WHICH SOURCES SHOULD PARTICIPATE? TARGET-SPECIFIC MULTI-SOURCE CONFIGURATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-source learning often treats source participation as fixed, either using all available sources or ranking sources independently for a target. We instead study source participation as a target-specific configuration problem over non-empty subsets of complete source domains. Empirical analysis reveals three recurring patterns. Adding sources can reduce target performance even when more training data become available, near-tied validation configurations frequently reverse on test data, and larger configurations do not reliably outperform smaller ones when validation support is comparable. These observations motivate ARCS, an Admissible Risk-aware Configuration Selection rule that constructs a near-optimal validation set, resolves ambiguity through parsimony, and uses worst-class validation performance only as a localized safeguard. We evaluate ARCS across multiple real-world heterogeneous multi-source benchmarks spanning wearable sensing, clinical multi-site data, and multimodal sentiment analysis, where it remains competitive while selecting compact source configurations. Further analyses show that the main gain comes from validation admissibility combined with parsimony, that performance remains stable under reduced target evidence, and that approximate candidate search can reduce configuration evaluation without changing the final selection principle. Overall, our results suggest that source participation should be treated as a target-specific configuration decision rather than a fixed assumption in multi-source learning.

1 INTRODUCTION

Modern learning increasingly takes place in a fragmented source landscape, where multiple datasets address the same prediction problem but are collected independently under different populations, devices, protocols, or modality configurations. Healthcare provides a representative example, as different medical centers may collect data for the same task under different patient cohorts, acquisition settings, and measurement modalities. Such datasets form a potentially valuable source pool for a new target, but their heterogeneity makes joint use nontrivial. A source that is relevant to the task may still be mismatched with the target, redundant with other sources, or harmful when combined with them. Before asking how multiple sources should be aligned, weighted, transferred, or fused, we therefore ask a more basic question. Which complete source domains should jointly participate for a particular target?

This is a configuration-level decision rather than an isolated property of each source. The effect of adding one source may depend on both the target and the sources already selected. Consequently, individually useful sources need not form the most effective joint source set, and independently ranking sources may miss interactions that emerge only after they are combined. We study this problem when only limited target evidence is available and the decision variable is a non-empty subset of complete source domains.

Existing lines of work address related but different decisions. Domain adaptation studies how to transfer or align knowledge once the source domains are given Mansour et al. (2009); Zhao et al. (2020; 2024), while domain generalization seeks representations that generalize from available sources to unseen domains Muandet et al. (2013); Zhou et al. (2022). Source weighting, mixture modeling, and transferability estimation modulate or assess the contribution of available sources Ahmed et al. (2021); Han et al. (2023), and multimodal methods study missing or selected

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

Table 1: Comparison with closely related problem settings. ✓ denotes an explicit or core setting, △ denotes partial involvement, and × denotes that the factor is typically not central to the problem formulation.

Problem Setting	Multiple Sources	Modality Variation	Whole-source Selection	Configuration-dependent	Core Question
Single-source domain adaptation	×	×	×	×	How to adapt one source to a target?
Multi-source domain adaptation	✓	△	×	△	How to transfer from multiple given sources?
Domain generalization	✓	△	×	△	How to generalize to unseen domains?
Source weighting / mixture	✓	△	△	△	How much should each given source contribute?
Source ranking / transferability	✓	△	△	×	Which individual source is more transferable?
Incomplete / missing-modality learning	△	✓	×	×	How to learn when some modalities are unavailable?
Modality selection / acquisition	△	✓	×	△	Which modalities should be observed or used?
Ours	✓	✓	✓	✓	**Which source domains should jointly participate for this target?**

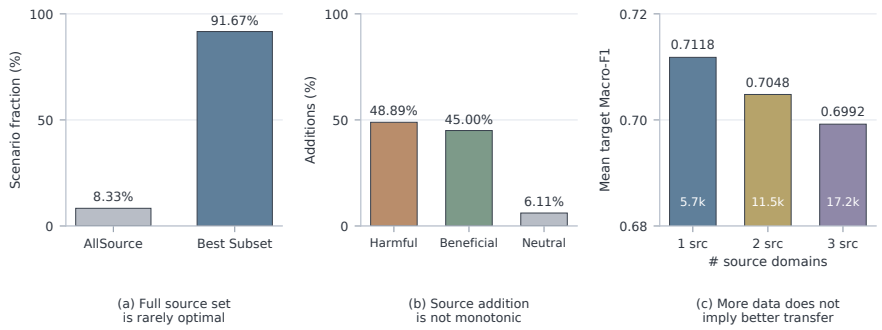


Figure 1: Problem motivation on HAR4-Natural. More participating sources and more source data do not consistently improve target performance, motivating target-specific source configuration selection.

modalities Wu et al. (2026); He et al. (2024). Our focus differs in the decision object. We select configurations of complete source domains for a specific target rather than source weights, independent source scores, or modality subsets. Table 1 summarizes this distinction.

