

Reliability Analysis and Scene-Aware Calibration of Federated Learning for Long-Tailed Autonomous Driving Scenes

Song Wen¹, Shuling Yin², Jun Zhu³, Xinying Wang⁴

School of Artificial Intelligence, HuBei Open University, Wuhan, Hubei, China^{1,2,3}

School of Computer Engineering, HuBei University of Arts and Science, Xiangyang, Hubei, China⁴

Abstract—This study examines the predictive reliability of FedRare weather classification on BDD100K across 18 weather-time-of-day scenes. Labeled logits are calibrated centrally using 2,653 observations and evaluated on 6,188 observations from a frozen scene-stratified split. In the primary analysis, Predicted Scene-Aware Temperature Scaling (Predicted-SATS) reduces negative log-likelihood (NLL) from 0.7108 under Global Temperature Scaling (GTS) to 0.6464 and Brier score from 0.3660 to 0.3308. Predicted-class scaling accounts for part of this reduction, with further improvement after adding time of day. A comparison of independently trained checkpoints with two seeds, evaluated on the same observations and frozen split, gives lower overall NLL, Brier score, and classwise calibration error under Predicted-SATS, but higher values in the rain/snow subset at night or dawn/dusk. Dirichlet calibration has lower overall NLL and Brier score than Predicted-SATS and higher losses in this subset. Confidence expected calibration error (ECE) and selective-error comparisons differ between the two seeds. Increasing the proportion of rainy/snowy observations across all times of day reduces the NLL advantage over GTS, while very small calibration subsets produce unstable temperature estimates. The results show that lower aggregate probability loss can coexist with higher subgroup loss, and that probability quality, calibration, and confidence-based prediction selection need separate evaluation.

Keywords—Autonomous driving; confidence calibration; long-tailed distribution; temperature scaling; predictive reliability

I. INTRODUCTION

Federated Learning (FL) enables multiple distributed clients to collaboratively train a shared model while keeping their raw local data decentralized [1]. This learning paradigm is particularly relevant when data are distributed across edge devices and direct centralization is constrained by privacy, communication, storage, or data-governance requirements [1], [2]. Practical FL systems, however, are characterized by statistical and system heterogeneity. Client datasets are often not independently and identically distributed (non-IID), differ in sample quantity and class composition, and contribute unevenly during optimization [2], [3]. These properties can affect convergence, global model quality, and the representation of underrepresented local data.

Autonomous driving is a representative application in which such heterogeneity is important. Vehicles operate across different geographical areas, traffic environments, weather conditions, and illumination periods, causing locally collected observations to differ among participants [4]. Existing studies

have investigated FL for connected and automated vehicles, including collaborative perception and object detection [4], [5], [6], [7]. While FL reduces the need to centralize raw vehicular observations, environmental heterogeneity creates an additional challenge: common driving conditions may dominate distributed training data, whereas rare but operationally important conditions can be underrepresented.

Long-tailed visual recognition studies have shown that imbalanced training distributions can bias learning toward frequently observed categories and reduce performance on less represented classes [8], [9], [10], [11]. A related problem occurs at the environmental-scene level in autonomous driving. BDD100K contains diverse observations spanning different weather and time-of-day conditions [12]. These combinations are not equally represented, and adverse conditions such as rain, snow, fog, and low illumination introduce additional perception challenges [13], [14]. Aggregate recognition performance can therefore conceal substantial differences among individual driving contexts.

FedRare was previously proposed for adaptive client selection in long-tailed autonomous-driving scene classification [15]. It uses scene rarity, criticality, and local update quality in client selection. The published study reported improved recognition for underrepresented driving conditions. The present study evaluates predictive confidence across driving scenes using two existing FedRare checkpoints, with the network weights fixed during calibration and evaluation.

Confidence reliability is distinct from classification accuracy. Modern neural networks can achieve high predictive accuracy while producing probability estimates that are systematically overconfident or underconfident [16]. A calibrated classifier should produce confidence values that correspond to empirical correctness frequencies. Guo et al. showed that modern neural networks can be miscalibrated and demonstrated that Temperature Scaling (TS) can adjust predictive confidence after training without changing the predicted class [16].

Calibration has been studied using a range of post-hoc methods and evaluation criteria [17], [18], [19], [20], [21], [22]. Expected Calibration Error (ECE) is commonly used to summarize the discrepancy between confidence and empirical accuracy, while Negative Log-Likelihood (NLL) and Brier score provide complementary assessments of probabilistic prediction quality. Calibration conclusions can depend on the metric and aggregation procedure used [18], [19]. A model with

satisfactory aggregate calibration may therefore still exhibit substantial reliability errors for particular classes or subgroups.

This issue is relevant to autonomous driving because reliability can vary as environmental conditions change, and predictive uncertainty estimates can become less reliable under distribution shift [23], [24]. Existing studies have examined uncertainty in autonomous-driving perception and object detection [14], [25], as well as calibration of predictive uncertainty for motion planning (CCTR) [26]. CCTR uses contextual and historical trajectory information to produce prediction-level temperatures for trajectory-prediction calibration. In contrast, the present analysis uses discrete weather-time-of-day groups and piecewise-constant temperatures to study scene-level heterogeneity and context-routing error in weather classification. Recent work in another safety-relevant machine-learning domain has similarly shown that aggregate calibration can conceal category-specific failures under changing operating conditions [27]. These findings motivate reliability analysis at a finer contextual level rather than reliance on a single aggregate calibration statistic.

For federated long-tailed scene recognition, this leads to an unresolved question. A model may improve recognition of rare or safety-critical scenes while still assigning unreliable confidence to those predictions. Calibration may also be heterogeneous across environmental contexts. A single global transformation may represent dominant scenes adequately while failing to capture less frequent contexts. Conversely, fine-grained scene-specific calibration requires reliable contextual information and sufficient labeled calibration observations.

This study investigates calibration heterogeneity, calibration context, and calibration-data availability in FedRare weather classification. We compare Global Temperature Scaling (GTS), Oracle Scene-Aware Temperature Scaling (Oracle-SATS), Predicted Scene-Aware Temperature Scaling (Predicted-SATS), simpler context controls, and Dirichlet calibration. Oracle-SATS uses the annotated weather-time context as a reference. Predicted-SATS uses predicted weather together with time-of-day metadata. The calibration stage operates centrally on logits exported from the previously trained models.

The analysis asks how recognition errors and calibration vary across weather-time scenes, and how context definition and calibration-data availability affect NLL and Brier score. It also examines whether the resulting probability changes reduce error when predictions are selected by confidence, and whether the score differences persist as scene proportions change. These questions are assessed separately through recognition statistics, calibration errors, probability scores, and coverage-error comparisons.

The experiments distinguish the gains from predicted-class scaling and additional time conditioning, trace the probability losses associated with predicted-context assignments, and compare temperature scaling with Dirichlet calibration. Calibration-sample experiments and a second training seed assess how these findings vary with the available data and the trained model.

The remainder of this paper is organized as follows. Section II reviews related work. Section III defines the reliability problem and metrics. Section IV presents the calibration settings. Section V describes the experimental setup. Section VI reports

the results. Section VII discusses implications and limitations. Section VIII concludes the paper.

II. RELATED WORK

A. Federated Learning and Heterogeneous Client Data

FedAvg established the basic iterative framework in which selected clients perform local optimization and a coordinating server aggregates model updates. FL has since been studied under communication constraints, statistical heterogeneity, partial participation, privacy, and robustness [2]. Non-IID local data can cause local objectives to diverge from the global objective. FedProx introduced a proximal term for heterogeneous local optimization [3], while Agnostic Federated Learning considered optimization with respect to mixtures of client distributions [28].

Client selection is also relevant when only a subset of participants can contribute in each round. Oort introduced guided participant selection based on training utility [29], and Cho et al. analyzed biased client selection and the trade-off between faster convergence and solution bias [30]. These studies mainly evaluate optimization, efficiency, or predictive performance; predictive reliability of the resulting global model is a separate question.

B. Federated Learning for Autonomous Driving

FL has been investigated for connected and automated vehicles, including object detection, collaborative perception, and vehicular crowdsensing [4], [5], [6], [7]. Different vehicles encounter different road, traffic, illumination, and weather conditions, so local datasets can be strongly heterogeneous [4]. FedRare addressed the resulting long-tailed scene-recognition problem through adaptive client selection. The present study does not modify that training procedure; it evaluates predictive reliability for fixed checkpoints of the resulting classifier.

C. Long-Tailed and Adverse Driving Conditions

Long-tailed recognition methods include label-distribution-aware margins, decoupled representation and classifier learning, bilateral architectures, and balanced objectives. BDD100K provides diverse autonomous-driving observations and environmental annotations. Adverse weather and illumination can affect sensing and perception [13], and uncertainty in autonomous-vehicle object detection has also been studied [14]. These observations motivate reliability analysis conditioned on environmental context rather than only aggregate recognition metrics.

