BFHD: Bidirectional Feature Harmonization Decomposition for Heterogeneous Clinical Assessments

Yuanhao Zhuo[1] , **Zixi Qin**[1] , **Ling Qin**[2] and **Wanqing Li**[1*]
[1]University of Wollongong, Wollongong, NSW 2522, Australia
[2]The People's Hospital of Guangxi Zhuang Autonomous Region, Nanning, Guangxi 530021, China
{yzhuo, wanqing}@uow.edu.au, zq181@uowmail.edu.au, lqin@gxams.org.cn

Abstract

Clinical assessments are often collected using heterogeneous assessment systems across centers and time, leading to records that mix different but related sets of measurements. This motivates harmonization beyond total-score linking. We formulate clinical harmonization at the measurement level as a bidirectional recoverability problem: given paired observations from two assessment systems, the goal is to identify which measurements can be reliably translated in both directions within an application-defined tolerance, while separating non-translatable components. We propose Bidirectional Feature Harmonization Decomposition (BFHD), a feasibility-driven framework that enforces bidirectionally coupled translation and uses feature-wise output gating to produce an explicit decomposition in the original measurement space. Experiments on synthetic data and real clinical assessment pairs show that BFHD achieves broader feasible harmonization coverage and improved subset stability compared to baselines.

1 Introduction

Clinical assessment of chronic and neuropsychiatric conditions relies on a wide range of rating scales, cognitive tests, and questionnaire-based assessments[American Psychiatric Association, 2022; Lui *et al.*, 2025]. In real clinical workflows, however, different centers, cohorts, and time periods often employ different combinations or versions of these assessment systems[Villalpando *et al.*, 2025]. Even when two assessment systems aim to measure similar clinical domains such as memory or executive function, their measurements (features) typically include both overlapping elements and system-specific components[Willcutt *et al.*, 2005]. As patients move between centers or as protocols evolve, longitudinal records naturally become mixtures of heterogeneous assessment systems rather than observations from a single fixed system[Lawrence *et al.*, 2017].

This heterogeneity creates a central challenge for clinical data integration and longitudinal analysis[Fortin *et al.*, 2017]. Beyond aligning summary scores, practical harmonization requires understanding how two assessment systems relate at the measurement level: which measurements can be reliably translated across systems, which encode system-specific information, and whether such relationships are consistent in both directions[Dorans, 2007; Kolen and Brennan, 2013]. These questions are critical for cross-protocol integration, retrospective cohort pooling, and interpretation of heterogeneous clinical records[Riley *et al.*, 2010].

Existing harmonization practices primarily address this problem through total-score linking and calibration methods, which estimate transformations between total scores or low-dimensional summaries[Kolen and Brennan, 2013]. While effective for aligning marginal outcomes, these approaches do not characterize bidirectional recoverability at the measurement level between heterogeneous assessment systems[Lord, 2012]. In practice, clinicians and researchers are often left without guidance on which individual measurements can be meaningfully harmonized and which should be treated as non-translatable[Hammarström *et al.*, 2016]. In contrast to feature selection approaches that prioritize features based on their utility for downstream prediction, clinical harmonization requires identifying which measurements can be meaningfully translated across assessments while explicitly preserving system-specific information[Molnar, 2020].

In this work, we formalize clinical harmonization at the measurement level as a bidirectional recoverability problem. Given paired observations from two assessment systems, we seek to decompose each system into (i) a harmonizable component consisting of measurements that admit reliable bidirectional translation under a bidirectionally coupled model that enforces synchronous updates across directions, and (ii) a complementary non-translatable component whose information cannot be recovered from the other system. Recoverability is defined as a model-relative, feasibility-based notion, extending classical total-score linking from summary transformations to multivariate, feature-level harmonization.

We introduce Bidirectional Feature Harmonization Decomposition (BFHD), a framework that formulates clinical harmonization as a decomposition of measurements into bidirectionally recoverable and non-translatable components, identified directly in the original measurement spaces. BFHD

*Corresponding author.

[1]Code and supplementary technical materials (SM) are available at https://github.com/Zhuo-Yuanhao/BFHD.

operationalizes this formulation through a bidirectionally coupled translation model that enforces consistent forward and backward relationships between assessment systems, together with feature-wise gating that determines which measurements must be reconstructed to satisfy cross-system feasibility within a clinically meaningful tolerance[Louizos *et al.*, 2017]. This yields an explicit and interpretable decomposition without task supervision or latent projections.

The contributions of this work are as follows:

- We formulate clinical harmonization as a bidirectional recoverability problem at the measurement level, and provide a principled, model-relative definition of harmonizable and non-translatable measurements.
- We develop a feasibility-driven, single-operator framework with feature-wise gating that identifies large bidirectionally recoverable subsets directly in the original measurement spaces.
- We demonstrate on synthetic and real clinical datasets that BFHD reveals meaningful harmonization structure and supports utilities including cross-protocol integration, temporal deployment under protocol drift, and robustness to test-time missingness.

2 Related Work

2.1 Clinical Harmonization and Task-Driven Baselines

Classical psychometric linking and equating methods, including total-score linking [Dorans, 2007; Kolen and Brennan, 2013], equipercentile equating [Kolen and Brennan, 2013], and IRT calibration [Cella *et al.*, 2007; Lord, 2012], map between assessment systems primarily at the level of total scores or low-dimensional traits. They support comparability across protocols but do not identify which individual measurements are bidirectionally recoverable between heterogeneous assessments.

