Valid and Efficient Uncertainty Quantification for Federated Joint Shift

Yuanjie Shi[1] **Peihong Li**[1] **Xuanyu Cao**[1] **Yan Yan**[1]

[1]School of EECS, Washington State University, Pullman, WA, USA

Abstract

Reliable uncertainty quantification (UQ) is critical for safety-sensitive federated learning (FL) applications, such as cross-hospital diagnosis and global sensor networks. In FL, privacy constraints prevent centralized data pooling, while client data exhibit joint distribution shifts across covariates, labels, and conditionals. These shifts violate the exchangeability assumption required by conformal prediction (CP), which otherwise guarantees distribution-free coverage under i.i.d. data. To address this gap, we propose Federated Conformal Prediction for Joint Shift (FCPJS), enabling valid CP under heterogeneous clients without sharing raw data. Specifically, the method operates in two stages: (i) the server constructs a privacy-preserving global sketch of nonconformity score distributions; (ii) each client performs importance-weighted calibration by contrasting its local scores with the global sketch. This distribution-ratio weighting corrects joint shifts in a unified manner. We prove that FCPJS attains valid marginal coverage across clients, up to an $O(\sqrt{\varepsilon_m} + 1/\sqrt{n})$ error from the finite number of calibration samples n and sketch size m. Experiments on four heterogeneous benchmarks show that FCPJS preserves coverage and improves predictive efficiency by 12.64% over the strongest baseline. To our knowledge, FCPJS is the first method providing provably valid and efficient conformal UQ for FL under joint distribution shift.

1 INTRODUCTION

Modern federated learning (FL) systems McMahan et al. [2017], Kairouz et al. [2021] power cross-hospital diagnostic models, planet-scale sensor networks, and privacy-preserving mobile assistants. In safety-critical applications, where prediction errors can lead to catastrophic consequences, it is essential not only to make accurate predictions but also to quantify uncertainty. Uncertainty quantification (UQ) Lu et al. [2023] is thus a fundamental requirement in practical deployment. However, most existing UQ techniques rely on a centrally aggregated calibration set, which is an assumption that directly conflicts with the privacy and communication constraints inherent to FL.

Conformal prediction (CP) Angelopoulos and Bates [2021], Vovk et al. [2005], Shafer and Vovk [2008] is a model-agnostic UQ framework that provides finite-sample coverage guarantees. Given a miscoverage level α, CP constructs prediction sets that contain the true label with probability at least $1-\alpha$. However, these guarantees rely on the exchangeability between calibration and test data Barber et al. [2023], Tibshirani et al. [2019]. In federated learning (FL), clients collect data from heterogeneous environments, inducing substantial cross-client distribution shifts Plassier et al. [2023, 2024]. Such heterogeneity violates exchangeability and renders standard CP unreliable, motivating distribution-aware conformal methods for FL.

To address the challenge of data heterogeneity in federated settings, several recent studies have proposed extensions of CP. One line of work applies CP locally on each client using its own calibration data, without any global coordination Humbert et al. [2023]. While this yields valid coverage per client, it sacrifices generalization across the system and fails to provide guarantees for new or low-data clients. Another line of work aims to estimate global quantiles centrally while preserving privacy. For instance, Lu et al. [2023] assume partial exchangeability, in which calibration and test samples are exchangeable with some unknown probability—a strong and unverifiable assumption in practice. More recently, Plassier et al. [2023, 2024] propose to correct for label shift and covariate shift, respectively, using importance-weighted quantiles. These methods estimate density ratios based on class priors or deep-feature similarity, and adjust the calibration score distribution accordingly.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:6232–6260.

However, these methods are fundamentally limited to correcting only one marginal component and fail when multiple types of shift occur simultaneously. In real-world FL deployments, clients often differ in both covariates and label marginals, and may even exhibit class-conditional differences. This setting, known as *joint distribution shift*, poses a severe challenge for existing federated CP methods. Our empirical findings (and Proposition 1) show that these prior importance-weighting schemes misestimate the test score distribution and fail to guarantee valid coverage under joint shift. This raises the key question: *How can we design a CP framework that is valid and efficient under joint distribution shift in FL?*

In this paper, we propose the **Federated Conformal Prediction for Joint Shift (FCPJS)** method, which produces valid and size-efficient prediction sets under full joint distribution shift. The key idea of FCPJS is to align each client's local calibration-score distribution with the global mixture distribution induced across clients. Concretely, each client uploads a compact sketch (e.g., a t-digest Dunning and Ertl [2019]) of its empirical score distribution, rather than raw calibration data or individual nonconformity scores. The server aggregates these sketches to obtain a global score distribution estimate and broadcasts it back to clients. For each calibration score, we construct importance weights as the density ratio between the aggregated global score distribution and the client's local score distribution. This weighting corrects discrepancies caused by joint shift and enables weighted quantile estimation across clients. We prove that FCPJS achieves $(1-\alpha)$ marginal coverage with high probability, with a finite-sample deviation bounded by $O(\sqrt{\varepsilon_m}+1/\sqrt{n})$, where n is the total number of calibration samples and m is the sketch size. Extensive experiments demonstrate that FCPJS consistently preserves coverage while reducing average prediction set sizes by 12.64%. To our knowledge, this is the first federated CP framework providing provable validity and efficiency under joint distribution shift.

Contributions. The key contributions of this paper include:

- We propose FCPJS, a communication-efficient and privacy-aware federated conformal framework that avoids sharing raw data and aligns client-level score distributions via sketch-based aggregation and density-ratio importance weighting.
- We prove that existing marginal reweighting methods fail under joint distribution shift, and establish that FCPJS achieves $(1-\alpha)$ marginal coverage with high probability, with a deviation bounded by $O\left(1/\sqrt{n}+\sqrt{\varepsilon_m}\right)$.
- We conduct extensive experiments on benchmark datasets of distribution shift, demonstrating that FCPJS achieves target coverage across all settings while improving predictive efficiency by 12.64% on average over the best baseline. The code is available at `https://github.com/YuanjieSh/FCPJS`

2 RELATED WORK

Federated Learning (FL) is a decentralized paradigm where multiple clients collaboratively train a global model without sharing raw data McMahan et al. [2017], Kairouz et al. [2021]. This framework suits privacy-sensitive domains such as commerce Jain and Jerripothula [2023], finance Long et al. [2020], Mammen [2021], and healthcare Antunes et al. [2022]. A central challenge in FL is *statistical heterogeneity* Li et al. [2020], Karimireddy et al. [2020], as data distributions often differ across clients due to user behavior or acquisition conditions Kairouz et al. [2021]. Privacy constraints further limit information exchange Zhang et al. [2023], making centralized calibration or validation infeasible. These challenges complicate uncertainty quantification, which typically assumes centralized i.i.d. data Lu et al. [2023], and motivate extending distribution-free tools like conformal prediction to the federated setting.

Conformal Prediction (CP). CP is a powerful framework for uncertainty quantification Angelopoulos and Bates [2021], Vovk et al. [2005], but it traditionally assumes data exchangeability Angelopoulos et al. [2024a], Huang et al. [2023]. This assumption breaks under covariate shift, temporal drift, or structured heterogeneity. To address covariate shift, several works use importance weighting to adjust calibration scores via estimated density ratios Tibshirani et al. [2019], Barber et al. [2023], Jin and Candès [2023]. For time-series and online settings, CP has been adapted using dynamic sets or martingale techniques to maintain valid coverage Zaffran et al. [2023], Angelopoulos et al. [2024b], Gibbs and Candès [2024]. CP has also been extended to handle label shift Podkopaev and Ramdas [2021], domain adaptation Lei and Candès [2021], and group-based heterogeneity Hore and Barber [2024].

CP in FL. Applying CP in FL introduces additional complexity: the inability to share data across heterogeneous clients breaks global exchangeability, complicating standard CP calibration strategies. Early work by Lu et al. [2023] proposed averaging locally computed quantiles across clients to construct prediction sets, but their method lacks formal coverage guarantees and performs poorly with small local datasets. Subsequent efforts have addressed these limitations by leveraging alternative aggregation strategies. For instance, Humbert et al. [2023] introduced a communication-efficient method that computes local quantiles on each client and then uses a quantile-of-quantiles aggregation at the central server, improving both coverage validity and robustness. Min et al. studies a client-level federated setting, where coverage is guaranteed for a specific target client's distribution. However, Humbert et al. [2023] leaves open the problem of achieving exact coverage guarantees. Other methods, such

as those by Plassier et al. [2023], explore CP under label shift in FL but often require multiple communication rounds. Ensemble-based approaches have also been proposed Gauraha and Spjuth [2021], but they typically assume access to a shared calibration set, which is impractical in real FL settings. Plassier et al. [2024] reweight calibration nonconformity scores using analytically tractable importance weights based on estimated density ratios between the target and calibration distributions, and employ a federated optimization algorithm with a regularized pinball loss to estimate the global quantile. However, their method assumes shared conditional label distributions across clients in order to estimate density ratios, which is often unrealistic in practice.

3 PROBLEM SETUP

Notations. Let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ be a data sample, where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = \{1, ..., C\}$ represent the feature and classification label, respectively. Let $\mathcal{P}_{XY}$ be the underlying distribution defined on $\mathcal{X} \times \mathcal{Y}$. Similarly, we define $\mathcal{P}_X$ and $\mathcal{P}_Y$ as the distributions of feature X and label Y, respectively. Define $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$ as the calibration set of n samples. K denotes the number of clients, and the set of all client indices is $[K]$. Let $\mathcal{I}^k$ be the index set for calibration data on client k. n^k denotes the number of calibration data samples on k-th client, where $n = \sum_{k=1}^K n^k$. We define the joint distribution on client k by $\mathcal{P}_{XY}^k$. Similarly, we denote $\mathcal{P}_X^k$ and $\mathcal{P}_Y^k$ as the distributions of the feature and the label on client k, respectively. Let $\mathcal{I}_y^k$ be the index set for calibration data with label y on client k, n_y^k denotes the number of calibration data with label y on k-th client.

Conformal prediction. Conformal prediction (CP) calibrates predictions based on a *non-conformity* scoring function. Let $V : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a non-conformity scoring function that maps an input–label pair to a real value (lower scores indicate higher conformity). For simplicity, we denote the non-conformity score of the i-th example as $V_i = V(X_i, Y_i)$ for $(X_i, Y_i) \in \mathcal{D}_{\text{cal}}$. Many non-conformity scoring functions are proposed in prior work Angelopoulos and Bates [2021], Shafer and Vovk [2008]. The specific design for non-conformity score is introduced in Appendix A.

Given the mis-coverage rate $\alpha \in (0, 1)$ and a testing input $X_{\text{test}} \sim \mathcal{P}_{XY}$, we construct a prediction set $\widehat{\mathcal{C}}(X_{\text{test}})$ as:

$$\mathcal{C}(X_{\text{test}}) = \{y \in \mathcal{Y} : V(X_{\text{test}}, y) \leq \widehat{Q}(1-\alpha; \{V_i\}_{i=1}^n)\},$$

where $\widehat{Q}(1-\alpha; \{V_i\}_{i=1}^n)$ is the $\lceil (1-\alpha)(n+1) \rceil$-th smallest value in $\{V_i\}_{i=1}^n$. As calibration and test samples are exchangeable, this simple rule guarantees marginal coverage:

$$\mathbb{P}\{Y_{\text{test}} \in \widehat{\mathcal{C}}(X_{\text{test}})\} \geq 1 - \alpha.$$

However, this guarantee relies on exchangeability between calibration and test data, which is frequently violated under covariate shift or heterogeneous federated settings, resulting in either conservative or under-covering prediction sets.

To address this issue, Tibshirani et al. [2019] employs the importance weighting method. Given a likelihood ratio $\omega(X_i) = d\mathcal{P}_X(X_i)/d\mathcal{P}_{X_{\text{test}}}(X_i)$ for data (X_i, Y_i), it will first compute the weighted empirical distribution:

$$u = \sum_{i=1}^n p_i^\omega(X_{\text{test}})\delta_{V_i} + p_{n+1}^\omega(X_{\text{test}})\delta_{+\infty}, \tag{1}$$

where δ_a is a point mass at a and weights are computed as:

$$p_i^\omega(X_{\text{test}}) = \frac{\omega(X_i)}{\sum_{j=1}^n \omega(X_j) + \omega(X_{\text{test}})},$$

$$p_{n+1}^\omega(X_{\text{test}}) = \frac{\omega(X_{\text{test}})}{\sum_{j=1}^n \omega(X_j) + \omega(X_{\text{test}})}.$$

Then, the prediction set $\widehat{\mathcal{C}}(X_{\text{test}})$ is constructed as:

$$\widehat{\mathcal{C}}(X_{\text{test}}) = \{y \in \mathcal{Y} : V(X_{\text{test}}, y) \leq \widehat{Q}(1-\alpha; u)\}. \tag{2}$$

Federated CP. In FL, clients often encounter significantly different data-generating processes due to variations in user behavior, environment, or hardware. This leads to **distribution shifts** across clients, which can take several forms:

- **Covariate shift:** the feature distribution varies, i.e., $\mathcal{P}_X^{k_1} \neq \mathcal{P}_X^{k_2}$ for $k_1 \neq k_2$, while the conditional distribution $\mathcal{P}_{Y|X}^k$ remains the same.
- **Label shift:** the label marginal differs, i.e., $\mathcal{P}_Y^{k_1} \neq \mathcal{P}_Y^{k_2}$ for $k_1 \neq k_2$, but the class-conditional distribution $\mathcal{P}_{X|Y}^k$ is invariant.
- **Joint distribution shift:** the full joint distribution $\mathcal{P}_{XY}^k$ differs across clients, encompassing shifts in covariates, labels, and class-conditional relationships.

Following the standard FL convention Mohri et al. [2019], Plassier et al. [2023, 2024], the test distribution is the mixture of client distributions, i.e., $P_{XY} = \sum_{k=1}^K \frac{n^k}{n} P_{XY}^k$.

These heterogeneities pose a serious challenge to CP. To adapt CP to FL settings, recent work Plassier et al. [2023, 2024] have extended CP to federated settings by correcting for specific types of shift through importance weighting.

Plassier et al. [2023] address label shift by estimating class-prior ratios across clients. Given label counts n_y^k on client k, they compute class-wise weights:

$$\omega_y^k = \frac{n \cdot n_y^k}{n^k \cdot \sum_{k \in [K]} n_y^k} \cdot \mathbb{1}\left[\sum_{k \in [K]} n_y^k \geq 1\right],$$

which are used to construct a label-specific weighted distribution u_y over calibration scores, and prediction sets are formed using the corresponding class-wise quantiles.

Plassier et al. [2024] focuses on covariate shift. Each client extracts feature embeddings $\phi(X)$ and fits Gaussian Mixture Models (GMMs) for each class. The sample-wise weight is then given by the ratio of the class-conditional likelihoods under client-local vs. global GMMs:

$$\omega_i = \frac{\sum_{y\in\mathcal{Y}} \frac{n_y^{k'}}{n^{k'}} \mathcal{N}(\phi(X_i); \mu_y^{k'}, \Sigma_y^{k'})}{\sum_{k\in[K]} \sum_{y\in\mathcal{Y}} \frac{n_y^k}{n^k} \mathcal{N}(\phi(X_i); \mu_y^k, \Sigma_y^k)},$$

where $\mathcal{N}$ denotes the Gaussian density, $i \in \mathcal{I}^{k'}$ and (μ_y^k, Σ_y^k) are class-conditional GMM parameters estimated from client k.

4 FEDERATED CP UNDER JOINT SHIFT

In this section, we first demonstrate the failure of marginal reweighting under joint shift, then introduce *Federated Conformal Prediction for Joint Shift (FCPJS)*, and finally establish its coverage guarantees.