D. Confidence Calibration in Machine Learning

A classifier is calibrated when predicted probabilities correspond to empirical outcome frequencies. Modern neural networks can be accurate yet poorly calibrated. TS learns one positive scalar temperature on held-out data and rescales all logits without changing their ordering. Other post-hoc approaches include Bayesian binning [17], Dirichlet calibration [20], ensemble and compositional calibration [21], and spline-based calibration [22].

Recent work has relaxed the single-temperature assumption. Joy et al. proposed sample-dependent adaptive temperature scaling, in which a different temperature is predicted for each input [31]. Balanya et al. studied a broader Adaptive Temperature Scaling family in which temperature is an input-dependent function of classifier outputs, and showed that more expressive calibration can become less reliable when calibration data are limited [32]. The SATS settings examined here fit a constant temperature within each weather–time partition. This permits comparison of environmental grouping, context assignment, and the calibration support available in each group.

Calibration measurement is itself nontrivial: ECE depends on binning and aggregation, and different metrics emphasize different properties. We therefore report confidence ECE, classwise ECE, NLL, and Brier score jointly, while keeping their interpretations separate.

E. Reliability Under Distribution and Context Shift

Deep ensembles provide a scalable approach to predictive uncertainty estimation [23], while evaluations under dataset shift show that uncertainty and calibration quality can deteriorate as the evaluation distribution changes [24]. Autonomous-driving studies have considered uncertainty in object detection, broader perception uncertainty [25], and calibrated trajectory uncertainty for motion planning. Per-category calibration under distribution shift has also been examined in intrusion detection [27]. Across these areas, aggregate probability quality does not necessarily describe reliability for minority or context-specific observations.

F. Research Gap

For long-tailed scene recognition, conditioning temperature on weather and time divides the calibration data into groups of unequal size. When weather is inferred by the classifier, its errors also determine which temperature is applied. The analysis therefore examines context assignment and calibration support together, using global, predicted-class, and joint weather–time scaling with fixed FedRare checkpoints.

III. PROBLEM FORMULATION AND RELIABILITY METRICS

A. Federated Scene Classification

For observation x_i , the fixed classifier produces logits $\mathbf{z}_i = [z_{i,1}, \dots, z_{i,C}]$, where, $C = 6$. The class probability is

$$p_{i,c} = \frac{\exp(z_{i,c})}{\sum_{j=1}^C \exp(z_{i,j})}. \quad (1)$$

The predicted class and confidence are $\hat{y}_i = \arg \max_c p_{i,c}$ and $q_i = \max_c p_{i,c}$. Accuracy measures whether $\hat{y}_i = y_i$, whereas calibration evaluates whether confidence corresponds to empirical correctness frequency [16].

B. Scene-Conditioned Reliability

Each observation has scene descriptor $s_i = (w_i, t_i)$, where, w_i is one of six weather categories and t_i is one of three time-of-day categories, yielding up to 18 scenes. We retain the memberships of the three FedRare evaluation groups: Standard, Diverse, and the rain/snow subset defined in Section V-C. For subset D_s ,

$$\text{Acc}(s) = \frac{1}{|D_s|} \sum_{i \in D_s} \mathbb{I}(\hat{y}_i = y_i), \quad \text{Conf}(s) = \frac{1}{|D_s|} \sum_{i \in D_s} q_i. \quad (2)$$

C. Expected Calibration Error

ECE summarizes the discrepancy between confidence and empirical accuracy [16], [18]. With M bins $B_1, \dots, B_M$,

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|. \quad (3)$$

We use $M = 15$ equal-width bins for the primary confidence ECE results. ECE depends on binning and aggregation [18], [19]. We also calculate classwise ECE [20]:

$$\text{cwECE} = \frac{1}{C} \sum_{c=1}^C \sum_{m=1}^M \frac{|B_{m,c}|}{N} \left| \overline{\mathbb{I}(y=c)}_{B_{m,c}} - \overline{p}_{c,B_{m,c}} \right|. \quad (4)$$

Here $B_{m,c}$ bins the probability of class c across all observations in the evaluated subset, including observations from other true classes. The two within-bin means are the observed class frequency and mean predicted probability. This assesses one-vs-rest calibration; it does not establish calibration conditional on the entire probability vector or on every scene. Sensitivity analyses use 5, 10, 15, and 20 equal-width or quantile-based bins, merging repeated quantile boundaries to keep identical probabilities together.

D. Negative Log-Likelihood

NLL is

$$\text{NLL} = -\frac{1}{N} \sum_{i=1}^N \log p_{i,y_i}. \quad (5)$$

It penalizes low probability assigned to the true class and is particularly sensitive to confident errors.

E. Brier Score

For multiclass classification,

$$\text{Brier} = \frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C (p_{i,c} - \mathbb{I}(y_i = c))^2. \quad (6)$$

It evaluates the full probability vector. NLL and Brier score assess probabilistic prediction quality and are influenced

by properties beyond calibration alone. A decrease in either score therefore does not by itself establish better classwise or scene-conditional calibration. Lower scores indicate better probabilistic predictions under these definitions.

F. Complementary Reliability Evaluation

Confidence ECE and cwECE assess binned calibration and depend on discretization [18], [19], [20]; NLL and Brier score assess the probability vector. Accuracy is also reported because scalar temperature scaling preserves the predicted class, whereas Dirichlet calibration can change it. The metrics are compared on the full evaluation subset and within each scene group.

IV. SCENE-AWARE RELIABILITY CALIBRATION

A. Global Temperature Scaling

For a positive scalar temperature T , calibrated probabilities are

$$p_{i,c}^{(T)} = \frac{\exp(z_{i,c}/T)}{\sum_{j=1}^C \exp(z_{i,j}/T)}. \quad (7)$$

GTS estimates one temperature T_G^* by minimizing NLL on the complete calibration subset [16]. Because a common positive scalar rescales all logits, the predicted class is unchanged.

B. Motivation for Scene-Aware Scaling

A single temperature assumes that the same confidence transformation is suitable across environmental contexts. The scene-level diagnosis in this study tests this assumption. If calibration errors differ by weather and time of day, context-specific temperatures provide a simple way to examine whether calibration heterogeneity can be reduced without retraining the classifier.

C. Oracle Scene-Aware Temperature Scaling

Oracle-SATS partitions calibration samples by their ground-truth scene $s = (w, t)$ and fits

$$T_s^* = \arg \min_{T>0} \sum_{i \in D_s^{\text{cal}}} -\log p_{i,y_i}^{(T)}. \quad (8)$$

An evaluation sample is scaled using the temperature of its ground-truth scene. Since weather is the classification target, this setting uses target information. It is therefore an oracle reference for studying the effect of correct contextual assignment, not a deployable calibration procedure.

D. Predicted Scene-Aware Temperature Scaling

Predicted-SATS removes ground-truth weather from context selection. For sample i ,

$$\hat{w}_i = \arg \max_c z_{i,c}, \quad \hat{s}_i = (\hat{w}_i, t_i), \quad (9)$$

where, t_i is the available time-of-day metadata. The same predicted-context definition is used in both calibration and

evaluation. A separate temperature is fitted for each predicted context by minimizing NLL over calibration samples assigned to that context. If a predicted context is absent from the calibration subset, the global temperature is used:

$$T_{\hat{s}} = \begin{cases} T_{\hat{s}}^*, & N_{\hat{s}}^{\text{cal}} > 0, \\ T_G^*, & N_{\hat{s}}^{\text{cal}} = 0. \end{cases} \quad (10)$$

This design uses predicted weather and supplied time-of-day metadata for context selection. Because predicted weather is the classifier's argmax, the Predicted-weather-only control is predicted-class-conditional temperature scaling. Adding time of day partitions those predicted classes further. Classification errors and temperature selection are linked through these shared outputs, which motivates the subsequent association analysis.

E. Calibration Protocol

All temperatures are fitted only on the calibration subset by minimizing NLL; the network parameters and exported logits remain fixed. Temperature is parameterized as $T = \exp(u)$ and optimized with a bounded one-dimensional search over $u \in [-6, 6]$. The scalar optimizer uses an absolute tolerance of 10^{-12} . Fitting uses a numerically stable log-sum-exp expression for NLL. Reported evaluation NLL clips the true-class probability below at 10^{-15} before taking the logarithm; this matters for the extreme probabilities in the small-budget experiment. Probabilities are computed using a numerically stable softmax. Calibration comparisons use only the frozen evaluation subset. Positive scalar temperatures preserve predicted classes, whereas the Dirichlet transformation below can change them.

F. Calibration-Data Sufficiency

Fine-grained calibration can become unstable when a scene contains few labeled calibration samples. We therefore repeat Oracle-SATS fitting with per-scene budgets $b \in \{1, 2, 5, 10, 20, 50, 100\}$. This experiment uses true scene grouping deliberately: it isolates temperature-estimation uncertainty from predicted-context assignment error.