In clinical ML, harmonization is often approached via predictive mappings [Fortin *et al.*, 2017; Dewey *et al.*, 2019; Dinsdale *et al.*, 2021] using tree-based models such as feature-wise random forests (FRF) [Breiman, 2001; Breiman *et al.*, 2017] or gradient boosted trees (XGB) [Chen, 2016]. These models can yield strong unidirectional prediction but their feature attributions depend on a chosen target and do not characterize intrinsic bidirectional recoverability. ComBat-style statistical harmonization [Johnson *et al.*, 2007; Beer *et al.*, 2020] removes site/batch effects as nuisance variation, which is misaligned with our goal of explicitly retaining non-translatable system-specific content.

2.2 Multiview Learning and Bidirectional Structure

Multiview learning relates multiple representations of the same individuals. CCA and its variants [Hotelling, 1935; Bach and Jordan, 2005; Yang *et al.*, 2019; Andrew *et al.*, 2013] and shared–private latent models such as multiview autoencoders and variational approaches [Kingma and Welling, 2013] capture cross-view dependence via latent projections or likelihood objectives. However, these methods typically operate in latent spaces and do not yield an explicit account of which observed measurements are recoverable across systems in the original measurement space.

Bidirectional mappings have also been explored in translation settings. Cycle-consistent models couple two directions through closed-loop reconstruction losses [Zhu *et al.*, 2017], while tied-weight autoencoder-style models enforce parameter sharing [Hinton and Salakhutdinov, 2006; Bengio *et al.*, 2013]. We implement a cycle-consistent MLP (CMLP) as a bidirectional mapping baseline without explicit subset selection. To realize strict single-operator coupling in our framework, which enforces shared constraints, we leverage invertible networks from the normalizing flow [Dinh *et al.*, 2014; Dinh *et al.*, 2016; Kingma and Dhariwal, 2018]. A tied-weight MLP (TWMLP) serves as a non-invertible alternative instantiation in our ablations.

2.3 Gating and Its Relation to Feature Selection

Feature selection methods typically identify features that improve task performance, commonly via sparsity-inducing penalties such as L_1 regularization [Tibshirani, 1996; Ng, 2004] or via neural gating mechanisms [Louizos *et al.*, 2017; Yamada *et al.*, 2020]. In contrast, our gates are optimized as continuous proxies for membership in a bidirectionally recoverable subset under a shared operator, using feasibility-style reconstruction constraints rather than task supervision. This enables an explicit decomposition into recoverable and non-translatable components aligned with clinical harmonization.

3 Problem Formulation

We consider paired observations from two clinical assessment systems, represented by random vectors $X \in R^{d_X}$ and $Y \in R^{d_Y}$. Let $I_X = \{1, \ldots, d_X\}$ and $I_Y = \{1, \ldots, d_Y\}$ denote feature indices. We seek recoverable subsets of measurements $S_X \subseteq I_X$ and $S_Y \subseteq I_Y$, whose complements are treated as non-translatable.

3.1 Bidirectional model class

Recoverability is defined relative to a bidirectional model class that reflects the mutual nature of harmonization. Given paired observations from two assessment systems, we consider two mappings parameterized by Φ:

$$f_\Phi : \mathcal{X} \to \mathcal{Y}, \qquad g_\Phi : \mathcal{Y} \to \mathcal{X},$$

A key requirement is that the two directions are not learned independently. Instead, updates to the mapping from X to Y must simultaneously constrain, and be constrained by, the mapping from Y to X. That is, learning the relationship between two assessment systems in one direction necessarily informs the relationship in the reverse direction.

We formalize this requirement by restricting f_Φ and g_Φ to be bidirectionally coupled: they are derived from a shared parameterization or structural constraint, such that optimizing one direction induces corresponding updates in the other, rather than allowing two unrelated functions linked only by loss terms.

3.2 ε-recoverability

Conceptual definition Noise and heterogeneity in clinical assessments make exact recoverability unrealistic. We therefore define a population-level, approximate notion of recoverability. A measurement X_i is said to be ε-recoverable from Y under a shared bidirectional model parameter Φ if

$$E\big[\, d(X_i, g_\Phi(Y)_i)\,\big] < \varepsilon, \tag{1}$$

where $d(\cdot,\cdot)$ is a feature-wise discrepancy measure and $\varepsilon > 0$ specifies an application-level tolerance for clinically acceptable agreement rather than a tuning parameter for predictive accuracy. Analogously, a measurement Y_j is ε-recoverable if

$$E\big[\, d(Y_j, f_\Phi(X)_j)\,\big] < \varepsilon. \tag{2}$$

Recoverability is defined relative to a chosen bidirectional model class $\mathcal{M}$, in the same way that classical total-score linking defines equivalence relative to a family of allowable linking functions. Smaller values of ε correspond to stricter harmonization requirements, while larger values admit broader but looser forms of cross-assessment translation. This definition is population-level and combinatorial: it assumes access to the true expectations and to the exact subsets of features whose recoverability constraints are satisfied.

Practical relaxation The population-level conditions in (1)–(2) cannot be evaluated directly: expectations are unobserved, and searching over all subsets (S_X, S_Y) is combinatorial. We therefore introduce a lightweight relaxation that preserves the semantics of recoverability while making the problem estimable from data.

Each feature is assigned a gate variable $G_{X,i}, G_{Y,j} \in [0,1]$ that acts as a continuous proxy for membership in the recoverable subsets. Only gated features contribute to empirical reconstruction losses, so the optimizer keeps a gate open when the shared bidirectional operator can reconstruct that feature with sufficiently small empirical error. This operationalizes the theoretical ε-recoverability condition using standardized feature-wise reconstruction error, which approximate the magnitude of the deviations appearing in (1)–(2).

Features whose reconstruction errors remain large force their gates toward zero and are treated as non-translatable. This provides a compact and data-driven approximation to the conceptual BFHD formulation while maintaining its intended semantics; the neural instantiation is described in Section 4.