This distinction has direct empirical consequences. We first examine an HAR4-Natural setting built from four heterogeneous activity-recognition datasets Anguita et al. (2013); Reiss & Stricker (2012); Banos et al. (2014); Kwapisz et al. (2011), where each source preserves its naturally available training samples and sensing setup. Using all available sources is test-optimal in only 8.33% of the evaluated scenarios, while a smaller source subset outperforms the full set in 91.67%. Moreover, 48.89% of legal source additions reduce target performance, and every source exhibits utility reversal across targets or source configurations. This behavior is not explained by having less source data. As the number of participating sources increases from one to three, the average amount of source data grows from 5,743.83 to 17,231.50 samples, while mean target Macro-F1 decreases from 0.7118 to 0.6992. Figure 1 therefore shows that more available source data do not imply a better source configuration.

We formalize this problem as *target-specific source configuration selection*. Given a pool of candidate source domains and limited target evidence, the goal is to select a non-empty subset of complete source domains that should jointly participate for a specific target. This problem introduces three challenges. First, limited target evidence makes fine-grained validation ranking unreliable. Second, source utility can be configuration-dependent, so independently assessing each source may miss harmful or redundant combinations. Third, the configuration space grows combinatorially with the number of candidate sources, making exhaustive evaluation increasingly costly.

Rather than designing a selector directly, we first study the empirical structure of these decisions. We find that near-tied validation configurations frequently change ordering on held-out target data, while larger configurations do not reliably outperform smaller ones when their validation support is comparable. Class-wise validation behavior provides additional risk information, but this signal is local and noisy. These observations motivate **ARCS**, an **Admissible Risk-aware Configuration Selection** rule. ARCS replaces exact validation ranking with a near-optimal admissible set, applies a localized worst-class safeguard within this set, and resolves the remaining ambiguity in favor of smaller configurations. For larger source pools, candidate search is separated from the final ARCS decision rule, allowing approximate search without redefining the selector. Our contributions are summarized as follows.

- We formulate target-specific source configuration selection as a distinct problem in heterogeneous multi-source learning. The decision object is a non-empty configuration of complete source domains whose utility can depend jointly on the target and the other participating sources.
- Through controlled empirical analysis, we identify validation-ranking instability, the absence of a consistent advantage for larger near-tied configurations, and only localized evidence from class-wise risk. These observations lead to ARCS, a simple selection rule built around validation admissibility, parsimony, and a conservative local safeguard.
- We conduct extensive evaluations across heterogeneous wearable-sensing, multi-site clinical, and multimodal settings, together with reduced-target-evidence, component, and search analyses. The results show that target-specific configuration selection remains effective across different source structures, while the main empirical benefit of ARCS comes from combining admissibility with parsimony rather than from aggressive risk screening.

2 Related Work

Multi-source domain adaptation and generalization. Domain adaptation and domain generalization study how knowledge from one or more source distributions can transfer to shifted or unseen domains. Multi-source domain adaptation has been developed through theoretical analysis, moment matching, source alignment, and deep multi-source transfer Mansour et al. (2009); Peng et al. (2019); Zhao et al. (2020; 2024), with recent extensions to multimodal settings Zhao et al. (2025); Huang et al. (2025a). These methods primarily address how to transfer, align, aggregate, or generalize from available sources. Our problem instead asks which complete source domains should participate for a target before such transfer or aggregation is performed.

Learning with incomplete or heterogeneous modalities. Multimodal learning has also studied incomplete observations, missing modalities, out-of-modal generalization, and modality selection Wu et al. (2024; 2026); Huang et al. (2025b); He et al. (2024). These methods relax the assumption that every example shares the same modality set, but their decision unit is typically a modality, missing-modality pattern, or acquisition choice. In our setting, heterogeneity occurs at the source-domain level, and the decision unit is a complete source domain collected under a particular observation configuration.

Source selection, weighting, and data valuation. Another line of work evaluates which training examples, datasets, sources, or pretrained knowledge are useful for a target. Representative approaches include target-aware data selection Xie et al. (2023); Engstrom et al. (2024) and source-importance or transferability estimation Han et al. (2023); Diniz et al. (2026). These methods establish that source relevance should be target dependent, but their outputs are often independent scores, rankings, weights, or single-source choices. Target-specific source configuration selection instead seeks a joint subset of complete source domains whose utility may depend on the other selected sources.