G. Dirichlet Calibration Baseline

We add Dirichlet calibration with off-diagonal and intercept regularization (Dirichlet-ODIR) [20]. For uncalibrated probabilities $\mathbf{p}_i$, the transformation is

$$\mathbf{p}'_i = \text{softmax}(W \log \mathbf{p}_i + \mathbf{b}), \quad (11)$$

with $W \in \mathbb{R}^{6 \times 6}$ and $\mathbf{b} \in \mathbb{R}^6$. We minimize mean calibration NLL plus

$$\frac{\lambda}{C(C-1)} \sum_{j \neq k} W_{jk}^2 + \frac{\mu}{C} \sum_j b_j^2. \quad (12)$$

Five-fold cross-validation stratified by true weather is confined to the 2,653 calibration observations, using random seed 12321 for the primary checkpoint and 3306 for

the additional checkpoint. Each penalty is selected from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$ by pooled out-of-fold NLL. For the primary archived-logit analysis, the selected values are $\lambda = 0.1$ and $\mu = 10^{-4}$; the transformation is then refitted on all calibration observations using L-BFGS-B. Evaluation labels are not used for fitting or penalty selection. The five calibration foggy observations contribute one observation to each validation fold. This baseline tests a more flexible transformation without retraining the feature extractor.

H. Scope

Calibration is performed centrally on labeled logits from the trained FedRare checkpoints. It assumes access to a labeled calibration collection. The comparisons examine post-hoc transformations of these outputs while keeping the network weights fixed.

V. EXPERIMENTAL SETUP

A. Dataset and Classification Task

Experiments use BDD100K and the same six-class weather task as FedRare: clear, partly cloudy, rainy, snowy, foggy, and overcast. Time of day is categorized as daytime, night, and dawn/dusk, yielding 18 weather–time scenes. After retaining observations with valid required labels, 8,841 samples are used.

BDD100K contains 100,000 video clips from more than 50,000 rides, with the frame at the tenth second annotated for image tasks. All 10,000 local validation annotations contain one frame with timestamp 10,000. Filtering by the required labels reproduces the 8,841 exported records in the same order. Calibration and evaluation share neither annotation identifiers nor byte-identical image files. One byte-identical pair with conflicting overcast/rainy labels occurs within evaluation; its exclusion sensitivity is reported in Section VI-H. Ride/session identifiers and absolute capture times were unavailable in the image and annotation files used for this analysis.

B. Federated Model

The primary checkpoint is the previously trained FedRare model, using a ResNet-18 classifier [33], 20 clients, and 30 communication rounds. We also evaluate an existing checkpoint trained with seed 3306; the archived training programs agree apart from the seed assignment and comments. Both checkpoints are the outputs of their 30-round runs and were not selected using the present calibration results. Network weights remain fixed throughout the post-hoc analyses, with no additional federated training or fine-tuning.

The original FedRare training procedure evaluated model performance on the BDD100K validation data during training. The present study reuses valid observations from that validation data for post-hoc reliability analysis. Consequently, the evaluation subset defined below is held out from the present calibration fitting, but it is not an external or previously unseen test set.

C. Driving-Scene Groups

The Standard group contains clear/daytime, clear/night, snowy/daytime, rainy/daytime, overcast/daytime, and partly cloudy/daytime. The rain/snow subset contains rainy/night, rainy/dawn/dusk, snowy/night, and snowy/dawn/dusk. It has the same membership as the group previously named Safety-Critical in the FedRare protocol; the descriptive name used here specifies its composition. The Diverse group contains the remaining eight scenes: clear/dawn/dusk, overcast/dawn/dusk, partly cloudy/dawn/dusk, partly cloudy/night, foggy/daytime, overcast/night, foggy/night, and foggy/dawn/dusk. The full set contains 7,533 Standard, 620 Diverse, and 688 rain/snow observations. The groups retain the original scene membership, with foggy observations assigned to Diverse. They describe weather and time-of-day composition; the available annotations do not include driving-risk outcomes.

D. Calibration and Held-Out Evaluation Split

The 8,841 observations are partitioned by scene into an approximately 30% calibration subset ($N_{\text{cal}} = 2653$) and a disjoint 70% evaluation subset ($N_{\text{eval}} = 6188$). The split uses NumPy `RandomState(12321)`. Scenes are traversed in their first-occurrence order in the exported logit file; for a scene with N_s observations, $N_s^{\text{cal}} = \max(1, \text{round}(0.30N_s))$ indices are sampled without replacement, and all remaining observations enter the evaluation subset. The exact row-level membership is stored in `split_manifest.csv` and is included in the accompanying reproducibility package.

Table I reports the per-scene counts. The single foggy/dawn/dusk observation is assigned to calibration by the $\max(1, \cdot)$ rule, leaving no evaluation observation for that scene. Oracle-SATS therefore has no foggy/dawn/dusk evaluation case to score. More generally, if an evaluation context has no fitted context-specific temperature, the implementation falls back to the GTS temperature; this rule is used for both oracle and predicted-context code paths. The complete 8,841-sample set is used only for descriptive diagnosis of the fixed uncalibrated model.

TABLE I. PER-SCENE CALIBRATION AND EVALUATION COUNTS

Scene	Cal.	Eval.	Scene	Cal.	Eval.
clear/dawn/dusk	92	215	clear/daytime	529	1235
clear/night	982	2292	foggy/dawn/dusk	1	0
foggy/daytime	2	3	foggy/night	2	5
overcast/dawn/dusk	55	127	overcast/daytime	312	727
overcast/night	5	13	partly cl./dawn/dusk	28	67
partly cl./daytime	191	447	partly cl./night	2	3
rainy/dawn/dusk	16	39	rainy/daytime	119	277
rainy/night	86	200	snowy/dawn/dusk	22	52
snowy/daytime	127	295	snowy/night	82	191
Total calibration: 2653			Total evaluation: 6188		

E. Calibration Settings and Metrics

The primary comparison includes Uncalibrated FedRare, GTS, Oracle-SATS, Predicted-SATS, and Dirichlet-ODIR. Predicted-weather-only and Time-only controls compare the joint grouping with each component. A sparse-context fallback ablation replaces a predicted-context temperature by the GTS temperature when its calibration count is below a threshold.

These comparisons concern post-hoc transformations of the same fixed network outputs.

Recognition is evaluated with accuracy and confusion matrices, probability quality with NLL and multiclass Brier score, and confidence and classwise calibration with ECE and cwECE. We report subgroup results and binning sensitivity alongside the overall metrics. Small-scene accuracy estimates are accompanied by sample counts and Wilson intervals.

F. Calibration-Data Sufficiency Experiment

For Oracle-SATS, per-scene budgets of 1, 2, 5, 10, 20, 50, and 100 labeled calibration samples are tested. If a scene contains fewer than the requested budget, all available samples are used. Each budget is repeated 30 times with random subsampling, and every fitted model is evaluated on the same held-out set. The “All” condition uses the complete calibration subset.

G. Bootstrap Analysis

Paired comparisons use 2,000 bootstrap resamples, with random seeds specified in Section V-J and the same sampled rows used for the compared methods. Point differences are computed on the original evaluation observations, and percentile 95% intervals are obtained from the resamples. The classifier, fitted calibrators, and selection thresholds remain fixed during resampling. This estimates evaluation-sample variability under the assumption of exchangeable rows. Intervals are pointwise, without multiplicity correction, and are computed separately for each training seed.

H. Prediction Selection and Scene Proportions

For the six settings that do not require true weather at inference, we accept a weather prediction when its maximum probability is at least τ , following confidence-based selective classification [34]. Coverage is the accepted fraction; selective error is the fraction of incorrect weather labels among accepted observations. We test fixed thresholds 0.70, 0.80, 0.90, and 0.95, and method-specific calibration-confidence quantiles targeting 50%, 70%, and 90% calibration coverage. These thresholds are fixed before evaluating labels and applied unchanged across subgroups. We report the resulting evaluation coverage. Quantiles use the same calibration observations as the fitted calibrators. Evaluation risk–coverage curves show how selective error changes as the threshold varies, with tied confidences accepted together. Risk here denotes the weather-classification error among accepted predictions.

To test composition sensitivity, we keep the calibrators fixed and assign weight 0.5, 2, or 4 to either night observations or rainy/snowy observations across all times of day, leaving all other weights at one. The two families are evaluated separately, together with the original equal-weight setting. Weighted NLL, Brier score, and accuracy are normalized weighted means; ECE uses weighted bin masses and weighted within-bin means. Paired bootstrap resampling retains these weights and renormalizes them in each replicate. This changes scene proportions while retaining the observed images and within-stratum distributions.

