FEATURE ATTRIBUTION FROM MULTI-SOURCE DATA WITH HETEROGENEOUS OBSERVATION PATTERNS

Anonymous authors
Paper under double-blind review

ABSTRACT

In multi-source learning, different sites typically observe different feature subsets, so pooling data across sites produces blockwise missingness and distributional shift. Shapley-based feature attribution has become a standard tool for model interpretability, yet existing estimators require fully observed evaluation data. We propose **DASH** (**D**ata-fused **A**ttribution via **Sh**apley), which derives the influence function of the constrained weighted-least-squares Shapley estimator and uses it to construct site-specific control variates from partially observed auxiliary sites, reducing variance without imputing missing features. A permutation-based screening step excludes incompatible sources, and data-adaptive calibration optimally weights each site. In simulations, DASH achieves 3–8$\times$ lower MSE than the single-site estimator and 2–3$\times$ lower error than imputation baselines, with 82% screening power at moderate misalignment. On environmental monitoring and multi-center clinical data, DASH reduces error by 3–7$\times$ over the single-site baseline, while imputation can increase error in the clinical setting.

1 INTRODUCTION

Model interpretability is central to trustworthy machine learning in high-stakes domains such as medicine, credit scoring, and criminal justice. Among approaches to post-hoc explanation, Shapley value (SV)-based feature attribution (Lundberg & Lee, 2017; Covert et al., 2020) has emerged as the principled standard, owing to its axiomatic foundation and model-agnostic applicability (Molnar, 2020; Mosca et al., 2022; Li et al., 2024; Salih et al., 2025). Given p input features and a predictive model f, SV methods define a cooperative game $v(S)$ that measures the predictive performance of f when only features in $S \subseteq \{1, \ldots, p\}$ are available, with absent features marginalized over a reference distribution. The SV ϕ_j of feature j is then the weighted average of j's marginal contribution $v(S \cup \{j\}) - v(S)$ across all 2^{p-1} coalitions not containing j; estimation of $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)$ requires evaluating $v(S)$ for many coalitions against a reference sample.

A fundamental assumption of existing population-level SV estimators, including KernelSHAP (Lundberg & Lee, 2017; Covert & Lee, 2021), SAGE (Covert et al., 2020), and FastSHAP (Jethani et al., 2022), is that the reference sample is drawn from a single, fully-observed data source. This assumption is routinely violated in multi-site studies across biomedicine, sensor networks, and scientific consortia, where different institutions measure different variables, producing systematic blockwise missingness. Two natural remedies are inadequate. Computing SV independently per site and averaging ignores distributional differences and offers no mechanism for detecting incompatible sources. Imputing missing features before SV estimation alters the covariance structure on which attributions depend, introducing coalition-dependent bias that does not diminish with improved imputation accuracy (Vo et al., 2025). We briefly review three relevant lines of work.

Shapley values under incomplete data. Computing SV requires specifying how absent features are handled within each coalition S; the choice among conditional, marginal, and baseline removal strategies materially affects the resulting attributions (Chen et al., 2023; Covert et al., 2021), and which convention is preferable remains open and context-dependent. The majority of recent methodological work has focused on improving the computational efficiency of SV estimation under a fixed removal convention (Covert & Lee, 2021; Mitchell et al., 2022), leaving the statistical challenge of heterogeneous, partially-observed evaluation data largely unaddressed.

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

A compounding difficulty arises when the evaluation data themselves contain missing entries. Standard imputation yields a surrogate distribution $\widetilde{P}_X \neq P_X$ (Shannon et al., 2026; Näf et al., 2026). Since all common SV formulations define the coalitional game through expectations with respect to P_X, this distributional error propagates directly into $\mathcal{V}(S)$ and hence into every attribution. Empirically, Vo et al. (2025) confirm that different imputation strategies produce systematically divergent Shapley attributions, yet no existing work corrects this bias under structured blockwise missingness.

Data fusion under blockwise missingness. Several recent works address estimation from multi-source data with heterogeneous variable coverage. Xue & Qu (2021) integrate conditional-mean imputations across block-missing pattern groups within a penalized generalized method of moments framework. Jin & Rothenhäusler (2023) leverage auxiliary variables under conditional independence to improve estimation efficiency. Li et al. (2025) develop a control-variate framework handling both blockwise missingness and distributional shift, and Xu et al. (2025) and Huang et al. (2025) extend similar ideas to broader parameter classes. In all such cases, the target is a single estimating equation or a small system of moment conditions. The Shapley vector $\boldsymbol{\phi} \in \mathbb{R}^p$, while also finite-dimensional, is defined through 2^p coalition-level value functions, a combinatorial structure absent from prior data-fusion targets.

Shapley values in multi-site and federated settings. Several works employ Shapley values in multi-site contexts but target fundamentally different estimands. One active line assigns a single SV to each *site*, quantifying how much each data source contributes to the overall model rather than to each *feature* within it (Wang et al., 2020; Zheng et al., 2023; Liu et al., 2022). Wu et al. (2021) use SV to explain performance disparities across clinical sites but treat site-level confounders as the players rather than model features. In all these formulations, each site has access to the same feature space; the heterogeneous feature coverage that defines blockwise missingness is absent.

We propose **DASH** (**D**ata-fused **A**ttribution via **Sh**apley) for estimating Shapley feature attributions from multi-site data with blockwise-missing covariates without imputation. (1) Extending the empirical findings of Vo et al. (2025), we show that imputing missing features before computing SV introduces systematic, coalition-dependent bias across the standard imputation methods considered. (2) We derive the influence function of the constrained WLS Shapley estimator, compressing coalition-level perturbations into a p-dimensional influence vector that enables control-variate fusion from blockwise-missing auxiliary sites without imputation. (3) We develop a permutation-based screening procedure that detects incompatible sites and a total-variance calibration that optimally weights each site. Across simulations and real data, DASH achieves 3–8$\times$ lower MSE than the single-site estimator and 2–3$\times$ lower than imputation baselines.

2 Problem Formulation

Let Y denote the outcome of interest and $\mathbf{X} = (X_1, \ldots, X_p)^\top$ a p-dimensional feature vector. We consider a multi-site setting with three types of data source. A *labeled complete* ($\mathcal{LC}$) of size n, where both $(Y, \mathbf{X})$ are jointly observed. R *labeled missing* ($\mathcal{LM}_r$, $r = 1, \ldots, R$) sites of size n_r, where the outcome Y and a subset $\mathbf{X}_{\Gamma_r}$ of covariates are observed, where $\Gamma_r \subsetneq \{1, \ldots, p\}$, but the remaining covariates $\mathbf{X}_{\Gamma_r^c}$ are entirely unobserved. An *Unlabeled complete* ($\mathcal{UC}$), of size $N \gg n$, where all covariates $\mathbf{X}$ are observed but Y is unavailable.

The $\mathcal{UC}$ sample defines the target population on which inference is desired. We assume centralized access to row-level data from all sources. Throughout, we denote by $\mathcal{D} = \{1, \ldots, p\}$ the full index set and $\rho_r = n_r/n$ the sample-size ratio of the r-th labeled-missing source to the complete source. Sources may differ in their marginal covariate distributions; identification relies on conditional alignment, formalized below.

Assumption 1 (Missing at random with sufficient alignment)**.** For each site $\mathcal{LM}_r$, $r = 1, \ldots, R$, there exists a sufficient alignment set $\Omega_r \subseteq \Gamma_r$ such that

$$p_{\mathcal{UC}}\big(Y, \mathbf{X}_{\Omega_r^c} \mid \mathbf{X}_{\Omega_r}\big) \;=\; p_{\mathcal{LM}_r}\big(Y, \mathbf{X}_{\Omega_r^c} \mid \mathbf{X}_{\Omega_r}\big)\,. \tag{1}$$

Assumption 1 requires that, conditional on the alignment variables $\mathbf{X}_{\Omega_r}$, the joint distribution of the outcome and remaining features is the same at site r and in the target population. This generalizes the missing-completely-at-random (MCAR) condition commonly adopted in the blockwise-missing literature (Xue & Qu, 2021; Jin & Rothenhäusler, 2023): when $\Omega_r = \varnothing$, equation 1 reduces to

MCAR; when $\Omega_r = \Gamma_r$, arbitrary marginal shift in $\mathbf{X}$ is permitted provided the conditional distributions agree (Li et al., 2025).

We now define the Shapley feature attribution estimand. Let $f : \mathbb{R}^p \to \mathbb{R}$ be a fixed, pre-trained predictive model and $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ a loss function. For each coalition $S \subseteq \mathcal{D}$, define the restricted prediction

$$\bar{f}_S(\mathbf{x}_S) \;=\; \mathbb{E}_{p_{\mathcal{UC}}}\big[f(\mathbf{x}_S, \mathbf{X}_{S^c})\big], \tag{2}$$

which marginalizes the absent features $\mathbf{X}_{S^c}$ over their marginal distribution under $p_{\mathcal{UC}}$, independently of $\mathbf{X}_S$. This is the marginal feature removal convention (Lundberg & Lee, 2017); the framework extends to the conditional alternative (see Appendix A).