3.3 Feasible decompositions

A triple (S_X, S_Y, Φ) is called ε-feasible if each selected feature satisfies its recoverability condition:

$$(S_X, S_Y, \Phi) \in \mathcal{F}_\varepsilon \iff \begin{cases} E[\, d(X_i, g_\Phi(Y)_i)\,] < \varepsilon, & \forall i \in S_X, \\ E[\, d(Y_j, f_\Phi(X)_j)\,] < \varepsilon, & \forall j \in S_Y\,. \end{cases} \tag{3}$$

This definition does not require exact reconstruction and admits feasible solutions under realistic clinical noise levels.

3.4 The BFHD optimization problem

The BFHD problem seeks the largest bidirectionally recoverable subsets:

$$\max_{S_X \subseteq I_X,\, S_Y \subseteq I_Y,\, \Phi \in \mathcal{M}} \quad w|S_X| + (1-w)|S_Y| \tag{4}$$

$$\text{subject to} \quad (S_X, S_Y, \Phi) \in \mathcal{F}_\varepsilon. \tag{5}$$

Here w is a fixed user-specified weight rather than an optimization variable. Unless otherwise stated, we use $w = 0.5$.

An optimal solution (S_X^*, S_Y^*, Φ^*) yields the decomposition

$$X_{\mathrm{rec}} = X_{S_X^*}, \qquad X_{\mathrm{nt}} = X_{I_X \setminus S_X^*}, \tag{6}$$

$$Y_{\mathrm{rec}} = Y_{S_Y^*}, \qquad Y_{\mathrm{nt}} = Y_{I_Y \setminus S_Y^*}. \tag{7}$$

When multiple maximal feasible solutions exist, BFHD recovers one such solution. In practice, stability across random initializations is used to assess identifiability.

4 Methodology

This section presents a practical realization of the BFHD framework. The goal is to construct a shared bidirectional operator for feature-level harmonization that jointly defines the mappings $X \to Y$ and $Y \to X$, together with feature-wise gating mechanisms that approximate the subsets appearing in the conceptual formulation in Section 3. The proposed method consists of three components: a shared bidirectional backbone with strict parameter sharing across directions (instantiated with an invertible architecture in our main implementation), feature-wise gates defining relaxed recoverable subsets, and cross-view reconstruction losses that implement the recoverability constraints in a differentiable manner.

4.1 Architecture Overview

Given paired observations (X, Y), BFHD constructs two cross-view mappings using a shared bidirectional backbone, which is instantiated as an invertible neural network (INN) in our main implementation. The $X \to Y$ and $Y \to X$ mappings correspond to the forward and inverse passes of the same backbone. For $X \to Y$ and $Y \to X$, we compute

$$\tilde{Y} = f_\Phi(X), \qquad \tilde{X} = g_\Phi(Y), \tag{8}$$

and apply feature-wise L_0 gates at the outputs:

$$\hat{Y}_j = G_{Y,j}\,\tilde{Y}_j, \qquad \hat{X}_i = G_{X,i}\,\tilde{X}_i. \tag{9}$$

The invertible backbone enforces strict bidirectional parameter sharing, while the feature-wise gates determine which measurements in each system are required to be reconstructed. Only gated outputs contribute to the loss, providing a differentiable approximation to the recoverability constraints in the BFHD formulation.

4.2 Bidirectional Backbone via Invertible Networks

A key requirement in the BFHD framework is that the mappings $X \to Y$ and $Y \to X$ share a common parameterization. To satisfy this requirement, we instantiate (f_Φ, g_Φ)

using an invertible neural architecture. Unlike conventional flow-based models, the invertible structure is not used for likelihood estimation or sampling; instead, it provides a single coupled operator that guarantees strict bidirectional parameter sharing.

Instantiation with a NICE-style architecture We adopt a NICE-style architecture because its additive coupling provides stable training for tabular, low-dimensional clinical data and avoids the scaling instability of affine coupling, which is unnecessary in our non-likelihood setting[Dinh *et al.*, 2014].

Because the NICE architecture is exactly invertible by construction, the two cross-view mappings used in BFHD correspond to a single operator F_Φ and its analytic inverse F_Φ^{-1}. This guarantees strict bidirectional parameter sharing and thus directly satisfies the requirement specified in the BFHD formulation. Importantly, the invertible structure serves not as a generative modeling assumption, but as a mechanism that enforces a tightly coupled forward–backward relationship between the two clinical systems.

Dimension alignment When $d_X \neq d_Y$, both features are embedded into a common dimension $D = \max(d_X, d_Y)$ by zero-padding the lower-dimensional system. Padding dimensions are never supervised in the reconstruction losses and do not interact with the gating mechanism. This ensures that the decomposition concerns only the original features.

Role of Invertibility in BFHD Invertibility is used solely to enforce strict parameter sharing without assuming the invertibility in clinical assessment systems. The X→Y and Y→X mappings are the forward and inverse passes of a single operator. This eliminates forward–backward drift common in cycle-consistent dual networks and improves subset stability. After output gating, the overall mapping is intentionally non-invertible. We use a NICE-style INN as a convenient instantiation; a non-invertible tied-weight backbone is evaluated in ablations.

4.3 Output L_0 Gating

The BFHD problem requires determining, for each feature of X and Y, whether it belongs to the bidirectionally recoverable subset. To obtain a differentiable relaxation of these binary decisions, we introduce feature-wise gates

$$G_X = (G_{X,1}, \ldots, G_{X,d_X}), \qquad G_Y = (G_{Y,1}, \ldots, G_{Y,d_Y}),$$

with values in $[0, 1]$. A gate value near 1 indicates that the corresponding feature is treated as recoverable, and a value near 0 indicates that the feature is assigned to the non-translatable component.

