap condition $\epsilon \le e(x) \le 1 - \epsilon$ on the target population; neither condition is created by the abductive encoder.

A.2 Analytic reduction and factual-risk recovery

Proposition 13 (Affine closure of independent Cauchy variables)**.** *If* $Z_j \sim \mathrm{Cauchy}(\mu_j, \gamma_j)$ *are mutually independent, then for constants* r *and* v_j *with* $\sum_j |v_j|\gamma_j > 0$*,*

$$r + \sum_{j=1}^{k} v_j Z_j \sim \mathrm{Cauchy}\left(r + \sum_{j=1}^{k} v_j \mu_j, \sum_{j=1}^{k} |v_j|\gamma_j \right). \tag{30}$$

If all stochastic coefficients vanish, the same characteristic-function calculation yields a point mass at $r + \sum_j v_j \mu_j$, not a positive-scale Cauchy law.

Proposition 14 (Analytic potential-outcome distributions)**.** *Suppose the coordinates in Equation* (4) *and the event noise are conditionally independent. Under Equation* (7)*,*

$$Y(t) \mid X = x \sim \mathrm{Cauchy}\big(m_t(x), s_t(x)\big), \tag{31}$$

with m_t *and* s_t *given by Equations* (8) *and* (9)*. Therefore the full family of conditional marginal quantile contrasts is given by Equation* (11)*.*

Corollary 3 (Closed-form causal location effect)**.** *Under Proposition 5,*

$$\tau_{\mathrm{loc}}(x) = b_0 + b^\top \mu_\phi(x). \tag{32}$$

Thus heterogeneity is carried by the interaction coefficient b *and the abducted unit location* $\mu_\phi(x)$*.*

The preceding statements are model algebra. The next result connects the unpenalized factual objective to the observable causal target.

Theorem 3 (Fisher consistency of the factual Cauchy risk)**.** *Assume Assumption 1. Suppose there is one shared* $\vartheta^\star = (\phi^\star, \theta^\star)$ *in the observable Cauchy AB-HTE family that simultaneously realizes both true conditional response kernels:*

$$p^\star(y \mid x, t) = \mathit{cauchy}\big(y; m_t^\star(x), s_t^\star(x)\big), \qquad s_t^\star(x) > 0. \tag{33}$$

Assume $R_\omega(\vartheta^\star) < \infty$*. Let* $p_\vartheta(y \mid x, t)$ *denote a candidate Cauchy AB-HTE response kernel and let* $\omega(X, T)$ *be either one or a measurable weight satisfying* $0 < c \le \omega(X, T) \le C < \infty$*. For the unpenalized population risk*

$$R_\omega(\vartheta) = \mathbb{E}[-\omega(X, T) \log p_\vartheta(Y \mid X, T)], \tag{34}$$

then every finite-risk global population minimizer induces $p_\vartheta(\cdot \mid x, t) = p^\star(\cdot \mid x, t)$ *almost surely for both arms. Hence* $m_t = m_t^\star$*,* $s_t = s_t^\star$*, and* $m_1 - m_0 = \tau_{\mathrm{loc}}^\star$ *almost surely. This conclusion concerns the observable kernels, not a unique latent decomposition.*

Theorem 1 does not apply unchanged to the experiment's fixed AdamW decay. An asymptotic consistency claim for a regularized estimator requires vanishing regularization, approximation and uniform-convergence conditions, or a separate analysis of the optimizer's population target.

A.3 THE LATENT FACTORIZATION IS NOT IDENTIFIED

Proposition 15 (Diagonal-affine latent reparameterization)**.** *Let D be any nonsingular diagonal matrix and $h \in \mathbb{R}^d$ for which the transformed encoder below remains representable in the declared model class and satisfies its scale floor. Define*

$$U' = DU + h, \qquad \mu'_\phi(x) = D\mu_\phi(x) + h, \qquad \gamma'_\phi(x) = |D|\gamma_\phi(x), \tag{35}$$

$$w'_t = D^{-\top} w_t, \qquad c'_t = c_t - {w'_t}^{\top} h, \qquad \sigma'_t = \sigma_t, \tag{36}$$

where $|D|$ denotes the diagonal matrix of absolute diagonal entries. The primed and unprimed parameterizations induce identical $m_t(x)$, $s_t(x)$, and factual likelihoods for both arms. Thus latent coordinates, mechanism coefficients, and their coordinate-wise scaling are not identified without anchors or additional measurements.

There is a separate, constructive allocation ambiguity. Fix $S > 0$, choose two distinct constants $g_1, g_2 > \gamma_{\min}$ with $g_k < S$, and take $d = 1$, $w_0 = w_1 = 1$, $\mu_\phi(x) = 0$, $\gamma_\phi(x) = g_k$, and $\sigma_0 = \sigma_1 = S - g_k$. Both choices induce the same two response kernels $\mathrm{Cauchy}(0, S)$, but assign different amounts of dispersion to the latent unit and event noise. Hence factual marginals do not identify that allocation either.

A.4 CROSS-WORLD EFFECT LAWS AND THEIR LIMITS

Proposition 16 (Effect laws under two declared couplings)**.** *Under Assumption 2, the independent-resampling contrast obeys*

$$\Delta_{\mathrm{ind}} \mid X = x \sim \mathrm{Cauchy}\big(\tau_{\mathrm{loc}}(x), \kappa_{\mathrm{ind}}(x)\big), \tag{37}$$

where κ_{ind} is defined in Equation (17)*. Under the shared-exogenous coupling with $E \perp U \mid X$, if $\kappa_{\mathrm{shared}}(x) > 0$,*

$$\Delta_{\mathrm{shared}} \mid X = x \sim \mathrm{Cauchy}\big(\tau_{\mathrm{loc}}(x), \kappa_{\mathrm{shared}}(x)\big). \tag{38}$$

If $\kappa_{\mathrm{shared}}(x) = 0$, then $\Delta_{\mathrm{shared}} = \tau_{\mathrm{loc}}(x)$ almost surely, and its benefit probability is $\mathbb{I}\{\tau_{\mathrm{loc}}(x) > 0\}$. For either declared coupling $c \in \{\mathrm{ind}, \mathrm{shared}\}$ with $\kappa_c(x) > 0$,

$$\mathbb{P}_\vartheta(\Delta_c > 0 \mid X = x) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{\tau_{\mathrm{loc}}(x)}{\kappa_c(x)}\right). \tag{39}$$

These are model-implied laws under different couplings, not two factual identification formulas.

Proposition 17 (The standard CATE need not exist)**.** *If the selected effect law in Proposition 7 is non-degenerate, then $\mathbb{E}(\Delta_c \mid X = x)$ does not exist. Hence $\tau_{\mathrm{loc}}(x)$ is not a conditional average treatment effect.*

The distinction is mathematical rather than terminological: a Cauchy location is also its median, but not its expectation. Reporting a principal value would not recover the standard causal estimand.

Proposition 18 (Cross-world non-identification)**.** *The two potential-outcome marginals do not generally identify the law of $Y(1) - Y(0)$, even within the shared-unit affine Cauchy family. In particular, two Cauchy AB-HTE parameterizations can induce identical Cauchy marginals for both arms but different effect laws under either coupling in Proposition 7.*

The constructive proof in Appendix B.9 makes this ambiguity explicit. Consequently, quantities such as probability of benefit are structural, coupling-dependent outputs.