4.1 MOTIVATION

While prior approaches (summarized in Section 3) extend CP to federated settings, they remain fundamentally limited, as each addresses only a single type of shift. In practice, FL often exhibits *joint distribution shift*, with simultaneous changes in $\mathcal{P}_X$, $\mathcal{P}_Y$, $\mathcal{P}_{X|Y}$, and $\mathcal{P}_{Y|X}$. The following proposition shows that these reweighting schemes fail to ensure valid coverage under such conditions.

Proposition 1 (Failure of Marginal Reweighting under Joint Shift)**.** *Assume the joint distribution exhibits both covariate shift and label shift, i.e., for all clients k,*

$$\mathcal{P}_{XY}^k \neq \mathcal{P}_{XY}, \quad \textit{and} \quad \mathcal{P}_X^k \neq \mathcal{P}_X, \quad \mathcal{P}_Y^k \neq \mathcal{P}_Y.$$

Let $d_{KS}(\mathcal{P}, u) := \sup_v |F_P(v) - F_u(v)|$ as Kolmogorov distance and $\beta(\alpha) := |F_u(v^*) - (1-\alpha)|$, where $v^* \in \arg\max_v |F_P(v) - F_u(v)|$. Then, the coverage of Plassier et al. [2023, 2024] deviates from the target $1-\alpha$ by:

$$\left|\mathbb{P}\left(Y_{test} \in \widehat{\mathcal{C}}(X_{test})\right) - (1-\alpha)\right| \geq \Omega(d_{KS}(\mathcal{P}, u)) - C_{reg}\beta(\alpha),$$

where C_{reg} is constant. Under mild regularity conditions (Appendix B.1), this deviation admits a lower bound in terms of the total variation distance between the induced weighted score distribution and the test-time score distribution.

Remark 1. *The quantity $d_{KS}(\mathcal{P}, u)$ captures the CDF mismatch between the weighted calibration scores and the true scores. Under single-shift settings (e.g., pure label shift or pure covariate shift), reweighting schemes that do not explicitly align the induced score distributions may remove certain distributional discrepancies (e.g., label or covariate shift). However, it generally fails to align the induced score distributions under joint shift, resulting in $d_{KS}(\mathcal{P}, u)$ remaining non-negligible. Proposition 1 further shows that when the target quantile level $1-\alpha$ lies near the region of maximal CDF discrepancy, the deviation becomes on the order of $d_{KS}(\mathcal{P}, u)$, leading to substantial under- or over-coverage.*

Algorithm 1 FCPJS: Federated Conformal Prediction under Joint Shift

Input: Calibration scores $\{V_i\}_{i\in\mathcal{I}^k}$ at each client $k \in [K]$, miscoverage level α
for each client k in parallel **do**
  Compute empirical CDF $\widehat{F}^k$
  Upload sketch $\widetilde{F}^k = h(\widehat{F}^k)$
end for
Server aggregates global CDF: $\widehat{F}(v) = \sum_{k=1}^K \frac{n_k}{n} \tilde{F}_k(v)$
Broadcast $\widehat{F}$ to clients
Compute weight $\widehat{\omega}_i = \dfrac{d\widehat{F}(V_i)}{d\widehat{F}^k(V_i)}$
Compute weighted $(1-\alpha)$-quantile:

$$\widehat{Q}(1-\alpha) = \inf\left\{ q : \sum_i \widehat{\omega}_i \mathbf{1}\{V_i \leq q\} \geq (1-\alpha) \sum_i \widehat{\omega}_i \right\}$$

Output: $\widehat{\mathcal{C}}_\alpha(x) = \{y : V(x,y) \leq \widehat{Q}(1-\alpha)\}$

Other federated CP methods remain limited. Lu et al. [2023] assume *partial exchangeability*, that client data are exchangeable with some unknown probability, a strong, unverifiable assumption under privacy constraints. Humbert et al. [2023] empirically study heterogeneity effects on coverage but provide no validity guarantees under distribution shift.

In summary, existing methods either correct only one component of distribution shift or rely on strong assumptions. Our goal is to develop a federated CP framework that achieves provable coverage under full joint distribution shift:

$$\mathbb{P}\{Y_{\text{test}} \in \widehat{\mathcal{C}}(X_{\text{test}})\} \geq 1-\alpha.$$

4.2 METHOD

Barber et al. [2023] proposes importance-weighted CP to restore marginal coverage beyond exchangeability. However, it requires access to all calibration scores, leaves the density-ratio weights unspecified, and assumes a uniform global shift. These assumptions break in FL, where raw scores cannot be shared and shifts are client-specific and joint.

To bridge this gap, we propose FCPJS to provide a coverage guarantee in FL. The core insight of FCPJS is to view federated CP as a score-space distribution alignment problem: each client reweights its local calibration scores toward a global target score distribution, defined as a mixture over clients. Concretely, we summarize score distributions using sketched empirical CDFs. By comparing the local and

global score CDFs, we construct sketch-based importance weights that approximately compensate for cross-client distributional discrepancies in the score space. This calibration enables reliable conformal quantile estimation under heterogeneous clients, with bounded approximation errors. We summarize the overall procedure in Algorithm 1.

Specifically, let $\{V_i\}_{i\in\mathcal{I}^k}$ be the calibration scores and $\widehat{F}^k(t)$ be their empirical Cumulative Distribution Function (CDF) for client k. Instead of sending raw scores, each client compresses $\widehat{F}^k$ with a quantile sketch function h (e.g. a t-digest) and uploads the sketch:

$$\widetilde{F}^k(v) = h(\widehat{F}^k(v)). \tag{3}$$

The sketch requires only $O(\log m)$ communication to approximate $\widehat{F}^k$ at m break-points, resulting in sub-kilobyte cost even for millions of calibration scores. By transmitting only coarse quantile summaries instead of raw scores, the sketch also limits per-example information exposure. Such summary-based mechanisms are standard in federated systems Bonawitz et al. [2019], Lu et al. [2023]. Our method is a general framework: if stronger guarantees are required, calibrated differential privacy (DP) noise can be injected into the sketch before transmission without modifying the subsequent alignment or quantile steps. A DP-enhanced variant is provided in Appendix C.3.

Next, the server forms an approximation of the global score CDF $\widehat{F}(v)$ by aggregating the K uploaded sketches, as sketch coordinates are additive:

$$\widehat{F}(v) = \sum_{k=1}^{K} \frac{n_k}{n} \tilde{F}_k(v). \tag{4}$$

After that, client k aligns its local score with $\widehat{F}$ via importance weights:

$$\widehat{\omega}_i^k = \frac{d\widehat{F}(V_i)}{d\widehat{F}^k(V_i)}, \quad i \in \mathcal{I}^k. \tag{5}$$

This ratio serves as a density correction in the score space, adjusting local distributions toward the global one. Since marginal coverage depends only on the distribution of $V(X_{\text{test}}, Y_{\text{test}})$, this score-space correction is sufficient to restore valid coverage under any form of joint distribution shift.

In implementation, Eq. (5) is operationalized through a finite-difference density estimator. For a bandwidth $h > 0$, we estimate the global and client-local score densities as

$$\tilde{f}(v) = \frac{\tilde{F}(v+h) - \tilde{F}(v-h)}{2h},$$
$$\tilde{f}_k(v) = \frac{\tilde{F}_k(v+h) - \tilde{F}_k(v-h)}{2h}.$$

The score-space density-ratio weight is then computed as

$$\hat{\omega}_i^k = \frac{\tilde{f}(V_i)}{\tilde{f}_k(V_i)}.$$

For finite-sample stability, we apply a small density floor $\lambda > 0$ to avoid division by nearly empty sketch bins. We also use a tempered log-density ratio

$$\hat{\omega}_i^k = \exp\left(t\left[\log(\tilde{f}(V_i) + \lambda) - \log(\tilde{f}_k(V_i) + \lambda)\right]\right),$$

where $t \in (0, 1]$ controls the strength of reweighting. The theoretical analysis below is stated for the untempered case $t = 1$, while the tempered version is used as a practical variance-control stabilizer in experiments. In all non-private experiments, we use $t = 0.001$ as the default temperature and select the finite-difference bandwidth according to $h \asymp \sqrt{\epsilon_m}$, following the bias–approximation tradeoff in Lemma 2. Sensitivity analyses for the temperature and density-estimation resolution are reported in Appendix C.3.

Finally, the weighted quantile is computed from the client-local calibration scores and their stabilized weights. This step can be implemented by aggregating weighted score summaries rather than transmitting individual calibration scores. Specifically, we estimate the $(1-\alpha)$-quantile $\widehat{Q}(1-\alpha)$ as the smallest threshold satisfying:

$$\sum_{i=1}^{n} \widehat{\omega}_i \cdot \mathbf{1}\{V_i \le \widehat{Q}(1-\alpha)\} \;\ge\; (1-\alpha) \cdot \sum_{i=1}^{n} \widehat{\omega}_i.$$

This corresponds to the quantile under a weighted distribution over the calibration scores. Then the prediction set is constructed as:

$$\widehat{\mathcal{C}}_\alpha(X) = \left\{y \in \mathcal{Y} : V(x, y) \le \widehat{Q}(1-\alpha)\right\}.$$

4.3 THEORETICAL ANALYSIS

Theorem 1 (Coverage Guarantee under Joint Shift)**.** *Assume that each F^k admits a density f^k on $\mathcal{V}$ and satisfies $\inf_{v\in\mathcal{V}} f^k(v) \ge c_0 > 0$ and the importance weights have finite second moment, i.e., $\mathbb{E}[\omega(V)^2] < \infty$. Let $\varepsilon_m := \max\left\{\|\widetilde{F} - F\|_\infty,\ \max_{k\in[K]} \|\widetilde{F}^k - F^k\|_\infty\right\}$ denote the CDF approximation error induced by the chosen sketching mechanism with sketch size m. Then, the following inequality holds with probability at least $1-\delta$:*

$$\left|\mathbb{P}\left(Y_{\text{test}} \in \widehat{\mathcal{C}}(X_{\text{test}})\right) - (1-\alpha)\right| \le C_1\sqrt{\frac{\log(1/\delta)}{n}} + C_2\sqrt{\varepsilon_m},$$

where $C_1, C_2 > 0$ are constants.

Remark 2. *Theorem 1 shows that our method achieves $(1-\alpha)$ marginal coverage up to two vanishing error terms:*

Partition	Methods	HPS		APS		RAPS	
		MCOV	MAPSS	MCOV	MAPSS	MCOV	MAPSS
RxRx1 ($\downarrow$ 11.07% on average, $\downarrow$ 2.8% IID, $\downarrow$ 19.34% non-IID)							
IID	FCP	0.64 ± 0.00	14.76 ± 0.032	$\mathbf{0.90 \pm 0.01}$	153.76 ± 11.18	$\mathbf{0.90 \pm 0.00}$	160.67 ± 3.58
	FCPLS	0.87 ± 0.00	156.55 ± 2.91	0.87 ± 0.00	167.65 ± 2.89	0.87 ± 0.00	190.66 ± 3.90
	FCPCS	$\mathbf{0.90 \pm 0.01}$	135.34 ± 5.53	$\mathbf{0.90 \pm 0.00}$	147.88 ± 2.23	$\mathbf{0.90 \pm 0.01}$	161.17 ± 3.72
	FCPJS	$\mathbf{0.91 \pm 0.01}$	$\mathbf{133.02 \pm 3.39}$ ($\downarrow$ 1.71%)	$\mathbf{0.90 \pm 0.00}$	$\mathbf{142.10 \pm 2.02}$ ($\downarrow$ 3.91%)	$\mathbf{0.90 \pm 0.00}$	$\mathbf{156.21 \pm 3.33}$ ($\downarrow$ 2.78%)
Non-IID	FCP	0.64 ± 0.01	14.81 ± 0.10	$\mathbf{0.91 \pm 0.00}$	163.05 ± 11.77	$\mathbf{0.91 \pm 0.00}$	170.61 ± 4.41
	FCPLS	0.70 ± 0.02	250.94 ± 19.14	0.70 ± 0.02	258.26 ± 19.20	0.70 ± 0.02	260.76 ± 18.81
	FCPJS	$\mathbf{0.90 \pm 0.00}$	$\mathbf{142.79 \pm 5.10}$ ($\downarrow$ 43.10%)	$\mathbf{0.90 \pm 0.00}$	$\mathbf{157.31 \pm 7.13}$ ($\downarrow$ 3.52%)	$\mathbf{0.90 \pm 0.00}$	$\mathbf{151.15 \pm 3.38}$ ($\downarrow$ 11.41%)
FMoW ($\downarrow$ 10.00% on average, $\downarrow$ 4.41% IID, $\downarrow$ 15.60% non-IID)							
IID	FCP	0.88 ± 0.00	6.49 ± 0.03	$\mathbf{0.92 \pm 0.01}$	11.30 ± 0.91	$\mathbf{0.91 \pm 0.01}$	8.73 ± 0.45
	FCPLS	$\mathbf{0.90 \pm 0.01}$	9.55 ± 0.43	$\mathbf{0.90 \pm 0.01}$	11.42 ± 0.41	$\mathbf{0.90 \pm 0.00}$	10.61 ± 0.48
	FCPJS	$\mathbf{0.92 \pm 0.00}$	$\mathbf{9.43 \pm 0.89}$ ($\downarrow$ 1.26%)	$\mathbf{0.91 \pm 0.01}$	$\mathbf{10.35 \pm 0.23}$ ($\downarrow$ 8.41%)	$\mathbf{0.90 \pm 0.01}$	$\mathbf{8.42 \pm 0.31}$ ($\downarrow$ 3.55%)
Non-IID	FCP	0.88 ± 0.00	6.50 ± 0.03	$\mathbf{0.91 \pm 0.00}$	11.32 ± 0.90	$\mathbf{0.91 \pm 0.00}$	8.89 ± 0.02
	FCPLS	$\mathbf{0.90 \pm 0.00}$	9.24 ± 0.32	$\mathbf{0.89 \pm 0.00}$	11.07 ± 0.33	$\mathbf{0.90 \pm 0.00}$	10.48 ± 0.54
	FCPJS	$\mathbf{0.90 \pm 0.00}$	$\mathbf{7.70 \pm 0.13}$ ($\downarrow$ 31.98%)	$\mathbf{0.90 \pm 0.00}$	$\mathbf{10.24 \pm 0.33}$ ($\downarrow$ 7.50%)	$\mathbf{0.90 \pm 0.00}$	$\mathbf{8.24 \pm 0.09}$ ($\downarrow$ 7.31%)
iWildCam ($\downarrow$ 26.70% on average, $\downarrow$ 26.42% IID, $\downarrow$ 26.98% non-IID)							
IID	FCP	0.80 ± 0.00	9.93 ± 0.03	$\mathbf{0.90 \pm 0.00}$	35.92 ± 0.15	$\mathbf{0.90 \pm 0.00}$	26.53 ± 0.04
	FCPLS	$\mathbf{0.90 \pm 0.00}$	108.05 ± 1.04	$\mathbf{0.90 \pm 0.01}$	115.56 ± 1.14	$\mathbf{0.90 \pm 0.00}$	108.59 ± 1.12
	FCPJS	$\mathbf{0.90 \pm 0.00}$	$\mathbf{23.48 \pm 0.69}$ ($\downarrow$ 78.27%)	$\mathbf{0.90 \pm 0.00}$	$\mathbf{35.85 \pm 0.27}$ ($\downarrow$ 0.19%)	$\mathbf{0.90 \pm 0.00}$	$\mathbf{26.32 \pm 0.28}$ ($\downarrow$ 0.79%)
Non-IID	FCP	0.80 ± 0.00	9.96 ± 0.05	$\mathbf{0.90 \pm 0.00}$	35.94 ± 0.20	$\mathbf{0.90 \pm 0.00}$	26.54 ± 0.05
	FCPLS	$\mathbf{0.89 \pm 0.00}$	107.73 ± 1.65	$\mathbf{0.90 \pm 0.00}$	115.35 ± 1.46	$\mathbf{0.90 \pm 0.00}$	108.24 ± 1.69
	FCPJS	$\mathbf{0.91 \pm 0.00}$	$\mathbf{24.68 \pm 0.39}$ ($\downarrow$ 77.09%)	$\mathbf{0.90 \pm 0.00}$	$\mathbf{35.78 \pm 0.43}$ ($\downarrow$ 0.45%)	$\mathbf{0.90 \pm 0.00}$	$\mathbf{25.64 \pm 0.24}$ ($\downarrow$ 3.39%)
Amazon Reviews ($\downarrow$ 2.78% on average, $\uparrow$ 4.91% IID, $\downarrow$ 10.48% non-IID)							
IID	FCP	0.53 ± 0.00	3.34 ± 0.00	$\mathbf{0.91 \pm 0.00}$	4.73 ± 0.00	$\mathbf{0.89 \pm 0.00}$	4.67 ± 0.01
	FCPLS	$\mathbf{0.90 \pm 0.00}$	$\mathbf{4.48 \pm 0.01}$	$\mathbf{0.90 \pm 0.01}$	$\mathbf{4.49 \pm 0.02}$	$\mathbf{0.90 \pm 0.00}$	$\mathbf{4.49 \pm 0.02}$
	FCPJS	$\mathbf{0.90 \pm 0.00}$	4.80 ± 0.00 ($\uparrow$ 7.14%)	$\mathbf{0.89 \pm 0.00}$	4.66 ± 0.00 ($\uparrow$ 3.79%)	$\mathbf{0.89 \pm 0.00}$	4.66 ± 0.00 ($\uparrow$ 3.79%)
Non-IID	FCP	$\mathbf{0.94 \pm 0.00}$	3.40 ± 0.05	$\mathbf{0.90 \pm 0.00}$	2.68 ± 0.07	$\mathbf{0.90 \pm 0.00}$	2.69 ± 0.06
	FCPLS	$\mathbf{0.90 \pm 0.00}$	4.52 ± 0.09	$\mathbf{0.90 \pm 0.00}$	4.37 ± 0.07	$\mathbf{0.90 \pm 0.00}$	4.37 ± 0.07
	FCPJS	$\mathbf{0.90 \pm 0.00}$	$\mathbf{2.47 \pm 0.02}$ ($\downarrow$ 27.35%)	$\mathbf{0.90 \pm 0.00}$	$\mathbf{2.63 \pm 0.05}$ ($\downarrow$ 1.86%)	$\mathbf{0.90 \pm 0.00}$	$\mathbf{2.63 \pm 0.06}$ ($\downarrow$ 2.23%)