The SAGE cooperative game (Covert et al., 2020) assigns to each coalition $S \subseteq \mathcal{D}$ the value

$$\mathcal{V}(S) \;=\; -\,\mathbb{E}_{p_{\mathcal{UC}}}\big[\,\ell\big(\bar{f}_S(\mathbf{X}_S),\; Y\big)\big], \tag{3}$$

the negated expected loss under coalition S, with larger values indicating better predictive performance. Encoding coalitions as binary vectors $\mathbf{z} \in \{0,1\}^p$ via $S(\mathbf{z}) = \{j : z_j = 1\}$, we write $\mathcal{V}(\mathbf{z})$ and $\mathcal{V}(S)$ interchangeably. The population value function admits a per-observation decomposition $\mathcal{V}(\mathbf{z}) = \mathbb{E}_{p_{\mathcal{UC}}}[\eta(\mathbf{z}, \mathbf{X}, Y)]$, where

$$\eta(\mathbf{z}, \mathbf{x}, y) \;=\; -\,\ell\big(\bar{f}_{S(\mathbf{z})}(\mathbf{x}_{S(\mathbf{z})}),\; y\big) \tag{4}$$

records the negated loss for a single observation $(\mathbf{x}, y)$ under coalition $\mathbf{z}$. The SV of feature j is the weighted average of its marginal contribution $\mathcal{V}(S\cup\{j\}) - \mathcal{V}(S)$ over all coalitions $S \not\ni j$, uniquely characterized by the efficiency, symmetry, linearity, and null-player axioms (Covert et al., 2020). Equivalently, $\bar{\boldsymbol{\phi}} = (\phi_1, \ldots, \phi_p)^\top$ is the solution to the constrained weighted least squares (WLS) problem (Lundberg & Lee, 2017; Covert & Lee, 2021)

$$\bar{\boldsymbol{\phi}} \;=\; \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \; \mathbb{E}_{\mu_{\mathrm{Sh}}}\Big[\big(\mathcal{V}(\mathbf{0}) + \mathbf{z}^\top\boldsymbol{\beta} - \mathcal{V}(\mathbf{z})\big)^2\Big] \quad \text{s.t.} \quad \mathbf{1}^\top\boldsymbol{\beta} = \mathcal{V}(\mathbf{1}) - \mathcal{V}(\mathbf{0}), \tag{5}$$

where μ_{Sh} is the Shapley kernel, the distribution over coalitions of intermediate size ($0 < |\mathbf{z}| < p$) with probability mass $\mu_{\mathrm{Sh}}(\mathbf{z}) \propto [\binom{p}{|\mathbf{z}|}\,|\mathbf{z}|\,(p - |\mathbf{z}|)]^{-1}$, and the efficiency constraint ensures that the attributions sum to the difference between full-model and null-model performance. The KKT conditions yield (Covert & Lee, 2021)

$$\bar{\boldsymbol{\phi}} \;=\; \Sigma^{-1}\left[\mathbf{b} + \mathbf{1}\,\frac{c - \mathbf{1}^\top\Sigma^{-1}\mathbf{b}}{\mathbf{1}^\top\Sigma^{-1}\mathbf{1}}\right], \tag{6}$$

with

$$\Sigma = \mathbb{E}_{\mu_{\mathrm{Sh}}}[\mathbf{z}\mathbf{z}^\top], \qquad \mathbf{b} = \mathbb{E}_{\mu_{\mathrm{Sh}}}\big[\mathbf{z}\big(\mathcal{V}(\mathbf{z}) - \mathcal{V}(\mathbf{0})\big)\big], \qquad c = \mathcal{V}(\mathbf{1}) - \mathcal{V}(\mathbf{0}). \tag{7}$$

Since Σ depends only on the Shapley kernel, the data-dependent part of $\bar{\boldsymbol{\phi}}$ enters entirely through $\mathbf{b}$ and c, both expectations under $p_{\mathcal{UC}}$ (expanded in Appendix C). Evaluating $\eta(\mathbf{z}, \mathbf{x}, y)$ requires both Y and all features $\mathbf{X}$, which only $\mathcal{LC}$ provides, yet its small sample size yields high variance. The $\mathcal{LM}_r$ sources provide additional labeled observations but lack features in Γ_r^c; $\mathcal{UC}$ provides complete features but no outcome. The central question is whether these partially-observed sources can reduce the variance of the $\mathcal{LC}$-only estimator without introducing bias.

3 METHOD

The goal is to estimate the population Shapley vector $\bar{\boldsymbol{\phi}}(p_{\mathcal{UC}}, f)$, defined with respect to the $\mathcal{UC}$ covariate distribution and the fixed, pre-trained model f, using data from the three source types described in Section 2.

3.1 PRELIMINARY ESTIMATOR AND INFLUENCE FUNCTION

Given m coalitions $\mathbf{z}_1, \ldots, \mathbf{z}_m$ drawn from the Shapley kernel μ_{Sh}, the value function $\mathcal{V}(\mathbf{z})$ is estimated from $\mathcal{LC}$ by the importance-weighted sample average

$$\widehat{\mathcal{V}}(\mathbf{z}) \;=\; \frac{1}{n}\sum_{i=1}^{n} \widehat{w}(\mathbf{x}_i)\;\eta(\mathbf{z},\, \mathbf{x}_i,\, y_i), \tag{8}$$

where the summation runs over $\mathcal{LC}$ observations. The density ratio $\widehat{w}(\mathbf{x}) = \widehat{p}_{\mathcal{UC}}(\mathbf{x}) / \widehat{p}_{\mathcal{LC}}(\mathbf{x})$ reweights $\mathcal{LC}$ to the target distribution $p_{\mathcal{UC}}$ and is estimated separately by training a binary classifier on $\mathcal{LC} \cup \mathcal{UC}$ with source indicators; $\widehat{w} \equiv 1$ when no covariate shift is present. The preliminary Shapley estimator $\widetilde{\boldsymbol{\phi}}$ is then the closed-form solution equation 6 with $(\Sigma, \mathbf{b}, c)$ replaced by their sample analogues

$$\widehat{\Sigma} = \tfrac{1}{m} \textstyle\sum_{j=1}^{m} \mathbf{z}_j \mathbf{z}_j^\top, \quad \widehat{\mathbf{b}} = \tfrac{1}{m} \textstyle\sum_{j=1}^{m} \mathbf{z}_j \big(\widehat{\mathcal{V}}(\mathbf{z}_j) - \widehat{\mathcal{V}}(\mathbf{0})\big), \quad \widehat{c} = \widehat{\mathcal{V}}(\mathbf{1}) - \widehat{\mathcal{V}}(\mathbf{0}). \tag{9}$$

The preliminary estimator $\widetilde{\boldsymbol{\phi}}$ depends on $p_{\mathcal{UC}}$ only through $\mathbf{b}$ and c in equation 7; the matrix Σ is determined by the Shapley kernel alone. Replacing the population expectation by the $\mathcal{LC}$ sample average perturbs $\mathbf{b}$ and c, and the resulting perturbation of $\bar{\boldsymbol{\phi}}$ can be expressed in terms of per-observation contributions via the chain rule.

For each observation $(\mathbf{x}, y)$ and coalition $\mathbf{z}$, let $\epsilon(\mathbf{z}, \mathbf{x}, y) = \eta(\mathbf{z}, \mathbf{x}, y) - \mathcal{V}(\mathbf{z})$ denote the residual of the per-observation value contribution about its population mean, and define $\mathbf{g}(\mathbf{x}, y) = \mathbb{E}_{\mu_{\mathrm{Sh}}}\big[\mathbf{z}\, \epsilon(\mathbf{z}, \mathbf{x}, y)\big]$. The influence function of the WLS Shapley estimator equation 6 is (the full derivation is given in Appendix C)

$$\bar{\boldsymbol{\psi}}(\mathbf{x}, y) \;=\; \Sigma^{-1} \big[\mathbf{g}(\mathbf{x}, y) + \lambda(\mathbf{x}, y)\, \mathbf{1}\big], \tag{10}$$

where

$$\lambda(\mathbf{x}, y) \;=\; \frac{\big(\eta(\mathbf{1}, \mathbf{x}, y) - \eta(\mathbf{0}, \mathbf{x}, y)\big) - c - \mathbf{1}^\top \Sigma^{-1} \mathbf{g}(\mathbf{x}, y)}{\mathbf{1}^\top \Sigma^{-1} \mathbf{1}} \tag{11}$$

is the Lagrange correction enforcing the efficiency constraint at the observation level. Since the closed-form solution equation 6 is linear in $(\mathbf{b}, c)$ with Σ fixed, the estimation error decomposes as

$$\widetilde{\boldsymbol{\phi}} - \bar{\boldsymbol{\phi}} \;=\; \frac{1}{n} \sum_{i \in \mathcal{LC}} \bar{\boldsymbol{\psi}}(\mathbf{x}_i, y_i) \;+\; O_p(m^{-1/2}), \tag{12}$$

where $\bar{\boldsymbol{\psi}}$ is given by equation 10–equation 11 and the remainder arises from approximating Σ and $\mathbf{b}$ with m sampled coalitions. The leading term is exact in the data-sampling component, as no higher-order remainder in n is incurred.

3.2 Variance reduction via partially-observed sites

The decomposition equation 12 reveals the structure that enables variance reduction. The estimation error is a sample average of per-observation contributions $\bar{\boldsymbol{\psi}}(\mathbf{x}_i, y_i)$ whose population mean is zero: $\mathbb{E}_{p_{\mathcal{UC}}}[\bar{\boldsymbol{\psi}}(\mathbf{X}, Y)] = \mathbf{0}$. Each $\bar{\boldsymbol{\psi}}(\mathbf{x}, y)$ depends on the full feature vector $\mathbf{x}$ and the outcome y, both partially available at $\mathcal{LM}_r$, which observes $(\mathbf{X}_{\Gamma_r}, Y)$. The component of $\bar{\boldsymbol{\psi}}$ predictable from these observed variables is the conditional expectation

$$\boldsymbol{\tau}_r^*(\mathbf{X}_{\Gamma_r}, Y) \;=\; \mathbb{E}\big[\bar{\boldsymbol{\psi}}(\mathbf{X}, Y) \;\big|\; \mathbf{X}_{\Gamma_r},\, Y\big]. \tag{13}$$

In practice $\boldsymbol{\tau}_r^*$ is unknown. We estimate it from the $\mathcal{LC}$ sample by regressing the estimated influence function $\widetilde{\boldsymbol{\psi}}(\mathbf{x}_i, y_i)$ on $(\mathbf{X}_{\Gamma_r}, Y)$ via cross-fitted ridge regression, yielding $\widehat{\boldsymbol{\tau}}_r(\mathbf{X}_{\Gamma_r}, Y)$.

The augmented estimator. The preliminary estimate is corrected by adding, for each site, the difference between the $\mathcal{LM}_r$ and $\mathcal{LC}$ averages of the estimated control variate:

$$\widehat{\boldsymbol{\phi}}_{\mathrm{aug}} \;=\; \widetilde{\boldsymbol{\phi}} \;+\; \sum_{r=1}^{R} \Bigg\{ \frac{1}{n_r} \sum_{(\mathbf{x}, y) \in \mathcal{LM}_r} \widehat{\boldsymbol{\tau}}_r(\mathbf{x}_{\Gamma_r}, y) \;-\; \frac{1}{n} \sum_{(\mathbf{x}, y) \in \mathcal{LC}} \widehat{\boldsymbol{\tau}}_r(\mathbf{x}_{\Gamma_r}, y) \Bigg\}. \tag{14}$$

When $p_{\mathcal{LC}} = p_{\mathcal{LM}_r} = p_{\mathcal{UC}}$, each correction term has population mean zero and the augmentation reduces variance without introducing bias. Under covariate shift, both averages require importance weighting to the target distribution $p_{\mathcal{UC}}$; the generalization is given below in equation 16.