I. Implementation Environment

The experiment used PyTorch on a workstation with an Intel Core i7 processor, 16 GB RAM, and an NVIDIA GeForce RTX 4060 GPU with 8 GB memory.

J. Training-Seed Comparison

We evaluate the checkpoints trained with seeds 12321 and 3306 on the same observations, labels, and frozen calibration/evaluation split. The seed-12321 logits are identical to those used in the primary analysis. Each checkpoint’s calibrators are fitted independently using its calibration observations. For the seed-3306 checkpoint, random seed 3306 is used for Dirichlet cross-validation and bootstrap comparisons, context-comparison bootstrap, and calibration-metric bootstrap. Other analysis seeds remain 12321.

VI. RESULTS AND ANALYSIS

Unless stated otherwise, the results below and Tables I–XIII use the primary archived seed-12321 logits. Section VI-L and Table XIV compare training seeds 12321 and 3306 on the same frozen split.

A. Reliability of the Uncalibrated FedRare Model

Table II reports descriptive results on all 8,841 valid observations. Overall accuracy is 74.51%, whereas mean confidence is 80.98%, giving a confidence–accuracy gap of 6.47 percentage points. The rain/snow recognition rate of 74.13% is the previously reported result for the fixed FedRare predictions; the present analysis adds the corresponding confidence and calibration measurements. In the prior group-level evaluation, the reported recognition rate/recall was computed as the number of correctly classified observations divided by the total number of observations in that scene group. For this single-label multiclass evaluation, that statistic is numerically the same as the group accuracy reported here. The larger gaps in the Diverse and rain/snow subsets show that recognition accuracy alone does not describe predictive reliability.

TABLE II. UNCALIBRATED RELIABILITY ON ALL 8,841 OBSERVATIONS

Group	<i>N</i>	Acc.	Conf.	ECE	NLL	Brier
Overall	8841	.7451	.8098	.0647	.7386	.3711
Standard	7533	.7573	.8111	.0558	.7050	.3566
Diverse	620	.6000	.7524	.1526	1.2138	.5680
Rain/snow	688	.7413	.8466	.1053	.6784	.3523

The group averages also conceal substantial scene variation. For example, partly cloudy/daytime has approximately 18.8% accuracy and ECE 0.520; partly cloudy/dawn/dusk has approximately 14.7% accuracy and ECE 0.520; and overcast/night has approximately 22.2% accuracy and ECE 0.536. By comparison, clear/night has approximately 79.4% accuracy and ECE 0.022. Rainy/night and snowy/night have accuracies of approximately 76.6% and 69.6%, with ECE values of approximately 0.106 and 0.125, respectively. Thus, miscalibration is not homogeneous across scenes. Fig. 1 summarizes the confidence–accuracy gap for all 18 scenes.

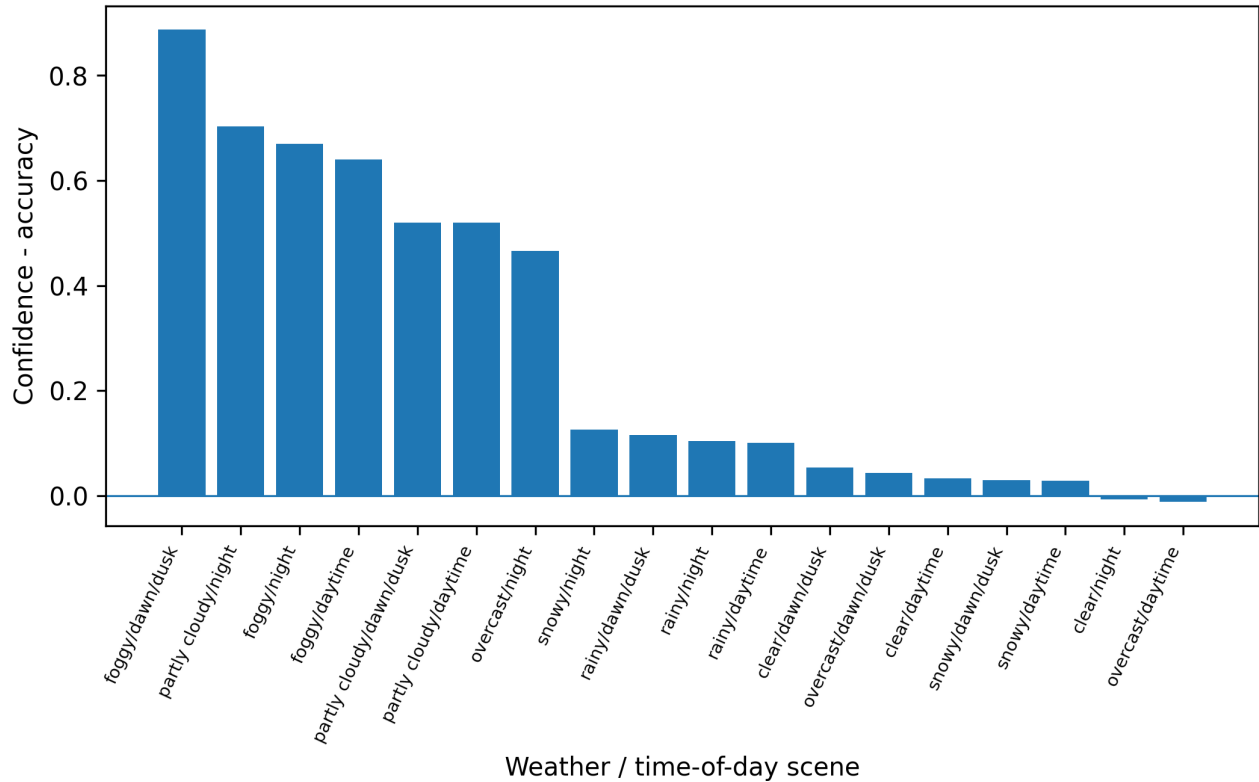


Fig. 1. Confidence–accuracy gap across the 18 weather–time-of-day scenes. Positive values indicate overconfidence.

B. Held-Out Calibration Comparison

Table III reports results on the frozen 6,188-sample evaluation set. The uncalibrated ECE (0.0646) differs slightly from Table II because the latter uses all 8,841 observations. All scalar-temperature settings retain accuracy 0.7471. Dirichlet-ODIR changes 1,038 predicted labels and increases overall accuracy to 0.8253.

TABLE III. CALIBRATION RESULTS ON THE FROZEN EVALUATION SUBSET ($N = 6188$)

Method	Accuracy	ECE	NLL	Brier
Uncalibrated	.7471	.0646	.7387	.3707
GTS	.7471	.0282	.7108	.3660
Oracle-SATS	.7471	.0253	.6464	.3361
Predicted-SATS	.7471	.0248	.6464	.3308
Dirichlet-ODIR	.8253	.0157	.5131	.2595

GTS reduces aggregate confidence ECE relative to the uncalibrated predictions. Oracle-SATS and Predicted-SATS further lower NLL and Brier score. Dirichlet-ODIR has the lowest overall ECE, NLL, and Brier values in Table III. Relative to Predicted-SATS, its NLL difference is -0.1333 (95% interval $[-0.1462, -0.1208]$) and Brier difference is -0.0713 ($[-0.0785, -0.0645]$). The confidence ECE comparison between Predicted-SATS and GTS is examined by bootstrap in Section VI-G.

C. Group-Wise Calibration Behavior

Table IV reports group-wise results and shows that aggregate results do not transfer uniformly to each group. Predicted-SATS improves Standard relative to GTS, but its rain/snow NLL and Brier score are worse than both GTS and Oracle-SATS. Dirichlet-ODIR further lowers the Standard and Diverse scores but has rain/snow NLL 1.1133 and Brier score 0.5861. In that subset, its accuracy falls from 0.7510 to 0.6141: 67 previously correct predictions become incorrect and one previously incorrect prediction becomes correct. The classifier remains fixed, but the flexible calibration transformation changes its output decisions.

Using the original 482 rain/snow evaluation observations, the observed Predicted-SATS minus GTS differences are $+0.0197$ for ECE, $+0.2161$ for NLL, and $+0.0652$ for Brier score. Paired bootstrap resampling is used only to form the 95% percentile intervals: $[-0.0126, 0.0609]$ for ECE, $[0.1630, 0.2723]$ for NLL, and $[0.0521, 0.0782]$ for Brier score. Thus, the NLL and Brier degradation remains positive under sample-level resampling, whereas the ECE interval crosses zero.

D. Context-Assignment Analysis

The predicted overcast/night context has only two calibration observations. None of the rain/snow evaluation observations is assigned to this context, so it does not contribute to the observed rain/snow losses.

We therefore trace rain/snow observations through all predicted contexts and compare their NLL contribution with GTS.

