TEMPORAL ELIGIBILITY ROUTING (TEAR) FOR TREATMENT-EFFECT ESTIMATION UNDER STRUCTURED COVARIATE CONTAMINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporal provenance and target-relative causal eligibility encode different information in structured covariate exports. Raw inclusion can expose treatment descendants, while temporal deletion can discard delayed records of pretreatment measurements. Temporal Eligibility Routing (TEAR) learns metadata-guided continuous gates and routes feature contributions through a finite-dimensional aggregate before conditional average treatment-effect estimation. A perturbation analysis connects routing-mass errors to representation and prediction perturbations for fixed embeddings and a fitted host. Controlled experiments show that inclusion and deletion favor opposite eligibility regimes, while gated aggregation helps when eligible and ineligible features coexist. Adding TEAR to an otherwise identical preprocessing candidate pool lowers test error measured by precision in estimation of heterogeneous effect (PEHE) in all eight dataset–host cells, with all paired 95% confidence intervals excluding zero. Observed-data selection recovers most of the PEHE-based improvement on MIMIC-P3, while provenance corruption drives gate discrimination toward chance and shifts selection toward Deletion. Code is available in an anonymous repository.

1 INTRODUCTION

Treatment-effect estimation relies on covariates that are admissible for the target estimand (Rosenbaum & Rubin, 1983; VanderWeele, 2019). Structured exports, including electronic health records (EHRs), can place baseline measurements, delayed records of pretreatment states, repeated measurements, and treatment descendants in the same static matrix (Stuart et al., 2013); we refer to this mixture as structured covariate contamination. Recording time and target-relative causal eligibility describe different properties of a feature: a delayed-neutral feature is recorded post-index while representing a pretreatment state, whereas a treatment descendant responds to treatment and can alter the admissible adjustment set. Because preprocessing determines the information available to every downstream treatment-effect estimator, structured covariate contamination creates two errors: exposing treatment descendants and discarding eligible delayed-neutral features.

Raw inclusion and temporal deletion define opposite preprocessing endpoints. Raw inclusion passes every recorded feature and can expose treatment descendants or other bad controls (Pearl, 2009; Cinelli et al., 2024), whereas deletion removes every schema-declared post-index feature and can discard delayed-neutral information. Predictive sparsity and masking retain features according to predictive utility. When delayed-neutral features and treatment descendants share post-index timing, eligibility remains feature specific. Temporal provenance provides uncertain evidence about that eligibility, motivating a continuous routing weight for each feature. TEAR applies the routing weights before finite-dimensional aggregation, so the downstream estimator receives a routed representation whose feature contributions are controlled by the gates.

We introduce Temporal Eligibility Routing (TEAR). TEAR derives a continuous eligibility gate from temporal provenance (timestamps and schema metadata) and applies it to learned feature embeddings. The gated embeddings form a fixed-dimensional host-visible aggregate supplied to the downstream conditional average treatment-effect (CATE) estimator, hereafter the host. A conservative temporal prior guides gate prewarming, and joint training then adapts the gate through the

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

training objective. We further formalize routing error through the visible mass assigned to ineligible features and the eligible mass withheld, and derive a perturbation bound connecting routing error to representation and prediction perturbations for fixed embeddings and a fitted prediction map.

Our experiments test this routing view from controlled eligibility disagreement to observed-data selection. Controlled regimes show that Raw and Deletion favor opposite eligibility conditions, while TEAR performs best when delayed-neutral features and treatment descendants coexist. Adding TEAR to the candidate family improves validation-selected PEHE in all eight dataset–host cells; for MIMIC-P3 with double machine learning (DML), PEHE decreases from 0.1983 to 0.1500. Observed-data selection recovers 69–78% of the PEHE-based improvement across the four MIMIC-P3 hosts. Under full provenance corruption, gate area under the receiver-operating-characteristic curve (AUROC) falls from 0.95 to 0.50, and selection shifts toward Deletion.

Our contributions are summarized below.

- **Target-relative eligibility routing.** We distinguish temporal provenance from target-relative causal eligibility and formulate structured-covariate preprocessing as continuous feature-level eligibility routing.
- **Gated aggregation and perturbation analysis.** We introduce metadata-guided finite-dimensional gated aggregation and derive a perturbation bound linking routing-mass error to representation and prediction perturbations.
- **Evaluation and observed-data selection.** We evaluate TEAR across controlled eligibility regimes, semi-synthetic EHR settings, multiple CATE hosts, observed-data selection, and provenance corruption, characterizing where eligibility routing improves downstream estimation and how its benefit changes with provenance quality.

2 Related Work

Heterogeneous treatment-effect estimation. Meta-learners reduce heterogeneous-effect estimation to combinations of treatment-specific outcome regressions and, for some variants, propensity-based weighting or transformation; DML and the R-learner use residualization to reduce sensitivity to nuisance estimation (Künzel et al., 2019; Chernozhukov et al., 2018; Nie & Wager, 2021). Causal forests and generalized random forests adapt local estimation to heterogeneous effects (Wager & Athey, 2018; Athey et al., 2019). Their adjustment assumptions concern the covariates passed to the estimator. Choosing a downstream estimator therefore leaves a separate upstream question of which information a temporally ambiguous export should make available.

Causal representation learning. TARNet learns a shared representation with treatment-specific outcome heads, whereas CFRNet augments this architecture with distribution-balancing regularization (Johansson et al., 2016; Shalit et al., 2017). DragonNet additionally couples outcome modeling with treatment-assignment prediction (Shi et al., 2019). TEDVAE models disentangled latent factors (Zhang et al., 2021); data-driven variable decomposition (D2VD) separates confounding and adjustment roles (Kuang et al., 2017). These approaches learn predictive or balanced representations from a supplied covariate set. Structured covariate contamination creates an earlier feature-routing decision: how much information from each feature reaches that representation.

Feature selection, masking, and bad controls. Sparsity and hard-concrete L_0 regularization learn which inputs or parameters to retain (Louizos et al., 2018). Predictive feature retention and causal admissibility solve different selection problems because eligibility is target-relative. Conditioning on descendants, colliders, or treatment-induced proxies can alter the target contrast (Angrist & Pischke, 2009; Hernán & Robins, 2020; Cinelli et al., 2024). A temporal deletion rule incorporates provenance but assigns the same status to every post-index feature. Structured covariate contamination therefore introduces a feature-level eligibility decision before downstream representation learning.

Longitudinal and proximal approaches. Marginal structural models and recurrent counterfactual methods address treatment histories and time-varying confounding (Robins et al., 2000; Lim et al., 2018; Bica et al., 2020). Proximal causal inference uses proxy and bridge assumptions to

identify effects with unmeasured confounding (Miao et al., 2018; Tchetgen Tchetgen et al., 2020; Cui et al., 2024). TEAR addresses static structured exports with column-level provenance, where preprocessing determines how temporally ambiguous covariates enter a downstream effect estimator.

3 Target-Relative Eligibility and Routing

For a declared target estimand and fixed admissible set, each feature has a fixed eligibility status. Temporal provenance supplies uncertain evidence about that status, and TEAR uses a continuous routing weight to control how much of each feature embedding reaches the downstream representation.

We observe $(X_i, T_i, Y_i)_{i=1}^n$, with $X_i \in \mathbb{R}^d$, binary treatment T_i, potential outcomes $Y_i(0), Y_i(1)$, and observed outcome $Y_i = Y_i(T_i)$. Fix the estimand and its admissible feature set $S^\star$. Write $X_{\text{elig}} = X_{S^\star}$ and $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X_{\text{elig}} = x]$. For feature j, temporal provenance is represented by timestamp t_j and schema metadata m_j, which provide evidence about eligibility.

We distinguish three quantities. The binary status $E_j = \mathbb{1}\{j \in S^\star\}$ is fixed by the target estimand. The belief p_j represents uncertainty about that status given provenance. The learned weight $h_j \in [0, 1]$ determines how much of the feature embedding enters the visible aggregate. In the controlled construction, U denotes a pretreatment baseline state, ρ its loading, and ε_p measurement noise. The delayed-neutral feature $X_{\text{delay}} = U$ remains eligible despite its post-index timestamp, whereas $X_{\text{desc}} = \rho U + \gamma T + \varepsilon_p$, with $\gamma \neq 0$, is a treatment descendant. Raw inclusion admits both, while temporal deletion removes both.

Eligibility uncertainty. For a simple decision model, let p_j be the posterior belief that feature j is eligible given (m_j, t_j), before its status is known. This is epistemic uncertainty about a fixed eligibility status. For a candidate routing mass $h \in [0, 1]$, omitting an eligible feature incurs cost $c_{\text{omit}}(1 - h)^2$, while exposing an ineligible feature incurs cost $c_{\text{exp}}h^2$. Minimizing the resulting expected loss gives

$$h_j^\star = \frac{c_{\text{omit}} p_j}{c_{\text{omit}} p_j + c_{\text{exp}}(1 - p_j)}. \tag{1}$$

Ambiguous eligibility yields an interior decision, and symmetric costs give $h_j^\star = p_j$. TEAR parameterizes h_j through the eligibility gate in Eq. (4). Further details appear in Appendix B.2.

