CAUSABRIDGE: AN IDENTIFICATION-GATED CAUSAL FOUNDATION MODEL FOR TRANSFERABLE EFFECT ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation models amortize inference across tasks, but causal effect estimation has a prerequisite absent from ordinary prediction: the requested effect must be identifiable under the declared causal model and available distributions before a numerical estimator is invoked. We introduce *CausaBridge*, an identification-gated causal foundation model. A typed task contract is compiled into a candidate identification certificate and independently verified; only an accepted point certificate can execute the contract-conditioned estimator. Verified bounds bypass the neural path, while non-identified, unsupported, or empirically unsupported queries return typed non-point results. We establish point-output soundness, noninterference of non-point branches, stability of registered functionals to nuisance error, conditional invariance under kernel-preserving adapters, and a task-distribution generalization bound. A prespecified common-target study evaluates 18 checkpoints on 256 unseen tasks (4,608 evaluations); diverse-mechanism pretraining reduces calibrated ATE error by 0.0180 relative to fixed-mechanism pretraining (95% crossed-bootstrap CI: 0.0117–0.0242; $p < 10^{-4}$) and satisfies a four-part decision rule specified before confirmatory evaluation. A 5,376-job comparison study and a matched 312-job selective evaluation on IHDP, ACIC, Twins, and Jobs produce 309 authorized point results and three typed refusals. On supported ACIC repetitions, CausaBridge reduces ATE error relative to CausalPFN by 0.0741 (95% repetition-bootstrap CI: 0.0419–0.1112). On nine authorized MIMIC-based semi-synthetic scenarios, CausaBridge obtains 0.1353 ATE MAE and 0.1756 root precision in estimation of heterogeneous effect (PEHE) under a prespecified multi-seed evaluation protocol; adapter-only transfer updates 5.8% of parameters and reduces both errors relative to full fine-tuning under the prespecified protocol. In five-seed structural replay, the complete executor produces no false point outputs and makes no neural calls on non-point branches. A linked real-cohort audit separately converts inadequate treatment support into a typed non-point outcome. CausaBridge thereby unifies exact-query authorization, typed non-point execution, and reusable causal effect estimation in a single verified pipeline.

1 INTRODUCTION

Foundation models replace one model per dataset with a reusable model that is pretrained across a broad task distribution and adapted at test time (Bommasani et al., 2021). That prospect is particularly attractive for causal inference: fitting and validating a separate estimator for every cohort, treatment, outcome, and modality is expensive, while structural causal models (SCMs) can generate an effectively unbounded pretraining curriculum. Recent prior-data fitted networks consequently amortize treatment-effect inference across simulated tasks (Zhang et al., 2024; Ma et al., 2026; Balazadeh et al., 2025; Robertson et al., 2025). Yet causal reuse raises a prerequisite absent from ordinary prediction. Before estimating an effect, one must establish that the requested interventional quantity is a functional of the distributions and assumptions made available to the system.

Identification and estimation are distinct. Graphical theory decides whether a causal query can be recovered under a stated graph and set of observational or experimental interfaces (Pearl, 2009; Shpitser & Pearl, 2008). Estimation approximates the resulting functional from finite samples. A

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.

universal approximator does not preserve this distinction: it can map every dataset to a number, including datasets for which no permitted functional exists. Balancing, latent-variable models, targeted regularization, and orthogonal nuisance estimation improve statistical behavior under their respective assumptions (Shalit et al., 2017; Louizos et al., 2017; Shi et al., 2019; Chernozhukov et al., 2018). CausaBridge addresses the orthogonal systems question of whether the current query is authorized under the declared contract before a reusable estimator is invoked.

The separation is acute in multimodal medicine. Electronic health records contain treatment histories, irregular measurements, missingness, and site artifacts; images provide high-dimensional evidence but may be acquired after treatment or selected by clinical workflow. MIMIC-IV and MIMIC-CXR provide a concrete test bed for linked tabular and radiographic evidence (Johnson et al., 2023; 2019). An image can improve an outcome model while being inadmissible for a particular causal query. Adaptation therefore needs both a representation interface and a structural authorization interface.

We present CausaBridge, which couples an identification control plane to an amortized effect estimator. A causal contract names the query, graph, temporal roles, domains, and available distributions. A compiler proposes a certificate; an independent verifier rechecks its functional and side conditions without calling the compiler. Only a verified point certificate authorizes the contract-conditioned foundation core. The core is pretrained across SCM task families and uses a cross-fitted *direct* task-effect head that learns task effects without imposing a fixed treated–control anchor. Typed adapters connect structured records and images without changing the gate. Bounds and refusals terminate before adapter materialization.

Our contributions are fourfold:

- We define a fail-closed interface between identification and amortized estimation: a point result requires an independently verified certificate for the exact query, contract, and runtime interfaces; bounds and refusals are first-class outputs.
- We introduce a contract-conditioned foundation estimator with supervision-separated SCM pretraining, cross-fitted direct effect prediction, and plug-in clinical adapters. The architecture supports new observation types without giving adapters authority to change identification status.
- We provide a conditional theory connecting software soundness to statistical transfer: gate soundness, compositional plug-in stability, adapter kernel preservation, and a uniform task-family generalization bound.
- We evaluate authorization and estimation accuracy as distinct axes. A prespecified common-target protocol evaluates mechanism-diversity transfer; a 5,376-job comparison study and a matched 312-job selective evaluation characterize public-benchmark risk and coverage; and a prespecified multi-seed MIMIC study tests plug-in multimodal interfaces and parameter-efficient reuse while a linked clinical cohort validates fail-closed empirical-support gating.

The scope of each claim is explicit: certificate validity is defined relative to the analyst-declared contract, and foundation transfer is evaluated over a prespecified task distribution. We align theorem-backed guarantees, executable safety tests, and empirical transfer evidence under a single auditable protocol.

2 Related Work

Identification and causal queries. Structural causal models distinguish associational distributions from interventional and counterfactual quantities. The ID family gives complete graphical criteria for deciding when interventional distributions can be obtained from observational information (Pearl, 2009; Shpitser & Pearl, 2008). CausaBridge uses this distinction as an explicit software boundary: the output of identification is not a prediction target for a second neural head, but an authorization object that constrains the estimator's execution.

Neural treatment-effect estimation. TARNet/CFR learn shared representations and treatment-specific heads under strong ignorability (Shalit et al., 2017); CEVAE introduces latent-variable

structure for proxy-assisted estimation (Louizos et al., 2017); Dragonnet adapts neural nuisance models using treatment-effect-specific regularization (Shi et al., 2019); and metalearners and doubly/debiased machine learning provide complementary estimator constructions and orthogonalization strategies (Kuenzel et al., 2019; Chernozhukov et al., 2018). These methods address estimation after an identification and estimand choice have been made. CausaBridge is complementary: it conditions estimator invocation on an explicit identification contract and adds typed non-point outcomes for queries outside the authorized branch.

Amortized causal inference. Prior-data fitted networks amortize Bayesian inference over a synthetic task prior (Müller et al., 2022). CInA relates optimal balancing to attention and performs zero-shot estimation across datasets (Zhang et al., 2024). CausalFM develops valid SCM priors and instantiates back-door, front-door, and instrumental-variable settings (Ma et al., 2026); CausalPFN and Do-PFN likewise amortize effect inference across simulated mechanisms (Balazadeh et al., 2025; Robertson et al., 2025). Building on this prior validity theory, CausaBridge contributes an orthogonal runtime mechanism that binds the current query to an explicit contract, independently verifies authorization, and returns typed bounds or refusals whenever point execution is not licensed.

Selective prediction and safe execution. Selective classifiers trade coverage for predictive risk by learning or calibrating a reject rule (Geifman & El-Yaniv, 2017). Our refusal is structurally different: it is a deterministic consequence of the declared graph and interfaces, not a confidence threshold learned from observed values. Statistical uncertainty remains attached to authorized point estimates and cannot override a failed identification check.

3 Problem Setting

Let X denote observed multimodal covariates, T a treatment, and Y an outcome. A causal query q specifies an estimand, treatment and outcome variables, target domain, and required assumptions. A causal contract κ additionally declares the graph, variable domains, available interfaces, and provenance of the observational batch $\mathcal{D}_{\text{obs}}$.

For a binary treatment, the point estimands considered by the model are

$$\tau_{\text{ATE}} = \mathbb{E}[Y(1) - Y(0)], \qquad \tau_{\text{CATE}}(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]. \tag{1}$$

Identification asks whether the target distribution is a functional of the distribution and assumptions available to the system. Estimation asks how accurately a statistical procedure approximates that functional from finite data. We maintain the two questions separately.

Result contract. Every execution returns one member of a typed result union:

$$R(q, \kappa, \mathcal{D}_{\text{obs}}) \in \{R_{\textsf{POINT}}, R_{\textsf{BOUND}}, R_{\textsf{REFUSED}}, R_{\textsf{UNSUPPORTED}}\}. \tag{2}$$

$R_{\textsf{POINT}}$ contains a point effect and uncertainty generated by the neural estimator. $R_{\textsf{BOUND}}$ contains only a verified symbolic interval from a registered bound family. $R_{\textsf{REFUSED}}$ records a non-identifiable query or failed verification. $R_{\textsf{UNSUPPORTED}}$ records a query or interface outside the registered method family. The latter three branches do not contain a neural point estimate.