Table 1: **Experiments on RxRx1, FMoW, iWildCam and Amazon Reviews.** We report the mean and standard deviation of all methods with three score functions and miscoverage rate $\alpha = 0.1$. As FCPCS incurs a 50 to $100\times$ computational cost compared to FCPJS(as shown in Figure 2), we report its results only on the RxRx1 dataset with IID partition, and provide an additional iWildCam non-IID comparison in Appendix C.3. The MCov results are not smaller than 0.9 with 0.01 tolerance, and the minimal MAPSS results while achieving valid coverage are in **bold**. The MCOV results show that FCPJS could consistently achieve a valid coverage guarantee under different score functions and on all datasets. The MAPSS results show that FCPJS consistently outperforms the best baseline with 12.64% reduction in terms of average prediction set size over four datasets.

a statistical error $O(1/\sqrt{n})$ arising from importance weighting, and a sketch approximation error $O(\sqrt{\varepsilon_m})$ induced by quantile compression. Both terms decrease as the calibration sample size n increases or the sketch approximation improves. For sketches satisfying a uniform CDF guarantee $\varepsilon_m = O(1/m)$ Chen et al. [2021], the second term reduces to $O(1/\sqrt{m})$. Importantly, the guarantee holds under heterogeneous client distributions, provided the target score distribution is well-defined and the density ratio exists. The condition $\inf_{v\in\mathcal{V}} f_k(v) \geq c_0 > 0$ is the score-space analogue of the standard overlap/positivity assumption in importance-weighted conformal prediction. It ensures that the global-to-local score-density ratio is well-defined and prevents the weighted quantile from being dominated by near-zero-density regions. When c_0 is very small, the constants in Theorem 1 may be large, so the bound should be interpreted as a rate-level validity guarantee under score-space overlap rather than a tight finite-sample certificate under severe support mismatch.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Baselines. We select federated CP methods that address nonexchangeable data in the FL setting: (i) FCP Lu et al. [2023], it assumes the partial exchangeability of global distribution, where data samples from different clients are exchangeable with probability; (ii) FCPLS Plassier et al. [2023], it investigates the impact of label shift in FL for CP; (iii) FCPCS Plassier et al. [2024], it investigates the impact of covariate shift in FL for CP; As FCPCS incurs

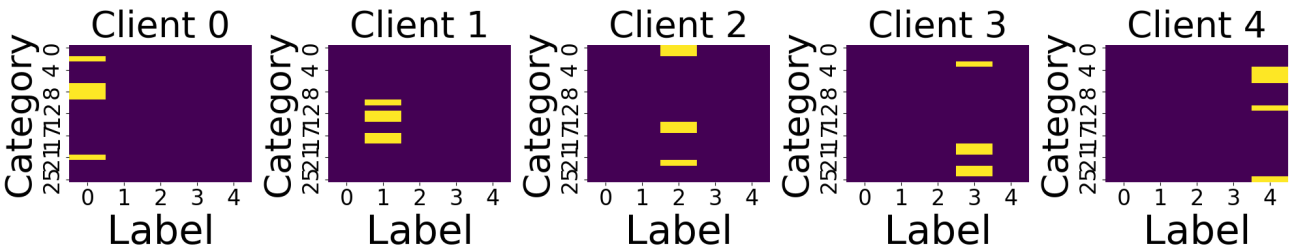

Figure 1: **Joint distribution heatmaps of groups (categories, y-axis) and labels (x-axis) across all clients under the non-IID partitioning** of the Amazon Reviews dataset. Each client is assigned disjoint product categories and disjoint label subsets, resulting in highly sparse and heterogeneous data distributions that reflect joint distribution shift.

Dataset	RxRx1	FMoW	iWildCam	Amazon Reviews
Number of calibration samples	17 212	11 048	21 391	23 475
Number of test samples	17 220	11 060	21 400	23 475
Number of classes	1 139	62	182	5
Number of clients	50	20	20	5
Type of shift	covariate	covariate	joint	joint
Partition groups	experiment	Region × year	location	category

Table 2: **Description of RxRx1, FMoW, iWildCam and Amazon Reviews datasets.** We highlight that while we use the same Dirichlet parameter (Dir $= 1.0$) across datasets, the resulting label imbalance varies due to differences in the number of classes and how labels are split among clients.

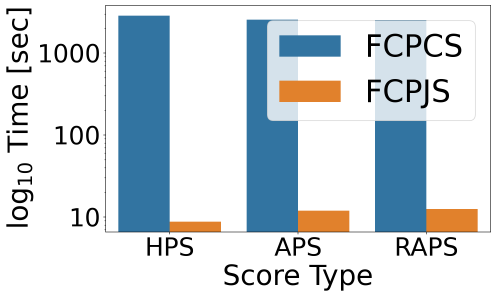

Figure 2: **Comparison of computational time (in log-scale seconds) between FCPCS and FCPJS** across three scores: HPS, APS, and RAPS. FCPCS incurs significantly higher runtime (exceeding 50–100×) due to centralized calibration and non-parallelizable optimization, while FCPJS is substantially more efficient.

a 50–100× computational cost compared to FCPJS, we report its results only on the RxRx1 dataset with IID partition. The comparison of computational cost on FCPCS and FCPJS is shown in Figure 2. It is clear that FCPCS requires 50–100× more computation time compared to FCPJS. We also include a non-private centralized oracle in Appendix C.3 to contextualize the remaining efficiency gap to pooled calibration.

Datasets and Deep models. To comprehensively evaluate the performance of all methods under different types of distribution shifts, we select four datasets from the WILDS repository Koh et al. [2021]: RxRx1 Taylor et al. [2019] for covariate shift, FMoW Christie et al. [2018] for covariate shift, iWildCam Beery et al. [2020] and Amazon Reviews Ni et al. [2019] for joint distribution shift. Following the WILDS benchmark settings, we use ResNet50 He et al. [2016] as the backbone model for RxRx1 and iWildCam, DenseNet121 Huang et al. [2017] for FMoW, and DistilBERT Sanh et al. [2019] for Amazon Reviews.

For each dataset, we construct IID and non-IID partitions to simulate homogeneous and heterogeneous distributions. In the IID case, data are uniformly split across clients without regard to groups or labels. In the non-IID case, we first partition data by disjoint groups (e.g., experiments in RxRx1, region×year in FMoW, locations in iWildCam, and category in Amazon Reviews) and disjoint label subsets to induce strong heterogeneity. Following Ma et al. [2025], Gao et al. [2022], we further apply a Dirichlet distribution with Dir $= 1.0$ to introduce within-client label imbalance. Table 2 summarizes dataset statistics, and we detail the shift sources and non-IID construction for each dataset.

RxRx1 contains fluorescent microscopy cell images from 51 experiments and 1139 genetic treatment labels. Because experimental protocols vary while labels do not, it naturally exhibits covariate shift. To simulate heterogeneity across clients, we partition the dataset into 50 clients by assigning disjoint experiment IDs. We then assign disjoint subsets of labels to each client and apply a Dirichlet distribution with Dir $= 1.0$ over the allowed labels to induce label imbalance within clients. We also report the joint distribution of groups (experiment IDs) and class labels across all 10 clients under IID partitioning strategy in Appendix C.

FMoW (Functional Map of the World) consists of satellite images labeled with one of 62 land-use or building categories. Each image is associated with a geographic region and year, introducing covariate variability due to differences in landscape, infrastructure, and imaging conditions over time. Since the label space remains consistent across regions and years, this results in covariate shift. We partition FMoW into 20 clients using disjoint region–year pairs, and assign

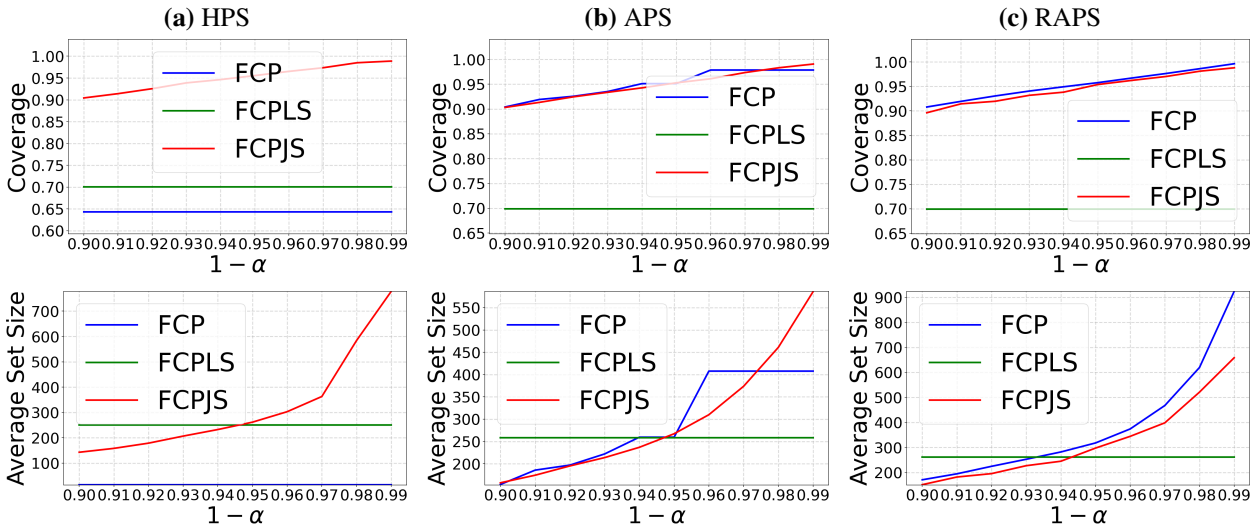

Figure 3: **Coverage (Top) and average set size (Bottom) results of all methods with three scores and under different confidence levels** on RxRx1 datasets, non-IID partition, where $1-\alpha \in \{0.90, \cdots, 0.99\}$ with range 0.01. It is clear that FCPJS consistently maintains valid coverage and smaller prediction sets across confidence levels with different scores.

disjoint label subsets to each client. Within each client, a Dirichlet distribution with Dir $= 1.0$ over the allowed labels controls label imbalance.

iWildCam contains camera trap images labeled with one of 182 animal species. Each sample is linked to a specific geographic location, which introduces substantial variability in both input distribution (due to differing environmental conditions) and label distribution (due to local species prevalence). Thus, iWildCam serves as a representative dataset for joint distribution shift. We simulate this shift by assigning disjoint locations to 20 clients and allocating disjoint subsets of labels to each. To induce label imbalance, we again apply a Dirichlet distribution with Dir $= 1.0$ over the allowed labels within each client.

Amazon Reviews consists of product review texts labeled with five sentiment categories. Each sample is associated with a product category, inducing variability in both the input distribution (domain-specific vocabulary and style) and the label distribution (category-dependent rating tendencies), thereby naturally exhibiting joint distribution shift. To simulate federated heterogeneity, we assign disjoint product categories to 5 clients, each representing a distinct domain. Within each client, all five sentiment labels remain but follow imbalanced proportions sampled from a Dirichlet distribution with Dir $= 1.0$, further amplifying the joint shift. Figure 1 visualizes the resulting category–label distributions: the sparse, localized heatmaps illustrate the heterogeneous joint distributions induced by our partitioning.

While we fix the Dirichlet concentration parameter (Dir $=$ 1.0) across datasets, the resulting label imbalance still differs due to variations in class numbers and label partitioning. Datasets with more classes (e.g., RxRx1) assign fewer labels to each client, leading to stronger skew under the same Dir. Thus, Dir $= 1.0$ ensures a consistent partitioning procedure while naturally producing dataset-specific imbalance levels.

Conformity scoring functions. We consider three nonconformity scoring functions: HPS Vovk et al. [2005], Lei et al. [2013] APS Romano et al. [2020], and RAPS Angelopoulos et al. [2021]. The detailed review of these scoring functions is in Appendix A.

Evaluation metrics. Our first evaluation metric is the marginal coverage (MCov), defined as $\text{MCov} = \frac{1}{|\mathcal{D}_{\text{test}}|}\sum_{i\in\mathcal{D}_{\text{test}}} \mathbb{1}[Y_i \in C_(X_i)]$. Our second evaluation metric is the marginal prediction set size (MAPSS), defined as $\text{MAPSS} = \frac{1}{|\mathcal{D}_{\text{test}}|}\sum_{i\in\mathcal{D}_{\text{test}}} |C_(X_i)|$. We repeat random calibration/test splits 10 times for FCP, FCPLS, and FCPJS, and 3 times for FCPCS.

5.2 RESULTS

FCPJS achieves the best trade-off between coverage validity and predictive efficiency. We evaluate the performance of all methods on four datasets under both IID and non-IID partitions. Each method is tested with three score functions, and we report both marginal coverage (MCOV) and average prediction set size (MAPSS) at the miscoverage rate of $\alpha = 0.1$. The results are summarized in Table 1. As shown, our proposed method FCPJS consistently achieves valid coverage, with MCOV values tightly concentrated around the target 0.90 across all datasets, score functions, and partition types. In contrast, baseline methods (e.g., FCP and FCPLS) often exhibit under- or over-coverage, especially under non-IID settings. Besides, FCPJS achieves the best MAPSS across all scenarios with an average 12.64%

reduction in terms of average prediction set size, with per-dataset average reductions of 11.07% on RxRx1, 10.00% on FMoW, 26.70% on iWildCam, and 2.78% on Amazon Reviews. The gains are largest in non-IID settings, where score-space alignment corrects heterogeneous client shifts. The main exception is Amazon Reviews under the IID partition, where FCPJS produces slightly larger sets than FCPLS. This setting is close to exchangeable and has only five classes, so score-space reweighting provides limited benefit while density-ratio estimation can introduce mild conservativeness. Under Amazon Reviews non-IID, FCPJS again reduces set size by 10.48%, consistent with the method being most useful under heterogeneous joint shift.

