RESFL: AN UNCERTAINTY-AWARE FRAMEWORK FOR RESPONSIBLE FEDERATED LEARNING BY BALANCING PRIVACY, FAIRNESS AND UTILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated Learning (FL) has gained prominence in machine learning applications across critical domains, offering collaborative model training without centralized data aggregation. However, FL frameworks that protect privacy often sacrifice fairness and reliability; differential privacy reduces data leakage but hides sensitive attributes needed for bias correction, worsening performance gaps across demographic groups. This work explores the trade-off between privacy and fairness in FL-based object detection and introduces RESFL, an integrated solution optimizing both. RESFL incorporates *adversarial privacy disentanglement* and *uncertainty-guided fairness-aware aggregation*. The adversarial component uses a gradient reversal layer to remove sensitive attributes, reducing privacy risks while maintaining fairness. The uncertainty-aware aggregation employs an *evidential neural network* to weight client updates adaptively, prioritizing contributions with lower fairness disparities and higher confidence. This ensures robust and equitable FL model updates. We demonstrate the effectiveness of RESFL in high-stakes autonomous vehicle scenarios, where it achieves high mAP on FACET and CARLA, reduces membership-inference attack success by 37%, reduces equality-of-opportunity gap by 17% relative to the FedAvg baseline, and maintains superior adversarial robustness. However, RESFL is inherently domain-agnostic and thus applicable to a broad range of application domains beyond autonomous driving.

1. INTRODUCTION

Federated Learning (FL) has emerged as a promising solution to privacy concerns by enabling decentralized model training, ensuring data remains on local devices. In contrast, only model updates, such as gradients or weight deltas, are shared for aggregation. This paradigm not only reduces the risk of raw data exposure but also supports collaborative learning across heterogeneous and sensitive data silos in domains like healthcare, finance, and smart cities. However, the inherent obfuscation of sensitive attributes such as demographic labels or personal identifiers introduces a critical trade-off: fairness interventions often require direct access to these attributes to detect and correct biases. By withholding sensitive information in pursuit of privacy preservation, FL frameworks inadvertently hamper bias mitigation strategies, leading to disparate model performance across groups defined by age, gender, or ethnicity (Kaplan, 2024; Zhang et al., 2024).

The problem is exacerbated by external uncertainties in real-world data collection and inference, which undermine model confidence and reliability. In safety-critical applications, input data are affected by sensor noise, environmental variability (e.g., lighting, weather, occlusions), and domain shift between training and deployment. Unmodeled, these uncertainties can disproportionately degrade performance for subpopulations, amplifying disparities. For example, under foggy or low-light conditions, object detection models

show higher false-negative rates for pedestrians with darker skin tones, compounding risks for vulnerable groups. Quantifying both epistemic and aleatoric uncertainty is therefore essential to ensure equitable reliability across demographic cohorts (Pathiraja et al., 2024).

While a growing body of work has advanced privacy-preserving techniques, such as differential privacy, secure multi-party computation, and homomorphic encryption, and fairness-aware methods, such as pre-processing transformations, in-processing regularizers, and post-processing adjustments, these solutions frequently optimize one objective at the expense of others. Differential privacy mechanisms can effectively limit membership inference and attribute leakage, but often degrade model utility and exacerbate fairness disparities by obscuring minority data patterns (Sun et al., 2021; Xin et al., 2020). Conversely, fairness-oriented re-weighting or regularization approaches can narrow demographic gaps but may inadvertently expose sensitive information if not carefully integrated. Centralized learning paradigms further magnify these tensions by aggregating unprotected data, while many federated solutions prioritize privacy guarantees without explicitly addressing equitable performance across groups (Chen et al., 2025; Ezzeldin et al., 2023; Yu et al., 2020). Post hoc fairness corrections or hard constraints also struggle to capture the nuanced interplay between privacy protection and bias mitigation in decentralized environments (Kim et al., 2024a).

To address these limitations, we propose `RESFL`, a domain-agnostic federated learning framework that jointly optimizes privacy and group fairness by integrating two complementary components within a single pipeline: (i) an adversarial representation module with gradient reversal that suppresses sensitive-attribute signals in shared representations, and (ii) an uncertainty-guided aggregation mechanism that leverages evidential uncertainty (via a scale-invariant uncertainty fairness metric (UFM)) to up-weight client updates exhibiting lower inter-group disparity and higher confidence. This unified design yields privacy-preserving, equitable, and reliable updates without sacrificing utility. We validate `RESFL` on autonomous vehicle (AV) scenarios to demonstrate its effectiveness in safety-critical and diverse environments. Empirically, `RESFL` delivers strong accuracy while reducing fairness gaps and privacy leakage, and remains robust under distribution shifts (weather variations). Across both FACET and CARLA, it consistently outperforms standard and state-of-the-art FL baselines on utility, fairness, and privacy.

2. Related Work

Federated Learning. Federated Learning (FL) is a decentralized training paradigm where multiple clients collaboratively train a shared model while keeping data on-device. This mitigates privacy risks of centralized aggregation but introduces challenges, particularly data heterogeneity, as clients typically hold non-IID data (Yang et al., 2023). Early research focused on improving communication efficiency and convergence under heterogeneous (non-IID) client data (Li et al., 2019; Karimireddy et al., 2020; Martinez et al., 2020). However, FL introduces new challenges beyond optimization, including privacy leakage, performance disparities across participants, and fairness across sensitive demographic groups (Wasif et al., 2025).

Privacy Preservation Techniques. Preserving user privacy is a core objective in FL. Differential Privacy (DP) is widely used, adding calibrated noise to model updates to ensure formal privacy guarantees (Dwork, 2006), though it can degrade model utility (Bagdasaryan et al., 2019). Alternative approaches, such as homomorphic encryption (HE) (Yi et al., 2014) and secure multi-party computation (SMC) (Tran et al., 2023), provide strong guarantees but suffer high computational costs and limited scalability (Chen et al., 2023; Xu et al., 2021). Consequently, research has shifted toward perturbation-based methods, including shuffler models, that strive to balance privacy, utility, and communication efficiency (Chen et al., 2024; Erlingsson et al., 2019; Kim et al., 2024b). However, most privacy solutions in FL neglect fairness, risking inequitable outcomes despite strong privacy protections.

Fairness in Federated Learning. Fairness in machine learning has been extensively explored in centralized settings, with numerous methods to mitigate biases against underrepresented groups (Hardt et al., 2016; Kairouz et al., 2021; Mehrabi et al., 2021). In FL, researchers distinguish between *client fairness*, which

aims for uniform model performance across data-silo clients (Yu et al., 2020; Karimireddy et al., 2020), and *group fairness*, which seeks equitable outcomes across sensitive demographic cohorts despite decentralized data (Kairouz et al., 2021). Traditional fairness strategies such as constrained optimization or regularization work well in centralized frameworks (Wu et al., 2018) but falter in FL, since the server lacks direct access to sensitive attributes for bias measurement (McMahan et al., 2017), and client-level equalization does not guarantee demographic parity, underscoring the need for FL-specific fairness mechanisms.

Privacy-preserving & Fair FL. Given the inherent tension between privacy and fairness, recent research has explored joint approaches to address both in FL. Differentially private algorithms can worsen fairness disparities by masking minority-group patterns (Zhang et al., 2021), while fairness-aware techniques may increase privacy risks by exposing sensitive attributes. Prior work includes FairDP-SGD and FairPATE for centralized settings (Yaghini et al., 2023), FPFL which enforces group fairness under DP guarantees but at high communication cost (Rodríguez-Gálvez et al., 2021), and two-step schemes that align a privacy-protected model with a fair proxy (Sun et al., 2023; Pujol et al., 2020). Pre- and post-processing defenses like (Pentyala et al., 2022) and (Corbucci et al., 2024) also integrate privacy and fairness but often incur computational overhead or limited scalability (Imteaj et al., 2021). Despite these advances, many approaches struggle to scale or balance the trade-offs effectively, motivating our unified `RESFL` framework.

Building on these insights, our proposed `RESFL` framework overcomes these limitations by integrating privacy preservation and *group fairness* optimization into a single, end-to-end FL algorithm.

3. METHODOLOGY

This section introduces our integrated privacy-preserving and fairness-aware Federated Learning framework, *responsible FL* (`RESFL`). Our approach tackles two key challenges: (i) preventing sensitive attribute leakage during training to ensure privacy and (ii) mitigating bias in client updates to ensure *group fairness*. To achieve this, we integrate adversarial privacy disentanglement with uncertainty-guided fairness-aware aggregation using an evidential neural network (ENN), enabling the estimates of epistemic uncertainty (Amini et al., 2020). The flow of the `RESFL` algorithm is depicted in Figure 1.

3.1. UNCERTAINTY FAIRNESS METRIC (UFM) FOR GROUP-FAIR AGGREGATION

Evidential Uncertainty Modeling. In `RESFL`, each client replaces its standard softmax detection head with an evidential output layer that predicts a nonnegative concentration vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_C)$ for C object classes. These concentration parameters parameterize a Dirichlet distribution over the categorical probability simplex, allowing closed-form computation of epistemic uncertainty without resorting to costly Monte Carlo sampling or deep ensembles. Formally, for each input x, the evidential head produces

$$p(\mathbf{p} \mid \boldsymbol{\alpha}) = \frac{\Gamma\Big(\sum_{c=1}^{C} \alpha_c\Big)}{\prod_{c=1}^{C} \Gamma(\alpha_c)} \prod_{c=1}^{C} p_c^{\alpha_c - 1}, \tag{1}$$

where C is the number of considered classes, $\Gamma(\cdot)$ denotes the Gamma function, and $\mathbf{p} = (p_1, \ldots, p_C)$ represents the class probabilities. The total evidence $\alpha_0 = \sum_{c=1}^{C} \alpha_c$ directly yields an analytic estimate of the approximate epistemic variance (Sensoy et al., 2018),

$$\sigma^2_{\mathrm{epi},c} = \mathbb{E}[p_c]\big(1 - \mathbb{E}[p_c]\big) \cdot \frac{1}{\alpha_0 + 1} = \frac{\alpha_c}{\alpha_0}\left(1 - \frac{\alpha_c}{\alpha_0}\right) \cdot \frac{1}{\alpha_0 + 1} \sim \frac{1}{\alpha_0}, \tag{2}$$

which faithfully reflects model confidence: higher α_0 implies lower epistemic uncertainty. Raw logits z_c are passed through a softplus-plus-one bias, $\alpha_c = 1 + \mathrm{softplus}(z_c)$, to ensure strict positivity and numerical stability. Training uses a composite Dirichlet negative log-likelihood augmented with a regularization

term that penalizes overconfident errors, thereby calibrating uncertainty estimates. At inference, each client computes epistemic variances in a single forward pass, enabling efficient uncertainty assessment on edge devices. This evidential formulation, integrated into the detection pipeline, provides a principled mechanism for quantifying per-detection confidence under data heterogeneity and environmental variability.

