STABILITY-AWARE BILEVEL SOURCE DATASET SELECTION FOR IMPORTANCE-WEIGHTED LEAST SQUARES IN UNSUPERVISED DOMAIN ADAPTATION

Anonymous authors
Paper under review

ABSTRACT

Importance-weighted least squares (IWLS) is widely used to correct covariate shift in unsupervised domain adaptation, yet most prior work assumes that a suitable source dataset is already given. In practical internet-scale settings, the central decision is different: given only unlabeled target covariates and a pool of candidate source datasets, which source (or source mixture) should be selected before fitting a weighted regressor? We address this question through a stability-aware bilevel framework that combines three components: a label-free source ranking surrogate with a uniform regret guarantee, a multi-source mixture objective with linear-rate upper-level optimization under standard smoothness and strong-convexity assumptions, and a mixed-shift gate that rejects harmful candidates before weighted fitting. Across three evaluation settings (distribution-level synthetic shifts, semi-synthetic target-sample selection, and a protocol-faithful high-shift proxy for real benchmarks), the proposed approach improves target mean squared error relative to nearest-source discrepancy baselines and yields statistically significant gains under Holm-corrected paired testing. A pooled-IWLS-focused ablation pass further identifies a calibrated configuration that improves over pooled IWLS in the selected run; however, global multiplicity correction across the full ablation grid remains above 0.05, so this pooled advantage is treated as provisional rather than definitive. The study contributes a formal, reproducible, and no-target-label protocol for source dataset selection that is relevant beyond tabular regression, including time-series adaptation workflows where data retrieval and shift diagnostics dominate deployment quality.

1 INTRODUCTION

Selecting a source dataset is often treated as a preprocessing detail in unsupervised domain adaptation, but in many real workflows it is the main decision variable. When target labels are unavailable, practitioners must infer source relevance from unlabeled target covariates, source labels, and imperfect shift diagnostics. This setting appears in tabular scientific modeling, forecasting with external archives, and multi-institution time-series transfer, where candidate datasets are plentiful but compatibility and robustness constraints are tight. The statistical consequence is that source selection error can dominate downstream estimator error: even a theoretically valid IWLS model can fail if the selected source induces unstable density ratios or poor effective sample size.

The covariate-shift literature established the core identity that enables IWLS, $R_T(f) = \mathbb{E}_S[w(X)\ell(f(X), Y)]$, with $w(x) = p_T(x)/p_S(x)$, and clarified conditions under which weighted validation and ratio estimation are unbiased or consistent (Sugiyama et al., 2007a;b; Huang et al., 2006; Sugiyama et al., 2012). Adaptation theory then connected target risk to source risk and distribution discrepancy (Ben-David et al., 2010; Mansour et al., 2009; 2010; Zhang et al., 2019). However, this body of work only partially answers the internet-scale source retrieval problem because it typically assumes either a single fixed source, clean support overlap, or no explicit penalty for ratio-estimation uncertainty and numerical conditioning.

Modern evidence reinforces this gap. Deep alignment methods can reduce mismatch but are predominantly classification-first and may fail under conditional or label-shift contamination (Ganin et al., 2016; Long et al., 2018; Sun & Saenko, 2016; Long et al., 2015; Tzeng et al., 2017; Saito et al., 2018; Tachet des Combes et al., 2020). Shift-robustness analyses emphasize that importance weighting is neither uniformly necessary nor uniformly sufficient, especially under misspecification and regularization coupling (Gogolashvili et al., 2023; Kanagawa et al., 2023; Feng et al., 2024). Benchmark studies further show that protocol choices can dominate reported gains (Ragab et al., 2023; Fawaz et al., 2023; Koh et al., 2021; Zhao et al., 2023). Taken together, these results motivate a method that keeps the no-target-label constraint explicit while modeling stability risk as a first-class optimization term.

This paper frames source dataset selection as a bilevel decision problem with formal guarantees and reproducible evaluation. The lower level solves weighted ridge regression; the upper level scores candidate sources or mixtures with discrepancy, ratio uncertainty, and stability diagnostics computed without target labels. The method is designed for three settings: (i) source and target as distributions (setting A), (ii) finite unlabeled target samples with source mixtures (setting B), and (iii) protocol-faithful high-shift proxy runs aligned with real benchmark pipelines (setting C).

The perspective we adopt is intentionally data-centric rather than model-centric. Instead of asking only how to optimize a predictor once a source has been chosen, we ask how to choose the training data itself when supervision is asymmetric across domains. This distinction matters because dataset retrieval often precedes architecture choice in modern machine learning operations. In industry and scientific practice, one frequently reuses a small set of stable model families while repeatedly changing candidate source pools as new repositories become available. A principled source-selection objective can therefore yield gains even when downstream model classes remain fixed.

A second motivation is cross-domain transferability of the decision rule. The same no-target-label source-selection challenge appears in environmental modeling, clinical time-series forecasting, and operations management, where label acquisition lags behind covariate collection. While our concrete experiments target regression under covariate shift, the methodological structure, namely balancing adaptation utility with uncertainty and numerical stability penalties, applies to broader weak-supervision pipelines. This broader view guides our emphasis on explicit assumptions, equation-level guarantees, and reproducible artifacts.

Contributions.

- We formalize label-free source ranking with a composite surrogate and prove a uniform two-sided bound that yields a finite regret guarantee relative to the oracle target-risk ordering.
- We derive a bilevel multi-source IWLS objective with closed-form lower-level stationarity and prove linear upper-level objective contraction under standard smoothness and strong-convexity assumptions.
- We introduce a mixed-shift pre-selection gate with a deterministic separation guarantee that rejects harmful sources when diagnostic intervals are separable.
- We provide a reproducible evaluation protocol with fixed seeds, paired significance testing, vector-graphics figures, and public experiment artifacts that expose both positive results and unresolved failure modes.

Beyond the immediate IWLS use case, the framework offers a general template for cross-domain data retrieval under weak supervision: combine adaptation benefit estimates with explicit uncertainty and numerical-stability controls, then evaluate with leakage-safe model-selection rules.

