Efficient Federated Conformal Prediction with Group-Conditional Guarantee

Haifeng Wen[1] **Osvaldo Simeone**[2] **Hong Xing**[1,3]

[1]IoT Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[2]Institute for Intelligent Networked Systems (INSI), Northeastern University London, London, UK
[3]Department of ECE, The Hong Kong University of Science and Technology, HK SAR

Abstract

Deploying trustworthy AI systems requires principled uncertainty quantification. Conformal prediction (CP) is a widely used framework for constructing prediction sets with distribution-free coverage guarantees. In many practical settings, including healthcare, finance, and mobile sensing, the calibration data required for CP are distributed across multiple clients, each with its own local data distribution. In this federated setting, data can often be partitioned into, potentially overlapping, groups, which may reflect client-specific strata or cross-cutting attributes such as demographic or semantic categories. We propose *group-conditional* federated conformal prediction (GC-FCP), a federated extension of conditional conformal calibration for a target mixture over prespecified groups. GC-FCP constructs mergeable, atom-stratified coresets from local calibration scores, enabling compact aggregation at the server when the number of active atoms is moderate. Experiments on synthetic and real-world datasets validate the performance of GC-FCP compared to centralized calibration baselines. The code of our work can be found at `https://github.com/HaifengWen/GC-FCP`.

1 INTRODUCTION

Deploying trustworthy AI systems critically depends on uncertainty quantification to enable reliability control at deployment time. Given a pretrained model, *conformal prediction (CP)* post-processes the model's outputs to construct prediction sets with finite-sample, distribution-free coverage guarantees. The *calibration* of the prediction set is carried out offline by leveraging held-out calibration data [Vovk et al., 2005, Angelopoulos and Bates, 2023, Barber et al., 2021, Simeone and Romano, 2025]. In many practical deployments, however, calibration data are inherently distributed and subject to privacy constraints, so that each client must retain its data on-site, e.g., hospitals, banks, or Internet-of-Things devices [McMahan et al., 2017].

Federated conformal prediction (FCP) [Lu et al., 2023] addresses the outlined distributed setting with the goal of preserving marginal coverage with respect to a distribution obtained by mixing local data distributions. Pursuing a similar marginal coverage guarantee, Humbert et al. [2023] and Humbert et al. [2024] proposed one-shot FCP schemes based on a quantile-of-quantiles estimator. Beyond calibration targeting the mixture distribution, FCP-Pro [Li et al., 2026] and personalized FCP [Min et al., 2025] address the heterogeneity of local data distributions by training additional models during the calibration process, providing marginal coverage on local distributions or on the distribution of a new client. To address robustness, Kang et al. [2024] studied Byzantine clients that may report arbitrary statistics. Considering more general connectivity constraints underlying communications, decentralized calibration via message passing over arbitrary graphs was proposed by [Wen et al., 2025], and distributed remote calibration protocols were studied in [Zhu et al., 2025] for wireless sensor networks.

The state of the art on federated calibration summarized above did not target *group-conditional* coverage guarantees. In federated settings, data can often be partitioned into, potentially overlapping, groups, which may reflect client-specific strata or cross-cutting attributes such as demographic or semantic categories (see Fig. 1). For centralized settings, Mondrian CP [Vovk et al., 2003] addresses group-conditional coverage over disjoint groups of covariates, while CondCP [Gibbs et al., 2025] allows for overlapping groups by reframing conditional coverage as simultaneous coverage over a class of covariate shifts. Kandinsky CP [Bairaktari et al., 2025] extended Mondrian CP and CondCP to groups that depend on both covariates and labels. Other related research directions for centralized scenarios include localized coverage guarantees [Guan, 2023]; learn-

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:7290–7313.

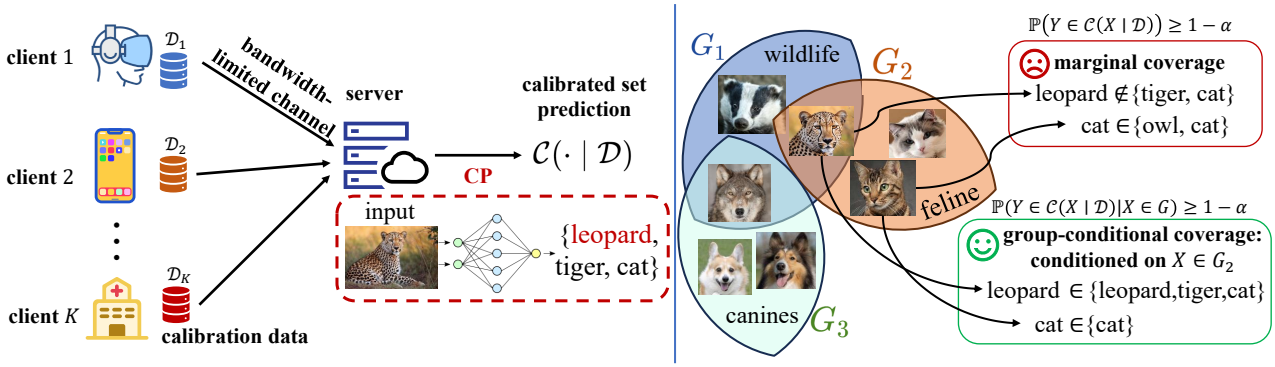


Figure 1: *Left*: In a federated system with heterogeneous clients, each client k holds local calibration data $\mathcal{D}_k$ and communicates over a bandwidth-limited channel to a central server. The server aggregates these summaries to perform conformal calibration (CP) and outputs a set-valued predictor $\mathcal{C}(\cdot \mid \mathcal{D})$. *Right*: GC-FCP targets group-conditional coverage for potentially overlapping groups $\mathcal{G} = \{G_1, G_2, G_3\}$, ensuring the inequality $\mathbb{P}(Y \in \mathcal{C}(X \mid \mathcal{D}) \mid X \in G) \geq 1 - \alpha$ for all groups $G \in \mathcal{G}$, whereas methods that only guarantee marginal coverage $\mathbb{P}(Y \in \mathcal{C}(X \mid \mathcal{D})) \geq 1 - \alpha$, such as FCP [Lu et al., 2023], may still under-cover within specific groups.

ing improved conformity scores to reduce conditional miscoverage [Xie et al., 2024]; as well as analyses of sample-conditional validity for split conformal methods [Duchi, 2025].

All this prior art on group-conditional CP has focused on centralized settings. To address this knowledge gap, in this paper, we propose a federated, atom-stratified extension of CondCP [Gibbs et al., 2025], termed *group-conditional federated conformal prediction (GC-FCP)*, which enables communication-efficient and computation-efficient prediction-set construction under mild conditions, such as when the overlap between groups is not too severe. Unlike FedCF [Srinivasan et al., 2025], GC-FCP does not treat different groups separately, seeking simultaneous coverage across all groups.

The main contributions are summarized as follows:

- We first propose *centralized GC-FCP*, an extension of CondCP [Gibbs et al., 2025] that achieves group-conditional coverage guarantees under a federated target mixture distribution.
- We develop *GC-FCP*, a federated protocol that compresses atom-stratified calibration scores into mergeable T-Digest [Dunning and Ertl, 2019] coresets, avoiding the loss of group-membership information caused by a single global sketch and double counting caused by overlapping group sketches.
- We establish group-conditional coverage bounds for GC-FCP, making it explicit how the coreset compression level affects the achieved coverage.
- We validate the proposed methods on synthetic and real-world benchmarks, demonstrating that GC-FCP attains group-conditional reliability while substantially reducing computational overhead.

2 PROBLEM SETTING

As illustrated in Fig. 1, we study a federated calibration setting that follows reference [Lu et al., 2023]. Accordingly, we consider a network consisting of K clients that communicate with a central server. In this setup, each client $k \in \{1, \ldots, K\} \triangleq [K]$ has access to a local calibration dataset $\mathcal{D}_k = \{(X_{i,k}, Y_{i,k})\}_{i=1}^{n_k}$, with data points $(X_{i,k}, Y_{i,k})$ drawn i.i.d. from a client-specific distribution P_k over $\mathcal{X} \times \mathcal{Y}$. We define a global calibration dataset as the union $\mathcal{D} = \bigcup_{k=1}^{K} \mathcal{D}_k$, with $n = \sum_{k=1}^{K} n_k$. As in [Lu et al., 2023], the goal is to calibrate a shared pre-trained model $f : \mathcal{X} \mapsto \mathcal{Y}$ through communication with the central server. Calibration is evaluated on test data (X_{n+1}, Y_{n+1}) drawn from the mixture

$$\mathrm{P} = \sum_{k=1}^{K} \pi_k \mathrm{P}_k, \tag{1}$$

for some arbitrary and known mixture weights $\pi_k \geq 0$ with $\sum_{k=1}^{K} \pi_k = 1$. In practice, the weights $\{\pi_k\}_{k=1}^{K}$ dictate the relative relevance of the data of each client k to the calibration of model f. The vector π should be interpreted as part of the deployment target, e.g., equal client, population, or policy-driven weighting, and it does not necessarily imply the empirical client-sample-size mixture.

Calibration aims at obtaining a set predictor

$$\mathcal{C}(\cdot \mid \mathcal{D}) : \mathcal{X} \to 2^{\mathcal{Y}}. \tag{2}$$

using the distributed calibration data via clients-to-server communication. Prior work [Lu et al., 2023] imposed the constraint that the prediction set (2) satisfies the marginal coverage condition

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1} \mid \mathcal{D})) \geq 1 - \alpha \tag{3}$$

with respect to the mixture distribution (1). In contrast, as explained next, we impose a more flexible conditional coverage condition based on grouping.

Let $\mathcal{G} \subseteq 2^{\mathcal{X}}$ be a finite collection of groups in the covariate space $\mathcal{X}$ and $\bigcup_{G\in\mathcal{G}} = \mathcal{X}$. Importantly, the groups are generally overlapping, i.e., there exist groups $G_i, G_j \in \mathcal{G}$ such that $G_i \bigcap G_j \neq \emptyset$ (see Fig. 1). Groups may or may not represent client-specific partitions. In fact, some groups G may correspond to data categories that are more representative of the distribution P_k of a given client k. However, it may also be that groups refer to categories that apply equally across all clients. We consider the set $\mathcal{G}$ to be arbitrary and given.

Given a target mis-coverage level $\alpha \in (0, 1)$, calibration aims to construct set prediction $\mathcal{C}(\cdot \mid \mathcal{D})$ such that, for every group $G \in \mathcal{G}$, the following group-conditional coverage requirement holds:

$$\mathbb{P}\left(Y_{n+1} \in \mathcal{C}(X_{n+1} \mid \mathcal{D}) \,|\, X_{n+1} \in G\right) \geq 1 - \alpha, \quad (4)$$

where the probability in (4) is taken over the calibration data in $\mathcal{D}$ and the test data (X_{n+1}, Y_{n+1}). For condition (4) to be meaningful, we assume that all groups $G \in \mathcal{G}$ satisfy the inequality $\mathbb{P}(X_{n+1} \in G) > 0$ under the mixture distribution (1). Note that this does not require that the inequality $\mathbb{P}_k(X \in G) > 0$ holds for every client k. Rather, it only requires that at least one component with probability $\pi_k > 0$ places positive mass on the group G.

The guarantee (4) applies to a prespecified finite group family. This may include public common groups for which all clients can compute the same membership vector. Private groups, as well as unknown or emergent groups, requiring separate discovery or auditing procedures, are outside the scope of this finite-group guarantee.

If the groups $\mathcal{G}$ were disjoint, the conditional coverage requirement (4) could be obtained by combining FCP [Lu et al., 2023] and Mondrian CP [Vovk et al., 2003]. Accordingly, one applies the FCP protocol separately for each group $G \in \mathcal{G}$. However, ensuring the conditions (4) become difficult when groups overlap, and we address this challenge in this paper.

3 PRELIMINARIES

To start, consider, for reference, a centralized split-conformal setting in which a server has access to i.i.d. calibration data $\{(X_i, Y_i)\}_{i=1}^n$ and a test covariate X_{n+1}. In this section, we review CondCP [Gibbs et al., 2025], which addresses the problem of satisfying the constraints (4) in such a centralized setting.

Let $s : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ be a score function derived from a predictive model f, and define the calibration scores $S_i = s(X_i, Y_i)$ with $i = 1, \ldots, n$. Furthermore, let $\ell_\alpha(\theta, S)$ denote the pinball loss at level $1 - \alpha$:

$$\ell_\alpha(\theta, S) = \begin{cases} (1-\alpha)(S-\theta), & S \geq \theta, \\ \alpha(\theta - S), & S < \theta. \end{cases} \quad (5)$$

Given the set of groups $\mathcal{G}$ (possibly overlapping), define the group-membership map $\Phi : \mathcal{X} \to \{0, 1\}^{|\mathcal{G}|}$ as

$$\Phi(x) = \big(\mathbf{1}\{x \in G\}\big)_{G\in\mathcal{G}} \in \{0, 1\}^{|\mathcal{G}|}, \quad (6)$$

which encodes membership of element $x \in \mathcal{X}$ belonging to each group of $\mathcal{G}$ as a binary indicator vector (where $\mathbf{1}\{\text{true}\} = 1$ and $\mathbf{1}\{\text{false}\} = 0$). Then, CondCP defines the set of functions

$$\mathcal{F}_\mathcal{G} = \left\{x \mapsto \beta^\top \Phi(x) : \beta \in \mathbb{R}^{|\mathcal{G}|}\right\}, \quad (7)$$

which corresponds to arbitrary linear combinations for the group indicators. Specifically, for any input test score $S \in \mathbb{R}$, CondCP solves the *augmented quantile regression* problem:

$$\hat{g}_S = \arg\min_{g\in\mathcal{F}_\mathcal{G}} \left\{ \frac{1}{n+1} \sum_{i=1}^{n} \ell_\alpha\big(g(X_i), S_i\big) + \frac{1}{n+1} \ell_\alpha\big(g(X_{n+1}), S\big) \right\}. \quad (8)$$

Note that the solution $\hat{g}_S$ is a function in class (7). By the properties of the pinball loss, the solution $\hat{g}_S$ represents a covariate (X_{n+1})-dependent version of the $(1-\alpha)$-empirical quantile of the augmented calibration scores $\{S_i\}_{i=1}^n \cup \{S\}$ [Gibbs et al., 2025]. The prediction set is then defined as the set of all labels $y \in \mathcal{Y}$ whose score $s(X_{n+1}, y)$ is no larger than the corresponding empirical quantile $\hat{g}_{s(X_{n+1},y)}(X_{n+1})$, i.e.,

$$\mathcal{C}(X_{n+1} \mid \mathcal{D}) = \left\{y \in \mathcal{Y} : s(X_{n+1}, y) \leq \hat{g}_{s(X_{n+1},y)}(X_{n+1})\right\}. \quad (9)$$

The prediction set (9) is constructed by addressing an equivalent formulation of the convex problem (8) via dual methods (see Appendix D for details). Under the assumption of i.i.d. calibration and test data, the set $\mathcal{C}(X_{n+1} \mid \mathcal{D})$ in (9) satisfies group-conditional coverage (4).

