Improving RCT-Based Treatment Effect Estimation Under Covariate Mismatch via Calibrated Alignment

Amir Asiaee[1] **Samhita Pal**[1]

[1]Department of Biostatistics, Vanderbilt University Medical Center, TN 37203, USA

Abstract

Randomized controlled trials (RCTs) are the gold standard for estimating treatment effects, yet they are often underpowered for detecting effect heterogeneity. Large observational studies (OS) can supplement RCTs for conditional average treatment effect (CATE) estimation, but a key barrier is *covariate mismatch*: the two sources measure different, only partially overlapping, covariates. We propose CALM (**C**alibrated **AL**ignment under covariate **M**ismatch), which learns embeddings that map each source's features into a common representation space. OS outcome models are transferred to the RCT embedding space and calibrated using trial data, preserving causal identification from randomization. Finite-sample risk bounds decompose into *alignment error*, *outcome-model complexity*, and *calibration complexity* terms, making explicit when the learned embedding is accurate enough to reduce variance. We instantiate CALM in two forms: a closed-form linear version, CALM-Lin, and a neural representation-learning version, CALM-NN. Across 51 simulation settings, calibration-based linear methods are effectively tied in linear-CATE regimes, while CALM-NN wins all 22 nonlinear-CATE settings by wide margins. Moreover, on two real-data studies CALM-NN delivers the largest gains over the trial-only baseline.

1 INTRODUCTION

Heterogeneous treatment effects (HTEs) are central to precision medicine: understanding how individual patient characteristics modulate treatment response helps clinicians match interventions to patients [Kosorok and Laber, 2019]. RCTs remain the gold standard for causal inference, providing unconfounded treatment comparisons that support consistent estimation of the conditional average treatment effect,

$$\tau^r(\boldsymbol{x}) = \mathbb{E}\big[Y(1) - Y(-1) \mid \boldsymbol{X} = \boldsymbol{x},\, S = r\big],$$

where $Y(a)$ denotes the potential outcome under treatment $a \in \{-1, 1\}$ and $S = r$ indicates the RCT population. In practice, however, most RCTs are powered for average effects rather than the fine-grained heterogeneity that guides individualized decisions [Wang et al., 2007, Kent et al., 2020]. At the same time, large-scale observational studies—electronic health records (EHRs), registries, and insurance claims—provide rich covariates and large sample sizes, but are vulnerable to confounding.

This contrast motivates borrowing from OS data to improve the precision of within-trial CATE estimation. Asiaee et al. [2025] formalize this $\mathcal{O} \to \mathcal{R}$ direction by introducing the *counterfactual mean outcome* (CMO) as the variance-minimizing augmentation function for pseudo-outcome regression, and propose R-OSCAR, a two-stage calibration procedure that learns outcome models from the OS, calibrates them to the RCT, and uses the calibrated predictions as an augmentation function. Thus, OS data are used to improve the CMO augmentation rather than to identify the treatment effect: better CMO estimates make the pseudo-outcome regression more efficient, while the CATE target remains anchored in the randomized trial.

The covariate mismatch barrier. A major obstacle to RCT–OS integration is that the covariates collected in trials and observational sources rarely align—a problem we call *covariate mismatch*. The RCT may contain behavioral or survey-based variables absent from the OS, while the OS may contain laboratory or utilization features not recorded in the trial [Rässler, 2002]. This differs fundamentally from *covariate shift* [Sugiyama and Kawanabe, 2012], where the same variables are observed but follow different distributions. Under mismatch, the outcome models trained on OS features cannot be directly evaluated on RCT units, so the transfer step in data-borrowing methods is no longer well defined. Indeed, for generalization ($\mathcal{R} \to \mathcal{O}$), Colnet et al.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:185–209.

[2022] show that oracle linear imputation of a key covariate observed in only one source does not remove the missing-covariate bias, even in a Gaussian linear-CATE setting.

Pal et al. [2026] adapt R-OSCAR to covariate mismatch by partitioning covariates into shared ($\boldsymbol{Z} \in \mathbb{R}^{p_z}$), RCT-only ($\boldsymbol{U} \in \mathbb{R}^{p_u}$), and OS-only ($\boldsymbol{V} \in \mathbb{R}^{p_v}$) blocks. Their imputation-based estimator, MR-OSCAR, learns a map from $\boldsymbol{Z}$ to $\boldsymbol{V}$ in the OS and fills in the structurally missing $\boldsymbol{V}$ block in the RCT before applying R-OSCAR. They also define SR-OSCAR (shared R-OSCAR), which applies R-OSCAR using only the shared covariates $\boldsymbol{Z}$.

Imputation is a harder problem than needed. The imputation approach requires reconstructing the entire $\boldsymbol{V}$ vector in the RCT—a potentially high-dimensional regression problem whose difficulty scales with p_v and the complexity of $P(\boldsymbol{V} \mid \boldsymbol{Z})$. For CATE estimation, however, the target is only a representation that is sufficient for (i) predicting outcomes and (ii) estimating discrepancy functions connecting outcome means across the two populations. When the outcome-relevant information in $\boldsymbol{V}$ lies on a low-dimensional manifold, the imputation approach pays an unnecessarily high price by attempting to recover the full $\boldsymbol{V}$ rather than its outcome-relevant projection.

Our contribution: embedding alignment. Our proposed **C**alibrated **AL**ignment under covariate **M**ismatch (CALM) replaces imputation with *representation alignment*. Instead of reconstructing missing covariates, we learn embedding functions $\phi^o : \mathbb{R}^{p_o} \to \mathbb{R}^d$ and $\phi^r : \mathbb{R}^{p_r} \to \mathbb{R}^d$ that map the heterogeneous feature spaces into a common d-dimensional representation space $\mathcal{H}$. Outcome models are trained in the OS embedding space and then transferred to the RCT embedding space, where they are calibrated and used for CATE estimation through the double-calibration pipeline of R-OSCAR. OS data only inform the pseudo-outcome used to reduce variance of the CATE estimate (see Figure 1 for an overview); causal identification rests entirely on randomization within the trial and remains intact.

Our contributions are:

1. **Embedding-alignment framework.** We introduce CALM, which integrates representation learning into the R-OSCAR calibration pipeline, replacing imputation with learned embeddings. The framework inherits the negative-transfer protection from double calibration.
2. **Finite-sample risk bounds.** We derive non-asymptotic bounds that decompose into alignment error, outcome-model, calibration, and CATE-class complexity terms, identifying conditions under which embedding outperforms imputation. In linear Gaussian models, CALM-Lin recovers oracle linear imputation as a special case.
3. **Extensive experiments.** We evaluate linear and neural-network instantiations across 51 simulation settings, a semi-synthetic benchmark using real trial covariates, and two real-data studies (the Greenlight Plus trial linked to an EHR cohort, and a semi-real Tennessee STAR analysis).

Relation to R-OSCAR and MR-OSCAR. R-OSCAR [Asiaee et al., 2025] borrows from observational data when both sources are measured in the same covariate space. With covariate mismatch, MR-OSCAR [Pal et al., 2026] first imputes the covariates missing from the trial and then applies this borrowing pipeline. CALM instead applies the same idea in a learned common embedding, avoiding full reconstruction of the missing covariates. Our theory and experiments show when this embedding route is preferable to imputation or to using only the shared covariates: when the embedding keeps outcome-relevant signal while reducing dimension and aligning the two sources well.

2 RELATED WORK

CATE estimation has been studied through meta-learners [Künzel et al., 2019, Kennedy, 2023, Nie and Wager, 2021], causal forests [Athey et al., 2019], Bayesian methods [Hahn et al., 2020], and representation-learning approaches [Johansson et al., 2016, Shalit et al., 2017, Yao et al., 2018, Shi et al., 2019, Hassanpour and Greiner, 2020]. Most RCT–OS integration work targets the $\mathcal{R} \to \mathcal{O}$ (generalizability) direction [Colnet et al., 2024, Degtiar and Rose, 2023, Dahabreh et al., 2020]; recent work in that direction also models outcome-mean shifts directly rather than only covariate shifts [Asiaee et al., 2026a,b]. In the opposite $\mathcal{O} \to \mathcal{R}$ direction, Cheng and Cai [2021] adaptively weight CATE estimates from both sources, Oberst et al. [2022] study optimal biased–unbiased combinations, and Karlsson et al. [2025] provide worst-case guarantees. The R-OSCAR framework of Asiaee et al. [2025] introduces double calibration for borrowing from OS data in a shared covariate space, while Pal et al. [2026] extend this route to covariate mismatch through imputation (MR-OSCAR) and shared-covariate borrowing (SR-OSCAR). We use these as baselines and introduce CALM, which borrows in a learned embedding rather than by imputing missing covariates. In heterogeneous transfer learning, standard domain-adaptation methods [Ben-David et al., 2010, Ganin et al., 2016, Long et al., 2015, Pan and Yang, 2010, Weiss et al., 2016] assume shared features. Cross-domain methods instead use CCA variants [Hardoon et al., 2004, Andrew et al., 2013] or shared-private encoders [Bousmalis et al., 2016]. In the causal setting, Bica and van der Schaar [2022] propose HTCE-learners, which require unconfoundedness in both domains, lack calibration-based negative-transfer protection, and provide no finite-sample guarantees. Sufficient dimension reduction [Cook, 2007, Li, 2018, Ma and Zhu, 2012, Ghosh et al., 2021] also targets outcome-relevant projections but does not handle cross-dataset covariate mismatch. Bayesian borrowing via power priors [Chen and Ibrahim, 2000] and commensurate

priors [Hobbs et al., 2011] discounts external data through a power parameter; a Bayesian, bias-limited version of this borrowing for the same covariate-mismatch setting is developed in [Asiaee and Pal, 2026], and we instead use a frequentist setting with calibration-based correction. Negative transfer is a well-documented risk [Rosenstein et al., 2005]; recent causal-inference work addresses it through optimal estimator combinations [Oberst et al., 2022], worst-case guarantees [Karlsson et al., 2025], and calibration [Asiaee et al., 2025]. CALM inherits this calibration-based protection and adds representation alignment.

3 PROBLEM SETUP AND BACKGROUND

3.1 NOTATION AND DATA STRUCTURE

There are two data sources: a randomized controlled trial ($S = r$) and a large observational study ($S = o$). Let Y denote the outcome of interest, $A \in \{-1, 1\}$ the binary treatment indicator, and $Y(a)$ the potential outcome under treatment a. We assume consistency: $Y = Y(A)$.

The two sources measure different subsets of baseline covariates: the RCT records $\boldsymbol{X}^r$ and the OS records $\boldsymbol{X}^o$. We denote the covariates measured in both sources by $\boldsymbol{Z} \in \mathbb{R}^{p_z}$ (shared), those exclusive to the RCT by $\boldsymbol{U} := \boldsymbol{X}^r \backslash \boldsymbol{Z} \in \mathbb{R}^{p_u}$, and those exclusive to the OS by $\boldsymbol{V} := \boldsymbol{X}^o \backslash \boldsymbol{Z} \in \mathbb{R}^{p_v}$, so that $\boldsymbol{X}^r = (\boldsymbol{U}, \boldsymbol{Z}) \in \mathbb{R}^{p_r}$ and $\boldsymbol{X}^o = (\boldsymbol{Z}, \boldsymbol{V}) \in \mathbb{R}^{p_o}$, where $p_r = p_u + p_z$ and $p_o = p_z + p_v$. We write $\boldsymbol{X} = (\boldsymbol{U}, \boldsymbol{Z}, \boldsymbol{V}) \in \mathbb{R}^p$ with $p = p_u + p_z + p_v$ for the complete covariate vector, and use lowercase $\boldsymbol{x}^r = (\boldsymbol{u}, \boldsymbol{z})$, $\boldsymbol{x}^o = (\boldsymbol{z}, \boldsymbol{v})$ for generic realizations.

The observed data are $\{(\boldsymbol{X}_i^r, A_i, Y_i)\}_{i=1}^{n^r}$ from the RCT and $\{(\boldsymbol{X}_j^o, A_j, Y_j)\}_{j=1}^{n^o}$ from the OS, with $n^r \ll n^o$ in the typical regime of interest. Let n_a^s denote the number of units in arm a under source s. The RCT propensity is $\pi_a^r(\boldsymbol{x}^r) = \mathbb{P}(A = a \mid \boldsymbol{X}^r = \boldsymbol{x}^r, S = r)$. For each source $s \in \{o, r\}$, write

$$\mu_a^s(\boldsymbol{x}^s) = \mathbb{E}[Y \mid \boldsymbol{X}^s = \boldsymbol{x}^s, A = a, S = s].$$

In the RCT, randomization implies $\mu_a^r(\boldsymbol{x}^r) = \mathbb{E}[Y(a) \mid \boldsymbol{X}^r = \boldsymbol{x}^r, S = r]$. Our inferential target is the CATE in the RCT population, marginalized over the unobserved $\boldsymbol{V}$:

$$\tau^r(\boldsymbol{x}^r) = \mu_1^r(\boldsymbol{x}^r) - \mu_{-1}^r(\boldsymbol{x}^r). \tag{1}$$

3.2 THE CMO-AUGMENTED PSEUDO-OUTCOME FRAMEWORK

Following Asiaee et al. [2025], we construct pseudo-outcomes using an augmentation function $m : \mathbb{R}^{p_r} \to \mathbb{R}$:

$$\tau_m(\boldsymbol{X}^r, A, Y) = \frac{A\big(Y - m(\boldsymbol{X}^r)\big)}{\pi_A^r(\boldsymbol{X}^r)}. \tag{2}$$

Under the standard RCT identification assumptions (SUTVA, ignorability, positivity), $\mathbb{E}[\tau_m \mid \boldsymbol{X}^r, S = r] = \tau^r(\boldsymbol{X}^r)$ for *any* m, so the pseudo-outcome has the CATE as its conditional mean. The CATE is then estimated by minimizing the empirical squared loss over a class of candidate functions $\mathcal{F}$:

$$\hat{\tau}^r \in \arg\min_{f \in \mathcal{F}} \frac{1}{n^r} \sum_{i:S_i=r} \big(\tau_m(\boldsymbol{X}_i^r, A_i, Y_i) - f(\boldsymbol{X}_i^r)\big)^2. \tag{3}$$

A key result of Asiaee et al. [2025] is that the augmentation m controls the CATE estimation risk. They show that the prediction risk decomposes as [Asiaee et al., 2025]:

$$R_m(\hat{\tau}) = \underbrace{\mathbb{E}_{\boldsymbol{X}^r}\big[\mathrm{Var}(\tau_m \mid \boldsymbol{X}^r)\big]}_{\text{irreducible error}} + \underbrace{\Delta_2^2(\hat{\tau}, \tau^r)}_{\text{CATE estimation error}}, \tag{4}$$

where $\Delta_2^2(f, g) = \mathbb{E}[(f(\boldsymbol{X}^r) - g(\boldsymbol{X}^r))^2 \mid S = r]$ is the mean squared error (MSE). They show that both terms are controlled by $\Delta_2^2(m, \tilde{\mu}^r)$, where $\tilde{\mu}^r$ is the *counterfactual mean outcome*:

$$\tilde{\mu}^r(\boldsymbol{x}^r) = \textstyle\sum_a \pi_{-a}^r(\boldsymbol{x}^r)\, \mu_a^r(\boldsymbol{x}^r). \tag{5}$$

In other words, the best strategy is to accurately estimate the CMO $\tilde{\mu}^r$, use it as the augmentation function m in (2), and then optimize the objective in (3). This motivates OS borrowing specifically for improved CMO estimation.

3.3 THE R-OSCAR PIPELINE AND IMPUTATION-BASED BASELINES

Here, we review the R-OSCAR pipeline [Asiaee et al., 2025] and the mismatch baselines of Pal et al. [2026].

R-OSCAR without covariate mismatch. When there is no covariate mismatch (i.e., $p_u = p_v = 0$), the R-OSCAR pipeline has three steps:

1. **OS outcome model:** For each arm a, learn the regression function $\hat{\mu}_a^o(\boldsymbol{x})$ from OS data.
2. **Outcome calibration:** Estimate a discrepancy function $\hat{\delta}_a^t(\boldsymbol{x})$ from RCT data so that $\hat{\mu}_a^o(\boldsymbol{x}) + \hat{\delta}_a^t(\boldsymbol{x}) \approx \mu_a^r(\boldsymbol{x})$. The calibrated CMO is $\hat{m}(\boldsymbol{x}) = \sum_a \pi_{-a}^r(\boldsymbol{x})[\hat{\mu}_a^o(\boldsymbol{x}) + \hat{\delta}_a^t(\boldsymbol{x})]$.
3. **CATE calibration:** Using $\hat{m}$ to form pseudo-outcomes in (2) and estimate a CATE correction $\hat{\delta}(\boldsymbol{x})$ in (3), yielding $\hat{\tau}_{\text{R-OSCAR}}(\boldsymbol{x}) = \sum_a a[\hat{\mu}_a^o(\boldsymbol{x}) + \hat{\delta}_a^t(\boldsymbol{x})] + \hat{\delta}(\boldsymbol{x})$.