Overall, prior work studies domain heterogeneity, incomplete observation structures, and source relevance from complementary perspectives. We focus on their intersection through a distinct decision object, namely which complete heterogeneous source domains should jointly participate for a specific target.

3 PROBLEM FORMULATION

We consider a heterogeneous multi-source learning setting with M candidate source domains and a target domain D^t. The source pool is $\mathcal{D}^s = \{D_1^s, \ldots, D_M^s\}$, where each D_i^s is an independently collected dataset for the same downstream prediction task as the target. Source domains may differ in population, acquisition condition, feature distribution, sensing setup, or modality availability. The target domain is partitioned into disjoint training, validation, and test splits, denoted by D_{train}^t, D_{val}^t, and D_{test}^t, respectively.

Decision object. The object of interest is a *source-domain configuration*. A configuration is a non-empty subset $S \subseteq \mathcal{D}^s$. The candidate space is $\mathcal{C} = 2^{\mathcal{D}^s} \setminus \{\emptyset\}$, where $2^{\mathcal{D}^s}$ denotes the power set of the source pool. The decision unit is therefore a complete source domain, and the output is a joint subset of such domains. It is not an individual sample, subject, modality, feature group, or continuous source weight.

The target-only setting is excluded from $\mathcal{C}$ because it does not answer which source domains should participate. It may be reported as an external diagnostic baseline, but it is never a candidate in source-configuration selection.

Target-specific configuration selection. Let L denote a fixed downstream learner and its training procedure. For each configuration $S \in \mathcal{C}$, L is trained or adapted using the selected source domains together with the allowed target training evidence. The resulting model is evaluated on the target validation split, giving

$$V(S) = F_{\text{val}}\left(L, S, D_{\text{train}}^t, D_{\text{val}}^t\right), \tag{1}$$

where F_{val} denotes the validation evaluation metric. In our experiments, the primary utility is Macro-F1.

A source-configuration selector π chooses

$$\hat{S} = \pi\left(\mathcal{C}, D_{\text{train}}^t, D_{\text{val}}^t, L\right), \qquad \hat{S} \in \mathcal{C}. \tag{2}$$

The selector may use only the permitted target training and validation evidence. The target test split is inaccessible during configuration selection and is reserved exclusively for final evaluation. This reflects the intended setting in which source participation must be decided from limited target evidence.

Oracle reference and regret. After selecting $\hat{S}$, we evaluate the corresponding learner on the held-out target test split. The target-test utility of configuration S is

$$U(S; D^t, L) = F_{\text{test}}\left(L, S, D_{\text{train}}^t, D_{\text{test}}^t\right), \tag{3}$$

where F_{test} denotes the same evaluation metric computed on the target test split.

For analysis, the oracle configuration within the evaluated candidate space is

$$S^* \in \arg\max_{S \in \mathcal{C}} U(S; D^t, L). \tag{4}$$

The oracle is used only as an evaluation reference and is never available to the selector. We summarize configuration-selection error by

$$\text{Regret}(\hat{S}) = U(S^*; D^t, L) - U(\hat{S}; D^t, L). \tag{5}$$

Selection goal. The goal is to select a configuration whose target-test utility approaches that of the oracle reference using only the allowed target evidence. The test utility $U(S; D^t, L)$ defines the desired outcome but is never observable during selection.

4 EMPIRICAL STRUCTURE OF SOURCE CONFIGURATION SELECTION

Before introducing ARCS, we examine how source configurations behave under limited target evidence. The purpose is not to compare final selectors, but to identify empirical properties that a selector should respect. We use HAR4-Fixed, where the total source-data budget is controlled

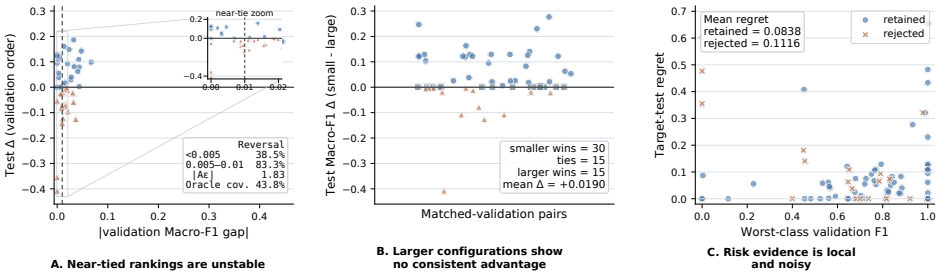


Figure 2: Empirical structure of source-configuration selection on HAR4-Fixed. Near-tied validation rankings are unstable, larger configurations show no consistent advantage, and class-level risk provides only a local signal.

across configurations so that performance differences primarily reflect source composition rather than source quantity. Figure 2 summarizes three observations. Validation is useful for identifying competitive configurations but unstable for fine-grained ranking, larger configurations do not consistently outperform smaller ones under comparable validation support, and class-level reliability provides only a noisy local signal. These findings directly motivate admissibility, parsimony, and localized risk screening.