UFM is computed solely from the *classification* Dirichlet evidence. For an image x, let $\mathcal{P}_\tau(x)$ be post-NMS detections above a fixed score threshold τ. We define the per-image average total evidence (set to 0 if $|\mathcal{P}_\tau(x)| = 0$) and then the group-wise mean:

$$\bar{\alpha}_{0,g} = \mathbb{E}_{x \in \mathcal{D}_g} \left[\frac{1}{\max(1, |\mathcal{P}_\tau(x)|)} \sum_{d \in \mathcal{P}_\tau(x)} \alpha_0^{(d)} \right]. \tag{3}$$

Here, $\mathcal{D}_g$ is the set of images that contain at least one ground-truth person instance of group g; the sum runs over post-NMS detections in x, and only detections matched to group-g instances contribute.

Group-Level Disparity Quantification. Using $\{\bar{\alpha}_{0,g}\}_{g=1}^{G}$ from Eq. 3, define the inter-group uncertainty gap and the normalized Uncertainty Fairness Metric (UFM) as

$$\Delta_u = \max_g \left(\frac{1}{\bar{\alpha}_{0,g}} \right) - \min_g \left(\frac{1}{\bar{\alpha}_{0,g}} \right), \qquad \mathrm{UFM} = \frac{\Delta_u}{\frac{1}{G} \sum_{g=1}^{G} \frac{1}{\bar{\alpha}_{0,g}} + \epsilon}, \tag{4}$$

with $\epsilon > 0$ for numerical stability. Higher UFM indicates greater disparity; lower UFM indicates better group fairness. See Appendix A.3 for detection-head details.

We formalize UFM as a scale-invariant measure of inter-group epistemic disparity and show (under bounded loss and standard evidential assumptions) that controlling it tightens confidence-adjusted group generalization terms (Appendix B, Theorem B.1, Corollary B.2). Because the evaluation distribution is a mixture of client distributions, global per-group confidence dispersion is a convex combination of client-level dispersions; consequently, reducing the aggregation-weighted UFM across participating clients tightens an upper bound on the global DI/EOP gaps. Under non-degenerate group coverage (each group appears on at least one active client) and standard client sampling, the weighting rule based on $\exp(-\beta\,\mathrm{UFM}_i)$ directly targets this bound: smaller β behaves like uniform averaging, while larger β emphasizes clients with tighter per-group disparities. This links the theory to practice and motivates UFM-guided aggregation for global fairness.

Aggregation Weighting Mechanism. On the server side, we aggregate client updates using a fairness-aware weighting scheme that dynamically adjusts to reported UFM values. Given each client i's update $\Delta\theta_i$ and corresponding UFM_i (see Figure 1), we assign weights via a temperature-scaled exponential:

$$\omega_i = \frac{\exp\big(-\beta\,\mathrm{UFM}_i\big)}{\sum_{j=1}^{N} \exp\big(-\beta\,\mathrm{UFM}_j\big)}, \tag{5}$$

where $\beta > 0$ controls the sharpness of fairness prioritization. As $\beta \to 0$, weights approach uniform averaging, while larger β concentrates updates on clients with minimal uncertainty disparity. The global model is then updated by:

$$\theta_G^{(t+1)} = \theta_G^{(t)} + \eta \sum_{i=1}^{N} \omega_i\, \Delta\theta_i, \tag{6}$$

ensuring that contributions from clients exhibiting both high confidence and equitable performance are amplified. This continuous reweighting adapts to temporal shifts and data heterogeneity, promoting robust convergence with reduced fairness gaps and preserved accuracy.

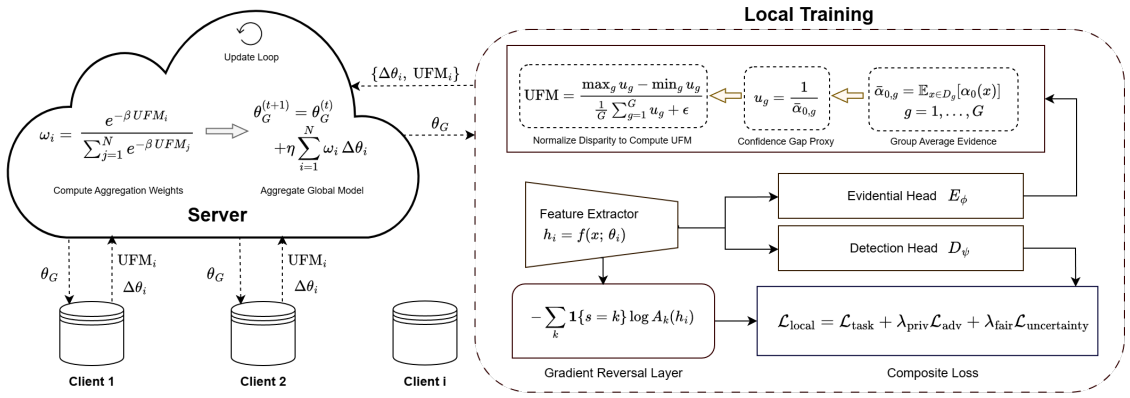


Figure 1: Overview of the RESFL framework. *Right:* Client i computes feature representation, applies a gradient reversal layer for adversarial privacy loss $\mathcal{L}_{\text{adv}}$, and an evidential head for uncertainty-fairness metric UFM_i, then forms the composite loss $\mathcal{L}_{\text{local}}$ to produce update $\Delta\theta_i$. *Left:* Server receives $\{\Delta\theta_i, \mathrm{UFM}_i\}$, computes aggregation weights $\omega_i \propto \exp(-\beta\, \mathrm{UFM}_i)$, and updates the global model θ_G.

3.2. Adversarial Privacy Disentanglement via Gradient Reversal

To mitigate sensitive attribute leakage during federated training, we augment the feature extractor $f(x;\theta) : \mathcal{X} \to \mathbb{R}^d$, which maps input data to a latent representation h, with an adversarial classifier $A(h;\phi) : \mathbb{R}^d \to [0,1]^K$. The adversary is trained to predict the sensitive attribute label $s \in \{1, \ldots, K\}$ from h, while the feature extractor is jointly optimized to make this prediction as difficult as possible, thus encouraging the learned representation to be invariant to s. During training, the classifier parameters ϕ seek to minimize the cross-entropy over the joint distribution of inputs and labels, while the feature extractor parameters θ are trained to maximize this same objective, thereby removing attribute-relevant signals. Concretely, we embed a Gradient Reversal Layer (GRL) $\mathcal{R}_{\lambda_{\text{adv}}}$ between f and A, which acts as the identity in the forward pass but multiplies incoming gradients by $-\lambda_{\text{adv}}$ in the backward pass. The resulting adversarial minimax objective is expressed as:

$$\min_{\theta} \max_{\phi}\ \mathbb{E}_{(x,s)\sim\mathcal{D}_i}\Big[-\lambda_{\text{adv}} \sum_{k=1}^{K} \mathbf{1}\{s=k\}\ \log\ A_k\big(\mathcal{R}_{\lambda_{\text{adv}}}(f(x;\theta));\ \phi\big)\Big], \tag{7}$$

where $\mathcal{D}_i$ denotes the local dataset of client i and $\mathbf{1}\{s=k\}$ is the indicator for class k and $A_k(\cdot)$ denotes the k-th output probability. While we do not claim (ε, δ)–DP guarantees, our objective has an information-theoretic interpretation: letting $H = f(X;\theta)$ denote the learned representation and S the sensitive attribute, maximizing $\mathcal{L}_{\text{adv}}$ reduces the mutual information $I(H;S)$; by Fano's inequality, as $I(H;S) \to 0$ any attribute-inference attack $\hat{S} = g(H)$ is driven to chance level, i.e., accuracy $\approx 1 - \frac{1}{K}$ (Appendix C).

Once the adversarial classifier is optimally trained for a fixed feature extractor, we obtain the induced privacy loss for θ by substituting the worst-case classifier parameters $\phi^*(\theta) = \arg\max_\phi \mathcal{L}_{\text{adv}}(\theta, \phi)$. The privacy-preserving gradient step for the feature extractor is then driven by:

$$\mathcal{L}_{\text{priv}}(\theta)\ =\ \lambda_{\text{adv}}\ \mathbb{E}_{(x,s)\sim\mathcal{D}_i}\Big[\sum_{k=1}^{K} \mathbf{1}\{s=k\}\ \log\ A_k\big(f(x;\theta);\ \phi^*(\theta)\big)\Big], \tag{8}$$

which, when differentiated through the GRL, enforces that feature representations h become invariant to s. In practice, we interleave updates of ϕ (maximization) and θ (minimization) within each local SGD step,

Table 1: Comparison of FL algorithms on the FACET dataset in detection performance (mAP), fairness ($|1\text{–DI}|$, ΔEOP), privacy (MIA, AIA success rates), and robustness (BA AD, DPA EODD)

Algorithm	Utility	Fairness		Privacy Attacks		Robustness Attacks	
	Overall mAP	$\lvert 1 - \text{DI}\rvert$	ΔEOP	MIA SR	AIA SR	BA AD	DPA EODD
FedAvg	0.6378	**0.2159**	**0.2362**	0.3341	0.4431	0.3125	0.0792
FedAvg-DP ($\epsilon = 0.1$)	0.2932	0.4521	0.3576	**0.1765**	**0.2154**	0.2833	0.1724
FedAvg-DP ($\epsilon = 0.5$)	0.4741	0.3869	0.2793	0.2286	0.2539	0.3019	0.1328
FairFed	**0.7013**	0.2496	0.2562	0.4409	0.5256	0.4139	**0.0566**
PrivFairFl-Pre	0.6154	0.2504	0.2659	0.3875	0.4038	0.3238	0.0953
PrivFairFl-Post	0.6119	0.2718	0.2505	0.2872	0.3159	0.3212	0.0937
PUFFLE	0.4192	0.3721	0.2976	0.2725	0.2909	**0.1439**	0.1360
PFU-FL	0.3952	0.3356	0.3446	0.2409	0.2546	0.2612	0.1459
Ours (`RESFL`)	**0.6654**	**0.2287**	**0.1959**	**0.2093**	**0.1832**	**0.1692**	**0.0674**

ensuring that the learned representation provably suppresses sensitive attribute information while retaining utility for the primary detection task.