TABLE IV. GROUP-WISE RESULTS ON THE FROZEN EVALUATION SUBSET

Group	Method	<i>N</i>	ECE	NLL	Brier
Standard	Uncal.	5273	.0562	.7088	.3574
	GTS	5273	.0352	.6879	.3549
	Oracle-SATS	5273	.0256	.6221	.3249
	Predicted-SATS	5273	.0287	.5930	.3086
	Dirichlet-ODIR	5273	.0267	.4263	.2125
Diverse	Uncal.	433	.1534	1.1864	.5623
	GTS	433	.0981	1.0923	.5437
	Oracle-SATS	433	.0838	.9875	.4900
	Predicted-SATS	433	.1016	1.0871	.5324
	Dirichlet-ODIR	433	.0827	.9025	.4686
Rain/snow	Uncal.	482	.1048	.6646	.3445
	GTS	482	.0598	.6184	.3276
	Oracle-SATS	482	.0488	.6054	.3207
	Predicted-SATS	482	.0795	.8345	.3928
	Dirichlet-ODIR	482	.1895	1.1133	.5861

Sixty-five rain/snow observations are assigned to predicted clear/night. For these observations, the fitted context temperature is 0.6217 and the sum of sample-wise NLL differences relative to GTS is 81.29. Predicted rainy/night contains 182 rain/snow observations and has a smaller sum of sample-wise NLL differences of 13.67. Other contexts make smaller positive or negative contributions.

A sparse-context fallback provides an additional check. Replacing context-specific temperatures by the global temperature for contexts with fewer than 3, 5, 10, or 20 calibration observations leaves the rain/snow metrics unchanged. A threshold of 50 changes rain/snow NLL from 0.8345 to 0.8250 and Brier score from 0.3928 to 0.3888. The limited recovery indicates that calibration sparsity alone does not explain the observed rain/snow degradation.

E. Context-Granularity Ablation

Table V compares the joint predicted weather–time context with simpler controls. Predicted-weather-only calibration reduces NLL and Brier score relative to GTS, while Time-only gives a lower ECE point estimate than GTS. Joint Predicted-SATS gives the lowest NLL and Brier score among these deployable context definitions, while its ECE is nearly identical to Time-only.

TABLE V. CONTEXT-GRANULARITY ABLATION ON THE FROZEN EVALUATION SUBSET

Method	ECE	NLL	Brier
GTS	.0282	.7108	.3660
Predicted-weather-only	.0294	.6759	.3423
Time-only	.0249	.6972	.3623
Predicted-SATS	.0248	.6464	.3308

Predicted-weather-only scaling conditions on the argmax class and is therefore a predicted-class temperature control. Relative to GTS, it reduces NLL by 0.0349 and Brier score by 0.0237. Adding time of day further reduces NLL by 0.0295 (95% interval $[-0.0361, -0.0236]$) and Brier score

by 0.0115 ($[-0.0132, -0.0098]$). The additional NLL changes are -0.0484 in Standard, $+0.0419$ in Diverse, and $+0.1139$ in the rain/snow subset; the corresponding Brier changes are -0.0163 , $+0.0080$, and $+0.0243$. Thus, the additional overall gain is concentrated in Standard. Along this comparison sequence, predicted-class scaling accounts arithmetically for 54.2% of the total NLL reduction and 67.4% of the Brier reduction from GTS to joint scaling. These fractions describe the sequence of transformations; adding time also changes the partitions and number of fitted temperatures.

Predicted-SATS has a Brier score 0.005326 below Oracle-SATS. Weighting each group’s mean loss difference by its evaluation fraction gives contributions of -0.013911 from Standard, $+0.002970$ from Diverse, and $+0.005615$ from the rain/snow subset. Their sum explains the overall difference. The overall NLL difference is $+0.000050$, with interval $[-0.0116, 0.0126]$. Oracle-SATS fits one scalar per true context by minimizing NLL. Its grouping and fitting objective therefore need not yield the lowest evaluation Brier score.

F. Calibration-Data Sufficiency

Table VI reports the Oracle-SATS sample-budget experiment using the frozen split. With one to five calibration observations per scene, the estimates are highly variable, with elevated NLL values under the probability-clipped NLL definition described above. The metrics generally improve with larger budgets. In this dataset and checkpoint, budgets of 50–100 observations per sufficiently represented scene approach the full-calibration result; this range is descriptive and is not proposed as a universal minimum.

TABLE VI. ORACLE-SATS CALIBRATION-DATA SUFFICIENCY (30 REPEATS)

Per-scene budget	ECE	NLL	Brier
1	.1944 ± .0439	4.2894 ± .8290	.4434 ± .0382
2	.1616 ± .0287	3.7223 ± .9983	.4234 ± .0262
5	.0990 ± .0366	2.3109 ± 1.1743	.3789 ± .0217
10	.0609 ± .0291	1.1526 ± .6161	.3544 ± .0128
20	.0432 ± .0157	.6943 ± .0762	.3427 ± .0056
50	.0302 ± .0061	.6542 ± .0066	.3381 ± .0018
100	.0289 ± .0075	.6499 ± .0037	.3371 ± .0012
All	.0253	.6464	.3361

The experiment shows that fine-grained temperature estimation is sensitive to labeled calibration-data availability, especially in the smallest-sample regime. Fig. 2 visualizes the ECE mean and standard deviation over the 30 repeats.

G. Paired Bootstrap Comparison

Table VII reports paired bootstrap differences for Predicted-SATS minus GTS over the full evaluation subset. The 95% intervals for NLL and Brier score are entirely below zero. The ECE interval crosses zero, so the small aggregate ECE difference is not stable under this resampling analysis.

The bootstrap results show that the aggregate improvement and the rain/snow degradation can coexist for the same fixed predictions.

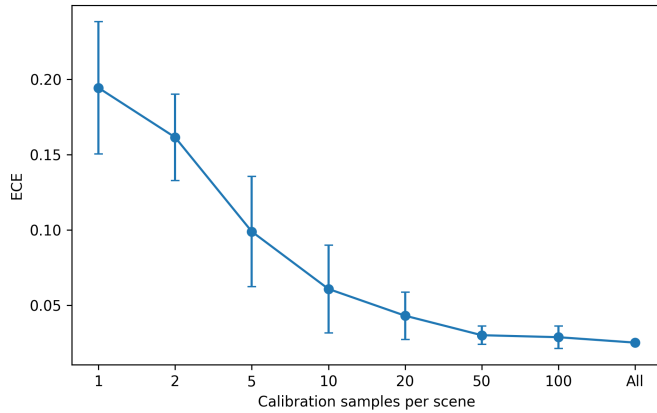


Fig. 2. Oracle-SATS ECE as a function of the per-scene calibration budget. Error bars show the standard deviation over 30 repeats.

TABLE VII. PAIRED BOOTSTRAP: PREDICTED-SATS MINUS GTS. DIFFERENCES ARE COMPUTED FROM UNROUNDED METRIC VALUES; DISPLAYED COMPONENT METRICS ELSEWHERE ARE ROUNDED INDEPENDENTLY

Metric	Difference	95% CI
ECE	-.0035	[-.0159, .0086]
NLL	-.0644	[-.0722, -.0567]
Brier	-.0351	[-.0377, -.0326]

H. Classification Errors and Annotation Ambiguity

Fig. 3 distinguishes full-set diagnostics from held-out evaluation. Among 638 partly cloudy/daytime observations, 120 are correct, 226 are predicted clear, and 241 overcast; the latter two classes account for 90.2% of its 518 errors. Among 18 overcast/night observations, four are correct, eight are predicted clear, and four rainy. All 13 foggy observations are misclassified (nine as rainy, two as clear, two as overcast), and the uncalibrated model never predicts foggy anywhere in the full set. Temperature scaling preserves these errors. On evaluation alone, partly cloudy/daytime accuracy is $83/447 = 18.57\%$, increasing to $227/447 = 50.78\%$ under Dirichlet-ODIR; both settings still have zero correct foggy predictions among eight observations.

We inspected the two highest-confidence errors and two errors randomly sampled from the remaining errors using a fixed seed in each of partly cloudy/daytime and overcast/night, plus all 13 foggy images. Some examples contain overlapping cues: a partly cloudy image predicted snowy shows roadside snow, blue sky, and thin clouds; an overcast/night image predicted rainy shows reflective pavement and little sky. Several foggy images also show wet roads, while nighttime examples offer limited distant-scene detail. These observations suggest single-label ambiguity but neither quantify labeling-error prevalence nor identify the cues used by the model.

The byte-identical evaluation images `c8c0c00c-39ef4894.jpg` and `c8c0c00c-5a90317a.jpg` have different weather annotations (overcast and rainy), establishing one annotation inconsistency. Excluding both records while retaining the fitted temperatures leaves 6,186 evaluation observations.