Communication overhead: Solving problem (8) at the server would require collecting all calibration scores and their group-membership features, yielding a communication overhead $\mathcal{O}(n)$.

Computational overhead: Solving problem (8) requires computational complexity of order $\mathcal{O}(n^{3/2}|\mathcal{G}|^2)$ using standard convex problem solvers [Renegar, 1988, Nesterov and Nemirovskii, 1994].

Table 1: Communication load and computational complexity. For GC-FCP, $\mathcal{A}_k^+$ is the set of non-empty atoms observed by client k, and $\mathcal{A}^+ = \bigcup_k \mathcal{A}_k^+$.

Method	Comm.	Comp.
CondCP & Centralized GC-FCP	$\mathcal{O}(n)$	$\mathcal{O}\left(n^{3/2}\lvert\mathcal{G}\rvert^2\right)$
GC-FCP	$\mathcal{O}\big(\delta \sum_k \lvert\mathcal{A}_k^+\rvert\big)$	$\mathcal{O}\big((\delta\lvert\mathcal{A}^+\rvert)^{3/2}\lvert\mathcal{G}\rvert^2\big)$

4 GROUP-CONDITIONAL FEDERATED CONFORMAL PREDICTION (GC-FCP)

This section introduces *group-conditional federated conformal prediction* (GC-FCP), a federated calibration procedure designed to achieve group-conditional coverage (4) over a prescribed, generally overlapping, collection of groups $\mathcal{G}$ under any mixture distribution (1). In GC-FCP, each client computes calibration scores locally and communicates to the server only a compact, group-stratified summary, while the server efficiently constructs the CP set for each test point without accessing raw client-level calibration scores.

4.1 FEDERATED AUGMENTED QUANTILE REGRESSION

To start, consider an ideal setting in which the entire calibration dataset $\mathcal{D}$, which includes the datasets $\mathcal{D}_k$ from all clients $k \in [K]$, is available at the central server. Assume also no computational limitations at the server. Even in this simplified setup, as explained in the previous section, CondCP would fail to guarantee the conditional coverage requirements (4), since the data points in dataset $\mathcal{D}$ are not i.i.d. with respect to the mixture distribution (1). We address this challenge in this section by proposing *centralized GC-FCP*, a centralized calibration scheme that guarantees the desired condition (4).

As in Section 3, let $s : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ be a score function derived from the shared predictive model f. We denote the calibration scores at client k as $S_{i,k} = s(X_{i,k}, Y_{i,k})$ with $i \in \{1, \ldots, n_k\}$ and $k \in [K]$. Furthermore, given a test point $(X_{n+1}, Y_{n+1}) \sim \mathrm{P}$, we denote its score by $S_{n+1} = s(X_{n+1}, Y_{n+1})$.

For a test covariate X_{n+1} and an input score value $S \in \mathbb{R}$, centralized GC-FCP solves the following *federated* augmented quantile regression estimator as a generalization of the CondCP problem (8):

$$\hat{g}_S = \operatorname*{argmin}_{g \in \mathcal{F}_\mathcal{G}} \sum_{k=1}^{K} \frac{\pi_k}{n_k + 1} \Bigg(\sum_{i=1}^{n_k} \ell_\alpha(g(X_{i,k}), S_{i,k}) + \ell_\alpha(g(X_{n+1}), S) \Bigg), \quad (10)$$

where the function set $\mathcal{F}_\mathcal{G}$ is defined in (7). The objective in problem (10) incorporates the scores of all K clients, with each k-th term weighted by the factor $\pi_k/(n_k + 1)$, thus matching the structure of the target mixture distribution (1). Using this definition, centralized GC-FCP obtains the prediction set as in (9).

Theorem 4.1 (Group-conditional coverage for centralized GC-FCP)**.** *For every group $G \in \mathcal{G}$, the set* (9) *with the empirical quantile* (10) *produced by centralized GC-FCP satisfies the conditional coverage condition* (4)*.*

Proof: See Appendix A.1.

As summarized in Table 1, the centralized GC-FCP shares the same communication and computational overheads as CondCP.

4.2 GROUP-CONDITIONAL FEDERATED CONFORMAL PREDICTION

The previous subsection focused on a centralized version of GC-FCP, which solves problem (10) using the calibration scores pooled from all clients. To mitigate the communication and computational overhead associated with this approach, we now introduce GC-FCP, a distributed protocol that replaces the full calibration dataset with mergeable atom-based coresets. As summarized in Table 1, this protocol reduces the communication load and enables a more efficient server-side optimization in conditions to be elaborated.

GC-FCP starts by applying a coreset construction based on T-Digest [Dunning and Ertl, 2019], which is a sketching algorithm for computing approximations of quantiles, and is elaborated in the sequel. The sketches produced by T-Digest are then used to solve the optimization problem (10) at the central server. While T-Digest was also used by FCP [Lu et al., 2023], GC-FCP must additionally account for the structure of the groups in the set $\mathcal{G}$, producing stratified structures (see Fig. 2).

4.2.1 T-Digest

Let $\{(R_i, w_i)\}_{i=1}^{\ell}$ be weighted real-valued samples with $R_i \in \mathbb{R}$ and weights $w_i > 0$, and define the total weight as $W = \sum_{i=1}^{\ell} w_i$. The associated weighted empirical cumulative distribution function (CDF) is $F(t) = \frac{1}{W} \sum_{i=1}^{\ell} w_i \, \mathbf{1}\{R_i \le t\}$, with generalized empirical quantile function $Q(u) = \inf\{t : F(t) \ge u\}$ for $u \in [0, 1]$.

T-Digest: A *T-Digest* [Dunning and Ertl, 2019] produces an ordered collection of $m < \ell$ clusters $\mathcal{R}_c \subseteq \{1, \ldots, \ell\}$, i.e., if $c_1 < c_2$, $R_i \le R_j$ for any $i \in \mathcal{R}_{c_1}$ and $j \in \mathcal{R}_{c_2}$. Cluster $\mathcal{R}_c$ is described by a cluster representative $\bar{R}_c$ and an aggregate weight W_c. Each cluster $\mathcal{R}_c$ includes a subset

of samples $\{R_i\}_{i=1}^{\ell}$ with total weight $W_c = \sum_{i\in\mathcal{R}_c} w_i$, and the weighted cluster means $\bar{R}_c = \frac{1}{W_c}\sum_{i\in\mathcal{R}_c} w_i R_i$. In summary, T-Digest returns the weights and the cluster means as

$$\mathsf{TD} = \left\{(\bar{R}_c, W_c)\right\}_{c=1}^{m}. \tag{11}$$

It is worth noting that T-Digest enforces the ordering $\bar{R}_1 \leq \bar{R}_2 \leq \cdots \leq \bar{R}_m$, producing ordered cluster means. To this end, it orders the original set $\{R_i\}_{i=1}^{\ell}$ such that we have $R_1 \leq R_2 \leq \cdots \leq R_\ell$. Clusters are then constructed greedily as follows. To elaborate, let $V_c = \sum_{i=1}^{c} W_i$ with $V_0 = 0$, and define the empirical left and right quantile boundaries of cluster $c \in \{1, \ldots, m\}$ as

$$q_c^{\mathrm{L}} = \frac{V_{c-1}}{W} \text{ and } q_c^{\mathrm{R}} = \frac{V_c}{W}. \tag{12}$$

Samples $R_1, R_2, \ldots, R_\ell$ are aggregated in an increasing order into the current cluster $\mathcal{R}_c$ as long as adding the next sample preserves the constraint

$$r\big(q_c^{\mathrm{R}}\big) - r\big(q_c^{\mathrm{L}}\big) \leq 1, \qquad c = 1, \ldots, m. \tag{13}$$

with scale function

$$r(q) = \frac{\delta}{2\pi}\arcsin(2q-1). \tag{14}$$

for some parameter $\delta > 0$ that controls the level of compression.

Intuitively, the condition (13) enforces finer resolution near the tails of the empirical distribution of the scalars $\{R_i\}_{i=1}^{\ell}$, while permitting larger clusters near the median. By this construction, the number of retained clusters scales as $m = \Theta(\delta)$ [Dunning and Ertl, 2019].

From the digest $\mathsf{TD} = \left\{(\bar{R}_c, W_c)\right\}_{c=1}^{m}$, the approximate quantile function for $u \in [0,1]$ is given by

$$\widehat{Q}(u) = \inf\{t : \widehat{F}(t) \geq u\}, \tag{15}$$

where

$$\widehat{F}(t) = \frac{1}{W}\sum_{c=1}^{m} W_c\, \mathbf{1}\{\bar{R}_c \leq t\}. \tag{16}$$

is an estimate of the empirical CDF $F(t)$. The quality of this estimate will be analyzed in Section 5.

Merging T-Digests: A key property of T-Digest is that digests can be merged and then re-compressed such that the number of retained clusters remains $\Theta(\delta)$ [Dunning and Ertl, 2019]. Let $\mathsf{TD}^{(j)} = \{(\bar{R}_c^{(j)}, W_c^{(j)})\}_{c=1}^{m_j}$ be digests obtained from disjoint datasets indexed by $j = 1, \ldots, J$. To merge them, one applies the same procedure discussed above to the pooled samples $\bigcup_{j=1}^{J} \mathsf{TD}^{(j)}$. We denote the merged digest as $\mathrm{Merge}(\mathsf{TD}^{(1)}, \ldots, \mathsf{TD}^{(J)})$.

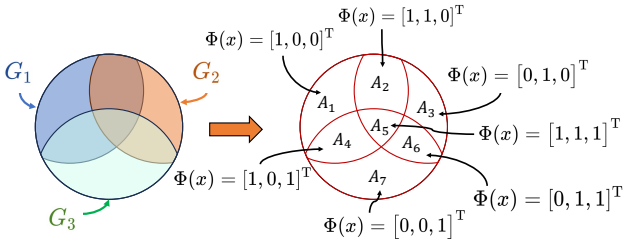


Figure 2: Illustration of the atom partitions applied by GC-FCP. Given the set of overlapping groups $\mathcal{G} = \{G_1, G_2, G_3\}$, the resulting 7 non-empty atoms $\mathcal{A} = \{A_1, \ldots, A_7\}$ are shown on the right, together with the corresponding group-membership vector (6).

4.2.2 GC-FCP

In order to address the group-conditional constraint (4), GC-FCP first stratifies the data at the clients into *disjoint atoms* based on unique group membership patterns. Independent T-Digests are then constructed per atom, and used at the server as coresets to approximate the solution of the quantile regression problem (10).

Atoms construction: As illustrated in Fig. 2, GC-FCP first partitions the input domain $\mathcal{X}$ into a collection $\mathcal{A} = \{A\}_{A\in\mathcal{A}}$ of atoms, which are defined as $\bigcup_{A\in\mathcal{A}} A = \mathcal{X}$ and $A \bigcap A' = \emptyset$, for any pair $A, A' \in \mathcal{A}$. The collection $\mathcal{A}$ corresponds to the smallest-cardinality partition of the set $\mathcal{X}$ such that all the sets $\mathcal{G}$ can be recovered via union operations on its atoms. Formally, for each subset of groups, $\mathcal{S} \subseteq \mathcal{G}$, define

$$A(\mathcal{S}) = \left(\bigcap_{G\in\mathcal{S}} G\right) \bigcap \left(\bigcap_{G\in\mathcal{G}\setminus\mathcal{S}} G^c\right) \tag{17}$$

as the intersection of all groups in subset $\mathcal{S}$ and the complement of all groups not in subset $\mathcal{S}$. The construction $\mathcal{A} = \{A(\mathcal{S})\}_{\mathcal{S}\subseteq\mathcal{G}}$ yields the collection of atoms [Kallenberg, 2002]. Although arbitrary overlaps can induce up to $2^{|\mathcal{G}|}$ atoms, GC-FCP is efficient for many practical group settings. For example, for one-dimension interval groups, the number of atoms scales linearly with $|\mathcal{G}|$.

By construction, each atom A corresponds to all points $x \in \mathcal{X}$ with the same membership vector $\Phi(x)$. Accordingly, for any $x \in A$, we can write $\Phi_A = \Phi(x)$. An example is illustrated in Fig. 2.

Local-score partition and digest: We now partition the set of local scores at client $k \in [K]$ into $|\mathcal{A}|$ subsets $\mathcal{D}_{k,A}$, one per atom A as

$$\mathcal{D}_{k,A} = \{s(X,Y) : X \in A,\ (X,Y) \in \mathcal{D}_k\}. \tag{18}$$

Define the active local and global atom sets as

$$\mathcal{A}_k^+ = \{A \in \mathcal{A} : \mathcal{D}_{k,A} \neq \emptyset\}, \qquad \mathcal{A}^+ = \bigcup_{k=1}^{K} \mathcal{A}_k^+, \quad (19)$$

which satisfy the inequalities $|\mathcal{A}^+| \leq \min\{2^{|\mathcal{G}|}, n\}$ and $|\mathcal{A}_k^+| \leq \min\{2^{|\mathcal{G}|}, n_k\}$. Empty atoms produce no digest. Each client $k \in [K]$ builds a separate digest $\mathsf{TD}_{k,A}$ upon the scores $\mathcal{D}_{k,A}$ for each active atom $A \in \mathcal{A}_k^+$ with equal weights $w = \pi_k/(n_k+1)$. Following Section 4.2, the resulting digest $\mathsf{TD}_{k,A}$ consists of a set of means and weights with $m_{k,A} = \Theta(\delta)$ clusters

$$\mathsf{TD}_{k,A} = \left\{\left(\bar{S}_{k,A,c},\, W_{k,A,c}\right)\right\}_{c=1}^{m_{k,A}}, \quad (20)$$

where $\bar{S}_{k,A,c} = \frac{1}{W_{k,A,c}} \sum_{i \in \mathcal{S}_{k,A,c}} \frac{\pi_k}{n_k+1} S_i$ is the weighted mean of the scores in cluster $\mathcal{S}_{k,A,c}$, and $W_{k,A,c} = |\mathcal{S}_{k,A,c}| \frac{\pi_k}{n_k+1}$ is the aggregated weight.