3.3. Joint Optimization of Privacy and Fairness

In each client's training loop, `RESFL` minimizes a composite loss that balances detection accuracy, attribute obfuscation, and uncertainty-based bias control. Formally, each client solves:

$$\mathcal{L}_{\text{local}}(\theta, \phi) = \mathcal{L}_{\text{task}}(\theta) + \lambda_{\text{priv}}\, \mathcal{L}_{\text{adv}}(\theta, \phi) + \lambda_{\text{fair}}\, \mathcal{L}_{\text{uncertainty}}(\theta), \tag{9}$$

where λ_{priv} scales the gradient reversal adversarial loss to limit information leakage, and λ_{fair} weights the evidential uncertainty term to reduce group disparity. By selecting $(\lambda_{\text{priv}}, \lambda_{\text{fair}})$ along the convex envelope of evaluated tradeoff points, practitioners obtain models that meet target privacy and fairness thresholds without unnecessary sacrifice of either objective.

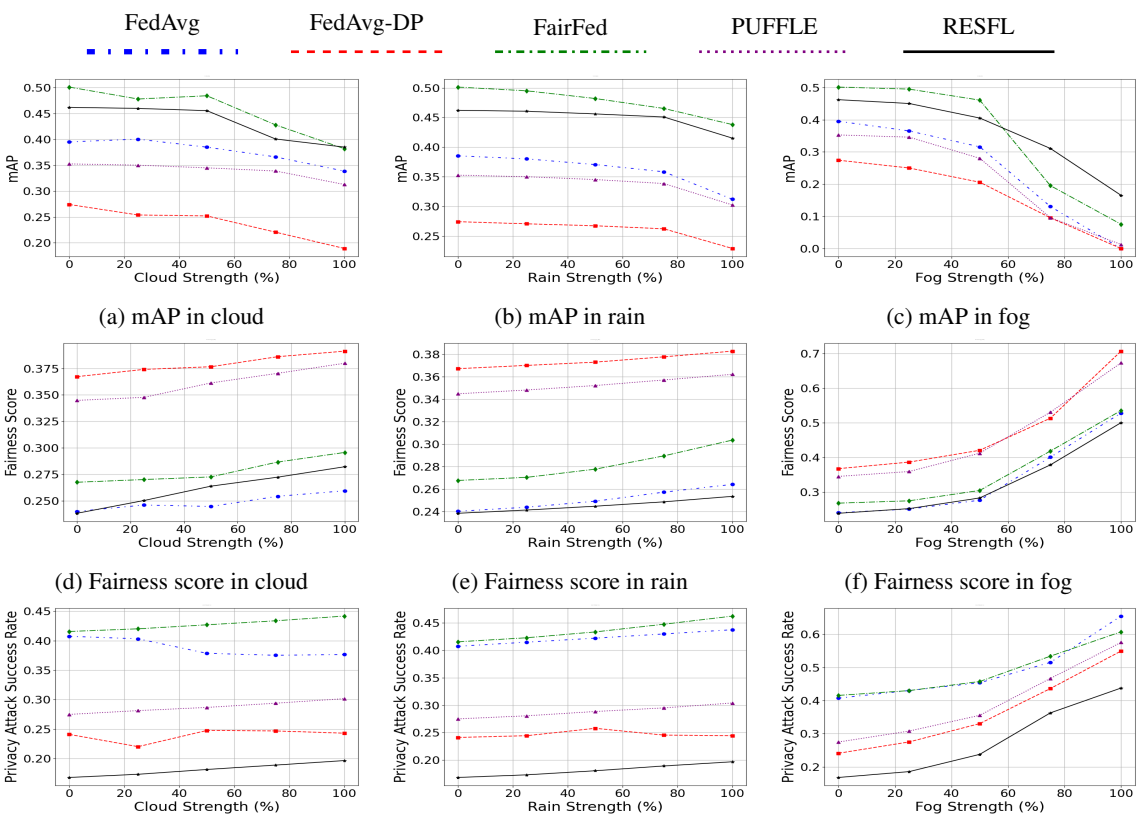
After local updates, each client computes its UFM_i and sends both the parameter update $\Delta\theta_i$ and UFM_i to the server. The server then aggregates via:

$$\theta_G^{(t+1)} = \theta_G^{(t)} + \eta \sum_{i=1}^{N} \frac{\exp\big(-\beta\, \text{UFM}_i\big)}{\sum_{j=1}^{N} \exp\big(-\beta\, \text{UFM}_j\big)}\, \Delta\theta_i, \tag{10}$$

where β controls the fairness weight. This alternating sequence of local composite-loss minimization and fairness-aware aggregation (Algorithm 1 in Appendix D) drives the global model to converge with robust detection, provable attribute privacy, and equitable treatment across sensitive groups.

4. Experimental Results & Analyses

We evaluate `RESFL` in an autonomous vehicle (AV) context using the FACET dataset and CARLA simulator to capture demographic variation and environmental perturbations. Given the safety-critical and privacy-sensitive nature of AV perception, we pose the following key questions: (1) *To what extent does `RESFL` balance utility, privacy, and fairness under standard AV operating conditions?* (2) *How resilient is `RESFL` to increased uncertainty caused by environmental variations such as changing weather and lighting?*



(a) mAP in cloud (b) mAP in rain (c) mAP in fog

(d) Fairness score in cloud (e) Fairness score in rain (f) Fairness score in fog

(g) Privacy attacks in cloud (h) Privacy attacks in rain (i) Privacy attacks in fog

Figure 2: Performance comparison of four state-of-the-art FL methods and RESFL across three weather conditions (cloud, rain, fog) at 0%–100% intensity. Rows represent performance metrics (accuracy, fairness, privacy attack), and columns correspond to weather conditions.

4.1. Experimental Setup

Datasets. The FACET benchmark (Gustafson et al., 2023) comprises 32,000 real-world images with over 50,000 person instances annotated for perceived skin tone on the ten-level Monk Skin Tone (MST) scale (MST = 1 lightest to MST = 10 darkest, see Figure 6 in Appendix E); we average multiple annotations per instance, discretize back to the ten MST levels, partition into ten cohorts, and split into four IID client shards to simulate cross-device heterogeneity and demographic variation without sharing raw data. Using the CARLA simulator (Dosovitskiy et al., 2017), we collect 6,000 clear-weather frames (600 per MST level) for fine-tuning and 7,800 evaluation frames across three urban layouts (Town01, Town03, and Town05) under clear, foggy, and rainy conditions at five intensities (0%, 25%, 50%, 75%, 100%). Each walker blueprint available in CARLA is manually assigned to a corresponding MST label through visual inspection based on appearance and attributes. Pedestrian bounding boxes are extracted via connected-component analysis on semantic segmentation masks, which serves as our ground truth (see Appendix E for details).

Comparing Schemes. We benchmark RESFL against standard and state-of-the-art federated learning methods. **FedAvg** serves as the canonical baseline, performing weighted averaging of client updates. We evaluate

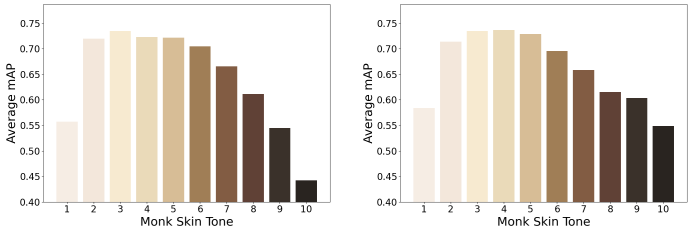
FedAvg-DP (Objective), which injects calibrated noise into the local objective under two privacy budgets ($\epsilon = 0.1$ and $\epsilon = 0.5$). **FairFed** (Ezzeldin et al., 2023) dynamically adjusts aggregation weights to penalize cross-client performance gaps. **PrivFairFL** (Pentyala et al., 2022) incorporates fairness constraints either before aggregation (**PrivFairFL-Pre**) or after local updates (**PrivFairFL-Post**). **PUFFLE** (Corbucci et al., 2024) unifies noise injection with fairness regularization in a joint optimization framework. Finally, **PFU-FL** (Sun et al., 2023) employs adaptive weighting to balance privacy, fairness, and utility objectives.

Metrics. We measure detection accuracy using mean Average Precision (mAP), which averages per-class AP to capture both object localization and classification quality. For fairness, we report the absolute disparate impact deviation $|1 - \mathrm{DI}|$, quantifying the ratio of favorable outcome rates between the most- and least-advantaged groups, and the equality of opportunity gap $\Delta\mathrm{EOP}$, the absolute difference in true positive rates across cohorts. Privacy is assessed by Membership Inference Attack Success Rate (MIA SR) and Attribute Inference Attack Success Rate (AIA SR), where lower values indicate stronger confidentiality. Robustness is measured by Byzantine Accuracy Degradation (BA AD), the relative per-condition mAP drop between clean and attacked runs, and Data Poisoning Attack Equalized Odds Difference Deviation (DPA EODD), the rise in fairness disparity, under a fixed protocol with a constant Byzantine-client fraction each round (sign-flip, ℓ_2-bounded) and a constant poisoning fraction over a specified block of local epochs.

Experimental Configuration. We implement `RESFL` using a modified YOLOv8 backbone with an evidential concentration-vector head. Each client trains for 100 epochs (batch size 64) using SGD (momentum 0.9, weight decay $1e^{-4}$) with an initial learning rate of 0.001, decayed by 0.1 at epochs 50 and 75. The FACET dataset (32k images) is split into four equal i.i.d. subsets. CARLA fine-tuning uses 6k neutral-weather frames and evaluates on 7.8k frames across 13 weather conditions. We set $\lambda_{\text{priv}} = 0.1$, $\lambda_{\text{fair}} = 0.01$, and aggregation temperature $\beta = 2.0$, and run 100 federated rounds with three random seeds. All hyperparameters were selected via an extensive grid search (more details in Appendix F).