Unlike conventional feature selection methods that gate inputs, BFHD applies gates at the outputs of the cross-view mappings. Thus the predicted vectors are masked according to the definitions in Eqs. (9), ensuring that the model always receives the full input X or Y while only the gated subset is required to be reconstructed. This matches the recoverability semantics of the BFHD formulation.

The gates are implemented using the hard-concrete relaxation of the L_0 penalty [Louizos *et al.*, 2017]. This parameterization yields stochastic binary gate variables during training while remaining differentiable in expectation, allowing gradient-based optimization of feature inclusion.

4.4 Gated Hinge Losses

The BFHD formulation defines ε-recoverability through measurement-level reconstruction error tolerances (Section 3). To approximate this conceptual feasibility constraint from data, we introduce a gated hinge loss with a stricter training tolerance $\varepsilon_{\text{train}}$, which serves as a conservative proxy for the target evaluation tolerance.

We first define a scale-normalized reconstruction error $\mathrm{err}(\cdot, \cdot)$ for individual features. In our implementation, we use feature-wise normalized mean squared error (NMSE). Normalization ensures that $\varepsilon_{\text{train}}$ has a consistent interpretation across heterogeneous features. Throughout this work, all assessment features are treated as numerical variables, and the same normalized squared-error-based discrepancy is used for all features. Ordinal items, when present, are modeled as numerical scores under this approximation.

Using err, we define the gated hinge losses for the two cross-view directions:

$$\mathcal{L}_{X \leftarrow Y} = \sum_{i=1}^{d_X} G_{X,i} \,\mathrm{ReLU}\big(\mathrm{err}(X_i, \tilde{X}_i) - \varepsilon_{\text{train}}\big), \tag{10}$$

$$\mathcal{L}_{Y \leftarrow X} = \sum_{j=1}^{d_Y} G_{Y,j} \,\mathrm{ReLU}\big(\mathrm{err}(Y_j, \tilde{Y}_j) - \varepsilon_{\text{train}}\big). \tag{11}$$

This loss has a direct feasibility interpretation. A gated feature incurs no penalty once its reconstruction error falls below $\varepsilon_{\text{train}}$, and therefore does not encourage over-optimization beyond the feasibility threshold. In practice, $\varepsilon_{\text{train}}$ is set substantially smaller than the evaluation tolerance $\varepsilon_{\text{cutoff}}$ to discourage borderline features from remaining gated open due to noise or finite-sample effects, thereby improving subset stability across random initializations. Penalties are applied only when the underlying ε-recoverability constraint cannot be satisfied under the shared bidirectional operator, creating pressure to either improve reconstruction or reduce the corresponding gate value.

4.5 Gate Reward Term

Using the same hard-concrete parameterization as above, we optimize gates to favor inclusion rather than sparsity. Unlike standard L_0 regularization, which penalizes gate activation to promote sparsity, BFHD requires the opposite behavior. Gates should remain open unless cross-view reconstruction forces them to close. Accordingly, we use a reward rather than a penalty:

$$\mathcal{L}_{\text{gate}} = -\lambda_{L_0} \left(\sum_{i=1}^{d_X} G_{X,i} + \sum_{j=1}^{d_Y} G_{Y,j} \right), \qquad \lambda_{L_0} > 0.$$

This term broadens the candidate recoverable subsets, while the gated cross-view losses apply opposing pressure by closing gates that lead to poor bidirectional reconstruction. Their interaction forms a continuous relaxation of the BFHD objective and identifies the largest empirically recoverable subsets.

4.6 Overall Objective

The complete training objective combines the gated cross-view reconstruction terms with the gate reward term:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{X \leftarrow Y} + \mathcal{L}_{Y \leftarrow X} + \mathcal{L}_{\text{gate}}.$$

Minimizing $\mathcal{L}_{\text{total}}$ jointly optimizes the shared bidirectional operator and the gating variables, producing an empirical approximation to the BFHD decomposition.

Training is performed using a staged procedure to stabilize optimization and avoid degenerate solutions; details in SM.

5 Experiments

5.1 Experimental Setup

Datasets: We use (1) a synthetic dataset generated using a known shared/private/noise-factor model (details and exact parameters in SM). This provides ground-truth recoverable and non-translatable measurement labels for evaluating structure recovery. (2) Alzheimer's Disease Neuroimaging Initiative (ADNI) cognitive assessments[Mueller *et al.*, 2005], and (3) a paired real-world ADHD assessment dataset consisting of two widely used symptom scales, collected under a controlled co-administration protocol and available upon reasonable request. All real-world datasets consist of paired, co-administered assessment systems with partially overlapping but non-identical measurement sets. To avoid subject-level leakage, each subject contributes at most one observation in all experiments unless otherwise stated.

Baselines: We compare our approach with:

1. Clinically-used harmonization and predictive baselines (IRT, FRF, XGB), which learn unidirectional or task-driven mappings;
2. Bidirectional learning models (CMLP, NICE) that enforce forward-backward consistency without explicit feasibility-based gating;
3. Ablated variants of our approach.

Implementation Details. Implementation details, model architectures, and training hyperparameters are reported in the SM. All experiments use 10 random seeds to assess stability ($s_{train} \in \{1, ..., 10\}$), and synthetic datasets are generated using fixed data-generation seeds ($s_{gen} \in \{1, ..., 5\}$). We distinguish between two tolerances: a training tolerance $\varepsilon_{\text{train}}$, used only for optimizing BFHD, and an evaluation tolerance $\varepsilon_{\text{cutoff}}$, used uniformly across all methods for reporting feasible measurements.