Next, client $k \in [K]$ transmits the digest $\mathsf{TD}_{k,A}$ along with the identifier of active atom $A \in \mathcal{A}_k^+$ to the server. For each atom $A \in \mathcal{A}^+$, the server merges the received digests $\{\mathsf{TD}_{k,A}\}_{k=1}^K$ to obtain a global digest

$$\begin{aligned} \mathsf{TD}_A &= \mathrm{Merge}\big(\mathsf{TD}_{1,A}, \ldots, \mathsf{TD}_{K,A}\big) \\ &= \left\{\left(\bar{S}_{A,c},\, W_{A,c}\right)\right\}_{c=1}^{m_A}, \end{aligned} \quad (21)$$

with $m_A = \Theta(\delta)$ clusters, where $\mathrm{Merge}(\cdot)$ denotes the merge operation on $\{\mathsf{TD}_{k,A}\}_{k=1}^K$. As a result, the union over the digests of all atoms yields the final coreset:

$$\widetilde{\mathcal{D}} = \left\{\left(A, \bar{S}, W\right) : \left(\bar{S},\, W\right) \in \mathsf{TD}_A,\ A \in \mathcal{A}\right\} \quad (22)$$

with $|\widetilde{\mathcal{D}}| = \Theta(\delta|\mathcal{A}^+|)$. Note that each entry $(A, \bar{S}, W)$ of the coreset includes the identifier of the atom A, the mean score $\bar{S}$, and the corresponding weight W. The uncompressed federated implementation would transmit all atom-stratified scores, and solve the same objective as centralized GC-FCP, thus requiring $\mathcal{O}(n)$ score records plus atom identifiers. GC-FCP replaces the full dataset with the atom-stratified mergeable coreset (22).

Set construction: Given a test score S, GC-FCP solves problem (10) using the coreset $\widetilde{\mathcal{D}}$ instead of the original pooled data $\mathcal{D}$, i.e.,

$$\begin{aligned} \tilde{\beta}(S) = \underset{\beta \in \mathbb{R}^{|\mathcal{G}|}}{\mathrm{argmin}} & \sum_{(A,\bar{S},W) \in \widetilde{\mathcal{D}}} W\, \ell_\alpha\left(\beta^\top \Phi_A,\, \bar{S}\right) \\ & + \left(\sum_{k=1}^{K} \frac{\pi_k}{n_k+1}\right) \ell_\alpha\left(\beta^\top \Phi(X_{n+1}),\, S\right). \end{aligned} \quad (23)$$

Finally, the prediction set is constructed as

$$\begin{aligned} &\mathcal{C}(X_{n+1} \mid \widetilde{\mathcal{D}}) \\ &\quad = \left\{y \in \mathcal{Y} : s(X_{n+1}, y) \leq \tilde{g}_{s(X_{n+1},y)}(X_{n+1})\right\}. \end{aligned} \quad (24)$$

Algorithm 1 GC-FCP

Input: Clients' calibration sets $\{\mathcal{D}_k\}_{k=1}^K$; groups $\mathcal{G}$; score $s(\cdot,\cdot)$; mis-coverage level α; mixture weights π_k; T-Digest compression parameter δ.

▷ `Client side (in parallel):`

for device $k \in \{1, \ldots, K\}$ **do**

Construct T-Digest $\mathsf{TD}_{k,A}$ for each active atom $A \in \mathcal{A}_k^+$ using $\mathcal{D}_{k,A}$.

Transmit T-Digests $\{\mathsf{TD}_{k,A}\}_{A \in \mathcal{A}_k^+}$ to the central server.

end for

▷ `Server side:`

Merge received digests for each active atom $A \in \mathcal{A}^+$ via (21).

Solve (23) and compute the prediction set (24).

Output: $\mathcal{C}(\cdot \mid \widetilde{\mathcal{D}})$

where $\tilde{g}_S(x) = \tilde{\beta}(S)^\top \Phi(x)$.

The proposed GC-FCP is summarized in Algorithm 1.

Communication overhead: Solving problem (23) at the server requires collecting active atom digests, yielding total communication load of order $\mathcal{O}(\delta \sum_{k=1}^K |\mathcal{A}_k^+|)$, compared to order $\mathcal{O}(n)$ for score-based communication of CondCP or centralized GC-FCP.

Computational overhead: On the server side, solving the dual of problem (23) requires computational complexity of order $\mathcal{O}(|\widetilde{\mathcal{D}}|^{3/2}|\mathcal{G}|^2) = \mathcal{O}\big((\delta|\mathcal{A}^+|)^{3/2}|\mathcal{G}|^2\big)$ by standard convex problem solvers [Renegar, 1988, Nesterov and Nemirovskii, 1994], compared to the order $\mathcal{O}(n^{3/2}|\mathcal{G}|^2)$ of CondCP or centralized GC-FCP. On client $k \in [K]$, membership computation costs $\mathcal{O}(n_k|\mathcal{G}|)$.

Memory overhead: As the digest memory scales as $\mathcal{O}(\delta|\mathcal{A}_k^+|)$, the server needs to store $\mathcal{O}(\delta|\mathcal{A}^+|)$ merged centroids.

Trust and Privacy: GC-FCP keeps raw calibration examples local, but does not provide any formal privacy guarantee. In particular, active membership patterns, atom-based counts, centroid weights, and approximate score distributions etc. may be exposed by T-Digests. The present analysis therefore assumes a trustworthy server. Privacy-preserving extensions possibly incorporate secure aggregation or differentially private quantile sketches [Dwork and Roth, 2014, Alabi et al., 2022] with potential compromise to coverage, which is left for future work.

5 COVERAGE GUARANTEES OF GC-FCP

The key property of GC-FCP is its ability to provide group-conditional coverage guarantees (4) for the prediction set. Deriving finite-sample group-conditional guarantees for GC-FCP is technically challenging, since one must simultane-

ously account for the non-exchangeability of the samples in the coreset and for the approximation error introduced by sketching via the local digest. To this end, we first analyze the quantile estimation accuracy of T-Digest, and then provide a group-conditional coverage bound for GC-FCP.

5.1 QUANTILE ACCURACY OF T-DIGEST

We start by deriving a relationship between the compression parameter δ used in the scale function (14) by T-Digest for compression, which dictates the number of clusters $m = \Theta(\delta)$, and the accuracy of the approximate quantile (15).

Lemma 5.1 (Uniform bound of T-Digest)**.** *Defining as $F(t)$ the true CDF of original samples, the CDF estimate* (16) *produced by T-Digest satisfies the uniform accuracy bound*

$$\sup_{t\in\mathbb{R}}\left|F(t)-\widehat{F}(t)\right| \leq \sin\left(\frac{\pi}{\delta}\right), \tag{25}$$

with $\delta \geq 2$. Moreover, for all $u \in [0,1]$, the quantile estimate (16) *satisfies the inequality $|F(\widehat{Q}(u)) - u| \leq \sin(\pi/\delta)$.*

Proof. See Appendix B.1. □

5.2 GROUP-CONDITIONAL COVERAGE GUARANTEES OF GC-FCP

Based on properties of T-Digest presented in Lemma 5.1, we can now prove the group-conditional coverage guarantees of GC-FCP.

Theorem 5.1 (Group-conditional coverage guarantees for GC-FCP)**.** *For each group $G \in \mathcal{G}$, the prediction set $\mathcal{C}(X_{n+1} \mid \widetilde{\mathcal{D}})$ produced by GC-FCP satisfies the inequality*

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1} \mid \widetilde{\mathcal{D}}) \mid X_{n+1} \in G) \geq 1-\alpha-\sin\left(\frac{\pi}{\delta}\right). \tag{26}$$

Proof: See Appendix B.2.

While Theorem 5.1 assumes that problems (23) are solved exactly, the next result accounts for suboptimality induced by numerically solving problem (23).

An upper bound of the group-conditional coverage of GC-FCP can be found in Appendix C.

6 EXPERIMENTS

We evaluate GC-FCP on (*i*) a synthetic regression benchmark [Romano et al., 2019]; (*ii*) CIFAR-10 image classification [Krizhevsky and Hinton, 2009]; and (*iii*) PathMNIST medical image classification [Yang et al., 2023]. Across all experiments, we compare the proposed scheme, GC-FCP, to the following benchmarks: (*i*) **vanilla centralized CP**, which provides the marginal guarantee (3); (*ii*) **FedCP** [Lu et al., 2023], which also satisfies (3); (*iii*) **centralized CondCP** [Gibbs et al., 2025]; and (*iv*) **centralized GC-FCP**, which is described in Section 4.1. CondCP and centralized GC-FCP are only evaluated on the small-scale synthetic dataset and CIFAR-10 due to their prohibitive computational complexity (see Table 1). Appendix E.2 further reports a large-scale ImageNet-1K experiment with federated and group-wise baselines, and additional ablation experiments over group settings and compression.

For each group $G \in \mathcal{G}$, we estimate group-conditional coverage on a test set $\mathcal{T}$ as $\widehat{\text{cov}}(G) = 1/|\mathcal{T}_G| \sum_{(x,y)\in\mathcal{T}_G} \mathbf{1}\{y \in \mathcal{C}(x \mid \mathcal{D})\}$, where $\mathcal{T}_G = \{(x,y) \in \mathcal{T} : x \in G\}$, and report the empirical coverage $\widehat{\text{cov}}(G)$ for all groups $G \in \mathcal{G}$. For classification, we also report the average prediction set size $(1/|\mathcal{T}|)\sum_{(x,y)\in\mathcal{T}} |\mathcal{C}(x \mid \mathcal{D})|$. Computational complexity is evaluated by the average wall-clock time required for each method to construct a prediction set on the same platform.

6.1 SYNTHETIC REGRESSION

For the synthetic regression task, we consider $K = 4$ clients with heterogeneous covariate distributions $P_{X,k}$ given by truncated normal distributions on the interval $[0,5]$ with mean $\mu_k = 0.5 + \frac{4(k-1)}{K-1}$ and variance $\sigma_k = 0.5 + 0.1(k-1)$.

Following [Romano et al., 2020], we generate responses as

$$Y_k \sim \text{Pois}(\sin^2(X)+0.1) + 0.03X\epsilon_1 + 25\mathbf{1}\{U < 0.01\}\epsilon_2 + \mathcal{N}(0, 0.01k^2), \tag{27}$$

with independent variables $\epsilon_1, \epsilon_2 \sim \mathcal{N}(0,1)$ and $U \sim \text{Unif}([0,1])$. We train a centralized linear regression model $f(\cdot)$ on a separate training dataset generated in the same way, and use the score $s(x,y) = |y - f(x)|$. Groups are given by overlapping intervals $\mathcal{G} = \{[0,2],[1,3],[2,4],[3,5]\}$. The miscoverage level α is set to $\alpha = 0.1$. Other parameters are summarized in Table 4 in Appendix E.1.

Fig. 3 (a)-(b) visualizes representative prediction sets, while Fig. 3 reports the mis-coverage rate across the four overlapping interval groups. Centralized CP and FedCP apply a single global threshold and therefore exhibit uneven group-wise miscoverage rates. In particular, some groups are under-covered (mis-coverage above α), reflecting shifts in the conditional score distribution across covariate regions. In contrast, CondCP, centralized GC-FCP, and GC-FCP yield substantially more uniform group-wise mis-coverage near the target level α across all groups (Fig. 3(c)) by adapting the threshold over the covariate space (Fig. 3(b)).

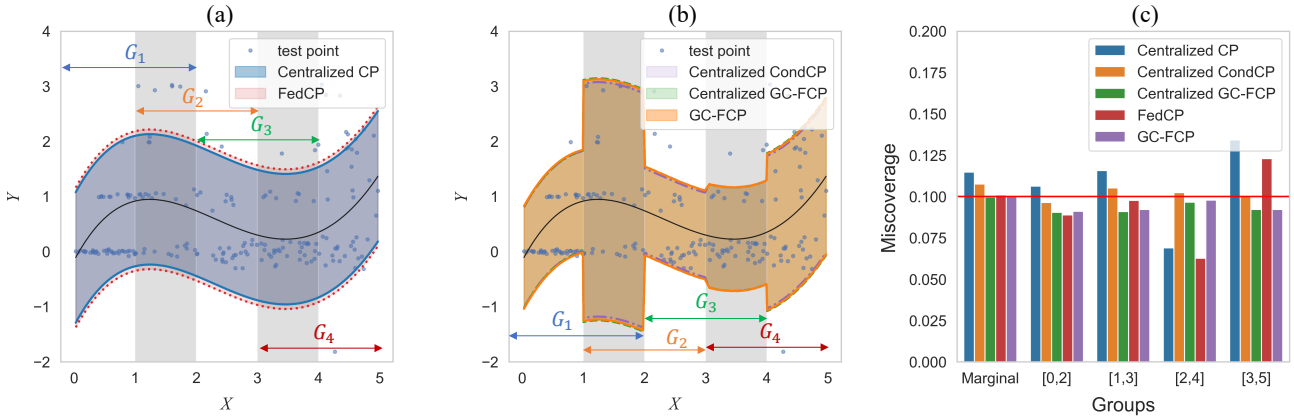


Figure 3: Visualization of prediction sets for the synthetic regression task for (a) centralized CP and FedCP [Lu et al., 2023], as well as for (b) centralized CondCP, centralized GC-FCP, and GC-FCP. (c) Per-group miscoverage rate.

Table 2: Coverage, set size, and computational load comparisons on CIFAR-10 [Krizhevsky and Hinton, 2009]. Each entry reports mean $\pm$ standard error over Monte Carlo repetitions. Standard errors for set size below 0.01 are omitted. A red star * indicates coverage below the target $1-\alpha = 0.9$.