4.2. Trade-off Analysis on the FACET Dataset

In this experiment, we evaluate the performance of various FL algorithms on the FACET dataset. Our objective is to compare the overall trade-offs of each method in a controlled setting. Table 1 reports results for baselines in Section 4.1, and our proposed `RESFL`, while Figure 3 illustrates per-skin-tone mAP distributions. `RESFL` attains 0.6654 mAP, close to FairFed (0.7013), and exceeds PUFFLE and PFU-FL. It maintains consistent accuracy across all ten MST cohorts via uncertainty-guided weighting. It yields $|1 - \mathrm{DI}| = 0.2287$ and $\Delta\mathrm{EOP} = 0.1959$, improves privacy (MIA 0.2093, AIA 0.1832) over FedAvg and its DP variants, and remains robust under attacks (BA AD 0.1692; DPA EODD 0.0674). See Appendix H for IID vs. non-IID results with 4 and 8 clients, where `RESFL` maintains strong accuracy with the best fairness–privacy profile. We also note its domain-agnostic performance in Appendix I.



(a) FedAvg (b) `RESFL`

Figure 3: Accuracy (mAP) across Monk Skin Tones on FACET: RESFL stays consistent; FedAvg drops on darker tones.

4.3. Resilience Analysis under Adverse Conditions in CARLA

We fine-tune and compare FACET baselines—FedAvg-DP (ϵ=0.1), FairFed, PUFFLE, and `RESFL`—on 2,600 CARLA frames under clear, cloud, rain, and fog at five intensities (0–100%). Figure 2 reports utility

(mAP), fairness (mean of $|1-\mathrm{DI}|$ and $\Delta\mathrm{EOP}$; lower is better), and privacy risk (mean MIA/AIA success rate; lower is better). Under clear weather (0%), RESFL achieves 0.46 mAP, 0.24 fairness, and 0.17 privacy. At 100% cloud, mAP drops to 0.39 (vs. 0.19 FedAvg-DP, 0.34 FedAvg), fairness rises to 0.28 (vs. 0.39 FedAvg-DP), and privacy to 0.24 (vs. 0.38 FedAvg). In heavy rain (100%), it retains 0.42 mAP, 0.25 fairness, and 0.20 privacy; under dense fog (100%), it maintains 0.17 mAP, 0.50 fairness, and 0.44 privacy, while others collapse below 0.10 mAP with fairness/privacy > 0.50. These results show that adversarial privacy disentanglement with uncertainty-guided aggregation enables RESFL to degrade gracefully in utility while preserving fairness and confidentiality under severe perturbations (see Appendix H).

4.4. Ablation Study with RESFL

We conduct an ablation study to examine the impact of two key hyperparameters in RESFL: the uncertainty-based fairness coefficient (λ_{fair}) and the adversarial privacy coefficient (λ_{priv}), while keeping the task loss coefficient fixed at one. In the first set of FACET experiments, we disable adversarial privacy ($\lambda_{\text{priv}} = 0$) and sweep λ_{fair}; as λ_{fair} increases, fairness disparities ($|1 - \mathrm{DI}|$, $\Delta\mathrm{EOP}$) decrease, but mAP also declines, reflecting the classical fairness–utility trade-off without privacy regularization. In the second set, we fix $\lambda_{\text{fair}} = 0.1$ and vary λ_{priv} to balance utility, fairness, and privacy. The optimal occurs at $\lambda_{\text{fair}} = 0.1, \lambda_{\text{priv}} = 0.01$, yielding mAP 0.6654, $|1-\mathrm{DI}|$ 0.2287, $\Delta\mathrm{EOP}$ 0.1959, and low MIA (0.2093) and AIA (0.1832) success. Increasing λ_{priv} beyond 0.01 degrades both detection accuracy and group equity, showing that overly strong adversarial signals disrupt the balance.

Table 2: RESFL's performance with varying uncertainty-based fairness and adversarial privacy coefficients (i.e., λ_{fair} and λ_{priv}).

Algorithm		**Utility**	**Fairness**		**Privacy**	
λ_{fair}	λ_{priv}	**mAP**	$\lvert 1-\mathrm{DI}\rvert$	$\Delta\mathrm{EOP}$	**MIA SR**	**AIA SR**
1	0	0.6278	**0.2258**	0.2062	0.3341	**0.1431**
0	1	0.5856	0.2571	0.2846	**0.1025**	0.1463
0.01	1	0.6056	0.2653	0.3459	0.1256	0.1668
0.1	1	0.6254	0.2538	0.2626	0.1477	0.1608
1	1	0.5953	0.2432	0.2513	0.2197	0.1782
0.1	0.01	**0.6654**	0.2287	**0.1959**	0.2093	0.1832
0.1	0.1	0.6430	0.2625	0.3143	0.1363	0.1474
0.1	1	0.5839	0.3862	0.4146	0.1176	0.1656

5. Conclusions & Future Work

This work introduced RESFL, a domain-agnostic federated learning framework that jointly improves utility, group fairness, and parameter privacy while remaining robust to adversarial perturbations and environmental variability. RESFL combines adversarial privacy disentanglement (via gradient reversal) with an evidential head that yields calibrated epistemic uncertainty and a scale-invariant Uncertainty Fairness Metric (UFM) for aggregation. This design suppresses sensitive-attribute signals in shared representations and adaptively upweights clients exhibiting lower inter-group uncertainty disparity, producing updates that are both confident and equitable. On FACET, RESFL matches or surpasses baselines in mAP while significantly reducing disparate impact and equal-opportunity gaps across Monk Skin Tones; on CARLA, it maintains accuracy under weather shifts and reduces membership/attribute inference success, indicating improved privacy.

Future work will focus on directions that reduce reliance on annotated sensitive labels by refining UFM with vacuity–dissonance decomposition and attribute-free proxies, as well as explore automated schedules for privacy/fairness temperatures to eliminate manual tuning. Finally, we also plan to extend to streaming and multimodal FL and evaluate deployments with secure aggregation or calibrated DP noise to strengthen end-to-end guarantees.

REPRODUCIBILITY STATEMENT

We release anonymized code and configuration files for data preprocessing, client partitioning (IID/Non-IID), training/evaluation for all baselines, and attack implementations. We fix random seeds, report hardware, and include full hyperparameter grids. Dataset construction, splits, and preprocessing pipelines are detailed in Appendix E; implementation specifics (models, losses, schedules, and runtime environment) are in Appendix F. Detection fairness metrics are defined for object detection in §A.5 (IoU threshold, matching, aggregation), and UFM is computed from Dirichlet classification evidence as specified in §3.1 and Appendix A.3. Exact evaluation settings (NMS, score thresholds, IoU thresholds) are fixed and documented.

ETHICS STATEMENT

We study privacy-, fairness-, and utility-aware federated detection using public datasets and simulation; no new human-subjects data are collected. Because we analyze perceived skin tone, we acknowledge risks of stereotyping or surveillance and restrict sensitive attributes to local clients, reporting only group-level statistics; we discourage any use that targets individuals or communities. Given the safety-critical AV context, prospective deployments should include stakeholder consultation, external auditing on the target population, and monitoring for distribution shift and unintended harms. We document datasets and implementation details to support independent verification (Appendix E, Appendix F).

REFERENCES

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in neural information processing systems*, 33:14927–14937, 2020.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32, 2019.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*, 2020.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 54–63, Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2007. URL https://www.aclweb.org/anthology/S19-2007.

Chunlu Chen, Ji Liu, Haowen Tan, Xingjian Li, Kevin I-Kai Wang, Peng Li, Kouichi Sakurai, and Dejing Dou. Trustworthy federated learning: Privacy, security, and beyond. *Knowledge and Information Systems*, 67(3):2321–2356, 2025.

E Chen, Yang Cao, and Yifei Ge. A generalized shuffle framework for privacy amplification: Strengthening privacy guarantees and enhancing utility. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11267–11275, 2024.

Huiqiang Chen, Tianqing Zhu, Tao Zhang, Wanlei Zhou, and Philip S Yu. Privacy and fairness in federated learning: on the perspective of tradeoff. *ACM Computing Surveys*, 56(2):1–37, 2023.

Luca Corbucci, Mikko A Heikkila, David Solans Noguero, Anna Monreale, and Nicolas Kourtellis. Puffle: Balancing privacy, utility, and fairness in federated learning. *arXiv preprint arXiv:2407.15224*, 2024.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Conference on robot learning*, pp. 1–16. PMLR, 2017.

Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pp. 1–12. Springer, 2006.

Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2468–2479. SIAM, 2019.

Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A Salman Avestimehr. Fairfed: Enabling group fairness in federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 7494–7502, 2023.

Laura Gustafson, Chloe Rolland, Nikhila Ravi, Quentin Duval, Aaron Adcock, Cheng-Yang Fu, Melissa Hall, and Candace Ross. Facet: Fairness in computer vision evaluation benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20370–20382, 2023.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Ahmed Imteaj, Urmish Thakker, Shiqiang Wang, Jian Li, and M Hadi Amini. A survey on federated learning for resource-constrained iot devices. *IEEE Internet of Things Journal*, 9(1):1–24, 2021.

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

Caelin Kaplan. *Inherent trade-offs in privacy-preserving machine learning*. PhD thesis, Université Côte d'Azur, 2024.

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

Dohyoung Kim, Hyekyung Woo, and Youngho Lee. Addressing bias and fairness using fair federated learning: A synthetic review. *Electronics*, 13(23):4664, 2024a.

Kibaek Kim, Krishnan Raghavan, Olivera Kotevska, Matthieu Dorier, Ravi Madduri, Minseok Ryu, Todd Munson, Rob Ross, Thomas Flynn, Ai Kagawa, et al. Privacy-preserving federated learning for science: Challenges and research directions. In *2024 IEEE International Conference on Big Data (BigData)*, pp. 7849–7853. IEEE, 2024b.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*, 2019.

Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In *International conference on machine learning*, pp. 6755–6764. PMLR, 2020.

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*, pp. 1273–1282. PMLR, 2017.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

Bimsara Pathiraja, Caleb Liu, and Ransalu Senanayake. Fairness in autonomous driving: Towards understanding confounding factors in object detection under challenging weather. *arXiv preprint arXiv:2406.00219*, 2024.

Sikha Pentyala, Nicola Neophytou, Anderson Nascimento, Martine De Cock, and Golnoosh Farnadi. Privfairfl: Privacy-preserving group fairness in federated learning. *arXiv preprint arXiv:2205.11584*, 2022.

David Pujol, Ryan McKenna, Satya Kuppam, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. Fair decision making using privacy-protected data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 189–199, 2020.