Theorem 4 (Makarov bounds from identified marginals)**.** *For continuous conditional marginals $F_0(\cdot \mid x)$ and $F_1(\cdot \mid x)$ and each fixed δ, define*

$$\underline{F}_\Delta(\delta \mid x) = \sup_{y \in \mathbb{R}} \big[F_1(y \mid x) - F_0(y - \delta \mid x)\big]_+, \tag{40}$$

$$\overline{F}_\Delta(\delta \mid x) = 1 + \inf_{y \in \mathbb{R}} \min\big\{F_1(y \mid x) - F_0(y - \delta \mid x), 0\big\}. \tag{41}$$

Every joint law with these marginals satisfies

$$\underline{F}_\Delta(\delta \mid x) \le F_\Delta(\delta \mid x) \le \overline{F}_\Delta(\delta \mid x), \tag{42}$$

and the bounds are pointwise sharp over unrestricted couplings (Fan & Park, 2010). Therefore

$$1 - \overline{F}_\Delta(0 \mid x) \le \mathbb{P}(\Delta > 0 \mid X = x) \le 1 - \underline{F}_\Delta(0 \mid x). \tag{43}$$

Any Cauchy AB-HTE coupling-specific CDF must lie inside these bounds; factual data alone do not select its position within them.

Table 7: Identification status of the main quantities. "Model-implied" means that factual data can fit compatible marginals but cannot validate the cross-world coupling or latent decomposition by themselves.

Quantity	Status
$F_t(y \mid x)$	Identified under Assumption 1
$m_t(x), s_t(x)$	Recovered by population log risk under correct Cauchy specification
$\tau_p(x)$ and $\tau_{\mathrm{loc}}(x)$	Identified marginal quantile contrasts
Joint law of $(Y(0), Y(1)) \mid x$	Not identified from factual marginals
F_Δ without a coupling	Partially identified by Theorem 2
$p_\vartheta(\Delta_c \mid x)$ and benefit probability	Model-implied under a declared coupling
Coordinates of U and latent/event allocation	Non-identified without anchors or extra measurements
Standard CATE	Undefined for a non-degenerate Cauchy effect law

A.5 Bounded residual scores, not unconditional bounded influence

Proposition 19 (Bounded Cauchy residual scores)**.** *For*

$$\ell(y; m, s) = \log(\pi s) + \log\left\{1 + \frac{(y-m)^2}{s^2}\right\}, \tag{44}$$

the location and log-scale derivatives satisfy

$$\frac{\partial \ell}{\partial m} = \frac{2(m-y)}{s^2 + (y-m)^2}, \qquad \left|\frac{\partial \ell}{\partial m}\right| \le \frac{1}{s}, \tag{45}$$

$$\frac{\partial \ell}{\partial \log s} = \frac{s^2 - (y-m)^2}{s^2 + (y-m)^2}, \qquad \left|\frac{\partial \ell}{\partial \log s}\right| \le 1. \tag{46}$$

If $s \ge s_{\min} > 0$, both residual scores are uniformly bounded in y.

Corollary 4 (A parameter-gradient bound under bounded Jacobians)**.** *Let $m = m_\vartheta(x,t)$ and $s = s_\vartheta(x,t)$. If $s \ge s_{\min} > 0$, $\|\nabla_\vartheta m\| \le G_m$, and $\|\nabla_\vartheta \log s\| \le G_s$, then*

$$\|\nabla_\vartheta \ell\| \le \frac{G_m}{s_{\min}} + G_s. \tag{47}$$

The location score redescends to zero as $|y - m| \to \infty$. This bounds the residual contribution conditional on the model Jacobians; it is not by itself a bounded-influence theorem for a neural estimator. Such a theorem additionally requires control of covariate leverage and a nonsingular estimating-equation Jacobian. The result also says nothing about confounding, overlap failure, treatment misclassification, or covariate shift.

A.6 What is and is not learned

Table 1 summarizes the claim boundary.

Accordingly, Cauchy AB-HTE's main empirical claims must be stated at the level of effect locations, marginal quantiles, calibration, and declared-coupling sensitivity— not exact recovery of U, a, b, or an assumption-free individual-effect distribution.

B Additional Derivations

B.1 Proof of affine Cauchy closure

Proof of Proposition 4. The characteristic function of $Z_j \sim \mathrm{Cauchy}(\mu_j, \gamma_j)$ is $\varphi_j(r) = \exp(\mathrm{i}\mu_j r - \gamma_j|r|)$. Independence gives

$$\begin{aligned}\varphi_{r_0+\sum_j v_j Z_j}(r) &= \exp(\mathrm{i}r_0 r)\prod_j \varphi_j(v_j r) \\ &= \exp\left[\mathrm{i}\left(r_0 + \sum_j v_j\mu_j\right) r - \left(\sum_j |v_j|\gamma_j\right)|r|\right],\end{aligned}$$

If the displayed scale is positive, this is the characteristic function of the Cauchy law in Equation (20). If every stochastic coefficient vanishes, the same calculation gives the point mass at the displayed location. □

B.2 Closed-form end-to-end factual likelihood

Let

$$c_t = a_0 + tb_0, \qquad w_t = a + tb,$$

and collect the treatment-mechanism parameters in $\theta = (a_0, a, b_0, b, \sigma_0, \sigma_1)$. Conditional on $X = x$, the abductive model is

$$q_\phi(u \mid x) = \prod_{j=1}^{d} \frac{1}{\pi\gamma_{\phi,j}(x)}\left[1 + \left\{\frac{u_j - \mu_{\phi,j}(x)}{\gamma_{\phi,j}(x)}\right\}^2\right]^{-1}. \tag{48}$$

Given (u, t), the structural mechanism and Cauchy event noise imply

$$p_\theta(y \mid u, t) = \frac{1}{\pi\sigma_t}\left[1 + \left\{\frac{y - c_t - w_t^\top u}{\sigma_t}\right\}^2\right]^{-1}, \tag{49}$$

for $\sigma_t > 0$. The conditional factual likelihood is obtained by reducing over the latent unit:

$$p_{\phi,\theta}(y \mid x, t) = \int_{\mathbb{R}^d} p_\theta(y \mid u, t) q_\phi(u \mid x)\, du. \tag{50}$$

We now evaluate this integral exactly. The characteristic function of $U_j \mid X = x$ is

$$\varphi_{U_j|x}(r) = \exp\{\mathrm{i}\mu_{\phi,j}(x)r - \gamma_{\phi,j}(x)|r|\},$$

and the characteristic function of E_t is $\varphi_{E_t}(r) = \exp(-|r|)$. Conditional independence therefore gives

$$\varphi_{Y|x,t}(r) = \exp(\mathrm{i}c_t r)\prod_{j=1}^{d}\varphi_{U_j|x}(w_{t,j}r)\varphi_{E_t}(\sigma_t r) \tag{51}$$

$$= \exp\left[\mathrm{i}\left\{c_t + w_t^\top\mu_\phi(x)\right\} r - \left\{\sum_{j=1}^{d}|w_{t,j}|\gamma_{\phi,j}(x) + \sigma_t\right\}|r|\right]. \tag{52}$$