Threat model and failure taxonomy. We treat the neural estimator, the certificate compiler, and the data provider as fallible components. The control plane must therefore tolerate an expressive estimator that would return a number for every input, a compiler that emits a syntactically valid but incorrect candidate, accidental inclusion of post-treatment or oracle fields, hash mismatches, and runtime failures during multimodal encoding. The protected property is narrow but testable: no point estimate may be emitted unless the independent verifier authorizes the exact query, contract, functional, and runtime interfaces. Structural non-identification, finite-sample overlap failure, unsupported interfaces, and implementation faults are separate result statuses. The analyst-declared graph remains an explicit, independently reviewable component of the contract, and the verifier establishes authorization relative to that declaration.

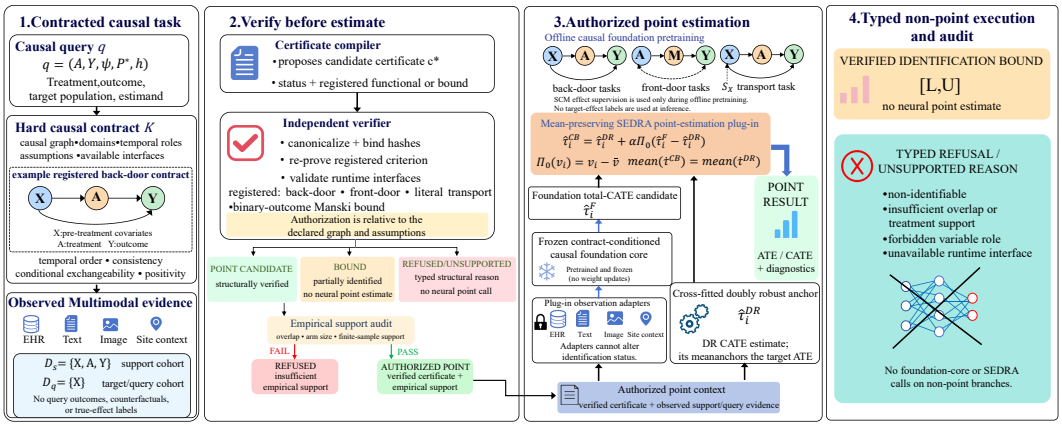


Figure 1: CausaBridge separates structural authorization from neural estimation. A point certificate is independently verified before the multimodal causal core is called. Verified bounds, refusals, and unsupported queries bypass the point-estimation path.

4 CausaBridge

4.1 Overview

Figure 1 summarizes the execution boundary. The system first compiles the query and contract, then independently verifies the candidate. Only the verified point branch reaches the multimodal adapters and foundation core. This ordering is a semantic invariant rather than a logging convention: the executor performs the guard and certificate checks before materializing an adapter batch or calling the neural core.

4.2 Typed causal contracts and certificates

The contract κ is a canonical, typed description of the causal task. It includes the treatment and outcome identifiers, target domain, graph edges and domains, hard assumptions, available data interfaces, query hash, and a contract hash. Canonical serialization sorts identifiers and edge descriptions before hashing, so equivalent input orderings produce the same contract identity while changes to a variable, direction, domain, or assumption change the identity.

The compiler maps (q, κ) to a candidate certificate. A certificate contains a candidate status, the canonical functional or bound description, required variables and domains, assumptions, compiler version, query hash, contract hash, and a certificate hash. The compiler is not trusted by the execution path. Instead, the verifier independently snapshots and canonicalizes the candidate, checks its envelope and hashes, validates its runtime interfaces, and re-proves the registered identification criterion. The current registered point forms cover the implemented back-door, front-door, and literal transport criteria; the registered partial-identification form is a narrow binary-outcome Manski bound. Algebraically equivalent but unregistered rewrites are refused by design.

The gate can be written as a typed partial function. Let $\mathsf{V}(q, \kappa, c)$ be the verifier predicate and let PointExec denote the neural point executor. Then

$$\mathsf{Exec}(\mathcal{D}_{\mathrm{obs}}, q, \kappa) = \begin{cases} \mathsf{PointExec}(\mathcal{D}_{\mathrm{obs}}, q, \kappa, c), & \mathsf{V}(q, \kappa, c) = 1 \ \wedge \ c.status = \mathsf{POINT}, \\ \mathsf{Bound}(c), & c.status = \mathsf{BOUND}, \\ \mathsf{Refuse}(c), & c.status \in \{\mathsf{REFUSED}, \mathsf{UNSUPPORTED}\}. \end{cases} \tag{3}$$

The first branch is unreachable for a rejected certificate. This is the semantic invariant proved in Appendix C; it is stronger than a post-hoc confidence threshold because the forbidden branch is never materialized.

4.3 CERTIFIED EXECUTOR

The certified executor is the only public inference control plane. It first rejects malformed or oracle-bearing inputs, then checks task counts, query and contract consistency, runtime interfaces, and the verifier decision. The verified point branch invokes the point executor exactly once and validates posterior shape, finiteness, dtype, device, and task alignment before constructing a result. Uncertainty is derived from the predicted potential-outcome scales and is not used to override the structural decision.

The bound branch returns the immutable verified interval without calling an adapter or the core. The refusal and unsupported branches return the candidate and decision metadata while fixing point fields to null. A neural runtime fault is propagated as a runtime fault; it is never relabeled as a causal refusal. This distinction prevents implementation failures from being mistaken for evidence that a query is non-identifiable.

4.4 CONTRACT-CONDITIONED CAUSAL FOUNDATION MODEL

Let a task $\xi = (\mathcal{D}_{\mathrm{obs}}, q, \kappa)$ be drawn from a distribution Π over SCMs, mechanisms, sample sizes, and registered queries. CausaBridge is pretrained once on many $\xi \sim \Pi_{\mathrm{train}}$ and evaluated on unseen tasks without updating the foundation core. This breadth and downstream reuse, rather than parameter count, is the foundation-model claim. A graph-contract encoder maps canonical nodes, temporal roles, domains, directed and bidirected edges, selection edges, assumptions, and the verified functional abstract syntax tree (AST) into deterministic task tokens. The estimator is thus conditioned on the structural object that authorizes the estimand, not on a true effect or an unverified identification label.

Clinical observations are represented by typed EHR tokens and optional image tokens. The image pathway uses a frozen radiology encoder and a trainable adapter/resampler; the structured pathway uses typed event, time, missingness, and reliability features. Each adapter satisfies the same token contract and has no access to compiler or verifier status. The posterior contains propensity, potential-outcome means and scales, CATE, verified-point ATE, overlap, and patient representations. Structural-status and functional-form prediction heads are intentionally absent: the neural core estimates an authorized functional but does not decide whether it is authorized.

4.5 ESTIMATOR INTERFACE AND TRAINING OBJECTIVE

For patient representation z_i, the core predicts propensity e_i, potential-outcome means μ_{i0}, μ_{i1}, log-scales s_{i0}, s_{i1}, and CATE δ_i. The estimator additionally forms two cross-fitted task summaries. Each fold concatenates treated and control representation means, their contrast, global task context, observed outcome summaries, and treatment/observation fractions. A learned head g_θ directly predicts the fold effect:

$$\widehat{\tau}^{(k)} = g_\theta(m_1^{(k)}, m_0^{(k)}, m_1^{(k)} - m_0^{(k)}, \bar{z}, r^{(k)}), \quad \widehat{\tau} = \tfrac{1}{2}\left(\widehat{\tau}^{(1)} + \widehat{\tau}^{(2)}\right). \tag{4}$$

The raw treated–control outcome difference occurs in $r^{(k)}$ as a learnable input but is not added with a fixed unit coefficient. This removes the bias of compulsory residual anchoring while preserving identical supervision and compute for fixed- and diverse-mechanism pretraining. The point-task objective is

$$\begin{aligned}\mathcal{L} = \lambda_Y \mathcal{L}_{\text{factual-NLL}} + \lambda_T \mathcal{L}_{\text{propensity}} + \lambda_C \mathcal{L}_{\text{CATE}} + \lambda_A \mathcal{L}_{\text{ATE}} \\ + \lambda_O \mathcal{L}_{\text{overlap}} + \lambda_R \mathcal{L}_{\text{representation}} + \lambda_D \mathcal{L}_{\text{dropout}} + \lambda_P \mathcal{L}_{\text{permutation}}.\end{aligned} \tag{5}$$

The CATE and ATE terms are masked by a verified point certificate whose query and contract hashes match the separate supervision certificate. For refused, bound, unsupported, or mismatched tasks these terms are exactly zero; factual and treatment nuisance losses remain available when their observations are valid. The representation term aligns the adapter output to the frozen-track reference, dropout consistency tests missing modalities, and permutation consistency prevents patient-order shortcuts. The implementation exposes each term separately in the evaluation record so that a lower aggregate loss cannot hide a failed safety term.

The resulting posterior is typed by both the structural branch and the estimator diagnostics:

$$\widehat{R} = \big(\text{status}, \widehat{\tau}_{\text{ATE}}, \widehat{\tau}_{\text{CATE}}, \widehat{\sigma}, \text{overlap}, \text{provenance}\big). \tag{6}$$

Only the point branch contains the first three numerical effect fields. This separation lets the evaluation report safety and estimation quality without collapsing them into one score.