Routing error and prediction perturbation. We isolate routing error while holding the feature maps and downstream predictor fixed. The total visible mass assigned to ineligible features and the total mass withheld from eligible features are

$$\epsilon_{\text{exp}} = \sum_{j:E_j=0} h_j, \qquad \epsilon_{\text{omit}} = \sum_{j:E_j=1} (1 - h_j).$$

These quantities measure visible ineligible mass and withheld eligible mass in the routed representation. For the same fixed fitted embeddings $e_{ij} \in \mathbb{R}^{d_e}$, where d_e is the embedding dimension, define the eligibility-masked reference and TEAR visible aggregate as

$$z_i^{\text{elig}} = \sum_j E_j e_{ij}, \qquad z_i^{\text{vis}} = \sum_j h_j e_{ij}.$$

Let $f : \mathbb{R}^{d_e} \to \mathbb{R}$ be the fixed prediction map of a fitted downstream host for the treatment-effect quantity being evaluated.

Proposition 1 (Routing-error perturbation)**.** *If $\|e_{ij}\|_2 \leq B$, then*

$$\|z_i^{\text{vis}} - z_i^{\text{elig}}\|_2 \leq B(\epsilon_{\text{exp}} + \epsilon_{\text{omit}}). \tag{2}$$

If f is L-Lipschitz on a neighborhood containing both representations, then

$$|f(z_i^{\text{vis}}) - f(z_i^{\text{elig}})| \leq LB(\epsilon_{\text{exp}} + \epsilon_{\text{omit}}). \tag{3}$$

Proposition 1 characterizes the core visible aggregate in Eq. (5). Routing-mass error bounds its representation perturbation, and Lipschitz continuity transfers this bound to the fixed fitted prediction map. The five-role extension preserves the same feature-level visible-mass budget; its routing identities are given in Appendix C. If the eligibility-masked reference also satisfies the balancing and heterogeneity-sufficiency conditions of Lemma B.1, it identifies the target CATE. The proof appears in Appendix B.3.

4 Temporal Eligibility Routing

4.1 Metadata-Guided Gated Aggregation

The core interface combines a metadata-guided eligibility gate and a finite-dimensional aggregate. Temporal provenance enters the eligibility gate and the metadata component of each feature embedding, while feature values enter the value component. Let σ denote the logistic sigmoid; $\mathrm{MLP}_{\mathrm{gate}}$ and $\mathrm{MLP}_{\mathrm{meta}}$ are multilayer perceptrons, and $v_j, b_j \in \mathbb{R}^{d_e}$ are learned feature parameters:

$$h_j = \sigma(\mathrm{MLP}_{\mathrm{gate}}([m_j, t_j])), \qquad e_{ij} = \mathrm{LayerNorm}(x_{ij} v_j + b_j + \mathrm{MLP}_{\mathrm{meta}}([m_j, t_j])), \tag{4}$$

$$z_i^{\mathrm{vis}} = \sum_{j=1}^{d} h_j e_{ij}, \qquad z_i^{\mathrm{hid}} = \sum_{j=1}^{d} (1 - h_j) e_{ij}. \tag{5}$$

Minimal TEAR exports the visible aggregate z_i^{vis} to the host. The non-exported aggregate z_i^{hid} collects the complementary routing mass. The gate uses temporal provenance $[m_j, t_j]$, and its parameters are trained jointly with the training objective. The host receives the fixed-dimensional visible aggregate as its routed input.

4.2 Training Objective

The TEAR adapter is trained with

$$\mathcal{L} = \mathcal{L}_T + \mathcal{L}_Y + \lambda_g \mathcal{L}_g + \lambda_b \mathcal{L}_b + \lambda_o \mathcal{L}_o. \tag{6}$$

Let a_T denote the auxiliary treatment head, with $\widehat{\pi}_i = \sigma(a_T(z_i^{\mathrm{vis}}))$, and let μ_t denote the factual outcome heads. $\mathcal{L}_T$ is treatment cross-entropy, and $\mathcal{L}_Y = n^{-1} \sum_i [Y_i - \mu_{T_i}(z_i^{\mathrm{vis}})]^2$. With t_j measured relative to the treatment index and $t_j < 0$ denoting pre-index timing, the gate-prior loss $\mathcal{L}_g$ is binary cross-entropy against the conservative temporal-prior labels $\bar{E}_j = \mathbb{1}(t_j < 0)$. Joint factual objectives update the gates together with the prior loss. The five-role configuration partitions routed mass into three visible subspaces and two non-exported subspaces; z^C denotes the visible subspace designated for treated–control balance. $\mathcal{L}_b$ is the radial-basis-function (RBF) kernel maximum mean discrepancy (MMD) between treated and control z^C distributions, and $\mathcal{L}_o$ penalizes cross-covariance among role-specific subspaces. The treatment head predicts assignment from the visible aggregate, and $\mathcal{L}_b$ regularizes treated–control discrepancy in z^C. Appendix C details the training losses.

Gate prewarming emphasizes $\mathcal{L}_g$ before joint optimization. Joint optimization then updates the gates using the factual, balance, and regularization objectives. Defaults are $(\lambda_g, \lambda_b, \lambda_o) = (0.5, 0.5, 0.1)$, $d_e = 32$, a $7 \to 32 \to 1$ gate MLP, and 40 epochs. Normalization, metadata encoding, and adapters are fitted on training data only. Cross-fitted hosts refit the adapter within each training fold; neural hosts fit it on the training split. The fitting sequence and configuration scope are detailed in Appendix C.2.

4.3 Extensions and Selection

Predeclared eligible baseline effect modifiers may use a direct pass-through; Appendix B gives the preservation condition. Two-role TEAR uses one visible and one non-exported path, while five-role TEAR uses the five-subspace decomposition defined above. Both preserve the total visible mass determined by the eligibility gate. Exact routing identities and host-specific exports appear in Appendix C.

Because the preferred preprocessing configuration can depend on provenance quality and host sensitivity, we define an observed-data selector over a fixed candidate pool. Let S_{risk} denote risky features

declared from timestamp and schema metadata before fitting, without oracle eligibility labels. For candidate c, let $w_j(c) \in [0,1]$ denote the effective visibility of feature j in the representation passed to the host: a deterministic inclusion indicator for fixed preprocessing rules and the learned mask or routing weight for learned methods. Its residual risky exposure is $\epsilon_{\text{risk}}(c) = \sum_{j \in S_{\text{risk}}} w_j(c)$. Let $R_{\text{fact}}(c)$ denote factual validation RMSE and $D_{\text{bal}}(c)$ treated–control RBF-MMD of the host-visible representation. The observed-data score is

$$\mathcal{J}(c) = \widetilde{R}_{\text{fact}}(c) + \alpha \widetilde{D}_{\text{bal}}(c) + \beta \widetilde{\epsilon}_{\text{risk}}(c). \tag{7}$$

Tildes denote the fixed within-run normalization across candidates. The normalization rule and weights α, β are determined on a separate development regime and frozen before evaluation; neither test PEHE nor counterfactual outcomes enter $\mathcal{J}(c)$. PEHE-based selection uses validation PEHE on the same candidate pool. Both selectors use $\{\text{Raw}, \text{Deletion}, L_0, \text{TEAR-B0}, \text{TEAR-B1}\}$ over 20 matched seeds per comparison. The fitting and evaluation protocols appear in Appendix D.

5 Experiments

5.1 Experimental Setup

Datasets and regimes. Synth-High introduces temporal contamination in a synthetic benchmark. A controlled generator separates delayed-neutral, descendant-dominant, and mixed eligibility regimes; eligibility labels are known by construction. MIMIC-P3 preserves MIMIC-IV covariates, timestamps, schema, and observed binary treatment while simulating potential outcomes (Johnson et al., 2023). MIMIC-P3-Causal generates treatment, descendant measurements, and potential outcomes structurally while preserving EHR-like covariate organization (Appendix E). Both MIMIC variants are semi-synthetic stress tests.

Hosts and candidate methods. DML, R-Learner, TARNet, and DragonNet are the primary hosts; CausalForest tests host dependence. Raw exposes all columns; Deletion removes schema-declared post-index columns. Learned L_0 masks use hard-concrete sparsity without temporal supervision. The TEAR candidate family contains two predeclared fixed operating points, B0 and B1. Dimension-matched random projection and Gate $\times$ Raw isolate finite-dimensional aggregation and coordinate-wise gating.

Evaluation protocols. Fixed comparisons evaluate named operating points. Nested candidate-pool comparisons use PEHE-based selection separately within the Base and Base+TEAR pools, then evaluate the selected candidate on the matched test split. Observed-data selection uses Eq. (7); PEHE-based selection uses validation PEHE on the same candidate pool. The provenance-corruption sweep permutes complete metadata records and retrains the gate, adapter, and host at every level.