The two-stage design decouples nuisance estimation from CATE estimation: misspecification in outcome models affects only the augmentation quality (variance), while CATE consistency depends only on correct specification of the CATE discrepancy $\hat{\delta}$.

Under covariate mismatch, we compare against the two baselines of Pal et al. [2026]. MR-OSCAR first imputes the

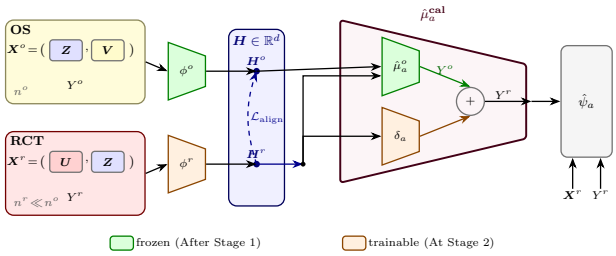


Figure 1: How OS data inform pseudo-outcome construction in CALM. OS and RCT data pass through source-specific encoders ϕ^o (frozen after Stage 1) and ϕ^r (trainable at Stage 2) into a shared embedding space $\boldsymbol{H} \in \mathbb{R}^d$. The calibrated outcome model $\hat{\mu}_a^r = \hat{\mu}_a^o + \delta_a^t$ combines the frozen OS outcome head with a learnable shift. Pseudo-outcomes $\hat{\psi}_a^r$ are then constructed from these calibrated predictions together with the RCT covariates and outcomes. The calibrated model $\hat{\mu}_a^r$ is also used in the subsequent CATE calibration stage (Stage 4 of Algorithm 1).

OS-only block by estimating $\hat{g} : \mathbb{R}^{p_z} \to \mathbb{R}^{p_v}$ in the OS, sets $\widehat{\boldsymbol{V}}_i = \hat{g}(\boldsymbol{Z}_i)$ for RCT units, and then runs R-OSCAR on $(\boldsymbol{X}^r, \widehat{\boldsymbol{V}})$. SR-OSCAR (shared R-OSCAR) skips imputation and runs R-OSCAR only on the shared covariates $\boldsymbol{Z}$. The relevant feature of the MR-OSCAR bound is that it pays both an imputation-error term in the missing block and outcome-model complexity in the full imputed covariate space; the comparison in Section 5 shows how CALM replaces these costs with alignment and sufficiency terms in a lower-dimensional embedding.

4 CALM: EMBEDDING-ALIGNED CATE ESTIMATION

This section presents CALM, our method for addressing covariate mismatch through representation alignment rather than imputation.

4.1 CORE IDEA: FROM IMPUTATION TO ALIGNMENT

The R-OSCAR pipeline requires not the raw covariates but rather *a common representation* in which (i) outcome models can be meaningfully evaluated for both OS and RCT units, and (ii) discrepancy functions can be estimated. Imputation constructs such a representation by reconstructing the missing $\boldsymbol{V}$ block, but this intermediate step solves a harder problem than necessary.

CALM instead learns embedding functions $\phi^o : \mathbb{R}^{p_o} \to \mathbb{R}^d$ (OS encoder) and $\phi^r : \mathbb{R}^{p_r} \to \mathbb{R}^d$ (RCT encoder) that map each source's full feature vector into a shared d-dimensional space $\mathcal{H}$. The OS outcome model is trained on the OS embeddings $\hat{\mu}_a^o(\phi^o(\boldsymbol{X}^o))$ and is evaluated on the RCT embeddings $\hat{\mu}_a^o(\phi^r(\boldsymbol{X}^r))$ during calibration. What makes this work is that the embeddings are outcome-supervised, not merely geometric. Stage 1 trains the OS encoder ϕ^o to predict the outcome from the *full* $\boldsymbol{X}^o = (\boldsymbol{Z}, \boldsymbol{V})$, so ϕ^o is forced to keep the outcome-relevant variation in $\boldsymbol{V}$ rather than just compress $\boldsymbol{Z}$. Stage 2 then aligns the RCT encoder ϕ^r to that outcome-aware target under a $\boldsymbol{Z}$-conditioned penalty. The operative notion of closeness is "produces similar outcome predictions on units with similar shared covariates," not "is geometrically close on the raw shared covariates": the latter would be satisfied by a degenerate $\boldsymbol{Z}$-only encoder, whereas the former would not, since the OS embedding carries $\boldsymbol{V}$-relevant signal that ϕ^r must reproduce. Under such an alignment, the OS outcome model transfers usefully to the RCT embeddings, and the calibration step corrects the residual discrepancy.

4.2 ASSUMPTIONS

We assume the standard internal validity of the RCT:

Assumption 1 (Internal validity of the RCT)**.** SUTVA holds; $(Y(1), Y(-1)) \perp\!\!\!\perp A \mid (\boldsymbol{X}^r, S = r)$; and there exists $\rho > 0$ such that $\rho \le \pi_a^r(\boldsymbol{X}^r) \le 1 - \rho$ almost surely for each a.

Unlike imputation-based borrowing, we need an embedding-level assumption:

Assumption 2 (Outcome-sufficient embedding)**.** For fixed encoders ϕ^s, the per-arm outcome means are well-approximated by functions of the embeddings: for each source $s \in \{o, r\}$ and arm a, there exists $\tilde{\mu}_a^s : \mathbb{R}^d \to \mathbb{R}$ in a function class $\mathcal{M}_a^s$ such that

$$\mathbb{E}\big[\big(\mu_a^s(\boldsymbol{X}^s) - \tilde{\mu}_a^s(\phi^s(\boldsymbol{X}^s))\big)^2 \mid S = s\big] \le \epsilon^2_{\text{suff}},$$

where ϵ^2_{suff} is the *sufficiency gap*.

This is weaker than requiring the embedding to be a sufficient statistic for the outcome: it permits an approximation error ϵ^2_{suff} that vanishes as d grows. This outcome-sufficient view has a clear lineage. Embeddings that keep only outcome-relevant directions go back to classical sufficient dimension reduction [Li, 1991, Cook, 2007] and its kernel extensions [Fukumizu et al., 2009], and they reappear in causal representation learning [Johansson et al., 2016, Shalit et al., 2017]. Assumption 2 departs from this line in two ways. It is stated per source—a separate condition on the OS and on the trial marginal—rather than as a single requirement on the joint distribution. And it asks the embedding to preserve the conditional outcome mean directly, rather than to balance the treatment groups as in the representation-learning approaches to CATE estimation.

Assumption 3 (Outcome shift in the embedding space)**.** For each arm a, there exists a discrepancy function $\delta_a : \mathbb{R}^d \to \mathbb{R}$ in a function class $\mathcal{D}_a$ of controlled complexity such that

$$\tilde{\mu}_a^r(\boldsymbol{h}) = \tilde{\mu}_a^o(\boldsymbol{h}) + \delta_a(\boldsymbol{h}),$$

where $\tilde{\mu}_a^s$ are the population-level outcome functions operating on the embedding $\boldsymbol{h} \in \mathbb{R}^d$.

This is the embedding-space analogue of the outcome-shift assumption in the R-OSCAR framework [Asiaee et al., 2025]. Outcome means and CATEs may still differ arbitrarily in the embedding space between the RCT and OS; the discrepancy is modeled and estimated from RCT data.

Assumption 4 (Alignment quality)**.** The alignment error between the two encoders, measured on the RCT distribution, is bounded:

$$r_\phi^2 := \mathbb{E}\Big[\big\|\phi^o(\boldsymbol{X}^{o,*}) - \phi^r(\boldsymbol{X}^r)\big\|^2 \;\Big|\; S = r\Big] < \infty,$$

where $\boldsymbol{X}^{o,*} = (\boldsymbol{Z}, \boldsymbol{V}^*)$ with $\boldsymbol{V}^*$ denoting the (unobserved) value that $\boldsymbol{V}$ would take for an RCT unit. Furthermore, the outcome model $\tilde{\mu}_a^o$ is L_μ-Lipschitz in its argument and the discrepancy δ_a is L_δ-Lipschitz.

Checking the assumptions in practice. The bound contains two population quantities that are not directly observable: r_ϕ^2, because $\boldsymbol{V}$ is unobserved in the RCT, and ϵ_{suff}^2, because it refers to the best outcome-preserving representation in the population. We therefore treat them as diagnostic targets rather than tuning parameters. Appendix D gives held-out proxies: a shared-covariate alignment proxy for r_ϕ^2 and an OS residual proxy for ϵ_{suff}^2. These proxies do not prove the assumptions, but they make the theory empirically checkable; in simulations the alignment proxy tracks the oracle error closely, and in the real-data studies we use the same diagnostics to qualify the observed gains. The Lipschitz constants in Assumption 4 translate embedding error into prediction error. Appendix D also reports spectral-product upper bounds for the outcome and calibration heads (Table 4); these are regularization-controlled diagnostics, not certified global constants for the full neural network. Exact certification for ReLU networks generally requires specialized norm-constrained architectures or verification methods [Anil et al., 2019, Fazlyab et al., 2019]. Training then reduces observable surrogates of r_ϕ^2 through the alignment objective in Section 4.4.

4.3 THE CALM ALGORITHM

The CALM procedure has four stages.

Stage 1: OS outcome model in embedding space. Train the OS encoder ϕ^o and per-arm outcome heads $\hat{\mu}_a^o : \mathbb{R}^d \to \mathbb{R}$ jointly on OS data:

$$(\hat{\phi}^o, \hat{\mu}_{-1}^o, \hat{\mu}_{+1}^o) \in \operatorname*{arg\,min}_{\phi^o, \mu_a} \sum_{a \in \{-1,1\}} \frac{1}{n_a^o} \times \sum_{i: A_i = a, S_i = o} \big(Y_i - \mu_a(\phi^o(\boldsymbol{X}_i^o))\big)^2 + \mathcal{P}_{\text{OS}}, \quad (6)$$

where $\mathcal{P}_{\text{OS}}$ is a penalty (e.g., ℓ_1, weight decay, or dropout).

Algorithm 1 CALM: Calibrated Alignment under Covariate Mismatch

Require: OS data $\{(\boldsymbol{X}_j^o, A_j, Y_j)\}_{j=1}^{n^o}$; RCT data $\{(\boldsymbol{X}_i^r, A_i, Y_i)\}_{i=1}^{n^r}$; embedding dimension d; alignment weight λ.
1: **Stage 1:** Train OS encoder $\hat{\phi}^o$ and outcome heads $\hat{\mu}_a^o$ via (6).
2: **Stage 2:** Learn RCT encoder $\hat{\phi}^r$ and discrepancies $\hat{\delta}_a^t$ via (7).
3: **Stage 3:** Compute calibrated CMO $\hat{m}$ via (9); form pseudo-outcomes $\hat{\psi}_i^r$.
4: **Stage 4:** Estimate CATE correction $\hat{\delta}$ via (10).
5: **return** $\hat{\tau}_{\text{CALM}}(\boldsymbol{x}^r) = \tilde{\tau}(\boldsymbol{x}^r) + \hat{\delta}(\boldsymbol{x}^r)$

Stage 2: RCT encoder alignment and outcome calibration. Fix $\hat{\phi}^o$ and $\hat{\mu}_a^o$ from Stage 1. Learn the RCT encoder ϕ^r and arm-specific discrepancy functions $\hat{\delta}_a^t : \mathbb{R}^d \to \mathbb{R}$ by minimizing a calibration loss with an alignment penalty:

$$(\hat{\phi}^r, \hat{\delta}_a^t) \in \operatorname*{arg\,min}_{\phi^r, d_a} \sum_a \frac{1}{n_a^r} \sum_{i: A_i = a, S_i = r} \Big(Y_i - \hat{\mu}_a^o\big(\phi^r(\boldsymbol{X}_i^r)\big) - d_a\big(\phi^r(\boldsymbol{X}_i^r)\big)\Big)^2 + \mathcal{P}_{\text{cal}}(d_a) + \lambda \mathcal{L}_{\text{align}}(\hat{\phi}^o, \phi^r). \quad (7)$$

Here, $\mathcal{P}_{\text{cal}}$ controls the complexity of the discrepancy (encouraging borrowing from the OS by shrinking d_a toward zero), and $\mathcal{L}_{\text{align}}$ is the alignment loss (see Section 4.4).

Stage 3: Calibrated CMO and pseudo-outcome construction. Compute the calibrated arm-specific predictions and the calibrated CMO for RCT units:

$$\hat{\mu}_a^{\text{cal}}(\boldsymbol{x}^r) = \hat{\mu}_a^o\big(\hat{\phi}^r(\boldsymbol{x}^r)\big) + \hat{\delta}_a^t\big(\hat{\phi}^r(\boldsymbol{x}^r)\big), \quad (8)$$

$$\hat{m}(\boldsymbol{x}^r) = \sum_a \pi_{-a}^r(\boldsymbol{x}^r)\, \hat{\mu}_a^{\text{cal}}(\boldsymbol{x}^r). \quad (9)$$

Form **pseudo-outcomes** $\hat{\psi}_i^r = A_i(Y_i - \hat{m}(\boldsymbol{X}_i^r))/\pi_{A_i}^r(\boldsymbol{X}_i^r)$ for each RCT unit.

Stage 4: CATE correction. The preliminary CATE is $\tilde{\tau}(\boldsymbol{x}^r) = \sum_a a\, \hat{\mu}_a^{\text{cal}}(\boldsymbol{x}^r)$. Estimate a CATE correction $\hat{\delta} : \mathbb{R}^{p_r} \to \mathbb{R}$ using RCT data:

$$\hat{\delta} \in \operatorname*{arg\,min}_{d \in \mathcal{D}} \frac{1}{n^r} \sum_{i: S_i = r} \big(\hat{\psi}_i^r - \tilde{\tau}(\boldsymbol{X}_i^r) - d(\boldsymbol{X}_i^r)\big)^2 + \mathcal{P}_\tau(d). \quad (10)$$

The CALM estimator is: $\boxed{\hat{\tau}_{\text{CALM}}(\boldsymbol{x}^r) = \tilde{\tau}(\boldsymbol{x}^r) + \hat{\delta}(\boldsymbol{x}^r).}$

The CATE calibration (Stage 4) operates on the *original* RCT covariates $\boldsymbol{X}^r$, not on the embedding. This keeps the final CATE estimate interpretable in terms of the original covariates; causal identification is preserved by the unbiasedness of the pseudo-outcomes under RCT randomization. The full procedure is summarized in Algorithm 1.

4.4 ALIGNMENT OBJECTIVES

The alignment loss $\mathcal{L}_{\text{align}}$ encourages the two encoders to produce compatible representations. Since $\boldsymbol{V}$ is un-

observed in the RCT and $\boldsymbol{U}$ in the OS, we cannot directly minimize $\|\phi^o(\boldsymbol{X}^o) - \phi^r(\boldsymbol{X}^r)\|^2$ on paired units; we therefore align through the shared covariates $\boldsymbol{Z}$. We present two generic alignment objectives; the experiments use the conditional-mean variant described in Appendix B. *Distribution-matching alignment* (MMD) uses the kernel maximum mean discrepancy [Gretton et al., 2012]:

$$\mathcal{L}_{\mathrm{MMD}} = \Big\| \tfrac{1}{n^o} \textstyle\sum_{j\in \mathrm{OS}} k\big(\hat{\phi}^o(\boldsymbol{X}_j^o), \cdot\big) - \tfrac{1}{n^r} \sum_{i\in \mathrm{RCT}} k\big(\phi^r(\boldsymbol{X}_i^r), \cdot\big) \Big\|^2_{\mathcal{H}_k}, \quad (11)$$

which matches the marginal embedding distributions (see Appendix B for the expanded objective and additional alignment variants). *$\boldsymbol{Z}$-conditioned contrastive alignment* matches units with similar shared covariates:

$$\mathcal{L}_{\mathrm{contr}} = \tfrac{1}{n^r} \textstyle\sum_{i\in \mathrm{RCT}} \tfrac{1}{|N_i|} \sum_{j\in N_i} \big\|\hat{\phi}^o(\boldsymbol{X}_j^o) - \phi^r(\boldsymbol{X}_i^r)\big\|^2, \quad (12)$$

where $N_i = \{j \in \mathrm{OS} : \|\boldsymbol{Z}_j - \boldsymbol{Z}_i\| \le \varepsilon\}$. Adversarial alignment [Ganin et al., 2016] is also applicable (Appendix B).

5 THEORETICAL ANALYSIS

This section derives finite-sample risk bounds for CALM and compares them to the MR-OSCAR bounds. Proofs for the theorem, corollaries, and propositions in this section are in Appendix A.