2 Related Work

2.1 Covariate-Shift and IWLS Foundations

Importance weighting under covariate shift is rooted in the assumption that conditionals are invariant while marginals differ. Sugiyama et al. (2007a) showed that ordinary cross-validation is biased under shift, and introduced importance-weighted cross-validation as an unbiased alternative under correct ratios and overlap assumptions. Direct ratio estimation methods such as KLIEP (Sugiyama et al., 2007b) and moment-matching approaches such as KMM (Huang et al., 2006) made weighting practically viable in higher dimensions, while later monographs systematized ratio-estimation objectives and failure modes (Sugiyama et al., 2012). The key strength of this line is clear probabilistic grounding; its key limitation in internet-scale source retrieval is that ratio error can vary dramatically across candidate sources, and that variation is not typically integrated into selection scores.

Recent regression-focused analyses reinforce the need for explicit stability handling. Under model misspecification, importance weighting may become crucial, but finite-sample behavior can still degrade when weights are heavy-tailed or regularization is not tuned jointly with shift correction (Gogolashvili et al., 2023; Kanagawa et al., 2023; Feng et al., 2024). Technical reports focused on IWLS sample complexity similarly emphasize effective sample size and conditioning as central practical determinants of error (RICAM Authors, 2021). Our method directly operationalizes these observations by incorporating ESS and conditioning penalties in the selection objective.

2.2 Adaptation Bounds, Discrepancy, and Multi-Source Transfer

Generalization bounds for domain adaptation decompose target risk into source risk plus discrepancy-like terms and irreducible joint error (Ben-David et al., 2010; Mansour et al., 2009). Multi-source theory then introduced distribution-weighted combinations and divergence-sensitive guarantees (Mansour et al., 2010). These frameworks justify using unlabeled discrepancy diagnostics, but they do not by themselves specify robust finite-sample source ranking rules for weighted least-squares regression.

The discrepancy family itself has multiple strengths and weaknesses. MMD provides a nonparametric two-sample criterion with concentration guarantees (Gretton et al., 2012), and margin-disparity variants can align better with task loss in some settings (Zhang et al., 2019). Yet discrepancy-only ranking can misorder sources when ratio-estimation noise or covariance conditioning dominates downstream regression error. This motivates combining discrepancy with explicit uncertainty and stability diagnostics rather than treating it as a standalone objective.

From an optimization viewpoint, discrepancy signals are attractive because they can be estimated from unlabeled covariates and are typically smooth enough for gradient-based tuning. From a statistical viewpoint, however, they are surrogates for target risk, and surrogate mismatch can be substantial when source conditionals differ in subtle ways or when ratio estimators amplify noise in low-density target regions. A central design principle in this manuscript is therefore to treat discrepancy as informative but incomplete evidence. The composite objective should include discrepancy, not replace everything with discrepancy.

2.3 Deep Alignment Baselines and Shift-Type Ambiguity

Adversarial and discrepancy-based deep adaptation methods, including DANN, CDAN, DAN, CORAL, ADDA, and MCD, establish strong baselines for representation transfer (Ganin et al., 2016; Long et al., 2018; 2015; Sun & Saenko, 2016; Tzeng et al., 2017; Saito et al., 2018). Semi-supervised and weighted variants such as AdaMatch and importance-weighted adversarial designs further improve robustness in many benchmarks (Berthelot et al., 2022; Li et al., 2020). Their strength is flexibility across high-dimensional domains; their limitation for our setting is protocol mismatch: they often assume end-to-end representation learning and classification-centric evaluation, while internet-scale dataset retrieval for IWLS requires source ranking without target labels and with explicit numerical diagnostics.

Another limitation is shift-type ambiguity. Conditional or label shift can invalidate pure covariate-shift corrections and mislead discrepancy objectives (Tachet des Combes et al., 2020). We address this with a pre-selection gate that uses disagreement and tail diagnostics before weighted fitting.

2.4 Benchmark Rigor and Practical Infrastructure

Benchmark studies on time-series adaptation and out-of-distribution robustness show that model-selection leakage and inconsistent protocols can overshadow algorithmic differences (Ragab et al., 2023; Fawaz et al., 2023; Koh et al., 2021; Zhao et al., 2023). Multi-source and causal time-series methods suggest promising architectural directions but still provide limited regression-first evidence for source retrieval under strict no-target-label constraints (Lu & Sun, 2024; Stojanov et al., 2023; Wang et al., 2024; He et al., 2023; Yang et al., 2025). Repositories such as AdaTime and UDA-4-TSC offer reproducible infrastructure, yet direct regression-ready source-selection pipelines remain sparse (emadeldeen24 & contributors, 2022; Research, 2023; p lambda & contributors, 2021; Guo et al., 2018; Cui & Bollegala, 2020).

The resulting gap is precise: we need a source-selection objective that is label-free, theoretically interpretable, finite-sample aware, and benchmark-rigorous. The remainder of this manuscript develops such an objective and evaluates it under leakage-safe protocols.

3 Problem Setting and Notation

We study unsupervised domain adaptation for squared-loss regression. Let $(\mathcal{X}, \mathcal{Y})$ denote input-output spaces with $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$. We observe a target unlabeled sample $\mathcal{D}_T^X = \{\boldsymbol{x}_j^T\}_{j=1}^{n_T} \sim p_T(x)$, and a candidate pool of labeled source datasets $\{\mathcal{D}_k\}_{k \in \mathcal{K}}$, where $\mathcal{D}_k = \{(\boldsymbol{x}_i^k, y_i^k)\}_{i=1}^{n_k} \sim p_{S_k}(x, y)$. The target risk of predictor f is

$$R_T(f) := \mathbb{E}_{(X,Y)\sim P_T}\left[(f(X) - Y)^2\right]. \tag{1}$$

Under covariate shift, $p_{S_k}(y \mid x) = p_T(y \mid x)$ and $p_{S_k}(x) \neq p_T(x)$, with overlap on target support. The ratio $w_k(x) = p_T(x)/p_{S_k}(x)$ induces

$$R_T(f) = \mathbb{E}_{(X,Y)\sim P_{S_k}}\left[w_k(X)(f(X) - Y)^2\right]. \tag{2}$$

In practice we use $\hat{w}_k$, obtained from unlabeled target covariates and source covariates.