FCPJS consistently maintains valid coverage and smaller prediction sets across varying confidence levels. To further evaluate the robustness of each method, we examine their performance across a range of confidence levels $(1-\alpha) \in \{0.90, 0.91, \cdots, 0.99\}$ with range 0.01 on the RxRx1 dataset with non-IID partition. Figure 3 shows the results for three score functions, plotting marginal coverage (top row) and average prediction set size (bottom row) across increasing confidence levels $1-\alpha$. Across all score types, FCPJS consistently maintains valid coverage, which closely tracks the diagonal coverage targets, while baseline methods either under-cover (FCPLS) or show unstable performance (FCP). Notably, FCPJS achieves significantly smaller prediction sets than FCP as confidence increases. This trend reflects the method's ability to adaptively balance coverage and efficiency under stricter reliability requirements. These results demonstrate that FCPJS provides strong coverage guarantees across confidence levels while maintaining competitive predictive efficiency.

Increasing the sketch size m reduces the sketch approximation error and improves predictive efficiency. To evaluate the impact of the sketch dimension, we vary $m \in \{10, 64, 100, 256, 512\}$ on the RxRx1 dataset under the non-IID partition with APS score. Figure 4 reports the coverage gap $|\mathrm{MCOV}(m) - (1-\alpha)|$ (left) and the corresponding MAPSS (right). As m increases, the coverage gap consistently decreases, indicating that larger sketches provide a more accurate approximation of the global score distribution. This trend is consistent with our theoretical analysis, which shows that the coverage deviation contains a sketch approximation component that diminishes as m grows. More accurate quantile estimation further translates into smaller prediction sets while maintaining valid coverage. Overall, these results demonstrate that enlarging the sketch size systematically improves both coverage reliability and efficiency.

6 LIMITATIONS AND FUTURE WORK

Several limitations remain. First, the formal guarantee is marginal over the federated mixture and does not imply per-client conditional coverage. Client-wise coverage is therefore best viewed as a deployment diagnostic rather than the formal target. Second, when the setting is nearly IID or the label space is very small, score-space reweighting may provide limited benefit and can be mildly conservative, as observed on Amazon Reviews IID. Third, FCPJS relies on stable score-density ratio estimation; extremely sparse clients or severe support mismatch near the target quantile may increase weight variance and reduce efficiency. Future work could develop personalized or client-conditional extensions that combine global efficiency with stronger local validity.

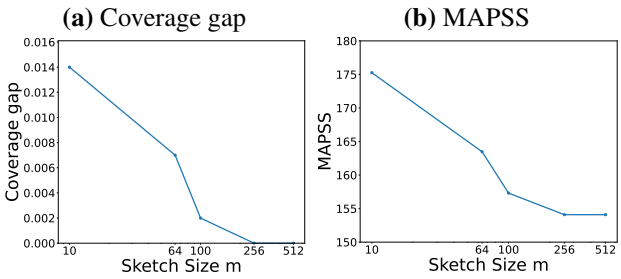

Figure 4: **Effect of sketch size m on coverage accuracy and efficiency** on RxRx1 datasets, non-IID partition with APS score. **Left:** Coverage gap ($|MCOV - (1-\alpha)|$) decreases as the sketch size increases, indicating that a larger sketch yields more accurate calibration, consistent with Theorem 1. **Right:** Corresponding MAPSS across different sketch sizes.

7 SUMMARY

We introduce a federated CP framework that achieves valid and efficient uncertainty quantification under joint distribution shift. While prior methods address either covariate or label shift using marginal importance weights, we show these are insufficient when both marginals and conditionals differ across clients—a common scenario in real-world federated learning. Our method estimates the global score distribution using importance-weighted calibration scores based on score-space density ratios, and aggregates information via compact quantile sketches. We provide theoretical guarantees on coverage, showing it holds up to a sketch error term of $O\left(\frac{1}{m}\right)$. Experiments on diverse benchmarks demonstrate that our method consistently maintains valid coverage under severe heterogeneity and improves predictive efficiency over existing federated CP approaches with an average 12.64% prediction set size reduction. These results suggest that score-space alignment is an effective abstraction for federated conformal calibration when clients differ in covariates, labels, and conditionals. Because FCPJS communicates only compact score summaries, it provides a practical path toward uncertainty quantification in federated systems where centralized calibration is infeasible.

ACKNOWLEDGMENTS

Yan Yan is supported by the USDA-NIFA funded AgAID Institute award 2021-67021-35344, and the NSF grant CNS-2312125, IIS-2443828, DUE-2519063. The views expressed are those of the authors and do not reflect the official policy or position of the USDA-NIFA and NSF.

References

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024a.

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.

Anastasios Nikolas Angelopoulos, Rina Barber, and Stephen Bates. Online conformal prediction with decaying step sizes. In *International Conference on Machine Learning*, pages 1616–1630. PMLR, 2024b.

Rodolfo Stoffel Antunes, Cristiano André da Costa, Arne Küderle, Imrana Abdullahi Yari, and Björn Eskofier. Federated learning for healthcare: Systematic review and architecture proposal. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4):1–23, 2022.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Sara Beery, Elijah Cole, and Arvi Gjoka. The iwildcam 2020 competition dataset. *arXiv preprint arXiv:2004.10340*, 2020.

Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečnỳ, Stefano Mazzocchi, Brendan McMahan, et al. Towards federated learning at scale: System design. *Proceedings of machine learning and systems*, 1:374–388, 2019.

Margarida Campos, António Farinhas, Chrysoula Zerva, Mário AT Figueiredo, and André FT Martins. Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 12:1497–1516, 2024.

Peiqing Chen, Yuhan Wu, Tong Yang, Junchen Jiang, and Zaoxing Liu. Precise error estimation for sketch-based flow measurement. In *Proceedings of the 21st ACM Internet Measurement Conference*, pages 113–121, 2021.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Ted Dunning and Otmar Ertl. Computing extremely accurate quantiles using t-digests. *arXiv preprint arXiv:1902.04023*, 2019.

Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. Feddc: Federated learning with non-iid data via local drift decoupling and correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10112–10121, 2022.

Niharika Gauraha and Ola Spjuth. Synergy conformal prediction. In *Conformal and Probabilistic Prediction and Applications*, pages 91–110. PMLR, 2021.

Isaac Gibbs and Emmanuel J Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Rohan Hore and Rina Foygel Barber. Conformal prediction with local weights: randomization enables robust guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae103, 2024.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Kexin Huang, Ying Jin, Emmanuel Candes, and Jure Leskovec. Uncertainty quantification over graph with conformalized graph neural networks. *Advances in Neural Information Processing Systems*, 36:26699–26721, 2023.

Pierre Humbert, Batiste Le Bars, Aurélien Bellet, and Sylvain Arlot. One-shot federated conformal prediction. In *International Conference on Machine Learning*, pages 14153–14177. PMLR, 2023.

Shreyansh Jain and Koteswar Rao Jerripothula. Federated learning for commercial image sources. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6534–6543, 2023.

Ying Jin and Emmanuel J Candès. Model-free selective inference under covariate shift via weighted conformal p-values. *arXiv preprint arXiv:2307.09291*, 2023.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021.

Jing Lei, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):911–938, 2021.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37 (3):50–60, 2020.

Guodong Long, Yue Tan, Jing Jiang, and Chengqi Zhang. Federated learning for open banking. In *Federated learning: privacy and incentive*, pages 240–254. Springer, 2020.

Charles Lu, Yaodong Yu, Sai Praneeth Karimireddy, Michael Jordan, and Ramesh Raskar. Federated conformal predictors for distributed uncertainty quantification. In *International Conference on Machine Learning*, pages 22942–22964. PMLR, 2023.

Zhiyong Ma, Yuanjie Shi, Yan Yan, and Jian Chen. Federated r\'enyi fair inference in federated heterogeneous system. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

Priyanka Mary Mammen. Federated learning: Opportunities and challenges. *arXiv preprint arXiv:2101.05428*, 2021.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

Yinjie Min, Chuchen Zhang, Liuhua Peng, and Changliang Zou. Personalized federated conformal prediction with localization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International conference on machine learning*, pages 4615–4625. PMLR, 2019.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

Art B. Owen. *Monte Carlo theory, methods and examples*. `https://artowen.su.domains/mc/`, 2013.

Vincent Plassier, Mehdi Makni, Aleksandr Rubashevskii, Eric Moulines, and Maxim Panov. Conformal prediction for federated uncertainty quantification under label shift. In *International Conference on Machine Learning*, pages 27907–27947. PMLR, 2023.

Vincent Plassier, Nikita Kotelevskii, Aleksandr Rubashevskii, Fedor Noskov, Maksim Velikanov, Alexander Fishkov, Samuel Horvath, Martin Takac, Eric Moulines, and Maxim Panov. Efficient conformal prediction under data heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pages 4879–4887. PMLR, 2024.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in Artificial Intelligence*, pages 844–853. PMLR, 2021.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114 (525):223–234, 2019.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Yam Tawachi and Bracha Laufer-Goldshtein. Multi-dimensional conformal prediction. In *The Thirteenth International Conference on Learning Representations*, 2025.

James Taylor, Berton Earnshaw, Ben Mabey, Mason Victors, and Jason Yosinski. Rxrx1: An image set for cellular morphological variation across many experimental batches. In *International Conference on Learning Representations (ICLR)*, volume 22, page 23, 2019.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Margaux Zaffran, Aymeric Dieuleveut, Julie Josse, and Yaniv Romano. Conformal prediction with missing values. In *International Conference on Machine Learning*, pages 40578–40604. PMLR, 2023.

Xinwen Zhang, Yihan Zhang, Tianbao Yang, Richard Souvenir, and Hongchang Gao. Federated compositional deep auc maximization. *arXiv preprint arXiv:2304.10101*, 2023.

A ADDITIONAL BACKGROUND DETAILS

Non-conformity scoring functions. The homogeneous prediction sets (HPS) Sadinle et al. [2019] scoring function is defined as follows:

$$S_f^{\text{HPS}}(X, Y) = 1 - f_\theta(X)_Y. \tag{6}$$

Romano et al. [2020] has proposed another conformity scoring function, Adaptive Prediction Sets (APS). APS scoring function is based on sorted probabilities. For a given input X, we sort the softmax probabilities for all classes $\{1, \cdots, K\}$ such that $1 \geq f_\theta(x)_{(1)} \geq \cdots \geq f_\theta(x)_{(K)} \geq 0$, and compute the cumulative confidence as follows:

$$S_f^{\text{APS}}(X, Y) = \sum_{l=1}^{r_f(X,Y)-1} f_\theta(X)_{(l)} + U \cdot f_\theta(X)_{(r_f(X,Y))}, \tag{7}$$

where $U \in [0, 1]$ is a random variable to break ties.

To reduce the probability of including unnecessary labels (i.e., labels with high ranks) and thus improve the predictive efficiency, Angelopoulos et al. [2020] has proposed Regularized Adaptive Prediction Sets (RAPS) scoring function. The RAPS score is computed as follows:

$$S_f^{\text{RAPS}}(X, Y) = \sum_{l=1}^{r_f(X,Y)-1} f_\theta(X)_{(l)} + U \cdot f_\theta(X)_{(r_f(X,Y))} + \lambda_{\text{RAPS}}(r_f(X, Y) - k_{\text{reg}})^+, \tag{8}$$

where λ_{RAPS} and k_{reg} are two hyper-parameters.

B TECHNICAL PROOFS

B.1 PROOF OF PROPOSITION 1

To quantify the coverage deviation described in Proposition 1, we introduce mild regularity conditions that allow translating discrepancies between score distributions into quantile and coverage differences. These assumptions are standard technical conditions on cumulative distribution functions and are used solely to derive an explicit lower bound; the qualitative miscalibration under joint shift does not rely on them.

Proposition (Re) 1 (Failure of Marginal Reweighting under Joint Shift)**.** *Define $F_{\mathcal{P}}(v) = \mathbb{P}_{(X,Y)\sim\mathcal{P}_{XY}}\big(V(X, Y) \leq v\big)$ and $F_u(v) = \mathbb{P}\big(u \leq v\big)$. Suppose that $F_{\mathcal{P}}(v)$ and $F_u(v)$ are $L_{\mathcal{P}}$-Lipschitz continuous and L_u-Lipschitz continuous, $F'_u(v) > c > 0$, and $1/2 - \frac{L_{\mathcal{P}}+L_u}{c} = O(1)$. We also suppose the calibration scores $\{V_i = V(X_i, Y_i)\}_{i=1}^n$ are distinct and drawn independently from a mixture of client distributions $\{\mathcal{P}_{XY}^k\}_{k=1}^K$. Assume the joint distribution exhibits both covariate shift and label shift, i.e., for all clients k,*

$$\mathcal{P}_{XY}^k \neq \mathcal{P}_{XY}, \quad \textit{and} \quad \mathcal{P}_X^k \neq \mathcal{P}_X, \quad \mathcal{P}_Y^k \neq \mathcal{P}_Y.$$

Let $d_{KS}(\mathcal{P}, u) := \sup_v |F_P(v) - F_u(v)|$ as Kolmogorov distance and $\beta(\alpha) := |F_u(v^*) - (1 - \alpha)|$, where $v^* \in \arg\max_v |F_P(v) - F_u(v)|$. Then, the coverage of Plassier et al. [2023, 2024] deviates from the target $1 - \alpha$ by:

$$\left|\mathbb{P}\left(Y_{test} \in \widehat{\mathcal{C}}(X_{test})\right) - (1 - \alpha)\right| \geq \Omega(d_{KS}(\mathcal{P}, u)) - \frac{L_P + L_u}{c}\,\beta(\alpha).$$

Proof. This proof consists of three steps: Step 1, we denote some notations to help the analysis; Step 2, we transfer the coverage into the outputs of $F_{\mathcal{P}_{XY}}(v)$ and $F_u(v)$; and Step 3, we prove the lower bound.

Step 1. Before proving proposition 1, we define the following notation.

Define $v^* \in \arg\max_v \big|F_{\mathcal{P}}(v) - F_u(v)\big|$ That is,

$$\big|F_{\mathcal{P}}(v^*) - F_u(v^*)\big| = \sup_v \big|F_{\mathcal{P}}(v) - F_u(v)\big| = d_{\text{KS}}(\mathcal{P}, u).$$

Denote $F_u^{-1}(v)$ as the inverse function of $F_u(v)$. Due to $F_u'(v) > c > 0$, $F_u^{-1}(v)$ is $1/c$-Lipschitz continuous.

Step 2. Recall that $X_{\text{test}} \sim \mathcal{P}_{XY}$ and the prediction sets of Plassier et al. [2023, 2024] are constructed by:

$$\widehat{\mathcal{C}}(X_{\text{test}}) = \{y \in \mathcal{Y} : V(X_{\text{test}}, y) \leq \widehat{Q}(1-\alpha; u)\}.$$

Thus, the coverage of of Plassier et al. [2023, 2024] could be rewritten as:

$$\begin{aligned}
&\mathbb{P}\left(Y_{\text{test}} \in \widehat{\mathcal{C}}(X_{\text{test}})\right) \\
=&\mathbb{P}\left(Y_{\text{test}} \in \{y \in \mathcal{Y} : V(X_{\text{test}}, y) \leq \widehat{Q}(1-\alpha; u)\}\right) \\
=&\mathbb{P}\left(V(X_{\text{test}}, Y_{\text{test}}) \leq \widehat{Q}(1-\alpha; u)\right) \\
=&F_{\mathcal{P}}\left(\widehat{Q}(1-\alpha; u)\right).
\end{aligned}$$

Similarly, given fixed calibration set $\mathcal{D}_{\text{cal}}$, we have $F_u\left(\widehat{Q}(1-\alpha; u)\right) = 1-\alpha$.