TABLE VIII. MEAN CLASSWISE ECE WITH 15 EQUAL-WIDTH BINS. RAIN/SNOW DENOTES THE FIXED NIGHT OR DAWN/DUSK SUBSET

Method	Overall	Standard	Diverse	Rain/snow
Uncalibrated	.0559	.0637	.0860	.0642
GTS	.0595	.0693	.0840	.0645
Oracle-SATS	.0580	.0667	.0937	.0675
Predicted-SATS	.0415	.0505	.0779	.0846
Dirichlet-ODIR	.0076	.0151	.0479	.1239

GTS NLL/Brier become 0.7100/0.3658, Oracle-SATS 0.6456/0.3360, and Predicted-SATS 0.6458/0.3306. The main comparisons among these temperature methods persist, and the rain/snow subset is unaffected.

I. Classwise Calibration and Small-Sample Sensitivity

Table VIII separates mean classwise calibration error from confidence ECE. GTS lowers overall confidence ECE but raises overall cwECE from 0.0559 to 0.0595. Predicted-SATS lowers overall cwECE relative to GTS by 0.0181 (95% interval $[-0.0190, -0.0166]$), while increasing it in the rain/snow subset by 0.0201 ($[0.0143, 0.0218]$). Dirichlet-ODIR achieves the smallest overall cwECE but the largest rain/snow value among these methods. Class-specific values in Table XIII evaluate every class probability against all 6,188 evaluation observations. For foggy, the low classwise error occurs with eight positive evaluation examples and zero recall.

The confidence ECE ordering of Oracle-SATS and Predicted-SATS changes across 5, 10, 15, and 20 equal-width or quantile-based bins. Complete binning results and scene counts accompany the analysis. For the full-set diagnostics, overcast/night accuracy 4/18 has a Wilson 95% interval of $[0.090, 0.452]$, while foggy/dawn/dusk accuracy 0/1 has interval $[0, 0.793]$. The foggy/dawn/dusk sample belongs to calibration and has no evaluation counterpart. Its ECE of 0.8874 describes a single observation and provides insufficient support for a scene-level calibration comparison.

J. Confidence-Based Weather Prediction Selection

Table IX reports the operating point targeting 70% coverage on each method's calibration confidences. Overall evaluation coverage is approximately 70% for all three displayed methods, but the accepted observations can differ. Predicted-SATS lowers overall selective error from 16.34% under GTS to 11.92%, with a paired difference of -4.42 percentage points (95% interval $[-5.08, -3.75]$). In the rain/snow subset, coverage is 76.97% for GTS and 75.31% for Predicted-SATS, while selective error rises from 14.29% to 19.28%; the difference is $+5.00$ points ($[2.44, 7.74]$). Dirichlet-ODIR has still lower overall error but higher rain/snow error and lower coverage. The same direction of the GTS–Predicted-SATS error comparison occurs at calibration targets of 50% and 90%, although subgroup coverages differ. Fig. 4 shows the full descriptive curves for all six applicable settings.

K. Sensitivity to Scene Proportions

Table X reports Predicted-SATS minus GTS under the specified weighting settings. Changing night weights from 0.5 to 2 and 4 produces night proportions of 27.96%, 60.82%, and

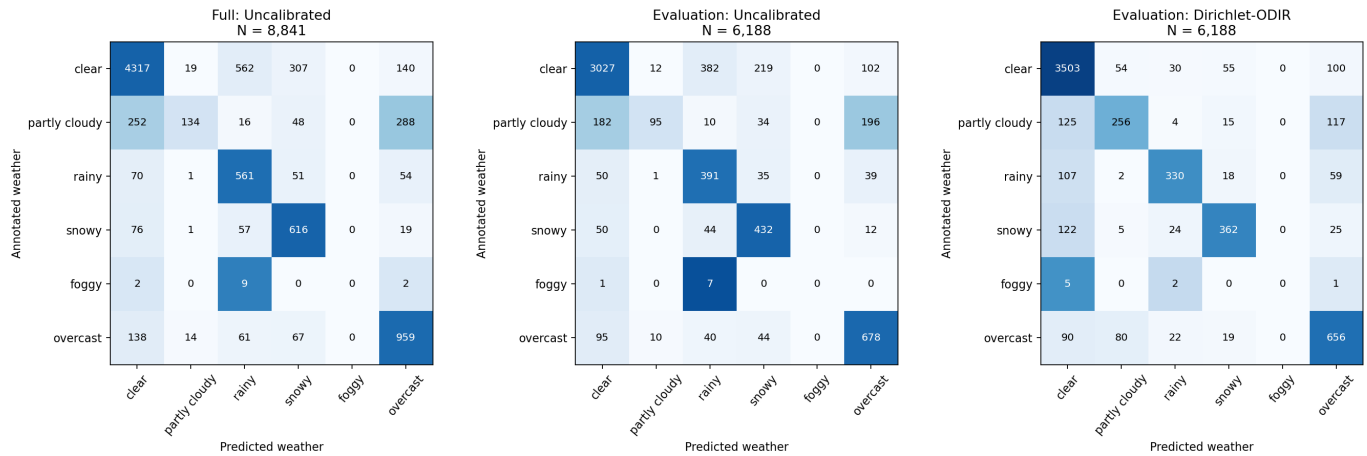


Fig. 3. Weather confusion matrices for the uncalibrated full set, uncalibrated evaluation subset, and Dirichlet-ODIR evaluation subset. Rows are annotated classes and columns are predictions. Numbers are counts; shading is normalized within each true class. Temperature scaling preserves the uncalibrated class assignments.

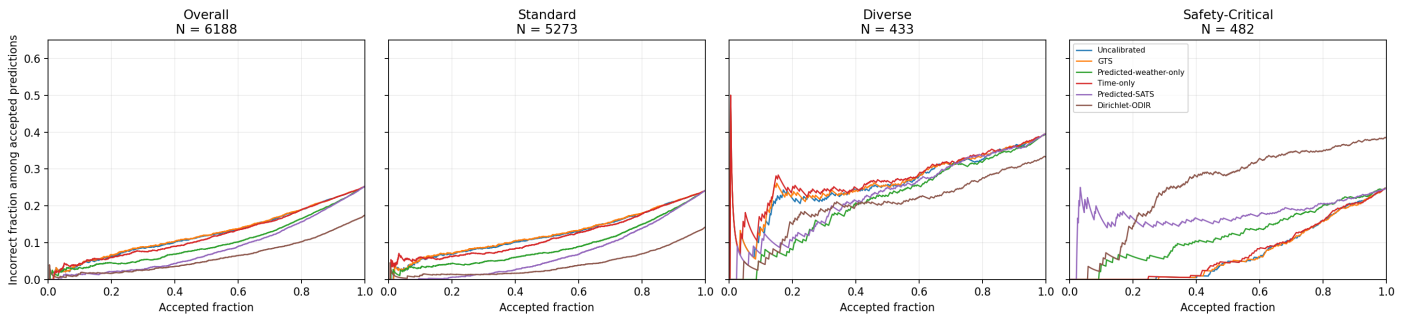


Fig. 4. Selective weather-classification error versus coverage on evaluation observations. Each curve varies the confidence threshold within the displayed subset and keeps ties together. These are descriptive curves; the operating points in Table IX instead use thresholds fixed from calibration confidences and shared across subsets.

TABLE IX. SELECTION WITH METHOD-SPECIFIC THRESHOLDS TARGETING 70% CALIBRATION COVERAGE. EVALUATION COVERAGE AND ERROR ARE PERCENTAGES; ACCEPTED COUNTS ARE SHOWN EXPLICITLY

Subset	Method	Accepted	Coverage	Error
Overall	GTS	4357/6188	70.41	16.34
Overall	Predicted-SATS	4344/6188	70.20	11.92
Overall	Dirichlet-ODIR	4362/6188	70.49	8.34
Rain/snow	GTS	371/482	76.97	14.29
Rain/snow	Predicted-SATS	363/482	75.31	19.28
Rain/snow	Dirichlet-ODIR	307/482	63.69	33.22

75.64%, compared with 43.70% originally; the overall NLL and Brier advantages persist. Changing rainy/snowy weights across all times of day produces combined proportions of 9.31%, 29.11%, and 45.09%, compared with 17.03% originally. Increasing this proportion reduces both advantages. At weight 4, the NLL difference is -0.0046 with interval $[-0.0173, 0.0090]$, while the Brier difference remains below zero. In that setting, Dirichlet-ODIR minus Predicted-SATS is $+0.0038$ for NLL $([-0.0155, 0.0224])$ and $+0.0054$ for Brier $([-0.0042, 0.0147])$; the intervals do not establish a ranking between them. Thus, the overall comparisons depend on scene

composition, even with the classifier and calibration functions fixed.