4.6 MEAN-PRESERVING RESIDUAL ADAPTATION

The foundation core is pretrained with the direct task-effect head in Equation 4; it is not DR-anchored during pretraining. For authorized downstream tabular transfer, the scale-equivariant doubly robust adapter (SEDRA) combines a cross-fitted doubly robust (DR) anchor with a frozen-core foundation residual. On the support cohort, let $\widehat{\phi}_i$ be the DR pseudo-outcome, $\widehat{\tau}_i^{\mathrm{DR}}$ the out-of-fold anchor, and $\widehat{r}_i$ an honest residual prediction from the adapted foundation representation to $\widehat{\phi}_i - \widehat{\tau}_i^{\mathrm{DR}}$. Ridge shrinkage selects

$$\alpha = \mathrm{clip}_{[0,1]} \frac{\sum_i \widehat{r}_i(\widehat{\phi}_i - \widehat{\tau}_i^{\mathrm{DR}})}{\sum_i \widehat{r}_i^2 + \lambda}. \tag{7}$$

Candidate adaptation budgets are selected by out-of-fold pseudo-outcome risk. For a target cohort of size n_q, the point estimate is

$$\widehat{\tau}_i^{\mathrm{CB}} = \widehat{\tau}_i^{\mathrm{DR}} + \alpha \Pi_0(\widehat{r}_i^F), \qquad \Pi_0(v_i) = v_i - \frac{1}{n_q}\sum_{j=1}^{n_q} v_j. \tag{8}$$

Thus the foundation residual can refine effect heterogeneity while its target mean is exactly zero, so the cohort ATE remains anchored to the cross-fitted DR estimate. Identification verification precedes this computation; empirical overlap failure returns a typed refusal and no point-valued residual is materialized.

4.7 SUPERVISION-SEPARATED TRAINING

Each provider returns an input batch and a separate supervision object. The input contains observations, contracts, and genuine point certificates. The supervision object may contain known SCM effects or semi-synthetic counterfactual labels, but those fields are not available to the public inference path. Effect losses are enabled only when the supervision certificate matches the task's query and contract hashes; otherwise the effect losses are exactly zero. This prevents oracle effect targets, structural labels, and failure reasons from entering the model as shortcut features.

The formal medical curriculum cycles through structured, image-only, and missing-not-at-random (MNAR) image scenarios. The real MIMIC protocol disables the text channel because the available report provenance is not sufficient for the declared temporal contract. Adapter experiments vary the configured target fraction and protocol mode, and every run records the fraction, mode, input hashes, freeze audit, and checkpoint provenance.

4.8 FORMAL GUARANTEES AND FOUNDATION TRANSFER

The design provides a software-level guarantee that is independent of the neural model's calibration. Under a sound registered verifier, deterministic canonicalization, and fail-closed control flow, a returned point effect implies an accepted certificate for the exact query and contract. A bound, refused, or unsupported result cannot call the point adapter or core, and changing a training-only supervision object cannot change inference for fixed observations and certificates. These statements are proved by case analysis over the executor and by function extensionality in Appendix C.

Theoretical synopsis. Under verifier soundness, collision-resistant canonical binding, and fail-closed control flow, **Theorem 1** states that every point output has a valid registered certificate for the exact current task; non-point branches cannot call the neural estimator. Under bounded registered kernels, **Theorem 2** controls functional error by the sum of nuisance-leaf errors. Under conditional-kernel preservation, **Theorem 4** shows that an adapter leaves the certified functional invariant; approximate preservation leaves an explicit remainder. These assumptions are proof obligations or diagnostics, not conclusions learned from clinical outcomes.

The representation theorem is conditional but operationally useful: if the adapter preserves every conditional kernel used by the certified functional and respects the contract's temporal and selection roles, replacing raw multimodal evidence with adapter tokens leaves the functional unchanged.

The target-effect error then decomposes into sample, functional, transport, and adapter terms. This decomposition turns a vague "multimodal gap" into four diagnostics that can be estimated by repeated splits, nuisance checks, held-out hospitals, and modality-dropout tests. The symbolic gate itself is cheap: canonicalization is $O(|V|\log|V| + |E|\log|E|)$, AST validation is linear in AST size, and graph separation is linear in the registered graph representation; neural encoding is never paid for refusal and bound branches. For an authorized point query, Appendix C additionally bounds the plug-in estimand error by the sum of registered nuisance-leaf errors, provides a held-out concentration bound, and proves that mismatched supervision has zero effect-loss gradient. Together, these results establish a compositional chain from identification authorization through fail-closed execution and nuisance estimation to an explicit error diagnostic, while treating clinical validity as a separately audited contract-level property.

The foundation claim concerns unseen tasks, not merely optimization. Let $\ell(h;\xi) \in [0, B]$ be authorized ATE loss for a contract-conditioned hypothesis h on task $\xi \sim \Pi$. For N i.i.d. pretraining tasks, Appendix C shows that, simultaneously for $h \in \mathcal{H}$ and with probability at least $1-\delta$,

$$R_\Pi(h) \le \widehat{R}_N(h) + 2\mathfrak{R}_N(\ell \circ \mathcal{H}) + B\sqrt{\frac{\log(1/\delta)}{2N}}. \tag{9}$$

The guarantee applies to authorized tasks drawn from Π, making the relevant task-distribution coverage explicit. This motivates evaluating every checkpoint on the same unseen task panel, which separates optimization progress from transfer across mechanism distributions.

5 Experiments

5.1 Questions and evidence domains

The evaluation separates five questions that require different evidence. First, does the certified executor prevent a point output when the declared query is not authorized? Second, does pretraining across diverse mechanisms improve transfer to unseen authorized tasks at matched compute? Third, how accurately does the estimator perform against classical, neural, and causal-foundation comparators on standard benchmarks? Fourth, can a frozen causal core and plug-in adapters transfer to MIMIC-based multimodal tasks? Fifth, does the linked clinical cohort provide sufficient empirical treatment support for the declared query? Each question is evaluated in its matching evidence domain: formal and executable safety checks, common-target transfer, matched benchmark accuracy, multimodal semi-synthetic evaluation, and clinical support auditing. Specifically, the common-target study evaluates transfer of the direct effect head, the public benchmarks evaluate identification-gated SEDRA on matched support, and the MIMIC study evaluates multimodal adaptation together with empirical-support refusal.

Table 1: Matched-support benchmarks on 309 authorized seed-indexed tasks (coverage: IHDP 285/285, ACIC 18/21, Twins 3/3, Jobs 3/3). Comparators use identical keys; Jobs reports clipped inverse-probability-weighted (IPW) policy value. Bold marks the best mean; inference clusters by repetition.

	ACIC		IHDP		Twins		Jobs
Method	ATEE ↓	PEHE ↓	ATEE ↓	PEHE ↓	ATEE ↓	PEHE ↓	Policy ↑
AIPW	0.5538	3.2869	0.6568	5.9238	0.0194	0.3297	6090.0820
Causal forest	0.1458	1.3839	0.5286	3.6291	0.0197	0.3196	5829.6758
TARNet	0.2637	3.0663	0.2751	2.1701	0.0280	0.5549	5409.6079
CFRNet	0.3756	3.7246	0.2759	2.1800	0.0210	0.5250	5246.6502
Dragonnet	0.4025	3.2831	0.9037	2.6820	0.0210	0.4957	**47489.5907**
CausalFM	1.5760	3.5137	2.3790	6.0222	0.0484	0.3211	5927.4360
CausalPFN	0.1375	1.2201	0.2866	**0.5938**	0.0117	0.3214	6116.2020
Do-PFN	2.7166	4.2444	2.5270	6.8335	0.1862	0.4233	5214.7543
CausaBridge + SEDRA	**0.0634**	**1.1284**	**0.1139**	2.4783	**0.0092**	**0.3178**	5632.9641

5.2 FOUNDATION-TRANSFER PROTOCOL

The transfer experiment crosses fixed-mechanism and diverse-mechanism curricula, three training seeds, and checkpoints at 1.5k, 7.5k, and 30k optimizer steps. For each of the 18 checkpoint cells, a two-parameter affine adapter is fitted on 128 calibration tasks. Evaluation uses a separate seed and the same ordered panel of 256 unseen confirmatory tasks for every cell, yielding 4,608 confirmatory evaluations. Calibration targets never enter the confirmatory panel.

At 30k steps let $d_{st} = |\widehat{\tau}_{st}^{\mathrm{diverse}} - \tau_t| - |\widehat{\tau}_{st}^{\mathrm{fixed}} - \tau_t|$. The decision rule, specified before confirmatory evaluation, requires a negative pooled mean and crossed-bootstrap upper endpoint, at least two negative seed means, and a negative error slope for every diverse-mechanism seed. The paired diverse-minus-fixed calibrated ATEE is -0.0180 (95% crossed seed–task bootstrap CI $[-0.0242, -0.0117]$; $p < 10^{-4}$), with seed means -0.0179, -0.0181, and -0.0180 and budget slopes -0.0187, -0.0097, and -0.0163. All four criteria pass (Table 4), establishing mechanism-diversity transfer on the prespecified common-target distribution.

5.3 PUBLIC BENCHMARKS AND COMPARATORS

We use the pinned public splits of IHDP, ACIC 2016, Twins, and Jobs (Hill, 2011; Dorie et al., 2019; Louizos et al., 2017; LaLonde, 1986).

The comparison study contains 5,376 completed jobs spanning 100 IHDP repetitions, 10 ACIC repetitions, one pinned evaluation each for Twins and Jobs, 16 methods, and three seed indices. The authorized CausaBridge evaluation uses the held-out repetitions IHDP 5–99 and ACIC 3–9 together with the pinned Twins and Jobs evaluations, for 312 candidate jobs. It returns 309 point results and three typed refusals, all corresponding to ACIC repetition 9 under the prespecified overlap rule. Thus the executor authorizes 309/312 candidate jobs (99.0%; 95% repetition-bootstrap CI $[0.971, 1.000]$).

IHDP, ACIC, and Twins report absolute ATE error (ATEE) and root PEHE; lower is better. Jobs lacks individual counterfactual labels and therefore reports IPW policy value; higher is better. Preprocessing records source hashes, original indices, treatment support, and the truth boundary. Missing counterfactual metrics remain unreported rather than being reconstructed from model outputs.