Metrics and inference. For test index set $\mathcal{I}_{\text{te}}$, we report precision in estimation of heterogeneous effect (PEHE), $\text{PEHE} = \sqrt{|\mathcal{I}_{\text{te}}|^{-1} \sum_{i \in \mathcal{I}_{\text{te}}} (\widehat{\tau}_i - \tau_i)^2}$, where $\widehat{\tau}_i$ and τ_i are the estimated and ground-truth CATE for test unit i; lower is better. In controlled regimes with known eligibility, we additionally report area under the receiver-operating-characteristic curve (AUROC), area under the precision–recall curve (AUPRC), and expected calibration error (ECE). Fixed, nested-pool, and objective comparisons use 50 matched seeds, observed-data selection uses 20, and corruption uses 100 per level. We report mean $\pm$ standard deviation (SD) for fixed causal and selector comparisons, and paired gaps with paired 95% confidence intervals for nested pools and objective ablations.

5.2 Eligibility Disagreement Favors Feature-Specific Routing

Raw inclusion and temporal deletion favor opposite eligibility regimes, while feature-specific routing performs best in the mixed regime (Table 1A). In the delayed-neutral regime, raw inclusion reaches 0.106 and TEAR 0.112, while deletion loses usable information and reaches 0.161. In the descendant regime, deletion reaches 0.119 and TEAR 0.121, compared with raw inclusion at 0.228. Mixed eligibility favors TEAR at 0.121, below L_0 at 0.148, deletion at 0.156, and raw inclusion at 0.193. The oracle admissible-set references are 0.101, 0.102, and 0.103.

Table 1: Controlled eligibility, gate diagnostics, and representation ablation with DML (50 matched seeds). Panels A–B use the same predeclared fixed TEAR configuration across eligibility regimes; panel C varies role count at a fixed operating point. Lower PEHE is better.

A. Controlled eligibility	Test PEHE ↓				
Regime	Oracle	Raw	Deletion	L_0	TEAR
Delayed-neutral	0.101	0.106	0.161	0.143	0.112
Descendant-dominant	0.102	0.228	0.119	0.141	0.121
Mixed	0.103	0.193	0.156	0.148	0.121

B. Gate diagnostics	Mixed regime	**C. Representation ablation**	Synth-High
Feature / metric	Value	Input / representation	PEHE ↓
Baseline eligible h	0.93	Raw	23.01
Delayed-neutral h	0.82	Gate × raw	21.60
Descendant h	0.11	Deletion	18.02
AUROC	0.95	Random projection	15.24
AUPRC	0.92	Two-role TEAR	9.29
ECE	0.05	Five-role TEAR	9.12

Table 2: Nested candidate-pool selection over 50 matched seeds across MIMIC-P3 and Synth-High.

Host	**Selected test PEHE ↓**		**Paired augmentation**	
	Base	Base + TEAR	Δ	95% CI
MIMIC-P3				
DML	0.1983	0.1500	-0.0483	$[-0.0639, -0.0327]$
R-Learner	0.1974	0.1601	-0.0373	$[-0.0470, -0.0276]$
TARNet	0.2477	0.2090	-0.0387	$[-0.0535, -0.0239]$
DragonNet	0.2416	0.2005	-0.0411	$[-0.0545, -0.0277]$
Synth-High				
DML	18.020	7.850	-10.170	$[-15.14, -5.20]$
R-Learner	4.266	2.990	-1.276	$[-1.96, -0.60]$
TARNet	8.535	7.880	-0.655	$[-1.05, -0.25]$
DragonNet	8.604	7.900	-0.704	$[-1.07, -0.33]$

Note. Base = $\{\mathrm{Raw}, \mathrm{Deletion}, L_0\}$; Base + TEAR additionally includes B0 and B1. Δ = Base + TEAR − Base; negative values favor augmentation. Intervals are paired 95% CIs.

Mixed-regime gate AUROC/AUPRC/ECE are 0.95/0.92/0.05. Mean h is 0.93 for baseline, 0.82 for delayed-neutral, and 0.11 for descendant features (Table 1B). The normalized exposure and omission rates appear in Appendix D.

5.3 Adding TEAR to the Candidate Pool Improves Validation-Selected PEHE

Adding TEAR to the preprocessing candidate family lowers selected test PEHE in all eight dataset–host cells (Table 2). The base pool contains Raw, Deletion, and L_0; the augmented pool adds TEAR-B0 and TEAR-B1. Every paired 95% confidence interval excludes zero. MIMIC-P3 DML decreases from 0.1983 to 0.1500, with paired gap -0.0483 and interval $[-0.0639, -0.0327]$. Synth-High DML decreases from 18.020 to 7.850.

At fixed operating points on MIMIC-P3-Causal, TEAR has the lowest non-oracle mean PEHE for DML, R-Learner, TARNet, and DragonNet (Table 3). DML reaches 0.149 ± 0.023, compared with L_0 at 0.174 ± 0.030 and deletion at 0.187 ± 0.033. CausalForest changes little across preprocessing choices in this regime.

Table 4 compares fixed B0 and B1 with two and five roles. MIMIC-P3 DML B0 is nearly tied across role counts (0.1881 versus 0.1875), while the preferred role count varies across hosts.

5.4 Eligibility-Guided Aggregation Provides Additional Gain

Table 1C separates coordinate-wise gating, random projection, and eligibility-guided aggregation on Synth-High DML. Gate × Raw yields 21.60 PEHE versus 23.01 for Raw.

Table 3: Fixed-method test PEHE on MIMIC-P3-Causal over 50 matched seeds. TEAR uses a predeclared fixed configuration; no candidate selection is applied.

Host	Oracle	TEAR (fixed)	L_0	Deletion	Raw
DML	0.139 ± 0.018	$\mathbf{0.149} \pm 0.023$	0.174 ± 0.030	0.187 ± 0.033	0.224 ± 0.040
R-Learner	0.144 ± 0.019	$\mathbf{0.157} \pm 0.024$	0.180 ± 0.031	0.193 ± 0.034	0.226 ± 0.041
TARNet	0.162 ± 0.024	$\mathbf{0.178} \pm 0.029$	0.199 ± 0.035	0.212 ± 0.037	0.237 ± 0.044
DragonNet	0.158 ± 0.023	$\mathbf{0.175} \pm 0.028$	0.196 ± 0.034	0.208 ± 0.036	0.233 ± 0.042
CausalForest	0.171 ± 0.025	0.181 ± 0.027	$\mathbf{0.180} \pm 0.027$	0.184 ± 0.028	0.187 ± 0.030

Note. Values are mean ± SD; Oracle uses the eligible covariate set. Lower is better; bold marks the best non-oracle mean for each host.

Table 4: Fixed TEAR operating points with five or two routing roles over 50 seeds. B0 and B1 denote the two predeclared TEAR configurations; entries are mean test PEHE ± SD (lower is better). Bold marks the lowest mean in each row.

Dataset	Estimator	Five-role B0	Five-role B1	Two-role B0	Two-role B1
MIMIC-P3	DML	**0.1875** ± 0.0558	0.3386 ± 0.0620	0.1881 ± 0.0510	0.2574 ± 0.0701
	DragonNet	0.3009 ± 0.0493	**0.2122** ± 0.0536	0.3155 ± 0.0736	0.3017 ± 0.0981
	R-Learner	0.1719 ± 0.0423	0.1856 ± 0.0585	0.1714 ± 0.0380	**0.1623** ± 0.0573
	TARNet	0.2971 ± 0.0432	**0.2083** ± 0.0466	0.3141 ± 0.0702	0.2877 ± 0.0901
Synth-High	DML	9.122 ± 7.390	7.840 ± 6.251	9.292 ± 7.322	**7.776** ± 6.464
	DragonNet	8.660 ± 12.090	8.266 ± 12.024	8.730 ± 11.637	**8.177** ± 11.838
	R-Learner	4.671 ± 1.044	**3.213** ± 0.805	4.905 ± 0.979	3.270 ± 0.728
	TARNet	8.723 ± 11.796	8.330 ± 11.832	8.719 ± 11.802	**8.262** ± 12.070

Random projection improves to 15.24, showing that finite-dimensional aggregation itself contributes substantially. Eligibility-guided aggregation improves further, reaching 9.29 with two roles and 9.12 with five roles. On MIMIC-P3 DML, the corresponding role-count values are 0.1881 and 0.1875. Two-role and five-role variants have similar performance in these comparisons.

An exploratory 10-seed control fixes $h_j = 1$ while retaining balance and orthogonality losses (Appendix Table 11). TEAR-B0 reaches 0.199 versus 0.213 PEHE on MIMIC-P3 DML and 17.48 versus 27.16 on Synth-High DML. Deletion leads the DragonNet cells. This exploratory comparison probes routing against decorrelation-only adaptation.

A separate 50-seed DML diagnostic tests whether the non-exported branch drives this gain. The full-minus-no-hidden PEHE gaps are -0.0006 on MIMIC-P3 and -0.0743 on Synth-High, with $p = 0.9144$ and $p = 0.7180$, respectively (Table 14 in the appendix). The non-exported aggregate contributes little to the measured DML gain in these comparisons; the visible aggregate accounts for nearly all of it.