This is the characteristic function of

$$Y \mid X = x, T = t \sim \mathrm{Cauchy}(m_t(x), s_t(x)), \tag{53}$$

where

$$m_t(x) = c_t + w_t^\top\mu_\phi(x), \tag{54}$$

$$s_t(x) = \sum_{j=1}^{d}|w_{t,j}|\gamma_{\phi,j}(x) + \sigma_t. \tag{55}$$

Hence Equation (50) has the closed-form density

$$p_{\phi,\theta}(y \mid x,t) = \frac{1}{\pi s_t(x)} \left[1 + \left\{ \frac{y - m_t(x)}{s_t(x)} \right\}^2 \right]^{-1}. \tag{56}$$

For i.i.d. observations $\mathcal{D}_n = \{(x_i, t_i, y_i)\}_{i=1}^n$, ordinary conditional maximum likelihood corresponds to $\omega_i = 1$. More generally, fixed nonnegative weights with $\sum_i \omega_i > 0$ define the weighted empirical risk

$$\mathcal{L}_{\mathrm{NLL}}(\phi,\theta) = \frac{1}{\sum_i \omega_i} \sum_{i=1}^n \omega_i \left[\log\{\pi s_i\} + \log\left\{ 1 + \frac{(y_i - m_i)^2}{s_i^2} \right\} \right], \tag{57}$$

where $m_i = m_{t_i}(x_i)$ and $s_i = s_{t_i}(x_i)$. The reported experiments use $\omega_i = 1$; AdamW decay 10^{-5} and norm-10 gradient clipping are optimizer choices, not extra terms in this derived likelihood.

B.3 Joint gradient propagation

For a single observation, write $r_i = y_i - m_i$ and let ℓ_i denote its unweighted term in Equation (57). Direct differentiation gives

$$g_{m,i} := \frac{\partial \ell_i}{\partial m_i} = \frac{2(m_i - y_i)}{s_i^2 + r_i^2}, \tag{58}$$

$$g_{s,i} := \frac{\partial \ell_i}{\partial s_i} = \frac{s_i^2 - r_i^2}{s_i(s_i^2 + r_i^2)}. \tag{59}$$

Because m_i and s_i depend on the encoder outputs, the exact encoder gradient is

$$\nabla_\phi \ell_i = g_{m,i} \sum_{j=1}^d w_{t_i,j} \nabla_\phi \mu_{\phi,j}(x_i) \tag{60}$$

$$+ g_{s,i} \sum_{j=1}^d |w_{t_i,j}| \nabla_\phi \gamma_{\phi,j}(x_i). \tag{61}$$

Thus every factual observation generally updates both the location and scale outputs of the shared abductive encoder.

For the treatment mechanism, let $v_i = \mathrm{sign}(w_{t_i}) \odot \gamma_\phi(x_i)$, where $\odot$ denotes componentwise multiplication. Away from zero-valued mechanism coefficients, the gradients are

$$\frac{\partial \ell_i}{\partial a_0} = g_{m,i}, \qquad \frac{\partial \ell_i}{\partial b_0} = t_i g_{m,i}, \tag{62}$$

$$\nabla_a \ell_i = g_{m,i} \mu_\phi(x_i) + g_{s,i} v_i, \qquad \nabla_b \ell_i = t_i \left\{ g_{m,i} \mu_\phi(x_i) + g_{s,i} v_i \right\}, \tag{63}$$

$$\frac{\partial \ell_i}{\partial \sigma_{t_i}} = g_{s,i}. \tag{64}$$

Equations (61) and (64) show that a single factual objective jointly trains ϕ and θ. There is no separate Abduction loss, no stage-wise freezing, and no Monte Carlo estimator of the latent integral. Each unit contributes only through its factual arm t_i, while observations from both arms jointly learn the shared encoder and mechanism across the population.

The absolute-value terms are nondifferentiable only when a component of w_t is exactly zero. The implementation retains the exact absolute value and uses its automatic-differentiation subgradient at zero. Enforcing $s_t(x) \geq s_{\min} > 0$ makes the displayed likelihood and gradients finite.

Trainability versus identification. The derivation proves closed-form marginalization and joint differentiability almost everywhere. It does not prove that ϕ and θ are separately identifiable: different latent parameterizations can induce the same observable response kernels. Nor does differentiability establish conditional exchangeability. End-to-end trainability is therefore a computational property of the model; causal interpretation still depends on Assumption 1, and parameter or kernel recovery requires additional statistical assumptions.

B.4 Proofs for marginal identification and exchangeability

Proof of Proposition 1. For each t and every continuity point y, consistency and conditional exchangeability give

$$\begin{aligned}\mathbb{P}(Y \le y \mid T = t, X = x) &= \mathbb{P}(Y(t) \le y \mid T = t, X = x) \\ &= \mathbb{P}(Y(t) \le y \mid X = x) = F_t(y \mid x).\end{aligned}$$

Positivity ensures that the observed conditional law exists for both arms on the target support. Applying the generalized inverse to each identified CDF identifies $Q_t(p \mid x)$ and their difference. □

Proof of Proposition 2. The pair $(Y(0), Y(1))$ is a measurable function of (X, U, E_0, E_1) under Equation (7). Conditional independence is preserved under measurable transformations, so $T \perp\!\!\!\perp (U, E_0, E_1) \mid X$ implies $T \perp\!\!\!\perp (Y(0), Y(1)) \mid X$. If the former independence fails, equality between the observed arm-conditional law and the potential-outcome marginal is not entailed; fitting the former therefore does not by itself recover the latter. □

B.5 Marginal quantiles versus effect quantiles

Proof of Proposition 3. Let $V \sim \mathrm{Uniform}(0, 1)$, set $Y(0) = V$, and define

$$Y(1) = (V + 1/4) \bmod 1.$$

Both potential outcomes are uniform on $(0, 1)$, so $Q_1(p) - Q_0(p) = 0$ for every p. Their within-unit difference equals $1/4$ with probability $3/4$ and $-3/4$ with probability $1/4$. Hence $\mathrm{med}\{Y(1) - Y(0)\} = 1/4$, while the difference of marginal medians is zero. Changing the coupling while preserving both marginals can therefore change the effect quantiles without changing any τ_p. □

B.6 Fisher consistency of the factual objective

Proof of Theorem 1. Condition on $(X, T) = (x, t)$. The theorem's shared correctly specified parameter $\vartheta^\star$ realizes both arms and has finite weighted cross-entropy. For any finite-risk candidate, subtracting the risk at $\vartheta^\star$ from the candidate risk gives

$$R_\omega(\vartheta) - R_\omega(\vartheta^\star) = \mathbb{E}_{X,T}[\omega(X, T) D_{\mathrm{KL}}\{p^\star(\cdot \mid X, T) \,\|\, p_\vartheta(\cdot \mid X, T)\}] .$$

The expression is nonnegative. Because $\omega \ge c > 0$, equality holds only when the conditional KL divergence is zero almost surely. Cauchy densities are equal almost everywhere only when both location and scale agree, giving $m_t = m_t^\star$ and $s_t = s_t^\star$. Positivity gives positive probability to both treatment arms on the target support. Assumption 1 then converts the recovered observed response kernels into the corresponding potential-outcome marginals, and their median difference is $m_1^\star - m_0^\star$. □