Observation 1. Validation supports screening, but not reliable fine-grained ranking. Target validation is the only direct evidence about the new target available during selection. The key question is therefore how precisely its induced ordering should be trusted. Figure 2A compares validation gaps with signed target-test differences, where negative values indicate that the validation ordering reverses on the test split. When the top-two validation gap is below 0.005, the test ordering reverses in 38.5% of cases, and the reversal rate reaches 83.3% for gaps between 0.005 and 0.01. Using $\epsilon = 0.01$, the average admissible-set size is 1.83, and the test-oracle configuration is contained in the admissible set in 43.8% of scenarios. Validation therefore remains useful for narrowing the candidate space, but small differences should not be treated as reliable evidence for exact ranking. This motivates replacing single-best validation selection with a near-optimal admissible set.

Observation 2. Larger configurations show no consistent advantage. Source cardinality is not a monotonic proxy for target utility. Figure 2B compares smaller and larger configurations with matched validation support. Among these pairs, the smaller configuration wins in 30 cases, ties in 15, and loses in 15, with an average smaller-minus-larger target-test Macro-F1 of $+0.0190$. Thus, once validation evidence is comparable, larger configurations do not exhibit a consistent target-test advantage. Additional sources may introduce useful information, but they may also add mismatch or redundancy. This motivates parsimony as a tie-breaking principle within the validation-competitive region rather than a global preference for fewer sources.

Observation 3. Class-level risk is local and noisy. Aggregate validation utility can hide poor behavior on individual classes. We therefore examine WORSTCLASS, the minimum class-wise validation F1, for candidates inside the admissible region. Figure 2C relates WORSTCLASS to target-test regret and distinguishes candidates retained or rejected by a median-based risk screen. Retained candidates have lower mean regret than rejected ones, 0.0838 versus 0.1116, indicating that weakest-class behavior contains some useful local information. However, the two groups overlap substantially, and their mean target-test Macro-F1 values are nearly identical, 0.7517 and 0.7534. Class-level risk is therefore too noisy to serve as a primary ranking criterion, but it can act as a localized safeguard among otherwise competitive configurations.

Design implications. These observations suggest a specific selection structure. A selector should first preserve configurations that are difficult to distinguish from the validation best, then resolve near-ties without assuming that more sources are preferable, and use class-level risk only as secondary evidence. ARCS implements this structure through validation admissibility, parsimonious configuration selection, and localized risk screening.

5 METHOD

5.1 OVERVIEW

ARCS selects a target-specific source configuration from the candidate space $\mathcal{C}$ defined in Section 3. Its design follows directly from the empirical structure identified in Section 4. Near-tied validation rankings are unstable, larger configurations show no consistent advantage when validation support is comparable, and class-level risk provides only a local and noisy signal.

These observations assign distinct roles to the three components of ARCS. Validation first defines a region of configurations that cannot be reliably distinguished at fine resolution. Within this region, class-level information is used only to screen locally unfavorable candidates. The remaining ambiguity is then resolved through parsimony, so additional source participation must be supported by a meaningful validation advantage rather than rewarded by default. Operationally, ARCS therefore follows validation admissibility, localized risk screening, and parsimonious final selection.

5.2 VALIDATION ADMISSIBILITY

Let $V(S)$ denote the target-validation utility of configuration $S \in \mathcal{C}$, as defined in Section 3. Selecting only the configuration with the largest observed $V(S)$ treats every validation difference as sufficient evidence for exact ordering. Observation 1 shows that this assumption becomes unreliable for near-tied configurations.

Let $V^* = \max_{S \in \mathcal{C}} V(S)$. ARCS replaces exact validation maximization with the admissible set

$$\mathcal{A}_\epsilon = \{S \in \mathcal{C} \mid V^* - V(S) \leq \epsilon\}. \tag{6}$$

We use $\epsilon = 0.01$ in all main experiments and evaluate its sensitivity in Appendix F. The set $\mathcal{A}_\epsilon$ does not assume that its members are equally useful. Instead, it contains configurations for which the available validation evidence does not justify a finer ordering at resolution ϵ. Validation is therefore used to define a competitive decision region rather than a single point estimate.