5.1 RISK BOUND FOR CALM

We use two standard learning-theory quantities in the bound. For a function class $\mathcal{A}$, define its approximation error relative to a target g by $\Delta_2^2(\mathcal{A}, g) = \inf_{f\in\mathcal{A}} \Delta_2^2(f, g)$, where $\Delta_2^2(f, g)$ is the RCT-population MSE defined in Section 3. We write $\mathfrak{R}_n(\mathcal{A})$ for the empirical Rademacher complexity of $\mathcal{A}$, $\mathfrak{R}_n(\mathcal{A}) = \mathbb{E}_\sigma\big[\sup_{f\in\mathcal{A}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(W_i)\big]$, where σ_i are independent Rademacher signs and W_i denotes the relevant inputs for the class under consideration (RCT covariates, OS embeddings, or RCT embeddings). This quantity measures the statistical complexity of a function class.

Theorem 1 (Risk bound for CALM)**.** *Suppose Assumptions 1–4 hold, the outcome and fitted-function classes have bounded envelopes, and all nuisance estimators are obtained via penalized empirical risk minimization with sample splitting or cross-fitting across stages. Let $\mathcal{D}$ be the function class for the CATE correction in Stage 4 (Algorithm 1), and let $\mathcal{F} = \{\tilde{\tau} + d : d \in \mathcal{D}\}$ be the induced class for the final CATE estimator on $\boldsymbol{X}^r$. Let $\mathcal{M}_a^o$ be the class for OS outcome models on $\mathbb{R}^d$ and $\mathcal{D}_a$ the class for arm-specific discrepancy on $\mathbb{R}^d$. Then, for any failure probability $\gamma \in (0, 1)$, there exists a constant $C > 0$ such that, with probability at least $1 - \gamma$:*

$$\begin{aligned}\Delta_2^2(\hat{\tau}_{\mathrm{CALM}}, \tau^r) \le{}& \Delta_2^2(\mathcal{F}, \tau^r) + C\frac{\log(1/\gamma)}{n^r} \\ &+ C\Bigg[\underbrace{\epsilon_{\mathrm{suff}}^2}_{\textit{suff.}} + \underbrace{(L_\mu + L_\delta)^2 r_\phi^2}_{\textit{align.}} + \underbrace{\textstyle\sum_a \mathfrak{R}_{n^o}^2(\mathcal{M}_a^o)}_{\textit{OS}} \\ &+ \underbrace{\textstyle\sum_a \mathfrak{R}_{n^r}^2(\mathcal{D}_a)}_{\textit{calib.}} + \underbrace{\mathfrak{R}_{n^r}^2(\mathcal{F})}_{\textit{CATE}}\Bigg], \end{aligned} \quad (13)$$

where L_μ, L_δ are the Lipschitz constants linking alignment error to prediction error.

Interpretation. The proof is in Appendix A. The first term is the approximation error of the CATE class, and $C\log(1/\gamma)/n^r$ is the finite-sample concentration cost. Inside the brackets, $\epsilon_{\mathrm{suff}}^2$ measures information lost by the embedding, $(L_\mu + L_\delta)^2 r_\phi^2$ is the prediction error induced by imperfect alignment, the two summed Rademacher terms are the OS outcome-model and RCT calibration complexities, and $\mathfrak{R}_{n^r}^2(\mathcal{F})$ is the complexity of the CATE regression.

5.2 COMPARISON WITH IMPUTATION-BASED BORROWING

Corollary 2 (When CALM improves over MR-OSCAR)**.** *Assume both estimators use function classes of comparable complexity for calibration [see Pal et al., 2026, for the* MR-OSCAR *bound]. Let r_{im}^2 denote the oracle risk of imputing $\boldsymbol{V}$ from $\boldsymbol{Z}$, and let $\mathcal{G}$ denote the imputation function class. Then* CALM *yields a tighter bound than* MR-OSCAR *whenever $\epsilon_{\mathrm{suff}}^2 + (L_\mu + L_\delta)^2 r_\phi^2 + \sum_a \mathfrak{R}_{n^o}^2(\mathcal{M}_a^o) < L^2 r_{\mathrm{im}}^2 + \sum_a \mathfrak{R}_{n^o}^2(\mathcal{M}_a^{o,\mathrm{im}}) + \mathfrak{R}_{n^o}^2(\mathcal{G})$. This holds when:*

1. *$d \ll p_o$, so outcome models in $\mathbb{R}^d$ have lower Rademacher complexity than those in $\mathbb{R}^{p_o}$;*
2. *$\boldsymbol{V}$ is hard to reconstruct but easy to summarize: $r_\phi^2 \ll r_{\mathrm{im}}^2$ because the outcome-relevant information in $\boldsymbol{V}$ lies on a low-dimensional manifold;*
3. *The sufficiency gap $\epsilon_{\mathrm{suff}}^2$ is small, i.e., the embedding captures most outcome-relevant variation.*

The bound exposes the trade-off: CALM gains from dimensionality reduction but pays for information loss ($\epsilon_{\mathrm{suff}}^2$) and alignment imperfection (r_ϕ^2), while MR-OSCAR avoids information loss but pays the full imputation cost (r_{im}^2 in $\mathbb{R}^{p_v}$). SR-OSCAR avoids imputation as well, but discards all information outside the shared block; it is therefore strongest when $\boldsymbol{Z}$ already captures the outcome-relevant variation.

5.3 NEGATIVE-TRANSFER PROTECTION

Proposition 3 (Safe borrowing)**.** *Under the conditions of Theorem 1, the pseudo-outcome $\hat{\psi}^r$ remains unbiased for*

$\tau^r(\boldsymbol{X}^r)$ regardless of augmentation quality (Eq. 2). Thus, both CALM*-Lin and* CALM*-NN preserve the RCT CATE target: poor augmentation can affect efficiency, but it does not change the estimand. When alignment error $(L_\mu + L_\delta)^2 r_\phi^2$ is large, the borrowed augmentation may provide little variance reduction, and Stage 4 must rely more heavily on RCT data through the correction $\hat{\delta}(\boldsymbol{X}^r)$. The practical difference is finite-sample stability:* CALM*-Lin has a more controlled fallback, while* CALM*-NN can suffer higher variance or approximation error under severe distributional shift if the learned encoder aligns poorly.*

5.4 SPECIALIZATION TO SPARSE LINEAR MODELS

We specialize CALM to a linear setting. Let the encoders be linear projections $\phi^o(\boldsymbol{X}^o) = \boldsymbol{W}^o(\boldsymbol{X}^o)^\top$ and $\phi^r(\boldsymbol{X}^r) = \boldsymbol{W}^r(\boldsymbol{X}^r)^\top$, where $\boldsymbol{W}^o \in \mathbb{R}^{d\times p_o}$ and $\boldsymbol{W}^r \in \mathbb{R}^{d\times p_r}$. Let the outcome be linear in the embedding: $\tilde{\mu}_a^o(\boldsymbol{h}) = \boldsymbol{\beta}_a^\top \boldsymbol{h}$.

Proposition 4 (Linear embedding & imputation)**.** *Under Gaussian covariates with $\boldsymbol{V} = \boldsymbol{\Lambda}\boldsymbol{Z} + \boldsymbol{\epsilon}_V$ and $\boldsymbol{U} \perp\!\!\!\perp \boldsymbol{V} \mid \boldsymbol{Z}$, the optimal linear embedding that minimizes alignment error subject to preserving outcome-relevant information satisfies*

$$\boldsymbol{W}^r_{\text{opt}} = \boldsymbol{W}^o \begin{pmatrix} \boldsymbol{0}_{p_z\times p_u} & \boldsymbol{I}_{p_z} \\ \boldsymbol{0}_{p_v\times p_u} & \boldsymbol{\Lambda} \end{pmatrix},$$

which is equivalent to imputing $\widehat{\boldsymbol{V}} = \boldsymbol{\Lambda}\boldsymbol{Z}$ and then projecting $(\boldsymbol{Z}, \widehat{\boldsymbol{V}})$ through $\boldsymbol{W}^o$.

In the linear case, the alignment-based approach with a d-dimensional embedding is equivalent to imputation followed by dimensionality reduction to $\mathbb{R}^d$. Thus, the embedding approach contains the linear imputation as a special case: when $d = p_o$, it recovers MR-OSCAR; when $d < p_o$, it can gain additional variance reduction through projection.

Corollary 5 (CALM-Lin risk bound)**.** *Under the conditions of Proposition 4, if the outcome model has sparsity s in the embedding space and the discrepancy has sparsity s_δ, then*

$$\Delta_2^2(\hat{\tau}_{\text{CALM}}, \tau^r) \lesssim \underbrace{\Delta_2^2(\mathcal{F}, \tau^r)}_{\textit{approx.}} + \underbrace{(L_\mu + L_\delta)^2 \text{tr}(\boldsymbol{\Sigma}_{\boldsymbol{V}|\boldsymbol{Z}})}_{\textit{irred. noise}} + \underbrace{\frac{s\log d}{n^o}}_{\textit{OS}} + \underbrace{\frac{s_\delta \log d}{n^r}}_{\textit{calib.}} + \underbrace{\frac{s_\tau \log p_r}{n^r}}_{\textit{CATE}},$$

where s_τ is the CATE sparsity. Compared to MR-OSCAR*, the outcome model and calibration complexities scale with $\log d$ instead of $\log p_o$, yielding tighter rates when $d \ll p_o$.*

6 PRACTICAL IMPLEMENTATION

In the linear instantiation (CALM-Lin), each stage reduces to penalized regression: Stage 1 fits a reduced-rank regression (or LASSO in a PCA-reduced space) of Y on $\boldsymbol{X}^o$ in the OS; Stage 2 constructs $\boldsymbol{W}^r = \boldsymbol{W}^o\hat{\boldsymbol{M}}$, where $\hat{\boldsymbol{M}}$ maps $(\boldsymbol{U}, \boldsymbol{Z})$ to $(\boldsymbol{Z}, \hat{\boldsymbol{\Lambda}}\boldsymbol{Z})$ and $\hat{\boldsymbol{\Lambda}}$ estimates $\mathbb{E}[\boldsymbol{V} \mid \boldsymbol{Z}]$ by ridge/LASSO, and calibrates via LASSO on $\boldsymbol{X}^r$; Stages 3–4 follow the standard R-OSCAR pipeline. In the neural instantiation (CALM-NN), the encoders ϕ^o and ϕ^r are MLPs with separate parameters and per-arm outcome heads; our experiments use the conditional-mean version of the $\boldsymbol{Z}$-conditioned alignment objective in Appendix B, while MMD (11) is a distribution-matching alternative. Stage 1 uses OS data alone, Stage 2 jointly optimizes ϕ^r and $\hat{\delta}_a^t$ with the alignment penalty (optionally initialized from ϕ^o restricted to shared features), Stage 4 uses LASSO on $\boldsymbol{X}^r$ for interpretability, and all nuisance estimates are cross-fitted over K RCT folds. The embedding dimension d controls the bias–variance trade-off: for CALM-Lin, d is set by cross-validated OS prediction error; for CALM-NN, d is selected by cross-validated CATE calibration error on the RCT.

Choosing between CALM-Lin and CALM-NN. Prefer CALM-Lin when the CATE is plausibly smooth or close to linear in $\boldsymbol{X}^r$, or when the trial is small ($n^r \lesssim 200$): there the safety guarantee of Proposition 3 applies in full, and the linear instantiation has the lowest variance among the calibration-family methods. Prefer CALM-NN when cross-validation shows that a nonlinear OS outcome model clearly outperforms a linear one and n^o is large enough to learn a useful encoder; the GPS and STAR analyses of Section 7.6 both fall in this regime. In borderline cases, we compare the cross-fitted Stage 2 calibration loss and final pseudo-outcome regression error on held-out RCT folds. If the nonlinear encoder does not improve these diagnostics, the linear version is the default because it borrows less aggressively and is easier to audit.

7 EXPERIMENTS

We evaluate CALM[1] with simulations that target the predictions of Section 5.

7.1 SIMULATION DESIGN

Data-generating process. We generate (OS, RCT) pair as follows. Let $p_z = 30$, $p_u = 10$, $p_v = 20$, and $d_{\text{true}} = 5$ denote the intrinsic dimension of the outcome-relevant signal. Shared covariates are drawn as $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{Z}\boldsymbol{Z}})$ with AR(1) correlation ($\rho = 0.5$); (i) OS-only covariates $\boldsymbol{V} = \boldsymbol{\Lambda}_{\boldsymbol{V}\boldsymbol{Z}}\boldsymbol{Z} + \boldsymbol{\epsilon}_V$, $\boldsymbol{\epsilon}_V \sim \mathcal{N}(\boldsymbol{0}, \sigma_V^2\boldsymbol{I})$, and (ii) RCT-only covariates $\boldsymbol{U} = \boldsymbol{\Lambda}_{\boldsymbol{U}\boldsymbol{Z}}\boldsymbol{Z} + \boldsymbol{\epsilon}_U$, $\boldsymbol{\epsilon}_U \sim \mathcal{N}(\boldsymbol{0}, \sigma_U^2\boldsymbol{I})$, are generated conditionally on $\boldsymbol{Z}$. Outcomes follow $Y^s(a) = f_a(\boldsymbol{P}^s\boldsymbol{X}^\top) + \delta_a^s(\boldsymbol{Z}) + \epsilon$, where $\boldsymbol{P}^s$ projects to a d_{true}-dimensional subspace, f_a is a nonlinear function (with linear and sinusoidal variants), and δ_a^s captures population-specific

[1]Code to reproduce all experiments is available at `https://github.com/AsiaeeLab/calm-cate`.

shifts. Treatment assignment in the OS uses a logistic model on 10 covariates; the RCT uses $\pi^r_{+1} = 0.5$.

Factors varied. Each experiment varies one factor while holding others at default values ($n^r = 500$, $\sigma^2_V = 1.0$, $d_{\text{true}} = 5$, linear outcome, shift magnitude 0.5). We vary six factors: imputation difficulty, $\sigma^2_V \in \{0.1, 0.25, 0.5, 1.0, 2.0\}$ (higher σ^2_V makes $\boldsymbol{V}$ harder to predict from $\boldsymbol{Z}$); intrinsic dimension, $d_{\text{true}} \in \{2, 3, 5, 10, 15, 20\}$; RCT sample size, $n^r \in \{100, 250, 500, 1{,}000, 2{,}000\}$ with $n^o = 10{,}000$ fixed; outcome nonlinearity (linear, quadratic, sinusoidal); outcome shift magnitude, $\|\delta^r_a - \delta^o_a\| \in \{0, 0.25, 0.5, 1.0, 2.0, 5.0\}$; and shared covariate proportion, $p_z/(p_z + p_v) \in \{0.3, 0.5, 0.7, 0.9\}$.

7.2 METHODS UNDER COMPARISON

We compare eight methods: **Naive**, RCT-only CATE estimation with no augmentation ($m = 0$); **RACER**, RCT-only augmentation using outcome models fitted on $\boldsymbol{X}^r$ in the RCT; **SR-OSCAR** [Pal et al., 2026], which borrows from OS using only shared covariates $\boldsymbol{Z}$; **MR-OSCAR** [Pal et al., 2026], imputation-based borrowing; **CALM-Lin**, the linear embedding version of CALM; **CALM-NN**, the neural network version of CALM (residual MLP encoders with an alignment loss); and **HTCE-T** and **HTCE-DR**, the transfer T-learner and DR-learner of Bica and van der Schaar [2022], respectively. For HTCE-T and HTCE-DR, we use the shared/private encoder architectures of Bica and van der Schaar [2022] and compute predictions by cross-fitting over RCT folds to avoid outcome leakage.

7.3 EVALUATION METRICS

We report the RMSE of CATE, $\left[(1/n^r)\sum_i(\hat{\tau}(\boldsymbol{X}^r_i) - \tau^r(\boldsymbol{X}^r_i))^2\right]^{1/2}$, averaged over 20 replicates.

7.4 RESULTS: LINEAR CATE REGIME

Unless noted, reported values are mean RMSE over 20 replicates. The baseline regime comprises 29 settings in which the CATE is linear in the target covariates and a single mismatch factor is varied at a time. Across these 29 settings, the lowest mean RMSE is achieved by RACER in 11, CALM-Lin in 9, MR-OSCAR in 6, SR-OSCAR in 2, and HTCE-T in 1. All four calibration-based methods (RACER, SR-OSCAR, MR-OSCAR, CALM-Lin) are effectively indistinguishable in this regime: pairwise differences in mean RMSE are below 10^{-3}, and which method achieves the argmin depends on the factor varied. Table 1 reports representative values.