The feasibility tolerance $\varepsilon_{\text{cutoff}}$ is system- and context-dependent rather than universal. Prior psychometric analyses report substantial intrinsic measurement uncertainty for NeuroBAT subtests in ADNI [Crane *et al.*, 2023] and SNAP-IV items in ADHD [Huang *et al.*, 2023]. When measurement-level variance estimates are available, these uncertainties can be mapped to the normalized error scale used in our experiments; otherwise, they anchor the order of magnitude of acceptable reconstruction error. We therefore adopt $\varepsilon_{\text{cutoff}} = 0.6$ as a conservative, measurement-anchored reference operating point across all experiments, with full protocol details and sensitivity analyses provided in the SM.

5.2 Experiment 1: Synthetic Feasibility Recovery under Ground-Truth ε

We first evaluate BFHD in a controlled synthetic setting where ground-truth recoverability labels are available by construction. This experiment focuses on feasibility structure recovery: whether a method can correctly identify which measurements satisfy the ε-recoverability condition, rather than comparing reconstruction accuracy.

Evaluation protocol. Recoverability is evaluated as a structure recovery problem, with ground-truth defined at the feature-index (measurement) level. For BFHD, recoverability is determined directly by the learned output gates after binarization. For baseline methods that do not perform gating, recoverability is inferred post hoc using a relaxed feasibility criterion: a feature is classified as recoverable if its reconstruction error falls below $2\varepsilon_{\text{gen}}$, where ε_{gen} is the noise tolerance used in synthetic data generation. This conservative relaxation compensates for finite-sample estimation error and favors baseline models that do not explicitly optimize feasibility.

We restrict baselines in this experiment to models that naturally admit bidirectional reconstruction and feature-level feasibility evaluation. Other architectures evaluated in real-world experiments are omitted here, as the synthetic data are designed to probe bidirectional feasibility recovery rather than predictive accuracy under task-specific assumptions.

We report structure recovery accuracy (Acc), defined as the proportion of feature indices whose recoverability status is correctly identified, separately for X and Y. To assess robustness, we additionally report the stability of the recovered measurement subsets across random seeds using Jaccard (J) similarity. Other diagnostic metrics are reported in the SM.

Model	Acc_X	Acc_Y	J_X	J_Y
Ours	$0.880_{\pm 0.024}$	$0.830_{\pm 0.024}$	0.933	0.924
CMLP	$0.855_{\pm 0.015}$	$0.810_{\pm 0.049}$	0.975	0.789
NICE	$0.735_{\pm 0.039}$	$0.725_{\pm 0.040}$	0.837	0.814

Table 1: Synthetic feasibility recovery under ground-truth ε (s_{gen}=5). Acc_X and Acc_Y denote index-level structure recovery accuracy for systems X and Y, respectively (mean $\pm_{std}$ across 10 seeds). J_X and J_Y report Jaccard similarity. For baseline models without gating, recoverability is inferred post hoc using a relaxed $2\varepsilon_{\text{gen}}$ feasibility criterion. Additional results are provided in the SM.

Results. Table 1 summarizes the results. Our approach achieves the best overall structure recovery accuracy with strong subset stability across seeds among bidirectional baselines. In particular, methods without explicit feasibility-driven gating exhibit unstable behavior on weakly recoverable measurements near the $2\varepsilon_{\text{gen}}$ boundary, leading to increased variability across random seeds.

5.3 Experiment 2: Real-World Harmonization

We next evaluate BFHD on real clinical assessment systems to assess its practical utility for cross-protocol harmonization. In contrast to Experiment 1, where ground-truth recoverability is known, real-world datasets do not admit a definitive

notion of measurement-level equivalence. Accordingly, the goal of this experiment is not to maximize reconstruction accuracy, but to characterize how much of each assessment system can be harmonized under a shared bidirectional model while maintaining feasibility and stability.

Evaluation protocol. For each paired assessment system and each mapping direction, we evaluate harmonization outcomes using one primary and three supporting metrics. The primary metric is the number of recoverable measurements (#Rec). For BFHD, recoverable measurements are reported using the learned output gates. For baseline methods, which do not produce an explicit measurement-level recoverability indicator, recoverability is inferred at the feature level using the feasibility criterion $err < \varepsilon_{\text{cutoff}}$. The Jaccard similarity (J) of the recovered measurement subsets across random seeds, which quantify harmonization coverage and stability, respectively. Diagnostic metrics summarize the strength of agreement among the selected measurements using concordance correlation coefficient (CCC) thresholds[Lawrence and Lin, 1989; Barnhart *et al.*, 2007]. Specifically, we report the counts of selected measurements with strong agreement ($CCC > 0.7$) and marginal/moderate agreement ($CCC > 0.5$). These CCC-based counts are not used to define feasibility; rather, they help interpret whether the additional measurements included under ϵ-feasibility correspond to clinically meaningful agreement levels.

Unless otherwise stated, all real-world results are reported at a fixed feasibility threshold $\varepsilon_{\text{cutoff}} = 0.6$. We further analyze sensitivity to this choice through an ε-sweep in Figure 1 (with additional settings reported in the SM).

Results. Table 2 summarizes the harmonization results across all dataset pairs. Across all settings, our approach consistently identifies larger recoverable subsets than baseline methods while maintaining comparable or improved stability across random seeds. In particular, our approach recovers substantially more measurements than traditional linking and predictive baselines, reflecting its objective of maximizing feasible harmonization coverage rather than optimizing pointwise reconstruction accuracy.

Importantly, the additional measurements identified by our approach are not dominated by poorly reconstructed measurements. Across datasets, our approach matches baselines on the number of strongly aligned measurements while consistently increasing the number and stability of marginally aligned measurements. This pattern supports our goal of expanding feasible harmonization coverage without sacrificing the clinically reliable core.