3.3 Screening and calibration

Screening for misaligned sites. The augmented estimator equation 14 benefits from site r only if the correction term $\widehat{\boldsymbol{\tau}}_r$ estimated on $\mathcal{LM}_r$ is consistent with the same quantity estimated on $\mathcal{LC}$.

When the two disagree systematically, whether due to distributional incompatibility between site r and the target population or because the regression $\widehat{\boldsymbol{\tau}}_r$ extrapolates poorly on $\mathcal{LM}_r$ data, including site r degrades rather than improves the estimate. For each site r, we compare the importance-weighted averages of $\widehat{\boldsymbol{\tau}}_r$ computed on $\mathcal{LM}_r$ and $\mathcal{LC}$. Define

$$\bar{\boldsymbol{\tau}}_r^{\mathcal{LM}} = \frac{1}{n_r} \sum_{(\mathbf{x},y)\in\mathcal{LM}_r} \widehat{w}_r(\mathbf{x}_{\Gamma_r})\, \widehat{\boldsymbol{\tau}}_r(\mathbf{x}_{\Gamma_r}, y), \qquad \bar{\boldsymbol{\tau}}_r^{\mathcal{LC}} = \frac{1}{n} \sum_{(\mathbf{x},y)\in\mathcal{LC}} \widehat{w}(\mathbf{x})\, \widehat{\boldsymbol{\tau}}_r(\mathbf{x}_{\Gamma_r}, y), \tag{15}$$

where $\widehat{w}_r(\mathbf{x}_{\Gamma_r}) = \widehat{p}_{\mathcal{UC}}(\mathbf{x}_{\Gamma_r}) \,/\, \widehat{p}_{\mathcal{LM}_r}(\mathbf{x}_{\Gamma_r})$ and $\widehat{w}(\mathbf{x}) = \widehat{p}_{\mathcal{UC}}(\mathbf{x}) \,/\, \widehat{p}_{\mathcal{LC}}(\mathbf{x})$ are the density ratios. Both reweight to $p_{\mathcal{UC}}$, so under Assumption 1 their difference $\mathbf{d}_r = \bar{\boldsymbol{\tau}}_r^{\mathcal{LM}} - \bar{\boldsymbol{\tau}}_r^{\mathcal{LC}} \in \mathbb{R}^p$ has population mean zero. To ensure that coordinates with noisier importance weights do not dominate the comparison, we studentize the difference. The test statistic is $T_r = \sum_{j=1}^{p} d_{r,j}^2 \,/\, \widehat{\mathrm{Var}}(d_{r,j})$, where $\widehat{\mathrm{Var}}(d_{r,j}) = \widehat{\mathrm{Var}}_{\mathcal{LC}}(\widehat{w}\,\widehat{\tau}_{r,j}) \,/\, n + \widehat{\mathrm{Var}}_{\mathcal{LM}_r}(\widehat{w}_r\,\widehat{\tau}_{r,j}) \,/\, n_r$ is the estimated variance of the j-th coordinate of the weighted mean difference.

Since the null distribution of T_r depends on the estimated importance weights and control variates in a complex way, we assess significance via a permutation test rather than a χ^2_p approximation. Let $\boldsymbol{\tau}_i^{\mathcal{LC}} = \widehat{w}(\mathbf{x}_i)\,\widehat{\boldsymbol{\tau}}_r(\mathbf{x}_{i,\Gamma_r}, y_i)$ denote the weighted control variate value for $\mathcal{LC}$ observation i, and define $\boldsymbol{\tau}_k^{\mathcal{LM}}$ analogously for $\mathcal{LM}_r$ observation k. The permutation procedure is as follows. (1) Pool $\{\boldsymbol{\tau}_i^{\mathcal{LC}}\}_{i=1}^{n}$ and $\{\boldsymbol{\tau}_k^{\mathcal{LM}}\}_{k=1}^{n_r}$ into a combined set of $n + n_r$ observations. (2) For $b = 1, \ldots, B_{\mathrm{perm}}$: randomly assign n observations to the $\mathcal{LC}$ group and n_r to the $\mathcal{LM}$ group; compute the corresponding studentized statistic. (3) The p-value is $\hat{p}_r = (1 + \sum_{b=1}^{B_{\mathrm{perm}}} \mathbf{1}\{T_r^{(b)} \geq T_r\}) \,/\, (1 + B_{\mathrm{perm}})$.

We use $B_{\mathrm{perm}} = 1{,}000$ throughout. This approach avoids parametric distributional assumptions on the test statistic. While exact exchangeability under estimated nuisance parameters is not formally guaranteed, we empirically assess the calibration and power of the resulting screening procedure under both aligned and misaligned sources in Section 4.2. Sites with $\hat{p}_r < \alpha$ are excluded from the summation in equation 16.

Calibration. The augmented estimator equation 14 applies unit weight to each site's correction. In practice, the optimal weight should depend on how well $\widehat{\boldsymbol{\tau}}_r$ predicts $\bar{\boldsymbol{\psi}}$ at site r. We therefore introduce a scalar calibration weight δ_r for each site:

$$\widehat{\boldsymbol{\phi}}_{\mathrm{aug}} \;=\; \widetilde{\boldsymbol{\phi}} \;+\; \sum_{r=1}^{R} \delta_r \Bigg\{ \frac{1}{n_r} \sum_{(\mathbf{x},y)\in\mathcal{LM}_r} \widehat{w}_r(\mathbf{x}_{\Gamma_r})\widehat{\boldsymbol{\tau}}_r(\mathbf{x}_{\Gamma_r}, y) \;-\; \frac{1}{n} \sum_{(\mathbf{x},y)\in\mathcal{LC}} \widehat{w}(\mathbf{x})\widehat{\boldsymbol{\tau}}_r(\mathbf{x}_{\Gamma_r}, y) \Bigg\}. \tag{16}$$

Under no covariate shift, δ_r is chosen to minimize the empirical variance of $\widehat{\boldsymbol{\phi}}_{\mathrm{aug}}$. For each feature j, this reduces to a quadratic program in R variables with closed-form solution $\delta_{r,j}^{*} = [A_j^{-1}\mathbf{b}_j]_r$ (derived in Appendix D), where

$$[A_j]_{rs} = \widehat{\mathrm{Cov}}_{\mathcal{LC}}(\widehat{\tau}_{r,j},\, \widehat{\tau}_{s,j}) + \mathbf{1}_{r=s}\, \frac{n}{n_r}\, \widehat{\mathrm{Var}}_{\mathcal{LM}_r}(\widehat{\tau}_{r,j}), \qquad [\mathbf{b}_j]_r = \widehat{\mathrm{Cov}}_{\mathcal{LC}}(\widetilde{\psi}_j,\, \widehat{\tau}_{r,j}). \tag{17}$$

The per-site weight minimizes the total variance $\sum_{j=1}^{p} \mathrm{Var}(\widehat{\phi}_{\mathrm{aug},j})$ directly:

$$\boldsymbol{\delta}^{*} \;=\; \Bigg(\sum_{j=1}^{p} A_j\Bigg)^{-1} \Bigg(\sum_{j=1}^{p} \mathbf{b}_j\Bigg), \tag{18}$$

a single R-dimensional linear system obtained by summing the per-feature quadratic objectives. The control variate $\widehat{\boldsymbol{\tau}}_r$ is cross-fitted on $\mathcal{LC}$ (Algorithm 1).

4 Simulations

We compare DASH against six baselines representing the principal strategies for blockwise-missing multi-site data. All methods target the same estimand ($\bar{\phi}$ on the $\mathcal{UC}$ population) and share the same $\mathcal{UC}$ background sample for equation 2, differing only in which labeled observations estimate $\mathcal{V}(\mathbf{z})$. The same baselines are used in Section 5. We summarize each below.

(A) Single-site. The WLS Shapley estimator applied to $\mathcal{LC}$ alone, with $\mathcal{UC}$ as the background sample for marginalizing absent features. **(B) Single-site with importance weighting.** As (A) but reweighted by $\hat{w}(x) = \hat{p}_{\mathcal{UC}}(x)/\hat{p}_{\mathcal{LC}}(x)$. **(C) Impute-then-estimate.** Missing features at each $\mathcal{LM}_r$ are imputed and pooled with $\mathcal{LC}$. We report mean imputation and MICE.[1] **(D) Per-site averaging.** Each site independently imputes, estimates $\widehat{\mathcal{V}}_r(\mathbf{z})$ from its own labeled observations, solves the WLS, and the resulting Shapley vectors are averaged weighted by sample size. **(E) Complete-case.** Only features observed at every site ($\bigcap_r \Gamma_r$) are retained; η is averaged over all labeled observations using this reduced feature set. **(F) Oracle.** All $\mathcal{LC}$ and $\mathcal{LM}_r$ observations are pooled with the missing features at each $\mathcal{LM}_r$ site treated as observed, yielding $n + \sum_r n_r$ labeled observations with complete feature vectors on which the WLS Shapley estimator is applied.

All baselines use the WLS characterization equation 5 with m coalitions from μ_{Sh}. The reference $\bar{\boldsymbol{\phi}}^{\mathrm{true}}$ is computed by exact enumeration over all 2^p coalitions, with each $\mathcal{V}(S)$ evaluated over an independent set ($n_{\mathrm{eval}} = 50{,}000$ from $p_{\mathcal{UC}}$, $K = 500$ background samples), computed once per configuration and held fixed across all B replications.

4.1 Simulation Setup

Data-generating process. The target population $\mathcal{UC}$ has features drawn as $\mathbf{X} \sim N(\mathbf{0}, I_p)$ with $p = 10$. The outcome is generated as $Y = f(\mathbf{X}) + \varepsilon$ with $\varepsilon \sim N(0, 0.25)$, independently of $\mathbf{X}$. We consider three outcome models of increasing complexity.

Model I (linear).

$$f_{\mathrm{I}}(\mathbf{x}) \;=\; \sum_{j=1}^{p} \beta_j \, x_j, \qquad \beta_j = \frac{p+1-j}{p}, \tag{19}$$

a linear predictor with monotonically decreasing coefficients. Under squared-error loss, the per-observation value contribution $\eta(\mathbf{z}, \mathbf{x}, y)$ is quadratic in $(\mathbf{x}, y)$, and the influence function $\bar{\boldsymbol{\psi}}$ inherits this polynomial structure.

Model II (sparse interactions).

$$f_{\mathrm{II}}(\mathbf{x}) \;=\; f_{\mathrm{I}}(\mathbf{x}) + 0.5\, x_1 x_2 + 0.3\, x_3 x_4 + 0.4\, x_5^2, \tag{20}$$

augmenting Model I with pairwise interactions and a quadratic term. The influence function retains polynomial dependence on the data, though of higher degree than in Model I.