Methods	**Marginal coverage**	**Group-conditional coverage (set size)**				**Comp. speedup**
		G_1	G_2	G_3	G_4	
Centralized CP	0.901 ± 0.001	$0.879^\star \pm 0.001$ (0.91)	$0.855^\star \pm 0.001$ (0.90)	0.906 ± 0.001 (0.93)	0.936 ± 0.001 (0.95)	N/A
Centralized CondCP	0.903 ± 0.001	0.901 ± 0.001 (0.96)	0.901 ± 0.001 (0.98)	0.902 ± 0.001 (0.94)	0.901 ± 0.001 (0.91)	1.00×
Centralized GC-FCP	0.906 ± 0.001	0.903 ± 0.001 (0.96)	0.903 ± 0.001 (0.99)	0.905 ± 0.001 (0.94)	0.905 ± 0.001 (0.92)	1.00×
FedCP	0.902 ± 0.001	$0.880^\star \pm 0.001$ (0.92)	$0.856^\star \pm 0.001$ (0.90)	0.906 ± 0.001 (0.93)	0.936 ± 0.001 (0.95)	N/A
GC-FCP ($\delta = 25$)	0.911 ± 0.001	$0.896^\star \pm 0.002$ (0.95)	$0.887^\star \pm 0.002$ (0.96)	0.919 ± 0.002 (0.95)	0.931 ± 0.002 (0.95)	**34.36**×
GC-FCP ($\delta = 250$)	0.906 ± 0.001	0.903 ± 0.001 (0.96)	0.902 ± 0.001 (0.98)	0.905 ± 0.001 (0.94)	0.906 ± 0.001 (0.92)	**19.73**×
GC-FCP ($\delta = 2500$)	0.906 ± 0.001	0.903 ± 0.001 (0.96)	0.903 ± 0.001 (0.99)	0.905 ± 0.001 (0.94)	0.905 ± 0.001 (0.92)	**3.96**×

6.2 CIFAR-10 EXPERIMENTS

We now adopt a pre-trained ResNet56 model $f(\cdot)$ for the CIFAR-10 image classification task with score function $s(x, y) = 1 - [f(x)]_y$, where y denotes the y-th entry of the softmax result [Sadinle et al., 2019]. We consider $K = 5$ clients with a non-i.i.d. label partition, so that each client holds $10/K$ disjoint classes. We set uniform mixture weights $\pi_k = 1/K$, and total calibration size $n = \sum_k n_k = 5000$. Groups are defined by overlapping predicted label classes $\hat{y}(x) = \arg\max_y [f(x)]_y$: $G_1 = \{x : \hat{y}(x) \in \{0, 1, 2, 3\}\}, G_2 = \{x : \hat{y}(x) \in \{2, 3, 4, 5\}\}, G_3 = \{x : \hat{y}(x) \in \{4, 5, 6, 7\}\}$, and $G_4 = \{x : \hat{y}(x) \in \{6, 7, 8, 9\}\}$. We set the miscoverage level $\alpha = 0.1$ and use a pre-trained ResNet56 to serve the model $f(\cdot)$ to construct the score function $s(x, y) = 1 - [f(x)]_y$ [Sadinle et al., 2019]. The report results represent the average after 50 Monte Carlo simulations, each with 5000 test points.

Table 2 reports marginal coverage, group-conditional coverage (with average set size in brackets), and computational speedup ("Comp. speedup"), with the latter being normalized with respect to the complexity of centralized CondCP. Centralized CP and FedCP attain the same marginal coverage near 0.9, but their group-conditional coverage varies substantially across groups (e.g., below 0.9 for groups G_1 and G_2), indicating that marginal calibration does not control errors uniformly across overlapping groups. Centralized CondCP and centralized GC-FCP achieve near-uniform group-conditional coverage across all groups, matching the intended behavior of augmented quantile calibration under grouping conditions.

GC-FCP closely tracks the centralized group-conditional performance while substantially improving computational

Table 3: Coverage and set size comparisons on PathMNIST. Each entry reports mean $\pm$ standard error over Monte Carlo repetitions. Standard errors of set size below 0.01 are omitted. A red star * indicates the coverage below the target $1-\alpha = 0.9$.

Methods	**Group-conditional coverage (set size)**				
(marginal coverage)	G_1	G_2	G_3	G_4	G_5
Centralized CP	$0.888^\star \pm 0.001$	$0.885^\star \pm 0.001$	$0.879^\star \pm 0.001$	0.935 ± 0.001	$0.895^\star \pm 0.001$
(0.901 ± 0.001)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)
FedCP	$0.892^\star \pm 0.001$	$0.889^\star \pm 0.001$	$0.883^\star \pm 0.001$	0.937 ± 0.001	0.900 ± 0.001
(0.905 ± 0.001)	(1.01)	(1.01)	(1.02)	(1.01)	(1.02)
GC-FCP ($\delta = 25$)	$0.891^\star \pm 0.002$	$0.890^\star \pm 0.003$	$0.880^\star \pm 0.003$	$0.895^\star \pm 0.004$	$0.853^\star \pm 0.003$
$(0.876^\star \pm 0.002)$	(1.12)	(1.14 ± 0.01)	(1.15 ± 0.09)	(1.07 ± 0.09)	(1.01 ± 0.05)
GC-FCP ($\delta = 250$)	0.913 ± 0.001	0.910 ± 0.001	0.911 ± 0.001	0.911 ± 0.001	0.898 ± 0.001
(0.908 ± 0.001)	(1.07)	(1.08)	(1.11)	(0.98)	(1.03)
GC-FCP ($\delta = 2500$)	0.913 ± 0.001	0.910 ± 0.001	0.910 ± 0.001	0.912 ± 0.001	0.900 ± 0.001
(0.909 ± 0.001)	(1.07)	(1.08)	(1.11)	(0.98)	(1.03)

efficiency. The computational gains are pronounced: depending on δ, GC-FCP yields from $3.96\times$ to $34.36\times$ per-test-point speedup relative to centralized CondCP, illustrating the accuracy–efficiency trade-off governed by the digest compression parameter. The value of the compression parameter affects coverage. For instance, with aggressive compression ($\delta = 25$), some groups are slightly under-covered (e.g., G_2), in a manner consistent with Theorem 5.1.

6.3 MEDICAL DATASET EXPERIMENTS

Finally, we evaluate GC-FCP on PathMNIST from MedMNIST (9 tissue classes) with images resized to $3 \times 28 \times 28$ [Yang et al., 2023]. We consider $K = 5$ clients, each holding samples from 2 disjoint classes, except the last client holds 1 class. We train a centralized CNN $f(\cdot)$ on the training split. We randomly shuffle the mixture validation and test data, split it into calibration and conformal test datasets equally, and allocate calibration samples to clients as in the CIFAR-10 experiments. Groups are defined by predicted-label classes: $G_1 = \{x : \hat{y}(x) \in \{0, 1, 2\}\}, G_2 = \{x : \hat{y}(x) \in \{1, 2, 3\}\}, G_3 = \{x : \hat{y}(x) \in \{2, 3, 4\}\}, G_4 = \{x : \hat{y}(x) \in \{3, 4, 5\}\}, G_5 = \{x : \hat{y}(x) \in \{4, 5, 6, 7, 8\}\}$. Other parameters remain the same as CIFAR-10 experiments.

Table 3 shows that Centralized CP and FedCP achieve marginal coverage near 0.9, but can deviate noticeably at the group level (e.g., lower coverage on G_3). GC-FCP with moderate compression ($\delta = 250$ or $\delta = 2500$) attains group-conditional coverage close to the target for all groups, while maintaining small average set sizes (near one label on average). In contrast, overly aggressive compression ($\delta = 25$) degrades calibration. Overall, these results corroborate that GC-FCP achieves group-conditional reliability in heterogeneous federated classification tasks, with accuracy controlled by the digest compression parameter δ.

7 CONCLUSION

We have introduced *Group-Conditional Federated Conformal Prediction (GC-FCP)*, a principled framework for constructing prediction sets with group-conditional guarantee in heterogeneous federated settings. GC-FCP leverages T-Digest to compress atom-stratified calibration scores into a mergeable coreset, achieving communication and computation efficiency when the number of active atoms is moderate. Under mild assumptions, we established group-conditional coverage guarantees for GC-FCP, and empirical results on synthetic regression and image classification benchmarks corroborated these findings while demonstrating substantial computational speedups.

GC-FCP assumes a prespecified finite group family and a trustworthy server for atom-based summaries. Future work includes extending the framework to richer conditional structures beyond finite group families, developing privacy-preserving sketches, as well as investigating robustness against more complex settings such as decentralized calibration [Wen et al., 2025], adversarial behavior [Kang et al., 2024], and online CP [Angelopoulos et al., 2024, Gasparin and Ramdas, 2024].

Acknowledgements

The work of H. Xing was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515010123, and in part by the Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things under Grant 2023B1212010007. The work of O. Simeone was supported by the European Research Council (ERC) under the European Union's Horizon Europe Programme (grant agreement No. 101198347), by an Open Fellowship of the EPSRC (EP/W024101/1), and by the EPSRC project (EP/X011852/1).

References

Daniel Alabi, Omri Ben-Eliezer, and Anamay Chaturvedi. Bounded space differentially private quantiles. *arXiv preprint arXiv:2201.03380*, 2022. URL https://arxiv.org/abs/2201.03380.

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101. URL https://doi.org/10.1561/2200000101.

Anastasios N. Angelopoulos, Stephen Bates, Michael I. 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 *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 1616–1630, Jul. 2024.

Konstantina Bairaktari, Jiayun Wu, and Steven Wu. Kandinsky conformal prediction: Beyond class- and covariate-conditional coverage. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 2581–2602, 13–19 Jul 2025.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

John Duchi. Sample-conditional coverage in split-conformal prediction. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Ted Dunning and Otmar Ertl. Computing extremely accurate quantiles using t-Digests. *arXiv preprint arXiv:1902.04023*, 2019. URL https://arxiv.org/abs/1902.04023.

Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*, volume 9 of *Foundations and Trends in Theoretical Computer Science*. Now Publishers, 2014.

Matteo Gasparin and Aaditya Ramdas. Conformal online model aggregation. *arXiv preprint arXiv:2403.15527*, 2024. URL https://arxiv.org/abs/2403.15527.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4):1100–1126, 2025. doi: 10.1093/jrsssb/qkaf008. URL https://doi.org/10.1093/jrsssb/qkaf008.

Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Pierre Humbert, Batiste Le Bars, Aurélien Bellet, and Sylvain Arlot. One-shot federated conformal prediction. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 14153–14177, Jul. 2023.

Pierre Humbert, Batiste Le Bars, Aurélien Bellet, and Sylvain Arlot. Marginal and training-conditional guarantees in one-shot federated conformal prediction. *arXiv preprint arXiv:2405.12567*, 2024. URL https://arxiv.org/abs/2405.12567.

Olav Kallenberg. *Foundations of Modern Probability*. Springer, New York, 2 edition, 2002.

Mintong Kang, Zhen Lin, Jimeng Sun, Cao Xiao, and Bo Li. Certifiably Byzantine-robust federated conformal prediction. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 23022–23057, Jul. 2024.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, Canada, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Guorui Li, Yanhui Zhang, Ying Wang, and Cong Wang. FCP-Pro: Federated conformal prediction algorithm based on prototype similarity. *Pattern Recognition*, 172:112514, 2026. doi: 10.1016/j.patcog.2025.112514. URL https://doi.org/10.1016/j.patcog.2025.112514.

Charles Lu, Yaodong Yu, Sai Praneeth Karimireddy, Michael Jordan, and Ramesh Raskar. Federated conformal predictors for distributed uncertainty quantification. In *Proceedings of the 40th International Conference on*

Machine Learning, volume 202, pages 22942–22964, Jul. 2023.

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*, 2025. URL https://openreview.net/forum?id=QQUhGPST45.

Yurii Nesterov and Arkadii Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994. doi: 10.1137/1.9781611970791.

Vincent Plassier, Mehdi Makni, Aleksandr Rubashevskii, Eric Moulines, and Maxim Panov. Conformal prediction for federated uncertainty quantification under label shift. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 27907–27947, Jul. 2023.

James Renegar. A polynomial-time algorithm, based on Newton's method, for linear programming. *Mathematical Programming*, 40(1):59–93, 1988. doi: 10.1007/BF01580724.

Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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.

Osvaldo Simeone and Yaniv Romano. Uncertainty quantification and data efficiency in AI: An information-theoretic perspective. *arXiv preprint arXiv:2512.05267*, 2025. URL https://arxiv.org/abs/2512.05267.

Anutam Srinivasan, Aditya T Vadlamani, Amin Meghrazi, and Srinivasan Parthasarathy. FedCF: Fair federated conformal prediction. *arXiv preprint arXiv:2509.22907*, 2025. URL https://arxiv.org/abs/2509.22907.

Vladimir Vovk, David Lindsay, Ilia Nouretdinov, and Alex Gammerman. Mondrian confidence machine. *Technical Report*, 2003.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Haifeng Wen, Hong Xing, and Osvaldo Simeone. Distributed conformal prediction via message passing. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=R7sAZDq0M9.

Ran Xie, Rina Barber, and Emmanuel Candes. Boosted conformal prediction intervals. *Advances in Neural Information Processing Systems*, 37:71868–71899, 2024.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1):41, 2023. doi: 10.1038/s41597-022-01721-8. URL https://doi.org/10.1038/s41597-022-01721-8.

Meiyi Zhu, Matteo Zecchin, Sangwoo Park, Caili Guo, Chunyan Feng, Petar Popovski, and Osvaldo Simeone. Conformal distributed remote inference in sensor networks under reliability and communication constraints. *IEEE Transactions on Signal Processing*, 73:1485–1500, 2025. doi: 10.1109/TSP.2025.3549222.

Efficient Federated Conformal Prediction with Group-Conditional Guarantee (Supplementary Material)

Haifeng Wen[1] **Osvaldo Simeone**[2] **Hong Xing**[1,3]

[1]IoT Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[2]Institute for Intelligent Networked Systems (INSI), Northeastern University London, London, UK
[3]Department of ECE, The Hong Kong University of Science and Technology, HK SAR

A PROOF OF SECTION 4

A.1 PROOF OF THEOREM 4.1

Proof. First assume there are no ties, i.e., $S_{i,k} \neq \hat{g}_{S_{n+1}}(X_{i,k})$ for all $i \in [n_k]$, $k \in [K]$, and $S_{n+1} \neq \hat{g}_{S_{n+1}}(X_{n+1})$. The tie case is handled at the end of the proof. Recall that $g(x) = \beta^\top \Phi(x)$ with $\beta \in \mathbb{R}^{|\mathcal{G}|}$ and $\Phi(\cdot)$ defined in (6), then the optimal function $\hat{g}_{S_{n+1}}(\cdot)$ solved by (10) reduces to a real-value vector given by

$$\beta^* = \underset{\beta \in \mathbb{R}^{|\mathcal{G}|}}{\operatorname{argmin}} \sum_{k=1}^{K} \left(\frac{\pi_k}{n_k+1} \sum_{i=1}^{n_k+1} \ell_\alpha \Big(\sum_{G \in \mathcal{G}} \beta_G \mathbf{1}\{X_{i,k} \in G\}, S_{i,k} \Big) \right), \tag{28}$$

where we denote $S_{n_k+1,k} = S_{n+1}$ for simplicity. Under the no-ties assumption, the first-order optimality condition implies that for every group $G \in \mathcal{G}$,

$$\sum_{k=1}^{K} \frac{\pi_k}{n_k+1} \sum_{i=1}^{n_k+1} \Big(\mathbf{1}\{X_{i,k} \in G\} \left(\mathbf{1}\{S_{i,k} < \theta^*_{i,k}\} - (1-\alpha) \right) \Big) = 0, \quad \forall G \in \mathcal{G} \tag{29}$$

where $\theta^*_{i,k} = \sum_{G \in \mathcal{G}} \beta^*_G \mathbf{1}\{X_{i,k} \in G\}$.