Table 1 reports point-risk means on exactly the keys where CausaBridge is authorized: 285/285 IHDP, 18/21 ACIC, 3/3 Twins, and 3/3 Jobs seed-indexed tasks. Every comparator is recomputed on these same keys; deterministic zero-shot predictions may repeat across seed indices and are not interpreted as independent stochastic fits. CausaBridge + SEDRA records the lowest displayed ACIC ATEE (0.0634) and PEHE (1.1284) at 18/21 point coverage; the repetition-bootstrap coverage interval is [0.571, 1.000]. Relative to CausalPFN, the paired ACIC ATEE difference is -0.0741 (95% CI $[-0.1112, -0.0419]$; $p < 10^{-4}$ at the 10,000-draw Monte Carlo resolution), whereas the PEHE difference of -0.0918 is not resolved (95% CI $[-0.2511, 0.0667]$; $p = 0.2702$). Results on the remaining benchmarks are metric dependent: CausaBridge records the lowest displayed IHDP ATEE (0.1139), whereas CausalPFN records the lowest displayed IHDP PEHE (0.5938); on Twins, CausaBridge records the lowest displayed ATEE (0.0092) and PEHE (0.3178). Its Jobs policy value is 5632.9641 and is not the highest displayed value. Accordingly, we do not claim uniform dominance across datasets or metrics. The complete matched matrix is reported in Appendix 10; Appendix Table 11 separates DR anchoring from the learned residual.

5.4 MIMIC-BASED TRANSFER AND PLUG-IN ADAPTATION

The multimodal evaluation uses linked MIMIC-IV/MIMIC-CXR support cohorts and nine point-authorized semi-synthetic test scenarios spanning structured, image-only, missing-image, site-shift, and measurement-error conditions. We evaluate independently trained models under a prespecified multi-seed protocol, producing matched scenario–seed point-estimation cells for each displayed method. Hidden-confounding and overlap-failure contracts remain non-point branches and are evaluated in the safety accounting rather than mixed into numerical error averages.

Table 2 reports the common semi-synthetic slice. CausaBridge achieves mean ATE MAE 0.1353 and mean root PEHE 0.1756, with lower descriptive errors than the displayed classical and causal-foundation comparators. The EHR-only (0.1302/0.1711), no-image (0.1304/0.1719), and no-site

(0.1367/0.1756) rows use identical checkpoints, seeds, scenarios, and support cohorts. These values are numerically close, but the EHR-only and no-image ablations are slightly lower than the full-input system. We therefore interpret the rows as an interface-sensitivity analysis rather than evidence that every additional modality monotonically improves estimation accuracy. Seed-specific results and the complete protocol matrix are provided in Appendix H.

The adapter matrix crosses zero-shot, adapter-only, full-finetune, and from-scratch modes with 5% and 100% target fractions. Adapter-only updates 2.98M of 51.36M parameters (5.8%). At the 100% target fraction it reduces ATE MAE by 49.5% and root PEHE by 42.5% relative to full fine-tuning. From-scratch training at the same target fraction attains lower descriptive errors (0.1214/0.1691) than adapter-only transfer (0.1353/0.1756). Thus, this comparison supports parameter-efficient reuse relative to the prespecified full-fine-tuning control, rather than a universal accuracy advantage over training from scratch. The frozen-core hash is unchanged for every adapter-only and zero-shot cell, documenting adapter-only transfer as a parameter-efficient reuse configuration.

5.5 Clinical empirical-support audit

The 360-minute MIMIC-IV/MIMIC-CXR landmark contract permits EHR variables and temporally eligible images, excluding report text. The 172-patient cohort has only two treated patients. Treated/control counts are 2/157 in the principal source environment and 1/127 and 2/126 in the two 128-patient query batches. All fail the prespecified minimum of 10 patients per arm, triggering typed refusals (Appendix Table 5).

5.6 Structural safety evidence

In five-seed replay across point and non-point contracts, mutations, and input checks, the complete executor achieves a 0% false-point rate, 100% refusal accuracy and semantic match, and zero non-point neural calls. Removing either the verifier or overlap audit yields a 25% false-point rate and 10 non-point neural calls; allowing unsafe variable roles yields a 50% false-point rate (Appendix Table 3).

6 Conclusion

CausaBridge couples reusable effect estimation with verified exact-query authorization, separating point estimates from bounds and refusals. Experiments support mechanism-diversity transfer, paired ACIC ATEE gains, and parameter-efficient reuse relative to the full-fine-tuning control. Structural ablations demonstrate the importance of verification, overlap auditing, and variable-role constraints for fail-closed execution.

Table 2: MIMIC-based semi-synthetic estimation over nine authorized scenarios under a prespecified multi-seed protocol. Results are aggregated over matched scenario–seed cells; rows share scenarios and support cohorts, bold identifies the principal system, and indented rows are input ablations.

Method	ATE MAE ↓	Root PEHE ↓
External and classical comparators		
AIPW	0.1948	0.6657
DR learner	0.1948	0.2955
Linear S/T/X learner	0.1934	0.2850
CausalPFN	0.2163	0.2949
Principal system and input ablations		
CausaBridge	0.1353	0.1756
EHR only	0.1302	0.1711
no image	0.1304	0.1719
no site	0.1367	0.1756

Ethics statement

This work uses credentialed, de-identified clinical resources and does not recruit participants or expose row-level data. MIMIC-IV/MIMIC-CXR access, storage, and use remain subject to their data-use agreements. The paper reports no patient-level examples and exports no direct identifiers. The method is not presented as a medical device or autonomous treatment policy. Potential risks include over-trust in retrospective effect estimates, site-specific bias, and misuse of a causal contract whose assumptions are clinically inappropriate; these risks motivate explicit refusal and empirical-support checks.

Reproducibility statement

The anonymous code supplement contains causal schemas, compiler/verifier tests, the certified executor, providers, training stages, formal configurations, manifest audits, and reproduction scripts. Appendices specify assumptions and proofs, task generation, target-panel construction, splits, seeds, adapter modes, hyperparameters, compute, and statistics. Restricted clinical data are not redistributed; reproduction requires credentialed access and locally configured paths. Numerical tables are generated only from evaluation records carrying input, environment, code, configuration, checkpoint, and protocol hashes together with passed completeness checks.

AI use statement

Generative AI tools were used to assist with language editing and polishing for readability, and with literature retrieval and discovery. All retrieved references were independently verified by the authors, and all AI-assisted edits were reviewed by the authors. The authors take full responsibility for the final content of this work.

References

Vahid Balazadeh, Hamidreza Kamkari, Valentin Thomas, Benson Li, Junwei Ma, Jesse C. Cresswell, and Rahul G. Krishnan. Causalpfn: Amortized causal effect estimation via in-context learning, 2025. URL https://arxiv.org/abs/2506.07918.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al. On the opportunities and risks of foundation models, 2021. URL https://arxiv.org/abs/2108.07258.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Aris Dufour, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019. doi: 10.1214/18-STS667.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011. doi: 10.1198/jcgs.2010.08162.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6:317, 2019. doi: 10.1038/s41597-019-0322-0.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo Anthony Celi, and Roger G. Mark. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 2023. doi: 10.1038/s41597-022-01899-x.

Soeren R. Kuenzel, 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.

Robert J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76(4):604–620, 1986.

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, Dennis Frauen, Emil Javurek, and Stefan Feuerriegel. Foundation models for causal inference via prior-data fitted networks. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=d2L1ndOKjq.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=KSugKcbNf9.

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

Jake Robertson, Arik Reuter, Siyuan Guo, Noah Hollmann, Frank Hutter, and Bernhard Schoelkopf. Do-pfn: In-context learning for causal effect estimation, 2025. URL https://arxiv.org/abs/2506.06039.

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.

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.

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9(64):1941–1979, 2008. URL https://www.jmlr.org/papers/v9/shpitser08a.html.

Jiaqi Zhang, Joel Jennings, Agrin Hilmkil, Nick Pawlowski, Cheng Zhang, and Chao Ma. Towards causal foundation model: on duality between optimal balancing and attention. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59042–59065, 2024. URL https://proceedings.mlr.press/v235/zhang24x.html.

A Certificate-gated execution algorithm

The certified executor below makes the ordering constraint explicit. Verification precedes every adapter or causal-core invocation, and the refusal and bound branches terminate before any point-estimation tensor is materialized.

B Certificate schema and independent verification

A point certificate contains the following immutable fields:

$$c = (\text{status}, \text{functional}, \text{required variables}, \text{required domains}, \\ \text{assumptions}, \text{compiler version}, H(q), H(\kappa), H(c)).$$

The canonical hash serializes identifiers as strict UTF-8 strings and sorts all order-insensitive sequences. The verifier rejects empty identifiers, duplicate or overlapping density/sum variables, invalid domains, malformed functional trees, cycles, depth above 64, or more than 4,096 functional

Certified executor.
Input: observational batch $\mathcal{D}_{\text{obs}}$, query q, contract κ.
1. Validate task count, typed fields, pointer ranges, and forbidden oracle fields.
2. Compile (q, κ) into candidate c.
3. Independently snapshot, canonicalize, hash, and verify c.
4. If status is BOUND, return the verified identification bound without invoking an adapter or the causal core.
5. If status is REFUSED or UNSUPPORTED, return the decision with point fields fixed to null.
6. If status is POINT, validate the verifier seal and query/contract hashes.
7. Materialize the authorized adapter batch and call the neural core exactly once.
8. Validate posterior shape, finiteness, dtype, device, and task alignment; return point effect and uncertainty.