Model III (non-polynomial).

$$f_{\mathrm{III}}(\mathbf{x}) \;=\; \sin(x_1 x_2) + 0.8 \max(x_3 + x_4,\, 0) + 0.5\, x_5^2 - 0.3\, x_6, \tag{21}$$

a non-polynomial model for which polynomial control variates provide only a partial approximation.

Multi-site data construction. From each model, we draw independent observations allocated to $\mathcal{LC}$ ($n = 300$), $R = 3$ labeled-missing sources ($n_r = 2{,}000$ each), and $\mathcal{UC}$ ($N = 5{,}000$). Site r observes all features except $\{2(r-1)+1,\, 2r\}$, so each site lacks 20% of the feature set. To reflect multi-site heterogeneity, features at source s are drawn as $\mathbf{X}^{(s)} \sim N(\delta_s \mathbf{1},\, \sigma_s^2 I_p)$ with $(\delta, \sigma) = (0, 1)$ for $\mathcal{UC}$, $(0.1, 1.05)$ for $\mathcal{LC}$, $(-0.15, 0.9)$ for $\mathcal{LM}_1$, $(0.2, 1.1)$ for $\mathcal{LM}_2$, and $(-0.1, 0.95)$ for $\mathcal{LM}_3$. The density ratio $\widehat{w}(\mathbf{x})$ is estimated via a gradient-boosted classifier with source indicators. In Experiments 1, 2, and 4, the shift applies to all p features including those declared missing, deliberately introducing moderate violations of Assumption 1. In Experiment 3, aligned sites satisfy Assumption 1 exactly (the shift applies only to observed features $\mathbf{X}_{\Gamma_r}$).

4.2 Experiments

All methods receive the same fixed, pre-trained predictive model and compute its Shapley attribution vector. In the simulation studies, this is the true data-generating function f; in the real-data applications (Section 5), this is a model trained on held-out data. Across B independent replications, we report the mean squared error $\mathrm{MSE} = \mathrm{Bias}^2 + \mathrm{Var}$, where $\mathrm{Bias}^2 = \|\bar{\hat{\boldsymbol{\phi}}} - \bar{\boldsymbol{\phi}}\|^2$ and

[1] Prior work and our simulations indicate that switching among standard imputers does not resolve attribution bias under blockwise missingness (Vo et al., 2025).

$\mathrm{Var} = \sum_j \widehat{\mathrm{Var}}_B(\hat{\phi}_j)$, together with Spearman's rank correlation ρ_s between each estimate and the ground truth to assess agreement in the feature importance ranking. Where appropriate, we report the variance ratio $\mathrm{VR} = \mathrm{Var}(\text{method}) \,/\, \mathrm{Var}(\text{LC-only})$.

Table 1: Experiment 1: MSE, (Bias^2 / Var), and Spearman ρ_s. $B = 100$, $p = 10$, $n = 300$, $n_r = 2{,}000$, $R = 3$. Best feasible method in bold.

Method	Model I (linear)			Model II (interact.)			Model III (non-poly.)		
(F) Oracle	0.032	.027/.004	1.00	0.065	.059/.006	0.99	0.002	.001/.001	0.96
DASH	**0.052**	.003/.049	0.98	**0.065**	.005/.060	0.98	**0.020**	.009/.010	0.94
(A) Single-site	0.175	.087/.088	0.98	0.550	.367/.183	0.97	0.061	.036/.025	0.95
(B) Single+IPW	0.092	.019/.074	0.98	0.109	.015/.093	0.97	0.025	.015/.011	0.95
(C) Impute-mean	0.118	.111/.008	0.99	0.129	.118/.011	0.98	0.062	.060/.002	0.86
(C) Impute-MICE	0.120	.112/.008	0.99	0.130	.119/.011	0.98	0.062	.060/.002	0.86
(D) Per-site avg	0.119	.111/.009	0.99	0.130	.118/.013	0.98	0.062	.060/.002	0.86
(E) Complete-case	2.492	2.49/.001	−.75	3.336	3.33/.002	−.75	0.374	.374/.000	—

Each cell: MSE (Bias^2/Var ρ_s).

Experiment 1 (imputation bias). Table 1 reports results over $B = 100$ replications. DASH achieves the lowest MSE among all feasible methods across all three models, improving over the single-site estimator by $3.4\times$ (Model I), $8.5\times$ (Model II), and $3.1\times$ (Model III). The gains are largest where the polynomial control variate is well-specified (Models I and II) and smallest under the non-polynomial Model III, yet DASH still achieves $3.1\times$ lower MSE than imputation. The bias-variance decomposition reveals that imputation baselines reduce variance but introduce substantial bias, while DASH achieves both low bias and moderate variance.

Experiment 2 (variance reduction). We examine how DASH's MSE depends on the $\mathcal{LM}$ sample size n_r, the number of sites R, and the $\mathcal{LC}$ sample size n (Figure 1). Three patterns are consistent across models. First, improvement increases with n_r but exhibits diminishing returns beyond $n_r \approx 1{,}000$. Second, adding sites monotonically reduces MSE. Third, $\mathcal{LC}$ sample size has a critical lower bound: at $n = 100$ the control variate regression overfits and DASH degrades, but for $n \geq 200$ DASH consistently improves upon LC-only. Variance reduction is largest under Model I, where the polynomial control variate is well-specified, and smallest under Model III.

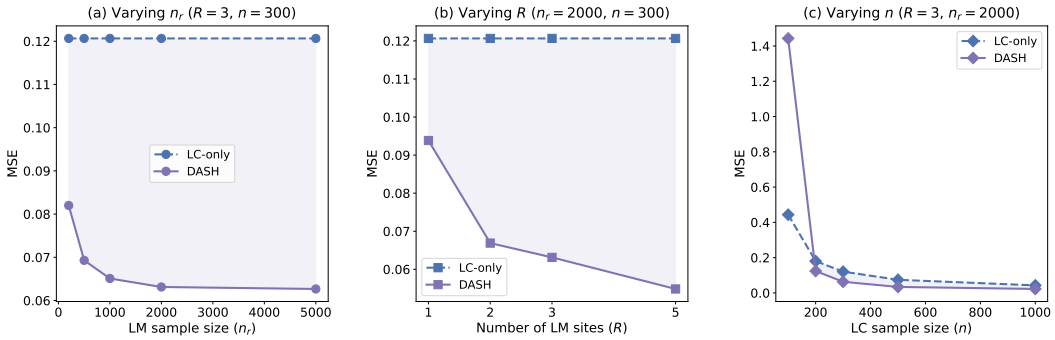
Experiment 3 (screening). Under Model I with $R = 4$ sites, the first three satisfy Assumption 1 exactly, with covariate shift applied only to observed features. The fourth site is misaligned, with its outcome generated under perturbed coefficients $\tilde{\beta}_j = \beta_j + \Delta$ for $j \in \Gamma_4^c$. Figure 2 reports the rejection rate and MSE as Δ varies from 0 to 3 over $B = 100$ replications. At $\Delta = 0$, the average rejection rate of aligned sites is 4.3%, close to the nominal $\alpha = 0.05$. The misaligned site is detected with 82% power at $\Delta = 0.25$ and 100% for $\Delta \geq 0.5$. Without screening, DASH's MSE degrades from 0.061 to 0.249, exceeding the single-site baseline. With screening, MSE stabilizes between 0.061 and 0.066, a roughly $2.5\times$ improvement (detailed per-site results in Table 4).

Experiment 4 (computational cost). DASH's overhead relative to the single-site estimator is modest (3–5×), arising from the control variate regression and calibration. Compared to impute-then-pool, DASH is 2.7–4.6× faster because it computes Shapley values on the small $\mathcal{LC}$ sample rather than the pooled dataset. Wall-clock timings are reported in Table 5 of Appendix E.

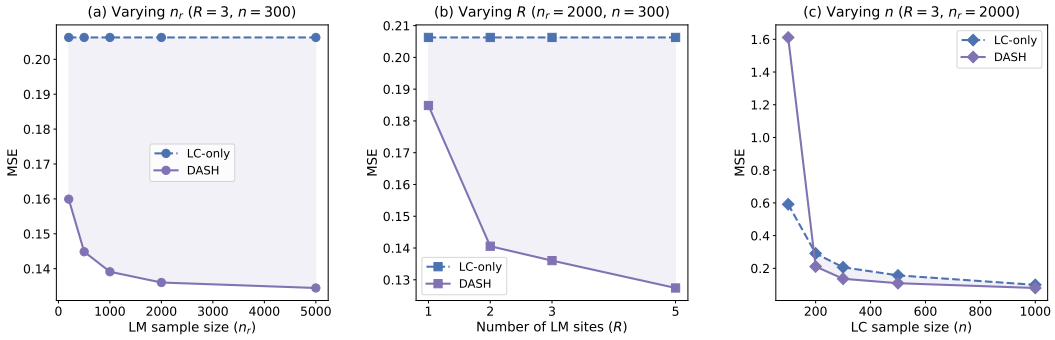
5 Real-data Applications

We evaluate DASH on two real-world datasets with controlled blockwise missingness imposed on the auxiliary sites, enabling quantitative comparison against a high-precision reference attribution vector. Since mean imputation and MICE produce nearly identical attributions under blockwise missingness (Vo et al., 2025), only mean imputation is reported.

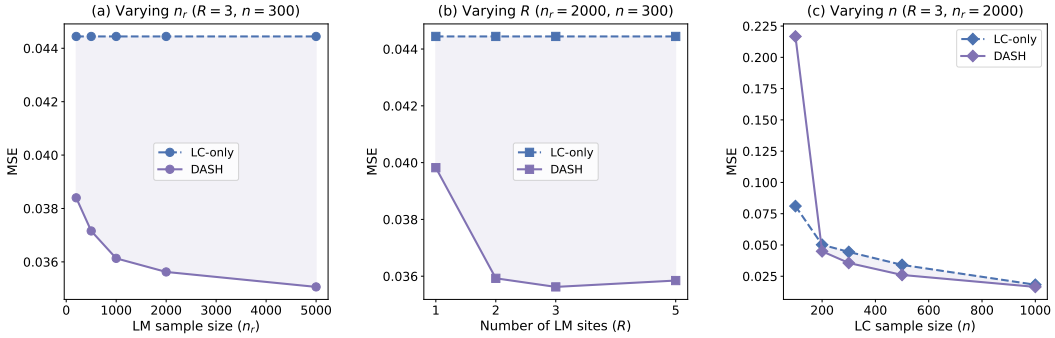
Beijing Multi-Site Air Quality. The Beijing Multi-Site Air Quality dataset (Zhang et al., 2017) records daily pollutant and meteorological measurements at 12 stations. We attribute a model for PM2.5 concentration using $p = 10$ features. One station (Dongsi, $n = 300$) serves as $\mathcal{LC}$, three



(a) Model I (linear)



(b) Model II (sparse interactions)



(c) Model III (non-polynomial)

Figure 1: Experiment 2: MSE of DASH (solid) and LC-only (dashed) as a function of n_r (left), R (center), and n (right), $B = 100$.