Let E_k be the event that (X_{n+1}, Y_{n+1}) is drawn from P_k. Conditioned on E_k, the (n_k+1) scores $\{S_{i,k}\}_{i=1}^{n_k+1}$ are exchangeable. Let $\mathcal{E}$ be the event that, for every $k \in [K]$, there exists a permutation σ_k such that

$$\big(S_{\sigma_k(1),k}, \ldots, S_{\sigma_k(n_k+1),k}\big) = \big(s_{1,k}, \ldots, s_{n_k+1,k}\big), \tag{40}$$

where $(s_{1,k}, \ldots, s_{n_k+1,k})$ denotes the realized values of $(S_{1,k}, \ldots, S_{n_k+1,k})$. Then, we have, for all $G \in \mathcal{G}$,

$$\begin{aligned}
&\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1} \mid \mathcal{D}) \mid X_{n+1} \in G, \mathcal{E}) \\
&= \frac{\mathbb{P}(S_{n+1} \leq \theta^*_{n+1}, X_{n+1} \in G \mid \mathcal{E})}{\mathbb{P}(X_{n+1} \in G \mid \mathcal{E})} \\
&= \frac{\sum_{k=1}^{K} \pi_k \mathbb{P}(S_{n+1} \leq \theta^*_{n+1}, X_{n+1} \in G \mid \mathcal{E}, \mathcal{E}_k)}{\sum_{k=1}^{K} \pi_k \mathbb{P}(X_{n+1} \in G \mid \mathcal{E}, \mathcal{E}_k)} \\
&\overset{(a)}{=} \frac{\sum_{k=1}^{K} \frac{\pi_k}{n_k+1} \sum_{i=1}^{n_k+1} \mathbf{1}\{S_{i,k} \leq \theta^*_{i,k}\} \mathbf{1}\{X_{i,k} \in G\}}{\sum_{k=1}^{K} \frac{\pi_k}{n_k+1} \sum_{i=1}^{n_k+1} \mathbf{1}\{X_{i,k} \in G\}} \\
&\overset{(b)}{=} \frac{(1-\alpha) \sum_{k=1}^{K} \frac{\pi_k}{n_k+1} \sum_{i=1}^{n_k+1} \mathbf{1}\{X_{i,k} \in G\}}{\sum_{k=1}^{K} \frac{\pi_k}{n_k+1} \sum_{i=1}^{n_k+1} \mathbf{1}\{X_{i,k} \in G\}} \\
&= 1 - \alpha,
\end{aligned} \tag{30}$$

where (a) follows exchangeability on client k under the event E_k and (b) follows the first-order condition.

This result implies marginal coverage under $G = \mathcal{X}$. If the assumption $S_i \neq \hat{g}_S(X_i)$ does not hold, we have

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1} \mid \mathcal{D}) \mid X_{n+1} \in G, \mathcal{E}) \geq 1 - \alpha.$$

The proof technique is similar to [Gibbs et al., 2025] and Appendix B.2 and is omitted here. By the tower rule, we obtain the desired result. □

B PROOFS FOR SECTION 5

B.1 RESULTS ON T-DIGEST

This section provides additional results on T-Digest and proves Lemma 5.1.

To begin, we state the following fact that holds in our digest construction described in Section 4.2.

B.1.1 Important Lemmas

Lemma B.1 (CDF error controlled by maximal cluster mass)**.** *Let $\{(R_i, w_i)\}_{i=1}^{\ell}$ be weighted samples with total weight W and empirical CDF F. Let $\mathsf{TD} = \{(\bar{R}_c, W_c)\}_{c=1}^{m}$ be a digest with induced CDF $\widehat{F}(t) = \frac{1}{W}\sum_{c=1}^{m} W_c \mathbf{1}\{\bar{R}_c \leq t\}$. Define the maximal normalized cluster mass*

$$\rho_{\max} = \max_{1 \leq c \leq m} \frac{W_c}{W}. \tag{31}$$

Then,

$$\sup_{t \in \mathbb{R}} \left|F(t) - \widehat{F}(t)\right| \leq \rho_{\max}. \tag{32}$$

Proof. Let $\{\mathcal{R}_c\}_{c=1}^{m}$ be the partition for the digest defined in Section 4.2, and define $L_c = \min_{i \in \mathcal{R}_c} R_i$ and $U_c = \max_{i \in \mathcal{R}_c} R_i$. Fix $t \in \mathbb{R}$ and consider any cluster c. If $t < L_c$, then cluster c contributes zero to both $F(t)$ and $\widehat{F}(t)$. If $t \geq U_c$, cluster c contributes W_c/W to both $F(t)$ and $\widehat{F}(t)$. If $t \in [L_c, U_c)$, discrepancy arises from the cluster c only because $\max_{i \in \mathcal{R}_c} R_i \leq \min_{i \in \mathcal{R}_{c'}} R_i$ for any $c' > c$, which yields

$$\left|F(t) - \widehat{F}(t)\right| \leq \frac{W_c}{W}. \tag{33}$$

Taking the supremum over t completes the proof. □

Lemma B.2 (Arcsine scale implies $\rho_{\max} \leq \sin(\pi/\delta)$)**.** *Suppose the digest is constructed with the scale $r(q) = \frac{\delta}{2\pi}\arcsin(2q-1)$ under the condition of* (13) *with $\delta \geq 2$. Then*

$$\rho_{\max} = \max_{1 \leq c \leq m} \frac{W_c}{W} \leq \sin\Big(\frac{\pi}{\delta}\Big) \leq \frac{\pi}{\delta}. \tag{34}$$

Proof. For each cluster c, let its empirical left/right quantile boundaries be $q_c^L = V_{c-1}/W$ and $q_c^R = V_c/W$, where $V_c = \sum_{i=1}^{c} W_i$ (with $V_0 = 0$). Then $q_c^R - q_c^L = W_c/W$. The well-formedness constraint gives

$$r(q_c^R) - r(q_c^L) \leq 1 \quad \Longleftrightarrow \quad \arcsin(2q_c^R - 1) - \arcsin(2q_c^L - 1) \leq \frac{2\pi}{\delta}. \tag{35}$$

Let $u_L = 2q_c^L - 1$ and $u_R = 2q_c^R - 1$, so that $u_L, u_R \in [-1, 1]$ and $q_c^R - q_c^L = (u_R - u_L)/2$. Under the constraint $\arcsin(u_R) - \arcsin(u_L) \leq 2\pi/\delta$, the difference $u_R - u_L$ is maximized by taking symmetry around zero: $\arcsin(u_R) = \pi/\delta$ and $\arcsin(u_L) = -\pi/\delta$, hence $u_R = \sin(\pi/\delta)$ and $u_L = -\sin(\pi/\delta)$. Therefore

$$u_R - u_L \leq 2\sin\Big(\frac{\pi}{\delta}\Big) \quad \Longrightarrow \quad q_c^R - q_c^L = \frac{u_R - u_L}{2} \leq \sin\Big(\frac{\pi}{\delta}\Big), \tag{36}$$

which is exactly $W_c/W \leq \sin(\pi/\delta)$. Taking the maximum over c yields the claim. □

Lemma B.3 (From $\|F - \widehat{F}\|_\infty$ to rank-accurate quantiles)**.** *Let $F, \widehat{F}$ be CDFs on $\mathbb{R}$ and define the generalized quantile function $\widehat{Q}(u) = \inf\{t : \widehat{F}(t) \geq u\}$. If $\sup_{t\in\mathbb{R}} |F(t) - \widehat{F}(t)| \leq \epsilon$, then for all $u \in [0, 1]$,*

$$F(\widehat{Q}(u)) \ \in \ [u - \epsilon, \ u + \epsilon]. \tag{37}$$

In particular, when F is an empirical CDF with total weight W, $\widehat{Q}(u)$ has empirical rank within $(u \pm \epsilon)W$.

Proof. Fix $u \in [0, 1]$ and let $t^\star = \widehat{Q}(u)$. By definition, $\widehat{F}(t^\star) \geq u$ and for any $t < t^\star$, $\widehat{F}(t) < u$. Using $\sup_t |F(t) - \widehat{F}(t)| \leq \epsilon$ gives $F(t^\star) \geq \widehat{F}(t^\star) - \epsilon \geq u - \epsilon$. For the upper bound, for any $\eta > 0$ we have $\widehat{F}(t^\star - \eta) < u$, hence $F(t^\star - \eta) \leq \widehat{F}(t^\star - \eta) + \epsilon < u + \epsilon$. Letting $\eta \to 0$ and using right-continuity of F yields $F(t^\star) \leq u + \epsilon$. □

B.1.2 Proof of Lemma 5.1

Proof. By Lemma B.1,

$$\sup_t |F(t) - \widehat{F}(t)| \leq \rho_{\max}. \tag{38}$$

By Lemma B.2, $\rho_{\max} \leq \sin(\pi/\delta)$, hence $\sup_t |F(t) - \widehat{F}(t)| \leq \sin(\pi/\delta)$. □

B.2 PROOF OF THEOREM 5.1

We begin with the following corollary that controls the error of T-Digest for each atom.

Corollary B.1 (Atom-wise sketch accuracy)**.** *Let p_A be its weighted empirical CDF and let $\widehat{p}_A$ be the step-CDF induced by the merged digest TD_A. Then, under the arcsine scale* (14)*,*

$$\sup_{t\in\mathbb{R}} \big|p_A(t) - \widehat{p}_A(t)\big| \leq \epsilon = \sin\Big(\frac{\pi}{\delta}\Big) \leq \frac{\pi}{\delta}. \tag{39}$$

Equivalently, the quantile query induced by TD_A is ϵ-accurate in the sense of Section 4.3.

Proof. This is a direct application of Lemma 5.1 to the weighted sample set $\mathcal{D}_A$. □

Step 1: First-order condition. The subgradient of (23) on $\hat{\beta}$ with respect to β is given by

$$\left\{ \sum_{A\in\mathcal{A}} \sum_{c=1}^{m_A} W_{A,c} v_{A,c} \Phi_A + w_{\text{test}} v_{\text{test}} \Phi(X_{n+1}) \right|$$

$$v_{A,c} = \begin{cases} \alpha & \text{if } \bar{S}_{A,c} < \hat{\beta}^T \Phi_A, \\ -(1-\alpha) & \text{if } \bar{S}_{A,c} > \hat{\beta}^T \Phi_A, \\ t_{A,c} & \text{if } \bar{S}_{A,c} = \hat{\beta}^T \Phi_A \end{cases}, v_{\text{test}} = \begin{cases} \alpha & \text{if } S < \hat{\beta}^T \Phi(X_{n+1}), \\ -(1-\alpha) & \text{if } S > \hat{\beta}^T \Phi(X_{n+1}), \\ t_{\text{test}} & \text{if } S = \hat{\beta}^T \Phi(X_{n+1}) \end{cases}, \tag{40}$$

where $w_{\text{test}} = \sum_{k=1}^K \lambda_k$ with $\lambda_k = \pi_k/(n_k + 1)$ and $t_{A,c}, t_{\text{test}} \in [\alpha - 1, \alpha]$.

The first-order condition for optimality implies that for each group $G \in \mathcal{G}$, there exist $t_{A,c} \in [\alpha - 1, \alpha]$ (for clusters with ties) and $t_{\text{test}} \in [\alpha - 1, \alpha]$ (if the test point has a tie) such that

$$\sum_{A:A\subseteq G} \sum_{c=1}^{m_A} W_{A,c} v_{A,c} + w_{\text{test}} v_{\text{test}} \mathbf{1}\{X_{n+1} \in G\} = 0,$$

where the $v_{A,c}$ and v_{test} are as defined above.

Step 2: Empirical CDF and T-Digest approximate. The atoms $\{A\}$ partition $\mathcal{X}$ disjointly, with membership vectors Φ_A. The true empirical CDF per atom is

$$p_A(t) = W_A^{-1} \sum_{i:X_i \in A} \lambda_{k(i)} \mathbf{1}\{S_i \le t\},$$

where $W_A = \sum_{i:X_i \in A} \lambda_{k(i)}$, $\lambda_{k(i)} = \pi_{k(i)}/(n_{k(i)}+1)$, and $k(i)$ is the index of the client that the sample X_i belongs to, for $i \in [n]$. The coreset per atom yields

$$\hat{p}_A(t) = W_A^{-1} \sum_{c=1}^{m_A} W_{A,c} \mathbf{1}\{\bar{S}_{A,c} \le t\},$$

with $\sup_t |p_A(t) - \hat{p}_A(t)| \le \epsilon$ by Corollary B.1.

Step 3: Coverage. With $(X_{n+1}, Y_{n+1}) \sim \sum_{k=1}^{K} \pi_k \mathrm{P}_k$, as in the proof of FCP [Lu et al., 2023], let E_k be the event that (X_{n+1}, Y_{n+1}) is drawn from P_k. For each $k \in [K]$, let Π_{n_k+1} denote the set of all permutations of $\{1, \ldots, n_k+1\}$, and define the event

$$\mathcal{E} = \left\{\forall k \in [K],\, \exists \sigma_k \in \Pi_{n_k+1} \text{ s.t. } \left(S_{\sigma_k(1),k}, \ldots, S_{\sigma_k(n_k+1),k}\right) = \left(s_{1,k}, \ldots, s_{n_k+1,k}\right)\right\}, \tag{41}$$

where $(s_{1,k}, \ldots, s_{n_k+1,k})$ denotes the realized values of $(S_{1,k}, \ldots, S_{n_k+1,k})$.