Imputation difficulty (Figure 2a). As σ^2_V increases, all calibration-based methods degrade together. The argmin shifts with the noise level: RACER is best at $\sigma^2_V \in \{0.1, 0.5, 2.0\}$ (RMSE 0.76, 0.91, and 1.45), CALM-Lin at $\sigma^2_V = 0.25$ (0.83), and MR-OSCAR at $\sigma^2_V = 1.0$ (1.03). CALM-NN has higher RMSE than the calibration-based methods in this linear-CATE setting.

Table 1: Mean RMSE across imputation difficulty (σ^2_V, linear outcome) and outcome nonlinearity ($\sigma^2_V = 1.0$), averaged over 20 replicates. **Bold**: lowest per column.

	Imputation difficulty (σ^2_V)				Outcome	
Method	0.1	0.5	1.0	2.0	Quad.	Sin.
Naive	1.16	1.22	1.30	1.65	1.70	1.29
Calib. group†	**0.76**	**0.91**	**1.03**	**1.45**	**1.45**	**1.06**
CALM-NN	0.88	1.02	1.18	1.59	1.54	1.15
HTCE-T	1.15	1.28	1.45	1.93	1.81	1.43
HTCE-DR	1.17	1.36	1.48	1.98	1.92	1.50

†RACER, SR-OSCAR, MR-OSCAR, CALM-Lin: single representative; pairwise ΔRMSE$<10^{-3}$.

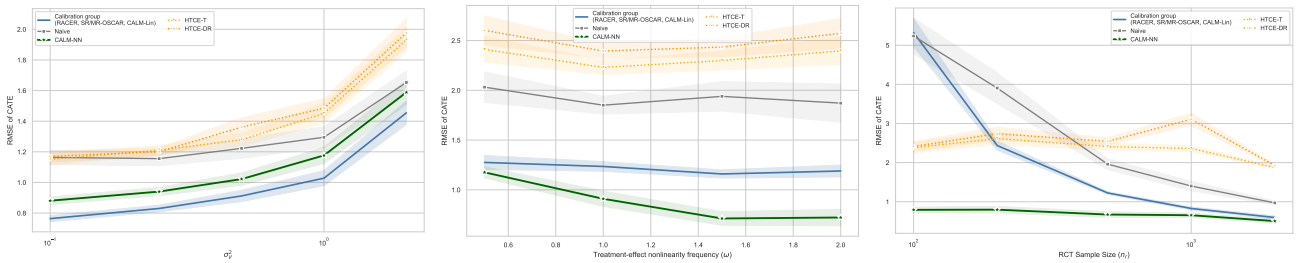
Additional linear-regime sweeps including intrinsic dimension, shared-covariate proportion, RCT sample size, outcome-model nonlinearity, and negative transfer appear in Appendix C.

7.5 RESULTS: NONLINEAR-CATE REGIME

We turn to a regime in which the CATE is nonlinear in the target covariates. We modify the DGP so that the outcome-relevant signal is mediated entirely by $\boldsymbol{V}$ through shared latent factors: $\boldsymbol{U}$ and $\boldsymbol{V}$ share a 5-dimensional latent factor $\boldsymbol{H}$ (scale 2.0 in both measurement models, $\sigma^2_V = \sigma^2_U = 0.1$), and the outcome depends only on $\boldsymbol{V}$ (signal weights $w_Z = w_U = 0$, $w_V = 2$). The CATE takes a nonlinear sinusoidal form with frequency parameter ω, so that $\tau^r(\boldsymbol{X}^r)$ is nonlinear after marginalization over $\boldsymbol{V}$. This regime, comprising 22 settings across the sweeps described below, is not included in the baseline grid. CALM-NN attains the lowest mean RMSE in all 22 settings, with margins ranging from 0.09 to 4.53.

Treatment-effect frequency (Figure 2b). Across $\omega \in \{0.5, 1.0, 1.5, 2.0\}$, CALM-NN attains the lowest RMSE in all four settings. At $\omega = 1.5$, CALM-NN achieves RMSE 0.71 compared with 1.16 for CALM-Lin; at $\omega = 2.0$, RMSE is 0.72 versus 1.19 for the best calibration-based method, a 39% reduction. This matches the regime predicted by Corollary 2: when outcome-relevant information is low-dimensional and the CATE is nonlinear, learned embeddings outperform imputation-based approaches.

Sample efficiency (Figure 2c). We vary $n^r \in \{100, 200, 500, 1000, 2000\}$ within the nonlinear-CATE DGP ($\omega = 1.5$, $w_Z = 0$). CALM-NN maintains RMSE in the range 0.51–0.79 across all sample sizes, whereas calibration-based methods degrade sharply at small n^r: RMSE 5.32 at $n^r = 100$ compared with 0.79 for CALM-NN. Even at $n^r = 2{,}000$, CALM-NN retains a modest advantage (0.51 vs. 0.60). This stability reflects the use of 10,000 OS



(a) Imputation difficulty (σ_V^2). (b) CATE nonlinearity (ω). (c) RCT sample size (n^r).

Figure 2: RMSE of CATE estimation across three experimental sweeps. Mean over 20 replicates. In all panels, the blue band groups four calibration-based methods (RACER, SR-OSCAR, MR-OSCAR, CALM-Lin) whose RMSEs are nearly identical; the band spans their min–max envelopes (mean $\pm$ SE). Individual lines show the remaining methods. **(a)** $n^r = 500$, $n^o = 10{,}000$, $d_{\text{true}} = 5$, linear outcome. **(b)** Shared-latent DGP where $\boldsymbol{X}^r$ carries information about $\boldsymbol{V}$ beyond $\boldsymbol{Z}$; CALM-NN separates from the calibration group as ω increases. **(c)** Nonlinear-CATE regime: CALM-NN maintains RMSE below 0.8 even at $n^r = 100$, where calibration-based methods exceed 5.3.

Table 2: GPS real-data evaluation: mean width of the per-unit 95% bootstrap CI ($B = 200$; $n^r = 330$, $n^o = 8{,}867$). Only CALM-NN improves materially over the trial-only RACER baseline.

Method	Mean 95% CI width	vs. RACER
Naive	2.25	-13.2%
RACER	1.98	—
SR-OSCAR	1.98	0.2%
MR-OSCAR	1.97	0.7%
CALM-Lin	1.98	0.0%
CALM-NN	**1.64**	**17.5%**
HTCE-T	3.49	-75.9%
HTCE-DR	4.12	-107.8%

samples for representation learning, making the quality of the learned features largely independent of n^r.

Further nonlinear-regime experiments including robustness to the shared-covariate signal weight, the CATE functional form, and the latent coupling strength are in Appendix C.

7.6 REAL-DATA EVALUATION

We complement the simulations with two analyses on real data, where the DGP is outside our control.

Greenlight Plus (GPS) with an EHR cohort. Following the covariate-mismatch protocol of Pal et al. [2026], we link the Greenlight Plus trial [Heerman et al., 2022] subset ($n^r = 330$) to EHR controls ($n^o = 8{,}867$) and mask the EHR-only insurance covariate from the trial as $\boldsymbol{V}$. The outcome is 24-month weight-for-length z-score; we report mean 95% bootstrap-CI width ($B = 200$). Table 2 shows that only CALM-NN materially tightens intervals over trial-only RACER (17.5%), while calibration variants stay within 1% and transfer-only neural baselines widen them.

Semi-real Tennessee STAR. Following Kallus et al. [2018], we build the OS by outcome-dependent subsampling of STAR students and mask kindergarten free-lunch from the trial side. Because all units come from STAR, we estimate per-unit ground-truth CATEs with a cross-fitted random-forest T-learner on the full first-grade cohort and report RMSE over 15 replicates at each trial fraction q. Table 3 shows that CALM-NN leads at $q = 0.25$ and $q = 1.0$, with ties in between.

Table 3: Semi-real STAR: RMSE against the ground-truth CATE (15 replicates) at trial fraction q. CALM-NN leads at the extremes and ties in between. Naive (≈ 1064) and the HTCE baselines (≈ 420–465) are omitted for readability.

q	RACER	SR-OSC.	MR-OSC.	C-Lin	C-NN
0.25	52.12	51.20	51.36	51.42	**50.61**
0.50	**31.33**	31.40	31.65	**31.33**	31.53
0.75	**25.87**	25.97	25.89	25.90	26.04
1.00	23.60	23.74	23.54	23.60	**22.30**

8 CONCLUSION

CALM shows that CATE estimation under covariate mismatch can borrow through an outcome-relevant aligned representation, rather than through full reconstruction of missing covariates. The risk bounds make this trade-off explicit by separating alignment error, embedding sufficiency, nuisance-model complexity, and final CATE-regression complexity. The empirical results mirror this pattern: linear calibration is stable when the CATE is simple, while CALM-NN gains when nonlinear OS-learned features improve held-out trial calibration. Because alignment quality is not directly observable in the RCT, the method is most useful when held-out calibration and residual diagnostics indicate that the learned representation carries outcome-relevant OS signal without overwhelming the trial. The method is therefore best viewed as calibrated borrowing, not a replacement for randomized identification.

Acknowledgements

This work was supported in part by the Patient-Centered Outcomes Research Institute (PCORI) under award ME-2023C1-32148. A.A. was partly supported by the National Institute of Mental Health under award R01MH139379. We thank Jared Huling for his support during the early development of this work.

References

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255, 2013.

Cem Anil, James Lucas, and Roger Grosse. Sorting out Lipschitz function approximation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 291–301, 2019.

Amir Asiaee and Samhita Pal. B-calm: Bias-limited bayesian borrowing for rct-anchored treatment effects under covariate mismatch. *arXiv preprint arXiv:2607.04036*, 2026.

Amir Asiaee, Chiara Di Gravio, Cole Beck, Yuting Mei, Samhita Pal, and Jared D. Huling. Improving precision of RCT-based CATE estimation using data borrowing with double calibration. *arXiv preprint arXiv:2306.17478*, 2025.

Amir Asiaee, Samhita Pal, Cole Beck, and Jared D Huling. Sharp bounds for treatment effect generalization under outcome distribution shift. *arXiv preprint arXiv:2602.09595*, 2026a.

Amir Asiaee, Samhita Pal, and Jared D Huling. Omitted-variable sensitivity analysis for generalizing randomized trials. *arXiv preprint arXiv:2603.27788*, 2026b.

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

Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3): 311–334, 2006.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175, 2010.

Ioana Bica and Mihaela van der Schaar. Transfer learning on heterogeneous feature spaces for treatment effects estimation. In *Advances in Neural Information Processing Systems*, volume 35, pages 37184–37198, 2022.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. *Advances in Neural Information Processing Systems*, 29, 2016.

Carly Lupton Brantner, Ting-Hsuan Chang, Trang Quynh Nguyen, Hwanhee Hong, Leon Di Stefano, and Elizabeth A Stuart. Methods for integrating trials and non-experimental data to examine treatment effect heterogeneity. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 38(4):640, 2023.

Ming-Hui Chen and Joseph G. Ibrahim. Power prior distributions for regression models. *Statistical Science*, 15(1): 46–60, 2000.

David Cheng and Tianxi Cai. Adaptive combination of randomized and observational data. *arXiv preprint arXiv:2111.15012*, 2021.

Bénédicte Colnet, Julie Josse, Gaël Varoquaux, and Erwan Scornet. Causal effect on a target population: A sensitivity analysis to handle missing covariates. *Journal of Causal Inference*, 10(1):372–414, 2022. doi: 10.1515/jci-2021-0059.

Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review. *Statistical Science*, 39(1):165–191, 2024.

R. Dennis Cook. Fisher lecture: Dimension reduction in regression. *Statistical Science*, 22(1):1–26, 2007.

Issa J. Dahabreh, Sarah E. Robertson, Jon A. Steingrimsson, Elizabeth A. Stuart, and Miguel A. Hernán. Extending inferences from a randomized trial to a new target population. *Statistics in Medicine*, 39(14):1999–2014, 2020. doi: 10.1002/sim.8426.

Irina Degtiar and Sherri Rose. A review of generalizability and transportability. *Annual Review of Statistics and Its Application*, 10:501–524, 2023.

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George J. Pappas. Efficient and accurate estimation of Lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pages 11423–11434, 2019.

Kenji Fukumizu, Francis R. Bach, and Michael I. Jordan. Kernel dimension reduction in regression. *The Annals of Statistics*, 37(4):1871–1905, 2009. doi: 10.1214/08-AOS637.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Trinetri Ghosh, Yanyuan Ma, and Xavier de Luna. Sufficient dimension reduction for feasible and robust estimation of average causal effect. *Statistica Sinica*, 31(2):821–842, 2021. doi: 10.5705/ss.202018.0416.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

P. Richard Hahn, Jared S. Murray, and Carlos M. Carvalho. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 15(3):965–1056, 2020.

David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16 (12):2639–2664, 2004.

Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2020.

Tobias Hatt, Jeroen Berrevoets, Alicia Curth, Stefan Feuerriegel, and Mihaela van der Schaar. Combining observational and randomized data for estimating heterogeneous treatment effects. *arXiv preprint arXiv:2202.12891*, 2022.

William J. Heerman, Eliana M. Perrin, H. Shonna Yin, Jonathan S. Schildcrout, Alan M. Delamater, Kori B. Flower, Lee Sanders, Charles Wood, Melissa C. Kay, Laura E. Adams, and Russell L. Rothman. The Greenlight Plus Trial: Comparative effectiveness of a health information technology intervention vs. health communication intervention in primary care offices to prevent childhood obesity. *Contemporary Clinical Trials*, 123: 106987, 2022. doi: 10.1016/j.cct.2022.106987.

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.

Brian P. Hobbs, Bradley P. Carlin, Sumithra J. Mandrekar, and Daniel J. Sargent. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, 67(3): 1047–1056, 2011.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pages 3020–3029, 2016.

Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Rickard Karlsson, Piersilvio De Bartolomeis, Issa J. Dahabreh, and Jesse H. Krijthe. Robust estimation of heterogeneous treatment effects in randomized trials leveraging external data. *arXiv preprint arXiv:2507.03681*, 2025.

Edward H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

David M. Kent, Jessica K. Paulus, David van Klaveren, Ralph D'Agostino, Steve Goodman, Rodney Hayward, John P. A. Ioannidis, Bray Patrick-Lake, Sally Morton, Michael Pencina, Gowri Raman, Joseph S. Ross, Harry P. Selker, Ravi Varadhan, Andrew Vickers, John B. Wong, and Ewout W. Steyerberg. The predictive approaches to treatment effect heterogeneity (PATH) statement. *Annals of Internal Medicine*, 172(1):35–45, 2020. doi: 10.7326/M18-3667.

Michael R. Kosorok and Eric B. Laber. Precision medicine. *Annual Review of Statistics and Its Application*, 6:263–286, 2019.

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.

Dasom Lee, Shu Yang, Lin Dong, Xiaofei Wang, Donglin Zeng, and Jianwen Cai. Improving trial generalizability using observational studies. *Biometrics*, 79(2):1213–1225, 2023.

Bing Li. *Sufficient dimension reduction: Methods and applications with R*. CRC Press, 2018.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991. doi: 10.1080/01621459.1991.10475035.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 97–105. JMLR.org, 2015.

Yanyuan Ma and Liping Zhu. A semiparametric approach to dimension reduction. *Journal of the American Statistical Association*, 107(497):168–179, 2012.

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

Michael Oberst, Alexander D'Amour, Minmin Chen, Yuyan Wang, David Sontag, and Steve Yadlowsky. Understanding the risks and rewards of combining unbiased and possibly biased estimators, with applications to causal inference. *arXiv preprint arXiv:2205.10467*, 2022.

Samhita Pal, Jared D. Huling, and Amir Asiaee. Improving RCT-based CATE estimation under covariate mismatch via double calibration. *arXiv preprint arXiv:2603.17066*, 2026.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

Susanne Rässler. *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. Springer, 2002.

Michael T. Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G. Dietterich. To transfer or not to transfer. In *NIPS 2005 Workshop on Inductive Transfer*, 2005.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085, 2017.

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

Masashi Sugiyama and Motoaki Kawanabe. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. MIT Press, 2012.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Rui Wang, Stephen W. Lagakos, James H. Ware, David J. Hunter, and Jeffrey M. Drazen. Statistics in medicine—reporting of subgroup analyses in clinical trials. *New England Journal of Medicine*, 357(21):2189–2194, 2007.

Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1): 9, 2016. doi: 10.1186/s40537-016-0043-6.

Shu Yang, Siyi Liu, Donglin Zeng, and Xiaofei Wang. Data fusion methods for the heterogeneity of treatment effect and confounding function. *Bernoulli*, 31(4), 2025. doi: 10.3150/24-bej1835.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Ruoqi Yu, Bikram Karmakar, Jessica Vandeleest, and Eleanor Bimla Schwarz. Using a two-parameter sensitivity analysis framework to efficiently combine randomized and nonrandomized studies. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2025. doi: 10.1093/jrsssb/qkaf079. Advance article qkaf079; arXiv:2412.03731.