Removing $\mathcal{L}_g$ produces the largest objective-ablation degradation: $+3.33$ PEHE on Synth-High (paired 95% CI $[2.14, 4.52]$) and $+0.0175$ on MIMIC-P3 ($[0.0098, 0.0252]$). Removing $\mathcal{L}_Y$ has the next largest effect; $\mathcal{L}_T$, $\mathcal{L}_b$, and gate prewarming make moderate contributions. Removing $\mathcal{L}_o$ changes PEHE by $+0.07$ ($[-0.14, 0.28]$) on Synth-High and $+0.0005$ ($[-0.0009, 0.0019]$) on MIMIC-P3, with both intervals crossing zero (Table 5).

5.5 Observed-Data Selection Recovers Most of the PEHE-Based Improvement

Observed-data selection recovers 69–78% of the improvement achieved by PEHE-based selection over Deletion across the four MIMIC-P3 hosts (Table 6). Both selectors use Raw, Deletion, L_0, B0, and B1; the observed-data score combines factual validation error, treated–control balance, and residual risky exposure. DML reaches 0.206 ± 0.084, compared with deletion at 0.306 ± 0.078 and PEHE-based selection at 0.178 ± 0.052.

On MIMIC-P3-Causal, the recovered fraction is 61–74%. TEAR is selected in 12–16 of 20 seeds per host across the MIMIC benchmarks, while Raw is selected in at most one (Table 7). These counts complement the test PEHE in Table 6.

Table 5: DML objective ablation at the fixed five-role TEAR-B0 operating point (50 matched seeds per dataset). Columns remove the indicated component; paired Δ is ablation minus full TEAR, with paired 95% confidence intervals.

A. Factual and balance losses				
Statistic	Full	$-\mathcal{L}_T$	$-\mathcal{L}_Y$	$-\mathcal{L}_b$
Synth-High				
PEHE	9.12	9.86	11.31	10.02
Paired Δ	—	$+0.74$	$+2.19$	$+0.90$
95% CI	—	$[+0.31, +1.17]$	$[+1.35, +3.03]$	$[+0.39, +1.41]$
MIMIC-P3				
PEHE	0.1875	0.1930	0.2020	0.1960
Paired Δ	—	$+0.0055$	$+0.0145$	$+0.0085$
95% CI	—	$[+0.0021, +0.0089]$	$[+0.0082, +0.0208]$	$[+0.0042, +0.0128]$

B. Gate and regularization terms				
Statistic	Full	$-\mathcal{L}_g$	No prewarm	$-\mathcal{L}_o$
Synth-High				
PEHE	9.12	12.45	10.41	9.19
Paired Δ	—	$+3.33$	$+1.29$	$+0.07$
95% CI	—	$[+2.14, +4.52]$	$[+0.64, +1.94]$	$[-0.14, +0.28]$
MIMIC-P3				
PEHE	0.1875	0.2050	0.1980	0.1880
Paired Δ	—	$+0.0175$	$+0.0105$	$+0.0005$
95% CI	—	$[+0.0098, +0.0252]$	$[+0.0051, +0.0159]$	$[-0.0009, +0.0019]$

Table 6: Observed-data model selection on MIMIC-P3 and MIMIC-P3-Causal over 20 matched seeds.

Host	Deletion (D)	Observed (O)	PEHE-based (P)	Recovered
MIMIC-P3				
DML	0.306 ± 0.078	0.206 ± 0.084	0.178 ± 0.052	**78%**
R-Learner	0.286 ± 0.071	0.197 ± 0.061	0.171 ± 0.046	**77%**
TARNet	0.264 ± 0.057	0.229 ± 0.055	0.213 ± 0.041	**69%**
DragonNet	0.258 ± 0.054	0.221 ± 0.052	0.204 ± 0.039	**69%**
MIMIC-P3-Causal				
DML	0.187 ± 0.041	0.159 ± 0.038	0.149 ± 0.030	**73.7%**
R-Learner	0.193 ± 0.045	0.167 ± 0.042	0.157 ± 0.033	**72.2%**
TARNet	0.211 ± 0.050	0.191 ± 0.047	0.178 ± 0.039	**60.6%**
DragonNet	0.208 ± 0.048	0.187 ± 0.044	0.175 ± 0.036	**63.6%**

Note. Values are test PEHE, mean $\pm$ SD. Both selectors use Raw, Deletion, L_0, B0, and B1. Recovered $= (D - O)/(D - P)$.

Table 7: Observed-data selection counts over 20 matched seeds per cell, using the same five candidates as Table 6. TEAR groups B0 and B1; every row sums to 20.

Dataset	Host	TEAR	Deletion	L_0	Raw
MIMIC-P3	DML	15	2	2	1
	R-Learner	14	3	2	1
	TARNet	12	4	3	1
	DragonNet	13	3	3	1
MIMIC-P3-Causal	DML	16	2	2	0
	R-Learner	15	2	2	1
	TARNet	13	3	3	1
	DragonNet	14	3	2	1

5.6 TEAR Performance Tracks Provenance Quality

We stress-test the controlled mixed regime with DML over 100 matched seeds per corruption level. At level q, complete timestamp/schema vectors are permuted among a fraction q of columns; feature values, treatments, and outcomes stay fixed. With hyperparameters and selector weights frozen, each seed retrains the gate, adapter, and host. The pool remains Raw, Deletion, L_0, B0, and B1.

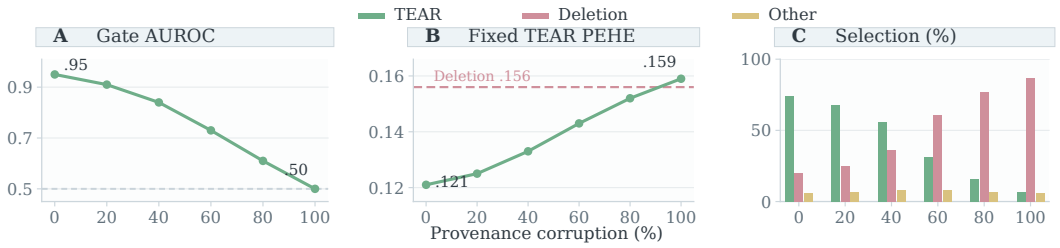

Figure 1: Controlled mixed-regime provenance-corruption stress test. Gate discrimination falls as PEHE for the predeclared fixed TEAR configuration approaches Deletion and selection shifts toward Deletion. Selection groups: TEAR $= \{\mathrm{B0}, \mathrm{B1}\}$; Other $= \{\mathrm{Raw}, L_0\}$.

As corruption increases from 0% to 100%, gate AUROC falls from 0.95 to 0.50 and fixed TEAR PEHE rises from 0.121 to 0.159, crossing deletion's 0.156 (Figure 1). TEAR selection falls from 74% to 7%, while deletion rises from 20% to 87%. Observed-selected PEHE converges to deletion at 0.156.

The validation decision shifts toward Deletion before fixed TEAR crosses it: at 60% corruption, TEAR PEHE is 0.143 versus Deletion's 0.156, yet Deletion is selected 61% of the time and TEAR 31%. Appendix Table 9 reports all levels and observed-selected PEHE.

Host and regime sensitivity. CausalForest changes little across MIMIC-P3-Causal preprocessing choices (TEAR 0.181 ± 0.027; Raw 0.187 ± 0.030). Deletion slightly outperforms TEAR in the descendant-dominant regime (0.119 versus 0.121); gains are larger when provenance distinguishes eligibility and preprocessing affects the host.

6 Discussion

The controlled regimes separate recording time from eligibility: Raw retains delayed-neutral features and descendants, Deletion removes both, and TEAR leads when they coexist. DML ablations show additional gains from eligibility-guided aggregation and the gate prior, with little difference between two and five roles. The protected-modifier diagnostic supports direct pass-through for predeclared eligible baseline modifiers (Appendix G.5).

Adding TEAR to the candidate pool improves validation-selected PEHE in all eight dataset–host cells under the same validation-PEHE selection rule, although fixed TEAR does not lead every MIMIC-P3 host. On evaluation splits, observed-data selection recovers most of the PEHE-based gain over Deletion without counterfactual validation outcomes. The Synth-High distribution appears in Appendix G.3.

Provenance corruption shifts both fixed performance and selection toward Deletion. On randomly contaminated Twins, fixed TEAR remains close to Deletion across hosts (Appendix G.4). CausalForest changes little across MIMIC-P3-Causal preprocessing choices. TEAR applies to declared estimands with feature-level provenance in static matrices, and its perturbation bound holds for fixed embeddings and a fixed Lipschitz host. The MIMIC results are semi-synthetic; clinical use requires external validation of sampling, measurement, selection, and fairness.

7 Conclusion

TEAR routes feature contributions by target-relative eligibility before CATE estimation. Across controlled and semi-synthetic settings, its gains concentrate where provenance distinguishes eligible features from descendants and host performance depends on preprocessing. Gated aggregation and validation-based candidate choice determine which covariate information reaches the CATE estimator.

ARTIFICIAL INTELLIGENCE USE STATEMENT

In this work, generative AI tools were not used for any research tasks requiring disclosure, including developing theoretical or conceptual frameworks, formulating or proving mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, generating or modifying experimental data, cleaning or reformatting datasets, performing qualitative or thematic analysis, interpreting results, or translating research materials. Survey- and interview-related tasks are not applicable to this work.