Effect of the feasibility tolerance. To analyze how harmonization coverage varies with the feasibility criterion, we sweep the tolerance ε and report the number of recoverable measurements ($\#\text{rec}$). Figure 1 shows the resulting curves for the EMBIC↔NeuroBAT pair, plotted separately for each assessment system. Across a broad range of thresholds, our approach consistently recovers more measurements than all baseline methods, demonstrating that it can achieve wider harmonization coverage while respecting the same feasibility constraint. This advantage is especially pronounced in the intermediate regime (approximately $\varepsilon_{\text{cutoff}} \in [0.55, 0.8]$), which

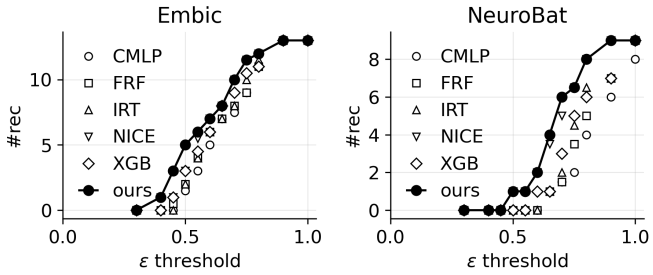

Figure 1: ε-sweep on EMBIC and NeuroBAT. Median number of recoverable measurements ($\#\text{rec}$) as a function of the feasibility threshold ε. CAS↔SNAP results are reported in the SM.

corresponds to realistic clinical settings where assessments share a partially overlapping core but differ in system-specific content. In contrast, baseline methods tend to be either overly conservative at lower thresholds or exhibit abrupt and unstable increases in $\#\text{rec}$ as $\varepsilon_{\text{cutoff}}$ is relaxed. These results highlight BFHD's ability to trace a smooth and interpretable coverage–feasibility frontier. Results for CAS↔SNAP and additional dataset pairs are deferred to the SM as they exhibit similar but less pronounced trends.

5.4 Experiment 3: Temporal Deployment Utility under Protocol Drift

While Experiment 2 evaluates harmonization outcomes under random splits from a fixed data distribution, real-world deployment requires applying harmonization decisions learned on earlier cohorts to data collected at a later time. Such temporal shifts may arise from changes in assessment procedures, rater behavior, cohort composition, or missingness patterns. Here we evaluate the temporal deployment utility of BFHD: whether measurement subsets identified as recoverable on early data remain feasible when deployed on later cohorts under potential protocol drift.

Temporal split and evaluation protocol. For each real-world dataset, we evaluate temporal deployment utility using a time-based split, where models are trained on early data and evaluated on a held-out later cohort; detailed split procedures and implementation details are provided in the SM.

Recoverable measurement subsets are defined on the training split and are not re-selected on the test split. For our approach, recoverable measurements are obtained by binarizing the learned output gates, while baseline methods infer recoverability on the training split using the same feasibility criterion as in Experiment 2. A training-period recoverable measurement is considered retained if it continues to satisfy $\text{err}_{\text{test}} < \varepsilon_{\text{cutoff}}$ on the held-out test split.

We report two deployment utility metrics: (i) the number of retained recoverable measurements on the test split (#Ret), and (ii) the retention rate (Ret%), defined as the proportion of training-period recoverable measurements that remain feasible at test time. Agreement strength of retained measurements is reported in the SM.

Results. Table 3 summarizes temporal deployment results for the representative EMBIC↔NeuroBAT assessment pair. Across methods, our approach retains a larger number of

Data	Method	$\#Rec_A$	J_A	$\#A > 0.7$	$\#A > 0.5$	$\#Rec_B$	J_B	$\#B > 0.7$	$\#B > 0.5$
	IRT	11	0.733	4.5	10.5	7	0.785	4	7
CAS(A)	XGB	9	0.710	6	9	7	0.842	5	7
$\leftrightarrow$	FRF	9	0.683	4.5	9	7.5	0.858	6	7.5
SNAP(B)	CMLP	9	0.640	5	9	7	0.764	6	7
	NICE	11	0.867	4	11	9	0.774	5	9
	Ours	11	0.930	4	11	9	0.817	5	9
	IRT	6	0.841	1	6	0	N/A	0	0
EMBIC(A)	XGB	6	0.806	1	6	1	0.4	0	1
$\leftrightarrow$	FRF	6	0.819	1	6	0	0.533	0	0
NeuroBAT	CMLP	5	0.739	1	5	0	0.491	0	0
(B)	NICE	7	0.806	1	7	2	0.523	0	1.5
	Ours	7	0.910	1	7	2	0.689	0	2

Table 2: Real-world harmonization results under $\varepsilon_{\text{cutoff}} = 0.6$. For each row, A and B correspond to the two assessment systems. $\#\text{Rec}_A$/$\#\text{Rec}_B$ are the median numbers of recoverable measurements and J_A/J_B are Jaccard stability across 10 seeds. $\#A$>0.7, $\#A$>0.5 (and analogously for B) count selected measurements whose CCC exceeds 0.7 or 0.5, used as diagnostic measures of feasibility.