B.7 Latent reparameterization non-identifiability

Proof of Proposition 6. Because D is diagonal and is restricted as stated in the proposition, the transformed encoder remains in the declared model class and respects its scale floor. The coordinates of $U' = DU + h$ remain conditionally independent Cauchy variables with locations and scales shown in Equation (22). Moreover,

$$\begin{aligned}c_t' + {w_t'}^\top \mu_\phi'(x) &= c_t - {w_t'}^\top h + {w_t'}^\top \{D\mu_\phi(x) + h\} \\ &= c_t + w_t^\top \mu_\phi(x), \\ \sum_j |w_{t,j}'| \gamma_{\phi,j}'(x) &= \sum_j \left| \frac{w_{t,j}}{D_{jj}} \right| |D_{jj}| \gamma_{\phi,j}(x) = \sum_j |w_{t,j}| \gamma_{\phi,j}(x).\end{aligned}$$

The event scale is unchanged, so both observable response kernels are identical. Finally, set $a' = w_0'$, $b' = w_1' - w_0'$, $a_0' = c_0'$, and $b_0' = c_1' - c_0'$ to recover the treatment-indexed parameterization. Thus a continuum of latent coordinates and mechanism parameters represents the same observable law. □

B.8 Proofs for the coupling-specific effect laws

Proof of Proposition 7. Under independent resampling, the mutual conditional independence in Assumption 2 permits repeated use of affine Cauchy closure, and

$$\Delta_{\text{ind}} = b_0 + b^\top U + \sigma_1 E_1 - \sigma_0 E_0.$$

Applying Proposition 4 gives location $b_0 + b^\top \mu_\phi(x) = \tau_{\text{loc}}(x)$ and scale $\sum_j |b_j| \gamma_{\phi,j}(x) + \sigma_1 + \sigma_0$. Under shared exogenous noise, $E \perp U \mid X$ likewise gives

$$\Delta_{\text{shared}} = b_0 + b^\top U + (\sigma_1 - \sigma_0) E,$$

which has the same location and scale $\sum_j |b_j| \gamma_{\phi,j}(x) + |\sigma_1 - \sigma_0|$. If this scale is positive, then for $Z \sim \text{Cauchy}(m, s)$, $\mathbb{P}(Z > 0) = 1/2 + \pi^{-1} \arctan(m/s)$, proving Equation (19). If the shared scale is zero, the contrast is a point mass at $\tau_{\text{loc}}(x)$ and the benefit probability is $\mathbb{I}\{\tau_{\text{loc}}(x) > 0\}$. □

Proof of Proposition 8. A non-degenerate Cauchy variable has $\int |z| \, dF(z) = \infty$ and therefore no expectation. Each selected effect law is Cauchy by Proposition 7; the claim follows whenever its scale is positive. □

B.9 Constructive cross-world non-identification

Proof of Proposition 9. Fix x, let $d = 1$, $U \sim \text{Cauchy}(0, 1)$, and take $E_0, E_1 \sim \text{Cauchy}(0, 1)$ with equal event scales $\sigma_0 = \sigma_1 = \sigma > 0$ and zero intercepts. Consider two models:

$$\mathcal{M}_A : \quad w_0 = w_1 = 1 \qquad \text{and} \qquad \mathcal{M}_B : \quad w_0 = 1, \; w_1 = -1.$$

Equivalently, $\mathcal{M}_A$ has $(a, b) = (1, 0)$ and $\mathcal{M}_B$ has $(a, b) = (1, -2)$. In both models and both treatment arms,

$$Y(t) \mid X = x \sim \text{Cauchy}(0, 1 + \sigma),$$

so the two factual marginal laws are identical. Under independent resampling, $\mathcal{M}_A$ has effect scale 2σ, whereas $\mathcal{M}_B$ has effect scale $2 + 2\sigma$. Under shared exogenous noise, $\mathcal{M}_A$ gives the point mass at zero, whereas $\mathcal{M}_B$ gives a positive-scale $\text{Cauchy}(0, 2)$ law. Thus neither pair of factual marginals determines the effect law under a chosen cross-world structural parameterization. □

B.10 Makarov bounds and bounded-score proofs

Justification of Theorem 2. For any y, the event $\{Y(1) \le y, Y(0) > y - \delta\}$ is contained in $\{Y(1) - Y(0) \le \delta\}$. The Fréchet inequality therefore gives the lower bound in Equation (41); applying the complementary Fréchet inequality gives the upper bound. Optimizing over y yields the displayed bounds. For each fixed δ, their pointwise sharpness over all couplings with continuous marginals is the classical Makarov result; a single coupling need not attain the entire envelope simultaneously. Evaluating at zero and taking complements gives Equation (23). □

Proof of Proposition 10 and Corollary 2. Direct differentiation gives Equation (24). With $r = y - m$, $2|r|/(s^2 + r^2) \le 1/s$ because $(|r| - s)^2 \ge 0$, and $|s^2 - r^2| \le s^2 + r^2$. The chain rule gives

$$\nabla_\vartheta \ell = \frac{\partial \ell}{\partial m} \nabla_\vartheta m + \frac{\partial \ell}{\partial \log s} \nabla_\vartheta \log s.$$

The triangle inequality and the assumed Jacobian bounds yield Equation (25). This controls outcome-residual contributions only; it does not establish bounded influence when covariate leverage or the estimating-equation inverse is uncontrolled. □

B.11 Cauchy quantiles

If $Z \sim \text{Cauchy}(m, s)$, then

$$F_Z(z) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{z - m}{s}\right), \qquad Q_Z(p) = m + s \tan\left(\pi \left(p - \tfrac{1}{2}\right)\right).$$

Substituting $(m_t(x), s_t(x))$ gives Equation 11.

B.12 SHARED-REALIZATION SENSITIVITY

If the same event-noise realization is reused in both potential outcomes,

$$Y(t) = a_0 + a^\top U + t(b_0 + b^\top U) + \sigma_t E,$$

then

$$\Delta_{\text{shared}} = b_0 + b^\top U + (\sigma_1 - \sigma_0)E.$$

Consequently, when $\sum_j |b_j|\gamma_{\phi,j}(x) + |\sigma_1 - \sigma_0| > 0$,

$$\Delta_{\text{shared}} \mid X = x \sim \text{Cauchy}\left(\tau_{\text{loc}}(x), \sum_j |b_j|\gamma_{\phi,j}(x) + |\sigma_1 - \sigma_0|\right).$$

When that scale is zero, the contrast is instead the point mass at $\tau_{\text{loc}}(x)$. When $\sigma_1 = \sigma_0$, the event noise cancels exactly. This is not the same as independent event-noise resampling, whose scale is given in Equation 17; the two are distinct couplings and must be declared rather than inferred from factual data.

B.13 WHY THE STANDARD MEAN EFFECT IS UNAVAILABLE

For $Z \sim \text{Cauchy}(m, s)$ with $s > 0$, the integral $\int |z|\, dF_Z(z)$ diverges. Hence neither $\mathbb{E}[Z]$ nor $\mathbb{E}[Y(1) - Y(0) \mid X = x]$ is defined whenever the induced effect is non-degenerate Cauchy. The location contrast in Equation 3 is a quantile functional and remains well-defined.