5.3 LOCALIZED RISK SCREENING AND PARSIMONIOUS SELECTION

Observation 3 indicates that class-level validation behavior contains useful but noisy information within the competitive region. Let $F_k^{\mathrm{val}}(S)$ denote the validation F1 score of target class k for configuration S. We define the worst-class statistic

$R(S) = \min_k F_k^{\mathrm{val}}(S)$,

where larger $R(S)$ indicates stronger worst-class validation performance. Within $\mathcal{A}_\epsilon$, let $\widetilde{R} = \mathrm{median}_{S \in \mathcal{A}_\epsilon} R(S)$. ARCS retains

$$\mathcal{A}_\epsilon^R = \left\{ S \in \mathcal{A}_\epsilon \;\middle|\; R(S) \geq \widetilde{R} \right\}. \tag{7}$$

This risk screen is intentionally local. It neither ranks the complete configuration space nor replaces validation utility. Its role is only to remove comparatively weak worst-class behavior after validation has already restricted the decision to competitive configurations.

Observation 2 further shows that larger configurations do not provide a consistent advantage once validation support is comparable. ARCS therefore resolves the remaining ambiguity by lexicographically minimizing

$$\hat{S} = \arg\min_{S \in \mathcal{A}_\epsilon^R} \left(|S|, -V(S), \mathrm{ord}(S) \right), \tag{8}$$

where $|S|$ is the number of participating source domains and $\mathrm{ord}(S)$ denotes the fixed lexicographic ascending order of the source-composition string. ARCS first selects the smallest retained configuration, then prefers higher validation utility among configurations of equal size, and uses $\mathrm{ord}(S)$ only to resolve an exact remaining tie.

The parsimony preference is conditional rather than global. ARCS does not assume that fewer sources are intrinsically better. A larger configuration is selected whenever its validation advantage is sufficient to exclude smaller alternatives from the admissible region. Parsimony is invoked only when the available target evidence does not clearly justify the additional participating sources.

All ARCS decisions use only training and validation information. Target-test data never enter admissibility, risk screening, or final configuration selection. The rule is also independent of how candidate configurations are generated, allowing exhaustive enumeration to be replaced by approximate search when the source pool becomes larger.

6 EXPERIMENT

6.1 EVALUATION PROTOCOL AND BENCHMARKS

We evaluate source-configuration selection under the formulation in Section 3. For each target domain D^t, the eligible non-target domains form $\mathcal{D}^s$, and each selector chooses one non-empty configuration $S \in \mathcal{C}$. Within every learner-target-seed unit, all selectors share the same target split and candidate configuration pool. TargetOnly is excluded from $\mathcal{C}$ and is reported only as an external diagnostic baseline. Target training, validation, and test data are disjoint, and the test split is reserved exclusively for final evaluation. All main experiments use three seeds, 2026–2028.

We evaluate three heterogeneous multi-source constructions together with a natural-source HAR4 variant used for motivation. HAR4 combines UCI-HAR, PAMAP2, MHEALTH, and WISDM Anguita et al. (2013); Reiss & Stricker (2012); Banos et al. (2014); Kwapisz et al. (2011). Each target has three candidate sources and seven non-empty configurations. HAR4-Natural preserves the available source sizes, whereas HAR4-Fixed controls the total source budget at 900 windows per configuration and serves as our primary controlled setting. PPMI Marek et al. (2011) provides a multi-site clinical benchmark with six candidate source sites and 63 configurations per target. MOSEI Bagher Zadeh et al. (2018) provides a multimodal sentiment benchmark constructed from four deterministic video groups, with three candidate sources and seven configurations per target. PPMI and MOSEI retain the available source sizes rather than imposing the HAR4-Fixed budget.

Each learner-target-seed combination defines one matched evaluation unit. We report target-test Macro-F1, test regret as defined in Section 3, and the average number of selected source domains. Detailed source construction, target splits, representations, and learner settings are provided in Appendix A.

6.2 MAIN RESULTS

We compare ARCS with two fixed participation strategies, AllSource and BestSingleton, and five literature-adapted selectors, DATE, A2XP, SEA/SUTE, PANDA, and PreSTA. All methods use the same candidate configurations and target evidence within each matched learner-target-seed unit. Tables 2–4 report learner-wise target-test Macro-F1 together with overall Macro-F1, test regret, and the average number of selected sources.

On HAR4-Fixed, ARCS achieves the highest overall Macro-F1 of 0.7465 while selecting 1.43 sources on average. Among the displayed methods, ARCS is strongest for BiATEN, ARITH, GGA, and MBCD, whereas PANDA is strongest for LR. BestSingleton is also competitive at 0.7347 overall, indicating that strong performance is not tied to uniformly using either more or fewer sources.