Borja Rodríguez-Gálvez, Filip Granqvist, Rogier van Dalen, and Matt Seigel. Enforcing fairness in private federated learning via the modified method of differential multipliers. *arXiv preprint arXiv:2109.08604*, 2021.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Kangkang Sun, Xiaojin Zhang, Xi Lin, Gaolei Li, Jing Wang, and Jianhua Li. Toward the tradeoffs between privacy, fairness and utility in federated learning. In *International Symposium on Emerging Information Security and Applications*, pp. 118–132. Springer, 2023.

Peng Sun, Haoxuan Che, Zhibo Wang, Yuwei Wang, Tao Wang, Liantao Wu, and Huajie Shao. Pain-fl: Personalized privacy-preserving incentive for federated learning. *IEEE Journal on Selected Areas in Communications*, 39(12):3805–3820, 2021.

Anh Tu Tran, The Dung Luong, and Xuan Sang Pham. A novel privacy-preserving federated learning model based on secure multi-party computation. In *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making*, pp. 321–333. Springer, 2023.

Dawood Wasif, Dian Chen, Sindhuja Madabushi, Nithin Alluru, Terrence J Moore, and Jin-Hee Cho. Empirical analysis of privacy-fairness-accuracy trade-offs in federated learning: A step towards responsible ai. *arXiv preprint arXiv:2503.16233*, 2025.

Yongkai Wu, Lu Zhang, and Xintao Wu. Fairness-aware classification: Criterion, convexity, and bounds. *arXiv preprint arXiv:1809.04737*, 2018.

Bangzhou Xin, Wei Yang, Yangyang Geng, Sheng Chen, Shaowei Wang, and Liusheng Huang. Private fl-gan: Differential privacy synthetic data generation based on federated learning. In *Icassp 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 2927–2931. IEEE, 2020.

Runhua Xu, Nathalie Baracaldo, and James Joshi. Privacy-preserving machine learning: Methods, challenges and directions. *arXiv preprint arXiv:2108.04417*, 2021.

Mohammad Yaghini, Patty Liu, Franziska Boenisch, and Nicolas Papernot. Learning with impartiality to walk on the pareto frontier of fairness, privacy, and utility. *arXiv preprint arXiv:2302.09183*, 2023.

Lei Yang, Jiaming Huang, Wanyu Lin, and Jiannong Cao. Personalized federated learning on non-iid data via group-based meta-learning. *ACM Transactions on Knowledge Discovery from Data*, 17(4):1–20, 2023.

Xun Yi, Russell Paulet, Elisa Bertino, Xun Yi, Russell Paulet, and Elisa Bertino. *Homomorphic encryption*. Springer, 2014.

Han Yu, Zelei Liu, Yang Liu, Tianjian Chen, Mingshu Cong, Xi Weng, Dusit Niyato, and Qiang Yang. A fairness-aware incentive scheme for federated learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 393–399, 2020.

Tao Zhang, Tianqing Zhu, Kun Gao, Wanlei Zhou, and S Yu Philip. Balancing learning model privacy, fairness, and accuracy with early stopping criteria. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5557–5569, 2021.

Yifei Zhang, Dun Zeng, Jinglong Luo, Xinyu Fu, Guanzhong Chen, Zenglin Xu, and Irwin King. A survey of trustworthy federated learning: Issues, solutions, and challenges. *ACM Transactions on Intelligent Systems and Technology*, 15(6):1–47, 2024.

Appendix to "RESFL: An Uncertainty-Aware Framework for Responsible Federated Learning by Balancing Privacy, Fairness and Utility"

A. PRELIMINARIES

This section presents the mathematical foundations and system specifications of our work. We detail the YOLOv8-based object detection model, describe the FL setup in the AV scenario, formalize threat models (privacy, robustness, and fairness attacks), and define evaluation metrics. We also provide a unified overview of the datasets used for training and testing.

A.1. SYSTEM MODEL: OBJECT DETECTION

Let $I \in \mathbb{R}^{H\times W\times C}$ denote an input image. Our object detection model, derived from YOLOv8, produces a set of detections:

$$\mathcal{P} = \{(b_i, c_i, s_i)\}_{i=1}^{N},$$

where each $b_i \in \mathbb{R}^4$ specifies the bounding box coordinates, $c_i \in \{1, \ldots, C\}$ is the predicted class label, and $s_i \in [0, 1]$ is the corresponding confidence score. The overall detection loss is given by:

$$\mathcal{L}_{\text{det}} = \lambda_{\text{cls}}\,\mathcal{L}_{\text{cls}} + \lambda_{\text{loc}}\,\mathcal{L}_{\text{loc}} + \lambda_{\text{conf}}\,\mathcal{L}_{\text{conf}}, \tag{11}$$

where $\mathcal{L}_{\text{cls}}$, $\mathcal{L}_{\text{loc}}$, and $\mathcal{L}_{\text{conf}}$ represent the classification, localization, and confidence losses, respectively, and $\lambda_{\text{cls}}, \lambda_{\text{loc}}, \lambda_{\text{conf}} \in \mathbb{R}^+$ are hyperparameters.

A.2. FEDERATED LEARNING SETUP AND NETWORK MODEL

Consider a set of N clients $\{C_i\}_{i=1}^{N}$, each possessing a local dataset $\mathcal{D}_i \subset \mathbb{R}^{H\times W\times C}$ and a local model with parameters θ_i. A central server maintains the global model θ_G. The FL process begins with the server initializing and distributing $\theta_G^{(0)}$ to all clients. Each client then updates its model by performing local stochastic gradient descent (SGD):

$$\theta_i^{(t+1)} \;=\; \theta_i^{(t)} \;-\; \eta\,\nabla\mathcal{L}_i\Big(\theta_i^{(t)}\Big)\,,$$

where $\eta > 0$ is the learning rate, t is the local iteration index, and $\mathcal{L}_i$ is the local loss (e.g., $\mathcal{L}_{\text{det}}$). The server aggregates the locally updated parameters via FedAvg:

$$\theta_G^{(t+1)} \;=\; \sum_{i=1}^{N} \frac{|\mathcal{D}_i|}{\sum_{j=1}^{N}|\mathcal{D}_j|}\,\theta_i^{(t+1)}.$$

This FL framework maintains privacy because raw data remain on devices; the server receives only model updates.

A.3. UNCERTAINTY QUANTIFICATION VIA EVIDENTIAL REGRESSION

Evidential head for detection (Dirichlet & NIG). Our detector uses a decoupled evidential head on top of YOLOv8. For every anchor/location, the *classification* branch outputs a nonnegative concentration vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_C)$ via $\alpha_c = 1 + \mathrm{softplus}(z_c)$, which parameterizes a Dirichlet over class probabilities. The *localization* branch outputs Normal–Inverse–Gamma (NIG) parameters for each box coordinate $q \in \{x, y, w, h\}$, $(\gamma_q, \nu_q, \alpha_q^{\text{nig}}, \beta_q)$, following (Amini et al., 2020). Concretely:

$$p(\mathbf{p} \mid \boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\alpha}), \qquad (\mu_q, \sigma_q^2) \sim \mathrm{NIG}\big(\gamma_q, \nu_q, \alpha_q^{\text{nig}}, \beta_q\big)\,.$$

For the Dirichlet, total evidence $\alpha_0 = \sum_{c=1}^{C} \alpha_c$ controls epistemic uncertainty; larger α_0 implies lower epistemic variance (Sensoy et al., 2018). For the NIG, the epistemic variance of the mean is $\mathrm{Var}[\mu_q] = \beta_q/(\nu_q(\alpha_q^{\mathrm{nig}} - 1))$.

Per-detection scalar uncertainties. We extract two scalar measures:

$$u_{\mathrm{cls}} = \frac{1}{\alpha_0 + 1} \quad \text{(classification epistemic; lower is more confident)},$$

$$u_{\mathrm{box}} = \frac{1}{4} \sum_{q \in \{x,y,w,h\}} \frac{\mathrm{Var}[\mu_q]}{s_q^2} \quad \text{(localization epistemic; normalized, lower is more confident)}.$$

Here $s_x = W,\ s_y = H,\ s_w = W,\ s_h = H$ are image-scale normalizers (width W and height H) so that u_{box} is dimensionless and comparable across resolutions. In practice we compute $\mathrm{Var}[\mu_q] = \beta_q/(\nu_q(\alpha_q^{\mathrm{nig}} - 1 + \epsilon))$ with a small ϵ for stability.

What feeds UFM. Unless otherwise specified, **UFM is computed *only* from the classification Dirichlet evidence**. For an image x with detections $\mathcal{P}(x)$, we collect u_{cls} for detections that pass NMS and a score threshold τ (we use $\tau = 0.25$; results are insensitive in $[0.2, 0.4]$). For a sensitive group g, we average the induced α_0 over that group:

$$\bar{\alpha}_{0,g} = \mathbb{E}_{x \in \mathcal{D}_g}\Big[\frac{1}{|\mathcal{P}_\tau(x)|} \sum_{d \in \mathcal{P}_\tau(x)} \alpha_0^{(d)} \Big],$$

and plug $\{\bar{\alpha}_{0,g}\}_{g=1}^{G}$ into Eq. (2) in the main paper to obtain $\mathrm{UFM} = \mathrm{UFM}_{\mathrm{cls}}$. This choice aligns UFM with fairness over *recognition* (who is detected/classified confidently), while the NIG head is used for robustness analyses and ablations.

Training losses. The classification branch is trained with the Dirichlet NLL plus an evidential regularizer that penalizes overconfident errors (Sensoy et al., 2018); the localization branch uses the NIG NLL with the regularizer from (Amini et al., 2020). This yields calibrated epistemic estimates for both branches while keeping the UFM definition unambiguous.

A.4. Threat Model

We study a cross-silo FL setting with an honest-but-curious server that observes client updates and may collude with a subset of clients. Training data remain on-device and are never shared; thus raw *data privacy* is enforced by the FL protocol. Our privacy goal is *parameter privacy*: reduce sensitive-attribute leakage from intermediate representations or model updates. We also evaluate *robustness* to malicious updates and *fairness* against bias amplification. The attacks below instantiate these goals.

A.4.1. Privacy Attacks

The selected privacy attacks assess whether an adversary can extract sensitive information from federated model updates.

Membership Inference Attack (MIA): MIA tests whether a sample $x \in \mathbb{R}^d$ was used in training. An adversarial client C_a trains a shadow model to mimic the global model M_t, queries M_t on member/non-member samples, and uses the resulting outputs to train a binary classifier $\mathcal{A}_{\mathrm{MIA}}$ that predicts membership:

$$\mathcal{A}_{\mathrm{MIA}}(x) = \begin{cases} 1, & x \in \mathcal{D}_{\text{train}} \\ 0, & x \notin \mathcal{D}_{\text{train}} \end{cases}.$$

We report the *MIA Success Rate* (binary accuracy):

$$S_{\text{MIA}} \;=\; \frac{TP+TN}{TP+TN+FP+FN}\,, \tag{12}$$

where TP, TN, FP, FN are counts over a balanced member/non-member evaluation set. Higher S_{MIA} indicates greater privacy leakage.