Then, we have, for all $G \in \mathcal{G}$,

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1} \mid \widetilde{\mathcal{D}}) \mid X_{n+1} \in G, \mathcal{E}) = \mathbb{E}_{E_k}\left[\mathbb{P}(S_{n+1} \le \hat{\beta}_{S_{n+1}}^T \Phi(X_{n+1}) \mid X_{n+1} \in G, \mathcal{E}, E_k)\right] = \frac{N}{D},$$

where

$$N = \sum_{A:A\subseteq G} W_A p_A(\hat{\theta}_A) + w_{\text{test}} \mathbf{1}\{S_{n+1} \le \hat{\theta}_{n+1}\} \mathbf{1}\{X_{n+1} \in G\},$$
$$D = \sum_{A:A\subseteq G} W_A + w_{\text{test}} \mathbf{1}\{X_{n+1} \in G\}$$

with $\hat{\theta}_A = \hat{\beta}_{S_{n+1}}^T \Phi_A$ and $\hat{\theta}_{n+1} = \hat{\beta}_{S_{n+1}}^T \Phi(X_{n+1})$

Analogously, define the GC-FCP approximate numerator and denominator as

$$\hat{N} = \sum_{A:A\subseteq G} W_A \hat{p}_A(\hat{\theta}_A) + w_{\text{test}} \mathbf{1}\{S_{n+1} \le \hat{\theta}_{n+1}\} \mathbf{1}\{X_{n+1} \in G\}, \quad \hat{D} = D.$$

By Corollary B.1, we have

$$N \ge \sum_{A:A\subseteq G} W_A(\hat{p}_A(\hat{\theta}_A) - \epsilon) + w_{\text{test}} \mathbf{1}\{S_{n+1} \le \hat{\theta}_{n+1}\} \mathbf{1}\{X_{n+1} \in G\} = \hat{N} - \epsilon \sum_{A:A\subseteq G} W_A.$$

By the first-order condition, for the component corresponding to G, there exist $t_{A,c}, t_{\text{test}} \in [\alpha - 1, \alpha]$ such that

$$\alpha \sum_{A:A\subseteq G} \sum_{c:\bar{S}_{A,c} < \hat{\theta}_A} W_{A,c} - (1-\alpha) \sum_{A:A\subseteq G} \sum_{c:\bar{S}_{A,c} > \hat{\theta}_A} W_{A,c} + \sum_{A:A\subseteq G} \sum_{c:\bar{S}_{A,c} = \hat{\theta}_A} W_{A,c} t_{A,c}$$
$$+ w_{\text{test}} v_{\text{test}} \mathbf{1}\{X_{n+1} \in G\} = 0$$

which implies

$$\sum_{A:A\subseteq G} \sum_{c=1}^{m_A} W_{A,c} \mathbf{1}\left\{\bar{S}_{A,c} < \hat{\theta}_A\right\} + w_{\text{test}} \mathbf{1}\left\{S_{n+1} < \hat{\theta}_{n+1}\right\} \mathbf{1}\{X_{n+1} \in G\} + (1-\alpha)\mathcal{T}_0 = (1-\alpha)D - \mathcal{T},$$

where

$$\mathcal{T}_0 = \sum_{A:A\subseteq G} \sum_{c=1}^{m_A} W_{A,c} \mathbf{1}\left\{\bar{S}_{A,c} = \hat{\theta}_A\right\} + w_{\text{test}} \mathbf{1}\left\{S_{n+1} = \hat{\theta}_{n+1}\right\} \mathbf{1}\{X_{n+1} \in G\}$$

$$\mathcal{T} = \sum_{A:A\subseteq G}\sum_{c=1}^{m_A} W_{A,c}t_{A,c}\mathbf{1}\left\{\bar{S}_{A,c} = \hat{\theta}_A\right\} + w_{\text{test}}t_{\text{test}}\mathbf{1}\{S_{n+1} = \hat{\theta}_{n+1}\}\mathbf{1}\{X_{n+1}\in G\}.$$

By the definition of $\hat{p}_A(t)$ and $\hat{N}$, we have

$$\hat{N} - \alpha\mathcal{T}_0 = (1-\alpha)D - \mathcal{T}$$

Because $t_{A,c}, t_{\text{test}} \in [\alpha - 1, \alpha]$,

$$\alpha\mathcal{T}_0 - \mathcal{T} \geq 0,$$

yielding

$$\hat{N} \geq (1-\alpha)D.$$

This yields

$$\begin{aligned}
\mathrm{P}(Y_{n+1}\in\mathcal{C}(X_{n+1}\mid\widetilde{\mathcal{D}})\mid X_{n+1}\in G,\mathcal{E}) &= \frac{N}{D}\\
&\geq \frac{\hat{N} - \epsilon\sum_{A:A\subseteq G}W_A}{D}\\
&\geq \frac{(1-\alpha)D - \epsilon\sum_{A:A\subseteq G}W_A}{D}\\
&\geq 1-\alpha-\epsilon,
\end{aligned}$$

where the last inequality is due to $\sum_{A:A\subseteq G} W_A < D$. Taking expectation on both sides w.r.t. $\mathcal{E}$ yields the desired result.

C GROUP-CONDITIONAL COVERAGE UPPER BOUND OF GC-FCP

Theorem C.1 (Upper bound coverage in the case of perfect quantile regression)**.** *Assume that the conditional score distribution $S \mid X$ is continuous and fix any confidence level $\eta \in (0,1)$. Then, with probability at least $1-\eta/2$ over the calibration data, for every group $G \in \mathcal{G}$, GC-FCP satisfies*

$$\mathbb{P}(Y_{n+1}\in\mathcal{C}(X_{n+1}\mid\widetilde{\mathcal{D}})\mid X_{n+1}\in G) \leq 1-\alpha+\frac{\pi}{\delta}+\Delta_G, \tag{42}$$

where the additive term

$$\Delta_G = |\mathcal{G}|\frac{\max\left\{\frac{\pi}{\delta}, \sum_{k=1}^{K}\frac{\pi_k}{n_k+1}\right\}}{\sum_k\left[\frac{\pi_k n_k}{n_k+1}\left(p_{k,G} - \sqrt{\frac{\log(2K/\eta)}{2n_k}}\right)\right]}, \tag{43}$$

with $p_{k,G} = \mathrm{P}_{X,k}(X\in G)$ being the probability that a covariate X drawn from client k belongs to group G.

Proof: See Appendix C.1.

Theorem C.1 shows that GC-FCP cannot be arbitrarily conservative. The group-wise coverage lies in a narrow band around $1-\alpha$, with band width controlled explicitly by the additive term Δ_G, which consists of the mis-coverage gap ϵ, calibration sizes n_k, and the mixture-weighted group mass $p_{k,G}$. For instance, when the mixture-weighted group mass $\sum_k \pi_k p_{k,G}$ is small, the denominator shrinks, and Δ_G grows, which suggests that an efficient prediction set for low-support groups requires more calibration samples.

C.1 PROOF OF THEOREM C.1

For any $G \in \mathcal{G}$, where $|\mathcal{G}| = d$, we begin with

$$\mathrm{P}(Y_{n+1}\in\mathcal{C}(X_{n+1}\mid\widetilde{\mathcal{D}})\mid X_{n+1}\in G,\mathcal{E}) = \frac{N}{D} \leq \frac{\hat{N}+\epsilon\sum_{A:A\subseteq G}W_A}{D} \leq \frac{\hat{N}}{D}+\epsilon$$

Record that $t_{A,c}, t_{\text{test}} \in [\alpha - 1, \alpha]$, which yields

$$\mathcal{T} \geq (\alpha - 1)\mathcal{T}_0, \quad \hat{N} = (1-\alpha)D - \mathcal{T} + \alpha\mathcal{T}_0 \implies \hat{N} \leq (1-\alpha)D + \mathcal{T}_0$$

Hence,

$$\mathrm{P}(Y_{n+1} \in \mathcal{C}(X_{n+1} \mid \widetilde{\mathcal{D}}) \mid X_{n+1} \in G, \mathcal{E}) \leq 1 - \alpha + \epsilon + \frac{\mathcal{T}_0}{D}$$

Now our target is to upper bound

$$\mathcal{T}_0 = \sum_{A:A\subseteq G} \sum_{c=1}^{m_A} W_{A,c} \mathbf{1}\left\{\bar{S}_{A,c} = \hat{\theta}_A\right\} + w_{\text{test}} \mathbf{1}\left\{S_{n+1} = \hat{\theta}_{n+1}\right\} \mathbf{1}\{X_{n+1} \in G\}$$

under the assumption that the distribution of $S \mid X$ is continuous.

To bound this term, we first simplify the notation by relabeling the pseudo-points with the test point with $i = 1, \ldots, m, m+1$, where $m = \sum_{A\in\mathcal{A}} m_A = \mathcal{O}(\delta|\mathcal{A}|)$ is the total number of clusters and $\bar{S}_{m+1} = S_{n+1}$.

Claim 1. Under the conditions of Theorem C.1, with probability 1, we have

$$\sum_{A:A\subseteq G} \sum_{c=1}^{m_A} \mathbf{1}\left\{\bar{S}_{A,c} = \hat{\theta}_A\right\} + \mathbf{1}\left\{S_{n+1} = \hat{\theta}_{n+1}\right\} \leq d.$$

Proof. See Appendix C.2 □

By Claim 1, with probability 1, we have

$$\frac{\mathcal{T}_0}{D} \leq \frac{d \cdot \max\left\{\max_{A,c:A\subseteq G, c\in[m_A]} W_{A,c}\ ,\ w_{\text{test}}\mathbf{1}\{X_{n+1}\in G\}\right\}}{\sum_{A:A\subseteq G} W_A + w_{\text{test}}\mathbf{1}\{X_{n+1}\in G\}}$$

Next, we bound RHS using the properties of T-Digest. Record that $w_{\text{test}} = \sum_{k=1}^{K} \lambda_k = \sum_{k=1}^{K} \frac{\pi_k}{n_k+1}$.

Claim 2. Assume that GC-FCP uses the scale function $r(q) = \frac{\delta}{2\pi}\arcsin(2q-1)$ for $q \in [0,1]$, where q is the normalized cumulative weight, i.e., quantile. Then, we have

$$\max_{A,c} W_{A,c} \leq W_A \sin\left(\frac{\pi}{\delta}\right) \leq \sin\left(\frac{\pi}{\delta}\right) \leq \frac{\pi}{\delta}.$$

Proof. This is a direct result of Lemma B.2 and $W_A \leq 1$. □

Claim 3. Assume that the mixture weights $\pi_k > 0$, with $p_{k,G} = P_{X,k}(X \in G) > 0$, for any $\eta > 0$, with probability larger than $1 - \eta/2$, we have

$$\sum_{A:A\subseteq G} W_A \geq \sum_k \frac{\pi_k n_k}{n_k+1}\left(p_{k,G} - \sqrt{\frac{\log(2K/\eta)}{2n_k}}\right). \tag{44}$$

Proof. See Appendix C.3. □

Combining Claims 1, 2, and 3, we have

$$\frac{\mathcal{T}_0}{D} \leq d\frac{\max\left\{\frac{\pi}{\delta}, \sum_{k=1}^{K}\frac{\pi_k}{n_k+1}\right\}}{\sum_k \frac{\pi_k n_k}{n_k+1}\left(p_{k,G} - \sqrt{\frac{\log(2K/\eta)}{2n_k}}\right)}$$

Record that

$$\mathrm{P}(Y_{n+1} \in \mathcal{C}(X_{n+1} \mid \widetilde{\mathcal{D}}) \mid X_{n+1} \in G, \mathcal{E}) \leq 1 - \alpha + \epsilon + \frac{\mathcal{T}_0}{D},$$

combing the bound of $\mathcal{T}_0/D$ and taking expectation over $\mathcal{E}$ yield the desired result.

C.2 PROOF OF CLAIM 1

Proof. Let $\mathcal{L}$ denote the set of all pseudo-points, consisting of the clusters across all atoms and the test point. Specifically, $\mathcal{L} = \{(A, c) : A \in \mathcal{A}, c \in [m_A]\} \cup \{\text{test}\}$, with $|\mathcal{L}| = m + 1$, where $m = \sum_{A \in \mathcal{A}} m_A = \mathcal{O}(\delta|\mathcal{A}|)$ is the total number of clusters at merged T-Digests. For each $l \in \mathcal{L}$, define the feature $\Psi_l = \Phi_A$ if $l = (A, c)$ is a cluster in atom A, or $\Psi_l = \Phi(X_{n+1})$ if $l = \text{test}$, and the score $T_l = \bar{S}_{A,c}$ if $l = (A, c)$, or $T_l = S_{n+1}$ if $l = \text{test}$.

The optimizer $\hat{\beta}_{S_{n+1}}$ is the solution to the weighted quantile regression problem (23) with $S = S_{n+1}$. A tie at pseudo-point l occurs if $T_l = \hat{\beta}^T_{S_{n+1}} \Psi_l$. Let $\mathcal{D}^{aug} = \{X_{i,k}\}_{i \in [n_k], k \in [K]} \cup \{X_{n+1}\}$.

As in [Gibbs et al., 2025], we calculate the probability

$$\mathrm{P}\left(\sum_{l \in \mathcal{L}} \mathbf{1}\left\{T_l = \hat{\beta}_S^T \Psi_l\right\} > d \mid \mathcal{D}^{aug}\right)$$

The event $\left\{\sum_{l \in \mathcal{L}} \mathbf{1}\left\{T_l = \hat{\beta}_S^T \Psi_l\right\} > d\right\}$ implies that there exists a subset $\mathcal{I} \subseteq \mathcal{L}$ with $|\mathcal{I}| = d + 1$ such that $T_l = \hat{\beta}_S^T \Psi_l$ for all $l \in \mathcal{I}$. Therefore,

$$\mathrm{P}\left(\sum_{l \in \mathcal{L}} \mathbf{1}\left\{T_l = \hat{\beta}_S^T \Psi_l\right\} > d \mid \mathcal{D}^{aug}\right) \leq \sum_{\mathcal{I} \subseteq \mathcal{L},\, |\mathcal{I}| = d+1} \mathbb{P}\left(\exists \beta \in \mathbb{R}^d \text{ s.t. } T_l = \beta^T \Psi_l \; \forall l \in \mathcal{I} \;\middle|\; \mathcal{D}^{aug}\right)$$

For a fixed $\mathcal{I}$, the event is that the vector $(T_l)_{l \in \mathcal{I}} \in \operatorname{span}\{(\Psi_l^T)_{l \in \mathcal{I}}\}$, where $\operatorname{span}\{(\Psi_l^T)_{l \in \mathcal{I}}\}$ is the image of the map $\beta \mapsto (\Psi_l^T \beta)_{l \in \mathcal{I}}$, which is a linear subspace of $\mathbb{R}^{d+1}$ with dimension at most d.