Figure 2: Execution order of the public causal inference control plane.

nodes. It checks that the claimed status matches the payload: a point certificate cannot carry a bound or failure field, while a bound certificate cannot carry a point functional. These checks are repeated after the candidate is copied into a trusted immutable snapshot.

The verifier operates independently of the compiler, observed values, supervision targets, model outputs, and provider internals. It re-proves the registered criterion using the generic graph separation primitive and the declared runtime interfaces, so every accepted candidate satisfies the execution path's independent checks.

Appendix C formalizes the registered software boundary, representation-preservation condition, transport decomposition, and proof obligations. The resulting guarantees are explicitly relative to the declared graph, interfaces, and finite-sample conditions, while clinical validity and estimator calibration are evaluated through separate empirical audits.

C Formal theory and proof obligations

This section states the guarantees that follow from the certificate and execution design. The guarantees condition on the declared causal graph and representation interface. The registered graph class is finite ADMGs with mixed observed variables and single-source/single-target selection diagrams; multi-source transport is evaluated as a separate empirical extension.

C.1 Scope and definitions

Let κ be a normalized hard causal contract, q a typed query, and c a canonical certificate. The contract fixes the graph, variable domains, temporal roles, available distributions, and analyst-declared assumptions. Let $\mathsf{V}(q, \kappa, c)$ be the independent verifier predicate and let $\mathsf{Exec}(D, q, \kappa)$ be the certified executor. The executor has four observable result branches: point, bound, refused, and unsupported.

We use the following assumptions.

(A1) **Verifier soundness.** If V accepts a point certificate, its functional is valid for the declared registered criterion and the hashes bind the query and contract snapshot.

(A2) **Canonical integrity.** Canonical serialization is deterministic and the hash is collision-resistant for the execution horizon.

(A3) **Fail-closed control flow.** The executor reaches the point constructor only after the verifier decision, hash checks, and interface checks have returned success.

(A4) **Hard-contract semantics.** Graph, temporal, measurement, and selection assumptions supplied to the verifier are treated as fixed semantics, not as quantities inferred from outcome values.

(A5) **Task sampling.** The N tasks used by the task-generalization statement are independent draws from the same declared distribution Π on which its population risk is evaluated. A shifted target distribution is handled by a separate transport argument.

(A6) **Bounded task loss.** Authorized task loss is measurable and lies in $[0, B]$. Non-point tasks use a separate status loss and are never assigned a fabricated numerical target.

The assumptions are an explicit scope boundary rather than hidden claims about clinical truth. A1 is audited with exhaustive finite-graph and mutated- certificate tests; A2 with canonicalization vectors and rehash tests; A3 with branch-call and forbidden-field tests; and A4 with analyst-facing contract review. Point-estimation regularity (bounded kernels, positivity, and nuisance convergence) is checked by the diagnostics attached to Theorem 2 below. A point result is authorized only when every applicable obligation is satisfied; otherwise, the executor emits the corresponding typed non-point result.

C.2 Registered identification forms

The verifier does not accept an arbitrary symbolic expression. It accepts only canonical forms whose side conditions are checked against the contract. For a back-door adjustment set Z, the registered functional is

$$p(y \mid \mathrm{do}(a)) = \sum_z p(y \mid a, z)p(z),$$

provided Z contains no descendant of A and blocks every back-door path. For a front-door mediator M, the registered functional is the usual two-stage expression

$$p(y \mid \mathrm{do}(a)) = \sum_m p(m \mid a) \sum_{a'} p(y \mid m, a')p(a'),$$

provided the declared front-door side conditions hold. Transport certificates are deliberately restricted to a single-source/single-target selection diagram. Let $D \in \{s, t\}$ denote source and target domains. The registered transport functional is

$$p_t(y \mid \mathrm{do}(a)) = \sum_z p_s(y \mid a, z)p_t(z),$$

provided that Z blocks source-domain back-door paths from A to Y, $Y(a) \perp D \mid Z$ (selection invariance), and the target distribution $p_t(z)$ and the required source conditional are available with the declared positivity conditions. An unregistered multi-source rewrite is refused. For binary outcomes, the only registered partial-identification form is the Manski bound. Writing $p_a = \Pr(Y = 1, A = a)$, it returns

$$\psi \in \big[p_1 - \min\{1, p_0 + \Pr(A = 1)\},\ \min\{1, p_1 + \Pr(A = 0)\} - p_0\big].$$

The symbolic AST stores density leaves and summation variables, while the verifier checks all graph, selection, positivity, and interface side conditions independently of the compiler.

Lemma 1 (AST evaluation correctness). For a well-formed finite AST, the functional evaluator returns the value of the AST under the supplied density interface, or fails before returning a value if a required density or summation variable is absent.

Proof. Induct on AST depth. A density leaf reads exactly one declared interface. A product multiplies the values of its children, and a sum-out node integrates its child over the declared variable. The validation rules reject cycles, malformed nodes, and missing interfaces before evaluation, so the induction is total on the accepted domain.

Lemma 2 (Relative soundness of registered forms). If the contract satisfies the back-door or front-door side conditions above, the corresponding registered AST equals the interventional distribution under the contract's SCM semantics. The same statement holds for the registered transport criterion relative to its selection diagram.

Proof. The back-door and front-door equalities follow from the standard do-calculus adjustment criteria; the transport case follows from the declared selection-diagram transport rule. Lemma 1 shows that the accepted AST executes exactly the density functional proved by the criterion. The result is relative to the hard graph and selection assumptions and does not infer their truth from the data.

Lemma 3 (Manski-bound validity). For binary Y and A, with no assumptions beyond consistency, the interval returned above contains every feasible average treatment effect $\psi = \mathbb{E}[Y(1) - Y(0)]$ compatible with the observed law.

Proof. Consistency gives $\Pr(Y(1) = 1) \geq \Pr(Y = 1, A = 1) = p_1$ and $\Pr(Y(1) = 1) \leq \min\{1, p_1 + \Pr(A = 0)\}$. The analogous inequalities for $Y(0)$ are $p_0 \leq \Pr(Y(0) = 1) \leq \min\{1, p_0 + \Pr(A = 1)\}$. Subtracting the upper and lower endpoints yields the stated interval.

C.3 Certificate and execution guarantees

Theorem 1 (Certificate-gate soundness). Under Lemmas 1–2 and (A1)–(A3), if $\mathsf{Exec}(D, q, \kappa)$ returns a point result, then the returned point was authorized by an accepted certificate whose canonical query and contract hashes match the current inputs, and the functional belongs to the registered verifier family.

Proof. By (A3), the point constructor is reachable only after the verifier, hash, and interface predicates have succeeded. The successful verifier decision is therefore $\mathsf{V}(q, \kappa, c) = 1$. By (A1), that decision implies validity of the registered functional and consistency of the certificate with the declared contract. By (A2), a changed normalized query or contract cannot silently reuse the accepted hash except with negligible hash collision probability. Hence every point result has the stated authorization relative to the analyst-declared assumptions.

Proposition 1 (Certified non-point execution and branch noninterference). For bound, refused, and unsupported statuses, the executor terminates in the corresponding typed non-point constructor before adapter materialization or neural-core execution. Point fields remain null on these mutually exclusive branches.

Proof. The executor dispatch is a case split on the immutable verifier status. The bound branch constructs its verified identification bound before any adapter materialization; the refused and unsupported branches set point fields to null and terminate. These branches are disjoint from the point branch, so their outputs cannot contain a point-estimation call.

Proposition 2 (Oracle noninterference at inference). Let S and S' be two supervision objects with arbitrary different true effects, counterfactual labels, or identification targets. If the observational batch, query, contract, and verifier-issued certificate are equal, the public inference output is equal for S and S', provided the runtime has no hidden global state. Thus supervision can affect training losses but not the deployed inference function.

Proof. The public inference signature is a function of (D, q, κ, c) and contains no supervision argument. The forbidden-field guard rejects an attempt to place S in the public batch. Function extensionality therefore gives $f(D, q, \kappa, c; S) = f(D, q, \kappa, c; S')$ for all admissible S, S'.

C.4 Estimation for authorized point queries

The preceding results authorize a functional but do not by themselves assert that a learned nuisance model estimates it accurately. The next result closes that interface for every registered finite AST. Let $\Phi_c(K_1, \ldots, K_m)$ be the scalar functional attached to a point certificate, where the K_j are its density, conditional-kernel, or bounded-moment leaves. Assume every leaf operator and its plug-in version has L^1 operator norm at most one under the contract's declared measure, sup-norm at most $B \geq 1$, and let $\epsilon_j = \|\widehat{K}_j - K_j\|_1$ in that same norm.

Theorem 2 (Certified functional plug-in stability). If the executor returns a point result $\widehat{\psi}_c = \Phi_c(\widehat{K}_1, \ldots, \widehat{K}_m)$, then

$$|\widehat{\psi}_c - \psi_c| \leq B^{m-1} \sum_{j=1}^{m} \epsilon_j, \qquad \psi_c = \Phi_c(K_1, \ldots, K_m).$$

Consequently, convergence of all registered nuisance leaves in L^1 implies consistency of the authorized plug-in estimand; non-point branches retain their corresponding bound, refusal, or unsupported semantics.

Proof. Replace the leaves one at a time and telescope the difference of the two AST evaluations. A product node contributes at most $B^{m-1}\epsilon_j$ by the L^1 product inequality. Sum, integration, and marginalization nodes are contractions under the unit-mass assumption, so the same bound propagates to the root. Summing over all leaves proves the claim.