B.14 REPRODUCIBILITY ARTIFACT CONTENTS

The checked experiment bundle records:

1. the exact synthetic data-generating equations and random seeds;
2. the treatment propensity and overlap diagnostics;
3. whether evaluation uses marginal quantiles or a selected cross-world coupling;
4. the scale floor and optimizer regularization;
5. per-run metrics and uncertainty summaries for every reported baseline, plus raw out-of-fold predictions for WAWS; and
6. controlled comparisons against direct S- and T-learners, with exact parameter counts.

Synthetic raw prediction tensors are not retained, and the planned Cauchy AB-HTE mechanism-mode ablations are not implemented. These are explicit artifact limitations rather than completed checklist items.

C FULL EMPIRICAL EVALUATION PROTOCOL

The evaluation is organized by what causal ground truth is available, rather than by the order in which experiments are implemented. This distinction is essential: fully synthetic data can reveal both potential outcomes and the stipulated cross-world coupling; semi-synthetic data can reveal designed response surfaces but not necessarily a realized individual effect; and real data reveal only factual outcomes. Numerical cells are populated only when produced by a checked dataset-to-command-to-artifact pipeline.

Across all layers, the learner receives only factual triples (X, T, Y). Oracle response surfaces, both potential outcomes, and coupling-dependent effect laws are evaluation objects, never training targets or model-selection criteria. The primary robustness study extends the label-corruption protocol of Causal Regression to binary treatment effects, but targets the complete fitted algorithm rather than the training loss alone. Training and validation receive independent draws from the same corruption operator and rate; the final causal evaluation population alone remains clean. A clean-validation oracle is kept only as a weaker decomposition ablation.

C.1 Research questions

RQ1: Clean effect recovery. Does Cauchy AB-HTE estimate $\tau_{\mathrm{loc}}(x)$ accurately on clean randomized data, both inside and outside the Cauchy–affine model family?

RQ2: Pipeline robustness. When the same contaminated record process generates training and validation data, how quickly does clean-test effect accuracy degrade after corrupted checkpoint selection?

RQ3: Source of robustness. How much is attributable to the Cauchy likelihood, distribution-valued abduction, event noise, and their analytic reduction, rather than to network capacity alone?

RQ4: Causal structure of corruption. Does robustness change when corruption breaks both treatment and covariate pairing, breaks covariate pairing only within treatment arms, corrupts treatment records, or affects the two arms differently?

RQ5: Marginal and coupled distributions. Are the two identified potential-outcome marginals calibrated, and how sensitive are nonidentified within-unit effect summaries to the stipulated cross-world coupling?

C.2 Three layers separated by available truth

Layer 1: fully synthetic data with oracle effects. This is the only layer in which individual counterfactual objects can be made fully known by construction. It contains three deliberately separated substudies.

First, a realizable sanity check generates data exactly from the Cauchy–affine SCM in Section 4. It verifies Cauchy scale propagation, end-to-end gradient flow, and recovery of $m_t(x)$, $s_t(x)$, and $\tau_{\mathrm{loc}}(x)$. Because the generator favors Cauchy AB-HTE by construction, this is engineering evidence and is not used as the primary evidence for superiority.

Second, the primary robustness study deliberately lies outside the Cauchy AB-HTE family. It draws $X \sim \mathcal{N}(0, I_6)$ and defines

$$g^\star(x) = .55 \sin x_1 + .20 x_2^2 - .30 x_3 x_4 + .35 \tanh x_5 - .15 x_6, \tag{65}$$

$$\tau^\star(x) = .10 + .35 \tanh(x_1 + .5 x_2) - .25 \sin x_3 + .15 x_4, \tag{66}$$

$$Y(t) = g^\star(X) + t\tau^\star(X) + \epsilon_t, \qquad \epsilon_0, \epsilon_1 \overset{\text{iid}}{\sim} \mathcal{N}(0, .35^2). \tag{67}$$

Treatment is randomized with $\mathbb{P}(T = 1) = 1/2$. Both potential outcomes and the oracle response locations are retained. Each confirmatory replication has $n = 3{,}000$ units and uses frozen generator seeds 100–109; the learner observes only

$$Y_i^{\mathrm{obs}} = T_i Y_i(1) + (1 - T_i) Y_i(0). \tag{68}$$

Third, the implemented observational stress test uses one prespecified propensity, not an overlap sweep:

$$e^\star(x) = \mathrm{clip}_{[.05,.95]}\left[\mathrm{logit}^{-1}(.65 x_1 - .55 x_2 + .30 x_3 x_4 + .25 \sin x_5)\right]. \tag{69}$$

For the hidden-confounding negative control, $H \sim \mathcal{N}(0, 1)$ is omitted from the learner, $1.25H$ is added to the propensity logit and to $Y(0)$, and $2H$ is added to $Y(1)$. Failure is expected and marks the identification boundary; no multiple-strength or restricted-population sweep is claimed.

The oracle synthetic record stores $m_0^\star(x)$, $m_1^\star(x)$, $\tau_{\mathrm{loc}}^\star(x)$, and $e^\star(x)$. A joint effect law or probability of benefit is treated as oracle truth only when the generator also specifies the joint noise and cross-world coupling. Merely retaining two marginal response surfaces does not create individual-effect ground truth.

Layer 2: semi-synthetic data with response-surface truth. This layer combines real covariates and, where appropriate, real treatment patterns with simulated outcomes from a documented benchmark response surface. Candidate designs include verified IHDP and current ACIC simulation tracks, subject to exact version, license, preprocessing, and target checks. For every benchmark we record whether it supplies conditional response locations, realized paired potential outcomes, or only factual and counterfactual simulation draws.

When the benchmark supplies $\mu_0^\star(x)$ and $\mu_1^\star(x)$, the valid oracle target is the response-surface contrast

$$\tau_\mu^\star(x) = \mu_1^\star(x) - \mu_0^\star(x). \tag{70}$$

It is not automatically the realized individual contrast $Y_i(1) - Y_i(0)$. Coupling-dependent metrics are used only if the simulator explicitly defines and exposes the joint potential-outcome law. Thus this layer tests recovery on realistic covariate and assignment distributions without silently upgrading simulated response-surface truth into real individual counterfactual truth.

Layer 3: real RCT and observational data with factual outcomes only. No real-data record is treated as if both potential outcomes were observed. For randomized trials, known assignment probabilities support design-based arm contrasts and honest off-policy evaluation. For observational data, corresponding causal summaries require consistency, conditional exchangeability, and positivity; propensity modeling does not remove hidden confounding.

Real-data evaluation is therefore restricted to held-out factual scoring and calibration, design-based marginal contrasts in trials, honest treatment-effect stratification, and cross-fitted policy evaluation. Individual-effect error, PEHE, probability-of-benefit calibration, and validation of a cross-world coupling are not reported as ground-truth metrics on real data. Coupling-based effect summaries may be shown only as model-implied sensitivity analyses, preferably alongside assumption-lean bounds.