Step 3. Next, we begin to prove Proposition 1. Recall that $F_u(\widehat{Q}(1-\alpha; u)) = 1-\alpha$.

$$\begin{aligned}
&\left|\mathbb{P}\left(Y_{\text{test}} \in \widehat{\mathcal{C}}(X_{\text{test}})\right) - (1-\alpha)\right| \\
=& \left|F_{\mathcal{P}}\left(\widehat{Q}(1-\alpha; u)\right) - F_u\left(\widehat{Q}(1-\alpha; u)\right)\right| \\
=& \left|F_{\mathcal{P}}\left(\widehat{Q}\right) - F_u\left(\widehat{Q}\right)\right| \\
=& \left|\big(F_{\mathcal{P}}(v^*) - F_u(v^*)\big) - \big(F_{\mathcal{P}}(v^*) - F_{\mathcal{P}}(\widehat{Q})\big) + \big(F_u(v^*) - F_u(\widehat{Q})\big)\right| \\
\geq& \left|F_{\mathcal{P}}(v^*) - F_u(v^*)\right| - \left|F_{\mathcal{P}}(v^*) - F_{\mathcal{P}}(\widehat{Q})\right| - \left|F_u(v^*) - F_u(\widehat{Q})\right| \\
\geq& \left|F_{\mathcal{P}}(v^*) - F_u(v^*)\right| - (L_{\mathcal{P}} + L_u)\left|v^* - \widehat{Q}\right| \\
=& \left|F_{\mathcal{P}}(v^*) - F_u(v^*)\right| - (L_{\mathcal{P}} + L_u)\left|F_u^{-1}\big(F_u(v^*)\big) - F_u^{-1}\big(F_u(\widehat{Q})\big)\right| \\
\geq& \left|F_{\mathcal{P}}(v^*) - F_u(v^*)\right| - \frac{L_{\mathcal{P}} + L_u}{c}\left|F_u(v^*) - F_u(\widehat{Q})\right| \\
=& \left|F_{\mathcal{P}}(v^*) - F_u(v^*)\right| - \frac{L_{\mathcal{P}} + L_u}{c}\left|F_u(v^*) - (1-\alpha)\right| \\
=& \, d_{\text{KS}}(\mathcal{P}, u) \; - \; \frac{L_{\mathcal{P}} + L_u}{c}\,\beta(\alpha),
\end{aligned}$$

where the first inequality follows from the triangle inequality, the second inequality follows from the Lipschitz continuity of $F_{\mathcal{P}}$ and F_u with constants $L_{\mathcal{P}}$ and L_u, and the third inequality follows from the fact that $F_u'(v) \geq c > 0$, which implies that F_u^{-1} is $1/c$-Lipschitz continuous.

Rearranging the above inequality, we have:

$$\left|\mathbb{P}\left(Y_{\text{test}} \in \widehat{\mathcal{C}}(X_{\text{test}})\right) - (1-\alpha)\right| \geq \Omega(d_{\text{KS}}\big(\mathcal{P}, u\big)) - \frac{L_P + L_u}{c}\,\beta(\alpha).$$

□

B.2 PROOF OF THEOREM 1

Theorem (Re) 1 (Coverage Guarantee under Joint Shift)**.** *Assume that Each F^k admits a density f^k on $\mathcal{V}$ and satisfies $\inf_{v\in\mathcal{V}} f^k(v) \geq c_0 > 0$ and the importance weights have finite second moment,i.e., $\mathbb{E}[\omega(V)^2] < \infty$. Let $\varepsilon_m := \max\Big\{\|\widetilde{F} -$*

$F\|_\infty,\ \max_{k\in[K]} \|\widetilde{F}^k - F^k\|_\infty \Big\}$ *denote the CDF approximation error induced by the chosen sketching mechanism with sketch size* m*. Then, the following inequality holds with probability at least* $1-\delta$*:*

$$\left|\mathbb{P}\Big(Y_{\text{test}} \in \widehat{\mathcal{C}}(X_{\text{test}})\Big) - (1-\alpha)\right| \le C_1\sqrt{\frac{\log(1/\delta)}{n}} + C_2\,\sqrt{\varepsilon_m},$$

where $C_1, C_2 > 0$ *are constants.*

Proof. Before proving Theorem 1, we present several technical lemmas. Recall that F denotes the true CDF of the test nonconformity scores, and F^k denotes the client-k score CDF. Let $\widetilde{F}$ and $\widetilde{F}^k$ be the sketched CDFs with sketch size m, and define.

Lemma 1 (Concentration of weighted empirical CDF)**.** *Assume the calibration scores are independent and the importance weights satisfy* $\mathbb{E}[\omega(V_i)^2] < \infty$*. Then for any* $\delta \in (0,1)$*, with probability at least* $1-\delta$*,*

$$\|\widehat{F} - F\|_\infty \le C_1\sqrt{\frac{\log(1/\delta)}{n}},$$

where $C_1 > 0$ *is a universal constant.*

Lemma 2 (Sketch-to-density approximation via finite difference)**.** *Fix* $h > 0$*. Assume* f *is* L*-Lipschitz on* $\mathcal{V}$ *(i.e.,* $|f(u) - f(v)| \le L|u-v|$*), and similarly each* f^k *is* L_k*-Lipschitz on* $\mathcal{V}$*. Define the finite-difference density surrogates*

$$\widetilde{f}(v) := \frac{\widetilde{F}(v+h) - \widetilde{F}(v-h)}{2h}, \qquad \widetilde{f}^k(v) := \frac{\widetilde{F}^k(v+h) - \widetilde{F}^k(v-h)}{2h}.$$

If $\|\widetilde{F} - F\|_\infty \le \varepsilon_m$ *and* $\|\widetilde{F}^k - F^k\|_\infty \le \varepsilon_m$*, then for all* $v \in \mathcal{V}$*,*

$$|\widetilde{f}(v) - f(v)| \le \frac{\varepsilon_m}{h} + Lh, \qquad |\widetilde{f}^k(v) - f^k(v)| \le \frac{\varepsilon_m}{h} + L_k h.$$

Lemma 3 (Uniform error of approximate importance weights)**.** *Assume* $\inf_{v\in\mathcal{V}} f^k(v) \ge c_0 > 0$ *for all* k*. Let* $\omega(v) := \frac{f(v)}{f^k(v)}$ *and* $\widehat{\omega}(v) := \frac{\widetilde{f}(v)}{\widetilde{f}^k(v)}$*, where* $\widetilde{f}, \widetilde{f}^k$ *are defined in Lemma 2. If* h *is chosen such that* $\frac{\varepsilon_m}{h} + L_k h \le \frac{c_0}{2}$*, then* $\inf_{v\in\mathcal{V}} \widetilde{f}^k(v) \ge c_0/2$ *and*

$$\|\hat{\omega} - \omega\|_\infty \le \frac{C}{c_0}\left(\frac{\varepsilon_m}{h} + (L + L_k)h\right),$$

for some universal constant $C > 0$*.*

Lemma 4 (CDF perturbation induced by weight approximation)**.** *Let* $\widehat{F}_{\hat{\omega}}$ *be the weighted empirical CDF computed using the approximate weights* $\hat{\omega}$*. Assume* $\omega_i, \hat{\omega}_i \ge 0$ *and* $\sum_{i=1}^n \omega_i \ge bn$ *for some* $b \in (0,1]$*. If* $\|\hat{\omega} - \omega\|_\infty \le \frac{b}{2}$*, then* $\sum_{i=1}^n \hat{\omega}_i \ge \frac{b}{2}n$ *and*

$$\|\widehat{F}_{\hat{\omega}} - \widehat{F}_\omega\|_\infty \le \frac{4}{b}\,\|\hat{\omega} - \omega\|_\infty.$$

Lemma 5 (Quantile stability)**.** *Let* $\widehat{q}$ *be the* $(1-\alpha)$ *quantile of* $\widehat{F}_{\widehat{\omega}}$*. If* $\|\widehat{F}_{\widehat{\omega}} - F\|_\infty \le \varepsilon$*, then*

$$\left|\mathbb{P}\Big(Y_{\text{test}} \in \widehat{\mathcal{C}}(X_{\text{test}})\Big) - (1-\alpha)\right| \le \varepsilon.$$

Now we prove Theorem 1. By triangle inequality,

$$\|\widehat{F}_{\hat{\omega}} - F\|_\infty \le \underbrace{\|\widehat{F}_{\hat{\omega}} - \widehat{F}_\omega\|_\infty}_{\text{bias from approximate weights}} + \underbrace{\|\widehat{F}_\omega - F\|_\infty}_{\text{statistical error}}.$$

Step 1. By Lemma 1, with probability at least $1-\delta$,

$$\|\widehat{F}_\omega - F\|_\infty \le C_1\sqrt{\frac{\log(1/\delta)}{n}}.$$

Step 2. Fix a bandwidth $h > 0$ and define the finite-difference density surrogates

$$\widetilde{f}(v) = \frac{\widetilde{F}(v+h) - \widetilde{F}(v-h)}{2h}, \qquad \widetilde{f}^k(v) = \frac{\widetilde{F}^k(v+h) - \widetilde{F}^k(v-h)}{2h}.$$

Assume f is L-Lipschitz on $\mathcal{V}$ and each f^k is L_k-Lipschitz on $\mathcal{V}$. Then Lemma 2 gives, uniformly over $v \in \mathcal{V}$,

$$\|\widetilde{f} - f\|_\infty \le \frac{\varepsilon_m}{h} + Lh, \qquad \|\widetilde{f}^k - f^k\|_\infty \le \frac{\varepsilon_m}{h} + L_k h.$$

Step 3. Assume $\inf_{v\in\mathcal{V}} f^k(v) \ge c_0 > 0$ for all k. If $\frac{\varepsilon_m}{h} + L_k h \le c_0/2$, then $\inf_v \widetilde{f}^k(v) \ge c_0/2$ and Lemma 3 yields

$$\|\widehat{\omega} - \omega\|_\infty \le \frac{C}{c_0}\left(\frac{\varepsilon_m}{h} + (L + L_k)h\right).$$

Step 4. By Lemma 4, we have

$$\|\widehat{F}_{\hat{\omega}} - \widehat{F}_{\omega}\|_\infty \le C'\|\hat{\omega} - \omega\|_\infty \le \frac{C''}{c_0}\left(\frac{\varepsilon_m}{h} + (L + L_k)h\right).$$

Step 5. Choose $h = \sqrt{\varepsilon_m}$. Then $\frac{\varepsilon_m}{h} = \sqrt{\varepsilon_m}$ and

$$\frac{\varepsilon_m}{h} + (L + L_k)h \;=\; \big(1 + L + L_k\big)\sqrt{\varepsilon_m}.$$

Moreover, the condition $\frac{\varepsilon_m}{h} + L_k h \le c_0/2$ reduces to $(1+L_k)\sqrt{\varepsilon_m} \le c_0/2$, which holds for sufficiently large m. Therefore,

$$\|\widehat{F}_{\hat{\omega}} - \widehat{F}_{\omega}\|_\infty \le C_2\,\sqrt{\varepsilon_m}.$$

Combining Step 1–5, with probability at least $1 - \delta$,

$$\|\widetilde{F}_{\hat{\omega}} - F\|_\infty \le C_1\sqrt{\frac{\log(1/\delta)}{n}} + C_2\,\sqrt{\varepsilon_m}.$$

Finally, applying Lemma 5 completes the proof:

$$\left|\mathbb{P}\Big(Y_{\text{test}} \in \widehat{\mathcal{C}}(X_{\text{test}})\Big) - (1-\alpha)\right| \le C_1\sqrt{\frac{\log(1/\delta)}{n}} + C_2\,\sqrt{\varepsilon_m}.$$

□

B.3 PROOF OF TECHNICAL LEMMAS

B.3.1 Proof of Lemma 1

Lemma (Re) 1 (Lemma 1 restated, concentration of weighted empirical CDF)**.** *Assume the calibration scores are independent and the importance weights satisfy $\mathbb{E}[\omega(V_i)^2] < \infty$. Then for any $\delta \in (0,1)$, with probability at least $1 - \delta$,*

$$\|\widehat{F} - F\|_\infty \le C_1\sqrt{\frac{\log(1/\delta)}{n}},$$

where $C_1 > 0$ is a universal constant.

Proof. (of Lemma 1)

Before proving Lemma 1, we first show the following technical lemma:

Lemma 6 (Importance weighting corrects joint shift)**.** *Under joint distribution shift, the importance weights satisfy*

$$\mathbb{E}[\omega(V_i)\mathbf{1}\{V_i \le t\}] = F(t).$$

Lemma 7 (SNIS concentration for threshold function class, Owen [2013]). *Assume $\{V_i\}_{i=1}^n$ are independent, $\omega(V_i) \geq 0$, $\mathbb{E}[\omega(V)] = 1$, and $\mathbb{E}[\omega(V)^2] < \infty$. Then there exists a universal constant $c > 0$ such that for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sup_{t\in\mathbb{R}} \left|A_n(t) - \mathbb{E}A_n(t)\right| \leq c\sqrt{\frac{\mathbb{E}[\omega(V)^2]\log(2/\delta)}{n}}, \qquad |B_n - \mathbb{E}B_n| \leq c\sqrt{\frac{\mathbb{E}[\omega(V)^2]\log(2/\delta)}{n}}.$$

The proof of Lemma 6 is deferred to the end of this proof. Now we begin to prove Lemma 1.

Recall the weighted empirical CDF as:

$$\widehat{F}(t) = \frac{\sum_{i=1}^n \omega_i \mathbf{1}\{V_i \leq t\}}{\sum_{i=1}^n \omega_i}, \qquad \omega_i := \omega(V_i).$$

Let

$$A_n(t) := \frac{1}{n}\sum_{i=1}^n \omega_i \mathbf{1}\{V_i \leq t\}, \qquad B_n := \frac{1}{n}\sum_{i=1}^n \omega_i,$$

so that $\widehat{F}(t) = A_n(t)/B_n$.

By Lemma 6,

$$\mathbb{E}[A_n(t)] = \mathbb{E}[\omega(V)\mathbf{1}\{V \leq t\}] = F(t), \qquad \mathbb{E}[B_n] = \mathbb{E}[\omega(V)] = 1.$$

Define the centered processes

$$\Delta_A(t) := A_n(t) - F(t), \qquad \Delta_B := B_n - 1.$$

Then for any t,

$$\widehat{F}(t) - F(t) = \frac{A_n(t)}{B_n} - F(t) = \frac{F(t) + \Delta_A(t)}{1+\Delta_B} - F(t) = \frac{\Delta_A(t) - F(t)\Delta_B}{1+\Delta_B}.$$

Hence, using $F(t) \in [0,1]$, we have:

$$|\widehat{F}(t) - F(t)| \leq \frac{|\Delta_A(t)| + |\Delta_B|}{1 - |\Delta_B|} \quad \text{whenever } |\Delta_B| \leq 1/2, \tag{9}$$

Let $\varepsilon_n(\delta) := c\sqrt{\frac{\mathbb{E}[\omega(V)^2]\log(2/\delta)}{n}}$. With probability at least $1-\delta$ (by a union bound over the two events in Lemma 7),

$$\sup_t |\Delta_A(t)| \leq \varepsilon_n(\delta/2), \qquad |\Delta_B| \leq \varepsilon_n(\delta/2).$$

Further, for n large enough such that $\varepsilon_n(\delta/2) \leq 1/2$, we can plug into (9) and obtain, uniformly over t,

$$\|\widehat{F} - F\|_\infty = \sup_t |\widehat{F}(t) - F(t)| \leq \frac{\varepsilon_n(\delta/2) + \varepsilon_n(\delta/2)}{1 - \varepsilon_n(\delta/2)} \leq 4\,\varepsilon_n(\delta/2).$$

Absorbing constants (including $\sqrt{\mathbb{E}[\omega(V)^2]}$) into a single constant $C_1 > 0$ yields

$$\|\widehat{F} - F\|_\infty \leq C_1\sqrt{\frac{\log(1/\delta)}{n}},$$

which proves Lemma 1.