Corollary 1 (Back-door ATE plug-in). For $\psi = \int(\mu_1(z) - \mu_0(z))p(z)\,dz$, bounded outcome regressions, and normalized $\widehat{p}$, the certified plug-in estimator obeys

$$|\widehat{\psi} - \psi| \le B_\mu\|\widehat{p} - p\|_1 + \|\widehat{\mu}_1 - \mu_1\|_{1,\widehat{p}} + \|\widehat{\mu}_0 - \mu_0\|_{1,\widehat{p}}.$$

The experiment therefore reports nuisance errors and overlap diagnostics separately rather than presenting a single aggregate training loss as a causal guarantee.

Theorem 3 (Held-out point-estimate concentration). Let a verified point certificate authorize a conditional effect $\delta(x)$, let a frozen model produce $\widehat{\delta}$ on an independent held-out fold of n patients, and assume both functions lie in $[-B, B]$. For $\widehat{\psi} = n^{-1}\sum_{i=1}^{n}\widehat{\delta}(X_i)$ and $\psi = \mathbb{E}[\delta(X)]$, with probability at least $1 - \alpha$,

$$|\widehat{\psi} - \psi| \le \mathbb{E}|\widehat{\delta}(X) - \delta(X)| + B\sqrt{\frac{2\log(2/\alpha)}{n}}.$$

The first term is a nuisance/representation approximation term and the second is the held-out sampling term. The bound is conditional on the certificate's identification assumptions, while clinical graph validity is assessed separately through domain review.

Proof. Add and subtract $\mathbb{E}[\widehat{\delta}(X)]$. The first term is bounded by the $L^1(P_X)$ difference. Conditional on the frozen model, the held-out summands lie in $[-B, B]$, so Hoeffding's inequality bounds the second term by the displayed quantity.

Proposition 3 (Effect-loss admissibility). Let M be one exactly when the point status and the query/contract hashes in the supervision certificate match, and zero otherwise. If $\mathcal{L}_{\text{eff}} = M(\lambda_C\mathcal{L}_{\text{CATE}} + \lambda_A\mathcal{L}_{\text{ATE}})$, then $\nabla_\theta\mathcal{L}_{\text{eff}} = 0$ whenever $M = 0$. Thus a refusal, bound, unsupported query, or mismatched oracle can train nuisance/representation terms but cannot train the effect heads with an unauthorized target.

Proof. M is a verifier-issued, stop-gradient gate determined before the model call. When $M = 0$, the effect loss is identically zero as a function of θ; differentiating gives the result.

Proposition 4 (Direct-head anchor freedom). Let a denote the observed treated–control outcome contrast and let the direct-head fold estimate be $g_\theta(u, a)$, where u collects all remaining task summaries. Its derivative with respect to a is $\partial g_\theta(u, a)/\partial a$ wherever the network is differentiable and is not constrained to one. By contrast, a residual estimator $a + r_\theta(u, a)$ has derivative $1 + \partial r_\theta/\partial a$ and must learn to cancel any bias in a.

Proof. This follows by differentiation of the two parameterizations. The architectural result isolates the direct head from a compulsory unit-slope path; estimator accuracy is evaluated over the pretraining and target-task distributions.

C.5 Representation and transport theory

Let E denote raw multimodal evidence and $Z = r(E)$ the representation emitted by the adapters. For a registered functional Φ, write $\Phi(P; E)$ for its value under distribution P and raw evidence, and $\Phi(P; Z)$ for the same functional after replacing raw evidence by Z.

Theorem 4 (Adapter representation preservation). Suppose that, for every conditional kernel used by the registered functional Φ, the adapter preserves that kernel: for almost every context u,

$$P(V_i \mid u, E) = P(V_i \mid u, Z).$$

Suppose also that r respects the contract's temporal, measurement, and selection roles. Then $\Phi(P; E) = \Phi(P; Z)$.

Proof. Evaluate the functional AST in topological order. Its leaves are the observed kernels and constants; the kernel-preservation assumption makes corresponding leaves equal. Each internal AST operation (product, sum, integration, or marginalization) preserves equality of its operands. Induction over the finite AST therefore gives equality at the root. The role-preservation assumption ensures that the same admissible kernel is used at each leaf. The theorem supplies the sufficient preservation condition assessed by the adapter-sufficiency diagnostics in the empirical protocol.

Corollary 2 (Approximate adapter preservation). If corresponding adapter and raw-evidence kernels differ by ϵ_j in the norms of Theorem 2 rather than being exactly equal, then $|\Phi(P; E) - \Phi(P; Z)| \leq B^{m-1} \sum_{j=1}^{m} \epsilon_j$.

Proof. Apply Theorem 2 to the two collections of functional leaves.

Theorem 5 (Transported-effect error decomposition). Let P_S and P_T denote source and target domains. Define $\psi^*_{T,E}$ as the target effect using raw evidence, $\psi^*_{T,Z}$ as the target effect using adapter evidence, $\psi^*_{S,Z}$ as the source effect using adapter evidence, and $\widetilde{\psi}_{S,Z}$ as the population plug-in value of the learned nuisance functions. Let $\widehat{\psi}$ be the finite-sample estimate. Then

$$|\widehat{\psi} - \psi^*_{T,E}| \leq \underbrace{|\widehat{\psi} - \widetilde{\psi}_{S,Z}|}_{\epsilon_{\text{sample}}} + \underbrace{|\widetilde{\psi}_{S,Z} - \psi^*_{S,Z}|}_{\epsilon_{\text{functional}}} + \underbrace{|\psi^*_{S,Z} - \psi^*_{T,Z}|}_{\epsilon_{\text{transport}}} + \underbrace{|\psi^*_{T,Z} - \psi^*_{T,E}|}_{\epsilon_{\text{adapter}}}.$$

Proof. Add and subtract the three intermediate quantities $\widetilde{\psi}_{S,Z}$, $\psi^*_{S,Z}$, and $\psi^*_{T,Z}$, then apply the triangle inequality. The resulting upper-bound decomposition assigns diagnostics to four components: repeated-sample variation for ϵ_{sample}, nuisance and functional checks for $\epsilon_{\text{functional}}$, held-out hospitals for $\epsilon_{\text{transport}}$, and representation sufficiency/dropout checks for $\epsilon_{\text{adapter}}$.

Proposition 5 (Safe addition of an optional observed domain). Let O be an optional observed distribution. If the registered identifier does not depend on O, adding O to the contract and recompiling preserves the same semantic certificate after the expected contract rehash. If a registered partial-identification bound uses O, the feasible model set can only shrink, so the resulting bound interval cannot widen.

Proof. When O is absent from the identifier dependency set, the functional AST and status are unchanged; only the contract identity and its derived hash are reissued. For a bound, adding an observed distribution adds constraints to the feasible set Θ, giving $\Theta_{\text{new}} \subseteq \Theta_{\text{old}}$. Taking the infimum and supremum over a subset cannot expand the interval.

Proposition 6 (Relative completeness of refusal). Assume the registered identification procedure is complete for the supported graph class and that the verifier rejects exactly when no registered point functional is available from the declared distributions and assumptions. If the procedure returns a non-identification witness, the executor cannot return a point estimate; it returns a verified bound when one is registered, and otherwise a refusal or unsupported status.

Proof. Completeness supplies the absence of a point functional within the registered class. The verifier therefore rejects the point certificate. Proposition 1 applies to the resulting branch, so execution terminates on the typed non-point path. The statement covers the registered causal-query class.

C.6 Task-distribution generalization

Let $\mathcal{H}$ be the class of contract-conditioned foundation estimators and let $\ell(h; \xi) \in [0, B]$ be the loss on an authorized task $\xi \sim \Pi$. Write $R_\Pi(h) = \mathbb{E}_{\xi \sim \Pi} \ell(h; \xi)$ and $\widehat{R}_N(h) = N^{-1} \sum_{i=1}^{N} \ell(h; \xi_i)$. The empirical Rademacher complexity of the induced loss class is

$$\widehat{\mathfrak{R}}_N(\ell \circ \mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} N^{-1} \sum_{i=1}^{N} \sigma_i \ell(h; \xi_i) \right].$$

Write $\mathfrak{R}_N(\ell \circ \mathcal{H}) = \mathbb{E}_{\xi_{1:N}} \widehat{\mathfrak{R}}_N(\ell \circ \mathcal{H})$ for its expectation over the sampled tasks.

Theorem 6 (Authorized task-family generalization). Under (A5)–(A6), with probability at least $1 - \delta$ over N pretraining tasks, every $h \in \mathcal{H}$ satisfies

$$R_\Pi(h) \le \widehat{R}_N(h) + 2\mathfrak{R}_N(\ell \circ \mathcal{H}) + B\sqrt{\frac{\log(1/\delta)}{2N}}.$$

The statement applies to authorized tasks from Π; target mechanisms outside the support of the declared task distribution require a separate transport analysis, while rejected queries follow their typed non-point semantics.

Proof. Symmetrization bounds the expected supremum of $R_\Pi(h) - \widehat{R}_N(h)$ by twice the Rademacher complexity of the loss class. Changing one task changes the supremum by at most B/N; McDiarmid's inequality adds the displayed confidence term. Combining the two inequalities and taking the simultaneous supremum over h proves the result.

Corollary 3 (Foundation-to-adapter error ledger). Suppose the bounded task loss upper-bounds the authorized effect error in the same units as Theorems 2–5 and the target task is drawn from Π. Then the error of an authorized adapted estimator is bounded by the sum of (i) empirical task risk and the complexity terms in Theorem 6, (ii) registered nuisance-leaf error from Theorem 2, (iii) target sampling error from Theorem 3, and (iv) adapter and transport terms from Theorems 4–5. This additive ledger is used diagnostically and is stated without independence or tightness assumptions; mechanism-distribution shift requires a separate transport argument.