The implemented real trial is WAWS (Lachowska et al., 2015). We first apply the official long-term analysis window 1986-W40–1987-W18, yielding $n = 4{,}218$ (1,606 exception-reporting and 2,612 pooled comparison units). This restriction is essential: exception reporting stopped after W18, so later public-use rows lack treatment positivity; the earliest cohort also lacks the common follow-up window. The documented pseudo-SSN last-digit rule implies $e = 3/8$ among the retained ER/control/NWS arms. Here action one is exception reporting (ER), whereas action zero is the design-induced historical comparator that mixes Control and NWS in a 3:2 ratio; it is not a single deterministic intervention. The full 7,059-row archive is never used as the primary randomized estimand. A conservative 25-variable pretreatment allowlist excludes union/standby status because the original study documents treatment-differential recording of that field.

C.3 Partitioning and leakage-free model selection

For every fully synthetic or semi-synthetic replication, the clean data are split $70/15/15$ into train, validation, and test partitions before any corruption. In the primary protocol, independently sampled corruption draws with a common operator and rate are then applied to the factual train and validation records:

$$\widehat{\tau}_\rho^{\,\mathrm{pipe}} = \mathcal{A}(C_\rho(D_{\mathrm{tr}}; \xi_{\mathrm{tr}}), C_\rho(D_{\mathrm{val}}; \xi_{\mathrm{val}}))\,, \qquad \xi_{\mathrm{tr}} \perp\!\!\!\perp \xi_{\mathrm{val}}, \tag{71}$$

$$\widehat{\tau}_\rho^{\,\mathrm{trusted}} = \mathcal{A}\big(C_\rho(D_{\mathrm{tr}}; \xi_{\mathrm{tr}}), D_{\mathrm{val}}^0\big)\,. \tag{72}$$

The same operator-specific transformation applies to Y or, for treatment misclassification, to T. Validation controls early stopping and checkpoint selection, but never exposes the unobserved arm. Hyperparameters are frozen on pilot data before the confirmatory seeds, so the implemented validation stage does not perform a new hyperparameter grid search. Any such search would also have to use $C_\rho(D_{\mathrm{val}})$ under the primary protocol.

Both fitted algorithms are evaluated once on the untouched D_{te}^0, whose oracle surface is accessed only after selection. Their difference is a useful model-selection-fragility diagnostic, but not an additive decomposition: changing validation corruption generally changes the selected checkpoint and its interaction with the corrupted training sample. Equation (72) therefore answers the weaker question of robustness given a trusted validation oracle and is reported only in the appendix.

Every method within a replication receives the same split and the same partition-specific corrupted indices, donors, and perturbations. Training and validation corruption seeds differ by a fixed offset of 10^6 and every audit record names its partition. Selection uses only the model's factual validation objective. Oracle effect error is never used for early stopping, hyperparameter selection, architecture selection, or selection of a corruption level. Pilot replications may be used to freeze search spaces and budgets; paper-level comparisons are then computed on separate confirmatory replications.

For real data, model selection and policy evaluation use outer sample splitting or cross-fitting. The effect model used to evaluate a unit is fitted without that unit. WAWS outer folds are functions only

of pretreatment assignment season, not the held-out unit's treatment or outcome; treatment-by-season stratification is used only for the inner training/validation split. A single fitted model is not reused both to learn a policy and to report its naive in-sample value.

C.4 Reference-aligned corruption operators

For each partition $p \in \{\mathrm{tr}, \mathrm{val}\}$, let $\mathcal{I}_p$ index its clean observations and independently draw $\mathcal{S}_{\rho,p} \subset \mathcal{I}_p$ uniformly with size $\lfloor \rho |\mathcal{I}_p| \rfloor$. We sweep

$$\rho \in \{0, .1, .2, .3, .4, .5\}. \tag{73}$$

The clean level $\rho = 0$ is included in every curve. The operator definition is identical across train and validation but all locations, donors, signs, and partition statistics are independently drawn or computed. No operator touches the test partition.

Global outcome shuffle. Within each partition, draw a random permutation π_p of all its indices and replace

$$\widetilde{Y}_i^{\mathrm{obs}} = \begin{cases} Y_{\pi_p(i)}^{\mathrm{obs}}, & i \in \mathcal{S}_{\rho,p}, \\ Y_i^{\mathrm{obs}}, & i \notin \mathcal{S}_{\rho,p}. \end{cases} \tag{74}$$

This is label permutation, not a shuffle of complete rows. It breaks selected labels' association with both X_i and T_i, changes treatment-arm outcome marginals, and asymptotically approximates

$$\widetilde{P}_\rho(Y \mid X, T) = (1 - \rho) P(Y \mid X, T) + \rho P(Y). \tag{75}$$

It tests robustness to combined covariate–outcome and treatment–outcome mismatch.

Within-arm outcome shuffle. Within each partition, selected outcomes are permuted among the selected indices separately inside $T = 0$ and $T = 1$. Unselected outcomes remain fixed. Consequently, the complete empirical outcome multiset in each arm is preserved exactly even for partial ρ, while the selected individual $X \mapsto Y(t)$ pairing. It therefore targets loss of prognostic and effect-modifier information without adding cross-arm label mismatch.

Treatment-record shuffle. As a separate causal diagnostic, selected treatment indicators are flipped while (X, Y^{obs}) are held fixed. This operator represents treatment misclassification rather than outcome corruption and is never pooled with the primary outcome-corruption curves. It is needed because the difference between global and within-arm outcome shuffle alone also contains a change in arm marginals and cannot uniquely isolate treatment-record error.

Outlier replacement. For $i \in \mathcal{S}_{\rho,p}$, the operator replaces the clean label by $\bar{Y}_p + d_i(3 s_{Y,p})$, where d_i is uniform on $\{-1, +1\}$ and $\bar{Y}_p, s_{Y,p}$ are computed from that clean partition before corruption.

Asymmetric shift. For selected observations add $+2 s_{Y,p}$ with probability .9 and $-2 s_{Y,p}$ with probability .1. This creates directional reporting error without changing the clean test estimand.

Systematic shift. For selected observations add the fixed offset $+1.5 s_{Y,p}$. This represents a consistent measurement drift in a random subset of model-development records.

Treatment-differential corruption. The implemented arm-differential operator selects a global fraction ρ and adds $-1.5 s_{Y,p}$ to selected control records and $+1.5 s_{Y,p}$ to selected ER records. It can manufacture a spurious contrast even when selection is random. More general unequal arm-specific rates are not part of the current artifact. All causal-specific settings remain separately named and are not pooled into a generic robustness number.

These operators alter the observed model-development law, not the clean causal target. We report complete robustness curves rather than selecting a single favorable corruption level; the 40% slice is used only for a compact table matching the predecessor presentation.

Table 8: Prespecified overlap diagnostics (means over ten seeds). ESS uses the known generator propensity; balance is mean absolute standardized difference before/after inverse-probability weighting.

Regime	$e_{.05}$–$e_{.95}$	ESS_0	ESS_1	SMD before	SMD after
Randomized	0.500–0.500	223.8	226.2	0.074	0.074
Observed confounding	0.185–0.813	186.0	181.2	0.237	0.083
Hidden confounding	0.080–0.927	120.4	132.3	0.184	0.113

C.5 Identification stress test and overlap diagnostics

The randomized synthetic benchmark is primary because it isolates outcome corruption from propensity estimation. A secondary observational generator draws treatment from the single nonlinear $e^\star(X)$ in Equation (69) while retaining conditional exchangeability. The hidden-confounding negative control adds the one fixed unobserved common-cause specification described above. These are three fixed regimes; the implementation does not claim a propensity-strength sweep.