Method	$\#\text{Ret}_A$	$\text{Ret}\%_A$	$\#\text{Ret}_B$	$\text{Ret}\%_B$
IRT	3	$42.9_{\pm 0.0}$	0	N/A
XGB	6	$85.7_{\pm 0.0}$	0.5	$62.5_{\pm 48.4}$
FRF	5	$56.0_{\pm 8.0}$	0	$0.0_{\pm 0.0}$
CMLP	6	$71.0_{\pm 7.5}$	1	$90.0_{\pm 10.0}$
NICE	6	$77.4_{\pm 9.9}$	2	$49.2_{\pm 10.5}$
Ours	7	$100.0_{\pm 0.0}$	3	$100.0_{\pm 0.0}$

Table 3: Temporal deployment utility under protocol drift on the EMBIC(A)↔NeuroBAT(B) assessment pair. $\varepsilon_{\text{cutoff}} = 0.6$. $\#\text{Ret}_A$/$\#\text{Ret}_B$ denote the numbers of training-period recoverable measurements that remain feasible on the held-out test split. $\text{Ret}\%_A$/$\text{Ret}\%_B$ report the corresponding retention rates. (mean $\pm_{std}$ across 10 seeds) Additional results in the SM.

training-period recoverable measurements under temporal shift while maintaining higher retention rates than baseline approaches. These results indicate that BFHD produces harmonization decisions that generalize more reliably across time and are therefore better suited for deployment in longitudinal and evolving clinical settings.

Additional utility analyses. Robustness under MCAR missingness[Little and Rubin, 2019; Rubin, 1976] and downstream transfer for prediction tasks are reported in the SM.

5.5 Experiment 4: Ablation and Stability Analysis

We examine how architectural and objective components contribute to the stability and behavior of BFHD.

Ablation settings. Starting from the full BFHD framework, we consider four ablations: (i) replacing the bidirectionally coupled invertible NICE backbone with a non-invertible tied weight backbone (TWMLP); (ii) switching the L_0 gate reward to a standard L_0 penalty; (iii) removing output gating; and (iv) replacing the hinge-based feasibility loss with a standard reconstruction loss. Each ablation modifies one component while keeping the others fixed.

Evaluation protocol. All variants are evaluated on representative real-world system pairs used in Experiment 2. We report the median number of recoverable measurements (#Rec) and the Jaccard similarity (J) across random seeds, which jointly capture coverage and stability.

Variant	$\#\text{Rec}_A$	J_A	$\#\text{Rec}_B$	J_B
Ours (full)	7	0.910	2	0.689
- TWMLP	7	0.886	1	0.600
- L_0 penalty	$0^\dagger$	N/A	$0^\dagger$	N/A
- No gating	7	0.806	2	0.523
- No hinge	$13^\dagger$	N/A	$9^\dagger$	N/A

Table 4: Ablation study on EMBIC (A) ↔ NeuroBAT (B) under $\varepsilon_{\text{cutoff}} = 0.6$. Notation are consistent with Table 2. Additional results in the SM. † indicate trivial gating solutions.

Results and observations. Table 4 summarizes the ablation results. Replacing the L_0 gate reward with a standard L_0 penalty consistently leads to degenerate solutions, confirming the necessity of explicitly favoring maximal feasible subsets. Removing output gating reduces subset stability, indicating that enforcing feasibility at the reconstruction level is critical for consistent measurement identification. Replacing the invertible backbone with a non-invertible tied-weight MLP preserves recoverable subset size, with dataset-dependent effects on stability. This suggests that strict bidirectional parameter sharing contributes to robustness, while invertibility itself is not a prerequisite for feasibility-based coverage.

Overall, these results show that effective BFHD instantiations depend critically on the joint use of a shared bidirectional backbone, output-level gating, and gated hinge losses.

6 Conclusion

We formulated clinical harmonization as a bidirectional recoverability problem and proposed BFHD, a feasibility-driven framework that identifies large bidirectionally recoverable measurement subsets under a shared operator. Across synthetic data and real clinical assessment pairs, BFHD consistently yields broader feasible harmonization coverage and higher subset stability than baselines. BFHD provides an explicit and interpretable decomposition for cross-protocol integration of heterogeneous clinical assessments.

Ethical Statement

This study uses de-identified datasets, including publicly available datasets and access-controlled datasets available upon reasonable request under appropriate approvals. No new data collection or human subject recruitment was conducted.

Acknowledgments

Data collection and sharing for the Alzheimer's Disease Neuroimaging Initiative (ADNI) is funded by the National Institute on Aging, National Institutes of Health Grant U19 AG024904. The grantee organization is the Northern California Institute for Research and Education.

Contribution Statement

Yuanhao Zhuo and Zixi Qin contributed equally to this work.

References

[American Psychiatric Association, 2022] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association, 2022.

[Andrew *et al.*, 2013] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR, 2013.

[Bach and Jordan, 2005] Francis R Bach and Michael I Jordan. A probabilistic interpretation of canonical correlation analysis. 2005.

[Barnhart *et al.*, 2007] Huiman X Barnhart, Michael J Haber, and Lawrence I Lin. An overview on assessing agreement with continuous measurements. *Journal of biopharmaceutical statistics*, 17(4):529–569, 2007.

[Beer *et al.*, 2020] Joanne C Beer, Nicholas J Tustison, Philip A Cook, Christos Davatzikos, Yvette I Sheline, Russell T Shinohara, Kristin A Linn, Alzheimer's Disease Neuroimaging Initiative, et al. Longitudinal combat: A method for harmonizing longitudinal multi-scanner imaging data. *Neuroimage*, 220:117129, 2020.

[Bengio *et al.*, 2013] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[Breiman *et al.*, 2017] Leo Breiman, Jerome Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Chapman and Hall/CRC, 2017.

[Breiman, 2001] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[Cella *et al.*, 2007] David Cella, Susan Yount, Nan Rothrock, Richard Gershon, Karon Cook, Bryce Reeve, Deborah Ader, James F Fries, Bonnie Bruce, Mattias Rose, et al. The patient-reported outcomes measurement information system (promis): progress of an nih roadmap cooperative group during its first two years. *Medical care*, 45(5):S3–S11, 2007.

[Chen, 2016] Tianqi Chen. Xgboost: A scalable tree boosting system. *Cornell University*, 2016.