Improving RCT-Based Treatment Effect Estimation Under Covariate Mismatch via Calibrated Alignment (Supplementary Material)

Amir Asiaee[1] **Samhita Pal**[1]

[1]Department of Biostatistics, Vanderbilt University Medical Center, TN 37203, USA

A PROOFS

Proof of Theorem 1. We decompose the error in three steps: the CATE calibration layer, the augmentation error, and the assembly.

Step 1 (CATE calibration layer). By the pseudo-outcome regression framework (Asiaee et al. [2025], Theorem 6), the CATE estimation error satisfies

$$\Delta_2^2(\hat{\tau}_{\text{CALM}}, \tau^r) \le \Delta_2^2(\mathcal{F}, \tau^r) + C_1\Big(1 + \textstyle\sum_a \Delta_{2,r}^2\big(\hat{\mu}_a^{\text{cal}}, \mu_a^r\big)\Big)\mathfrak{R}_{n^r}^2(\mathcal{F}) + C_2\frac{\log(1/\gamma)}{n^r}, \tag{14}$$

where μ_a^r is the RCT arm-specific mean from Section 3, and $\hat{\mu}_a^{\text{cal}}(\boldsymbol{x}^r) = \hat{\mu}_a^o(\hat{\phi}^r(\boldsymbol{x}^r)) + \hat{\delta}_a^t(\hat{\phi}^r(\boldsymbol{x}^r))$ is the calibrated prediction from Stage 2. The first term is the approximation error of $\mathcal{F}$; the second captures how augmentation quality changes the effective noise in pseudo-outcome regression. Cross-fitting ensures that the nuisance estimates $\hat{\mu}_a^{\text{cal}}$ are independent of the pseudo-outcome regression sample, allowing us to bound the stochastic term via Rademacher complexity and a concentration inequality.

Step 2 (Augmentation error decomposition). The per-arm augmentation error $\Delta_{2,r}^2\big(\hat{\mu}_a^{\text{cal}}, \mu_a^r\big)$ is decomposed into estimation/alignment error and embedding sufficiency error:

$$\hat{\mu}_a^{\text{cal}}(\boldsymbol{x}^r) - \mu_a^r(\boldsymbol{x}^r) = \underbrace{\hat{\mu}_a^o(\hat{\phi}^r) + \hat{\delta}_a^t(\hat{\phi}^r) - \tilde{\mu}_a^r(\hat{\phi}^r)}_{\text{(I): estimation and alignment}} + \underbrace{\tilde{\mu}_a^r(\hat{\phi}^r) - \mu_a^r(\boldsymbol{x}^r)}_{\text{(II): sufficiency}},$$

where $\hat{\phi}^r = \hat{\phi}^r(\boldsymbol{x}^r)$ is shorthand.

Term (I): By Assumption 3, $\tilde{\mu}_a^r = \tilde{\mu}_a^o + \delta_a$. By cross-fitting, $\hat{\mu}_a^o$ is estimated on independent OS data and $\hat{\delta}_a^t$ on independent RCT data. Standard empirical process arguments (e.g., Bartlett and Mendelson [2006], Wainwright [2019]) give the OS outcome-model and RCT discrepancy complexities. Because the OS outcome model is trained on $\phi^o(\boldsymbol{X}^o)$ but evaluated on $\phi^r(\boldsymbol{X}^r)$, Assumption 4 and the Lipschitz constants in Assumption 4 add the transfer term $(L_\mu + L_\delta)^2 r_\phi^2$. Thus

$$\mathbb{E}\Big[(\hat{\mu}_a^o(\hat{\phi}^r(\boldsymbol{X}^r)) + \hat{\delta}_a^t(\hat{\phi}^r(\boldsymbol{X}^r)) - \tilde{\mu}_a^r(\hat{\phi}^r(\boldsymbol{X}^r)))^2 \mid S = r\Big] \le C\{\mathfrak{R}_{n^o}^2(\mathcal{M}_a^o) + \mathfrak{R}_{n^r}^2(\mathcal{D}_a) + (L_\mu + L_\delta)^2 r_\phi^2\}.$$

Term (II): Assumption 2 bounds the RCT-side embedding sufficiency error by ϵ_{suff}^2.

Combining Terms (I)–(II) via the Cauchy–Schwarz inequality:

$$\Delta_{2,r}^2\big(\hat{\mu}_a^{\text{cal}}, \mu_a^r\big) \le C\big[\epsilon_{\text{suff}}^2 + (L_\mu + L_\delta)^2 r_\phi^2 + \mathfrak{R}_{n^o}^2(\mathcal{M}_a^o) + \mathfrak{R}_{n^r}^2(\mathcal{D}_a)\big]. \tag{15}$$

Step 3 (Assembly). Substituting (15) into (14) yields the multiplicative term $\{\sum_a \Delta_{2,r}^2\big(\hat{\mu}_a^{\text{cal}}, \mu_a^r\big)\}\mathfrak{R}_{n^r}^2(\mathcal{F})$. The bounded-envelope condition implies $\mathfrak{R}_{n^r}^2(\mathcal{F}) = O(1)$, so this product is bounded by a constant multiple of the augmentation-error

terms in (15); the leading $\mathfrak{R}^2_{n^r}(\mathcal{F})$ term remains as the CATE-regression complexity. Summing over arms gives (13). The $\log(1/\gamma)/n^r$ term arises from a union bound and Bernstein-type concentration applied to the empirical process in the CATE calibration stage. □

Proof of Corollary 2. The CALM bound in Theorem 1 differs from the MR-OSCAR bound of Pal et al. [2026] only in the terms used to make OS information usable under covariate mismatch, up to the comparable calibration and final CATE-regression complexities assumed in the corollary. For CALM, these terms are

$$\epsilon^2_{\text{suff}} + (L_\mu + L_\delta)^2 r^2_\phi + \textstyle\sum_a \mathfrak{R}^2_{n^o}(\mathcal{M}^o_a).$$

For MR-OSCAR, the corresponding terms are the imputation error, the outcome-model complexity after imputation, and the complexity of the imputation class,

$$L^2 r^2_{\text{im}} + \textstyle\sum_a \mathfrak{R}^2_{n^o}(\mathcal{M}^{o,\text{im}}_a) + \mathfrak{R}^2_{n^o}(\mathcal{G}).$$

Subtracting the common terms gives the displayed sufficient condition. □

Proof of Proposition 3. Condition on the training folds used to construct the augmentation $\hat{m}$, so that $\hat{m}(\boldsymbol{X}^r)$ is fixed for the held-out RCT unit. For any such augmentation,

$$\mathbb{E}\left[\frac{A(Y - \hat{m}(\boldsymbol{X}^r))}{\pi^r_A(\boldsymbol{X}^r)} \;\middle|\; \boldsymbol{X}^r, S = r\right] = \{\mu^r_1(\boldsymbol{X}^r) - \hat{m}(\boldsymbol{X}^r)\} - \{\mu^r_{-1}(\boldsymbol{X}^r) - \hat{m}(\boldsymbol{X}^r)\} = \tau^r(\boldsymbol{X}^r),$$

where the first equality uses RCT randomization and positivity. Thus augmentation quality affects the variance of the pseudo-outcome, but not its conditional mean. □

Proof of Proposition 4. Let

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{0}_{p_z \times p_u} & \boldsymbol{I}_{p_z} \\ \boldsymbol{0}_{p_v \times p_u} & \boldsymbol{\Lambda} \end{pmatrix}.$$

Under the Gaussian model and $\boldsymbol{U} \perp\!\!\!\perp \boldsymbol{V} \mid \boldsymbol{Z}$, the L_2-optimal linear prediction of the OS covariate vector from the RCT covariates is

$$\mathbb{E}[(\boldsymbol{Z}, \boldsymbol{V})^\top \mid \boldsymbol{U}, \boldsymbol{Z}] = \boldsymbol{M}(\boldsymbol{U}, \boldsymbol{Z})^\top,$$

because $\mathbb{E}[\boldsymbol{V} \mid \boldsymbol{U}, \boldsymbol{Z}] = \mathbb{E}[\boldsymbol{V} \mid \boldsymbol{Z}] = \boldsymbol{\Lambda}\boldsymbol{Z}$. For a fixed OS projection $\boldsymbol{W}^o$, the best RCT-side linear embedding is therefore the conditional mean of the OS embedding given $\boldsymbol{X}^r$:

$$\mathbb{E}[\boldsymbol{W}^o(\boldsymbol{Z}, \boldsymbol{V})^\top \mid \boldsymbol{X}^r] = \boldsymbol{W}^o \boldsymbol{M}(\boldsymbol{X}^r)^\top.$$

Thus $\boldsymbol{W}^r_{\text{opt}} = \boldsymbol{W}^o\boldsymbol{M}$, which is exactly the embedding obtained by first imputing $\widehat{\boldsymbol{V}} = \boldsymbol{\Lambda}\boldsymbol{Z}$ and then applying $\boldsymbol{W}^o$. □

Proof of Corollary 5. Apply Theorem 1 to the linear embedding in Proposition 4. The residual alignment error is the conditional variation in the missing block after linear imputation, so it is controlled, up to the linear projection constants absorbed by $\lesssim$, by $\text{tr}(\boldsymbol{\Sigma}_{\boldsymbol{V}|\boldsymbol{Z}})$. Standard Rademacher bounds for s-sparse linear classes give squared complexities of order $s \log d/n^o$ for the OS outcome model, $s_\delta \log d/n^r$ for the calibration discrepancy, and $s_\tau \log p_r/n^r$ for the final CATE regression. Substitution gives the displayed rate. □

B ALIGNMENT OBJECTIVE DETAILS

Throughout this section we write $\boldsymbol{w}^o_j = \hat{\phi}^o(\boldsymbol{X}^o_j)$ for the frozen OS embeddings and $\boldsymbol{w}^r_i = \phi^r(\boldsymbol{X}^r_i)$ for the learnable RCT embeddings, both in $\mathbb{R}^d$.

B.1 DISTRIBUTION-MATCHING ALIGNMENT (MMD)

Instantiating (11) with the Gaussian RBF kernel $k(\boldsymbol{w}, \boldsymbol{w}') = \exp\big(-\|\boldsymbol{w} - \boldsymbol{w}'\|^2/(2\sigma^2)\big)$ and expanding the squared RKHS norm yields the unbiased U-statistic estimate

$$\mathcal{L}_{\mathrm{MMD}} = \frac{1}{n^o(n^o-1)} \sum_{\substack{j,j' \in \mathrm{OS} \\ j \neq j'}} \exp\Big(-\frac{\|\boldsymbol{w}_j^o - \boldsymbol{w}_{j'}^o\|^2}{2\sigma^2}\Big) - \frac{2}{n^o n^r} \sum_{j \in \mathrm{OS}} \sum_{i \in \mathrm{RCT}} \exp\Big(-\frac{\|\boldsymbol{w}_j^o - \boldsymbol{w}_i^r\|^2}{2\sigma^2}\Big) + \frac{1}{n^r(n^r-1)} \sum_{\substack{i,i' \in \mathrm{RCT} \\ i \neq i'}} \exp\Big(-\frac{\|\boldsymbol{w}_i^r - \boldsymbol{w}_{i'}^r\|^2}{2\sigma^2}\Big). \tag{16}$$

The bandwidth σ is set via the median heuristic: $\sigma = \sqrt{\mathrm{median}\{\|\boldsymbol{w}_i - \boldsymbol{w}_j\|^2 : i \neq j\}}$, computed over the pooled embeddings from both sources [Gretton et al., 2012]. Since $\hat{\phi}^o$ is frozen, only the cross-term and the RCT–RCT term contribute gradients with respect to ϕ^r.

B.2 $\boldsymbol{Z}$-CONDITIONED CONTRASTIVE ALIGNMENT

The contrastive loss (12) pairs each RCT unit i with its OS neighbours $N_i = \{j \in \mathrm{OS} : \|\boldsymbol{Z}_j - \boldsymbol{Z}_i\| \leq \varepsilon\}$ in shared-covariate space and penalizes the squared embedding distance:

$$\mathcal{L}_{\mathrm{contr}} = \frac{1}{n^r} \sum_{i \in \mathrm{RCT}} \frac{1}{|N_i|} \sum_{j \in N_i} \|\boldsymbol{w}_j^o - \boldsymbol{w}_i^r\|^2. \tag{17}$$

This objective is already in closed form; the neighbourhood radius ε trades off bias (large ε matches dissimilar units) against variance (small ε yields few or no neighbours).

B.3 ADVERSARIAL DOMAIN ALIGNMENT

Following Ganin et al. [2016], we introduce a domain discriminator $g_\omega : \mathbb{R}^d \to [0, 1]$ trained to distinguish OS from RCT embeddings, while the RCT encoder ϕ^r is trained to fool it. Let $s_j = 1$ for OS units and $s_i = 0$ for RCT units. The adversarial alignment objective is

$$\mathcal{L}_{\mathrm{adv}} = -\frac{1}{n^o} \sum_{j \in \mathrm{OS}} \log g_\omega(\boldsymbol{w}_j^o) - \frac{1}{n^r} \sum_{i \in \mathrm{RCT}} \log\big(1 - g_\omega(\boldsymbol{w}_i^r)\big), \tag{18}$$

which is maximized over ω and minimized over ϕ^r via a gradient reversal layer [Ganin et al., 2016]. At the minimax equilibrium, $g_\omega \equiv 1/2$, implying that the embedding distributions are indistinguishable. We omit adversarial alignment from the experiments because it introduces a second network and alternating optimization, making training less stable; we include it here for completeness.

B.4 IMPLEMENTATION NOTES

In the experiments, CALM-NN uses a *conditional-mean alignment* variant, which fits a ridge regression $\boldsymbol{Z} \mapsto \mathbb{E}[\hat{\phi}^o(\boldsymbol{Z}, \boldsymbol{V}) \mid \boldsymbol{Z}]$ on the OS and uses the predicted embeddings as alignment targets for ϕ^r:

$$\mathcal{L}_{\mathrm{cond}} = \frac{1}{n^r} \sum_{i \in \mathrm{RCT}} \big\|\boldsymbol{w}_i^r - \hat{h}(\boldsymbol{Z}_i)\big\|^2, \quad \hat{h} = \operatorname*{arg\,min}_{h \in \mathcal{H}_{\mathrm{ridge}}} \frac{1}{n^o} \sum_{j \in \mathrm{OS}} \|h(\boldsymbol{Z}_j) - \boldsymbol{w}_j^o\|^2 + \alpha \|h\|^2. \tag{19}$$

This variant avoids kernel bandwidth selection and is computationally cheaper but assumes a linear relationship between $\boldsymbol{Z}$ and the OS embeddings. The alignment weight λ is annealed during Stage 2 training: held constant for the first 60% of epochs, then linearly decayed to 0.2λ over the remaining 40%.

C ADDITIONAL EXPERIMENTAL RESULTS

C.1 BASELINE REGIME

This subsection collects additional sweeps from the baseline (linear-CATE) regime described in Section 7. The DGP uses $p_z = 30$, $p_u = 10$, $p_v = 20$, linear outcomes, and default parameters $n^r = 500$, $n^o = 10{,}000$, $\sigma_V^2 = 1.0$, $d_{\text{true}} = 5$, shift magnitude 0.5. Each figure varies one factor while keeping the remaining settings at defaults. Mean RMSE is computed over 20 replicates. Across all settings, the four calibration-based methods (RACER, SR-OSCAR, MR-OSCAR, CALM-Lin) remain effectively indistinguishable (pairwise RMSE differences below 10^{-3}).

Intrinsic dimension and shared proportion (Figures 3, 6). No single method dominates as d_{true} varies; the winner alternates among RACER ($d_{\text{true}} \in \{2, 15\}$), SR-OSCAR (10 and 20), MR-OSCAR (3), and CALM-Lin (5), with small gaps throughout. Similarly, as the shared covariate proportion increases, all methods improve and the identity of the best calibration-based method shifts: CALM-Lin at low and moderate overlap (0.3 and 0.5), MR-OSCAR at intermediate overlap (0.7), and RACER at high overlap (0.9).

RCT sample size (Figure 4). At $n^r = 100$ (baseline linear-CATE regime), HTCE-T yields the lowest RMSE (2.84), while the OS-borrowing calibration variants (SR-OSCAR, MR-OSCAR, CALM-Lin) show instability. For $n^r \geq 250$ the calibration-based methods converge; at $n^r = 2{,}000$, RACER, SR-OSCAR, MR-OSCAR, and CALM-Lin all achieve RMSE ≈ 0.53.