L. Repetition Across Training Seeds

Table XIV shows that Predicted-SATS lowers overall NLL and Brier score relative to GTS while increasing both in the rain/snow subset for each checkpoint. All of these within-seed paired intervals exclude zero. Overall cWECE also decreases and rain/snow cWECE increases for both seeds. Predicted-class scaling explains part of the overall gains, with additional NLL/Brier reductions after time conditioning in both repetitions. Dirichlet-ODIR has better overall scores but larger rain/snow losses for both checkpoints; at seed 3306 its overall NLL/Brier are $0.5111 / 0.2596$ and rain/snow values are $1.1323 / 0.5892$, with rain/snow accuracy falling from 0.7635 to 0.5934 .

The confidence and selection results are less consistent. Seed 3306 has an overall Predicted-SATS minus GTS confidence ECE difference of $+0.0245$ (95% interval $[0.0115, 0.0358]$); the corresponding seed-12321 interval spans zero. At 70% target calibration coverage, overall selective error decreases for both seeds, but the rain/snow difference is $+5.00$ percentage points for seed 12321 $([2.44, 7.74])$ and $+0.40$ points for seed 3306 $([-2.32, 3.12])$. Thus, rain/snow selection-

TABLE X. CONTROLLED SCENE-PROPORTION CHANGES: PREDICTED-SATS MINUS GTS. WEIGHTS APPLY TO ONE NAMED STRATUM AT A TIME; OTHER RECORDS RETAIN WEIGHT ONE. INTERVALS USE PAIRED ROW BOOTSTRAP WITH FIXED CALIBRATORS

Weighted stratum	NLL difference	95% interval	Brier difference	95% interval
Original composition	-.0644	[-.0721, -.0569]	-.0351	[-.0377, -.0327]
Night $\times 0.5$	-.0489	[-.0562, -.0420]	-.0256	[-.0280, -.0234]
Night $\times 2$	-.0812	[-.0904, -.0717]	-.0455	[-.0486, -.0426]
Night $\times 4$	-.0958	[-.1066, -.0845]	-.0545	[-.0582, -.0511]
Rainy/snowy $\times 0.5$	-.0808	[-.0876, -.0738]	-.0417	[-.0442, -.0394]
Rainy/snowy $\times 2$	-.0387	[-.0483, -.0287]	-.0248	[-.0278, -.0219]
Rainy/snowy $\times 4$	-.0046	[-.0173, .0090]	-.0112	[-.0149, -.0075]

error degradation is not established in the second checkpoint. At rainy/snowy weight 4, the overall NLL-difference interval spans zero for both seeds, while the Brier difference remains negative. The two checkpoints show the same direction of NLL and Brier changes, while confidence ECE and selective error differ.

VII. DISCUSSION

A. Recognition Performance and Confidence Reliability Are Distinct

FedRare was designed to improve recognition of rare and safety-critical scenes. The present analysis shows that this recognition objective does not imply calibrated confidence. This is consistent with the broader finding that accurate neural networks can remain miscalibrated. The distinction is visible in the Diverse and rain/snow confidence–accuracy gaps and supports reporting reliability separately from recognition performance.

B. Reliability is Context-Dependent

BDD100K contains heterogeneous environmental conditions. In the present data, the imbalance is explicit: the Standard group contains 7,533 of 8,841 observations (Table II), so aggregate results are weighted primarily by common scenes. The scene-level results therefore provide information that is not visible in the aggregate values.

C. Different Metrics Give Different Conclusions

Dirichlet-ODIR has the lowest overall probability scores and calibration errors in the primary comparison, but the largest rain/snow values. Predicted-SATS also lowers overall NLL, Brier score, and cwECE relative to GTS while increasing these measures in the rain/snow subset for both training seeds. Confidence ECE behaves differently: its overall difference from GTS is uncertain for seed 12321 and positive for seed 3306. These results distinguish changes in probability quality from changes in binned calibration error.

D. Context Assignment

Oracle-SATS and Predicted-SATS differ in how weather context is obtained. In the primary analysis, observations assigned to predicted clear/night make the largest contribution to the increase in rain/snow NLL under Predicted-SATS. Replacing temperatures for sparse contexts with the global temperature changes the losses only slightly. Context assignment therefore needs to be examined alongside calibration

support. Because weather prediction and context assignment use the same outputs, their separate contributions to the loss differences cannot be isolated by this comparison.

E. Calibration Granularity and Data Sufficiency

Predicted-class scaling explains part of the overall improvement. Adding time of day lowers overall NLL and Brier score further, but increases these losses in Diverse and the rain/snow subset for both checkpoints. Finer partitions also leave fewer observations for each temperature estimate. The sample-budget experiment shows that estimates are particularly variable at the smallest calibration budgets.

F. Implications for Federated Autonomous Driving

Confidence-based selection changes both the fraction of accepted predictions and their error rate within each scene group. In the primary analysis, overall selective error decreases under Predicted-SATS while rain/snow error increases; the rain/snow difference is uncertain in the second checkpoint. The weighting experiment also shows that overall score differences vary with scene composition. Reporting these outcomes alongside recognition and calibration measurements helps assess how weather probabilities behave when used for prediction selection, complementing uncertainty analysis in driving perception [14], [25], [26].

G. Limitations

The evaluation covers two FedRare training seeds and six-class weather recognition on BDD100K. More training runs and other models are needed to assess variation beyond this comparison. The validation source was used for training monitoring, so the evaluation is held out from calibration fitting rather than being a previously unseen test set. Ride/session identifiers were unavailable, so possible overlap of recording sessions between calibration and evaluation could not be checked. The bootstrap assumes exchangeable images and conditions on the fitted models; it does not include training or calibration-fit variability or potential session dependence. Rare scenes provide limited support, particularly fog, for which both checkpoints have zero evaluation recall before and after Dirichlet calibration. The decision experiment evaluates weather-prediction selection; vehicle planning, control, and accident outcomes were not evaluated.

H. Future Work

Future work can examine probabilistic rather than hard context assignment, regularized or hierarchical calibration that shares information across sparse scenes, online calibration under changing conditions, and evaluation across additional independently trained checkpoints and federated methods. Reliability measurements could also be examined as inputs to client-selection or aggregation objectives.

VIII. CONCLUSION

The analysis of two FedRare training seeds shows lower overall NLL, Brier score, and cwECE under Predicted-SATS than GTS, together with higher values in the rain/snow subset at night or dawn/dusk. Predicted-class scaling accounts for part of the overall reduction, and additional time conditioning lowers the scores further. Dirichlet calibration has lower overall probability losses but higher rain/snow losses and lower recognition accuracy in that subset. Increasing the proportion of rainy/snowy observations across all times of day reduces the overall NLL advantage over GTS for both seeds. Confidence ECE and selective error vary between checkpoints, and small calibration subsets produce unstable temperature estimates. The results support evaluating overall probability scores together with calibration, subgroup performance, and the coverage and error of selected predictions.

SUPPLEMENTARY RESULTS

Table XI and Table XII report the complete uncalibrated scene-level statistics underlying Fig. 1. These are descriptive full-set results, including calibration observations. The very small foggy and partly-cloudy/night scenes require interpretation with their counts and do not support scene-level ECE rankings. Table XIII reports class-specific calibration errors on evaluation only. Table XIV reports the comparison across training seeds.

TABLE XI. COMPLETE SCENE-LEVEL RELIABILITY STATISTICS, PART I

Scene	<i>N</i>	Acc.	Conf.	Gap	ECE	NLL	Brier
clear/dawn/dusk	307	0.7394	0.7932	0.0538	0.0715	0.6832	0.3682
clear/daytime	1764	0.8447	0.8779	0.0332	0.0379	0.4341	0.2267
clear/night	3274	0.7941	0.7869	-0.0073	0.0218	0.5309	0.2963
foggy/dawn/dusk	1	0.0000	0.8874	0.8874	0.8874	4.8775	1.7760
foggy/daytime	5	0.0000	0.6404	0.6404	0.6404	3.2560	1.3023
foggy/night	7	0.0000	0.6702	0.6702	0.6702	3.7478	1.5026
overcast/dawn/dusk	182	0.6978	0.7411	0.0433	0.1119	0.8890	0.4426
overcast/daytime	1039	0.7969	0.7846	-0.0123	0.0558	0.6507	0.3250
overcast/night	18	0.2222	0.6876	0.4654	0.5356	3.4617	1.1414

TABLE XII. COMPLETE SCENE-LEVEL RELIABILITY STATISTICS, PART II

Scene	<i>N</i>	Acc.	Conf.	Gap	ECE	NLL	Brier
partly cl./dawn/dusk	95	0.1474	0.6676	0.5202	0.5202	2.6579	1.1844
partly cl./daytime	638	0.1881	0.7076	0.5195	0.5195	2.5776	1.1813
partly cl./night	5	0.0000	0.7034	0.7034	0.7034	3.7663	1.3355
rainy/dawn/dusk	55	0.7091	0.8242	0.1151	0.1600	0.9202	0.4257
rainy/daytime	396	0.7652	0.8652	0.1001	0.1073	0.7836	0.3450
rainy/night	286	0.7657	0.8699	0.1041	0.1063	0.6004	0.3299
snowy/dawn/dusk	74	0.8378	0.8668	0.0290	0.0657	0.3362	0.1932
snowy/daytime	422	0.8626	0.8911	0.0285	0.0497	0.4178	0.2099
snowy/night	273	0.6960	0.8212	0.1253	0.1253	0.8041	0.4041