[Crane *et al.*, 2023] Paul K Crane, Seo-Eun Choi, Michael Lee, Phoebe Scollard, R Elizabeth Sanders, Brandon Klinedinst, Connie Nakano, Emily H Trittschuh, Jesse Mez, Andrew J Saykin, et al. Measurement precision across cognitive domains in the alzheimer's disease neuroimaging initiative (adni) data set. *Neuropsychology*, 37(4):373, 2023.

[Dewey *et al.*, 2019] Blake E Dewey, Can Zhao, Jacob C Reinhold, Aaron Carass, Kathryn C Fitzgerald, Elias S Sotirchos, Shiv Saidha, Jiwon Oh, Dzung L Pham, Peter A Calabresi, et al. Deepharmony: A deep learning approach to contrast harmonization across scanner changes. *Magnetic resonance imaging*, 64:160–170, 2019.

[Dinh *et al.*, 2014] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

[Dinh *et al.*, 2016] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[Dinsdale *et al.*, 2021] Nicola K Dinsdale, Emma Bluemke, Stephen M Smith, Zobair Arya, Diego Vidaurre, Mark Jenkinson, and Ana IL Namburete. Learning patterns of the ageing brain in mri using deep convolutional networks. *NeuroImage*, 224:117401, 2021.

[Dorans, 2007] Neil J Dorans. Linking scores from multiple health outcome instruments. *Quality of Life Research*, 16(Suppl 1):85–94, 2007.

[Fortin *et al.*, 2017] Jean-Philippe Fortin, Drew Parker, Birkan Tunç, Takanori Watanabe, Mark A Elliott, Kosha Ruparel, David R Roalf, Theodore D Satterthwaite, Ruben C Gur, Raquel E Gur, et al. Harmonization of multi-site diffusion tensor imaging data. *Neuroimage*, 161:149–170, 2017.

[Hammarström *et al.*, 2016] Anne Hammarström, Hugo Westerlund, Kaisa Kirves, Karina Nygren, Pekka Virtanen, and Bruno Hägglöf. Addressing challenges of validity and internal consistency of mental health measures in a 27-year longitudinal cohort study–the northern swedish cohort study. *BMC medical research methodology*, 16(1):4, 2016.

[Hinton and Salakhutdinov, 2006] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[Hotelling, 1935] Harold Hotelling. The most predictable criterion. *Journal of educational Psychology*, 26(2):139, 1935.

[Huang *et al.*, 2023] Xia Huang, Hui-Qin Li, Alan Simpson, Jia-Jun Xu, Wan-Jie Tang, and Yuan-Yuan Li. Differences among fathers, mothers, and teachers in symptom assessment of adhd patients. *Frontiers in Psychiatry*, 14:1029672, 2023.

[Johnson *et al.*, 2007] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[Kingma and Dhariwal, 2018] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Kolen and Brennan, 2013] Michael J Kolen and Robert L Brennan. *Test equating: Methods and practices*. Springer Science & Business Media, 2013.

[Lawrence and Lin, 1989] I Lawrence and Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989.

[Lawrence *et al.*, 2017] Emma Lawrence, Carolin Vegvari, Alison Ower, Christoforos Hadjichrysanthou, Frank De Wolf, and Roy M Anderson. A systematic review of longitudinal studies which measure alzheimer's disease biomarkers. *Journal of Alzheimer's Disease*, 59(4):1359–1379, 2017.

[Little and Rubin, 2019] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2019.

[Lord, 2012] Frederic M Lord. *Applications of item response theory to practical testing problems*. Routledge, 2012.

[Louizos *et al.*, 2017] Christos Louizos, Max Welling, and Diederik P Kingma. Learning sparse neural networks through l_0 regularization. *arXiv preprint arXiv:1712.01312*, 2017.

[Lui *et al.*, 2025] Forshing Lui, Ischel Gonzalez Kelso, and Marjorie Launico. Cognitive assessment. *StatPearls*, 2025.

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

[Mueller *et al.*, 2005] Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. The alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics*, 15(4):869–877, 2005.

[Ng, 2004] Andrew Y Ng. Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78, 2004.

[Riley *et al.*, 2010] Richard D Riley, Paul C Lambert, and Ghada Abo-Zaid. Meta-analysis of individual participant data: rationale, conduct, and reporting. *Bmj*, 340, 2010.

[Rubin, 1976] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[Tibshirani, 1996] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

[Villalpando *et al.*, 2025] Juan Manuel Villalpando, Bernard-Simon Leclerc, Minh Tri Le, Carol Hudon, Aline Bolduc, Marie-Jeanne Kergoat, and Consortium for the Early Identification of Alzheimer's Disease-Quebec (CIMA-Q). A comparison of clinical diagnostic classification criteria used in longitudinal cohort studies of the alzheimer's disease continuum: A systematic review. *Neuropsychology Review*, pages 1–22, 2025.

[Willcutt *et al.*, 2005] Erik G Willcutt, Alysa E Doyle, Joel T Nigg, Stephen V Faraone, and Bruce F Pennington. Validity of the executive function theory of attention-deficit/hyperactivity disorder: a meta-analytic review. *Biological psychiatry*, 57(11):1336–1346, 2005.

[Yamada *et al.*, 2020] Yutaro Yamada, Ofir Lindenbaum, Sahand Negahban, and Yuval Kluger. Feature selection using stochastic gates. In *International conference on machine learning*, pages 10648–10659. PMLR, 2020.

[Yang *et al.*, 2019] Xinghao Yang, Weifeng Liu, Wei Liu, and Dacheng Tao. A survey on canonical correlation analysis. *IEEE Transactions on Knowledge and Data Engineering*, 33(6):2349–2368, 2019.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.