On PPMI, ARCS again achieves the highest overall Macro-F1 of 0.6772 and the lowest average regret of 0.0950 among the displayed methods, while selecting 2.57 of six available sources. The learner-wise ranking remains heterogeneous, with AllSource strongest for LR and A2XP for MBCD. Against DATE, the strongest competing selector in overall Macro-F1, ARCS improves by 0.0175 on average, with a paired-bootstrap 95% confidence interval of $[-0.0021, 0.0375]$; we therefore do not claim statistically significant superiority.

MOSEI provides a complementary case. AllSource obtains the highest overall Macro-F1 of 0.6325, while ARCS reaches the second-best overall result of 0.6254 using 1.85 sources on average. The learner-wise optimum varies across methods, showing that target-specific configuration selection does not imply that a subset must always be preferred.

Table 2: Main source-configuration selection results on HAR4-Fixed. Results include learner-wise and overall target-test Macro-F1, regret, and average selected source count.

Method	LR	BiATEN AAAI'24	ARITH CVPR'25	GGA CVPR'25	MBCD AAAI'26	Overall	Regret ↓	#Src.
AllSource	0.7435	0.7178	0.6333	0.7392	0.7291	0.7126	0.0622	3.00
BestSingleton	0.7683	0.7438	0.6766	0.7484	0.7363	0.7347	0.0401	1.00
DATE (AAAI'23)	0.7058	0.7272	0.6171	0.7336	0.7464	0.7060	0.0687	2.00
A2XP (CVPR'24)	0.7354	0.7376	0.6758	0.7422	0.7323	0.7247	0.0501	1.65
SEA/SUTE (TPAMI'25)	0.7337	0.7452	0.6616	0.7425	0.7260	0.7218	0.0529	1.98
PANDA (AAAI'26)	**0.7715**	0.7435	0.6059	0.7378	0.7294	0.7176	0.0571	2.00
PreSTA (CVPR'26)	0.7440	0.7417	0.6702	0.7390	0.7250	0.7240	0.0508	1.65
ARCS (Ours)	0.7657	**0.7533**	**0.7011**	**0.7597**	**0.7528**	**0.7465**	**0.0282**	1.43

Table 3: Main source-configuration selection results on PPMI across five underlying learners and six candidate source domains. Results include learner-wise and overall target-test Macro-F1, regret, and average selected source count.

Method	LR	BiATEN AAAI'24	ARITH CVPR'25	GGA CVPR'25	MBCD AAAI'26	Overall	Regret ↓	#Src.
AllSource	**0.7053**	0.6353	0.5789	0.6566	0.6959	0.6544	0.1178	6.00
BestSingleton	0.6632	0.6531	0.5305	0.6938	0.6488	0.6379	0.1343	1.00
DATE (AAAI'23)	0.7032	0.6754	0.5688	0.6520	0.6989	0.6597	0.1125	2.97
A2XP (CVPR'24)	0.6657	0.6802	0.5539	0.6592	**0.6995**	0.6517	0.1205	2.75
SEA/SUTE (TPAMI'25)	0.6851	0.6581	0.5452	0.6782	0.6623	0.6458	0.1264	2.00
PANDA (AAAI'26)	0.6989	0.6722	0.5422	0.6747	0.6669	0.6510	0.1212	2.00
PreSTA (CVPR'26)	0.6123	0.5781	0.4518	0.6366	0.5861	0.5730	0.1992	2.63
ARCS (Ours)	0.6946	**0.6974**	**0.6239**	**0.6956**	0.6743	**0.6772**	**0.0950**	2.57

Overall, the preferred participation pattern varies across benchmarks and learners. ARCS performs strongest in aggregate on HAR4-Fixed and PPMI, while MOSEI favors full participation. This variation supports the central problem formulation: source participation should be decided for the target rather than fixed a priori.

6.3 Limited Target Evidence

We examine whether ARCS remains effective when less target evidence is available for configuration selection. On HAR4-Fixed, we retain 25%, 50%, 75%, or 100% of the target evidence and repeat each setting with five resampling seeds while keeping the target test split fixed.

Table 5 shows that ARCS degrades only mildly as target evidence is reduced. Macro-F1 decreases from 0.7465 with full evidence to 0.7429 at 25%, while the average selected source count remains between 1.43 and 1.47. ARCS is not uniformly best at every evidence level. ValidationBest is higher by 0.0005 at 25%, whereas ARCS is numerically best at 50%, 75%, and 100%; the margin at 50% is negligible.