Generative AI tools were used only for manuscript-level assistance, including editing text for clarity and readability, organizing author-provided experimental results for presentation in the manuscript, and checking textual consistency such as abbreviation definitions and usage. The tools did not generate new experimental findings or scientific interpretations. The authors reviewed all AI-assisted edits and take responsibility for the final content of the work, including all text and presentation choices made with the aid of generative AI.

ETHICS STATEMENT

The MIMIC-based evaluations use deidentified EHR structure with semi-synthetic outcomes or structural generators. They assess estimation procedures in semi-synthetic EHR settings. Clinical treatment-effect claims and treatment recommendations require external clinical validation. Sampling, measurement, and fairness remain separate validation considerations.

REPRODUCIBILITY STATEMENT

Sections 3–5 and Appendices B–E document assumptions, model fitting, data construction, candidate pools, and evaluation protocols. Appendices F–G provide supporting results. Code and figure-generation scripts are available in an anonymous repository.

REFERENCES

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

Ioana Bica, Ahmed M. Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/2002.04083.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Carlos Cinelli, Andrew Forney, and Judea Pearl. A crash course in good and bad controls. *Sociological Methods & Research*, 53(3):1071–1104, 2024.

Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 119(546):1348–1359, 2024.

Miguel A Hernán and James M Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pp. 3020–3029. PMLR, 2016.

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

Kun Kuang, Peng Cui, Bo Li, Meng Jiang, Shiqiang Yang, and Fei Wang. Treatment effect estimation with data-driven variable decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. doi: 10.1609/aaai.v31i1.10480.

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.

Bryan Lim, Ahmed M Alaa, and Mihaela Van Der Schaar. Forecasting treatment responses over time using recurrent marginal structural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models, 2017. URL https://arxiv.org/abs/1705.08821.

Christos Louizos, Max Welling, and Diederik P. Kingma. Learning sparse neural networks through L_0 regularization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1Y8hhg0b.

Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

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

Judea Pearl. *Causality*. Cambridge university press, 2009.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

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.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085. PMLR, 2017.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Elizabeth A Stuart, Eva DuGoff, Michael Abrams, David Salkever, and Donald Steinwachs. Estimating causal effects in observational studies using electronic health data: challenges and (some) solutions. *Egems*, 1(3):1038, 2013.

Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Tyler J VanderWeele. Principles of confounder selection. *European journal of epidemiology*, 34(3): 211–219, 2019.

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.

Weijia Zhang, Lin Liu, and Jiuyong Li. Treatment effect estimation with disentangled latent factors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10923–10930, 2021.

A Evaluation Protocols and Result Definitions

Eligibility is membership in the oracle admissible set for a declared estimand. A post-index timestamp is one provenance attribute used to reason about eligibility. B0 and B1 denote two predeclared fixed TEAR operating points used consistently within the reported protocols. A selected-system mean records a validation decision over a declared candidate pool. Rounded scores should be interpreted within their corresponding fitting and selection protocols.

The nested comparison selects separately from $\{\mathrm{Raw}, \mathrm{Deletion}, L_0\}$ and its augmentation by B0 and B1. Fixed methods use 50 matched seeds without selection. Observed-data and PEHE-based selectors use the same five candidates over 20 matched seeds. Paired confidence intervals summarize within-seed differences, while standard deviations summarize variation in PEHE across seeds.

B Theory Details

B.1 Identification through an Eligible Representation

Fix the estimand and admissible set $S^\star$, and let $X_{\mathrm{elig}} = X_{S^\star}$. Define

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X_{\mathrm{elig}} = x], \qquad Z^\star = \phi^\star(X_{\mathrm{elig}}).$$

Assume consistency, ignorability $(Y(0), Y(1)) \perp T \mid X_{\mathrm{elig}}$, and overlap. Suppose $Z^\star$ balances the eligible covariates, $T \perp X_{\mathrm{elig}} \mid Z^\star$, and preserves heterogeneity, $\tau(X_{\mathrm{elig}}) = g^\star(Z^\star)$.

Lemma B.1 (Identification through an eligible representation)**.** *For almost every z in the induced overlap region,*

$$g^\star(z) = \mathbb{E}[Y \mid T = 1, Z^\star = z] - \mathbb{E}[Y \mid T = 0, Z^\star = z]. \tag{8}$$

Proof. For a bounded measurable function a and $s, t \in \{0, 1\}$, conditioning on X_{elig} gives

$$\begin{aligned}\mathbb{E}[a(Y(t)) \mid T = s, Z^\star = z] &= \int \mathbb{E}[a(Y(t)) \mid X_{\mathrm{elig}} = x]\, dP(x \mid T = s, Z^\star = z) \\ &= \int \mathbb{E}[a(Y(t)) \mid X_{\mathrm{elig}} = x]\, dP(x \mid Z^\star = z).\end{aligned}$$

The first equality uses ignorability and the fact that $Z^\star$ is a function of X_{elig}. The second uses balancing. The final expression is independent of s, so $Y(t) \perp T \mid Z^\star$. Consistency then gives

$$\mathbb{E}[Y \mid T = t, Z^\star = z] = \mathbb{E}[Y(t) \mid Z^\star = z].$$

Subtracting treatment arms yields $\mathbb{E}[\tau(X_{\mathrm{elig}}) \mid Z^\star = z] = g^\star(z)$ by heterogeneity sufficiency. □

B.2 Soft Routing under Eligibility Uncertainty

The binary E_j is fixed for a declared estimand. Before its status is known, the decision model represents epistemic uncertainty by a belief $p_j = \Pr(E_j = 1 \mid m_j, t_j)$. The probability represents epistemic uncertainty about the fixed eligibility status. The learned TEAR gate h_j directly parameterizes the routing decision.

If a feature is eligible, withholding mass $1 - h$ incurs omission cost $c_{\mathrm{omit}}(1 - h)^2$. If it is ineligible, exposing mass h incurs cost $c_{\mathrm{exp}}h^2$. With $c(1) = c_{\mathrm{omit}} > 0$ and $c(0) = c_{\mathrm{exp}} > 0$, the expected squared decision loss is

$$\begin{aligned}\mathcal{R}_j(h) &= \mathbb{E}_{\mathrm{belief}}[c(E_j)(h - E_j)^2 \mid m_j, t_j] \\ &= c_{\mathrm{omit}}p_j(1 - h)^2 + c_{\mathrm{exp}}(1 - p_j)h^2.\end{aligned} \tag{9}$$

Differentiating and solving $\mathcal{R}_j'(h) = 0$ gives

$$h_j^\star = \frac{c_{\mathrm{omit}}p_j}{c_{\mathrm{omit}}p_j + c_{\mathrm{exp}}(1 - p_j)}.$$

The positive quadratic coefficient makes this the unique minimum. For $0 < p_j < 1$, the solution is interior and has lower value under this stylized risk than either binary endpoint. Symmetric costs give $h_j^\star = p_j$.

B.3 Routing Error and Representation Perturbation

For fixed fitted embeddings e_{ij}, the eligibility-masked reference $z_i^{\text{elig}} = \sum_j E_j e_{ij}$ uses the true eligibility mask with exactly the same feature maps as $z_i^{\text{vis}} = \sum_j h_j e_{ij}$. It becomes target preserving when it also meets the assumptions of Lemma B.1. Proposition 1 bounds changes relative to this reference using embedding boundedness and, for prediction, a fixed Lipschitz map f.

Proof of Proposition 1. The difference consists of visible mass assigned to ineligible features and mass omitted from eligible features:

$$z_i^{\text{vis}} - z_i^{\text{elig}} = \sum_{j:E_j=0} h_j e_{ij} - \sum_{j:E_j=1} (1-h_j) e_{ij}.$$

Bounding each embedding by B and applying the triangle inequality gives

$$\|z_i^{\text{vis}} - z_i^{\text{elig}}\|_2 \le B \sum_{j:E_j=0} h_j + B \sum_{j:E_j=1} (1-h_j) = B(\epsilon_{\text{exp}} + \epsilon_{\text{omit}}).$$

Lipschitz continuity of f gives the prediction bound. □

B.4 Protected Baseline Pass-Through

Lemma B.2 (Adding predeclared eligible modifiers)**.** *Let $W = q(X_{\text{elig}})$ be fixed before outcome-based fitting. If $Z^\star$ balances X_{elig} and is heterogeneity-sufficient, then $(Z^\star, W)$ has both properties.*

Proof. Because W is a function of X_{elig} and $Z^\star$ balances X_{elig}, the conditional treatment laws satisfy

$$\begin{aligned} P(T \mid X_{\text{elig}}, Z^\star, W) &= P(T \mid X_{\text{elig}}, Z^\star) \\ &= P(T \mid Z^\star) = P(T \mid Z^\star, W). \end{aligned}$$

Since $W = q(X_{\text{elig}})$, the balancing relation also gives $T \perp W \mid Z^\star$. Heterogeneity sufficiency is unchanged because $g^\star(Z^\star)$ is also a function of $(Z^\star, W)$. □

The argument applies when W is a predeclared function of eligible baseline covariates. Temporally ambiguous features remain subject to the eligibility gate.