Outcome-model nonlinearity and negative transfer (Figures 5, 7). Under outcome nonlinearity, RACER is best for linear outcomes (RMSE 1.15), CALM-Lin for quadratic (1.45), and MR-OSCAR for sinusoidal (1.06); CALM-NN does not improve over the strongest linear baselines here. Under increasing outcome shift, calibration-based methods remain stable (e.g., RMSE ≈ 1.11 for CALM-Lin at shift 5.0), whereas Naive inflates to 3.84 and HTCE-T to 3.32. CALM-NN is less robust at extreme shift (RMSE 2.25 at shift 5.0), consistent with the limitation that the alignment objective can distort the learned representation when the domain gap is large.

C.2 NONLINEAR-CATE REGIME

This subsection collects additional sweeps from the nonlinear-CATE regime described in Section 7.5. The DGP uses a shared-latent structure: $\boldsymbol{U}$ and $\boldsymbol{V}$ share a 5-dimensional latent factor ($\sigma_V^2 = \sigma_U^2 = 0.1$), outcomes depend only on $\boldsymbol{V}$ ($w_Z = w_U = 0$, $w_V = 2$), and the CATE takes a sinusoidal form with frequency ω. Mean RMSE is computed over 20 replicates. CALM-NN has the lowest RMSE in all settings reported here.

Robustness to signal routing and CATE functional form (Figures 8, 10). To test whether the advantage depends on the outcome signal being routed entirely through unobserved covariates $\boldsymbol{V}$, we vary the shared-covariate signal weight $w_Z \in \{0, 0.5, 1.0, 1.5, 2.0\}$ while holding $w_V = 2$ and $\omega = 1.5$. CALM-NN wins all 5 settings with margins of 0.45–0.59 over calibration-based methods, and the advantage does not diminish as w_Z increases (RMSE 0.62 vs. 1.12 at $w_Z = 2$). This pattern suggests that the gain stems from the ability to model nonlinear CATEs, not from exploiting $\boldsymbol{V}$-specific information inaccessible to imputation-based approaches. The advantage also generalizes across CATE functional forms (absolute-value, quadratic; Figure 10).

Latent coupling strength (Figure 9). We vary $\alpha_U \in \{0.5, 1.0, 2.0, 3.0, 4.0\}$, controlling how much $\boldsymbol{U}$ encodes information about $\boldsymbol{V}$ through the shared latent. The largest margin appears at $\alpha_U = 0.5$ (RMSE 0.57 vs. 1.58); as coupling increases, calibration-based methods improve (from 1.58 to 1.27) while CALM-NN slightly increases (0.57 to 0.87), but CALM-NN remains the best method throughout. This pattern is consistent with the theory: when the latent channel is narrow, imputation is difficult and the nonlinear encoder has a greater relative advantage.

Under absolute-value CATEs, CALM-NN achieves RMSE 2.41 versus 2.86 for the best calibration-based method (margin 0.45); under quadratic CATEs, 15.60 versus 18.83 (margin 3.23). The quadratic form has high RMSE for all methods due to the unbounded square function, but CALM-NN still attains the lowest RMSE (sinusoidal: 0.63 vs. 1.18, margin 0.55).

C.3 SEMI-SYNTHETIC BENCHMARK

We construct a semi-synthetic benchmark from the IHDP dataset [Hill, 2011], splitting covariates into $\boldsymbol{Z}$, $\boldsymbol{U}$, and $\boldsymbol{V}$ and bootstrap-sampling units to form OS and RCT datasets. Outcomes follow a structural model with a treatment effect $\tau(\boldsymbol{X}^r)$

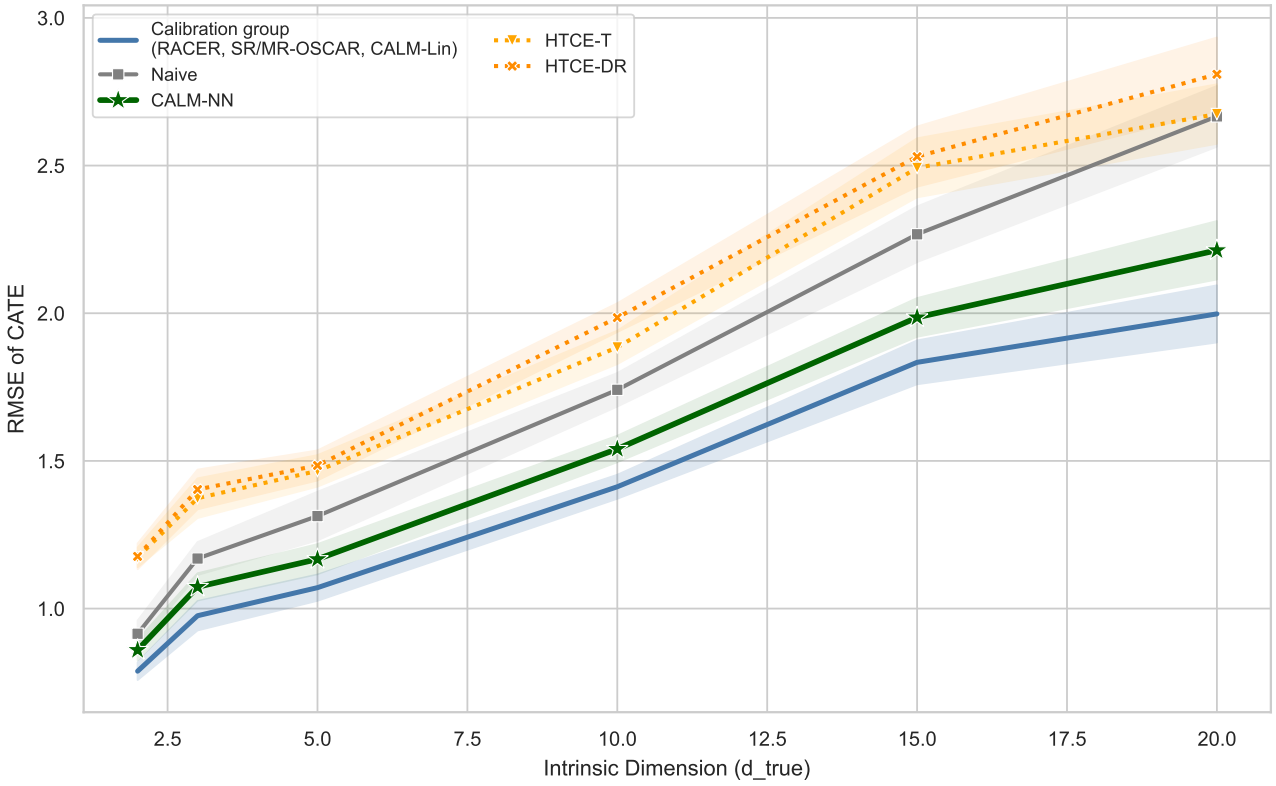


Figure 3: RMSE of CATE estimation as a function of intrinsic dimension d_{true}, with $\sigma_V^2 = 1.0$, $n^r = 500$, and linear outcome model. Mean over 20 replicates. Blue band: calibration-group envelope (see Figure 2 caption). No single method dominates uniformly across intrinsic dimensions.

and an RCT-only shift $\delta_a^r(\boldsymbol{Z})$ to induce domain mismatch. All eight methods from Section 7 are evaluated over 50 replicates. Results show the calibration-based equivalence observed in the linear-CATE simulations: all four calibration methods achieve RMSE ≈ 0.205, with CALM-NN at 0.23 (Figure 11).

D ADDITIONAL THEORY FOR OBSERVABLE DIAGNOSTICS

D.1 OBSERVABLE DIAGNOSTICS FOR THE ALIGNMENT AND SUFFICIENCY TERMS

The leading nonstandard terms in Theorem 1 are r_ϕ^2 and ϵ_{suff}^2. The first is not directly observable because $\boldsymbol{V}$ is unmeasured in the trial; the second is a population approximation error. Under additional conditions, the following held-out quantities upper bound or empirically track the corresponding terms in the bound.

Lemma 6 (Alignment proxy for r_ϕ^2)**.** *Assume Assumption 1. Suppose that, under the target RCT population, $\boldsymbol{U} \perp\!\!\!\perp \boldsymbol{V}^* \mid \boldsymbol{Z}$; the support of $\boldsymbol{Z}$ is compact; the encoders ϕ^o and ϕ^r are Lipschitz on the relevant compact covariate supports; the OS and RCT conditional laws of $\boldsymbol{V} \mid \boldsymbol{Z}$ agree for the counterfactual value $\boldsymbol{V}^*$; and the OS and RCT $\boldsymbol{Z}$ marginals coincide, or else the OS conditional expectations below are reweighted to the RCT $\boldsymbol{Z}$ marginal. Let*

$$m_o(\boldsymbol{z}) = \mathbb{E}\{\phi^o(\boldsymbol{Z}, \boldsymbol{V}) \mid \boldsymbol{Z} = \boldsymbol{z}, S = o\}, \qquad m_r(\boldsymbol{z}) = \mathbb{E}\{\phi^r(\boldsymbol{Z}, \boldsymbol{U}) \mid \boldsymbol{Z} = \boldsymbol{z}, S = r\}.$$

Define

$$\begin{aligned} B_{\text{contr}} &= \mathbb{E}\big[\|\phi^r(\boldsymbol{X}^r) - m_o(\boldsymbol{Z})\|^2 \mid S = r\big], \\ B_{\text{OS}} &= \mathbb{E}\big[\|\phi^o(\boldsymbol{X}^o) - m_o(\boldsymbol{Z})\|^2 \mid S = o\big], \\ B_{\text{RCT}} &= \mathbb{E}\big[\|\phi^r(\boldsymbol{X}^r) - m_r(\boldsymbol{Z})\|^2 \mid S = r\big]. \end{aligned}$$

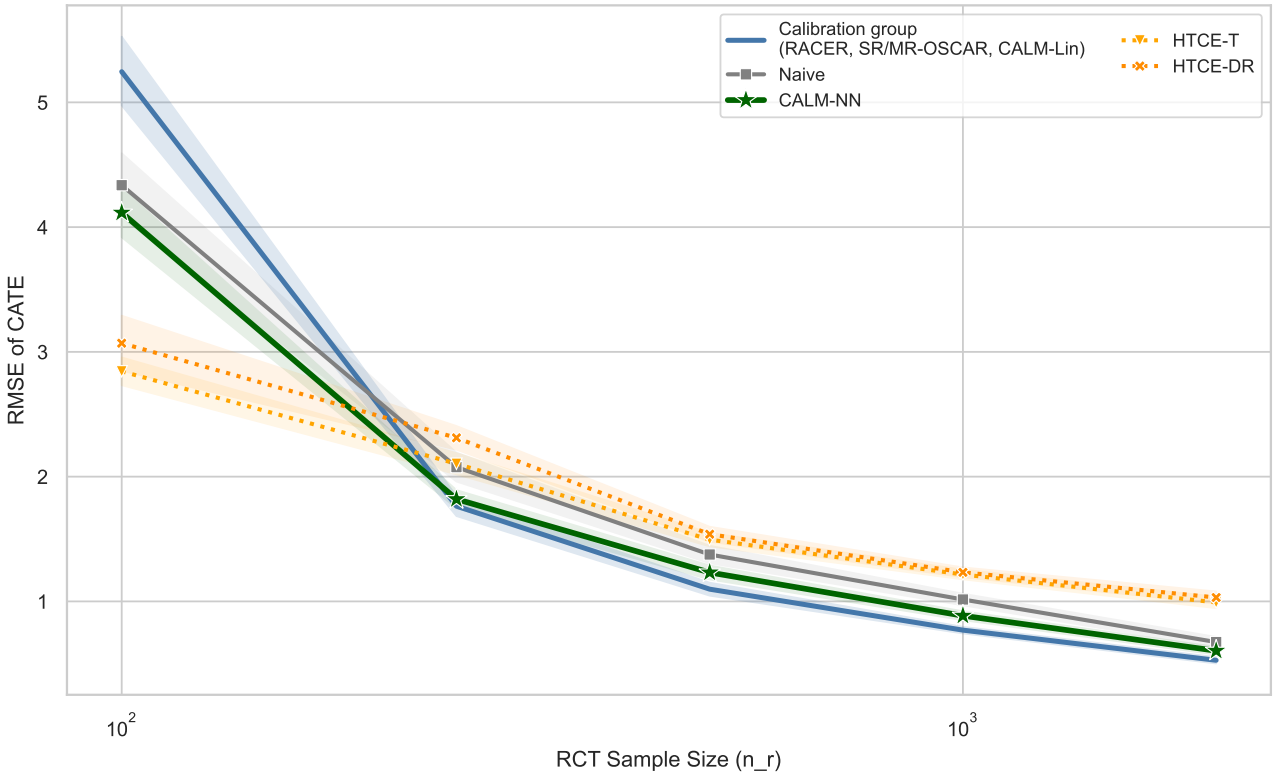


Figure 4: RMSE of CATE estimation as a function of RCT sample size n^r, with $n^o = 10{,}000$ fixed, $\sigma_V^2 = 1.0$, $d_{\text{true}} = 5$, and linear outcome model. Mean over 20 replicates. Blue band: calibration-group envelope (see Figure 2 caption). HTCE methods perform best at the smallest n^r, while calibration-based methods converge as n^r grows.

Then

$$r_\phi^2 \le 3\{B_{\text{contr}} + B_{\text{OS}} + B_{\text{RCT}}\}.$$

If the $\boldsymbol{Z}$ marginals differ across sources, the same display holds with B_{OS} evaluated under the RCT $\boldsymbol{Z}$ marginal, equivalently by weighting OS units by $dP(\boldsymbol{Z} \mid S = r)/dP(\boldsymbol{Z} \mid S = o)$ when this ratio exists and is bounded.

Proof. For an RCT unit, couple the unobserved $\boldsymbol{V}^*$ through its conditional law given $\boldsymbol{Z}$. By conditional transportability of $\boldsymbol{V}^* \mid \boldsymbol{Z}$ and the definition of m_o,

$$\mathbb{E}\big[\|\phi^o(\boldsymbol{Z}, \boldsymbol{V}^*) - m_o(\boldsymbol{Z})\|^2 \mid S = r\big] = B_{\text{OS}},$$

up to the density-ratio weighting noted in the statement. Now write

$$\phi^o(\boldsymbol{Z}, \boldsymbol{V}^*) - \phi^r(\boldsymbol{X}^r) = \{\phi^o(\boldsymbol{Z}, \boldsymbol{V}^*) - m_o(\boldsymbol{Z})\} + \{m_o(\boldsymbol{Z}) - \phi^r(\boldsymbol{X}^r)\}.$$

The two-term inequality $\|u + v\|^2 \le 2\|u\|^2 + 2\|v\|^2$ gives

$$r_\phi^2 \le 2(B_{\text{OS}} + B_{\text{contr}}) \le 3(B_{\text{OS}} + B_{\text{contr}} + B_{\text{RCT}}),$$

where the last step adds the nonnegative RCT residual term. This third term is not needed for the minimal upper bound, but it gives a symmetric diagnostic for RCT-private variation in the learned embedding. □

Empirical diagnostic. We estimate B_{contr} as the held-out distance from each RCT embedding to the OS conditional mean embedding $m_o(\boldsymbol{Z})$, B_{OS} as the OS residual around $m_o(\boldsymbol{Z})$, and B_{RCT} as the RCT residual around $m_r(\boldsymbol{Z})$. The unscaled sum $B_{\text{contr}} + B_{\text{OS}} + B_{\text{RCT}}$ had Pearson correlation 0.9952 with the oracle r_ϕ^2 over 102 CALM-Lin and CALM-NN settings.