For each observational setting we report treatment-specific propensity distributions, common-support summaries, clipping fractions, maximum weights, and arm-wise effective sample size

$$\mathrm{ESS}_t = \frac{\left(\sum_{i:T_i=t} w_i\right)^2}{\sum_{i:T_i=t} w_i^2}. \tag{76}$$

We also report covariate balance before and after weighting on the original target population. No overlap-restricted population is used in the current artifact. Propensity calibration is a diagnostic, not evidence that hidden confounding is absent.

The hidden-confounding regime tests whether the method fails at the advertised identification boundary. Hidden confounding is not diagnosed from observed fit; the synthetic oracle reveals the failure only because the DGP is known. No real observational sensitivity analysis is presented in this paper.

C.6 Baselines and controlled comparisons

The implemented synthetic comparison set mirrors the robustness logic of the predecessor while respecting the new estimand:

- Gaussian AB-HTE is architecture and initialization matched to Cauchy AB-HTE;
- Huber and median-pinball S-learners provide robust point controls; and
- direct Cauchy S- and T-learners test whether the likelihood alone is sufficient without Cauchy AB-HTE's shared abductive representation.

IHDP and WAWS additionally include TARNet and a CFR-style linear-MMD learner (Shalit et al., 2017). Causal forests, R-learners, Dragonnet, and broader conditional-distributional learners are not implemented and are not described as empirical controls.

The implemented contrasts have distinct interpretations:

1. Cauchy AB-HTE versus Gaussian AB-HTE isolates the distributional choice;
2. Cauchy AB-HTE (1,556 parameters) versus direct Cauchy S (1,378 parameters) tests whether shared abduction helps beyond a robust likelihood, subject to their small but explicit capacity difference; and
3. matched-corrupted versus clean validation tests pipeline selection robustness versus access to a trusted validation oracle.

All neural comparisons use the same 120-epoch maximum, patience 18, batch size 128, learning rate 10^{-3}, AdamW decay 10^{-5}, norm-10 gradient clipping, split, and partition-specific corruption draws. The latent dimension is four. Mechanism-mode, interaction-removal, and Monte-Carlo-reduction ablations remain future work rather than missing table cells.

C.7 LAYER-VALID METRICS

Synthetic and semi-synthetic response-surface metrics. The primary causal robustness metric is median absolute location-effect error on a clean oracle test population:

$$\mathrm{MedAE}_\tau = \mathrm{median}_{i\in\mathcal{I}_{\mathrm{te}}} \left|\widehat{\tau}_{\mathrm{loc}}(x_i) - \tau^\star_{\mathrm{loc}}(x_i)\right|. \tag{77}$$

Secondary metrics are integrated absolute location error, effect RMSE when the location-error target has a finite second moment, Spearman rank correlation, sign accuracy, and oracle location-policy regret

$$\mathrm{Regret}_{\mathrm{loc}} = \frac{1}{|\mathcal{I}_{\mathrm{te}}|}\sum_{i\in\mathcal{I}_{\mathrm{te}}}\left[\max_{t\in\{0,1\}} m^\star_t(x_i) - m^\star_{\widehat{\pi}(x_i)}(x_i)\right], \qquad \widehat{\pi}(x) = \mathbb{I}\{\widehat{\tau}_{\mathrm{loc}}(x) > 0\}. \tag{78}$$

This is explicitly a location-based decision criterion, not an expected-outcome value for a Cauchy variable.

Observed-arm response distributions are assessed using held-out factual log score and PIT/CDF calibration. For a Cauchy target, finite-moment scoring rules are not given a population-risk interpretation. Interval coverage and a full proper-scoring-rule suite are not reported in the current artifact.

Effect-law quantile accuracy, effect scale, and probability-of-benefit calibration are reported only when the synthetic generator specifies the true cross-world coupling. A semi-synthetic response surface without such a coupling supports location-effect evaluation but not these metrics.

For each method and operator we report both absolute error and degradation from its own clean baseline,

$$\Delta_\rho = \mathrm{MedAE}_\tau(\rho) - \mathrm{MedAE}_\tau(0), \qquad R_\rho = \frac{\mathrm{MedAE}_\tau(\rho)}{\mathrm{MedAE}_\tau(0)}. \tag{79}$$

Real-data factual and design-based metrics. On both trials and observational data we report only observable held-out metrics: factual log score, arm-wise MedAE, arm-wise PIT/CDF calibration, and treatment counts. In randomized trials, marginal arm quantiles and equal-count strata formed from cross-fitted effect scores can additionally receive randomized arm contrasts. Those contrasts are honest with respect to the held- out outcomes, but the reported first-order uncertainty conditions on the learned scores and does not integrate training variability. The summaries assess factual calibration and subgroup ranking; they are not individual-effect ground truth.

Policy evaluation uses a bounded utility $u(Y)$ chosen before examining policy results, so the value exists even when the Cauchy outcome mean does not. For an outer-fold policy $\widehat{\pi}^{(-k)}$ in a trial with known assignment probability e, the implemented Horvitz–Thompson estimator is

$$\widehat{V}(\widehat{\pi}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}\{T_i = \widehat{\pi}_i\}u(Y_i)}{\widehat{\pi}_i e + (1-\widehat{\pi}_i)(1-e)}, \tag{80}$$

where $\widehat{\pi}_i = \widehat{\pi}^{(-k(i))}(X_i)$. It fits neither a propensity nor an outcome nuisance.

For WAWS, $e = 3/8$, the rule is the sign of the out-of-fold location score, and $u(y) = \tanh(y/20{,}000)$. Action one assigns ER; action zero assigns the historical pooled comparator intervention. Conditional on a fixed cross-fitted rule, Equation (80) is design unbiased under the documented trial assignment, but it is not a score for an unobserved individual effect. We report the point differences from assign-ER and assign-comparator. The artifact also stores iid contribution standard errors conditional on the learned fold scores; because these omit overlapping cross-fit training and optimizer variability, we do not present them as formal learned-policy confidence intervals.

If $\widehat{\pi}$ is defined by the sign of a location contrast but evaluated under a different bounded utility, Equation (80) estimates the value of that fixed rule; it does not establish that the rule is optimal for the bounded utility. No observational policy-value result is reported in the current artifact.

C.8 REPLICATION AND REPORTING

Fully synthetic results are aggregated over at least ten independently generated datasets and corruption draws, with paired method comparisons, means, medians, and interval estimates. Semi-synthetic

Table 9: Model-selection robustness decomposition at $\rho = 0.4$ (clean-test effect MedAE, mean $\pm$ SE). Pipeline columns corrupt train and validation independently; trusted-validation columns corrupt train only. The difference is diagnostic, not an additive decomposition.