C Exact Routing Budget and Method Interface

Let $S_{\text{vis}} = \{C, I, Y\}$ and $S_{\text{hid}} = \{P, R\}$ denote the optional visible and non-visible sub-role sets. For a metadata-only role softmax $\widetilde{g}_j$, define

$$\begin{aligned} g_j^k &= h_j \frac{\widetilde{g}_j^k}{\sum_{q \in S_{\text{vis}}} \widetilde{g}_j^q}, && k \in S_{\text{vis}}, \\ g_j^r &= (1-h_j) \frac{\widetilde{g}_j^r}{\sum_{q \in S_{\text{hid}}} \widetilde{g}_j^q}, && r \in S_{\text{hid}}. \end{aligned}$$

The corresponding role aggregates are

$$z_i^k = \sum_j g_j^k e_{ij}, \qquad k \in S_{\text{vis}} \cup S_{\text{hid}}.$$

For hosts using the five-role visible export, write

$$z_i^{\text{role}} = [z_i^C; z_i^I; z_i^Y].$$

Because the role weights are normalized within the visible and non-visible branches,

$$\sum_{k \in S_{\text{vis}}} g_j^k = h_j, \qquad \sum_{r \in S_{\text{hid}}} g_j^r = 1 - h_j.$$

Role logits redistribute mass within a branch but cannot change the total visible mass assigned by h_j.

For a declared risky set S_{risk}, the residual visible exposure is therefore exactly

$$\epsilon_{\text{risk}} = \sum_{j \in S_{\text{risk}}} \sum_{k \in S_{\text{vis}}} g_j^k = \sum_{j \in S_{\text{risk}}} h_j.$$

Role names denote predictive parameterizations within the routing architecture.

C.1 Training Objective and Default Adapter

The metadata-only gate is

$$h_j = \sigma(\text{MLP}_{\text{gate}}([m_j, t_j])),$$
$$\text{MLP}_{\text{gate}} : \mathbb{R}^7 \to \mathbb{R}^{32} \to \mathbb{R}.$$

Feature values do not enter this gate input. Its parameters are trained jointly with the training objective. Values enter only their embeddings,

$$u_{ij} = x_{ij} v_j + b_j + \text{MLP}_{\text{meta}}([m_j, t_j]),$$
$$e_{ij} = \text{LayerNorm}(u_{ij}).$$

The minimal export passes only the finite-dimensional visible aggregate $z_i^{\text{vis}} = \sum_j h_j e_{ij}$ to the host. The complementary aggregate records non-exported routing mass. This interface differs from coordinate-wise scalar rescaling because the downstream host receives the aggregate rather than every original coordinate.

The losses have distinct roles. $\mathcal{L}_T$ retains treatment-assignment information in the visible representation, and $\mathcal{L}_Y$ preserves factual outcome information. $\mathcal{L}_g$ anchors the gate to the conservative temporal prior. $\mathcal{L}_b$ balances the designated z^C subspace, while $\mathcal{L}_o$ regularizes redundancy among optional subspaces. The adapter minimizes

$$\mathcal{L} = \mathcal{L}_T + \mathcal{L}_Y + \lambda_g \mathcal{L}_g + \lambda_b \mathcal{L}_b + \lambda_o \mathcal{L}_o,$$

where $\mathcal{L}_T$ and $\mathcal{L}_Y$ are auxiliary factual treatment and outcome losses, $\mathcal{L}_g$ is binary cross-entropy against the conservative temporal prior $\bar{E}_j$, $\mathcal{L}_b$ is RBF-MMD between treated and control z^C representations, and $\mathcal{L}_o$ penalizes cross-covariance among optional subspaces. Only $\mathcal{L}_g$ uses the temporal prior label; all differentiable terms can update the gate. Defaults are $(\lambda_g, \lambda_b, \lambda_o) = (0.5, 0.5, 0.1)$, $d_e = 32$, and 40 epochs.

The factual and gate-prior losses are

$$\widehat{\pi}_i = \sigma(a_T(z_i^{\text{vis}})),$$
$$\mathcal{L}_T = -\frac{1}{n} \sum_i \{T_i \log \widehat{\pi}_i + (1 - T_i) \log(1 - \widehat{\pi}_i)\},$$
$$\mathcal{L}_Y = \frac{1}{n} \sum_i \{Y_i - \mu_{T_i}(z_i^{\text{vis}})\}^2,$$
$$\mathcal{L}_g = -\frac{1}{d} \sum_j \{\bar{E}_j \log h_j + (1 - \bar{E}_j) \log(1 - h_j)\}.$$

The role-structured balance term compares the treated and control z^C distributions with an RBF kernel, while the assignment head takes the visible representation. Balancing a designated subspace and retaining assignment information in the visible representation are different optimization targets. The cross-covariance term $\mathcal{L}_o$ regularizes optional subspaces without identifying causal roles. The minimal two-role aggregate omits role-orthogonality regularization.

C.2 Fitting Sequence and Split Scope

For each declared training unit (a full training split for a neural host or one training fold for a cross-fitted host), the fitting sequence is:

Table 8: Evaluation protocols, fitting scope, and seed counts.

Comparison	Selection and fitting	Seeds
Controlled regimes / gate recovery	DML, fixed operating points; no selection.	50
Fixed methods / ablations	Named configuration; no candidate selection.	50
Nested candidate pools	Validation PEHE selects separately from Base and Base + B0/B1.	50
Observed-data selector	Observed score and validation-PEHE benchmark use Raw, Deletion, L_0, B0, B1.	20
Provenance corruption	DML, mixed regime; retrain gate, adapter, and host at every level and seed. Same five candidates.	100/level
Exploratory controls	Decorrelation-only diagnostic (P4), fixed methods.	10

1. fix the treatment index, schema metadata, risky-feature declaration, candidate pool, and any protected eligible pass-through variables;
2. estimate normalization statistics on that training unit only;
3. initialize a new metadata encoder, eligibility gate, feature embeddings, optional role router, and auxiliary heads;
4. prewarm the gate with emphasis on $\mathcal{L}_g$;
5. jointly optimize the full factual objective for the declared epochs;
6. export only the visible representation, optionally concatenated with the predeclared eligible pass-through block;
7. fit the downstream host on the corresponding training representation and apply the frozen adapter to validation or held-out observations;
8. select a candidate on validation information only, then evaluate that frozen choice once on the matched test split.

For DML and the R-learner, a separate adapter is trained inside every cross-fit fold and applied to that fold's held-out observations. The representation for each held-out observation is learned entirely from its corresponding training fold. TARNet and DragonNet fit the adapter on the training split and receive concatenated visible subspaces. CausalForest uses a pre-specified visible export fit on its training split or fold. The perturbation bound applies directly to the minimal/two-role export; a host-specific five-role export requires the same balancing and heterogeneity-sufficiency conditions.

D Experimental Protocol Registry

The experimental unit is a matched seed within a dataset–host cell. Fixed operating points and selected systems answer different questions and are evaluated separately.

Candidate pools and evaluation. Base is $\{\mathrm{Raw}, \mathrm{Deletion}, L_0\}$; Base + TEAR adds TEAR-B0 and TEAR-B1. Validation PEHE selects once within each pool, followed by one evaluation on the matched test split. Both selectors in the multi-host observed-data comparison use the augmented five-candidate pool. PEHE-based selection is a counterfactual-assisted benchmark. Protocol labels P1 (fixed), P2 (nested), P3 (observed-data), and P4 (exploratory) in supporting tables denote these evaluation types; P3 is unrelated to the dataset suffix.

Provenance permutation and refitting. For each $q \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$, randomly select a proportion q of feature-level records, then randomly permute complete timestamp/schema metadata vectors among the selected columns. The multiset of metadata records and X, T, Y remain unchanged. The association of metadata with feature eligibility is progressively disrupted. The controlled mixed-eligibility data-generating process (DGP) uses DML and 100 matched seeds per level. Hyperparameters are frozen before the sweep; selector weights are frozen on an independent development regime. For each level and seed, the gate, adapter, and downstream host are retrained, including the fold-specific fitting in Appendix C.2. The pool is Raw, Deletion, L_0, B0, B1. TEAR selection groups B0/B1; Other groups Raw/L_0.

Statistical summaries. Nested-pool gaps and objective-ablation changes are paired by seed and accompanied by paired 95% confidence intervals. Fixed MIMIC-P3-Causal and observed-data comparisons report mean ± SD over 50 and 20 matched seeds, respectively. Gate recovery and controlled-regime means use 50 matched seeds.