□

Now we begin to prove Lemma 6

Proof. (of Lemma 6)

By the definition of the importance ratio,

$$\mathbb{E}[\omega(V)\mathbf{1}\{V \leq t\}] = \int \frac{dF}{dF_k}(v)\mathbf{1}\{v \leq t\} dF_k(v) = \int_{v \leq t} dF(v) = F(t).$$

□

B.3.2 Proof of Lemma 2

Lemma (Re) 2 (Lemma 2 restated, sketch-to-density approximation via finite difference)**.** *Fix $h > 0$. Assume f is L-Lipschitz on $\mathcal{V}$ (i.e., $|f(u) - f(v)| \le L|u - v|$), and similarly each f^k is L_k-Lipschitz on $\mathcal{V}$. Define the finite-difference density surrogates*

$$\widetilde{f}(v) := \frac{\widetilde{F}(v+h) - \widetilde{F}(v-h)}{2h}, \qquad \widetilde{f}^k(v) := \frac{\widetilde{F}^k(v+h) - \widetilde{F}^k(v-h)}{2h}.$$

If $\|\widetilde{F} - F\|_\infty \le \varepsilon_m$ and $\|\widetilde{F}^k - F^k\|_\infty \le \varepsilon_m$, then for all $v \in \mathcal{V}$,

$$|\widetilde{f}(v) - f(v)| \le \frac{\varepsilon_m}{h} + Lh, \qquad |\widetilde{f}^k(v) - f^k(v)| \le \frac{\varepsilon_m}{h} + L_k h.$$

Proof. Define $f_h(v) := \frac{F(v+h)-F(v-h)}{2h} = \frac{1}{2h}\int_{v-h}^{v+h} f(u)\,du$. Then

$$|\widetilde{f}(v) - f(v)| \le |\widetilde{f}(v) - f_h(v)| + |f_h(v) - f(v)|.$$

For the first term,

$$|\widetilde{f}(v) - f_h(v)| = \frac{1}{2h}\Big|(\widetilde{F} - F)(v+h) - (\widetilde{F} - F)(v-h)\Big| \le \frac{\varepsilon_m}{h}.$$

For the second term, by Lipschitzness,

$$|f_h(v) - f(v)| = \left|\frac{1}{2h}\int_{v-h}^{v+h}(f(u) - f(v))\,du\right| \le \frac{1}{2h}\int_{v-h}^{v+h} L|u - v|\,du = Lh.$$

The proof for $\widetilde{f}^k$ is identical. □

B.3.3 Proof of Lemma 3

Lemma (Re) 3 (Lemma 3 restated, uniform error of approximate importance weights)**.** *Assume $\inf_{v\in\mathcal{V}} f^k(v) \ge c_0 > 0$ for all k. Let $\omega(v) := \frac{f(v)}{f^k(v)}$ and $\widehat{\omega}(v) := \frac{\widetilde{f}(v)}{\widetilde{f}^k(v)}$, where $\widetilde{f}, \widetilde{f}^k$ are defined in Lemma 2. If h is chosen such that $\frac{\varepsilon_m}{h} + L_k h \le \frac{c_0}{2}$, then $\inf_{v\in\mathcal{V}} \widetilde{f}^k(v) \ge c_0/2$ and*

$$\|\hat{\omega} - \omega\|_\infty \le \frac{C}{c_0}\left(\frac{\varepsilon_m}{h} + (L + L_k)h\right),$$

for some universal constant $C > 0$.

Proof. Fix an arbitrary client k and an arbitrary $v \in \mathcal{V}$. Recall the definitions

$$\omega(v) = \frac{f(v)}{f^k(v)}, \qquad \hat{\omega}(v) = \frac{\tilde{f}(v)}{\tilde{f}^k(v)}.$$

We first show that the denominator $\tilde{f}^k(v)$ is uniformly bounded away from zero. By Lemma 2, for all $v \in \mathcal{V}$,

$$|\tilde{f}^k(v) - f^k(v)| \le \frac{\varepsilon_m}{h} + L_k h.$$

Under the condition $\frac{\varepsilon_m}{h} + L_k h \le c_0/2$ and the assumption $\inf_{v\in\mathcal{V}} f^k(v) \ge c_0$, we have

$$\tilde{f}^k(v) \ge f^k(v) - |\tilde{f}^k(v) - f^k(v)| \ge c_0 - \frac{c_0}{2} = \frac{c_0}{2}.$$

Since v is arbitrary, this implies $\inf_{v\in\mathcal{V}} \tilde{f}^k(v) \ge c_0/2$.

Next, we bound the pointwise weight error $|\hat{\omega}(v) - \omega(v)|$. Start from a common-denominator expansion:

$$\hat{\omega}(v) - \omega(v) = \frac{\tilde{f}(v)}{\tilde{f}^k(v)} - \frac{f(v)}{f^k(v)} = \frac{\tilde{f}(v)f^k(v) - f(v)\tilde{f}^k(v)}{\tilde{f}^k(v)\,f^k(v)}.$$

Add and subtract $f(v)f^k(v)$ in the numerator:

$$\begin{aligned}\tilde{f}(v)f^k(v) - f(v)\tilde{f}^k(v) &= \big(\tilde{f}(v)f^k(v) - f(v)f^k(v)\big) + \big(f(v)f^k(v) - f(v)\tilde{f}^k(v)\big) \\ &= \big(\tilde{f}(v) - f(v)\big)f^k(v) + f(v)\big(f^k(v) - \tilde{f}^k(v)\big).\end{aligned}$$

Taking absolute values and applying the triangle inequality gives

$$|\hat{\omega}(v) - \omega(v)| \le \frac{|\,\tilde{f}(v) - f(v)\,|\,f^k(v) + |f(v)|\,|\,f^k(v) - \tilde{f}^k(v)\,|}{\tilde{f}^k(v)\,f^k(v)}.$$

Using the lower bounds $f^k(v) \ge c_0$ and $\tilde{f}^k(v) \ge c_0/2$ derived above, we obtain

$$\begin{aligned}|\hat{\omega}(v) - \omega(v)| &\le \frac{|\,\tilde{f}(v) - f(v)\,|\,f^k(v)}{(c_0/2)\,f^k(v)} + \frac{|f(v)|\,|\,f^k(v) - \tilde{f}^k(v)\,|}{(c_0/2)\,f^k(v)} \\ &= \frac{2}{c_0}\,|\,\tilde{f}(v) - f(v)\,| + \frac{2|f(v)|}{c_0\,f^k(v)}\,|\,\tilde{f}^k(v) - f^k(v)\,| \\ &\le \frac{2}{c_0}\,|\,\tilde{f}(v) - f(v)\,| + \frac{2\|f\|_\infty}{c_0^2}\,|\,\tilde{f}^k(v) - f^k(v)\,|,\end{aligned}$$

where in the last inequality we used $|f(v)| \le \|f\|_\infty$ and $f^k(v) \ge c_0$.

Finally, apply Lemma 2 to bound the density approximation errors: for all $v \in \mathcal{V}$,

$$|\,\tilde{f}(v) - f(v)\,| \le \frac{\varepsilon_m}{h} + Lh, \qquad |\,\tilde{f}^k(v) - f^k(v)\,| \le \frac{\varepsilon_m}{h} + L_k h.$$

Substituting these bounds into the previous inequality yields

$$\begin{aligned}|\hat{\omega}(v) - \omega(v)| &\le \frac{2}{c_0}\left(\frac{\varepsilon_m}{h} + Lh\right) + \frac{2\|f\|_\infty}{c_0^2}\left(\frac{\varepsilon_m}{h} + L_k h\right) \\ &= \left(\frac{2}{c_0} + \frac{2\|f\|_\infty}{c_0^2}\right)\frac{\varepsilon_m}{h} + \left(\frac{2L}{c_0} + \frac{2\|f\|_\infty L_k}{c_0^2}\right)h.\end{aligned}$$

Taking supremum over $v \in \mathcal{V}$ on both sides gives

$$\|\hat{\omega} - \omega\|_\infty \le \left(\frac{2}{c_0} + \frac{2\|f\|_\infty}{c_0^2}\right)\frac{\varepsilon_m}{h} + \left(\frac{2L}{c_0} + \frac{2\|f\|_\infty L_k}{c_0^2}\right)h.$$

Absorbing the coefficients into a constant $C > 0$ depending only on c_0 and $\|f\|_\infty$ (and using that L and L_k are fixed smoothness constants) yields the stated bound:

$$\|\hat{\omega} - \omega\|_\infty \le \frac{C}{c_0}\left(\frac{\varepsilon_m}{h} + (L + L_k)h\right).$$

This completes the proof. □

B.3.4 Proof of Lemma 4

Lemma (Re) 4 (Lemma 4, CDF perturbation induced by weight approximation)**.** *Let $\widehat{F}_{\hat{\omega}}$ be the weighted empirical CDF computed using the approximate weights $\hat{\omega}$. Assume $\omega_i, \hat{\omega}_i \ge 0$ and $\sum_{i=1}^n \omega_i \ge bn$ for some $b \in (0, 1]$. If $\|\hat{\omega} - \omega\|_\infty \le \frac{b}{2}$, then $\sum_{i=1}^n \hat{\omega}_i \ge \frac{b}{2}n$ and*

$$\|\widehat{F}_{\hat{\omega}} - \widehat{F}_{\omega}\|_\infty \le \frac{4}{b}\,\|\hat{\omega} - \omega\|_\infty.$$

Proof. For any t, write $A(t) = \sum_i \omega_i \mathbf{1}\{V_i \le t\}$, $B = \sum_i \omega_i$ and $\hat{A}(t) = \sum_i \hat{\omega}_i \mathbf{1}\{V_i \le t\}$, $\hat{B} = \sum_i \hat{\omega}_i$. Then

$$|\widehat{F}_{\hat{\omega}}(t) - \widehat{F}_{\omega}(t)| = \left|\frac{\hat{A}(t)}{\hat{B}} - \frac{A(t)}{B}\right| \le \frac{|\hat{A}(t) - A(t)|}{\hat{B}} + \frac{A(t)}{B\hat{B}}|\hat{B} - B|.$$

Since $\omega_i, \hat{\omega}_i \geq 0$, we have $A(t) \leq B$. Moreover, $|\hat{A}(t) - A(t)| \leq \sum_i |\hat{\omega}_i - \omega_i| \leq n\|\hat{\omega} - \omega\|_\infty$ and $|\hat{B} - B| \leq n\|\hat{\omega} - \omega\|_\infty$. Under $B \geq bn$ and $\|\hat{\omega} - \omega\|_\infty \leq b/2$, we have $\hat{B} \geq bn/2$. Plugging in yields

$$|\widehat{F}_{\hat{\omega}}(t) - \widehat{F}_{\omega}(t)| \leq \frac{n\|\hat{\omega} - \omega\|_\infty}{bn/2} + \frac{n\|\hat{\omega} - \omega\|_\infty}{bn/2} = \frac{4}{b}\|\hat{\omega} - \omega\|_\infty.$$

Taking supremum over t completes the proof. □

B.3.5 Proof of Lemma 5

Lemma (Re) 5 (Lemma 5 restated, quantile stability)**.** *Let $\widehat{q}$ be the $(1-\alpha)$ quantile of $\widehat{F}_{\widehat{\omega}}$. If $\|\widehat{F}_{\widehat{\omega}} - F\|_\infty \leq \varepsilon$, then*

$$\left|\mathbb{P}\Big(Y_{\text{test}} \in \widehat{\mathcal{C}}(X_{\text{test}})\Big) - (1-\alpha)\right| \leq \varepsilon.$$

Proof. (of Lemma 5)

By the definition of the quantile,

$$\widehat{F}_{\widehat{\omega}}(\widehat{q}) \geq 1 - \alpha.$$

Therefore,

$$F(\widehat{q}) \geq \widehat{F}_{\widehat{\omega}}(\widehat{q}) - \varepsilon \geq 1 - \alpha - \varepsilon,$$

and similarly

$$F(\widehat{q}) \leq \widehat{F}_{\widehat{\omega}}(\widehat{q}) + \varepsilon \leq 1 - \alpha + \varepsilon.$$

□

C ADDITIONAL EXPERIMENTS

C.1 TRAINING DETAILS

We conduct experiments on the RxRx1 dataset Taylor et al. [2019] from the WILDS benchmark Koh et al. [2021], which is designed to evaluate model robustness under distribution shift.

RxRx1 consists of high-resolution fluorescence microscopy images of human cells subjected to $1,339$ different genetic treatments, collected across 51 experiments and 4 cell types, each representing a distinct domain. Due to experimental variability, the dataset exhibits significant covariate shift across domains. Following the WILDS protocol, we utilize a ResNet50 model pre-trained by the WILDS authors on data from 37 experiments using supervised learning with a cross-entropy loss, where the task is to classify each image into one of the $1,139$ treatment labels. The top-1 accuracy is 29.9%. The remaining 14 experiments, which represent unseen domains, are partitioned into separate calibration and test sets to evaluate generalization under distribution shift.

FMoW is a satellite image classification dataset comprising over 360,000 images spanning diverse geographies and time periods, with each image labeled according to its land use or functional category. Due to variations in geographic regions and temporal acquisition, the dataset exhibits substantial domain shift across locations. Following the WILDS protocol, we use a DenseNet121 model pre-trained by the WILDS authors using supervised learning with a cross-entropy loss, where the task is to classify each image into one of the 62 functional categories. The top-1 accuracy is 55.5%. The evaluation is conducted on held-out regions not seen during training, to assess model generalization under distribution shift.

iWildCam is a wildlife image dataset consisting of camera trap photos collected from conservation areas across various geographic locations and years. The dataset presents significant domain shift due to differences in camera hardware, lighting, and animal species distribution across locations. We follow the WILDS benchmark setup and use a ResNet50 backbone model trained with cross-entropy loss on labeled data from training domains to classify each image into one of the 182 animal species. The top-1 accuracy is 75.7%. The test domains consist of held-out camera trap locations, used to evaluate model robustness under spatial and temporal distribution shifts.

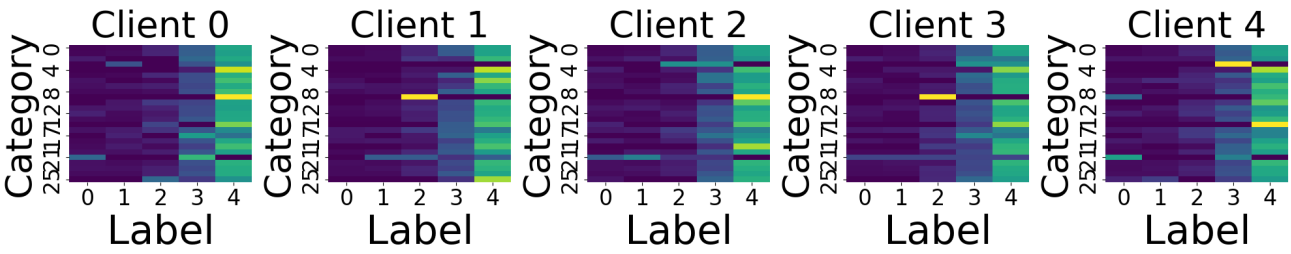

Figure 5: **Joint distribution heatmaps of groups (categories, y-axis) and labels (x-axis) cross all clients under the IID partitioning** of the Amazon Reviews dataset. Each client receives a uniformly sampled subset of the global data, resulting in homogeneous data distributions of each client.