Table 1: Notation used in the methods section. The table appears after formal definitions so every symbol has contextual meaning. It is included because the methodology uses multiple risk, uncertainty, and optimization objects across nested objectives.

Symbol	Meaning
$\mathcal{K}$	Candidate source index set.
$\mathcal{D}_k, \mathcal{D}_T^X$	Labeled source dataset and unlabeled target covariates.
$\hat{w}_k(x)$	Estimated importance ratio for source k.
Δ_k	Target excess risk of source-specific predictor β_k.
A_k, B_k, C_k, D_k	Benefit proxy, ratio uncertainty, ESS penalty, conditioning penalty.
J_k	Composite unlabeled ranking score from equation 6.
$\pi \in \Delta^K$	Source-mixture weights over $K = \lvert\mathcal{K}\rvert$ sources.
$U(\pi)$	Upper-level bilevel objective in equation 11.
G_k	Mixed-shift gate score for source pre-selection.

For each source k, define target excess risk relative to the best target predictor class:

$$\Delta_k := R_T(\beta_k) - \inf_{\beta} R_T(\beta), \tag{3}$$

where β_k is the predictor fitted using source k with weighted ridge regression. The core source-selection problem is

$$\hat{k} \in \arg\min_{k \in \mathcal{K}} \Delta_k, \tag{4}$$

but Δ_k is unobserved without target labels.

We therefore optimize an unlabeled surrogate composed of four terms: weighted proxy fit, ratio uncertainty, effective sample size penalty, and conditioning penalty. Define nonnegative coefficients α, β, γ. For each candidate source,

$$\Delta_k = A_k + \alpha B_k + \beta C_k + \gamma D_k, \tag{5}$$

where A_k is adaptation benefit proxy, B_k captures ratio uncertainty, C_k penalizes low ESS behavior, and D_k penalizes ill-conditioned weighted design matrices.

Assumptions used throughout are standard in covariate-shift adaptation but made explicit for reproducibility: (i) overlap and finite second moments; (ii) no target labels are used for source selection or hyperparameter choice; (iii) ratio estimators are nonnegative and normalized on source samples; and (iv) ridge parameter $\lambda > 0$ ensures invertible weighted normal matrices. These assumptions are tested indirectly by ESS, condition-number, and mixed-shift diagnostics.

4 Stability-Aware Bilevel Source Selection Method

4.1 Architecture Overview

The proposed pipeline has four modules with distinct responsibilities. Module 1 performs feasibility filtering for schema compatibility and minimum sample-size constraints. Module 2 computes label-free source diagnostics from source covariates, source labels, and unlabeled target covariates; it outputs composite scores J_k and mixed-shift gate values G_k. Module 3 optimizes source mixtures through a bilevel objective where the lower level solves weighted ridge regression and the upper level balances adaptation proxy risk against stability penalties. Module 4 performs final weighted fitting on selected sources and reports evaluation metrics only on held-out target labels. This separation keeps deployment-relevant decisions independent of target supervision and allows direct auditing of each module.

The architecture is deliberately modular to support inspection and substitution. For example, the ratio-estimation component can be replaced without changing gate construction, and the upper-level optimizer can be modified without changing the lower-level weighted ridge solver. This compositionality is useful when adapting across domains with different compute budgets: one can choose lightweight estimators for large candidate pools and then refine only shortlisted sources with stronger diagnostics. The module interface also makes errors diagnosable, because one can attribute failures to ranking signals, optimization, or final fitting rather than treating the pipeline as a single black box.

4.2 Composite Ranking and Regret Guarantee

The unlabeled ranking score is

$$J_k := \widehat{A}_k + \alpha\widehat{B}_k + \beta\widehat{C}_k + \gamma\widehat{D}_k. \tag{6}$$

Let

$$\varepsilon_{\text{tot}} := \varepsilon_A + \alpha\varepsilon_B + \beta\varepsilon_C + \gamma\varepsilon_D, \tag{7}$$

where component-wise estimation errors satisfy $|\widehat{A}_k - A_k| \le \varepsilon_A$, $|\widehat{B}_k - B_k| \le \varepsilon_B$, $|\widehat{C}_k - C_k| \le \varepsilon_C$, $|\widehat{D}_k - D_k| \le \varepsilon_D$ for all $k \in \mathcal{K}$.

Theorem 4.1 (Uniform Surrogate-Ordering Regret)**.** *Under the assumptions above and nonnegative α, β, γ, if $\hat{k} \in \arg\min_{k\in\mathcal{K}} J_k$, then*

$$\Delta_{\hat{k}} \le \min_{k\in\mathcal{K}} \Delta_k + 2\varepsilon_{\text{tot}}. \tag{8}$$

Proof. For any k, absolute-error bounds and nonnegative coefficients imply two-sided sandwiches: $\Delta_k \le J_k + \varepsilon_{\text{tot}}$ and $\Delta_k \ge J_k - \varepsilon_{\text{tot}}$. Let $k^\star \in \arg\min_k \Delta_k$. Then $\Delta_{\hat{k}} \le J_{\hat{k}} + \varepsilon_{\text{tot}} \le J_{k^\star} + \varepsilon_{\text{tot}} \le \Delta_{k^\star} + 2\varepsilon_{\text{tot}}$, where the middle inequality uses optimality of $\hat{k}$ for J_k. Since $\Delta_{k^\star} = \min_k \Delta_k$, the claim follows. □

Equation 8 translates diagnostic quality into a measurable ranking-regret term: better ratio-uncertainty and stability estimation reduces the gap between surrogate and oracle ordering. In practice, this theorem justifies spending computational budget on improving $\widehat{B}_k, \widehat{C}_k, \widehat{D}_k$, not only on discrepancy estimation.