Method	Global, pipeline	Global, trusted val.	Outlier, pipeline	Outlier, trusted val.
Cauchy AB-HTE (ours)	0.103 ± 0.010	0.096 ± 0.008	0.086 ± 0.005	0.081 ± 0.003
Gaussian AB-HTE	0.186 ± 0.010	0.128 ± 0.009	0.206 ± 0.005	0.217 ± 0.006
Cauchy S	0.099 ± 0.010	0.091 ± 0.007	0.084 ± 0.005	0.082 ± 0.005
Huber S	0.131 ± 0.013	0.111 ± 0.008	0.128 ± 0.007	0.128 ± 0.006
Pinball S	0.107 ± 0.012	0.099 ± 0.009	0.092 ± 0.004	0.091 ± 0.005
Cauchy T	0.125 ± 0.007	0.131 ± 0.006	0.119 ± 0.007	0.124 ± 0.006

Table 10: Causal-specific record corruption at $\rho = 0.4$ (effect MedAE, mean $\pm$ SE).

Method	Treatment flip	Arm-differential
Cauchy AB-HTE (ours)	0.198 ± 0.005	0.646 ± 0.013
Gaussian AB-HTE	0.203 ± 0.007	0.935 ± 0.009
Cauchy S	0.179 ± 0.005	0.700 ± 0.009
Huber S	0.185 ± 0.005	0.927 ± 0.008
Pinball S	0.179 ± 0.006	0.727 ± 0.012
Cauchy T	0.195 ± 0.006	0.685 ± 0.012

results are aggregated over documented response-surface replications. A single random seed is insufficient evidence for a robustness claim.

All primary plots show clean accuracy together with degradation under corruption. Global, within-arm, treatment-record, and arm-differential corruptions remain separately labeled. Results from hidden-confounding negative controls are displayed as identification stress tests rather than included in an optimizer leaderboard.

C.9 Implementation status and causal-specific results

The clean-room package now implements the realizable Cauchy sanity check, the nonlinear finite-noise robustness sweep, global and within-arm shuffles, three shift/outlier operators, treatment-label flips, arm-differential shifts, observed- and hidden-confounding stress tests, overlap diagnostics, declared- coupling evaluation, and the hash-pinned IHDP-1000 loader. Every corruption draw records its partition, independent seed, selected indices, donors, and realized perturbations. The primary confirmatory sweep corrupts both training and validation; a same-seed clean-validation sweep is retained only as the trusted-oracle ablation. WAWS persists every out-of-fold factual location, scale, effect score, fold identity, and source-row index without manufacturing counterfactual labels.

The current implementation does not yet include all planned Cauchy AB-HTE uncertainty modes, causal forests/R-learners, or ACIC generation. These remain implementation limits rather than blank empirical claims.

D Extended Related Work

Heterogeneous treatment-effect learning. Potential-outcome methods estimate causal contrasts from observational or randomized data under explicit identification assumptions (Rubin, 1974; Rosenbaum & Rubin, 1983; Imbens & Rubin, 2015). Classical and modern learners include causal forests (Wager & Athey, 2018), meta-learners (Künzel et al., 2019), and the R-learner (Nie & Wager, 2021). Representation-learning approaches such as TARNet and CFRNet share features across treatment arms and regularize treated–control imbalance (Shalit et al., 2017). Dragonnet incorporates propensity estimation and targeted regularization (Shi et al., 2019). The taxonomy of Curth & van der Schaar (2021) clarifies when shared and treatment-specific components help different effect learners. Cauchy AB-HTE also shares a representation across arms, but its target is a Cauchy location/quantile contrast and its unit representation is distribution-valued. Its intended distinguishing properties are analytic propagation, an explicit cross-world structural coupling, and outcome-contamination

Table 11: Five-fold out-of-fold evaluation on the official WAWS RCT window ($n = 4{,}218$). The randomized mean ITT is 118 ± 366 dollars (estimate $\pm$ SE). Outer folds are treatment blind. Factual metrics, randomized contrasts, and policy point values use no individual-effect truth; formal learned-policy inference is not claimed.

Method	Factual MedAE ↓	NLL ↓	PIT err. ↓	Policy value ↑	Policy − all ER
Cauchy AB-HTE (ours)	5918	10.725	0.034	0.486	0.005
Gaussian AB-HTE	6308	10.522	0.022	0.473	−0.008
Cauchy S	5943	10.731	0.036	0.474	−0.007
Huber S	6243	–	–	0.472	−0.010
Pinball S	6016	–	–	0.478	−0.003
Cauchy T	6082	10.741	0.037	0.475	−0.006
TARNet	6310	–	–	0.494	0.012
CFR-style MMD	6314	–	–	0.494	0.013

robustness. These distinctions must be tested against similar-capacity direct Cauchy learners, with exact parameter counts reported, rather than inferred from architecture.

Generative counterfactual models and abduction. GANITE uses generative adversarial learning to impute counterfactual outcomes (Yoon et al., 2018), while CEVAE introduces latent variables to address hidden confounding under a generative model (Louizos et al., 2017). Deep structural causal models formalize abduction–action–prediction with deep generative mechanisms (Pawlowski et al., 2020). Cauchy AB-HTE does not claim to be the first latent or abductive counterfactual model. It studies a deliberately narrow Cauchy–affine family in which reduction over latent-unit uncertainty and event noise is closed form. The latent coordinates themselves are not generally identified by (X, T, Y) alone; our causal claims concern observable response kernels and explicitly stated structural contrasts, not recovery of a unique "true" representation.

Distributional treatment effects. Conditional distributional effects capture treatment-induced changes beyond the mean. Kernel conditional mean embeddings compare potential-outcome distributions (Park et al., 2021); collaborating causal networks learn potential-outcome laws for decision making (Zhou et al., 2022); robust and agnostic learners target conditional quantiles, super-quantiles, and risk functionals (Kallus & Oprescu, 2023); and DiffPO uses diffusion models with orthogonal losses (Ma et al., 2024). Cauchy AB-HTE sacrifices distributional flexibility for exact Cauchy quantiles, scales, and likelihoods. Here "robustness" refers to bounded scores under outcome contamination, whereas the robustness of orthogonal distributional learners concerns nuisance-model misspecification. The two are complementary.

Cross-world dependence and individual-effect distributions. Marginal laws of $Y(0)$ and $Y(1)$ do not determine their joint law or the law of $\Delta = Y(1) - Y(0)$. Sharp partial-identification bounds follow from Fréchet–Hoeffding/Makarov arguments (Fan & Park, 2010), and recent neural work estimates uncertainty bounds without asserting a point-identified effect distribution (Melnychuk et al., 2024). Latent-variable models can identify benefit/harm rates only under additional restrictions (Yin et al., 2018); latent rank invariance with proxies is another route (Shin, 2024). Cauchy AB-HTE instead declares a shared-unit structural coupling and derives the effect law it implies. We present that law alongside its non-identification boundary and propose comparison with assumption-lean Makarov bounds.

Heavy tails and robust causal estimation. Robust regression and quantile regression provide classical tools for limiting the impact of extreme outcomes (Huber, 1964; Koenker & Bassett, 1978). Weighted absolute-deviation methods have been adapted to heterogeneous treatment effects (Li et al., 2023), and recent work studies calibrated heavy-tailed effect learning with robust pseudo-likelihoods (Uehara, 2026). Cauchy AB-HTE's claim is therefore not that robust HTE is unexplored. Its narrower contribution is the alignment of a Cauchy abductive distribution, treatment-indexed affine structural mechanism, analytic potential-outcome/effect propagation, and bounded factual scores.