C.2 CALIBRATION DETAILS

We randomly split the validation set of each dataset into calibration and testing datasets. We repeat calibration experiments for 3 times for FCPCS, and 10 times for the remaining methods to compute the mean and standard deviation. We use three commonly used scoring functions (HPS, APS, and RAPS), as non-conformity scores, the detailed function is introduced in Appendix A. Following Ding et al. [2024], Tawachi and Laufer-Goldshtein [2025], Campos et al. [2024], we set $\lambda_{\text{RAPS}} = 0.01$ and $k_{\text{reg}} = 5$ as fixed hyperparameters for RAPS score, which is the commonly used hyperparameter combination for RAPS. We set $\alpha = [0.01, 0.1]$ with range 0.01 as various mis-coverage rates.

C.3 ADDITIONAL RESULTS

C.3.1 Partition Diagnostics

Our partitioning strategy could simulate IID setting and non-IID setting with joint distribution shift.

To illustrate the data distribution under the *IID* setting, Figure 5 displays the joint distribution of experiment identifiers and treatment labels for all 10 clients in the RxRx1 dataset. Each client receives a uniformly sampled subset of the global data, with no restriction on experiment or label selection. The resulting heatmaps show a dense and homogeneous coverage over both axes across clients, indicating that the local data distributions are representative of the overall distribution.

To illustrate the data distribution under the *non-IID* setting, Figure 6 and 7 display the joint distribution of experiment identifiers and treatment labels for all 20 clients in the FMoW and iWildCam datasets, respectively. Each client is assigned a disjoint subset of experiments (introducing covariate shift in FMoW or joint shift in iWildCam) and labels (introducing label shift), resulting in sparse and localized distributions in the heatmaps. This clearly demonstrates that our partitioning simulates a joint distribution shift, effectively capturing the heterogeneity characteristic of real-world FL scenarios.

C.3.2 Sensitivity to Confidence Levels

FCPJS consistently maintains valid coverage and smaller prediction sets across varying confidence levels.

To further evaluate the robustness of each method, we examine their performance across a range of confidence levels $(1-\alpha) \in \{0.90, 0.91, \cdots, 0.99\}$ with range 0.01 on the three datasets with non-IID partition in Figure 3, 9, and 10, respectively. We also compare the performance of each method across a range of confidence levels $(1-\alpha) \in \{0.90, 0.91, \cdots, 0.99\}$ with range 0.01 on the RxRx1 with IID partition in Figure 8, Figure 3, 8, 9, and 10 show the results for three score functions, plotting marginal coverage (top row) and average prediction set size (bottom row) across increasing confidence levels $1-\alpha$. Across all score types, FCPJS consistently maintains valid coverage, which closely tracks the diagonal coverage targets, while baseline methods either undercover (FCPLS) or show unstable performance (FCP). Notably, FCPJS achieves significantly smaller prediction sets than FCP as confidence increases. This trend reflects the method's ability to adaptively balance coverage and efficiency under stricter reliability requirements. These results demonstrate that FCPJS provides strong coverage guarantees across confidence levels while maintaining competitive predictive efficiency.

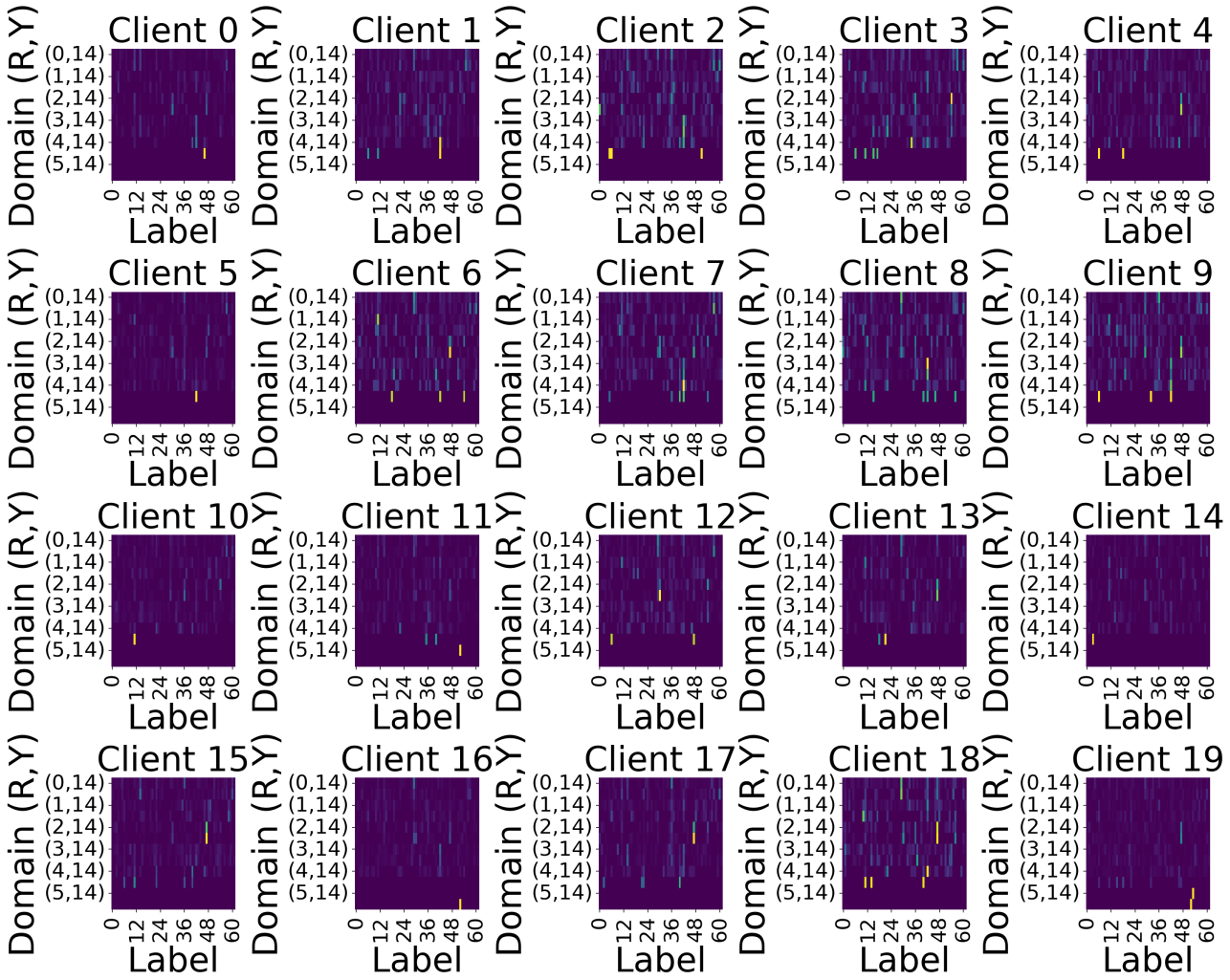

Figure 6: **Joint distribution heatmaps of groups (Region $\times$ year, y-axis) and labels (x-axis) cross all clients under the non-IID partitioning** of the FMoW dataset. Each client is assigned disjoint experiments and disjoint label subsets, resulting in highly sparse and heterogeneous data distributions that reflect joint covariate and label shift.

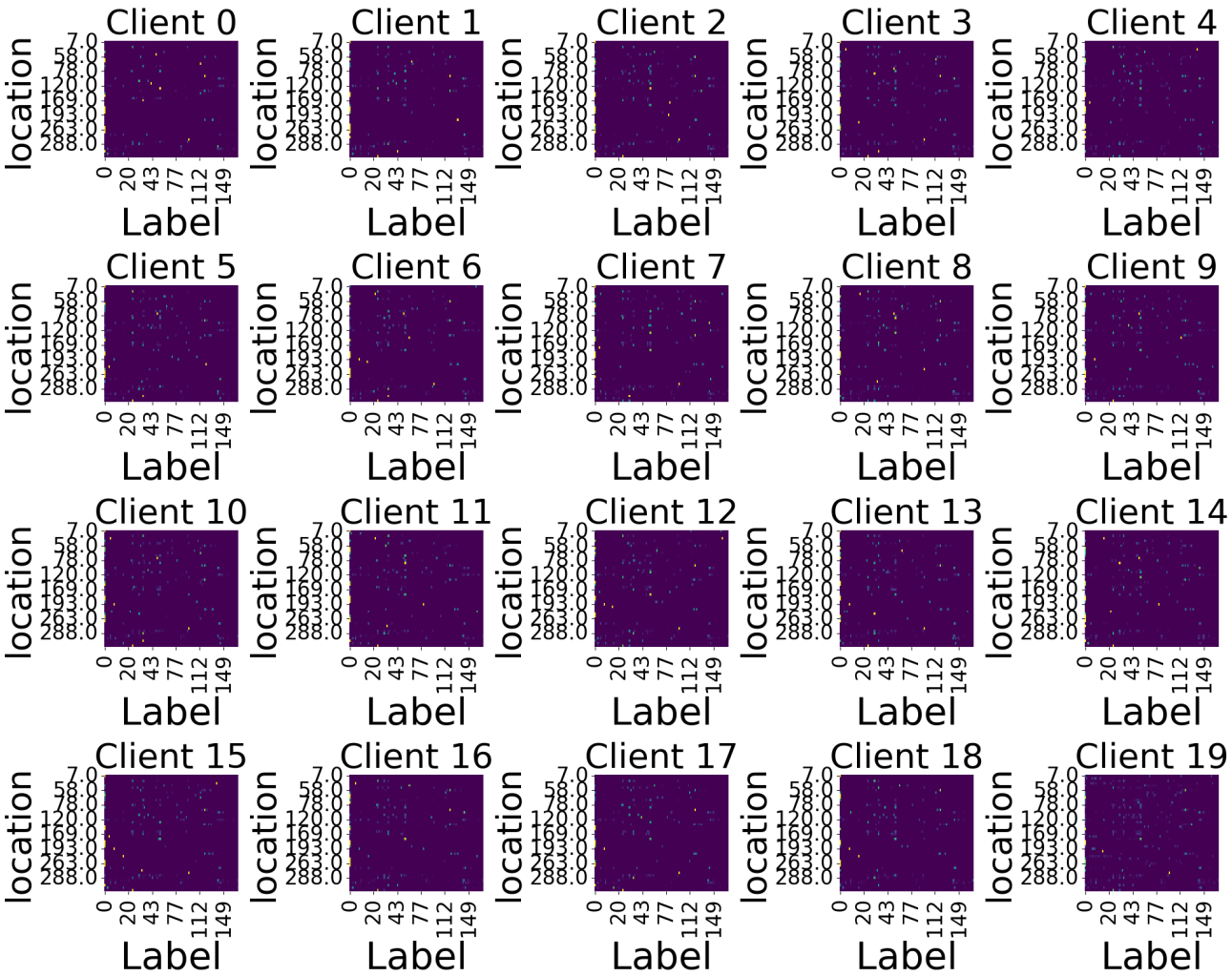

Figure 7: **Joint distribution heatmaps of groups (Location, y-axis) and labels (x-axis) cross all clients under the non-IID partitioning** of the iWildCam dataset. Each client is assigned disjoint experiments and disjoint label subsets, resulting in highly sparse and heterogeneous data distributions that reflect joint covariate and label shift.

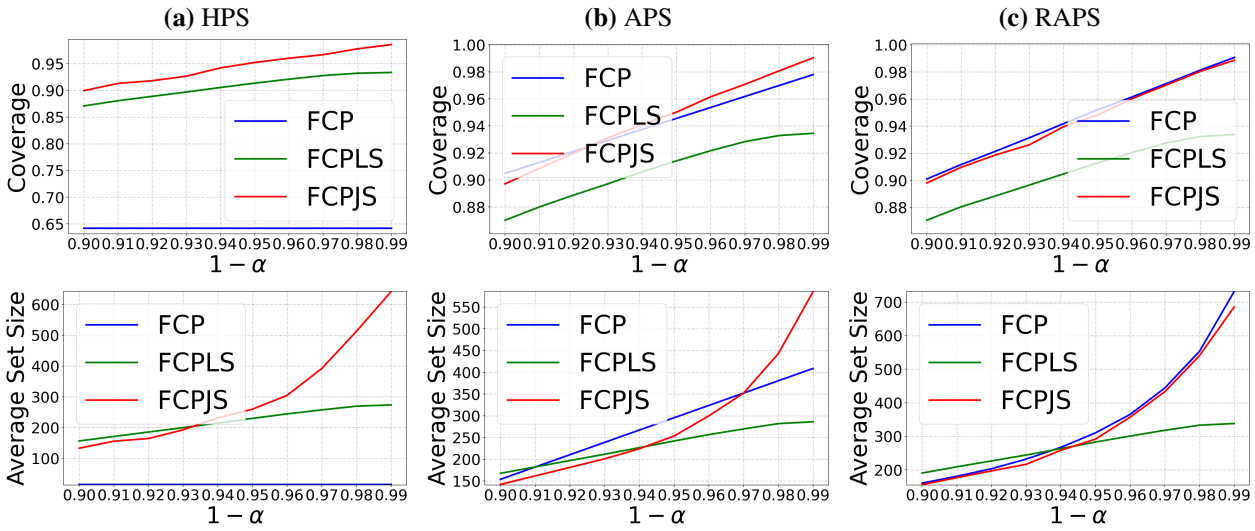

Figure 8: **Coverage (Top) and average set size (Bottom) results of all methods with three scores and under different confidence levels on RxRx1 datasets, IID partition**, where $1-\alpha \in \{0.90, 0.91, \cdots, 0.99\}$ with range 0.01. It is clear that FCPJS consistently maintains valid coverage and smaller prediction sets across varying confidence levels with different scores.

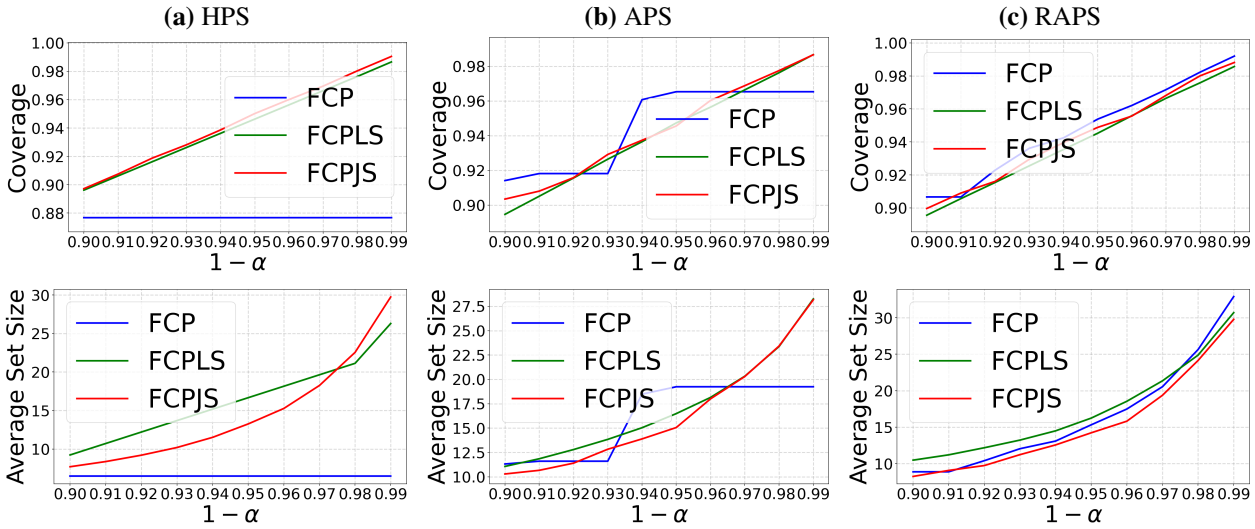

Figure 9: **Coverage (Top) and average set size (Bottom) results of all methods with three scores and under different confidence levels on FMoW datasets, non-IID partition**, where $1-\alpha \in \{0.90, 0.91, \cdots, 0.99\}$ with range 0.01. It is clear that FCPJS consistently maintains valid coverage and smaller prediction sets across varying confidence levels with different scores.