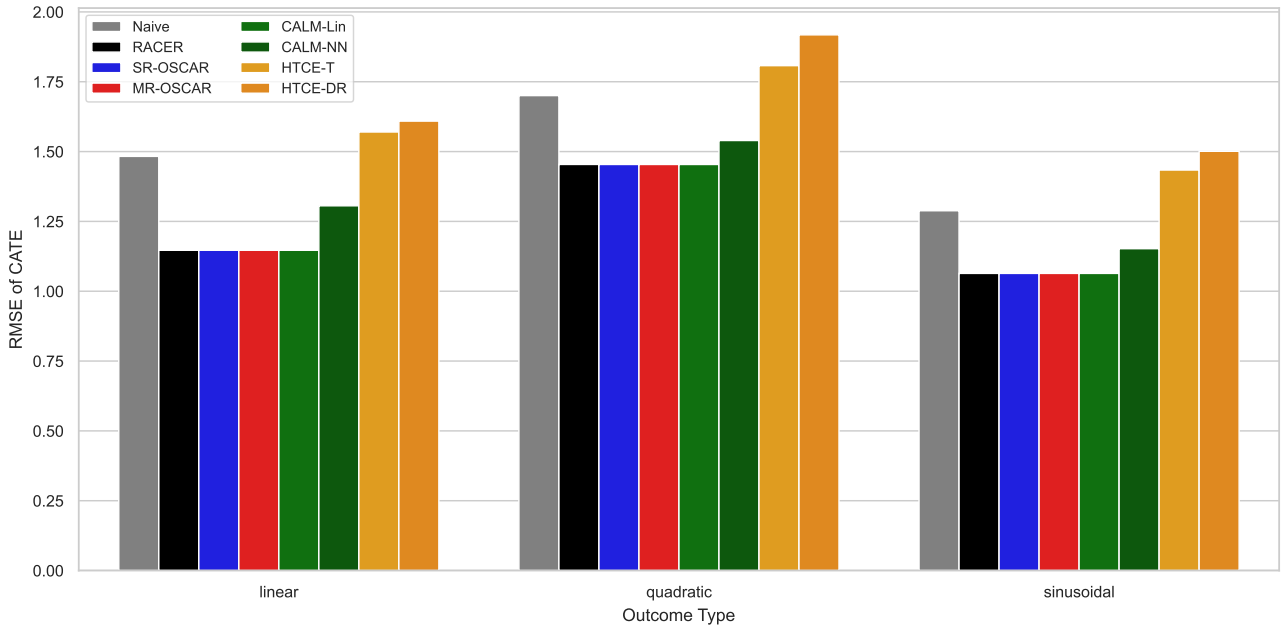


Figure 5: RMSE across outcome model types (linear, quadratic, sinusoidal), with $\sigma_V^2 = 1.0$, $n^r = 500$, $d_{\text{true}} = 5$. Bars show mean RMSE over 20 replicates. The best method depends on the outcome type.

Lemma 7 (Cross-source residual proxy)**.** *Fix an arm a. Let*

$$R^2_{\text{raw},a} = \mathbb{E}\Big[\{Y - \hat{\mu}^o_a(\hat{\phi}^r(\boldsymbol{X}^r))\}^2 \mid S = r, A = a\Big]$$

and

$$R^2_{\text{cal},a} = \mathbb{E}\Big[\{Y - \hat{\mu}^o_a(\hat{\phi}^r(\boldsymbol{X}^r)) - \hat{\delta}^t_a(\hat{\phi}^r(\boldsymbol{X}^r))\}^2 \mid S = r, A = a\Big],$$

both estimated on held-out RCT folds. Suppose Assumptions 1–4 hold, the nuisance estimates are $L_2(P_r)$ consistent for their population limits, and

$$\sigma^2_{\text{noise},a} = \mathbb{E}[\text{Var}\{Y(a) \mid \boldsymbol{X}^r, S = r\} \mid S = r, A = a] < \infty.$$

Then

$$R^2_{\text{raw},a} - \sigma^2_{\text{noise},a} \le 2\mathbb{E}[\delta^2_a(\phi^r(\boldsymbol{X}^r)) \mid S = r, A = a] + 4\epsilon^2_{\text{suff}} + 4(L_\mu + L_\delta)^2 r^2_{\phi,a} + o_p(1),$$
$$R^2_{\text{cal},a} - \sigma^2_{\text{noise},a} \le 2\epsilon^2_{\text{suff}} + 2(L_\mu + L_\delta)^2 r^2_{\phi,a} + o_p(1),$$

where

$$r^2_{\phi,a} = \mathbb{E}[\|\phi^o(\boldsymbol{X}^{o,*}) - \phi^r(\boldsymbol{X}^r)\|^2 \mid S = r, A = a].$$

Proof. Work at the population nuisance limits; the held-out consistency assumptions add the $o_p(1)$ terms. For the raw residual, Assumption 3 gives the exact three-term decomposition

$$Y(a) - \tilde{\mu}^o_a(\phi^r(\boldsymbol{X}^r)) = \underbrace{Y(a) - \mu^r_a(\boldsymbol{X}^r)}_{\eta_a} + \underbrace{\mu^r_a(\boldsymbol{X}^r) - \tilde{\mu}^r_a(\phi^r(\boldsymbol{X}^r))}_{e^r_{\text{suff},a}} + \underbrace{\tilde{\mu}^r_a(\phi^r(\boldsymbol{X}^r)) - \tilde{\mu}^o_a(\phi^r(\boldsymbol{X}^r))}_{\delta_a(\phi^r(\boldsymbol{X}^r))}.$$

The noise term η_a has conditional mean zero given $\boldsymbol{X}^r, S = r, A = a$, and therefore contributes $\sigma^2_{\text{noise},a}$. Thus

$$R^2_{\text{raw},a} - \sigma^2_{\text{noise},a} \le 2\mathbb{E}[(e^r_{\text{suff},a})^2 \mid S = r, A = a] + 2\mathbb{E}[\delta^2_a(\phi^r(\boldsymbol{X}^r)) \mid S = r, A = a] + o_p(1).$$

Assumption 2 controls the RCT-side sufficiency residual. The displayed bounds keep the alignment contribution from Theorem 1 as a nonnegative diagnostic term, which allows the residual check to be read on the same scale as the main bound. Substitution gives the raw-residual bound. For the calibrated residual, the discrepancy term is subtracted:

$$Y(a) - \{\tilde{\mu}^o_a(\phi^r(\boldsymbol{X}^r)) + \delta_a(\phi^r(\boldsymbol{X}^r))\} = \eta_a + e^r_{\text{suff},a}.$$

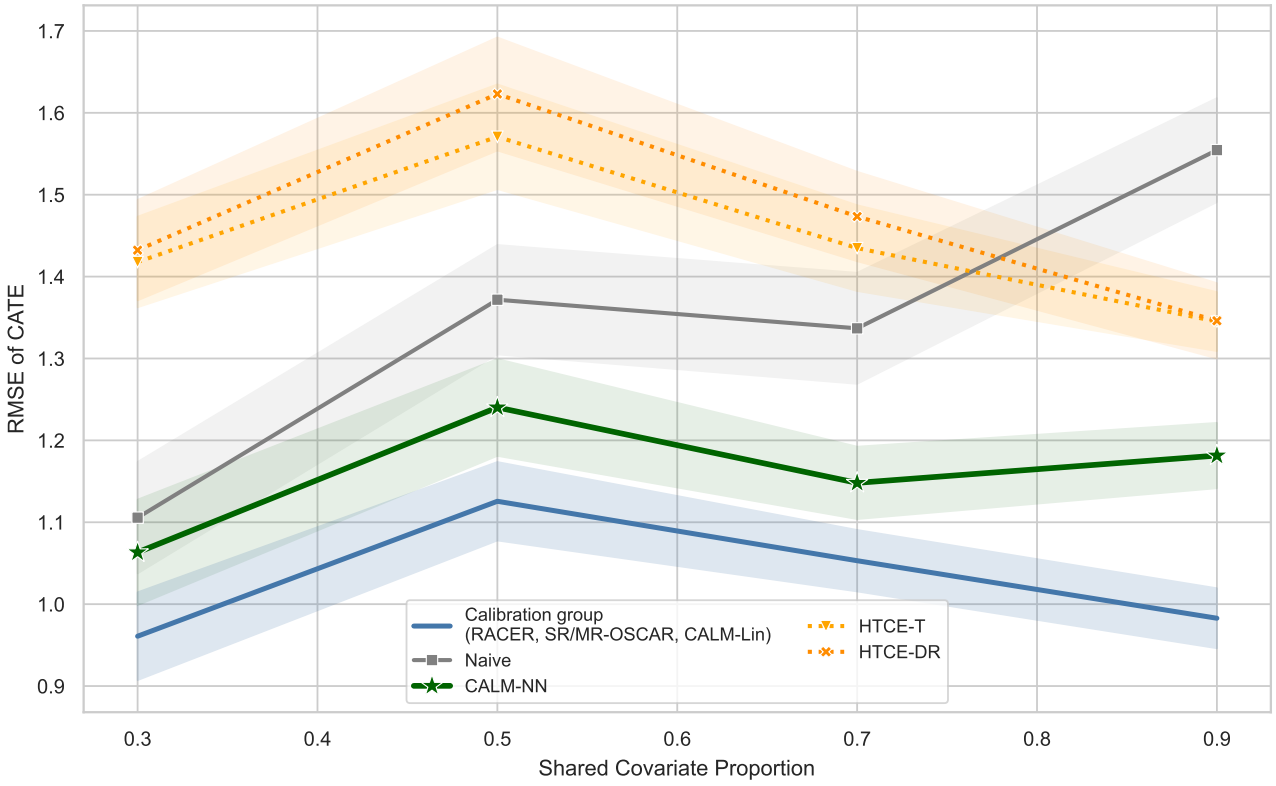


Figure 6: RMSE as a function of shared covariate proportion $p_z/(p_z + p_v)$, with $\sigma_V^2 = 1.0$, $n^r = 500$, $d_{\text{true}} = 5$, and linear outcome model. Mean over 20 replicates. Blue band: calibration-group envelope (see Figure 2 caption). Performance improves as shared proportion increases and the mismatch problem weakens.

The same argument for $e^r_{\text{suff},a}$ gives the second display. No equality up to $o_p(1)$ is claimed, since these residual components are not orthogonal in general. □

Empirical diagnostic. We report the raw and calibrated held-out RCT residuals $R^2_{\text{raw},a}$ and $R^2_{\text{cal},a}$ to check whether transferred OS predictions remain accurate after calibration. Their alignment component is read alongside the three-term alignment proxy above, whose unscaled sum had Pearson correlation 0.9952 with the oracle r^2_ϕ over 102 settings.

Lemma 8 (OS residual proxy for ϵ^2_{suff})**.** *Fix an arm a. Let*

$$R^2_{\text{suff,OS},a} = \mathbb{E}\Big[\{Y - \hat{\mu}^o_a(\hat{\phi}^o(\boldsymbol{X}^o))\}^2 \mid S = o, A = a\Big],$$

estimated on a held-out OS fold independent of the fold used to train $(\hat{\phi}^o, \hat{\mu}^o_a)$. Suppose Assumption 2 holds for the OS source, $\hat{\mu}^o_a \circ \hat{\phi}^o$ is $L_2(P_o)$ consistent for $\tilde{\mu}^o_a \circ \phi^o$, and

$$\sigma^2_{\text{noise},o,a} = \mathbb{E}[\text{Var}\{Y(a) \mid \boldsymbol{X}^o, S = o\} \mid S = o, A = a] < \infty.$$

Then

$$R^2_{\text{suff,OS},a} - \sigma^2_{\text{noise},o,a} = \mathbb{E}\big[\{\mu^o_a(\boldsymbol{X}^o) - \tilde{\mu}^o_a(\phi^o(\boldsymbol{X}^o))\}^2 \mid S = o, A = a\big] + o_p(1) \le \epsilon^2_{\text{suff}} + o_p(1).$$

Proof. On the held-out fold,

$$Y(a) - \tilde{\mu}^o_a(\phi^o(\boldsymbol{X}^o)) = \{Y(a) - \mu^o_a(\boldsymbol{X}^o)\} + \{\mu^o_a(\boldsymbol{X}^o) - \tilde{\mu}^o_a(\phi^o(\boldsymbol{X}^o))\}.$$

The first term has conditional mean zero given $\boldsymbol{X}^o, S = o, A = a$, so its cross-product with the second term has expectation zero. The first term therefore contributes $\sigma^2_{\text{noise},o,a}$, and the second is the embedding sufficiency residual. Held-out $L_2(P_o)$ consistency replaces the population predictor by $\hat{\mu}^o_a \circ \hat{\phi}^o$ at an $o_p(1)$ cost, and Assumption 2 bounds the residual by ϵ^2_{suff}. □

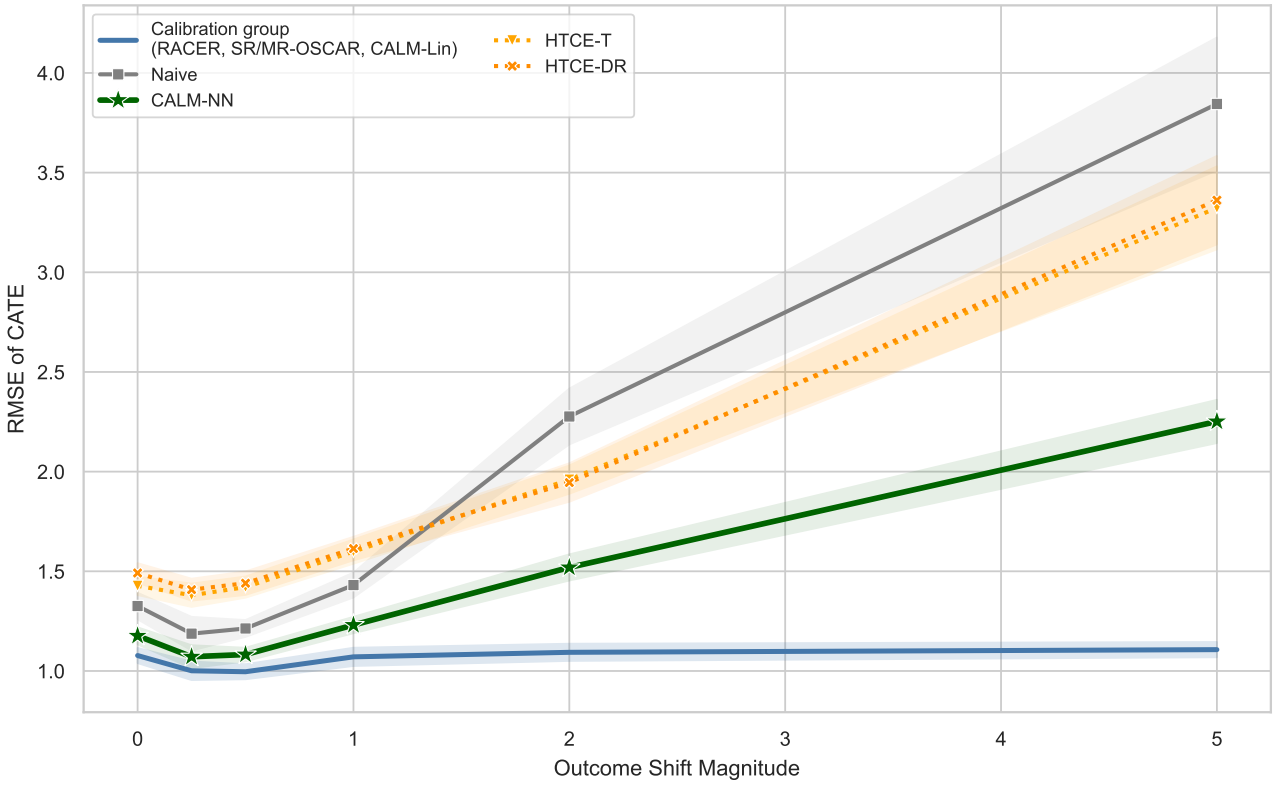


Figure 7: RMSE under increasing outcome shift between OS and RCT, with $\sigma_Y^2 = 1.0$, $n^r = 500$, and linear outcome model. Mean over 20 replicates. Blue band: calibration-group envelope (see Figure 2 caption). Calibration-based methods remain stable under shift, while Naive and HTCE methods inflate under large shift.

Empirical diagnostic. This held-out OS outcome residual is the sufficiency-side diagnostic read alongside the calibrated RCT residual. It does not itself target r_ϕ^2, which is checked by the alignment proxy above.