Proof. Apply Theorem 6 to the amortized task loss and use the triangle inequalities in Theorems 2, 3, and 5. Theorem 4 sets the adapter term to zero under exact kernel preservation; otherwise it remains an empirical remainder.

C.7 Complexity and auditability

Let n_V and n_E be the numbers of graph variables and edges, and let n_{AST} be the number of nodes in the functional AST. Sorting the canonical variable and edge lists costs $O(n_V \log n_V + n_E \log n_E)$; hash computation is linear in the canonical byte string; AST validation and evaluation are $O(n_{\mathrm{AST}})$ plus the cost of the declared density interface. The graph-separation primitive is linear in the graph representation for each registered query. Thus the certificate gate adds a deterministic symbolic cost before the multimodal encoder, whose cost is independent of the number of clinical rows in the refusal/bound branches. This separation is why the audit log can report gate latency and neural latency separately.

C.8 Scope of the guarantees

The guarantees are conditional on graph specification, empirical overlap, adapter preservation, and finite-sample calibration. CausaBridge records these obligations explicitly and evaluates them through the corresponding structural and empirical audits described in the main text.

D Structural component ablation

We replay the prespecified executor policy under five seeds while removing one authorization component at a time. False-point rate counts non-point tasks that incorrectly produce a numerical estimate; a non-point neural call records any invocation of the foundation core on a bound, refused, or unsupported branch. The complete system makes neither error. Removing the independent verifier or the empirical-overlap audit creates both false point outputs and unauthorized neural execution, whereas relaxing variable-role constraints yields the largest false-point rate. Table 3 therefore tests the causal control plane directly rather than inferring safety from predictive accuracy.

Table 3: Five-seed prespecified structural-policy replay. Lower is better for false-point rate and non-point neural calls; higher is better for refusal accuracy and semantic match.

Variant	False-point rate ↓	Refusal accuracy ↑	Semantic match ↑	Non-point neural calls ↓
Full CausaBridge	**0.000**	**1.000**	**1.000**	**0**
Compiler only (no verifier)	0.250	0.750	0.000	10
No empirical-overlap audit	0.250	0.750	1.000	10
Unsafe variable roles allowed	0.500	1.000	0.500	0

E TRAINING STAGES AND PROTOCOL MATRIX

Foundation scaling uses six paired lineages: fixed-mechanism and diverse-mechanism curricula for seeds 2027, 2028, and 2029. Each lineage saves nested checkpoints at 1,500, 7,500, and 30,000 optimizer steps. Architectures, optimizer settings, patient counts, update counts, and evaluation tasks are matched across curricula. Each checkpoint receives an affine calibration adapter fitted on 128 tasks. A disjoint 256-task panel is then shared by all 18 cells for confirmatory evaluation.

Medical adaptation starts from a certified foundation checkpoint and separates multimodal alignment, joint causal training, and calibration. The matrix crosses zero-shot, adapter-only, full-finetune, and from-scratch controls with target fractions 0.05 and 1.0 where supported. Every stage fingerprint binds the input-package hash, preprocessing hash, encoder revision, model configuration, seed, checkpoint parent, adapter mode, and target fraction.

The core input is a `FoundationInputBatch` containing observations, a causal contract, and a genuine verifier-issued point certificate. Known SCM effects, counterfactual labels, and identification targets are stored in a separate `CausalSupervision` object. The provider and trainer check matching task, query, and contract hashes before enabling effect losses. This makes the training-time oracle boundary auditable and prevents a clinical row from acquiring a counterfactual truth field merely because a semi-synthetic provider supports one.

Table 4: Certified foundation-transfer protocol. Δ is diverse-mechanism minus fixed-mechanism ATEE, so negative favors broader pretraining.

Protocol component	Value	Evidence status
Curricula	Fixed mechanism; diverse mechanism	Complete
Training seeds and budgets	3 seeds; 1.5k, 7.5k, and 30k steps	18 checkpoint cells
Calibration panel	128 tasks per checkpoint; seed disjoint from evaluation	Complete
Confirmatory panel	256 shared unseen tasks	4,608 evaluations
Paired inference	10,000-draw crossed seed–task bootstrap	Complete
30k calibrated ATEE	Fixed 0.1957; diverse 0.1777	$\Delta = -0.0180$
Paired diverse-minus-fixed Δ	-0.0180; 95% CI $[-0.0242, -0.0117]$; $p < 10^{-4}$	
Per-seed Δ	-0.0179, -0.0181, -0.0180	3/3 favor diverse
Diverse-curriculum slopes	-0.0187, -0.0097, -0.0163	3/3 negative
Prespecified confirmatory criteria	Pooled $\Delta < 0$; CI upper < 0; ≥ 2 seeds win; all diverse slopes < 0	4/4 passed
Foundation-transfer decision	**Supported under the prespecified task distribution**	

F DATA PROCESSING AND SPLIT POLICY

The MIMIC image–study–encounter package is an immutable Parquet manifest with separate image, study, and encounter-link tables plus audit, schema, and provenance metadata. The manifest is joined to MIMIC-IV by the declared encounter contract. Patient-level development splits are used for model selection, while the fixed test split is used only for final evaluation. The materializer records the landmark query, missingness representation, source hashes, row counts, and split hashes without exporting direct identifiers.

The real protocol disables clinical report text because temporal provenance is not sufficient for the declared landmark contract. Missing image and missing EHR observations are typed rather than silently imputed as observed values. The MIMIC materializer is isolated from the public tabular benchmarks, preventing public benchmark repetitions or outcomes from augmenting the clinical cohort.

G EVALUATION DEFINITIONS

For known-truth semi-synthetic tasks, we use

$$\text{ATE error} = |\widehat{\tau}_{\text{ATE}} - \tau_{\text{ATE}}|, \tag{10}$$

$$\text{PEHE} = \sqrt{\mathbb{E}_x[(\widehat{\tau}_{\text{CATE}}(x) - \tau_{\text{CATE}}(x))^2]}, \tag{11}$$

$$\text{false-ID} = \frac{\#\{\text{non-point tasks with point output}\}}{\#\{\text{non-point tasks}\}}. \tag{12}$$

Coverage is reported for intervals with a declared target and separately from the verified Manski identification interval. For clinical data, authorized numerical metrics are reported whenever an admissible reference target is available; status correctness, overlap, calibration, transport gap, typed non-point rates, and provenance completeness are reported for all applicable queries.

H MIMIC MULTI-SEED EVALUATION

The evaluation follows a prespecified multi-seed protocol and is reported as descriptive evidence across independently trained seeds. It contains 498 result rows and 16,146 patient-level prediction rows: 378 point outcomes, 88 typed refusals, 32 unsupported outcomes, and no failed rows. Result, prediction, environment, and three shard-completion hashes agree with the evaluation manifest. The main semi-synthetic table filters to `task_kind=semisynthetic`; adapter rows are summarized only in the adapter matrix and are never pooled into the main method average.

Table 5: Treatment-support audit for the linked real-clinical MIMIC cohort. The empirical gate returns a typed refusal under the prespecified minimum arm size.

Cohort component	Patients	Treated	Control	Decision
All linked patients	172	2	170	Empirical-support audit
Principal source environment	159	2	157	Typed refusal triggered
Other source environments	13	0	13	Typed refusal triggered
Prespecified query batch, seed 2027	128	1	127	Typed refusal triggered
Prespecified query batch, seed 2028	128	2	126	Typed refusal triggered
Minimum-arm requirement	–	At least 10 per arm		Point execution not authorized
CausaBridge output	Fail-closed non-point output			**TYPED REFUSAL**

Table 6 reports the principal method separately for each evaluated training seed. Numerical errors are computed only on the nine authorized semi-synthetic scenarios. The treatment counts and decisions in the last three columns refer to the separate real-clinical query branch. The clinical branch does not inherit semi-synthetic truth fields and emits no point estimate.

Table 6: Seed-specific MIMIC evidence under the prespecified multi-seed protocol. ATE MAE and root PEHE summarize the nine authorized semi-synthetic scenarios; treatment counts and the final column summarize the distinct real-clinical support audit.

Seed	ATE MAE ↓	Root PEHE ↓	Treated	Control	Clinical branch
2027	0.1319	0.1756	1	127	REFUSED
2028	0.1387	0.1757	2	126	REFUSED

Table 7 gives the complete adaptation matrix. The adapter-only model updates 2,978,816 of 51,362,823 parameters and preserves the frozen-core hash in every evaluated cell. Its paired ATE-MAE ratio relative to full fine-tuning is 0.5046, below the prespecified noninferiority margin of 1.2. Full fine-tuning and from-scratch training update all parameters; zero-shot updates none.

Table 7: Complete MIMIC adaptation matrix over nine authorized semi-synthetic scenarios under the prespecified multi-seed protocol. "Trainable params." is the fraction of the 51.36M-parameter system updated for the target task.

Adaptation mode	Target fraction	Trainable params.	ATE MAE ↓	Root PEHE ↓
Zero-shot	5%	0.0%	0.1295	0.1781
Zero-shot	100%	0.0%	0.1506	0.1923
Adapter-only	5%	5.8%	0.1352	0.1851
Adapter-only	100%	5.8%	0.1353	0.1756
Full fine-tune	5%	100.0%	0.2681	0.2998
Full fine-tune	100%	100.0%	0.2682	0.3057
From scratch	5%	100.0%	0.1302	0.1716
From scratch	100%	100.0%	**0.1214**	**0.1691**

The protocol also records diagnostics that are not collapsed into the headline error averages (Table 8). Paired effect-scale and modality-sensitivity values are reported over the fixed scenario panel. These diagnostics keep representation behavior separate from authorization and do not alter structural or empirical-support decisions.