Table 2: Beijing Air Quality: MSE and (Bias^2 / Var), $\times 10^{-3}$, $B = 50$. Best feasible method in bold.

Method	Linear	RF	GBM	MLP
(A) Single-site	42.7 (13.4/29.2)	42.1 (13.3/28.7)	39.4 (12.4/27.0)	39.0 (12.2/26.8)
(B) Single+IPW	20.9 (11.9/9.0)	22.4 (14.9/7.6)	22.0 (14.4/7.6)	25.0 (14.1/10.9)
(C) Impute-mean	12.1 (10.7/1.4)	11.8 (10.6/1.2)	12.7 (11.6/1.2)	23.3 (20.4/2.8)
(D) Per-site avg	11.8 (10.6/1.2)	11.6 (10.6/1.0)	12.5 (11.5/1.0)	22.8 (20.7/2.1)
(E) Complete-case	62.7 (61.6/1.1)	79.7 (79.5/0.2)	77.1 (76.8/0.2)	136.1 (134.7/1.5)
DASH	**5.8** (4.2/1.7)	**7.3** (5.5/1.8)	**7.4** (5.5/1.8)	**8.8** (5.5/3.3)

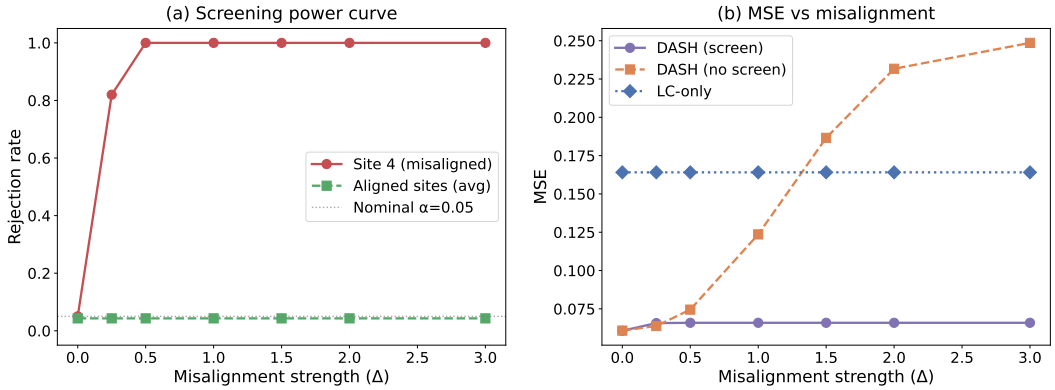


Figure 2: Experiment 3: screening power and MSE ($B = 100$, $\alpha = 0.05$). Left: rejection rate of the misaligned site and aligned sites (average). Right: MSE of DASH with and without screening.

serve as $\mathcal{LM}$ sites each missing a different feature pair (Appendix E.3), and eight form $\mathcal{UC}$. Four models are trained on 15,646 non-$\mathcal{LC}$ observations and held fixed.

Table 2 reports the results. DASH achieves the lowest MSE across all four models, improving over the single-site estimator by 4.4–7.4× and over the best imputation baseline by 1.6–2.6×. The bias-variance decomposition reveals that imputation baselines reduce variance but introduce substantial bias, while DASH achieves both low bias and low variance.

Table 3: NACC Alzheimer's: MSE and (Bias^2 / Var), $B = 50$. Best feasible method in bold.

Method	Ridge	RF	GBM	MLP
(A) Single-site	6.83 (0.80/6.03)	8.54 (1.83/6.71)	7.35 (0.97/6.38)	7.07 (0.61/6.47)
(C) Impute-mean	6.23 (6.12/0.11)	16.25 (16.12/0.13)	10.67 (10.54/0.13)	21.53 (21.09/0.45)
(D) Per-site avg	6.22 (6.11/0.11)	16.26 (16.13/0.13)	10.67 (10.54/0.13)	21.59 (21.15/0.44)
DASH	**1.03** (0.18/0.85)	**3.25** (1.55/1.70)	**2.79** (0.93/1.86)	**1.79** (0.36/1.43)

NACC Alzheimer's Disease. The NACC dataset aggregates multi-center clinical assessments of Alzheimer's patients. We predict MMSE scores from $p = 9$ demographic and clinical features. A random sample of $n = 300$ patients serves as $\mathcal{LC}$, three $\mathcal{LM}$ sites have complementary synthetic missingness, and the remainder forms $\mathcal{UC}$ (Appendix E.4). Complete-case analysis is infeasible because no features are common to all sites. Since $\mathcal{LC}$ is a random population sample, Baseline (B) is omitted. Four models are trained on 16,900 non-$\mathcal{LC}$ patients and held fixed.

Table 3 reports the results. DASH achieves the lowest MSE across models, improving over the single-site estimator by 2.6–6.6× and over the best imputation baseline by 3.8–12.0×. Notably, for three of the four models imputation performs worse than the single-site estimator, with MSE up to 3.0× higher (MLP), confirming that imputation bias can outweigh its variance reduction. The bias-variance pattern mirrors Beijing, with DASH again achieving both low bias and reduced variance.

6 Discussion

We proposed DASH, a method for estimating global Shapley feature attributions from multi-site data with blockwise missingness. By deriving the influence function of the constrained WLS Shapley estimator and constructing site-specific control variates, DASH reduces variance without imputation, avoiding the systematic coalition-dependent bias that imputation introduces. A permutation-based screening procedure protects against incompatible sites, and data-adaptive calibration weights each site proportional to its informativeness. Two extensions merit future work: relaxing the assumption that $\mathcal{UC}$ has complete feature coverage, which would require additional identification assumptions, and adapting the framework to privacy-preserving distributed settings, which is natural since the augmentation and calibration depend only on site-level summary statistics.

REPRODUCIBILITY STATEMENT

The data-generating processes, multi-site construction, and evaluation protocol are specified in Sections 4–5 and Appendices E.3–E.4. The complete algorithm is given in Algorithm 1. The influence function and calibration weight derivations appear in Appendices C and D. The Beijing Multi-Site Air Quality dataset is publicly available from the UCI Machine Learning Repository (Zhang et al., 2017). NACC data are available upon request through the NACC data request process (`https://naccdata.org/data-request-process/`). Anonymous source code is available at `https://anonymous.4open.science/r/DASH-87BE`.

ETHICS STATEMENT

This work uses the publicly available Beijing Air Quality dataset and de-identified clinical data from NACC. The NACC database operates under IRB approval at each contributing Alzheimer's Disease Research Center. No additional human subjects research was conducted for this study.

ACKNOWLEDGMENTS

Acknowledgments are withheld for anonymous review.

AI USE STATEMENT

We used generative AI tools (ChatGPT and Claude) for editing and proofreading manuscript text. All AI-assisted output was reviewed and revised by the authors. We take responsibility for the final content of this work.

REFERENCES

Hugh Chen, Ian C Covert, Scott M Lundberg, and Su-In Lee. Algorithms to estimate shapley value feature attributions. *Nature Machine Intelligence*, 5(6):590–601, 2023.

Ian Covert and Su-In Lee. Improving KernelSHAP: practical Shapley value estimation via linear regression. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

Ian Covert, Scott Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021.

Jingyue Huang, Huiyuan Wang, Yuqing Lei, and Yong Chen. Efficient semiparametric inference for distributed data with blockwise missingness, 2025. URL `https://arxiv.org/abs/2508.16902`.

Neil Jethani, Mukund Sudarshan, Ian Covert, Su-In Lee, and Rajesh Ranganath. FastSHAP: real-time Shapley value estimation. In *International Conference on Learning Representations (ICLR)*, 2022.

Ying Jin and Dominik Rothenhäusler. Modular regression: Improving linear models by incorporating auxiliary data, 2023. URL `https://arxiv.org/abs/2211.10032`.

Meng Li, Hengyang Sun, Yanjun Huang, and Hong Chen. Shapley value: from cooperative game to explainable artificial intelligence. *Autonomous Intelligent Systems*, 4(1):2, 2024.

Yiming Li, Ying Wei, and Molei Liu. Adaptive learning with blockwise missing and semi-supervised data. *arXiv preprint arXiv:2405.18722v3*, 2025.

Zelei Liu, Yuanyuan Chen, Han Yu, Yang Liu, and Lizhen Cui. GTG-Shapley: efficient and accurate participant contribution evaluation in federated learning. *ACM Transactions on Intelligent Systems and Technology*, 13(4):1–21, 2022.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Rory Mitchell, Joshua Cooper, Eibe Frank, and Geoffrey Holmes. Sampling permutations for shapley value estimation. *Journal of Machine Learning Research*, 23(43):1–46, 2022.

Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. Shap-based explanation methods: a review for nlp interpretability. In *Proceedings of the 29th international conference on computational linguistics*, pp. 4593–4603, 2022.

Jeffrey Näf, Erwan Scornet, and Julie Josse. What is a good imputation under mar missingness?, 2026. URL https://arxiv.org/abs/2403.19196.

Ahmed M Salih, Zahra Raisi-Estabragh, Ilaria Boscolo Galazzo, Petia Radeva, Steffen E Petersen, Karim Lekadir, and Gloria Menegaz. A perspective on explainable artificial intelligence methods: Shap and lime. *Advanced Intelligent Systems*, 7(1):2400304, 2025.

Luke Shannon, Song Liu, and Katarzyna Reluga. Distribution shift in missing data imputation: A risk-based perspective and importance-weighted correction under mar, 2026. URL https://arxiv.org/abs/2602.06713.

Tuan L. Vo, Thu Nguyen, Luis M. Lopez-Ramos, Hugo L. Hammer, Michael A. Riegler, and Pal Halvorsen. Explainability of machine learning models under missing data, 2025. URL https://arxiv.org/abs/2407.00411.

Tianhao Wang, Johannes Rausch, Ce Zhang, Ruoxi Jia, and Dawn Song. A principled approach to data valuation for federated learning, 2020. URL https://arxiv.org/abs/2009.06192.