TABLE XIII. ONE-Vs-REST CLASS ECE ON ALL 6,188 EVALUATION OBSERVATIONS, WITH 15 EQUAL-WIDTH BINS. N_+ IS THE TRUE-CLASS COUNT; IT IS NOT THE DENOMINATOR OF CLASS ECE

Class	N_+	GTS	Predicted-SATS	Dirichlet-ODIR
clear	3742	.1289	.0768	.0116
partly cloudy	517	.0472	.0430	.0083
rainy	516	.0878	.0594	.0088
snowy	538	.0563	.0403	.0060
foggy	8	.0195	.0170	.0006
overcast	867	.0173	.0122	.0105

TABLE XIV. COMPARISON ACROSS TRAINING SEEDS: PREDICTED-SATS MINUS GTS. OVERALL $N=6188$; RAIN/SNOW $N=482$. INTERVALS ARE CONDITIONAL WITHIN-SEED PAIRED BOOTSTRAP INTERVALS

Seed	Subset	Score	Difference	95% interval
12321	Overall	NLL	-0.0644	[-0.0722, -0.0567]
12321	Overall	Brier	-0.0351	[-0.0377, -0.0326]
12321	Rain/snow	NLL	0.2161	[0.1630, 0.2723]
12321	Rain/snow	Brier	0.0652	[0.0521, 0.0782]
3306	Overall	NLL	-0.0814	[-0.0943, -0.0677]
3306	Overall	Brier	-0.0709	[-0.0747, -0.0673]
3306	Rain/snow	NLL	0.3215	[0.2131, 0.4309]
3306	Rain/snow	Brier	0.0518	[0.0339, 0.0696]

DECLARATIONS

Declaration on Generative AI: Generative AI tools were used for language refinement, manuscript preparation, and assistance with supplementary analysis code and interpretation. The authors take responsibility for the experimental protocol, verification of the analyses, and final interpretation and conclusions.

ACKNOWLEDGMENT

This work was supported by Hubei Province Science and Technology Department Project(2025DJC021), Natural Science Foundation Innovation and Development Joint Fund of Hubei Province in China(2025AFD031), Hubei Province Science and Technology Department Project(2025DJC020).

REFERENCES

- [1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, PMLR, vol. 54, 2017, pp. 1273–1282.
- [2] P. Kairouz, H. B. McMahan, B. Avent, *et al.*, "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, nos. 1–2, pp. 1–210, 2021, doi: 10.1561/22000000083.
- [3] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Machine Learning and Systems (MLSys)*, vol. 2, 2020.
- [4] V. P. Chellapandi, L. Yuan, C. G. Brinton, S. H. Žak, and Z. Wang, "Federated learning for connected and automated vehicles: A survey of existing approaches and challenges," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 119–137, 2024, doi: 10.1109/TIV.2023.3332675.
- [5] D. Jallepalli, N. C. Ravikumar, P. V. Badarinath, S. Uchil, and M. A. Suresh, "Federated learning for object detection in autonomous vehicles," in *Proc. IEEE 7th Int. Conf. Big Data Computing Service and Applications (BigDataService)*, 2021, pp. 107–114, doi: 10.1109/BigDataService52369.2021.00018.

- [6] S. Wang, C. Li, D. W. K. Ng, Y. C. Eldar, H. V. Poor, Q. Hao, and C. Xu, "Federated deep learning meets autonomous vehicle perception: Design and verification," *IEEE Network*, vol. 37, no. 3, pp. 16–25, 2023, doi: 10.1109/MNET.104.2100403.
- [7] Y. Ma, S. Hu, Z. Fang, Y. Ji, Y. Deng, and Y. Fang, "Sense4FL: Vehicular crowdsensing enhanced federated learning for object detection in autonomous driving," *IEEE Transactions on Mobile Computing*, vol. 25, no. 8, pp. 13004–13018, 2026, doi: 10.1109/TMC.2026.3674333.
- [8] K. Cao, C. Wei, A. Gaidon, N. Aréchiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [9] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2020.
- [10] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen, "BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9716–9725, doi: 10.1109/CVPR42600.2020.00974.
- [11] J. Ren, C. Yu, S. Sheng, X. Ma, H. Zhao, S. Yi, and H. Li, "Balanced Meta-Softmax for long-tailed visual recognition," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [12] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2633–2642, doi: 10.1109/CVPR42600.2020.00271.
- [13] Y. Zhang, A. Carballo, H. Yang, and K. Takeda, "Perception and sensing for autonomous vehicles under adverse weather conditions: A survey," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 196, pp. 146–177, 2023, doi: 10.1016/j.isprsjprs.2022.12.021.
- [14] L. Peng, H. Wang, and J. Li, "Uncertainty evaluation of object detection algorithms for autonomous vehicles," *Automotive Innovation*, vol. 4, pp. 241–252, 2021, doi: 10.1007/s42154-021-00154-0.
- [15] S. Wen and X. Wang, "An adaptive client selection method for long tailed scene classification in autonomous driving," *Scientific Reports*, 2026, doi: 10.1038/s41598-026-62562-8.
- [16] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Machine Learning (ICML)*, PMLR, vol. 70, 2017, pp. 1321–1330.
- [17] M. P. Naeini, G. F. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using Bayesian binning," in *Proc. 29th AAAI Conf. Artificial Intelligence*, vol. 29, no. 1, 2015, pp. 2901–2907, doi: 10.1609/aaai.v29i1.9602.
- [18] J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran, "Measuring calibration in deep learning," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 38–41.
- [19] A. Kumar, P. Liang, and T. Ma, "Verified uncertainty calibration," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [20] M. Kull, M. Perello-Nieto, M. Kängsepp, T. Silva Filho, H. Song, and P. Flach, "Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [21] J. Zhang, B. Kailkhura, and T. Y.-J. Han, "Mix-n-Match: Ensemble and compositional methods for uncertainty calibration in deep learning," in *Proc. 37th Int. Conf. Machine Learning (ICML)*, PMLR, vol. 119, 2020, pp. 11117–11128.
- [22] K. Gupta, A. Rahimi, T. Ajanthan, T. Mensink, C. Sminchisescu, and R. Hartley, "Calibration of neural networks using splines," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.
- [23] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [24] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [25] B. Araújo, J. F. Teixeira, J. Fonseca, R. Cerqueira, and S. C. Beco, "The road to safety: A review of uncertainty and applications to autonomous driving perception," *Entropy*, vol. 26, no. 8, Art. no. 634, 2024, doi: 10.3390/e26080634.
- [26] C. Cao, X. Chen, J. Wang, Q. Song, R. Tan, and Y.-H. Li, "CCTR: Calibrating trajectory prediction for uncertainty-aware motion planning in autonomous driving," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 38, no. 19, pp. 20949–20957, 2024, doi: 10.1609/aaai.v38i19.30085.
- [27] K. Alalawi, "When calibrated detectors meet new attacks: Per-category reliability of machine-learning intrusion detection under distribution shift," *International Journal of Advanced Computer Science and Applications*, vol. 17, no. 8, 2026, doi: 10.14569/IJACSA.2026.0170812.
- [28] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *Proc. 36th Int. Conf. Machine Learning (ICML)*, PMLR, vol. 97, 2019, pp. 4615–4625.
- [29] F. Lai, X. Zhu, H. V. Madhyastha, and M. Chowdhury, "Oort: Efficient federated learning via guided participant selection," in *Proc. 15th USENIX Symp. Operating Systems Design and Implementation (OSDI)*, 2021, pp. 19–35.
- [30] Y. J. Cho, J. Wang, and G. Joshi, "Towards understanding biased client selection in federated learning," in *Proc. 25th Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, PMLR, vol. 151, 2022, pp. 10351–10375.
- [31] T. Joy, F. Pinto, S.-N. Lim, P. H. S. Torr, and P. K. Dokania, "Sample-dependent adaptive temperature scaling for improved calibration," *Proc. AAAI Conf. Artificial Intelligence*, vol. 37, no. 12, pp. 14919–14926, 2023, doi: 10.1609/aaai.v37i12.26742.
- [32] S. A. Balanya, J. Maroñas, and D. Ramos, "Adaptive temperature scaling for robust calibration of deep neural networks," *Neural Computing and Applications*, vol. 36, pp. 8073–8095, 2024, doi: 10.1007/s00521-024-09505-4.
- [33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
- [34] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.