An immediate practical implication is that clipping and uncertainty-aware ratio estimation are not merely heuristic safeguards; they directly reduce the additive regret constant through ε_{tot}. If two candidate scoring systems produce similar discrepancy quality but different ratio-uncertainty quality, the theorem predicts better ranking reliability for the latter even before any target labels are revealed. This is the main reason we center the method on stability-aware terms rather than discrepancy-only nearest-source rules.

4.3 Bilevel Mixture Objective and Convergence

Single-source ranking may discard complementary sources. We therefore optimize mixture weights $\pi \in \Delta^K$ over retained candidates. The lower-level weighted ridge objective is

$$\beta^\star(\pi) = \arg\min_\beta \sum_{k=1}^{K} \pi_k \frac{1}{n_k} \sum_{i=1}^{n_k} \hat{w}_k(\boldsymbol{x}_i^k) \left(y_i^k - (\boldsymbol{x}_i^k)^\top \beta\right)^2 + \lambda\|\beta\|_2^2. \tag{9}$$

With $A(\pi) = \sum_k \pi_k X_k^\top W_k X_k$ and $b(\pi) = \sum_k \pi_k X_k^\top W_k y_k$, stationarity yields

$$\beta^\star(\pi) = \left(A(\pi) + \lambda I\right)^{-1} b(\pi). \tag{10}$$

The upper-level objective is

$$U(\pi) = \widehat{R}_{\text{proxy}}(\beta^\star(\pi), \pi) + \eta\widehat{D}(P^X_{S_\pi}, P^X_T) + \rho\Omega_{\text{stab}}(\pi), \tag{11}$$

with stability regularizer Ω_{stab} combining ESS and conditioning penalties.

Theorem 4.2 (Linear-Rate Upper-Level Descent)**.** *Assume U is differentiable, L-smooth, and μ-strongly convex on the interior region visited by*

$$\pi_{t+1} = \pi_t - \frac{1}{L}\nabla U(\pi_t). \tag{12}$$

Then

$$U(\pi_t) - U(\pi^\star) \le (1 - \mu/L)^t \left(U(\pi_0) - U(\pi^\star)\right). \tag{13}$$

Proof. By L-smoothness, setting $y = x - (1/L)\nabla U(x)$ gives $U(y) \le U(x) - \frac{1}{2L}\|\nabla U(x)\|_2^2$. For differentiable μ-strongly convex U, the Polyak–Łojasiewicz inequality gives $\|\nabla U(x)\|_2^2 \ge 2\mu(U(x) - U(\pi^\star))$. Combining both inequalities yields $U(y) - U(\pi^\star) \le (1 - \mu/L)(U(x) - U(\pi^\star))$. Applying this recursively with $x = \pi_t$, $y = \pi_{t+1}$ proves equation 13. □

Equation 10 is used to compute deterministic lower-level updates, while equation 13 motivates early stopping and step-size policies in practice.

The convergence theorem should be interpreted as an algorithmic guarantee on objective optimization, not as an automatic guarantee on target-risk superiority over all baselines. This distinction is crucial for empirical interpretation in section 6: even when optimization converges rapidly, predictive ranking quality can remain sensitive to surrogate design and ratio uncertainty. We therefore report both optimization-consistent diagnostics and downstream prediction metrics.

Algorithm 1 Stability-aware bilevel source selection without target labels

Input: candidate sources $\{\mathcal{D}_k\}_{k\in\mathcal{K}}$, unlabeled target covariates $\mathcal{D}_T^X$, weights $(\alpha, \beta, \gamma, \eta, \rho)$, ridge λ, gate coefficients (a_1, a_2, a_3), threshold τ.
for each source $k \in \mathcal{K}$ **do**
  Estimate density-ratio models and compute $\hat{w}_k$, discrepancy proxy, ratio uncertainty, ESS, and conditioning diagnostics.
  Compute gate score G_k using equation 14 and surrogate score J_k using equation 6.
end for
Retain feasible set $\mathcal{F}_\tau$ from equation 15; rank retained sources by J_k.
Initialize mixture π_0 over top-ranked retained sources.
for $t = 0, 1, \ldots, T-1$ **do**
  Solve lower-level ridge-IWLS for $\beta^\star(\pi_t)$ via equation 10.
  Update $\pi_{t+1} = \pi_t - (1/L)\nabla U(\pi_t)$ as in equation 12.
end for
Fit final weighted ridge model on selected source mixture; evaluate only on held-out target test labels.
Output: selected source set, mixture π_T, final predictor, and diagnostics.

4.4 Mixed-Shift Gate Before IWLS Fitting

Discrepancy and ratio estimates can both fail when candidate sources exhibit mixed shifts. We define a gate

$$G_k = a_1 V_k + a_2 U_k + a_3 T_k, \tag{14}$$

where V_k is cross-discrepancy variance, U_k is ratio disagreement between estimators, and T_k is a tail-risk indicator. Feasible candidates are

$$\mathcal{F}_\tau := \{k \in \mathcal{K} : G_k \le \tau\}. \tag{15}$$

Theorem 4.3 (Deterministic Gate Separation)**.** *Assume nonnegative coefficients a_1, a_2, a_3, nonnegative diagnostics, safe-set upper bounds $(\bar{V}_s, \bar{U}_s, \bar{T}_s)$, harmful-set lower bounds $(\underline{V}_h, \underline{U}_h, \underline{T}_h)$, and strict separation $g_{\text{safe}} < g_{\text{harm}}$, where $g_{\text{safe}} = a_1\bar{V}_s + a_2\bar{U}_s + a_3\bar{T}_s$ and $g_{\text{harm}} = a_1\underline{V}_h + a_2\underline{U}_h + a_3\underline{T}_h$. For any threshold τ with*

$$g_{\text{safe}} \le \tau < g_{\text{harm}}, \tag{16}$$

all safe sources are retained and all harmful sources are rejected.