Eric Wu, Kevin Wu, and James Zou. Explaining medical ai performance disparities across sites with confounder shapley value analysis, 2021. URL https://arxiv.org/abs/2111.08168.

Qi Xu, Lorenzo Testa, Jing Lei, and Kathryn Roeder. Blockwise missingness meets ai: A tractable solution for semiparametric inference, 2025. URL https://arxiv.org/abs/2509.24158.

Fei Xue and Annie Qu. Integrating multisource block-wise missing data in model selection. *Journal of the American Statistical Association*, 116(536):1914–1927, 2021.

Shuyi Zhang, Bin Guo, Anlan Dong, Jing He, Ziping Xu, and Song Xi Chen. Cautionary tales on air-quality improvement in Beijing. *Proceedings of the Royal Society A*, 473(2205):20170457, 2017.

Shuyuan Zheng, Yang Cao, and Masatoshi Yoshikawa. Secure shapley value for cross-silo federated learning. *Proceedings of the VLDB Endowment*, 16(7):1657–1670, March 2023. ISSN 2150-8097. doi: 10.14778/3587136.3587141. URL http://dx.doi.org/10.14778/3587136.3587141.

A Additional Remarks

Remark 1 (Feature removal convention)**.** equation 2 adopts the marginal removal convention, in which $\mathbf{X}_{S^c}$ is drawn independently of $\mathbf{X}_S$. The conditional alternative $\bar{f}_S(\mathbf{x}_S) = \mathbb{E}_{p_{\mathcal{UC}}}[f(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S]$ preserves feature dependencies but requires estimating high-dimensional conditional distributions. The estimation framework in Section 3 is agnostic to this choice: it requires only that $\mathcal{V}(\mathbf{z}) = \mathbb{E}_{p_{\mathcal{UC}}}[\eta(\mathbf{z}, \mathbf{X}, Y)]$ for some per-observation function η, a property satisfied under either convention.

Remark 2 (Extension to covariate shift)**.** The influence function in equation 10–equation 11 is derived under $p_{\mathcal{LC}} = p_{\mathcal{UC}}$. Suppose instead that the target covariate distribution is absolutely continuous with respect to the $\mathcal{LC}$ distribution, with density ratio $w(\mathbf{x}) = p_{\mathcal{UC}}(\mathbf{x}) \,/\, p_{\mathcal{LC}}(\mathbf{x})$. When w is known, the same derivation applies after replacing $\eta(\mathbf{z}, \mathbf{x}, y)$ by $w(\mathbf{x})\,\eta(\mathbf{z}, \mathbf{x}, y)$. In particular, the weighted residual is $\epsilon_w(\mathbf{z}, \mathbf{x}, y) = w(\mathbf{x})\,\eta(\mathbf{z}, \mathbf{x}, y) - \mathcal{V}(\mathbf{z})$, and $\mathbf{g}$, λ, and $\bar{\boldsymbol{\psi}}$ are defined analogously using ϵ_w.

In practice, w is replaced by an estimate $\widehat{w}$ obtained from the $\mathcal{LC}$ and $\mathcal{UC}$ covariates. The linearization in equation 12 continues to hold with the same first-order influence function whenever the contribution from density-ratio estimation is asymptotically negligible. A sufficient condition is $\|\widehat{w} - w\|_\infty = o_p(n^{-1/2})$, together with appropriate moment and regularity conditions on η. We treat this condition as an assumption in the present analysis; more generally, when density-ratio estimation contributes at first order, an orthogonal/debiased construction is required to account for this additional nuisance-estimation error.

B Summary of the DASH Algorithm

C Derivation of the Influence Function

The population Shapley vector equation 6 is a function of $(\mathbf{b}, c)$:

$$\bar{\boldsymbol{\phi}} = h(\mathbf{b}, c) = \Sigma^{-1}\left[\mathbf{b} + \frac{c - \mathbf{1}^\top\Sigma^{-1}\mathbf{b}}{\mathbf{1}^\top\Sigma^{-1}\mathbf{1}}\,\mathbf{1}\right], \tag{22}$$

where $\Sigma = \mathbb{E}_{\mu_{\mathrm{Sh}}}[\mathbf{z}\mathbf{z}^\top]$ depends only on p and the Shapley kernel. For notational convenience, define

$$\mathbf{s} = \Sigma^{-1}\mathbf{1}, \qquad \kappa = \mathbf{1}^\top\mathbf{s} = \mathbf{1}^\top\Sigma^{-1}\mathbf{1}. \tag{23}$$

Note that $\mathbf{b}$ and c are both expectations under $p_{\mathcal{UC}}$:

$$\mathbf{b} = \mathbb{E}_{\mathbf{z}}\big[\mathbf{z}\,\mathbb{E}_{(\mathbf{X},Y)\sim p_{\mathcal{UC}}}[\eta(\mathbf{z}, \mathbf{X}, Y)]\big] - \mathcal{V}(\mathbf{0})\,\mathbb{E}_{\mathbf{z}}[\mathbf{z}], \tag{24}$$

$$c = \mathbb{E}_{(\mathbf{X},Y)\sim p_{\mathcal{UC}}}[\eta(\mathbf{1}, \mathbf{X}, Y) - \eta(\mathbf{0}, \mathbf{X}, Y)]. \tag{25}$$

Expanding equation 22:

$$\begin{aligned} h(\mathbf{b}, c) &= \Sigma^{-1}\mathbf{b} + \frac{c}{\kappa}\,\mathbf{s} - \frac{\mathbf{s}^\top\mathbf{b}}{\kappa}\,\mathbf{s} \\ &= \underbrace{\left(\Sigma^{-1} - \frac{\mathbf{s}\mathbf{s}^\top}{\kappa}\right)}_{M_1}\mathbf{b} \;+\; \underbrace{\frac{\mathbf{s}}{\kappa}}_{M_2}\,c. \end{aligned} \tag{26}$$

Since M_1 and M_2 are constant matrices (depending only on Σ), the map h is linear in $(\mathbf{b}, c)$. The first-order expansion is therefore exact with no higher-order remainder. The partial derivatives follow from equation 26:

$$\frac{\partial h}{\partial \mathbf{b}} = M_1 = \Sigma^{-1} - \frac{\mathbf{s}\mathbf{s}^\top}{\kappa}, \tag{27}$$

$$\frac{\partial h}{\partial c} = M_2 = \frac{\mathbf{s}}{\kappa} = \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}^\top\Sigma^{-1}\mathbf{1}}. \tag{28}$$

Algorithm 1 DASH: Variance-Reduced Shapley Attribution

Require: $\mathcal{LC}$, $\{\mathcal{LM}_r\}_{r=1}^{R}$, $\mathcal{UC}$, model f, loss ℓ, significance level α

Step 1: Preliminary estimator
Estimate density ratio $\widehat{w}(\mathbf{x}) = \widehat{p}_{\mathcal{UC}}(\mathbf{x})/\widehat{p}_{\mathcal{LC}}(\mathbf{x})$ via gradient-boosted classifier on $\mathcal{LC} \cup \mathcal{UC}$
Sample m coalitions $\mathbf{z}_1, \ldots, \mathbf{z}_m \sim \mu_{\mathrm{Sh}}$
Compute importance-weighted value functions $\widehat{\mathcal{V}}(\mathbf{z}_j) = n^{-1} \sum_{i \in \mathcal{LC}} \widehat{w}(\mathbf{x}_i)\, \eta(\mathbf{z}_j, \mathbf{x}_i, y_i)$
Solve WLS via equation 6 $\rightarrow$ $\widetilde{\boldsymbol{\phi}}$

Step 2: Influence function
Compute Σ^{-1} from the Shapley kernel
Compute weighted residuals $\epsilon_w(\mathbf{z}_j, \mathbf{x}_i, y_i) = \widehat{w}(\mathbf{x}_i)\, \eta(\mathbf{z}_j, \mathbf{x}_i, y_i) - \widehat{\mathcal{V}}(\mathbf{z}_j)$
Compute $\mathbf{g}_w(\mathbf{x}_i, y_i) = m^{-1} \sum_{j=1}^{m} \mathbf{z}_j\, \epsilon_w(\mathbf{z}_j, \mathbf{x}_i, y_i)$
Compute $\lambda_w(\mathbf{x}_i, y_i)$ via equation 11
Compute $\widetilde{\boldsymbol{\psi}}_w(\mathbf{x}_i, y_i) = \Sigma^{-1}[\mathbf{g}_w(\mathbf{x}_i, y_i) + \lambda_w(\mathbf{x}_i, y_i)\, \mathbf{1}]$

Step 3: Control variates
for $r = 1, \ldots, R$ **do**
  Estimate per-site density ratio $\widehat{w}_r(\mathbf{x}_{\Gamma_r})$ via classifier on $\mathcal{LM}_r \cup \mathcal{UC}$ using features Γ_r
  Cross-fit $\widehat{\boldsymbol{\tau}}_r$: ridge regression of $\widetilde{\boldsymbol{\psi}}$ on $\mathrm{poly}(\mathbf{X}_{\Gamma_r}, Y)$ with 5-fold CV on $\mathcal{LC}$
  Refit on all $\mathcal{LC}$; evaluate $\widehat{\boldsymbol{\tau}}_r$ on $\mathcal{LM}_r$
end for

Step 4: Screening
for $r = 1, \ldots, R$ **do**
  Compute studentized T_r as defined in Section 3.3; assess significance by permutation test ($B_{\mathrm{perm}} = 1{,}000$)
  if $\hat{p}_r < \alpha$ **then**
    Exclude site r
  end if
end for

Step 5: Calibration
for $j = 1, \ldots, p$ **do**
  Compute A_j and $\mathbf{b}_j$ via equation 17
  Add adaptive ridge: $A_j \leftarrow A_j + \lambda_{\mathrm{reg}} I_R$
end for
Solve $\boldsymbol{\delta}^* = \big(\sum_{j=1}^{p} A_j\big)^{-1} \big(\sum_{j=1}^{p} \mathbf{b}_j\big)$ via equation 18

Output: $\widehat{\boldsymbol{\phi}}_{\mathrm{aug}} = \widetilde{\boldsymbol{\phi}} + \sum_{r \in \mathrm{aligned}} \delta_r^* \big\{ \bar{\boldsymbol{\tau}}_r^{\mathcal{LM}} - \bar{\boldsymbol{\tau}}_r^{\mathcal{LC}} \big\}$