C.3.3 Differentially Private Variant.

To assess robustness under privacy constraints, we consider a record-level (ε, δ)-differentially private (DP) variant of our method based on the Gaussian mechanism. We map scores to a bounded domain $[0, 1]$ via a fixed monotone transformation and partition the interval into B bins. The resulting histogram count vector

$$h_k \in \mathbb{R}^B, \qquad (h_k)_b = \sum_{i=1}^{n_k} \mathbf{1}\{v_i \in \text{bin } b\},$$

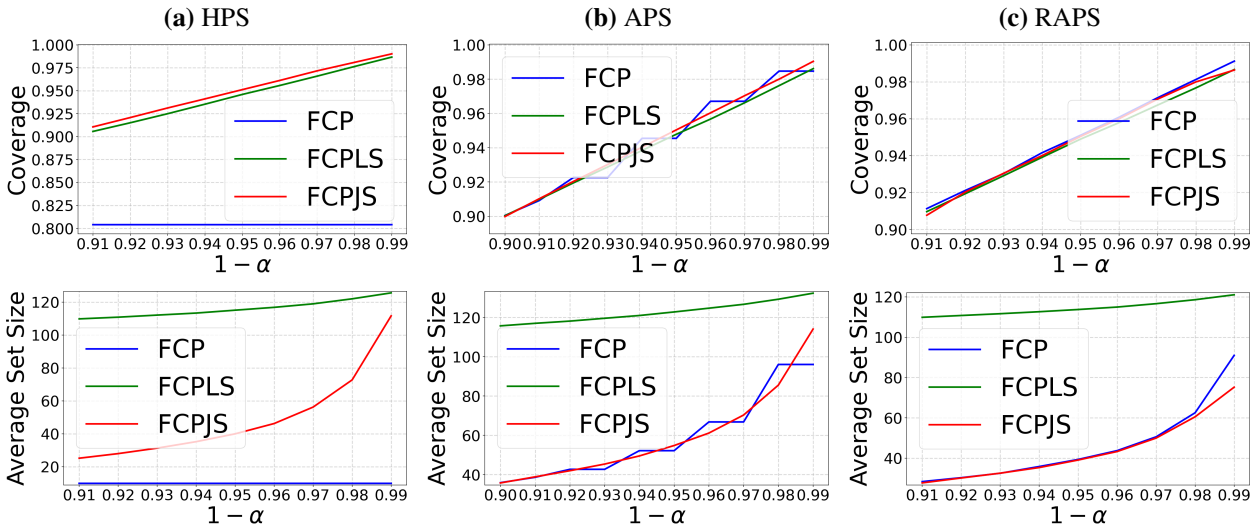

Figure 10: **Coverage (Top) and average set size (Bottom) results of all methods with three scores and under different confidence levels on iWildCam datasets, non-IID partition**, where $1-\alpha \in \{0.90, 0.91, \cdots, 0.99\}$ with range 0.01. It is clear that FCPJS consistently maintains valid coverage and smaller prediction sets across varying confidence levels with different scores.

has ℓ_2-sensitivity 1 under record-level adjacency. We therefore release

$$\tilde{h}_k = h_k + \mathcal{N}(0, \sigma^2 I), \qquad \sigma = \frac{\sqrt{2\ln(1.25/\delta)}}{\varepsilon},$$

which guarantees (ε, δ)-DP for each client.

The server constructs privatized source and target density estimates from the noisy histograms and computes importance weights using these densities. Each client privatizes its own histogram locally before transmission, and the server only accesses noisy statistics. All subsequent steps, including weighted conformal quantile computation, are post-processing and incur no additional privacy loss. Importantly, privacy noise is applied only to the calibration score distribution, while the conformal prediction procedure itself remains unchanged.

DP variant of FCPJSmaintains valid coverage while exhibiting stable efficiency across a wide range of privacy budgets. We evaluate the DP variant of FCPJSon the RxRx1 dataset and construct a heterogeneous (non-IID) federated setting following the dataset's experimental metadata. Specifically, we set the number of histogram bins to $B = 400$ and use a squashing map to bound scores to $[0, 1]$. We sweep the privacy budget over $\varepsilon \in \{0.1, 1.0, 10.0\}$ with $\delta = 10^{-5}$, and keep all remaining hyperparameters identical to the non-private setting.

As shown in Table 3, the coverage remains at or above the target level $1 - \alpha = 0.9$ for most score functions across $\varepsilon \in \{0.1, 1.0, 10.0\}$. HPS and RAPS show negligible variation in both coverage and MAPSS across privacy budgets. Although APS exhibits larger fluctuation at $\varepsilon = 1.0$, valid coverage is recovered as ε increases. Overall, the method maintains stable performance under moderate calibration-level DP noise in the non-IID federated setting.

C.3.4 Additional Comparison with FCPCS

FCPJS is more efficient than FCPCS on iWildCam non-IID while maintaining valid coverage.

FCPCS is the closest prior method based on feature-space density-ratio correction, but it is substantially more computationally expensive. To strengthen the comparison beyond RxRx1, Table 4 reports an additional comparison on iWildCam under the non-IID partition across HPS, APS, and RAPS.

Both FCPCS and FCPJS maintain valid coverage. However, FCPJS reduces MAPSS by 11.16% under HPS, 14.71% under APS, and 9.66% under RAPS. This indicates that score-space alignment is not only cheaper than feature-space density-ratio correction, but can also produce smaller prediction sets under joint shift.

Partition	Score	$\varepsilon = 0.1$		$\varepsilon = 1.0$		$\varepsilon = 10.0$	
		MCOV	MAPSS	MCOV	MAPSS	MCOV	MAPSS
Non-IID	HPS	0.90 ± 0.01	149.96 ± 8.43	0.90 ± 0.01	148.55 ± 5.96	0.90 ± 0.00	147.58 ± 7.84
	APS	0.91 ± 0.01	161.24 ± 11.15	0.90 ± 0.02	158.64 ± 11.48	0.90 ± 0.01	158.54 ± 11.58
	RAPS	0.90 ± 0.02	166.90 ± 9.54	0.90 ± 0.01	166.42 ± 9.73	0.90 ± 0.01	166.26 ± 9.91

Table 3: **DP robustness of FCPJS** on RxRx1 under non-IID partition with $\varepsilon \in \{0.1, 1.0, 10.0\}$ and across three scoring functions (HPS, APS, RAPS). We report the mean and std results of coverage (MCOV) and mean average prediction set size (MAPSS). Across different score functions and privacy budgets, the method maintains valid coverage while exhibiting stable efficiency.

Table 4: Additional comparison with FCPCS on iWildCam under the non-IID partition. Both methods achieve valid marginal coverage, while FCPJS yields smaller prediction sets.

Method	MCOV (HPS)	MAPSS (HPS)	MCOV (APS)	MAPSS (APS)	MCOV (RAPS)	MAPSS (RAPS)
FCPCS	0.92	27.78	0.91	41.95	0.92	28.38
FCPJS	0.91	24.68	0.90	35.78	0.90	25.64

Therefore, the additional FCPCS comparison supports both the computational and statistical motivation for FCPJS.

C.3.5 Centralized Oracle Comparison

FCPJS approaches the non-private centralized oracle on HPS and APS, while preserving the federated protocol.

We compare FCPJS with a centralized oracle on RxRx1. The oracle pools all calibration scores and computes a single conformal threshold. This violates the federated privacy constraint and is not a deployable FL baseline, but it provides a useful reference for the efficiency gap to pooled calibration.

FCPJS is close to the centralized oracle on HPS and APS, with only a small increase in MAPSS. On RAPS, the centralized oracle remains more efficient, indicating that some room remains relative to the non-private pooled-calibration limit. Importantly, FCPJS achieves these results without centralizing raw calibration data or individual scores.

This comparison shows that FCPJS can recover much of the benefit of pooled calibration while respecting the federated communication constraint.

C.3.6 Weight-Stability and Overlap Diagnostics

The score-density weights are variable under heterogeneous clients, but they are not catastrophically concentrated in our main benchmark.

The coverage guarantee relies on score-space overlap: each client should have sufficient local score density in regions that matter for the global conformal quantile. Since this condition cannot be verified exactly from finite samples, we report practical weight-stability diagnostics. Table 6 reports the normalized effective sample size,

$$\mathrm{ESS}/n = \frac{(\sum_i w_i)^2}{n \sum_i w_i^2},$$

where smaller values indicate stronger weight concentration.

The ESS ratios remain a nontrivial fraction of the calibration set for all three scoring functions. This suggests that the weighted quantile is not dominated by only a few samples. Table 7 further reports the raw density-ratio weight moments.

The raw weights show non-negligible variability, especially for HPS and APS. This confirms that weight degeneracy is a relevant diagnostic in heterogeneous federated calibration. However, the ESS results in Table 6 remain nontrivial and the corresponding coverage stays near the target. In implementation, we also apply a small density floor before computing ratios to avoid division by nearly empty bins.

Table 5: Comparison with a non-private centralized oracle on RxRx1. The centralized oracle violates the federated constraint by pooling calibration scores.

Method	MCOV (HPS)	MAPSS (HPS)	MCOV (APS)	MAPSS (APS)	MCOV (RAPS)	MAPSS (RAPS)
Centralized	0.90 ± 0.00	131.40 ± 2.30	0.90 ± 0.00	141.25 ± 2.19	0.90 ± 0.00	136.24 ± 2.71
FCPJS (IID)	0.91 ± 0.01	133.02 ± 3.39	0.90 ± 0.00	142.10 ± 2.02	0.90 ± 0.00	156.21 ± 3.33

Table 6: Normalized effective sample size on RxRx1 under the non-IID partition. Larger values indicate less severe weight concentration.

Metric	HPS	APS	RAPS
ESS$/n$	0.367 ± 0.08	0.412 ± 0.11	0.624 ± 0.06

Overall, these diagnostics support the practical stability of the score-density reweighting step in our evaluated regimes, while also clarifying that extreme support mismatch remains a possible failure mode.

C.3.7 Overlap Stress Test

When local overlap is artificially reduced, coverage degrades smoothly rather than collapsing.

To directly stress the score-space overlap condition, we thin calibration scores near the empirical $(1-\alpha)$-quantile. The keep rate denotes the fraction of calibration scores retained in a local window around the target quantile. Table 8 reports both MCOV and ESS$/n$ under different keep rates on RxRx1 non-IID.

Coverage decreases as overlap is reduced, which is expected because weaker local support makes density-ratio estimation harder. However, the degradation is gradual: even at keep rate 0.1, the largest coverage drop is about 0.02. The ESS ratios also remain nontrivial, suggesting that the weighted quantile does not collapse to a few calibration points.

This stress test confirms the role of the overlap assumption and shows that FCPJS remains stable under the evaluated overlap perturbations.

C.3.8 Sensitivity to Density-Ratio Stabilization

FCPJS is stable across small temperature values, while large temperature values make the weighted quantile more conservative.

The implementation tempers log-density ratios before exponentiation. This temperature parameter controls a finite-sample bias–variance tradeoff: smaller values shrink extreme density ratios and reduce weight variance, while $t = 1$ corresponds to the untempered density-ratio estimator. Table 9 reports a sensitivity study on RxRx1 non-IID with APS.

Coverage remains valid for all tested values of t. For small temperatures, especially $t \in \{0.0001, 0.001, 0.01\}$, MAPSS is relatively stable and close to its minimum. As t increases, tempering becomes weaker and the finite-sample variance of the density-ratio weights increases, leading to more conservative prediction sets. We therefore use $t = 0.001$ as the default because it preserves target coverage with near-minimal set size.

These results show that the empirical conclusions are not artifacts of one specific temperature value.

FCPJS is also insensitive to the practical density-estimation resolution used for score-density ratios.

We next vary the density-estimation resolution used to approximate score densities from sketches. Let B denote the number of bins, equivalently an effective resolution $\Delta = 1/B$. Table 10 reports the results on RxRx1 non-IID with HPS.

MCOV remains exactly at the target level across all tested resolutions, while MAPSS changes only slightly. This suggests that the finite-difference density-ratio estimation step is not overly sensitive to the chosen resolution in this setting.

Together, the temperature and density-resolution ablations show that the stabilized score-density-ratio implementation is robust to reasonable hyperparameter choices.

Table 7: Raw density-ratio weight moments on RxRx1 under the non-IID partition.

Metric	HPS	APS	RAPS
Weight	27.39 ± 33.87	16.16 ± 22.16	1.477 ± 1.264

Table 8: Overlap stress test on RxRx1 under the non-IID partition. The keep rate controls the fraction of calibration scores retained near the empirical $(1-\alpha)$-quantile.

Score	Metric	Keep=1.0	Keep=0.75	Keep=0.5	Keep=0.25	Keep=0.1
HPS	MCOV	0.90	0.90	0.90	0.90	0.89
	ESS$/n$	0.367	0.378	0.379	0.410	0.426
APS	MCOV	0.90	0.90	0.89	0.89	0.88
	ESS$/n$	0.412	0.408	0.404	0.390	0.382
RAPS	MCOV	0.90	0.89	0.89	0.88	0.88
	ESS$/n$	0.624	0.621	0.620	0.615	0.614

Table 9: Sensitivity to the temperature parameter t on RxRx1 under the non-IID partition with APS.

Metric	$t = 0.0001$	$t = 0.001$	$t = 0.01$	$t = 0.1$	$t = 1$
MCOV	0.90 ± 0.01	0.90 ± 0.00	0.90 ± 0.01	0.92 ± 0.01	0.95 ± 0.01
MAPSS	155.29 ± 8.11	157.31 ± 7.13	160.40 ± 8.23	188.94 ± 13.45	280.17 ± 34.87

Table 10: Sensitivity to density-estimation resolution on RxRx1 under the non-IID partition with HPS.

Metric	$B = 100$	$B = 200$	$B = 400$	$B = 800$
MCOV	0.90 ± 0.00	0.90 ± 0.00	0.90 ± 0.00	0.90 ± 0.00
MAPSS	141.74 ± 7.51	141.84 ± 7.56	142.79 ± 5.10	143.04 ± 8.11

Table 11: Client-wise diagnostics on RxRx1 under the non-IID partition with HPS. CCOV and CAPSS denote client-wise coverage and client-wise average prediction set size.

Method	CCOV-mean	CCOV-std	CCOV-median	CAPSS-mean	CAPSS-std	CAPSS-median
FCP	0.62	0.18	0.61	15.02	1.82	15.03
FCPLS	0.72	0.08	0.74	292.12	7.76	292.44
FCPJS	0.89	0.10	0.91	138.57	28.83	139.89

C.3.9 Client-Wise Coverage Diagnostics

Although FCPJS guarantees marginal coverage over the federated mixture, it also substantially improves client-wise coverage diagnostics.

The formal target of FCPJS is marginal coverage over the federated mixture, not per-client conditional coverage. Nevertheless, client-wise coverage is useful for understanding deployment behavior under heterogeneous clients. Table 11 reports client-wise coverage and client-wise average prediction set size on RxRx1 non-IID with HPS. We denote client-wise coverage by CCOV and client-wise average prediction set size by CAPSS.

FCP severely under-covers individual clients, with CCOV-mean 0.62. FCPLS improves coverage but remains below

the target and requires very large prediction sets. FCPJS achieves CCOV-mean 0.89 and CCOV-median 0.91, while producing substantially smaller client-wise prediction sets than FCPLS. The nonzero CCOV standard deviation reflects client heterogeneity and does not contradict the formal guarantee, which is marginal rather than per-client conditional.

These diagnostics show that FCPJS improves client-level behavior in practice, while preserving the correct interpretation of its theoretical guarantee.