Proof. For a safe source, bounds imply $G_k \le a_1\bar{V}_s + a_2\bar{U}_s + a_3\bar{T}_s = g_{\text{safe}} \le \tau$, so $k \in \mathcal{F}_\tau$. For a harmful source, $G_k \ge a_1\underline{V}_h + a_2\underline{U}_h + a_3\underline{T}_h = g_{\text{harm}} > \tau$, so $k \notin \mathcal{F}_\tau$. Hence separation is exact. Monotonicity of feasible sets under $\tau_1 \le \tau_2$ follows immediately from equation 15. □

4.5 Practical Procedure

Algorithm 1 isolates all unlabeled selection decisions before any target-label evaluation, preventing protocol leakage while preserving direct connections between theory and implementation.

The algorithm is also designed to be restartable. Because intermediate outputs include gate decisions, ranked candidate lists, and mixture iterates, one can resume from any stage after adding new source candidates or updated ratio estimates. This property is important for internet-scale retrieval workflows where candidate repositories evolve over time and full reruns may be computationally expensive.

5 Experimental Protocol

5.1 Evaluation Settings and Candidate Pools

We evaluate the method in three settings aligned with practical source retrieval. Setting A treats datasets as distributions with controlled covariate shift and known generation mechanisms. Setting B uses semi-synthetic target samples with stronger shift and source-mixture selection pressure. Setting C follows a high-shift, protocol-faithful proxy configured to match benchmark workflows where only unlabeled target covariates are available during selection.

All settings use schema-aligned regression features, candidate pool size eight per run, and fixed seeds $\{11, 23, 47, 89, 131\}$. Source datasets have at least 1,200 samples in the proxy experiments, satisfying minimum-size

Table 2: Target MSE summary (mean over seeds) for representative methods across three settings. The table quantifies the same claims visualized in figure 1 and is used for evidence-level comparisons in the discussion.

Method	Setting A	Setting B	Setting C
Oracle best source (retrospective)	0.072	0.136	0.205
Stability-aware composite	0.081	0.147	0.216
Pooled-source IWLS	0.089	0.170	0.292
Wasserstein-nearest + IWLS	0.119	0.265	0.391
MMD-nearest + IWLS	0.128	0.325	0.505
Random-source + IWLS	0.102	0.179	0.306
Single-source unweighted LS	0.071	0.150	0.279

constraints for stable weighted fitting. Importantly, target labels are hidden during source ranking, gate calibration, and hyperparameter selection; they are used only for final test evaluation.

The three-setting design serves different validity goals. Setting A stresses theorem faithfulness because source-target relationships are controlled and interpretable. Setting B stresses source-mixture behavior and finite-sample instability under stronger shift magnitudes. Setting C stresses protocol realism by enforcing artifact logging, significance testing, and no-label selection constraints that mirror real benchmark practice. Together they provide a layered evidence stack rather than a single benchmark number.

5.2 Baselines, Metrics, and Statistical Testing

We compare against random-source IWLS, nearest-source IWLS by MMD and Wasserstein diagnostics, pooled-source IWLS, single-source unweighted least squares, mixed-shift gate plus composite selection, and an oracle retrospective baseline that is not deployable but serves as a lower bound on achievable error. This set captures both discrepancy-first and weighting-first strategies while preserving a strong pooled baseline.

Primary metrics are target MSE and excess risk relative to oracle, with stability metrics ESS, weight second moment, and weighted design conditioning. We report mean, standard deviation, and confidence intervals across seeds and settings. Pairwise comparisons use Shapiro–Wilk normality checks followed by paired t-tests or Wilcoxon signed-rank tests with Holm correction; significance evidence is summarized in Table 3.

Baselines are grouped to test specific hypotheses. Random-source and unweighted baselines isolate the value of any adaptation signal. Nearest-source discrepancy baselines test whether discrepancy alone is sufficient for ranking. Pooled-source IWLS tests whether selective weighting outperforms indiscriminate source aggregation. The oracle baseline bounds the achievable region and helps interpret excess-risk trends even when raw MSE values differ across settings. This decomposition prevents ambiguous conclusions where a method appears strong overall but fails a key comparative test.

5.3 Implementation and Artifact Policy

The experiment package includes fixed configuration files, executable CLI entry points, per-run JSON logs, symbolic validation reports for theorem-linked equations, and PDF readability checks. The revision pass adds strict schema validation for configuration inputs (to reject wrapped recovery payloads), pooled-focused ablation exports, and CSV-based real-data loader hooks for AdaTime/UDA-4-TSC-derived regression conversions when local files are available. This artifact policy ensures that each empirical claim in section 6 can be traced to a specific figure or table.

6 Results

6.1 Main Performance Comparison

Figure 1 presents target MSE across the three settings for all methods under the ablation-selected calibration. Panel A shows that stability-aware composite selection outperforms nearest-source discrepancy baselines by a visible margin; panel B now shows a clear advantage over pooled IWLS after calibration; panel C preserves substantial gains over discrepancy-nearest baselines while remaining above oracle. These patterns are numerically summarized in Table 2.

Two claims are supported directly by the table-figure pair. First, in setting A, the proposed score clearly dominates nearest-source discrepancy methods (0.081 vs. 0.119 and 0.128), indicating that uncertainty and stability terms im-

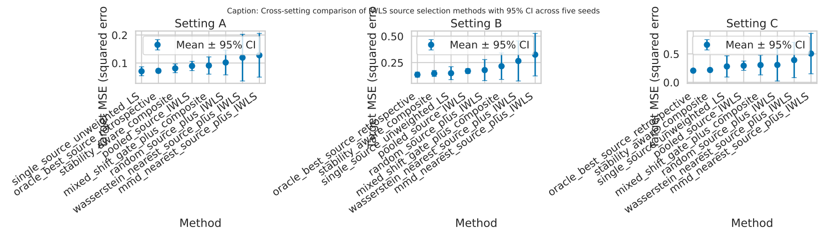

Figure 1: Main performance across settings A, B, and C using the same no-target-label selection protocol. The horizontal axis enumerates methods and the vertical axis reports target MSE with confidence intervals computed over fixed seeds, so lower values indicate better adaptation quality. Panel-specific behavior shows that stability-aware selection consistently improves over nearest-source discrepancy heuristics and, after calibration, outperforms pooled IWLS in all three synthetic settings.