D.1 Metrics and Their Roles

For test set $\mathcal{I}_{\mathrm{te}}$,

$$\mathrm{PEHE} = \sqrt{\frac{1}{|\mathcal{I}_{\mathrm{te}}|} \sum_{i \in \mathcal{I}_{\mathrm{te}}} (\widehat{\tau}_i - \tau_i)^2},$$

and absolute CATE bias is

$$\left| \frac{1}{|\mathcal{I}_{\mathrm{te}}|} \sum_{i \in \mathcal{I}_{\mathrm{te}}} (\widehat{\tau}_i - \tau_i) \right|.$$

Mean PEHE is the primary effect-estimation metric. Median and upper-tail quantiles show distributional behavior across seeds. Gate AUROC/AUPRC and calibration error are used only where feature eligibility is known by construction. Residual risky exposure audits routed mass on the declared risky set, while PEHE measures treatment-effect estimation error. Factual root mean squared error (RMSE) and treated–control balance serve as observed-data selection signals.

Supporting paired gaps and win rates compare matched test seeds. Statistical significance is interpreted with effect size and seed count; standard deviations and confidence intervals represent different quantities.

Normalized exposure and omission. For $d_0 = |\{j : E_j = 0\}|$ ineligible and $d_1 = |\{j : E_j = 1\}|$ eligible columns, the normalized diagnostics are $r_{\mathrm{exp}} = \epsilon_{\mathrm{exp}}/d_0$ and $r_{\mathrm{omit}} = \epsilon_{\mathrm{omit}}/d_1$ when the respective denominators are positive. They summarize soft routing mass per column. The unnormalized total entering Proposition 1 is $d_0 r_{\mathrm{exp}} + d_1 r_{\mathrm{omit}}$. The controlled mixed regime has $r_{\mathrm{exp}} = 0.09$ and $r_{\mathrm{omit}} = 0.13$.

D.2 Observed-Data Selector Audit

For candidate c, let $w_j(c)$ denote the effective visibility of feature j in the host input: an inclusion indicator for fixed rules and the learned mask or routing weight for learned methods. With the provenance-declared risky set S_{risk}, define $\epsilon_{\mathrm{risk}}(c) = \sum_{j \in S_{\mathrm{risk}}} w_j(c)$. The selector minimizes

$$\mathcal{J}(c) = \widetilde{R}_{\mathrm{fact}}(c) + \alpha \widetilde{D}_{\mathrm{bal}}(c) + \beta \widetilde{\epsilon}_{\mathrm{risk}}(c),$$

Here R_{fact} is factual validation RMSE and D_{bal} is treated–control RBF-MMD of the host-visible representation on the validation split. The risky set is declared from timestamp and schema metadata before fitting, without oracle eligibility labels. Tildes denote the fixed within-run normalization across candidates; its rule and α, β are fixed on a separate development regime before evaluation and held constant across reported test comparisons. Neither test PEHE nor counterfactual outcomes enter this score. The PEHE-based selector is a counterfactual-assisted benchmark. The recovered fraction is $(\mathrm{Deletion} - \mathrm{Observed})/(\mathrm{Deletion} - \mathrm{PEHE\text{-}selected})$, evaluated within each dataset–host cell.

E Data-Generating Constructions and Scope

E.1 Exact Temporal-Mismatch Example

Let U be a baseline state determining treatment and potential outcomes. Record $X_{\mathrm{delay}} = U$ after treatment assignment and define $X_{\mathrm{desc}} = \rho U + \gamma T + \varepsilon_p$, $\gamma \neq 0$. The late record equals the baseline state and has no structural dependence on T, so it is treatment-neutral; the second feature is a descendant. Raw preprocessing admits both, deletion removes both, and the oracle retains only X_{delay}. Noisy delayed records of pretreatment measurements are used only in controlled experiments, where approximation rather than exact identification is evaluated.

Table 9: Provenance permutation in the controlled mixed-eligibility DGP with DML: 100 matched seeds per level. The gate, adapter, and host are retrained at every level and seed, with hyperparameters and selector weights frozen. TEAR groups B0/B1 and Other groups Raw/L_0. Higher AUROC and lower PEHE are better.

	Gate	PEHE			Selection frequency (%)		
Corruption	AUROC	Fixed TEAR	Deletion	Observed	TEAR	Deletion	Other
0%	0.95	0.121	0.156	0.124	74	20	6
20%	0.91	0.125	0.156	0.128	68	25	7
40%	0.84	0.133	0.156	0.136	56	36	8
60%	0.73	0.143	0.156	0.146	31	61	8
80%	0.61	0.152	0.156	0.152	16	77	7
100%	0.50	0.159	0.156	0.156	7	87	6

Table 10: Controlled eligibility generator, fixed comparison with DML (50 matched seeds). Each cell reports PEHE / absolute CATE bias; lower is better. Oracle uses the admissible set.

Method	Delayed-neutral	Descendant	Mixed
Oracle	0.101 / 0.006	0.102 / 0.006	0.103 / 0.006
Raw	0.106 / 0.010	0.228 / 0.119	0.193 / 0.083
Deletion	0.161 / 0.014	0.119 / 0.010	0.156 / 0.014
L_0	0.143 / 0.013	0.141 / 0.024	0.148 / 0.022
TEAR	0.112 / 0.009	0.121 / 0.012	0.121 / 0.011

The delayed-neutral, descendant-dominant, and mixed regimes vary which error matters. Raw inclusion is expected to be competitive at the delayed-neutral endpoint, deletion at the descendant endpoint, and continuous routing in the mixture. This design evaluates the predicted ordering across eligibility regimes.

E.2 MIMIC-P3 Construction and Intended Use

MIMIC-P3 uses 10,000 admissions and 25 covariates: 15 baseline or pre-index variables and 10 post-index variables. It preserves observed binary treatment assignments and the MIMIC-IV covariate, timestamp, and schema structure while simulating potential outcomes so that PEHE is observable. MIMIC-P3-Causal instead generates

$$T \sim \text{Bernoulli}(\pi(X_{\text{pre}}, U)), \qquad 0 < \pi(X_{\text{pre}}, U) < 1,$$

with delayed records of pretreatment measurements X_{delay}, descendant potential measurements $X_{\text{desc}}(t)$, and

$$Y(t) = f_Y(X_{\text{pre}}, U) + t\tau(X_{\text{pre}}) + \varepsilon_Y.$$

Treatment, descendant measurements, and potential outcomes are generated structurally while preserving EHR-like covariate organization. The role of U and the eligible covariates must be interpreted under the declared generator; the identification statement in Section 3 separately assumes ignorability for the oracle eligible set.

Both datasets are semi-synthetic stress tests of estimation behavior. The controlled generator provides known eligibility labels, while MIMIC-P3 preserves EHR-like covariate and provenance structure. Clinical deployment would require external validation of treatment effects, sampling and measurement processes, fairness, and selector behavior.

F Additional Experimental Results

Table 9 lists every provenance-corruption level, including fixed and observed-selected performance. Table 10 gives PEHE and absolute CATE bias for the controlled eligibility regimes.

G Expanded Experimental Evidence

G.1 Fixed Operating Points and Architectural Ablations

Table 4 compares the five-role refinement with the two-role routing interface. B0 values are close for five-role and two-role routing on Synth-High DML (9.12 and 9.29) and MIMIC-P3 DML (0.1875 and 0.1881). The visible/non-exported routing interface captures nearly all of the measured DML effect in these comparisons.

The 10-seed P4 decorrelation-only block in Table 11 checks a different alternative explanation. Fixing $h_j = 1$ preserves the balance and orthogonality losses but removes eligibility routing. TEAR-B0 is lower than this ablation in the four reported cells, although deletion is lower for the two DragonNet cells. This P4 block is reported as an exploratory diagnostic.

Table 11: Exploratory decorrelation-only ablation over 10 seeds (P4). *Decorr-only* fixes all gates at $h_j = 1$ and retains the balance and orthogonality losses. Lower PEHE is better. This 10-seed exploratory block isolates the contribution of decorrelation from eligibility routing.

Dataset	Estimator	Method	Mean PEHE	SD
MIMIC-P3	DML	TEAR-B0 (routing + gate)	**0.199**	0.038
MIMIC-P3	DML	Decorr-only (h_j=1, no routing)	0.213	0.052
MIMIC-P3	DML	Deletion	0.308	0.078
MIMIC-P3	DragonNet	Deletion	**0.213**	0.037
MIMIC-P3	DragonNet	TEAR-B0 (routing + gate)	0.296	0.040
MIMIC-P3	DragonNet	Decorr-only (h_j=1, no routing)	0.305	0.043
Synth-High	DML	TEAR-B0 (routing + gate)	**17.48**	10.56
Synth-High	DML	Decorr-only (h_j=1, no routing)	27.16	23.63
Synth-High	DML	Deletion	28.58	19.63
Synth-High	DragonNet	Deletion	**12.63**	11.00
Synth-High	DragonNet	TEAR-B0 (routing + gate)	14.33	11.60
Synth-High	DragonNet	Decorr-only (h_j=1, no routing)	14.53	11.64

G.2 Local Sensitivity and Optimization Diagnostics

Table 12 reports the complete one-at-a-time grids behind the local-sensitivity statement. The sensitivity summary of a 0.0048 range for a twentyfold λ_g change on MIMIC-P3 DML is the difference between 0.1906 and 0.1858 in this table. This sweep measures local one-at-a-time stability.

Table 13 reports gate-gradient magnitudes for the prior and factual loss components at the fitted solution. Gradient norms depend on loss scaling and achieved parameter values.