Under the continuity assumption, the conditional distribution of $(S_{i,k})_{i,k} \cup \{S_{n+1}\}$ given the X's is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^{n+1}$.

The mapping from the original scores $S_{i,k}$ (and S_{n+1}) to the coreset scores $\{T_l\}_{l \in \mathcal{L}}$ is piecewise affine. For each atom A, the score space $\mathbb{R}^{n_A}$ is partitioned into finitely many open cones $\mathcal{R}_\pi$ indexed by permutations $\pi \in \mathcal{S}_{n_A}$, where $\mathcal{R}_\pi = \{\mathbf{s} \in \mathbb{R}^{n_A} : s_{\pi(1)} < \cdots < s_{\pi(n_A)}\}$, where boundaries have measure zero. Within each $\mathcal{R}_\pi$, the sorted scores $\mathbf{s}^{(\pi)}$ determine fixed quantile positions q_l based on permuted weights, leading to fixed cluster assignments via the deterministic T-Digest merge rules. Thus, each $\bar{S}_{A,c}$ is an affine function of $\mathbf{s}^{(\pi)}$ (weighted average over fixed indices), and hence affine in $\mathbf{s}$. The full map to $(T_l)_{l \in \mathcal{I}}$ is therefore piecewise affine across atoms and the test score, preserving measure-zero sets under the continuous distribution of $S \mid X$.

Since the subspace has Lebesgue measure zero in $\mathbb{R}^{d+1}$, its pre-image under the affine map also has measure zero. For a fixed $\mathcal{I}$, we have

$$\mathrm{P}\left(\exists \beta \in \mathbb{R}^d \text{ s.t. } T_l = \beta^T \Psi_l \; \forall l \in \mathcal{I} \;\middle|\; \mathcal{D}^{aug}\right) = 0.$$

Since there are at most $\binom{m+1}{d+1}$ many such $\mathcal{I}$, the union bound yields

$$\mathrm{P}\left(\sum_{l \in \mathcal{L}} \mathbf{1}\left\{T_l = \hat{\beta}_S^T \Psi_l\right\} > d \mid \mathcal{D}^{aug}\right) = 0$$

Because

$$\left\{\sum_{A : A \subseteq G} \sum_{c=1}^{m_A} \mathbf{1}\left\{\bar{S}_{A,c} = \hat{\theta}_A\right\} + \mathbf{1}\left\{S_{n+1} = \hat{\theta}_{n+1}\right\} > d\right\} \subseteq \left\{\sum_{l \in \mathcal{L}} \mathbf{1}\left\{T_l = \hat{\beta}_S^T \Psi_l\right\} > d\right\},$$

we have

$$\mathrm{P}\left(\sum_{A : A \subseteq G} \sum_{c=1}^{m_A} \mathbf{1}\left\{\bar{S}_{A,c} = \hat{\theta}_A\right\} + \mathbf{1}\left\{S_{n+1} = \hat{\theta}_{n+1}\right\} > d \mid \mathcal{D}^{aug}\right) \leq \mathbb{P}\left(\sum_{l \in \mathcal{L}} \mathbf{1}\left\{T_l = \hat{\beta}_S^T \Psi_l\right\} > d \mid \mathcal{D}^{aug}\right) = 0.$$

Marginalizing $\mathcal{D}^{aug}$ yields the desired result. □

C.3 PROOF OF CLAIM 3

Proof. The calibration weight in G is $\sum_{A:A\subseteq G} W_A = \sum_k \lambda_k \cdot \#\{i : X_{i,k} \in G\}$, where $\lambda_k = \pi_k/(n_k+1)$. Let $Z_k = \#\{i : X_{i,k} \in G\}$, then $\sum_{A:A\subseteq G} W_A = \sum_k \lambda_k Z_k$, and

$$Z_k \sim \mathrm{Bin}(n_k, p_{k,G}), \text{ with } p_{k,G} = P_{X,k}(X \in G)$$

and the expected weight is

$$E\left[\sum_{A:A\subseteq G} W_A\right] = \sum_k \lambda_k n_k p_{k,G} = \sum_k \frac{\pi_k n_k}{n_k+1} p_{k,G}.$$

By Hoeffding's inequality, for each k,

$$\mathrm{P}\left(Z_k \le n_k p_{k,G} - t_k\right) \le \exp\left(\frac{-2t_k^2}{n_k}\right),$$

for $t_k > 0$. Set $t_k = \sqrt{(n_k/2)\log(2K/\eta)}$ with any $\eta > 0$, we have

$$\mathrm{P}(Z_k \le n_k p_{k,G} - t_k) \le \frac{\eta}{2K}.$$

By union bound over k, with probability larger than $1-\eta/2$,

$$\sum_k \lambda_k Z_k \ge \sum_k \lambda_k (n_k p_{k,G} - t_k) = \sum_k \frac{\pi_k n_k}{n_k+1}\left(p_{k,G} - \sqrt{\frac{\log(2K/\eta)}{2n_k}}\right).$$

We have, with probability larger than $1-\eta/2$,

$$\sum_{A:A\subseteq G} W_A = \sum_k \lambda_k Z_k \ge \sum_k \frac{\pi_k n_k}{n_k+1}\left(p_{k,G} - \sqrt{\frac{\log(2K/\eta)}{2n_k}}\right).$$

□

D DUAL CONSTRUCTION FOR (9) UNDER THE OBJECTIVE (10)

Reformulate the primal: Let $\lambda_k = \pi_k/(n_k+1)$. The primal objective of (10) is rewritten as:

$$\sum_{k=1}^{K} \lambda_k \sum_{i=1}^{n_k} \ell_\alpha(g(X_{i,k}), S_{i,k}) + \left(\sum_{k=1}^{K} \lambda_k\right) \ell_\alpha(g(X_{n+1}), S).$$

The pinball loss is given by $\ell_\alpha(u,s) = (1-\alpha)(s-u)_+ + \alpha(u-s)_+$, where $(a)_+ = \max\{a, 0\}$. To reformulate the primal problem (10), we first introduce the following claim to rewrite the pinball loss as an optimization problem.

Claim 4 (Pinball Loss Reformulation)**.** Fix any scalar residual $r = s-u$. Consider the following problem

$$\min_{p,q\ge 0} (1-\alpha)p + \alpha q \quad \text{s.t.} \quad r = p - q. \tag{45}$$

Then, the optimal value of (45) equals $\ell_\alpha(u,s)$.

Proof. Since $r = p-q$, we have $p = r+q$. Together with $p \ge 0$ and $q \ge 0$, feasible pairs satisfy $q \ge \max\{-r, 0\} = (-r)_+$ and then $p = r+q \ge 0$. Substitute $p = r+q$ into the objective:

$$(1-\alpha)p + \alpha q = (1-\alpha)(r+q) + \alpha q = (1-\alpha)r + q.$$

Thus, minimizing over feasible q is equivalent to

$$\min_{q \geq (-r)_+} (1-\alpha)r + q,$$

whose minimum is attained at $q^* = (-r)_+$, giving value

$$(1-\alpha)r + (-r)_+.$$

If $r \geq 0$, then $(-r)_+ = 0$, so the value is $(1-\alpha)r = (1-\alpha)(s-u)$. If $r < 0$, then $(-r)_+ = -r$, so the value is $(1-\alpha)r - r = -\alpha r = \alpha(u-s)$. Therefore, the optimal value is exactly

$$(1-\alpha)(s-u)_+ + \alpha(u-s)_+ = \ell_\alpha(u,s).$$

□

By Claim 4, we introduce auxiliary $p_{i,k}, q_{i,k} \geq 0$ for each calibration point (i,k) and $p_k^{\text{test}}, q_k^{\text{test}} \geq 0$ for each virtual test copy per client k and rewrite the problem (10) as

$$\begin{aligned} \text{(P0)}: \quad \min_{g \in \mathcal{F}_\mathcal{G}, p, q} \quad & \sum_{k=1}^{K}\sum_{i=1}^{n_k} \lambda_k \left((1-\alpha)p_{i,k} + \alpha q_{i,k}\right) + \sum_{k=1}^{K} \lambda_k \left((1-\alpha)p_k^{\text{test}} + \alpha q_k^{\text{test}}\right), \\ \text{s.t.} \quad & S_{i,k} - g(X_{i,k}) = p_{i,k} - q_{i,k}, S - g(X_{n+1}) = p_k^{\text{test}} - q_k^{\text{test}}, \quad \forall i \in [n_k],\ \forall k \in [K] \\ & p_{i,k}, q_{i,k}, p_k^{\text{test}}, q_k^{\text{test}} \geq 0, \quad \forall i \in [n_k],\ \forall k \in [K] \end{aligned} \tag{46}$$

Dual problem: Introduce dual variables $\eta_{i,k}$ for calibration constraints and η_k^{test} for each test copy. Using standard calculations, the dual problem is given by the following linear programming (LP) problem:

$$\text{(P1)}: \quad \max_{\{\eta_{i,k}\}, \{\eta_k^{\text{test}}\}} \quad \sum_{k=1}^{K}\sum_{i=1}^{n_k} \eta_{i,k} S_{i,k} \ + \ S \sum_{k=1}^{K} \eta_k^{\text{test}} \tag{47}$$

$$\text{s.t.} \quad -\lambda_k \alpha \leq \eta_{i,k} \leq \lambda_k (1-\alpha), \quad \forall i \in [n_k],\ \forall k \in [K], \tag{48}$$

$$-\lambda_k \alpha \leq \eta_k^{\text{test}} \leq \lambda_k (1-\alpha), \quad \forall k \in [K], \tag{49}$$

$$\sum_{k=1}^{K}\sum_{i=1}^{n_k} \eta_{i,k} \Phi(X_{i,k}) \ + \ \sum_{k=1}^{K} \eta_k^{\text{test}} \Phi(X_{n+1}) = 0. \tag{50}$$

where $\Phi(x) = \left(\mathbf{1}\{x \in G\}\right)_{G \in \mathcal{G}} \in \{0,1\}^{|\mathcal{G}|}$. The last constraint is to enforce the linearity of $g(x) = \beta^T \Phi(x)$, which arises mathematically from the stationary condition with respect to the primal weights β.

Set construction via KKT conditions: In [Gibbs et al., 2025, Section 4], the conformal prediction set is constructed using the dual solution η_S (for input score S) and KKT conditions. The optimal $\hat{g}_S(X_{n+1})$ satisfies: $\eta_S^{n+1} = 1-\alpha$ if $S > \hat{g}_S(X_{n+1})$, $\eta_S^{n+1} = -\alpha$ if $S < \hat{g}_S(X_{n+1})$, and $\eta_S^{n+1} \in [-\alpha, 1-\alpha]$ if $S = \hat{g}_S(X_{n+1})$.

Similarly, let $\eta_S^{\text{test}} = \sum_{k=1}^{K} \eta_{k,S}^{\text{test}}$ be the sum of dual variables over test copies (since the test is duplicated). The set is:

$$\hat{C}_{\text{dual}}(X_{n+1}) = \left\{ y : \eta_{s(X_{n+1},y)}^{\text{test}} \leq \left(\sum_{k=1}^{K} \lambda_k\right)(1-\alpha) \right\},$$

using the weighted upper bound.

Binary search: We compute $\hat{C}_{\text{dual}}(X_{n+1})$ using the following two-step procedure. First, using Algorithm 1 of [Gibbs et al., 2025], we binary search for the largest value of S^* such that $\eta_{S^*}^{n+1} < 1-\alpha$. Second, we output all y such that $s(X_{n+1}, y) \leq S^*$ [Gibbs et al., 2025, Theorem 4].

Table 4: Experimental setup for synthetic regression.

Parameter	**Value**
T-Digest parameter δ	$\delta = 250$
Local calibration sizes n_k	$n_1 = 1000$, $n_k = 333$ for $k > 1$
Mixture weights π_k	$\pi_k = 1/K$ for all k
Miscoverage level α	$\alpha = 0.1$
Model $f(\cdot)$	linear regression
Score $s(x, y)$	$\lvert y - f(x)\rvert$
Test points n_{test}	200
Monte Carlo runs	100

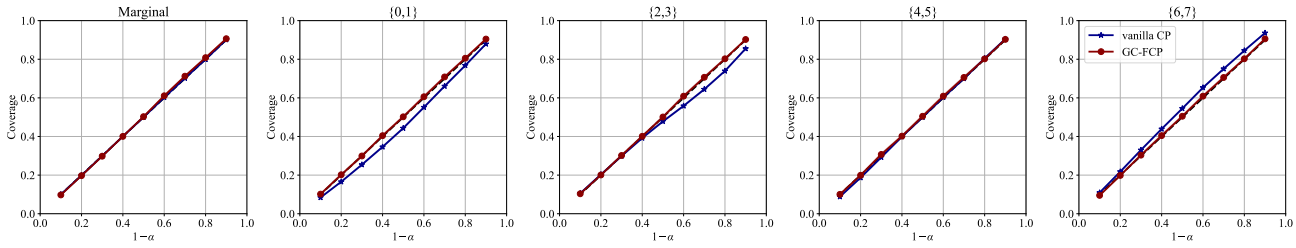
E ADDITIONAL EXPERIMENTAL RESULTS

E.1 EXPERIMENTAL SETTINGS AND RESULTS OF SECTION 6

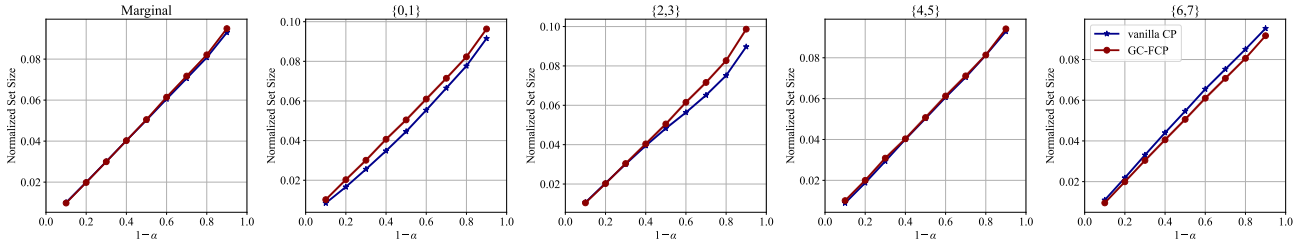
Other parameters for the synthetic regression can be found in Table 4.