Attribute Inference Attack (AIA): AIA tests whether a sensitive attribute $s \in \mathcal{S}$ can be inferred from update-derived features. The adversary extracts features I from observed gradients/updates and trains $\mathcal{A}_{\text{AIA}}$ to predict s:

$$\hat{s} \;=\; \mathcal{A}_{\text{AIA}}(I)\,. \tag{13}$$

We report the *AIA Success Rate* as top-1 accuracy over M instances:

$$S_{\text{AIA}} \;=\; \frac{1}{M}\sum_{i=1}^{M} \mathbf{1}\{\hat{s}_i = s_i\}\,. \tag{14}$$

(For binary attributes, S_{AIA} reduces to standard binary accuracy.)

A.4.2. **Robustness Attack**

We assess the FL system's resilience to malicious modifications of model updates using a robustness attack.

Byzantine Attack: In a Byzantine attack, a subset of clients manipulates their model updates before sending them to the central aggregator. Let θ_k be the legitimate update from client k, and let the adversary introduce a perturbation δ_k, yielding a modified update:

$$\tilde{\theta}_k = \theta_k + \delta_k, \quad \text{with} \quad \|\delta_k\| \gg 0. \tag{15}$$

A sufficiently large δ_k disrupts training, leading to model divergence or severe performance degradation. The attack's impact is quantified by comparing global model accuracy without malicious interference (A_{clean}) to accuracy under Byzantine updates ($A_{\text{Byzantine}}$), measured as:

$$D_{\text{Byz}} = A_{\text{clean}} - A_{\text{Byzantine}}. \tag{16}$$

A larger D_{Byz} indicates a stronger attack and greater vulnerability of the federated learning system to such perturbations.

Data Poisoning Attack: This attack injects manipulated samples $\Delta\mathcal{D}$ into a client's local dataset, altering its distribution. The poisoned dataset is defined as:

$$\mathcal{D}'_k = \mathcal{D}_k \cup \Delta\mathcal{D}. \tag{17}$$

The adversary selects injected samples to skew feature distributions, favoring one demographic group over another and shifting the global model's decision boundaries. The impact on fairness is measured using the Equalized Odds Difference (EOD), which quantifies disparities in true positive rates (TPR) and false positive rates (FPR) between protected and unprotected groups:

$$EOD = \left|\text{TPR}_{\text{protected}} - \text{TPR}_{\text{unprotected}}\right| \quad + \left|\text{FPR}_{\text{protected}} - \text{FPR}_{\text{unprotected}}\right|. \tag{18}$$

To assess the attack's effect, we compute the Equalized Odds Difference Deviation (EODD) as the change in EOD between the poisoned and clean datasets:

$$EODD = EOD_{\text{poisoned}} - EOD_{\text{clean}}. \tag{19}$$

A larger $EODD$ indicates greater fairness violation, confirming the attack's success in introducing bias. This highlights the dual threat of data poisoning, which compromises both model accuracy and equity across user groups.

A.5. Detection-based Fairness & Privacy Metrics

Setup and matching protocol. For each image x, let $\mathcal{G}(x) = \{(b_k^*, y_k, g_k)\}_{k=1}^{N_x}$ be ground-truth person instances with box b_k^*, class $y_k = \texttt{person}$, and sensitive group $g_k \in \{1, \dots, G\}$ (e.g., MST). Let $\mathcal{P}(x) = \{(\hat{b}_i, \hat{c}_i, \hat{s}_i)\}_{i=1}^{M_x}$ be predicted boxes, classes, and confidences after standard NMS (we use IoU NMS=0.5 and score threshold $\tau_{\text{score}} = 0.25$; same across all methods). We evaluate fairness at a fixed IoU threshold $\tau_{\text{fair}} = 0.5$ (distinct from mAP's COCO sweep).

We perform a one-to-one greedy match between $\mathcal{G}(x)$ and $\mathcal{P}(x)$ by descending $\hat{s}_i$: a prediction $(\hat{b}_i, \hat{c}_i, \hat{s}_i)$ matches a ground-truth (b_k^*, y_k, g_k) iff $\hat{c}_i = y_k$ and $\text{IoU}(\hat{b}_i, b_k^*) \geq \tau_{\text{fair}}$ and neither has been matched yet. Matched pairs count as true positives (TP); unmatched predictions are false positives (FP); unmatched ground truths are false negatives (FN). Multiple predictions for the same ground-truth are penalized as FP except the highest-scoring matched one.

Per-group rates (micro-averaged). Let $\text{TP}_g = \sum_x \sum_{k:g_k=g} \mathbf{1}\{\text{GT } k \text{ is matched}\}$ and $\text{FN}_g = \sum_x \sum_{k:g_k=g} \mathbf{1}\{\text{GT } k \text{ is unmatched}\}$. Define the per-group detection true positive rate (recall)

$$\text{TPR}_g(\tau_{\text{fair}}) = \frac{\text{TP}_g}{\text{TP}_g + \text{FN}_g}.$$

In our fairness metrics, the *favorable outcome* is a *correct detection of a person instance* (i.e., contributing to TP_g at $\text{IoU} \geq \tau_{\text{fair}}$ with correct class). Counts are aggregated *over all instances* in the cohort (micro average across images).

Disparate Impact (DI) for detection. With multiple cohorts, we compute the best and worst group detection rates and form a ratio:

$$\text{DI}(\tau_{\text{fair}}) = \frac{\min_g \text{TPR}_g(\tau_{\text{fair}})}{\max_g \text{TPR}_g(\tau_{\text{fair}})} \in [0, 1], \qquad |1 - \text{DI}| \text{ is reported (lower is better).}$$

This "rate ratio" view is standard for multi-group DI and equals 1 under perfect parity.

Equality of Opportunity gap (ΔEOP) for detection. We report the max range of per-group TPRs:

$$\Delta\text{EOP}(\tau_{\text{fair}}) = \max_g \text{TPR}_g(\tau_{\text{fair}}) \; - \; \min_g \text{TPR}_g(\tau_{\text{fair}}) \in [0, 1],$$

which is 0 under perfect parity (lower is better). Note that both DI and ΔEOP use the *same* matching protocol and τ_{fair}.

Relation to mAP. mAP is computed with the COCO protocol ($\text{IoU} \in \{0.50 : 0.05 : 0.95\}$, class-aware). Fairness metrics use the *single* threshold $\tau_{\text{fair}} = 0.5$ defined above so that TP/FP/FN (and thus TPR) are unambiguous and reproducible.

Reproducibility keys for fairness on detection (*we fix these in all experiments*):
- IoU NMS $= 0.5$; score threshold $\tau_{\text{score}} = 0.25$; max dets per image $= 300$.
- Fairness IoU threshold $\tau_{\text{fair}} = 0.5$ for TP/FP/FN and TPR.
- Greedy one-to-one matching by descending score; ties broken by higher IoU.
- Per-group TPR is micro-averaged over all instances with that group's ground-truth.

Membership Inference Attack (MIA) for detection. We use a black-box shadow-model attack tailored to detectors. Let $\mathcal{M}$ be a set of *member* images used in training and $\bar{\mathcal{M}}$ a disjoint *non-member* set of the

same size from the same distribution. For any image x, we compute an image-level feature vector from the model's post-NMS outputs:

$$\varphi(x) = \big[\ \#\text{dets},\ \bar{\hat{s}},\ \max \hat{s},\ \overline{\alpha_0},\ \overline{u_{\text{cls}}}\ \big],$$

where #dets is the number of predicted person boxes with $\hat{s} \geq \tau_{\text{score}}$, $\bar{\hat{s}}$ is their mean confidence, and $\overline{\alpha_0}$ and $\overline{u_{\text{cls}}}$ are the mean Dirichlet total evidence and its derived epistemic scalar (Sec. A.3) over those detections (empty sets use zeros). We train a logistic-regression (or two-layer MLP) shadow attacker on a disjoint shadow split to predict membership from $\varphi(x)$ and report the *attack success rate*

$$\text{MIA SR} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$

evaluated on $\mathcal{M} \cup \bar{\mathcal{M}}$ with a 50/50 prior.

Attribute Inference Attack (AIA) for detection. We probe sensitive attributes from per-instance features. For each matched detection (as above), we apply ROIAlign to the model's neck feature map at the matched box to get a fixed-size tensor, global-average–pool it to a vector h, and train a two-layer MLP (on a disjoint shadow split) to predict the instance's group $g \in \{1, \ldots, G\}$. We report the *top-1 accuracy* over all matched instances:

$$\text{AIA SR} = \frac{\#\text{correct group predictions}}{\#\text{matched instances}}.$$

Note: Images without matched persons contribute nothing to AIA; they still contribute to MIA.

B. Theoretical Analysis of the Uncertainty Fairness Metric

In federated learning, each client holds a distinct local distribution, resulting in unequal representation of sensitive groups (e.g. demographic cohorts or environmental conditions). Epistemic uncertainty (quantifying a model's ignorance about its predictions) naturally reflects these imbalances: groups with fewer or more variable examples yield higher uncertainty, while well-represented, homogeneous groups yield lower uncertainty. We formalize this insight and show that controlling the dispersion of group-wise uncertainties effectively enforces fairness.

Let each client compute, for each sensitive group $g \in \{1, \ldots, G\}$, an epistemic variance σ_g^2. In evidential models, $\sigma_g^2 = 1/\alpha_g$, where α_g accumulates "evidence" proportional to effective sample size and signal-to-noise ratio. Concretely,

$$\alpha_g \ \propto\ n_g \,\text{SNR}_g \quad \Longrightarrow \quad \sigma_g^2 \ \approx\ \frac{1}{n_g \,\text{SNR}_g}\,.$$

Thus σ_g^2 is large when group g is under-sampled or noisy, and small otherwise.

To measure fairness, we track the relative spread of $\{\sigma_g^2\}$. We define the Uncertainty Fairness Metric (UFM) as

$$\text{UFM} \ =\ \frac{\max_g \sigma_g^2 \ -\ \min_g \sigma_g^2}{\frac{1}{G}\sum_{g=1}^{G} \sigma_g^2 + \epsilon}\,,$$

with $\epsilon > 0$ for stability. By normalizing by the mean uncertainty, UFM is scale-invariant, takes value zero when all groups share equal confidence, and grows smoothly as disparities arise.