Recall that $\mathbf{b}$ is defined as a weighted average over coalitions $\mathbf{b} = \mathbb{E}_{\mu_{\mathrm{Sh}}}\big[\mathbf{z}\big(\mathcal{V}(\mathbf{z}) - \mathcal{V}(\mathbf{0})\big)\big]$. When each $\mathcal{V}(\mathbf{z})$ is perturbed by $\epsilon(\mathbf{z}, \mathbf{x}, y)$ (including the null coalition $\mathcal{V}(\mathbf{0})$ which is perturbed by $\epsilon(\mathbf{0}, \mathbf{x}, y)$), the perturbed $\mathbf{b}$ becomes

$$\begin{aligned}\mathbf{b} + \delta\mathbf{b} &= \mathbb{E}_{\mu_{\mathrm{Sh}}}\Big[\mathbf{z}\Big(\big(\mathcal{V}(\mathbf{z}) + \epsilon(\mathbf{z}, \mathbf{x}, y)\big) - \big(\mathcal{V}(\mathbf{0}) + \epsilon(\mathbf{0}, \mathbf{x}, y)\big)\Big)\Big] \\ &= \underbrace{\mathbb{E}_{\mu_{\mathrm{Sh}}}\big[\mathbf{z}\big(\mathcal{V}(\mathbf{z}) - \mathcal{V}(\mathbf{0})\big)\big]}_{\mathbf{b}} + \mathbb{E}_{\mu_{\mathrm{Sh}}}\big[\mathbf{z}\big(\epsilon(\mathbf{z}, \mathbf{x}, y) - \epsilon(\mathbf{0}, \mathbf{x}, y)\big)\big].\end{aligned}$$

Hence

$$\delta\mathbf{b} = \mathbb{E}_{\mu_{\mathrm{Sh}}}\big[\mathbf{z}\big(\epsilon(\mathbf{z}, \mathbf{x}, y) - \epsilon(\mathbf{0}, \mathbf{x}, y)\big)\big]. \tag{29}$$

Since $\epsilon(\mathbf{0}, \mathbf{x}, y) = \eta(\mathbf{0}, \mathbf{x}, y) - \mathcal{V}(\mathbf{0})$ does not depend on the coalition $\mathbf{z}$, it is a scalar that factors out of the expectation:

$$\delta\mathbf{b} = \underbrace{\mathbb{E}_{\mu_{\mathrm{Sh}}}\big[\mathbf{z}\, \epsilon(\mathbf{z}, \mathbf{x}, y)\big]}_{\mathbf{g}(\mathbf{x}, y)} - \underbrace{\mathbb{E}_{\mu_{\mathrm{Sh}}}[\mathbf{z}]}_{\bar{\mathbf{z}}}\, \epsilon(\mathbf{0}, \mathbf{x}, y). \tag{30}$$

Similarly, from $c = \mathcal{V}(\mathbf{1}) - \mathcal{V}(\mathbf{0})$, the perturbation of c is

$$\begin{aligned}\delta c &= \big(\mathcal{V}(\mathbf{1}) + \epsilon(\mathbf{1},\mathbf{x},y)\big) - \big(\mathcal{V}(\mathbf{0}) + \epsilon(\mathbf{0},\mathbf{x},y)\big) - \underbrace{\big(\mathcal{V}(\mathbf{1}) - \mathcal{V}(\mathbf{0})\big)}_{c} \\ &= \epsilon(\mathbf{1},\mathbf{x},y) - \epsilon(\mathbf{0},\mathbf{x},y) \\ &= \big[\eta(\mathbf{1},\mathbf{x},y) - \mathcal{V}(\mathbf{1})\big] - \big[\eta(\mathbf{0},\mathbf{x},y) - \mathcal{V}(\mathbf{0})\big] \\ &= \big[\eta(\mathbf{1},\mathbf{x},y) - \eta(\mathbf{0},\mathbf{x},y)\big] - \underbrace{\big[\mathcal{V}(\mathbf{1}) - \mathcal{V}(\mathbf{0})\big]}_{=c}. \end{aligned} \tag{31}$$

Define the influence function $\bar{\boldsymbol{\psi}}(\mathbf{x}, y)$ as the per-observation contribution to the estimation error $\widetilde{\boldsymbol{\phi}} - \bar{\boldsymbol{\phi}}$, so that

$$\widetilde{\boldsymbol{\phi}} - \bar{\boldsymbol{\phi}} = \frac{1}{n} \sum_{i \in \mathcal{LC}} \bar{\boldsymbol{\psi}}(\mathbf{x}_i, y_i).$$

Since $h(\mathbf{b}, c) = M_1\mathbf{b} + M_2 c$ is linear, each observation's contribution is

$$\bar{\boldsymbol{\psi}}(\mathbf{x}, y) = M_1\, \delta\mathbf{b}(\mathbf{x}, y) + M_2\, \delta c(\mathbf{x}, y) = \frac{\partial h}{\partial \mathbf{b}}\, \delta\mathbf{b} + \frac{\partial h}{\partial c}\, \delta c. \tag{32}$$

Substituting equation 27–equation 31 and writing $\epsilon_0 = \epsilon(\mathbf{0}, \mathbf{x}, y)$:

$$\begin{aligned}\bar{\boldsymbol{\psi}} &= \left(\Sigma^{-1} - \frac{\mathbf{s}\mathbf{s}^\top}{\kappa}\right)\big(\mathbf{g} - \bar{\mathbf{z}}\,\epsilon_0\big) + \frac{\mathbf{s}}{\kappa}\,\delta c \\ &= \Sigma^{-1}\big(\mathbf{g} - \bar{\mathbf{z}}\,\epsilon_0\big) - \frac{\mathbf{s}\,\mathbf{s}^\top(\mathbf{g} - \bar{\mathbf{z}}\,\epsilon_0)}{\kappa} + \frac{\mathbf{s}\,\delta c}{\kappa} \\ &= \Sigma^{-1}\left[\mathbf{g} - \bar{\mathbf{z}}\,\epsilon_0 + \frac{\delta c - \mathbf{s}^\top(\mathbf{g} - \bar{\mathbf{z}}\,\epsilon_0)}{\kappa}\,\mathbf{1}\right], \end{aligned} \tag{33}$$

where the last equality uses $\mathbf{s} = \Sigma^{-1}\mathbf{1}$ to factor Σ^{-1} from the second and third terms.

The Shapley kernel $\mu_{\mathrm{Sh}}(\mathbf{z})$ depends on $\mathbf{z}$ only through $|\mathbf{z}| = \mathbf{1}^\top\mathbf{z}$. Since each feature appears symmetrically across all coalitions of a given size,

$$\bar{\mathbf{z}} = \mathbb{E}_{\mu_{\mathrm{Sh}}}[\mathbf{z}] = \tfrac{1}{2}\,\mathbf{1}. \tag{34}$$

Substituting into equation 33, the expression inside the brackets becomes

$$\mathbf{g} \;-\; \tfrac{1}{2}\mathbf{1}\,\epsilon_0 \;+\; \frac{\delta c - \mathbf{s}^\top(\mathbf{g} - \tfrac{1}{2}\mathbf{1}\,\epsilon_0)}{\kappa}\,\mathbf{1}.$$

Since $-\tfrac{1}{2}\mathbf{1}\,\epsilon_0$ is proportional to $\mathbf{1}$, it can be merged with the last term. To do so, first expand $\mathbf{s}^\top(\mathbf{g} - \tfrac{1}{2}\mathbf{1}\,\epsilon_0)$:

$$\mathbf{s}^\top\big(\mathbf{g} - \tfrac{1}{2}\mathbf{1}\,\epsilon_0\big) = \mathbf{s}^\top\mathbf{g} - \tfrac{1}{2}\underbrace{(\mathbf{s}^\top\mathbf{1})}_{=\kappa}\,\epsilon_0 = \mathbf{s}^\top\mathbf{g} - \tfrac{\kappa}{2}\,\epsilon_0. \tag{35}$$

Now collect all scalar multiples of $\mathbf{1}$ from both terms and denote their sum by λ^*:

$$\begin{aligned}\lambda^* &= -\tfrac{1}{2}\,\epsilon_0 \;+\; \frac{\delta c - \mathbf{s}^\top\mathbf{g} + \frac{\kappa}{2}\,\epsilon_0}{\kappa} \\ &= -\tfrac{1}{2}\,\epsilon_0 \;+\; \frac{\delta c - \mathbf{s}^\top\mathbf{g}}{\kappa} \;+\; \frac{\frac{\kappa}{2}\,\epsilon_0}{\kappa} \\ &= \frac{\delta c - \mathbf{s}^\top\mathbf{g}}{\kappa}. \end{aligned} \tag{36}$$

Since all ϵ_0 terms have canceled, the non-$\mathbf{1}$ part of equation 33 reduces to $\mathbf{g}$ alone. Expanding δc and $\mathbf{s}^\top \mathbf{g}$ in equation 36 using equation 31 and $\mathbf{s}^\top = \mathbf{1}^\top \Sigma^{-1}$:

$$\begin{aligned}\lambda^* &= \frac{\delta c - \mathbf{s}^\top \mathbf{g}}{\kappa} \\ &= \frac{\big[\eta(\mathbf{1}, \mathbf{x}, y) - \eta(\mathbf{0}, \mathbf{x}, y)\big] - c - \mathbf{1}^\top \Sigma^{-1} \mathbf{g}(\mathbf{x}, y)}{\mathbf{1}^\top \Sigma^{-1} \mathbf{1}}.\end{aligned} \tag{37}$$

Substituting back into equation 33:

$$\boxed{\bar{\boldsymbol{\psi}}(\mathbf{x}, y) = \Sigma^{-1}\big[\mathbf{g}(\mathbf{x}, y) + \lambda(\mathbf{x}, y)\, \mathbf{1}\big],} \tag{38}$$

with $\lambda(\mathbf{x}, y)$ given by equation 37, recovering equation 10–equation 11 in the main text. □

D Derivation of the Calibration Weights

We derive the optimal calibration weights $\boldsymbol{\delta}^*_{\cdot,j} = A_j^{-1} \mathbf{b}_j$ stated in equation 17 under no covariate shift. Under covariate shift, we retain the same calibration rule while importance-weighting the augmentation terms as in equation 16.