Figure 4 reports the empirical group-wise coverage and the average set size over confidence level α on CIFAR-10. Vanilla CP achieves competitive marginal coverage but exhibits group-wise miscalibration under non-i.i.d. client partitions. GC-FCP reduces these discrepancies, pushing group-wise coverage closer to the target across $G \in \mathcal{G}$. This improvement comes with a moderate increase in set size, reflecting the standard trade-off when enforcing conditional validity over overlapping groups.

Figure 5 reports the empirical group-wise coverage and the average set size over confidence level α on PathMNIST. As expected, vanilla CP exhibits nonuniform conditional coverage across groups and different values of α. GC-FCP improves stability across groups while keeping prediction sets competitive in federated settings where raw examples remain local.



(a) Average coverage over different groups.

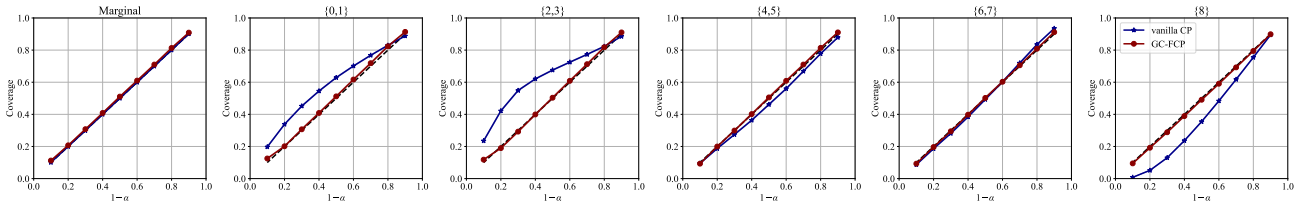


(b) Average set size over different groups.

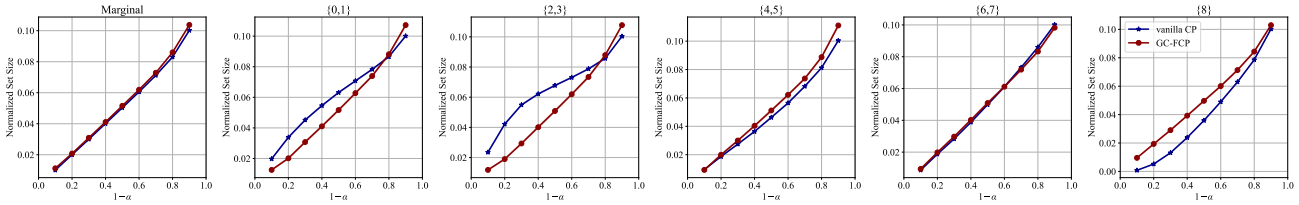
Figure 4: Average coverage and set size versus coverage level $1-\alpha$ with vanilla CP and the proposed GC-FCP on CIFAR-10.

E.2 IMAGENET-1K EXPERIMENTS

We additionally evaluate GC-FCP on the ImageNet-1K validation set [Deng et al., 2009] using a pretrained ResNet-50 classifier [He et al., 2016]. In each Monte Carlo repetition, we split the validation data into $n_{\text{cal}} = 40{,}000$ calibration samples and $n_{\text{test}} = 10{,}000$ evaluation samples. The calibration set is distributed across $K = 50$ clients by a class-wise Dirichlet allocation with concentration parameter 0.3, enforcing at least 200 calibration samples per client, and we use



(a) Average coverage over different groups.



(b) Average set size over different groups.

Figure 5: Average coverage and set size versus coverage level $1 - \alpha$ with vanilla CP and the proposed GC-FCP on PathMNIST.

uniform mixture weights. We consider three standard classification scores: THR/HPS [Sadinle et al., 2019], APS [Romano et al., 2020], and RAPS [Angelopoulos et al., 2021].

We design 10 semantic groups plus 5 ambiguity groups, which tests coverage across both category and model uncertainty, following ImageNet's semantic hierarchy [Deng et al., 2009] and standard confidence-based uncertainty measures in calibrated classification [Guo et al., 2017]. Let $\mathcal{Y} = \{1, \ldots, 1000\}$ denote the ImageNet-1K label space, and let $\{\mathcal{C}_m\}_{m=1}^{10}$ be a fixed partition of labels into ten semantic meta-categories. For classifier probabilities $p_\theta(\cdot \mid x)$, define

$$P_m(x) = \sum_{c \in \mathcal{C}_m} p_\theta(c \mid x), \qquad M(x) = \arg \max_{m \in \{1,\ldots,10\}} P_m(x),$$

which induces semantic groups $G_m^{\text{sem}} = \{x : M(x) = m\}$. To capture prediction ambiguity, let $p_{(1)}(x) \geq p_{(2)}(x)$ be the two largest class probabilities and define the margin $A(x) = p_{(1)}(x) - p_{(2)}(x)$. Using calibration-split quantiles of $A(x)$, we form five ambiguity bins G_b^{amb}. The final family is $\mathcal{G} = \{G_m^{\text{sem}}\}_{m=1}^{10} \cup \{G_b^{\text{amb}}\}_{b=1}^{5}$. Each example belongs to one semantic and one ambiguity group, yielding at most $10 \times 5 = 50$ semantic–ambiguity atoms after removing empty intersections.

We compare GC-FCP against centralized CP, FedCP [Lu et al., 2023], Mondrian-FedCP, a FedCF-style group-wise baseline [Srinivasan et al., 2025], and importance-weighted FCP [Plassier et al., 2023]. Since the group-wise baselines do not natively produce a simultaneous predictor under overlapping groups, overlapping test points are assigned by a fixed minimum-index rule.

Figures 6 and 7 report the main ImageNet comparison using the RAPS score. GC-FCP with $\delta \in \{50, 250, 500\}$ achieves near-target coverage across the semantic–ambiguity groups, whereas marginal and group-wise baselines under-cover several difficult groups.

Table 5 isolates the effect of group design with $\delta = 250$ fixed. Across THR, APS, and RAPS, semantic, ambiguity, and semantic–ambiguity groups all maintain worst-group coverage near the target, while normalized set size changes with the conditioning family and score.

Figure 8 isolates the compression–accuracy trade-off by sweeping $\delta \in \{25, 50, 250, 500, 1000, 2500\}$ under the semantic–ambiguity group family. The results support the theoretical trend in Theorem 5.1: aggressive compression can reduce worst-group coverage, while moderate δ values recover the target level. The effect of federated aggregation is also reflected in the main CIFAR-10 and PathMNIST comparisons between centralized GC-FCP and compressed GC-FCP, where moderate compression closely tracks the exact atom-stratified reference while substantially reducing computation.

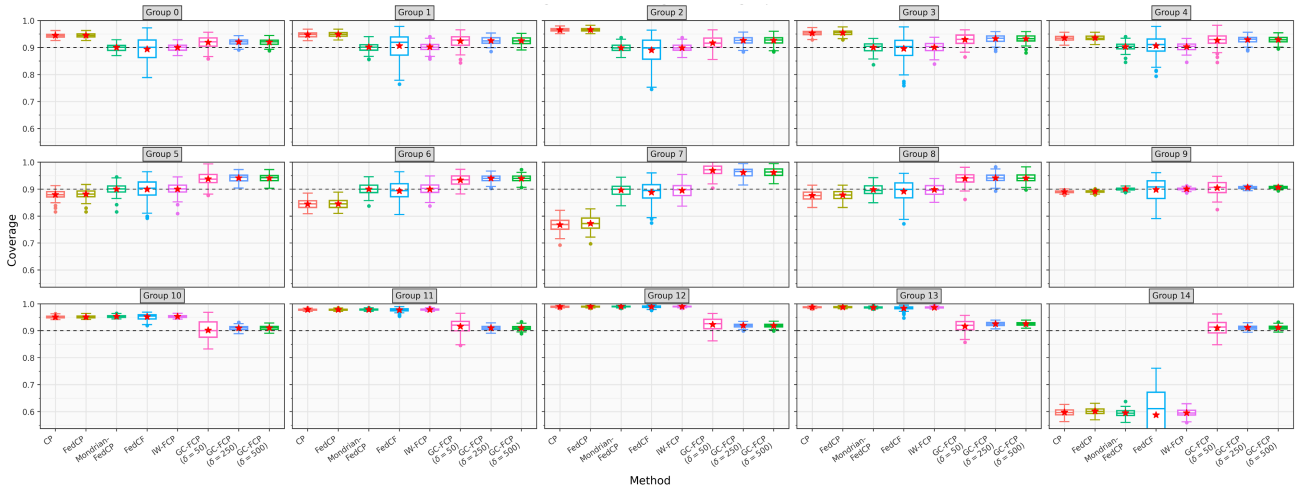


Figure 6: Group-conditional coverage of baselines and GC-FCP on ImageNet-1K with the RAPS score.

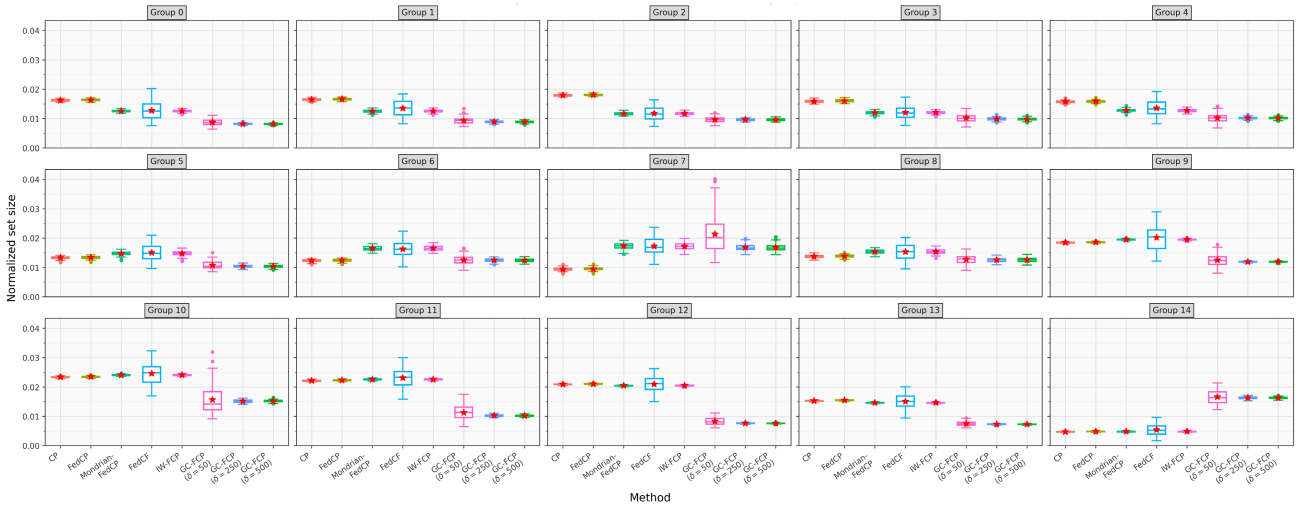


Figure 7: Group-conditional normalized set size of baselines and GC-FCP on ImageNet-1K with the RAPS score.

Table 5: Ablation of the group design used by GC-FCP on ImageNet-1K under three classification scores. Marginal coverage, average group coverage, and normalized set size are reported as mean ± standard error over 10 Monte Carlo repetitions; worst-group coverage is computed from pooled group-wise coverages across repetitions.

Score	Group design	Marg. cov.	Worst-group cov.	Avg. group cov.	Norm. set size
APS	ambiguity	0.905 ± 0.001	0.901	0.905 ± 0.001	0.0875 ± 0.0008
APS	semantic	0.911 ± 0.001	0.902	0.927 ± 0.001	0.1779 ± 0.0015
APS	semantic–ambiguity	0.915 ± 0.001	0.906	0.926 ± 0.001	0.0953 ± 0.0005
RAPS	ambiguity	0.907 ± 0.001	0.902	0.907 ± 0.001	0.0110 ± 0.0000
RAPS	semantic	0.913 ± 0.001	0.905	0.927 ± 0.002	0.0184 ± 0.0001
RAPS	semantic–ambiguity	0.916 ± 0.001	0.908	0.925 ± 0.002	0.0113 ± 0.0000
THR	ambiguity	0.906 ± 0.001	0.900	0.906 ± 0.001	0.0023 ± 0.0000
THR	semantic	0.912 ± 0.001	0.903	0.929 ± 0.002	0.0017 ± 0.0000
THR	semantic–ambiguity	0.916 ± 0.001	0.907	0.924 ± 0.001	0.0096 ± 0.0005

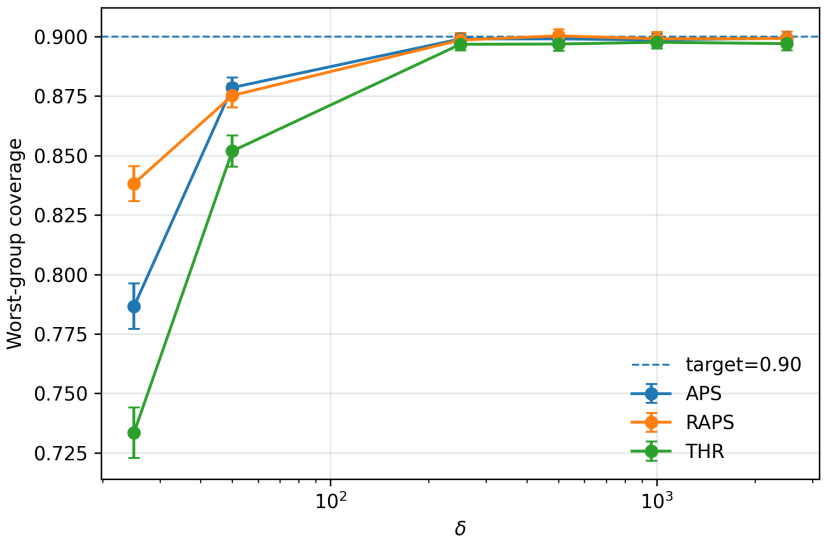


Figure 8: Worst-group coverage of GC-FCP versus T-Digest compression parameter δ on ImageNet-1K.