Statistical Rationale. Classical generalization bounds for group g involve its sample size n_g and model complexity. A simplified high-probability bound is

$$\mathcal{L}^{(g)} \ \leq\ \hat{\mathcal{L}}^{(g)} + \mathcal{O}\Big(\sqrt{\tfrac{\log(1/\delta)}{n_g}}\Big),$$

where $\hat{\mathcal{L}}^{(g)}$ is the empirical loss. Since $\sigma_g^2 \approx 1/(n_g \,\mathrm{SNR}_g)$, the uncertainty gap $\max_g \sigma_g^2 - \min_g \sigma_g^2$ upper-bounds the disparity in confidence-adjusted generalization terms. Minimizing UFM therefore tightens and balances each group's bound, improving equity in expected loss.

Fair Aggregation. In federated aggregation, each client i reports its UFM_i. The server assigns weights

$$w_i \;=\; \frac{\exp(-\beta\,\mathrm{UFM}_i)}{\sum_j \exp(-\beta\,\mathrm{UFM}_j)},$$

so that clients with lower internal disparity (smaller UFM) contribute more. This bias toward uniformly confident updates automatically re-balances the global model as data distributions evolve, without exposing sensitive attributes centrally.

Notation alignment. Throughout, evidential classification uses Dirichlet evidence with total evidence $\alpha_0(x) = \sum_{c=1}^{C} \alpha_c(x)$. For group g, let $\bar{\alpha}_{0,g} = \mathbb{E}_{x\in\mathcal{D}_g}[\alpha_0(x)]$ and define the group epistemic variance proxy $\sigma_g^2 := \mathbb{E}_{x\in\mathcal{D}_g}[1/\alpha_0(x)] \approx 1/\bar{\alpha}_{0,g}$. We set $\epsilon = 10^{-6}$ in UFM for numerical stability.

Assumptions. (A1) *Bounded loss:* the per-sample loss $\ell \in [0,1]$. (A2) *Evidential calibration:* there exist constants $0 < s_{\min} \le s_{\max}$ such that $\frac{1}{n_g s_{\max}} \le \sigma_g^2 \le \frac{1}{n_g s_{\min}}$ for each group g (i.e., evidence scales with effective sample size and signal-to-noise). (A3) *Group-wise mixing:* samples within a group are i.i.d. under a fixed distribution.

Theorem B.1 (Confidence-adjusted generalization disparity)**.** *Under (A1)–(A3), for any $\delta \in (0,1)$ there exists a constant $C > 0$ (depending only on $s_{\min}, s_{\max}$) such that, with probability at least $1-\delta$, for every group g,*

$$\mathcal{L}^{(g)} \;\le\; \hat{\mathcal{L}}^{(g)} \;+\; C\sqrt{\sigma_g^2\,\log(1/\delta)}.$$

Consequently,

$$\max_g\Big(\mathcal{L}^{(g)} - \hat{\mathcal{L}}^{(g)}\Big) - \min_g\Big(\mathcal{L}^{(g)} - \hat{\mathcal{L}}^{(g)}\Big) \;\le\; C\sqrt{\log(1/\delta)}\,\big(\sqrt{\max_g \sigma_g^2} - \sqrt{\min_g \sigma_g^2}\big).$$

Proof sketch. Hoeffding's inequality yields $\mathcal{L}^{(g)} \le \hat{\mathcal{L}}^{(g)} + O\big(\sqrt{\frac{\log(1/\delta)}{n_g}}\big)$ under (A1),(A3). By (A2), $1/n_g$ is sandwiched by constants times σ_g^2, giving the per-group term $O\big(\sqrt{\sigma_g^2 \log(1/\delta)}\big)$. The disparity bound follows by subtracting the best/worst groups. □

Corollary B.2 (UFM controls disparity)**.** *Let $\bar{\sigma}^2 = \frac{1}{G}\sum_{g=1}^{G} \sigma_g^2$. Then there exist constants $C_1, C_2 > 0$ such that*

$$\max_g\Big(\mathcal{L}^{(g)} - \hat{\mathcal{L}}^{(g)}\Big) - \min_g\Big(\mathcal{L}^{(g)} - \hat{\mathcal{L}}^{(g)}\Big) \;\le\; C_1\sqrt{\log(1/\delta)}\,\bar{\sigma}\;\mathrm{UFM} \;\le\; C_2\sqrt{\log(1/\delta)}\,\mathrm{UFM},$$

i.e., minimizing UFM tightens a normalized upper bound on the group disparity in confidence-adjusted generalization.

Aggregation limits and practice. We compute UFM_i per client on a *held-out local validation split* and report $w_i \propto \exp(-\beta\,\mathrm{UFM}_i)$. As $\beta \to 0$ we recover *uniform* averaging; as $\beta \to \infty$ the aggregator concentrates on clients with smallest UFM. To reduce noise, we use an exponential moving average of UFM_i across rounds.

Integrity of reported UFM We assume an honest-but-curious server. Because the aggregation weights ω_i in Eq. 5–Eq. 6 require *per-client* inputs, each client transmits its scalar $u_i = \mathrm{UFM}_i$ *in clear* after applying a publicly announced clipping rule,

$$u_i \;\leftarrow\; \min\big(\max(u_i, a),\, b\big), \quad (a, b) \text{ fixed.}$$

Model updates $\Delta\theta_i$ continue to use secure aggregation; only the clipped scalar u_i is visible to the server. The server then computes $\omega_i \propto \exp(-\beta\, u_i)$ and updates θ_G as in Eq. 5–Eq. 6. Further implementation details appear in Appendix F.

C. Information-Theoretic Guarantees of Adversarial Privacy Disentanglement

We analyze how adversarial training with a gradient reversal layer and an attribute classifier limits the leakage of a sensitive attribute S from representations $H = f_\theta(X)$. Throughout, we allow H to be a (possibly stochastic) mapping of X (e.g., due to dropout); all logarithms are natural (nats).

Adversarial objective and conditional entropy. Consider the minimax problem

$$\min_\theta \max_\phi \; \mathcal{L}_{\mathrm{adv}}(\theta, \phi), \qquad \mathcal{L}_{\mathrm{adv}}(\theta, \phi) \;=\; -\mathbb{E}_{(X,S)}\big[\log A_\phi(S \mid H)\big],$$

where $A_\phi(\cdot \mid H)$ is the attribute classifier fed by the representation $H = f_\theta(X)$. If the adversary family is universally expressive and the supremum is attained, then

$$\sup_\phi \mathcal{L}_{\mathrm{adv}}(\theta, \phi) \;=\; H(S \mid H).$$

In general (with approximation/optimization error), we have the one–sided relation

$$\sup_\phi \mathcal{L}_{\mathrm{adv}}(\theta, \phi) \;\le\; H(S \mid H).$$

Consequently,

$$I(H; S) = H(S) - H(S \mid H) \;\le\; H(S) - \sup_\phi \mathcal{L}_{\mathrm{adv}}(\theta, \phi),$$

so maximizing $\mathcal{L}_{\mathrm{adv}}$ (for fixed θ) *minimizes* an upper bound on $I(H; S)$. By the data–processing inequality, reducing $I(H; S)$ weakens any inference from H about S.

Attack error via Fano. Let an attacker output $\widehat{S} = g(H)$ over K attribute classes. Fano's inequality yields

$$P_e \;\ge\; 1 - \frac{I(H; S) + \log 2}{\log K}.$$

Hence as $\sup_\phi \mathcal{L}_{\mathrm{adv}}(\theta, \phi) \;\to\; H(S \mid H)$ (i.e., $I(H; S) \;\to\; 0$), the minimum achievable error satisfies $P_e \to 1 - 1/K$, driving attribute inference toward chance level.

Privacy–utility frontier and tuning. Incorporating the adversarial term with coefficient λ_{priv} into the local objective,

$$\mathcal{L}_{\text{local}}(\theta, \phi) = \mathcal{L}_{\text{task}}(\theta) + \lambda_{\text{priv}}\, \mathcal{L}_{\mathrm{adv}}(\theta, \phi),$$

traces a (piecewise) convex frontier in the $\big(I(H; S),\, \mathcal{L}_{\text{task}}\big)$ plane under standard regularity conditions. By the envelope theorem,

$$-\frac{d\, I(H; S)}{d\, \mathcal{L}_{\text{task}}} \;=\; \frac{\partial_{\lambda_{\text{priv}}} \mathcal{L}_{\text{task}}}{\partial_{\lambda_{\text{priv}}} I(H; S)},$$

so λ_{priv} directly tunes the privacy–utility balance: larger λ_{priv} increases the pressure to maximize $\mathcal{L}_{\mathrm{adv}}$, thereby decreasing $I(H; S)$ (stronger privacy) at the potential cost of task loss.

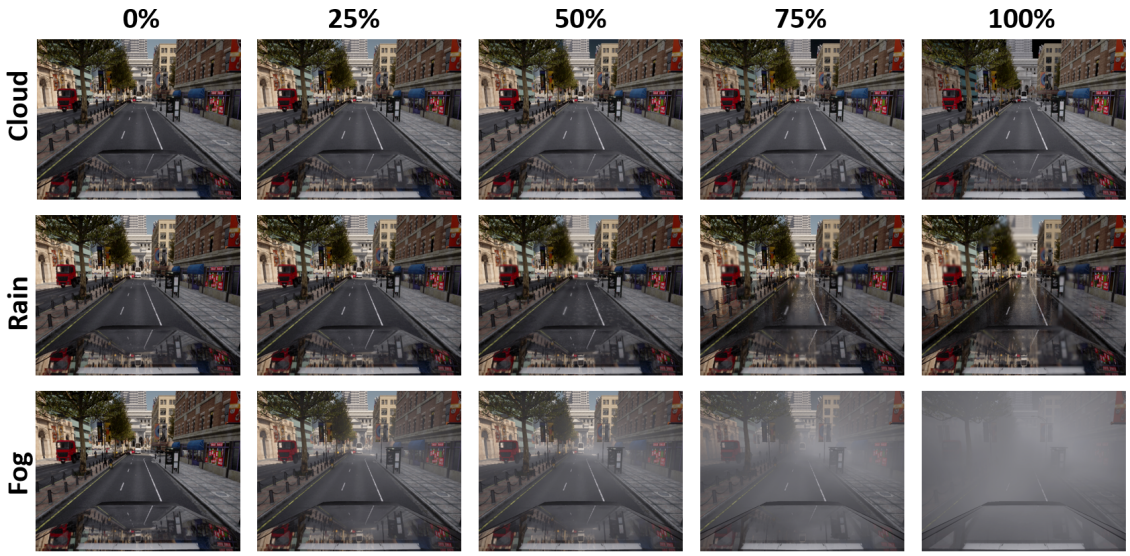


Figure 4: Sample visualization of weather conditions (cloud, rain, and fog) at increasing intensity levels (0%, 25%, 50%, 75%, 100%) using the CARLA simulation, illustrating how environmental severity gradually impacts visibility and scene clarity.