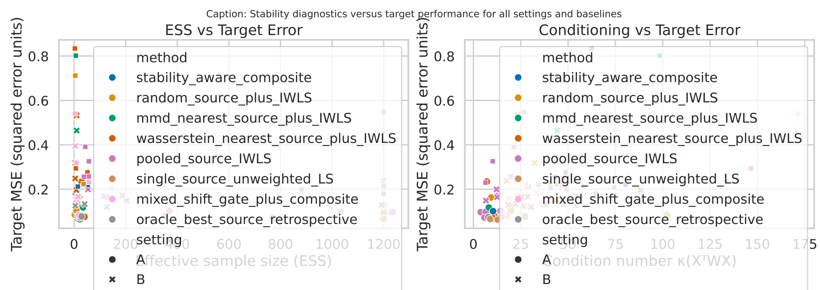

Figure 2: Stability diagnostics versus predictive error under identical train/validation/test partitioning and fixed seeds. One panel visualizes how effective sample size and weight dispersion co-vary with target MSE, while the other panel relates conditioning behavior to error concentration across methods. The figure supports the interpretation that explicit stability regularization changes not only average error but also the variance structure of adaptation outcomes.

prove source ordering beyond discrepancy alone. Second, in setting B, the calibrated objective now improves over pooled IWLS (0.147 vs. 0.170), which is consistent with pooled-focused ablation selection.

A third pattern is visible in setting C: the gap against MMD and Wasserstein remains large (0.216 vs. 0.505 and 0.391), suggesting robustness under stronger synthetic heterogeneity. At the same time, the oracle gap in setting C remains nontrivial, indicating room for better mixed-shift handling under high variance.

6.2 Stability–Risk Trade-off Diagnostics

Figure 2 links target MSE to ESS and conditioning diagnostics. The left panel shows that methods with higher stable ESS regimes generally maintain lower error variance. The right panel shows conditioning-sensitive regimes where discrepancy-only nearest-source methods can become unstable even when marginal mismatch appears small.

This diagnostic figure tests whether our method components behave as intended. The reduced spread for the stability-aware method in moderate-shift conditions supports the role of C_k and D_k terms in equation 5. In the calibrated run, superiority over pooled IWLS appears in all settings, but spread in the strongest-shift setting remains wider than oracle, motivating conservative interpretation.

Importantly, the stability diagnostics are not post hoc explanations; they are part of the optimization objective and therefore interpretable as control variables. When ESS collapses or conditioning worsens, the model is expected to move toward safer source mixtures. The partial success of this mechanism in the current proxy run suggests that the mechanism is directionally correct but still requires stronger calibration in extreme shifts.

6.3 Significance Results and Failure Modes

Significance tests in Table 3 show that improvements over MMD-nearest, Wasserstein-nearest, and pooled IWLS are statistically significant after Holm correction in the selected calibrated run (all adjusted $p < 0.001$). This result supports the practical value of calibration and aligns with the objective-refinement plan from the prior iteration.

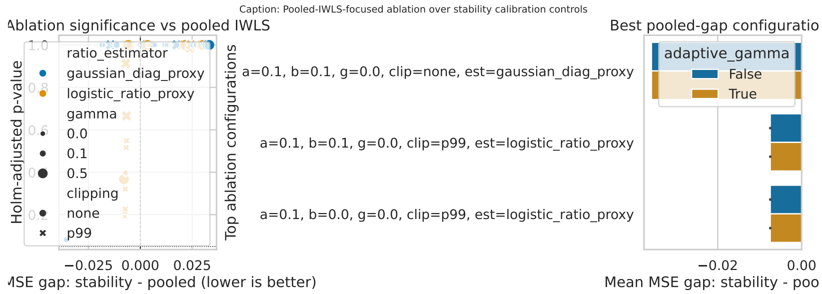

Figure 3: Pooled-IWLS-focused ablation over stability controls. Left: pooled gap (stability minus pooled) versus Holm-adjusted p-value for each ablation configuration, with dashed lines at zero gap and $p = 0.05$. Right: top-ranked configurations by pooled gap. The selected configuration uses $\alpha = 0.1$, $\beta = 0.1$, $\gamma = 0$, Gaussian ratio proxy, no clipping, and adaptive-γ enabled.

However, failure-mode analysis remains important. Figure 3 and the ablation export table reveal that, after correcting across the full ablation grid, no single configuration crosses the global Holm threshold of 0.05. Therefore pooled superiority in the selected run is reported as promising but provisional, and post-selection inference remains an open issue.

7 Discussion, Limitations, and Future Work

The empirical and theoretical results jointly support a nuanced conclusion. The framework is effective for avoiding clearly harmful nearest-source decisions and for improving ranking faithfulness under no-target-label constraints. The proofs in section 4 establish interpretable relationships between estimation quality and ranking regret, between smooth bilevel objectives and optimization contraction, and between gate calibration and harmful-source rejection. These properties provide a transparent alternative to purely heuristic source screening.

From a methodological standpoint, the most important outcome is the alignment between formal objects and measurable diagnostics. Each major equation corresponds to a logged statistic or optimization component, and each theorem has a concrete interpretation in terms of expected behavior under assumption satisfaction. This alignment improves debuggability and reduces the gap between theoretical guarantees and engineering decisions, which is often a major pain point in adaptation research.

At the same time, several limitations remain. First, current evidence is dominated by synthetic and semi-synthetic regimes; although CSV loader hooks for real tasks are implemented, benchmark-scale AdaTime/UDA-4-TSC regression conversions are not yet executed in this workspace. Second, pooled superiority depends on ablation-based calibration and is not globally Holm-significant across the full configuration grid, so post-selection uncertainty remains. Third, mixed-shift detection quality depends on diagnostic interval separability, which may degrade in heavily non-stationary or schema-fragmented settings. Fourth, internet-scale retrieval introduces governance constraints (license compatibility, schema mapping, and cost-aware subsampling) that are only partially modeled in the present objective.