Direct validation selection also becomes less stable as evidence decreases. Its selection stability drops from 1.000 at full evidence to 0.827 at 25%. This behavior is consistent with the motivation for admissibility: reduced target evidence weakens fine-grained validation ordering without causing a corresponding collapse in ARCS.

6.4 Component Analysis

We next examine how the ARCS decision rule emerges from its individual design choices. Table 6 progresses from direct validation maximization to validation admissibility, admissibility with parsimony, and the complete ARCS rule.

ValidationBest and Admissibility Only both obtain 0.7422 Macro-F1, showing that constructing a near-optimal set alone does not improve the final decision. Introducing parsimony raises Macro-F1

Table 4: Main source-configuration selection results on the constructed MOSEI benchmark across five underlying learners. Results include learner-wise and overall target-test Macro-F1, regret, and average selected source count.

Method	LR	BiATEN AAAI'24	ARITH CVPR'25	GGA CVPR'25	MBCD AAAI'26	Overall	Regret ↓	#Src.
AllSource	0.6430	**0.6335**	0.6030	**0.6442**	**0.6388**	**0.6325**	**0.0262**	3.00
BestSingleton	0.6231	0.6120	**0.6091**	0.6328	0.5941	0.6142	0.0445	1.00
DATE (AAAI'23)	0.6451	0.6204	0.5703	0.6302	0.6161	0.6164	0.0423	2.00
A2XP (CVPR'24)	0.6324	0.6261	0.5832	0.6272	0.6182	0.6174	0.0413	1.57
SEA/SUTE (TPAMI'25)	**0.6496**	0.6248	0.5838	0.6328	0.6216	0.6225	0.0362	2.00
PANDA (AAAI'26)	0.6365	0.6208	0.5800	0.6407	0.6297	0.6216	0.0371	2.00
PreSTA (CVPR'26)	0.6065	0.6006	0.5802	0.6315	0.6111	0.6060	0.0527	1.32
ARCS (Ours)	0.6436	0.6210	0.6026	0.6343	0.6257	0.6254	0.0333	1.85

Table 5: ARCS under reduced target evidence on HAR4-Fixed, averaged over five resampling seeds.

Evidence	Macro-F1 ↑	Regret ↓	#Src.
25%	0.7429	0.0318	1.45
50%	0.7454	0.0293	1.47
75%	0.7448	0.0299	1.45
100%	0.7465	0.0282	1.43

Table 6: Progressive component analysis of ARCS on HAR4-Fixed.

Selection Rule	Macro-F1 ↑	Regret ↓	#Src.
ValidationBest	0.7422	0.0325	1.52
Admissibility Only	0.7422	0.0325	1.55
Admissibility + Parsimony	0.7448	0.0299	1.40
ARCS	**0.7465**	**0.0282**	1.43

to 0.7448 and reduces the average selected source count from 1.55 to 1.40. The complete ARCS rule further reaches 0.7465.

These results identify admissibility plus parsimony as the main structural mechanism. Admissibility defines where fine-grained validation ordering should no longer be trusted, while parsimony resolves the resulting ambiguity. The WORSTCLASS gate provides a more localized refinement, changing the admissibility-plus-parsimony selection in only 3 of 60 evaluation units. Two interventions improve target-test performance and one is unfavorable. This profile is consistent with its intended role as a local safeguard rather than the primary source of gain.

7 CONCLUSION AND FUTURE WORK

We study target-specific source configuration selection in heterogeneous multi-source learning. Our empirical analysis shows that source participation should not be treated as fixed: validation rankings can be unstable, and larger source sets are not consistently better. These observations motivate ARCS, which combines validation admissibility, localized risk screening, and parsimony. Across heterogeneous wearable, clinical, and multimodal settings, ARCS provides competitive performance while selecting target-dependent source configurations. Component analysis identifies admissibility plus parsimony as the main mechanism, with risk screening acting only locally.

The current formulation operates at the domain level and is evaluated on moderate source pools. Future work should study scalable configuration search for larger pools and hierarchical selection that combines domain-level participation with finer-grained within-domain adaptation.

REFERENCES

Sk Miraj Ahmed, Dripta S. Raychaudhuri, Sujoy Paul, Samet Oymak, and Amit K. Roy-Chowdhury. Unsupervised multi-source domain adaptation without access to source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10103–10112, 2021.

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pp. 437–442, 2013.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1208.

Oresti Banos, Rafael Garcia, Juan A. Holgado-Terriza, Miguel Damas, Hector Pomares, Ignacio Rojas, Alejandro Saez, and Claudia Villalonga. mhealthdroid: A novel framework for agile development of mobile health applications. In *Proceedings of the International Work-Conference on Ambient Assisted Living*, pp. 91–98. Springer, 2014.