For feature j, the estimation error of the calibrated estimator equation 16 is approximately

$$\widehat{\phi}_{\mathrm{aug},j} - \bar{\phi}_j \;\approx\; \frac{1}{n} \sum_{i \in \mathcal{LC}} \left[\widetilde{\psi}_j(\mathbf{x}_i, y_i) - \sum_{r=1}^{R} \delta_r\, \widehat{\tau}_{r,j}(\mathbf{x}_i, y_i) \right] \;+\; \sum_{r=1}^{R} \frac{\delta_r}{n_r} \sum_{k \in \mathcal{LM}_r} \widehat{\tau}_{r,j}(\mathbf{x}_k, y_k). \tag{39}$$

Since $\mathcal{LC}$ and $\mathcal{LM}_r$ are independent, the variance decomposes as

$$\mathrm{Var}(\widehat{\phi}_{\mathrm{aug},j}) \;=\; \frac{1}{n} \mathrm{Var}\left(\widetilde{\psi}_j - \sum_r \delta_r\, \widehat{\tau}_{r,j} \right) \;+\; \sum_{r=1}^{R} \frac{\delta_r^2}{n_r} \mathrm{Var}_{\mathcal{LM}_r}(\widehat{\tau}_{r,j}). \tag{40}$$

Expanding the first term:

$$\begin{aligned}\mathrm{Var}\left(\widetilde{\psi}_j - \sum_r \delta_r\, \widehat{\tau}_{r,j} \right) &= \mathrm{Var}(\widetilde{\psi}_j) - 2 \sum_r \delta_r\, \mathrm{Cov}(\widetilde{\psi}_j,\, \widehat{\tau}_{r,j}) \\ &\quad + \sum_r \sum_s \delta_r\, \delta_s\, \mathrm{Cov}(\widehat{\tau}_{r,j},\, \widehat{\tau}_{s,j}).\end{aligned} \tag{41}$$

Substituting equation 41 into equation 40 and combining the δ_r^2 terms:

$$\begin{aligned}V_j(\boldsymbol{\delta}) &= \frac{1}{n} \mathrm{Var}(\widetilde{\psi}_j) - \frac{2}{n} \sum_r \delta_r\, \mathrm{Cov}(\widetilde{\psi}_j,\, \widehat{\tau}_{r,j}) \\ &\quad + \frac{1}{n} \sum_r \sum_s \delta_r\, \delta_s\, \mathrm{Cov}(\widehat{\tau}_{r,j},\, \widehat{\tau}_{s,j}) + \sum_r \frac{\delta_r^2}{n_r} \mathrm{Var}_{\mathcal{LM}_r}(\widehat{\tau}_{r,j}).\end{aligned} \tag{42}$$

This is quadratic in $\boldsymbol{\delta}$. Taking the derivative with respect to δ_r and setting it to zero:

$$\frac{\partial V_j}{\partial \delta_r} = -\frac{2}{n} \mathrm{Cov}(\widetilde{\psi}_j,\, \widehat{\tau}_{r,j}) + \frac{2}{n} \sum_s \delta_s\, \mathrm{Cov}(\widehat{\tau}_{r,j},\, \widehat{\tau}_{s,j}) + \frac{2\delta_r}{n_r} \mathrm{Var}_{\mathcal{LM}_r}(\widehat{\tau}_{r,j}) = 0. \tag{43}$$

Dividing by $2/n$ and rearranging:

$$\sum_{s=1}^{R} \delta_s \left[\mathrm{Cov}(\widehat{\tau}_{r,j},\, \widehat{\tau}_{s,j}) + \mathbf{1}_{r=s}\, \frac{n}{n_r} \mathrm{Var}_{\mathcal{LM}_r}(\widehat{\tau}_{r,j}) \right] = \mathrm{Cov}(\widetilde{\psi}_j,\, \widehat{\tau}_{r,j}). \tag{44}$$

This is the linear system $A_j \boldsymbol{\delta}_{\cdot,j} = \mathbf{b}_j$ with

$$\begin{aligned}[A_j]_{rs} &= \mathrm{Cov}_{\mathcal{LC}}(\widehat{\tau}_{r,j},\, \widehat{\tau}_{s,j}) + \mathbf{1}_{r=s}\, \frac{n}{n_r} \mathrm{Var}_{\mathcal{LM}_r}(\widehat{\tau}_{r,j}), \\ [\mathbf{b}_j]_r &= \mathrm{Cov}_{\mathcal{LC}}(\widetilde{\psi}_j,\, \widehat{\tau}_{r,j}),\end{aligned}$$

recovering equation 17 in the main text. When a single per-site scalar δ_r is used across all features, the total variance $\sum_{j=1}^{p} V_j(\boldsymbol{\delta})$ is minimized by summing equation 44 over j, yielding $\big(\sum_j A_j\big)\boldsymbol{\delta} = \sum_j \mathbf{b}_j$. □

E Additional Experimental Results

E.1 Simulation Experiment 3: Detailed Screening Results

Table 4 reports per-site rejection rates and MSE across all misalignment strengths tested in Experiment 3. Panel (a) confirms that the three aligned sites maintain rejection rates between 2% and 7% across all values of Δ, consistent with the nominal $\alpha = 0.05$. Panel (b) shows that DASH with screening stabilizes MSE between 0.061 and 0.066 across all misalignment strengths, while DASH without screening degrades beyond the single-site baseline for $\Delta \geq 1$.

Table 4: Experiment 3: per-site rejection rates and MSE across misalignment strengths ($\alpha = 0.05$, $B = 100$). Aligned sites satisfy Assumption 1 exactly.

(a) Rejection rates (%)

Δ	Site 1 (aligned)	Site 2 (aligned)	Site 3 (aligned)	Site 4 (misaligned)
0	4	7	2	5
0.25	4	7	2	82
0.5	4	7	2	100
1.0	4	7	2	100
1.5	4	7	2	100
2.0	4	7	2	100
3.0	4	7	2	100

(b) MSE comparison

Δ	LC-only	DASH (no screen)	DASH (screen)
0	0.164	0.061	0.061
0.25	0.164	0.064	0.066
0.5	0.164	0.075	0.066
1.0	0.164	0.124	0.066
1.5	0.164	0.187	0.066
2.0	0.164	0.232	0.066
3.0	0.164	0.249	0.066

E.2 Simulation Experiment 4: Computational Cost

Table 5: Experiment 4: Wall-clock time in seconds (mean $\pm$ std over 3 runs, $n = 300$, $n_r = 2{,}000$, $R = 2$).

p	LC-only	Impute-pool	DASH
4	0.47 ± 0.01	6.79 ± 0.08	2.52 ± 0.02
6	0.67 ± 0.02	9.39 ± 0.00	2.79 ± 0.02
8	0.77 ± 0.01	11.65 ± 0.07	3.04 ± 0.04
10	0.99 ± 0.01	14.43 ± 0.18	3.41 ± 0.03
12	1.10 ± 0.02	16.48 ± 0.20	3.62 ± 0.12

E.3 Beijing Air Quality

The Beijing Multi-Site Air Quality dataset (Zhang et al., 2017) contains hourly measurements from 12 monitoring stations over March 2013 to February 2017. We aggregate to daily averages and retain observations complete on the variables used by each source.

The target variable is daily mean PM2.5 concentration. The $p = 10$ predictor features are: PM10, SO2, NO2, CO, O3 (pollutants) and TEMP, PRES, DEWP, RAIN, WSPM (meteorological variables).

Table 6: Beijing Air Quality: data partition.

Source	Station	n	Missing
$\mathcal{LC}$	Dongsi	300	—
$\mathcal{LM}_1$	Changping	1,450	CO, O3
$\mathcal{LM}_2$	Huairou	1,445	SO2, NO2
$\mathcal{LM}_3$	Shunyi	1,384	DEWP, PRES
$\mathcal{UC}$	8 stations	11,375	—

To construct a controlled multi-site scenario with complementary blockwise missingness, we designate one urban station (Dongsi) as $\mathcal{LC}$ with all features observed ($n = 300$ days subsampled per replication). Three suburban stations serve as $\mathcal{LM}$ sites, each with a different pair of features artificially removed: CO and O3 at Changping, SO2 and NO2 at Huairou, and DEWP and PRES at Shunyi. The remaining eight stations form $\mathcal{UC}$ ($N = 11{,}375$ days) with complete feature coverage but no outcome variable used during Shapley estimation. This design ensures complementary missingness patterns across sites while preserving real inter-station distributional heterogeneity.

All features and the outcome are standardized using $\mathcal{UC}$ means and standard deviations. Four regression models (linear, random forest, GBM, MLP) are trained on $15{,}646$ observations from all non-$\mathcal{LC}$ stations and held fixed during Shapley estimation. The reference attribution vector is computed from $2{,}000$ held-out training observations.

E.4 NACC Alzheimer's Data

Table 7: NACC Alzheimer's: data partition.

Source	n	Missing
$\mathcal{LC}$ (random)	300	—
$\mathcal{LM}_1$	3,756	WEIGHT, HEIGHT, NACCLIVS
$\mathcal{LM}_2$	3,756	ALCOHOL, TOBAC100, CDRLANG
$\mathcal{LM}_3$	3,755	SEX, RACE, NACCAGE
$\mathcal{UC}$	5,633	—

The National Alzheimer's Coordinating Center (NACC) Uniform Data Set (UDS) aggregates clinical assessments from over 40 Alzheimer's Disease Research Centers across the United States. We use the investigator dataset which contains demographic, clinical, and cognitive variables collected under heterogeneous protocols across centers.

The outcome variable is the Mini-Mental State Examination (MMSE) score, a continuous measure of cognitive function ranging from 0 (severe impairment) to 30 (no impairment). The $p = 9$ predictor features are: sex, race, age at visit (NACCAGE), weight, height, living situation (NACCLIVS), alcohol use history, tobacco use history (TOBAC100), and CDR language domain score (CDRLANG). Observations with missing values in any feature or the outcome are excluded, yielding $17{,}200$ complete cases from 34 centers.

To construct a controlled multi-site scenario with complementary blockwise missingness, we randomly sample $n = 300$ patients from the pooled complete cases to serve as $\mathcal{LC}$. The remaining patients are partitioned into three $\mathcal{LM}$ sites and one $\mathcal{UC}$ pool. Each $\mathcal{LM}$ site has a different triplet of features artificially removed (Table 7), ensuring complementary missingness patterns. This design satisfies Assumption 1 by construction, as all partitions are drawn from the same population.

Continuous features (NACCAGE, WEIGHT, HEIGHT, CDRLANG) are standardized using the $\mathcal{UC}$ means and standard deviations. Four regression models (ridge, random forest, GBM, MLP) are trained on all non-$\mathcal{LC}$ patients ($n_{\text{train}} = 16{,}900$) and held fixed during Shapley estimation. The reference attribution vector is computed from all non-$\mathcal{LC}$ patients ($n_{\text{train}} = 16{,}900$) with $m = 300$ sampled coalitions, providing an evaluation target independent of the $\mathcal{LC}$ sample used for attribution estimation.