Table 8: MIMIC protocol diagnostics retained under the prespecified multi-seed evaluation. The modality statistic is full-input minus no-image ATE MAE.

Diagnostic	Observed value	Protocol interpretation
Effect-scale direction	Consistent across evaluated seeds	Direction recovered
Effect-scale response ratio	0.0623	Paired low–high scale sensitivity
Full minus no-image ATE MAE, image-focused tasks	$+0.0109$	Full input has higher ATE MAE
Prediction diversity	9/9 unique within each evaluated seed	No scenario-level collapse
Real-clinical support	1/127; 2/126	Typed refusal across evaluated seeds

Because the MIMIC evaluation is conducted across independently trained seeds, cross-seed means and seed-specific values are the primary report. Bootstrap intervals and multiple-comparison-adjusted values remain in the evaluation records for audit, but are not promoted to confirmatory claims in the manuscript. The public benchmark and common-target experiments retain their separately prespecified three-seed protocols.

Calibrated common-target foundation test. For each checkpoint, an affine map $\widehat{\tau}^{\mathrm{cal}} = a + b\widehat{\tau}$ is fitted by ridge regression on 128 calibration tasks generated with seed 6,000,000. The calibration panel is disjoint from training and from the confirmatory panel. A generator with seed 7,000,000 creates 256 confirmatory tasks indexed 0–255. The runner requires the same ordered task hashes and true ATEs in every curriculum, training seed, and checkpoint cell. The confirmatory matrix has $2 \times 3 \times 3 \times 256 = 4{,}608$ rows. At 30,000 steps define $d_{st} = |\widehat{\tau}_{st}^{\mathrm{diverse}} - \tau_t| - |\widehat{\tau}_{st}^{\mathrm{fixed}} - \tau_t|$. Each of 10,000 crossed-bootstrap draws independently resamples three training seeds and 256 tasks before averaging the Cartesian submatrix. The prespecified decision rule requires $\bar{d} < 0$, a negative 95% interval upper endpoint, at least two negative seed means, and a negative error slope across budgets for all diverse-mechanism seeds.

I Comparator scope and complete public-benchmark results

The complete matrix below is regenerated on the exact 309 seed-indexed tasks for which the authorized CausaBridge estimator returned a point result. The candidate evaluation contains 312 tasks in total; all three non-point outcomes are overlap-based refusals for ACIC repetition 9. Comparator predictions come from the complete 5,376-job comparison but are filtered to the same authorized keys before averaging. Thus accuracy is conditional on reported coverage, rather than treating refused tasks as missing numerical errors.

Table 9: Scope comparison. "Runtime verification" means that the current task's declared query and interfaces are checked independently before point execution; graph validity remains an explicit analyst-declared contract condition.

Method	Amortized across tasks	Explicit causal setting	Runtime verification	Typed non-point output	Plug-in multimodal adapter
CInA (Zhang et al., 2024)	Yes	Balancing assumptions	–	–	–
CausalFM (Ma et al., 2026)	Yes	BD/FD/IV prior	–	–	–
CausalPFN (Balazadeh et al., 2025)	Yes	Ignorability prior	–	–	–
Do-PFN (Robertson et al., 2025)	Yes	Learned SCM prior	–	–	–
CausaBridge	Yes	Typed task contract	Yes	Verified bound / typed refusal	Yes

"–" denotes an execution interface outside the published method specification.

Table 10: Complete matched-support public-benchmark means. Point-risk values use the exact keys on which CausaBridge + SEDRA is authorized; coverage is reported in the main text. Jobs reports clipped-IPW policy value.

Dataset	Method	ATEE ↓	PEHE ↓	Policy value ↑
ACIC	AIPW	0.5538	3.2869	–
ACIC	CausaBridge + SEDRA	0.0634	1.1284	–
ACIC	Causal forest	0.1458	1.3839	–
ACIC	CausalFM	1.5760	3.5137	–
ACIC	CausalPFN	0.1375	1.2201	–
ACIC	CFRNet	0.3756	3.7246	–
ACIC	Difference in means	1.0611	3.4605	–
ACIC	Do-PFN	2.7166	4.2444	–
ACIC	Dragonnet	0.4025	3.2831	–
ACIC	Linear DR-learner	0.5777	2.6867	–
ACIC	Linear S-learner	0.5245	2.6405	–
ACIC	Linear T-learner	0.5379	2.6433	–
ACIC	Linear X-learner	0.5361	2.6376	–
ACIC	RF DR-learner	0.4810	2.0405	–
ACIC	RF X-learner	0.3483	2.1247	–
ACIC	TARNet	0.2637	3.0663	–
IHDP	AIPW	0.6568	5.9238	–
IHDP	CausaBridge + SEDRA	0.1139	2.4783	–
IHDP	Causal forest	0.5286	3.6291	–
IHDP	CausalFM	2.3790	6.0222	–
IHDP	CausalPFN	0.2866	0.5938	–
IHDP	CFRNet	0.2759	2.1800	–
IHDP	Difference in means	0.7492	5.9357	–
IHDP	Do-PFN	2.5270	6.8335	–
IHDP	Dragonnet	0.9037	2.6820	–
IHDP	Linear DR-learner	0.2935	2.5518	–
IHDP	Linear S-learner	0.2979	2.5546	–
IHDP	Linear T-learner	0.2789	2.4957	–
IHDP	Linear X-learner	0.2831	2.5032	–
IHDP	RF DR-learner	0.4282	3.2275	–
IHDP	RF X-learner	0.5037	4.4364	–
IHDP	TARNet	0.2751	2.1701	–
Twins	AIPW	0.0194	0.3297	–
Twins	CausaBridge + SEDRA	0.0092	0.3178	–
Twins	Causal forest	0.0197	0.3196	–
Twins	CausalFM	0.0484	0.3211	–
Twins	CausalPFN	0.0117	0.3214	–
Twins	CFRNet	0.0210	0.5250	–
Twins	Difference in means	0.0139	0.3175	–
Twins	Do-PFN	0.1862	0.4233	–
Twins	Dragonnet	0.0210	0.4957	–
Twins	Linear DR-learner	0.0201	0.3219	–
Twins	Linear S-learner	0.0200	0.3219	–
Twins	Linear T-learner	0.0200	0.3219	–
Twins	Linear X-learner	0.0200	0.3219	–

Continued on next page

Dataset	Method	ATEE ↓	PEHE ↓	Policy value ↑
Twins	RF DR-learner	0.0198	0.3225	–
Twins	RF X-learner	0.0208	0.3197	–
Twins	TARNet	0.0280	0.5549	–
Jobs	AIPW	–	–	6090.0820
Jobs	CausaBridge + SEDRA	–	–	5632.9641
Jobs	Causal forest	–	–	5829.6758
Jobs	CausalFM	–	–	5927.4360
Jobs	CausalPFN	–	–	6116.2020
Jobs	CFRNet	–	–	5246.6502
Jobs	Difference in means	–	–	6090.0820
Jobs	Do-PFN	–	–	5214.7543
Jobs	Dragonnet	–	–	47489.5907
Jobs	Linear DR-learner	–	–	4286.8027
Jobs	Linear S-learner	–	–	4692.6779
Jobs	Linear T-learner	–	–	4692.6779
Jobs	Linear X-learner	–	–	4692.6779
Jobs	RF DR-learner	–	–	6508.5367
Jobs	RF X-learner	–	–	6045.8957
Jobs	TARNet	–	–	5409.6079

J SEDRA SAFETY AND ATTRIBUTION ANALYSIS

Table 11 separates the cross-fitted doubly robust (DR) anchor from the pretrained residual and from an otherwise matched from-scratch residual. Across all 309 authorized public-benchmark tasks, the mean-preserving correction changes the mean CATE by at most 3.55×10^{-15}, numerically preserving the DR ATE anchor. The selected residual weight α has median 0.0113, mean 0.0189, and maximum 0.1763. Consequently, the complete estimator remains close to the anchor when the learned correction is not supported, while retaining a nonzero individualized correction on supported tasks. This analysis localizes the contribution of SEDRA to mean-preserving individualization on top of a safe DR anchor and directly supports the claimed anchor-preservation and fallback properties.

Table 11: SEDRA component analysis on the 309 authorized public-benchmark tasks. Lower is better. Equal ATE errors for the DR anchor, from-scratch SEDRA, and CausaBridge+SEDRA follow from the mean-preserving construction; differences appear in individual-effect error.

	ACIC (6 repetitions)		IHDP (95 repetitions)	
Estimator component	ATE MAE ↓	Root PEHE ↓	ATE MAE ↓	Root PEHE ↓
DR anchor	**0.0634**	**1.1267**	**0.1139**	2.4742
Foundation residual only	0.7522	10.1785	2.9389	12.7964
From-scratch SEDRA	**0.0634**	1.1271	**0.1139**	**2.4622**
CausaBridge + SEDRA	**0.0634**	1.1284	**0.1139**	2.4783

K IMPLEMENTATION AND ENVIRONMENT

The formal configuration records Python 3.11, four 48-GB A6000 devices, bfloat16 image encoding, fixed encoder revisions, batches, seeds, budgets, adapters, and data fractions. Preflight and final records bind input, split, environment, source, configuration, checkpoint, and gate hashes; only passed runs enter scientific aggregation.