Raphaella Diniz, Jackson de Faria, and Martin Ester. Pas: Estimating the target accuracy before domain adaptation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Logan Engstrom, Axel Feldmann, and Aleksander Madry. Dsdm: Model-aware dataset selection with datamodels. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12491–12526, 2024.

Zhongyi Han, Zhiyan Zhang, Fan Wang, Rundong He, Wan Su, Xiaoming Xi, and Yilong Yin. Discriminability and transferability estimation: A bayesian source importance estimation approach for multi-source-free domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7811–7820, 2023. doi: 10.1609/aaai.v37i6.25946.

Yifei He, Runxiang Cheng, Gargi Balasubramaniam, Yao-Hung Hubert Tsai, and Han Zhao. Efficient modality selection in multimodal learning. *Journal of Machine Learning Research*, 25(47): 1–39, 2024.

Hai Huang, Yan Xia, Sashuai Zhou, Hanting Wang, Shulei Wang, and Zhou Zhao. Bridging domain generalization to multimodal domain generalization via unified representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22488–22498, 2025a.

Zhuo Huang, Gang Niu, Bo Han, Masashi Sugiyama, and Tongliang Liu. Towards out-of-modal generalization without instance-level modal correspondence. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Jennifer R. Kwapisz, Gary M. Weiss, and Samuel A. Moore. Activity recognition using cell phone accelerometers. *ACM SIGKDD Explorations Newsletter*, 12(2):74–82, 2011. doi: 10.1145/1964897.1964918.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In *Advances in Neural Information Processing Systems*, volume 21, 2009.

Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Christopher Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, Werner Poewe, Brit Mollenhauer, Todd Sherer, Mark Frasier, Claire Meunier, Alice Rudolph, Cindy Casaceli, John Seibyl, Susan Mendick, Norbert Schuff, Ying Zhang, Arthur Toga, Karen Crawford, Alison Ansbach, Pauline De Blasio, Massimo Piovella, John Trojanowski, Leslie Shaw, Andrew Singleton, Keith Hawkins, Jamie Eberling, David Brooks, David Russell, Laura Leary, Stewart Factor, Beth Sommerfeld, Penelope Hogarth, Elizabeth Pighetti, Kelly Williams, David Standaert, Sharon Guthrie, Robert Hauser, Helen Delgado, Joseph Jankovic, Christine Hunter, Matthew Stern, Beverly Tran,

James Leverenz, Melissa Baca, Steven Frank, Catherine-Ann Thomas, Irene Richard, Cathy Deeley, Linda Rees, Falk Sprenger, Erich Lang, Holly Shill, Svetlana Obradov, Hubert Fernandez, Amanda Winters, Daniela Berg, Kristina Gauss, Douglas Galasko, Daniel Fontaine, Zoltan Mari, Mark Gerstenhaber, David Brooks, Sophie Malloy, Paolo Barone, Katia Longo, Thomas Comery, Bernard Ravina, Igor Grachev, Kelly Gallagher, Michael Collins, Kieran Widnell, Stefan Ostrowizki, Paulo Fontoura, Tony Ho, John Luthman, Marcel van der Brug, Alastair Reith, and Paul Taylor. The parkinson progression marker initiative (ppmi). *Progress in Neurobiology*, 95(4): 629–635, 2011. doi: 10.1016/j.pneurobio.2011.09.005.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18, 2013.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1406–1415, 2019.

Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *Proceedings of the 16th International Symposium on Wearable Computers*, pp. 108–109, 2012. doi: 10.1109/ISWC.2012.13.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *Transactions on Machine Learning Research*, 2026.

Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Michael Nalls, Jimeng Sun, and Faraz Faghri. Multimodal patient representation learning with missing modalities and labels. In *The Twelfth International Conference on Learning Representations*, 2024.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1482.

Sicheng Zhao, Bo Li, Colorado Reed, Pengfei Xu, and Kurt Keutzer. Multi-source domain adaptation in the deep learning era: A systematic survey. *arXiv preprint arXiv:2002.12169*, 2020.

Sicheng Zhao, Hui Chen, Hu Huang, Pengfei Xu, and Guiguang Ding. More is better: Deep domain adaptation with multiple sources. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 8354–8362, 2024. doi: 10.24963/ijcai.2024/923.

Sicheng Zhao, Jing Jiang, Wenbo Tang, Jiankun Zhu, Hui Chen, Pengfei Xu, Björn W. Schuller, Jianhua Tao, Hongxun Yao, and Guiguang Ding. Multi-source multi-modal domain adaptation. *Information Fusion*, 117:102862, 2025. doi: 10.1016/j.inffus.2024.102862.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2022.