Table 14 tests the non-exported branch separately. The non-exported-branch diagnostic shows small mean gaps across the reported cells.

G.3 Tail Behavior and External Method Controls

The separate 50-seed Synth-High side-by-side comparison reports the full PEHE distribution (Table 15). TEAR-B0 has the lowest median and P95 among the reported methods, reducing P95 from Deletion's 61.09 to 23.77. TEDVAE has the lower maximum, 36.75 versus TEAR-B0's 40.11. This fixed comparison uses a different protocol from the nested-pool selected mean of 7.85.

G.4 Weak-Provenance Boundary

Twins is a treatment-effect benchmark derived from U.S. twin births, with treatment defined by heavier-versus-lighter birth weight and first-year mortality as the outcome (Louizos et al., 2017). We use it as an additional weak-provenance boundary by randomly contaminating the provenance information supplied to the routing mechanism. Table 16 supplies the full four-host values behind the contaminated-Twins boundary statement. Under randomly contaminated Twins, TEAR remains close to deletion across hosts while retaining a larger advantage over the timestamp rule. This protocol represents a weak-provenance boundary for the routing mechanism.

Table 12: One-at-a-time hyperparameter sensitivity over 50 seeds. For each parameter, the table reports the tested values, best and worst mean PEHE, their range, and the best tested value. Defaults are $\lambda_g = 0.5$, $\lambda_b = 0.5$, $\lambda_o = 0.1$, and prewarm fraction 0.1. Lower PEHE and smaller ranges indicate greater local stability.

Dataset	Estimator	Sweep parameter	Values tested	Best PEHE	Worst PEHE	Range	Best value
MIMIC-P3	DML	λ_b	0, 0.1, 0.5, 1	0.1783	0.1985	0.0202	1
		λ_g	0.1, 0.25, 0.5, 1, 2	0.1858	0.1906	0.0048	0.1
		λ_o	0, 0.05, 0.1, 0.5	0.1844	0.1886	0.0042	0.5
		prewarm fraction	0, 0.05, 0.1, 0.2	0.1872	0.1889	0.0017	0.05
	DragonNet	λ_b	0, 0.1, 0.5, 1	0.3009	0.3269	0.0260	0.5
		λ_g	0.1, 0.25, 0.5, 1, 2	0.3002	0.3055	0.0053	2
		λ_o	0, 0.05, 0.1, 0.5	0.2984	0.3019	0.0036	0.5
		prewarm fraction	0, 0.05, 0.1, 0.2	0.3003	0.3019	0.0016	0.2
Synth-High	DML	λ_b	0, 0.1, 0.5, 1	8.9775	9.2945	0.3171	1
		λ_g	0.1, 0.25, 0.5, 1, 2	9.0690	9.1322	0.0632	2
		λ_o	0, 0.05, 0.1, 0.5	9.0690	9.2299	0.1609	0
		prewarm fraction	0, 0.05, 0.1, 0.2	8.9747	9.2618	0.2871	0.2
	DragonNet	λ_b	0, 0.1, 0.5, 1	8.5190	8.8034	0.2844	1
		λ_g	0.1, 0.25, 0.5, 1, 2	8.6200	8.6474	0.0278	2
		λ_o	0, 0.05, 0.1, 0.5	8.5863	8.7777	0.1913	0
		prewarm fraction	0, 0.05, 0.1, 0.2	8.5119	8.6426	0.1307	0.2

On MIMIC-P3, λ_g, λ_o, and the prewarm fraction have ranges at most 0.0053, while λ_b has the largest range (0.0202 for DML and 0.0260 for DragonNet). On Synth-High, ranges are larger in the scale of that benchmark: up to 0.3171 for DML and 0.2844 for DragonNet. The default settings lie inside the tested grids. These one-at-a-time sweeps characterize local sensitivity; joint-setting robustness requires a multivariate sweep.

Table 13: Gate-gradient diagnostic over 50 seeds at the fitted solution. Values are per-feature ℓ_2 gradient norms at the post-index gate coordinates, together with the mean post-index gate value and residual visible mass. In these schemas the post-index set is the declared risky set, so $\epsilon_{\text{risk}} = \epsilon_{\text{post}}$.

Dataset	$\overline{h}_{\text{post}}$	ϵ_{risk}	$\|\nabla\mathcal{L}_g\|_2$	$\|\nabla\mathcal{L}_T\|_2$	$\|\nabla\mathcal{L}_Y\|_2$	$\|\nabla\mathcal{L}\|_2$
MIMIC-P3	0.387	3.868	0.1039	0.0062	0.0692	0.0801
Synth-High	0.463	16.665	0.0576	0.0391	1.2333	1.2286

On MIMIC-P3, the gate-prior gradient ($\|\nabla\mathcal{L}_g\| = 0.1039$) exceeds the total gradient ($\|\nabla\mathcal{L}\| = 0.0801$). The smaller total is possible because component gradients need not align. The reported values show the relative local gradient magnitudes of the listed loss components at the fitted solution; residual risky mass is 3.87.
On Synth-High, the outcome-loss gradient ($\|\nabla\mathcal{L}_Y\| = 1.2333$) is larger than the gate-prior gradient ($\|\nabla\mathcal{L}_g\| = 0.0576$) in this fitted configuration. Residual risky mass is reported on each dataset's schema scale; the Synth-High and MIMIC-P3 totals therefore use different normalizations.
The relative gradient magnitudes depend on dataset properties, loss scaling, and the achieved gate values.

Table 14: Non-exported-branch diagnostic over 50 seeds with the DML host. Gap = full TEAR − ablation; negative values favor full TEAR.

Dataset	Comparison	Gap (full − ablation)	Win Rate	p-value	Cohen's d
MIMIC-P3	Full vs. No-hidden	−0.0006	0.50	0.9144	−0.015
	Full vs. Gate-only	−0.0005	0.46	0.9515	−0.009
Synth-High	Full vs. No-hidden	−0.0743	0.60	0.7180	−0.051
	Full vs. Gate-only	−0.1692	0.72	0.4442	−0.109

Removing the non-exported branch produces small mean changes in this diagnostic. The reported gaps and uncertainty indicate little measurable contribution from that branch in these comparisons. This exploratory diagnostic is separate from the objective-removal block.

Table 15: Synth-High DML distribution over 50 fixed-method seeds. P90 and P95 denote the 90th and 95th percentiles.

Method	Mean PEHE	Median PEHE	P90	P95	Max
TEAR-B0	**9.23**	**6.15**	**18.64**	**23.77**	40.11
TEDVAE variant	10.20	7.66	22.42	26.65	**36.75**
D2VD-nonlinear	17.71	9.50	38.88	55.59	135.39
D2VD-linear	17.71	8.34	38.62	57.51	128.77
Deletion	18.02	8.66	37.89	61.09	127.29
Raw inclusion	23.01	12.05	49.54	70.12	142.22

Table 16: Randomly contaminated Twins weak-provenance boundary. The best fixed TEAR-family operating point remains within -0.1% to -2.2% of deletion across four hosts and is lower than the timestamp rule in every cell.

Estimator	Deletion	Timestamp rule	Best TEAR	TEAR vs deletion	TEAR vs rule	p (vs rule)
DML	0.1631	0.1669	0.1629	-0.1%	-2.4%	2.64×10^{-11}
R-Learner	0.1630	0.1711	0.1624	-0.4%	-5.1%	8.83×10^{-3}
TARNet	0.1662	0.1945	0.1660	-0.1%	-14.7%	5.50×10^{-6}
DragonNet	0.1663	0.1887	0.1627	-2.2%	-13.8%	6.31×10^{-8}

G.5 Additional Diagnostics

The separate frozen-hyperparameter metadata-shift test changes metadata between domains, while Figure 1 permutes feature-level provenance within the controlled generator. TEAR PEHE changes from 0.136 in-domain to 0.149 under shift, a 9.6% degradation. The corresponding degradations are 19.9% for L_0, 13.0% for Raw, and 3.8% for Deletion.

The protected-modifier experiment with $\tau(\mathrm{age}) = 1 + 0.8\,\mathrm{age}$ reports recovered slope 0.79 ± 0.08 and $R^2 = 0.82$ for pass-through TEAR, versus 0.70 ± 0.12 for compressed TEAR, 0.72 ± 0.11 for deletion, and 0.91 ± 0.18 for raw inclusion.

In the Synth-High R-learner diagnostic, raw and deletion PEHE are 3.17 and 7.88. Relative to deletion, TEAR changes normalized held-out outcome-nuisance RMSE from 1.00 to 0.86 and the P99 pseudo-outcome magnitude from 1.00 to 0.72, with overlap reported as essentially unchanged. These diagnostics localize sensitivity to upstream preprocessing without identifying the cause of every host difference.

The runtime evaluation reports 1.4 seconds per Synth-High adapter fit and 4.6 seconds per MIMIC-P3 fit on an A100 80 GB GPU; five-fold MIMIC fitting takes 24.8 seconds. The adapter accounts for 14.2% of the DML pipeline and 8.7% of DragonNet, with 2.1 GB peak memory.