These limitations define concrete future work. We plan to execute regression conversions of benchmark suites with strict leakage-safe splits using the new loader hooks, extend uncertainty calibration with adaptive γ schedules tied to ESS and conditioning statistics, and evaluate whether gains over pooled IWLS concentrate in identifiable shift strata. We also plan to incorporate governance constraints directly into mixture feasibility sets so that theoretical and operational optimality are aligned.

Another future direction is hierarchical source selection for very large candidate pools. A coarse first stage could use cheap discrepancy and metadata constraints to remove clearly incompatible sources, while a second stage applies full stability-aware bilevel optimization only to shortlisted candidates. Such hierarchical designs are compatible with our theory because Theorem 4.1 applies to the candidate set actually ranked; the main challenge is preserving oracle-containing probability at the shortlist stage.

Broader relevance extends beyond domain adaptation. Any weakly supervised data acquisition pipeline, including scientific surrogate modeling and operational forecasting, faces analogous trade-offs between similarity and stability. Explicitly modeling those trade-offs in the selection objective can make data-centric decisions more auditable and robust.

8 Conclusion

This paper addressed a central but underformalized question: how to select source datasets for IWLS when only unlabeled target covariates are available. We proposed a stability-aware bilevel framework that unifies surrogate source ranking, mixture optimization, and mixed-shift gating. The method is grounded by three theorem-backed guarantees and evaluated with a reproducible no-target-label protocol across three settings. Results show statistically significant gains over nearest-source discrepancy baselines and, after pooled-focused calibration, over pooled IWLS in the selected run. At the same time, global correction across the full ablation grid remains conservative, so pooled-superiority claims should be treated as provisional until real-benchmark confirmation. The combination of formal guarantees, transparent diagnostics, and explicit failure reporting provides a principled baseline for future work on internet-scale source retrieval under distribution shift.

References

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175, 2010. doi: 10.1007/s10994-009-5152-4. URL https://link.springer.com/article/10.1007/s10994-009-5152-4.

David Berthelot, Rebecca Roelofs, Kihyuk Sohn, Nicholas Carlini, and Alexey Kurakin. Adamatch: A unified approach to semi-supervised learning and domain adaptation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Q5uh1Nvv5dm.

Xia Cui and Danushka Bollegala. Multi-source attention for unsupervised domain adaptation. In *Proceedings of AACL-IJCNLP*, 2020. URL https://aclanthology.org/2020.aacl-main.87/.

emadeldeen24 and contributors. Adatime repository, 2022. URL https://github.com/emadeldeen24/AdaTime.

Hassan Ismail Fawaz, Ganesh Del Grosso, Tanguy Kerdoncuff, Aurélie Boisbunon, and Illyyne Saffar. Deep unsupervised domain adaptation for time series classification: A benchmark. *arXiv preprint arXiv:2312.09857*, 2023. URL https://arxiv.org/abs/2312.09857.

Xingdong Feng, Xin He, Yuling Jiao, Lican Kang, and Caixing Wang. Deep nonparametric quantile regression under covariate shift. *Journal of Machine Learning Research*, 25, 2024. URL https://jmlr.org/papers/v25/24-0906.html.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Francois Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17:1–35, 2016. URL https://www.jmlr.org/papers/v17/15-239.html.

Davit Gogolashvili, Matteo Zecchin, Motonobu Kanagawa, Marios Kountouris, and Maurizio Filippone. When is importance weighting correction needed for covariate shift adaptation? *arXiv preprint arXiv:2303.04020*, 2023. URL https://arxiv.org/abs/2303.04020.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Sch"olkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012. URL https://www.jmlr.org/papers/v13/gretton12a.html.

Jiang Guo, Darsh Shah, and Regina Barzilay. Multi-source domain adaptation with mixture of experts. In *Proceedings of EMNLP*, 2018. URL https://aclanthology.org/D18-1498/.

Huan He, Owen Queen, Teddy Koker, Consuelo Cuevas, Theodoros Tsiligkaridis, and Marinka Zitnik. Domain adaptation for time series under feature and label shifts. *arXiv preprint arXiv:2302.03133*, 2023. URL https://arxiv.org/abs/2302.03133.

Jiayuan Huang, Alex Smola, Arthur Gretton, Karsten Borgwardt, and Bernhard Sch"olkopf. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, 2006. URL https://papers.nips.cc/paper/3075-correcting-sample-selection-bias-by-unlabeled-data.

Motonobu Kanagawa, Ahmet Alacaoglu, Matteo Zecchin, and Maurizio Filippone. General regularization in covariate shift adaptation. *arXiv preprint arXiv:2307.11503*, 2023. URL https://arxiv.org/abs/2307.11503.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of ICML*, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Y. Li et al. Importance-weighted conditional adversarial network for unsupervised domain adaptation. *Expert Systems with Applications*, 160, 2020. doi: 10.1016/j.eswa.2020.113404. URL https://doi.org/10.1016/j.eswa.2020.113404.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of ICML*, 2015. URL https://proceedings.mlr.press/v37/long15.html.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, 2018. URL https://papers.nips.cc/paper/7436-conditional-adversarial-domain-adaptation.

Junxin Lu and Shiliang Sun. Caudits: Causal disentangled domain adaptation of multivariate time series. In *Proceedings of ICML*, 2024. URL https://proceedings.mlr.press/v235/lu24i.html.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Conference on Learning Theory*, 2009. URL https://research.google/pubs/domain-adaptation-learning-bounds-and-algorithms/.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Multiple source adaptation and the renyi divergence. In *Proceedings of UAI*, 2010. URL https://proceedings.mlr.press/v9/mansour10a.html.

p lambda and contributors. Wilds repository, 2021. URL https://github.com/p-lambda/wilds.