D.2 CALIBRATION OF THE CONSTANT C

In the sparse linear setting of Corollary 5, the constant in Theorem 1 can be calibrated using standard bounded-design and sub-Gaussian-noise quantities. Suppose that, for each arm a, $Y(a) - \mu_a^r(\boldsymbol{X}^r)$ is conditionally σ_Y^2-sub-Gaussian, the RCT overlap constant in Assumption 1 satisfies $\rho \le \pi_a^r(\boldsymbol{X}^r) \le 1 - \rho$, and the normalized linear design has empirical Gram eigenvalues bounded away from zero and infinity. A Bernstein inequality for the inverse-propensity-weighted pseudo-outcome noise yields, with probability at least $1 - \gamma$,

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n^r} \sum_{i:S_i=r} \{f(\boldsymbol{X}_i^r) - \tau^r(\boldsymbol{X}_i^r)\}\xi_i \right| \le c_0 \frac{\sigma_Y}{\rho} \sqrt{\frac{\log(p_o + p_r) + \log(1/\gamma)}{n^r}},$$

where ξ_i is the centered pseudo-outcome noise and c_0 is a universal constant. Squaring and absorbing the usual empirical-process constants gives

$$C \le c \, \frac{\sigma_Y^2}{\rho^2} \log(p_o + p_r).$$

For the representative simulation configuration

$$\sigma_Y \approx 1, \qquad \rho = 0.5, \qquad p_o = 50, \qquad p_r = 40, \qquad n^r = 500,$$

we have $\log(p_o + p_r) = \log(90) \approx 4.50$. Taking the conservative constant $c = 16$ gives

$$C \le 16 \cdot \frac{1}{0.5^2} \cdot 4.50 \approx 288.$$

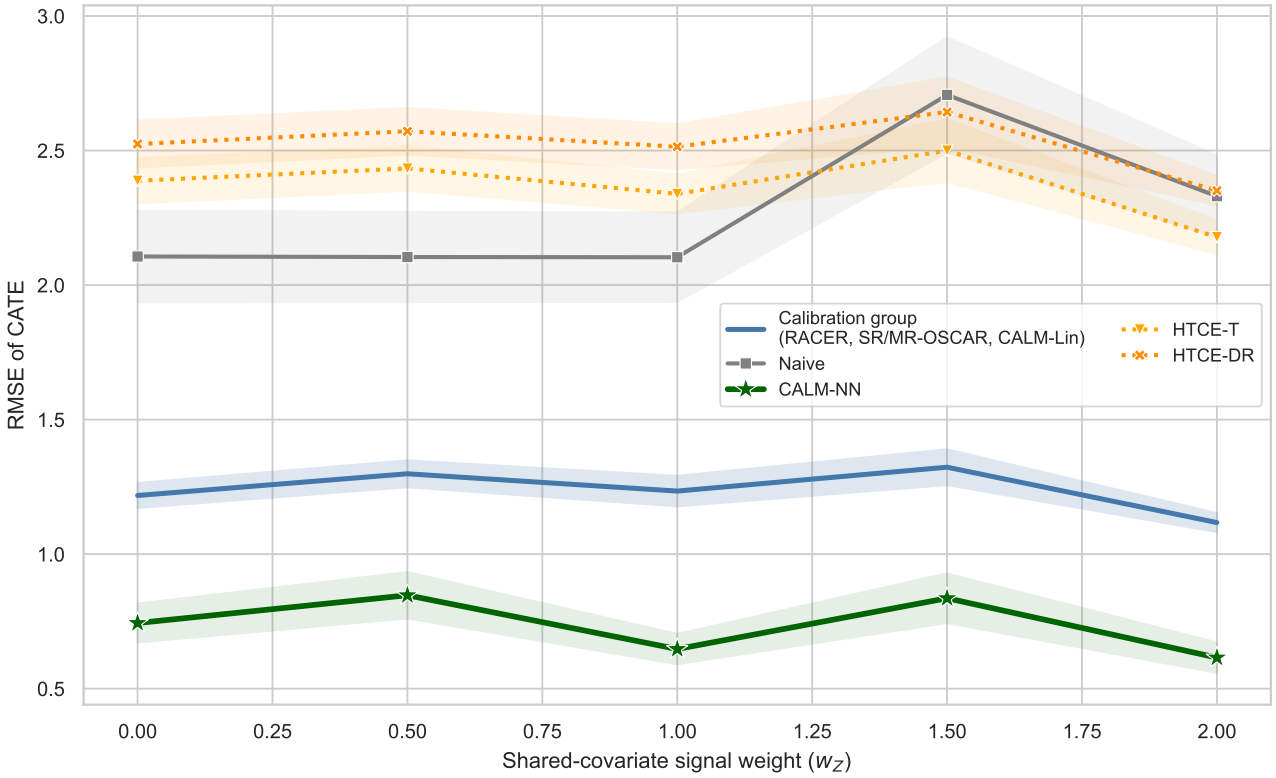


Figure 8: RMSE as a function of shared-covariate signal weight w_Z, with $w_V = 2$, $\omega = 1.5$, and the nonlinear-CATE DGP. Mean over 20 replicates. Blue band: calibration-group envelope (see Figure 2 caption). CALM-NN's advantage persists even when $\boldsymbol{Z}$ contributes substantially to the outcome signal.

At $\gamma = 0.05$, the confidence contribution in Theorem 1 is therefore

$$C\frac{\log(1/\gamma)}{n^r} \leq 288\frac{\log(20)}{500} \approx 1.73.$$

This calibration is intentionally conservative: it includes the worst-case $1/\rho^2$ overlap factor, a coordinatewise union bound, and a non-sharp universal constant.

D.3 REPORTING THE LIPSCHITZ CONSTANTS

The constants in Assumption 4 are prediction-head constants: L_μ controls the OS outcome head and L_δ controls the calibration head. For the diagnostics in this paper we report spectral-product upper bounds for those heads only. The neural encoders also have formal spectral-product bounds, but those values are vacuous in the residual-MLP architecture because conservative LayerNorm handling multiplies by $\max|\gamma|/\sqrt{\varepsilon}$; we therefore do not report them as L_μ or L_δ.

These are upper bounds, not exact Lipschitz constants. The neural stages use Adam with weight decay 10^{-4} but no spectral normalization, so the values should be interpreted as regularization-controlled diagnostics rather than certified global constants. Encoder smoothness is controlled only indirectly by the same weight decay and could be tightened with norm-constrained architectures. Exact Lipschitz computation for ReLU networks is NP-hard in general; post-hoc certified bounds would require methods such as those of Anil et al. [2019] and Fazlyab et al. [2019].

D.4 EXTENSION TO MULTIPLE OBSERVATIONAL SOURCES

The same argument extends to one RCT and K observational sources. Assume the K OS sources are statistically independent, as is standard when they are distinct studies; otherwise the bound gains cross-source covariance terms. Let source k have

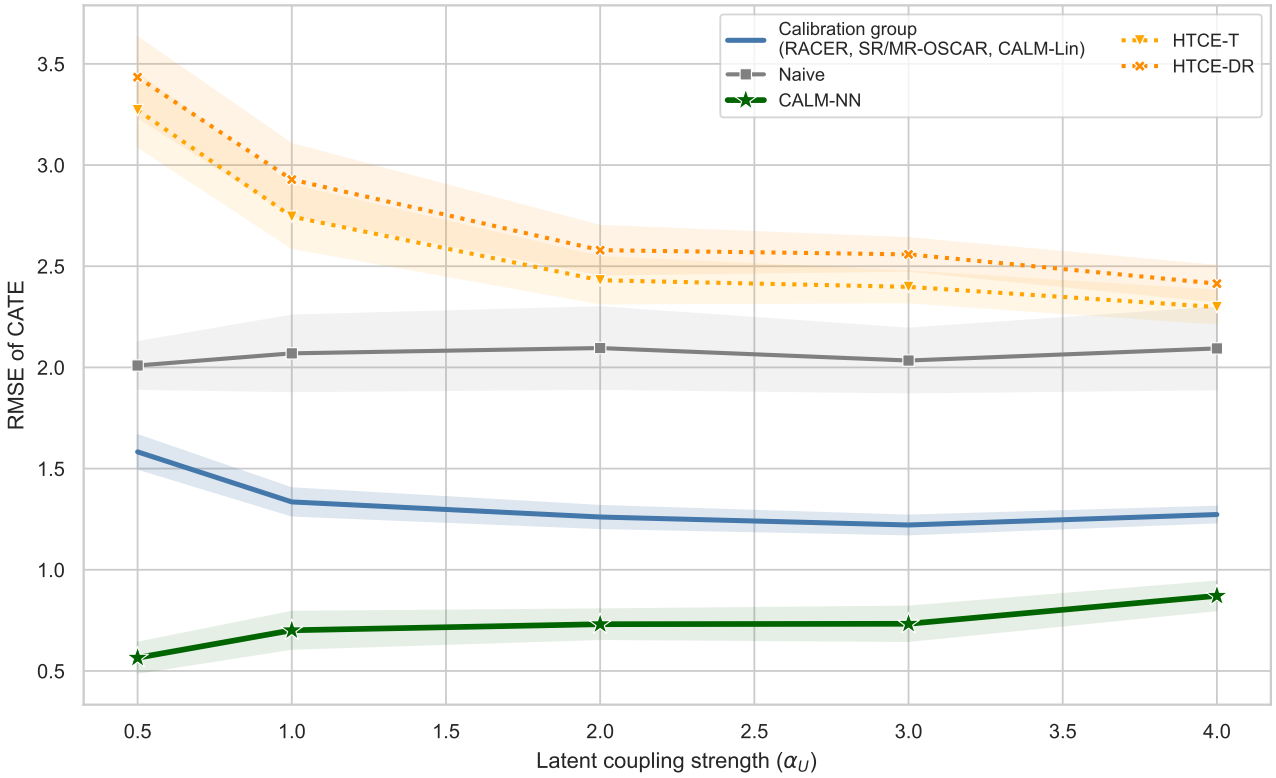


Figure 9: RMSE as a function of latent coupling strength α_U, which controls how much $\boldsymbol{U}$ encodes information about $\boldsymbol{V}$ through shared latent factors. Nonlinear-CATE DGP with $\omega = 1.5$, $w_Z = 0$. Mean over 20 replicates. Blue band: calibration-group envelope (see Figure 2 caption).

encoder $\phi^{o,k}$, outcome classes $\mathcal{M}^o_{a,k}$, discrepancy classes $\mathcal{D}_{a,k}$, sufficiency gap $\epsilon^2_{\text{suff},k}$, alignment error $r^2_{\phi,k}$, and Lipschitz constants $L_{\mu,k}$ and $L_{\delta,k}$ with respect to the shared RCT encoder ϕ^r. For weights $w_k \geq 0$ with $\sum_{k=1}^K w_k = 1$, define the aggregated calibrated arm-specific prediction

$$\hat{\mu}^{\text{multi}}_a(\boldsymbol{X}^r) = \sum_{k=1}^K w_k \{\hat{\mu}^{o,k}_a(\hat{\phi}^r(\boldsymbol{X}^r)) + \hat{\delta}^{t,k}_a(\hat{\phi}^r(\boldsymbol{X}^r))\}.$$

Using this prediction in the CMO construction and final pseudo-outcome regression, the proof of Theorem 1 gives, with high probability,

$$\begin{aligned}\Delta_2^2(\hat{\tau}^{\text{multi}}_{\text{CALM}}, \tau^r) \lesssim \Delta_2^2(\mathcal{F}, \tau^r) + \frac{\log(1/\gamma)}{n^r} \\ + \sum_{k=1}^K w_k^2 \left[\epsilon^2_{\text{suff},k} + (L_{\mu,k} + L_{\delta,k})^2 r^2_{\phi,k} + \sum_a \mathfrak{R}^2_{n^{o,k}}(\mathcal{M}^o_{a,k})\right] \\ + \sum_{k=1}^K \sum_a \mathfrak{R}^2_{n^r}(\mathcal{D}_{a,k}) + \mathfrak{R}^2_{n^r}(\mathcal{F}).\end{aligned}$$

The w_k^2 factor is the usual variance scaling for a convex combination of independent source-specific augmentation errors. If v_k denotes the bracketed source-specific error plus its estimated finite-sample variance, the oracle inverse-variance weights are

$$w_k^\star = \frac{v_k^{-1}}{\sum_{\ell=1}^K v_\ell^{-1}}.$$

This aggregation connects to the broader literature on combining one RCT with multiple observational sources [Brantner

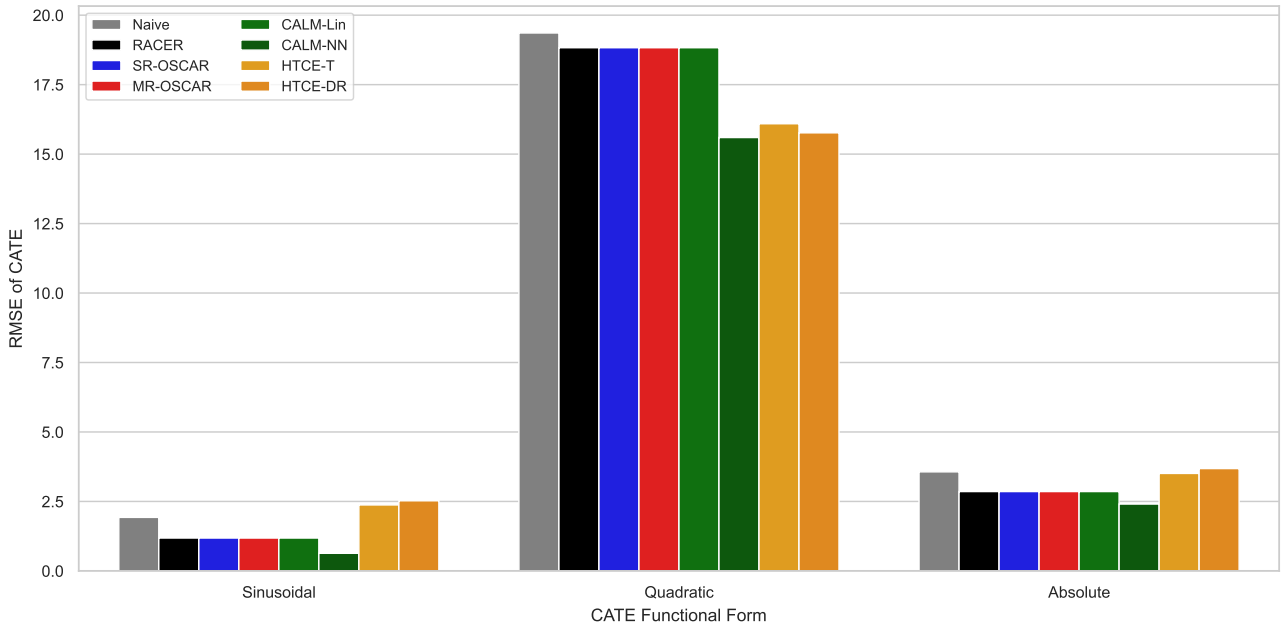


Figure 10: RMSE across three nonlinear CATE functional forms (sinusoidal, quadratic, absolute value) in the nonlinear-CATE DGP ($\omega = 1.5$, $w_Z = 0$). Bars show mean RMSE over 20 replicates. CALM-NN (dark green) achieves the lowest RMSE across all three forms.

Table 4: Spectral-product upper bounds for the outcome-head (L_μ) and calibration-head (L_δ) maps in CALM-NN. Treated corresponds to $a = +1$ and control to $a = -1$. On GPS the observational cohort has a single (control) arm, so the treated-arm outcome head is not learned from OS data and is marked n/a.

Setting	$L_{\mu,+1}$	$L_{\mu,-1}$	$L_{\delta,+1}$	$L_{\delta,-1}$
Synthetic baseline linear	0.936	0.866	0.650	0.591
Synthetic nonlinear private-$\boldsymbol{V}$	0.949	0.847	0.780	0.690
GPS real data, EHR-restricted	n/a	0.548	0.914	0.541

et al., 2023, Colnet et al., 2024, Yang et al., 2025, Lee et al., 2023, Hatt et al., 2022]; per-source bias can be controlled with a sensitivity-analysis module such as the two-parameter framework of Yu et al. [2025].

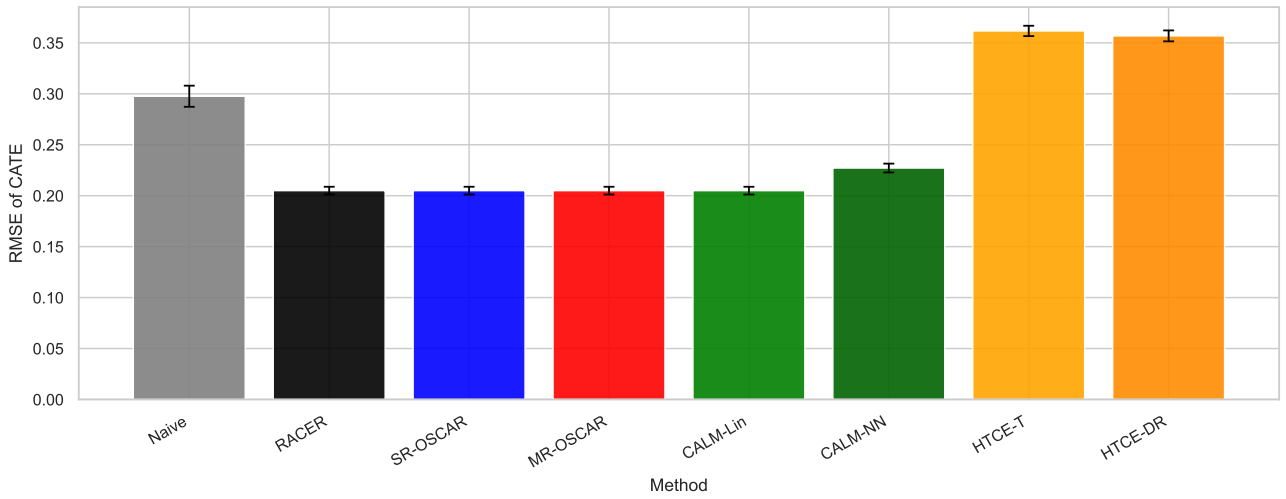


Figure 11: Semi-synthetic benchmark using IHDP covariates: mean RMSE of CATE estimation over 50 replicates (error bars: ± 1 SE).