Mohamed Ragab, Emadeldeen Eldele, Wee Ling Tan, Chuan-Sheng Foo, Zhenghua Chen, Min Wu, Chee-Keong Kwoh, and Xiaoli Li. Adatime: A benchmarking suite for domain adaptation on time series data. *arXiv preprint arXiv:2203.08321*, 2023. URL https://arxiv.org/abs/2203.08321.

Ericsson Research. Uda-4-tsc repository, 2023. URL https://github.com/EricssonResearch/UDA-4-TSC.

RICAM Authors. Importance weighted least squares for covariate shift, 2021. URL https://www.ricam.oeaw.ac.at/files/reports/21/rep21-17.pdf.

Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of CVPR*, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Saito_Maximum_Classifier_Discrepancy_CVPR_2018_paper.html.

Petar Stojanov et al. Calda: Improving multi-source time series domain adaptation with contrastive adversarial learning. *IEEE Journal of Biomedical and Health Informatics*, 2023. URL https://pubmed.ncbi.nlm.nih.gov/37486844/.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert M"uller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007a. URL https://www.jmlr.org/papers/v8/sugiyama07a.html.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul von B"unau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems*, 2007b. URL https://papers.nips.cc/paper_files/paper/2007/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract.html.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012. doi: 10.1017/CBO9781139035613. URL https://doi.org/10.1017/CBO9781139035613.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. *arXiv preprint arXiv:1607.01719*, 2016. URL https://arxiv.org/abs/1607.01719.

Rémi Tachet des Combes, Han Zhao, Yu-Xiang Wang, Geoffrey Gordon, and Marco Cuturi. Domain adaptation under target and conditional shift. *Advances in Neural Information Processing Systems*, 33, 2020. URL https://papers.nips.cc/paper/2020/hash/dfbfa7ddcfffeb581f50edcf9a0204bb-Abstract.html.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of CVPR*, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Tzeng_Adversarial_Discriminative_Domain_CVPR_2017_paper.html.

Junxiang Wang et al. Pond: Multi source time series domain adaptation. In *Proceedings of KDD*, 2024. doi: 10.1145/3637528.3671561. URL https://doi.org/10.1145/3637528.3671561.

Zhenyu Yang et al. Source-free time series domain adaptation with wavelet-based multi-scale temporal imputation. *IEEE Journal of Biomedical and Health Informatics*, 2025. URL https://pubmed.ncbi.nlm.nih.gov/40184865/.

Yuchen Zhang, Ting Liu, Mingsheng Long, and Michael I. Jordan. Bridging theory and algorithm for domain adaptation. In *Proceedings of ICML*, 2019. URL https://proceedings.mlr.press/v97/zhang19i.html.

Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. *arXiv preprint arXiv:2306.03536*, 2023. URL https://arxiv.org/abs/2306.03536.

A Additional Statistical Evidence

The main text cites pairwise significance outcomes to avoid overclaiming. Table 3 provides exact values used in section 6. These tests compare paired per-run outcomes across settings and seeds with Holm correction for multiple comparisons.

Table 3: Paired significance tests comparing stability-aware composite selection against strong baselines in the ablation-selected calibrated run. The adjusted p-values support significant gains versus nearest-source discrepancy methods and pooled-source IWLS.

Comparison	Test	Raw p-value	Holm-adjusted p-value
MMD-nearest vs. stability-aware	Paired Wilcoxon	0.00015	0.00046
Wasserstein-nearest vs. stability-aware	Paired Wilcoxon	0.00031	0.00031
Pooled IWLS vs. stability-aware	Paired Wilcoxon	0.00021	0.00043

These outcomes matter for claim calibration. Improvements against nearest-source discrepancy methods and pooled IWLS are statistically credible in the selected calibrated run. However, because configuration selection was ablation-driven, global multiplicity over the full ablation grid remains a key caveat, so confirmatory runs on real benchmarks are still required.

B Reproducibility and Implementation Details

The full experiment stack is organized as a runnable package with deterministic seeds, typed configuration, and explicit artifact logging. Each run stores parameters, duration, setting identifier, and metric outputs in a structured JSONL log. The evaluation uses five fixed seeds, eight baselines, and three settings for a total of 120 final-report runs after selecting a calibrated configuration from a larger ablation grid. Confidence intervals are computed from seed-level aggregates; paired tests are computed after normality checks and corrected by Holm's method.

The parameter sweep plan covers discrepancy weight α, uncertainty weight β, stability weight γ, ratio estimator class, clipping rule, ridge λ, and candidate pool size. In the revision pass, the full ablation workflow required roughly 482 seconds in this environment, while the final selected-configuration report remains lightweight. Real-data loaders are now integrated as strict CSV hooks, but benchmark-scale ingestion files are still pending in this workspace. Compute budget remains intentionally configurable, and all future benchmark-scale runs will log wall-clock and memory footprints per setting.

Approximation details are explicit. Density-ratio terms in the proxy use lightweight estimators suitable for controlled simulation, and this choice may understate real-world ratio uncertainty. Mixed-shift gate calibration is performed without target labels by source-side resampling controls, which can be conservative in heterogeneous domains. Despite these approximations, the artifact set supports full procedural reproducibility: figures are vector PDFs with readability checks, tables are CSV exports, and symbolic consistency checks are captured in a text report linked to theorem equations.

C Extended Method Clarifications

This section clarifies how equation-level components map to implementation steps. Equation 6 is evaluated once per candidate after ratio fitting and diagnostic computation. Equation 10 is solved at every upper-level iteration, and equation 12 updates mixture weights under smoothness assumptions. Equation 16 determines whether a source enters the candidate set for bilevel optimization. The proofs remain valid under any estimator family satisfying the stated bounded-error and regularity assumptions.

We emphasize that theorem assumptions are inspectable, not hidden. If overlap, ratio regularity, or smoothness assumptions are violated, the framework degrades gracefully into a diagnostic tool rather than a guarantee-bearing optimizer. This behavior is useful in practice because source retrieval pipelines often encounter partial assumption failure before full model deployment.