Algorithm 1 `RESFL` Training with Adversarial Privacy and Uncertainty-Guided Aggregation

Input: global model θ_G, adversary $A(x;\phi)$, client data $\{\delta_i\}$, weights $\lambda_{\text{priv}}, \lambda_{\text{fair}}$, temperature β, learning rates η, η_ϕ, rounds T
for $t = 0 \to T-1$ **do**
  Server broadcasts $\theta_G^{(t)}$ to all clients
  for each client i **in parallel do**
    Initialize: $\theta_i \leftarrow \theta_G^{(t)}$, $\phi_i \leftarrow \phi$
    for each local step **do**
      Compute $L_{\text{task}}, L_{\text{adv}}, L_{\text{unc}}$
      $\phi_i \leftarrow \phi_i - \eta_\phi \nabla_{\phi_i} L_{\text{adv}}$
      $\theta_i \leftarrow \theta_i - \eta \nabla_{\theta_i} [L_{\text{task}} + \lambda_{\text{priv}} L_{\text{adv}} + \lambda_{\text{fair}} L_{\text{unc}}]$ ▷ Composite local loss (Eq. 9)
    end for
    Compute UFM_i (Eq. 4 and $\Delta\theta_i = \theta_i - \theta_G^{(t)}$)
    Client sends $\{\Delta\theta_i, \text{UFM}_i\}$ to server
  end for
  Server computes $\omega_i \propto \exp(-\beta \cdot \text{UFM}_i)$ (Eq. 5)
  Update: $\theta_G^{(t+1)} = \theta_G^{(t)} + \sum_{i=1}^{N} \omega_i \Delta\theta_i$ ▷ Aggregation (Eq. 6)
end for
Output: final global model $\theta_G^{(T)}$

Scope of the guarantee. These are *information-theoretic* guarantees on representation leakage $I(H;S)$; they are not (ε, δ)–DP guarantees on the training algorithm. In practice, one monitors $\mathcal{L}_{\text{adv}}$ (or a calibrated surrogate) and adjusts λ_{priv} to meet a target inference–error bound while limiting degradation in $\mathcal{L}_{\text{task}}$.

D. Joint Trade-off Analysis of Privacy and Fairness

Our RESFL framework simultaneously addresses three competing objectives—detection utility, attribute privacy, and demographic fairness—by integrating adversarial privacy disentanglement with uncertainty-guided aggregation (summarized in Algorithm 1. The adversarial module employs a gradient reversal layer and attribute classifier to suppress sensitive information in each client's features, effectively minimizing the mutual information between latent representations and protected attributes. This enforces a controllable privacy constraint without degrading task performance unduly. In parallel, each client's evidential uncertainty head estimates per-group epistemic variances, from which we compute an Uncertainty Fairness Metric (UFM) that quantifies disparities in model confidence across sensitive cohorts. During aggregation, clients report both their parameter updates and UFM scores; the server then weights each update by a softmax of the negative UFM, amplifying contributions from clients with more uniform confidence and down-weighting those with high disparity. By tuning the adversarial strength λ_{priv} and the uncertainty coefficient λ_{fair}, RESFL effectively scalarizes a convex multi-objective problem, tracing out the full Pareto frontier in the space of utility, privacy leakage, and fairness gap. Unlike single-objective baselines—which either sacrifice accuracy for privacy protection or apply fixed fairness regularizers—RESFL dynamically balances both axes: stronger adversarial signals tighten privacy guarantees, while uncertainty-based weights correct emerging fairness imbalances. Empirically and theoretically, this joint mechanism dominates pure DP or fairness-only schemes by exploring descent directions unavailable to one-dimensional fixes, yielding models that maintain high mean average precision, provably low attribute-inference risk, and minimal performance disparity across sensitive groups.

E. Datasets

Our experiments leverage two complementary data sources: FACET for federated training and CARLA for controlled evaluation, to measure object detection utility, demographic fairness, privacy resilience, and robustness under diverse conditions. In this section, we detail dataset composition, annotation processing, domain-specific partitioning, and preprocessing pipelines.

Figure 5: Example images from the FACET dataset. Each red bounding box denotes a detected person instance, annotated with its corresponding Monk Skin Tone (MST) label (e.g. MST #2, #3, #4, #6). These samples illustrate the range of skin-tone levels (1 = lightest to 10 = darkest) used for fairness evaluation in our object detection experiments.

E.1. FACET Dataset

The FACET dataset (Gustafson et al., 2023) provides 32 000 real-world images with over 50 000 annotated person instances, each labeled with a bounding box and multiple attributes (perceived skin tone, hair type,

person class). We concentrate on perceived skin tone, a sensitive attribute strongly correlated with performance gaps in detection model (Pathiraja et al., 2024). FACET adopts the Monk Skin Tone (MST) scale with ten discrete levels $g \in \{1, \ldots, 10\}$, where $g = 1$ is the lightest and $g = 10$ the darkest tone, as shown in Figure 6. To mitigate annotation noise due to lighting or labeler variance, each instance receives n independent MST labels $s_1, \ldots, s_n$. We aggregate these as

$$s^* = \frac{1}{n} \sum_{i=1}^{n} s_i,$$

then discretize s^* by rounding to the nearest integer in $\{1, \ldots, 10\}$. This yields a robust single-toned label per instance, denoted $\mathrm{MST}(b)$.

We group the dataset into $G = 10$ MST cohorts. Let $\mathcal{D} = \{(x_k, b_k, s_k^*)\}_{k=1}^{N}$ be the full set of image–instance pairs ($N \approx 50\,000$). Define

$$\mathcal{D}_g = \big\{(x, b, s^*) : \mathrm{MST}(b) = g\big\}, \quad g = 1, \ldots, 10,$$

so that $\sum_{g=1}^{10} |\mathcal{D}_g| = N$. In practice, each $|\mathcal{D}_g|$ ranges from approximately 4 000 to 6 000 instances, ensuring sufficient representation across the skin-tone spectrum.

To simulate federated clients, we partition the 32 000 FACET images into $K = 4$ i.i.d. subsets $\{\mathcal{I}_i\}_{i=1}^{4}$, each containing 8 000 images and all associated instances. Formally, $\mathcal{I}_i \cap \mathcal{I}_j = \emptyset$ for $i \neq j$, $\bigcup_{i=1}^{4} \mathcal{I}_i$ covers all images, and each split preserves the MST distribution:

$$\forall\, g,\ \big|\{(x, b) \in \mathcal{D}_g : x \in \mathcal{I}_i\}\big| \approx \tfrac{1}{4}|\mathcal{D}_g|.$$

Clients share only model updates (gradients $\Delta\theta_i$ and a scalar UFM per round), never raw images or labels. A few samples are visible in Figure 5.

Preprocessing and Augmentation. Each image is resized to 640×640 pixels using bicubic interpolation. We apply standard YOLOv8 augmentations: random horizontal flip (probability 0.5), brightness and contrast jitter ($\pm 20\%$), and random hue shift ($\pm 10\%$). Pixel values are normalized to $[0, 1]$ and then standardized using ImageNet channel means $\mu = [0.485, 0.456, 0.406]$ and standard deviations $\sigma = [0.229, 0.224, 0.225]$. During training, we further apply mosaic augmentation by stitching four images into a 1×1 grid with random scaling in $[0.5, 1.5]$.

E.2. CARLA Simulation Dataset

The CARLA simulator (Dosovitskiy et al., 2017) v0.9.13 generates synthetic driving scenarios to evaluate model robustness under controlled environmental and urban variations. We select three canonical maps: Town01 (suburban streets), Town03 (dense downtown), and Town05 (mid-density mixed-use) to capture a broad spectrum of road geometry, building density, and occlusion patterns.

An autopilot-enabled ego vehicle equipped with an RGB camera (1920×1080, 100° FOV) and a semantic segmentation sensor (same specs) records frames every 3s. We retain only frames containing at least one pedestrian. Pedestrian bounding boxes

$$b = (x_{\min},\, y_{\min},\, x_{\max},\, y_{\max} \in \mathbb{R}^4$$

are extracted via connected-component analysis on semantic masks, discarding detections with fewer than 50 pixels.

Skin-tone assignment. Each synthetic pedestrian blueprint is manually mapped to a Monk Skin Tone (MST) label by visual inspection, ensuring consistency with the FACET scale. Let

$$\mathcal{C} = \{\mathrm{Town01},\ \mathrm{Town03},\ \mathrm{Town05}\}, \quad \mathcal{S} = \{1, \ldots, 10\}.$$

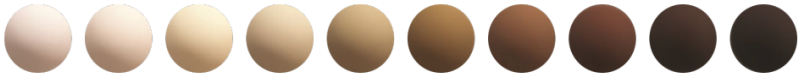


Figure 6: The Monk Skin Tone (MST) scale (Pathiraja et al., 2024) ranges from MST=1, representing the lightest skin tone, to MST=10, representing the darkest skin tone.

Each pedestrian instance receives a pair $(c, s) \in \mathcal{C} \times \mathcal{S}$.

Domain adaptation fine-tuning. To reduce domain shift, we fine-tune the federated global model on clear-weather CARLA frames. For each $c \in \mathcal{C}$ and $s \in \mathcal{S}$ we sample 200 frames, yielding

$$N_{\text{tune}} = |\mathcal{C}| \times |\mathcal{S}| \times 200 = 3 \times 10 \times 200 = 6000$$

tuning samples. Fine-tuning uses the same SGD hyperparameters as federated local updates.

Adverse-weather evaluation. We evaluate under 13 conditions: a clear baseline plus fog and rain at intensities $\alpha \in \{0, 25, 50, 75, 100\}\%$. For each c, s, and weather condition w we capture 20 frames,

$$N_{\text{eval}} = |\mathcal{C}| \times |\mathcal{S}| \times 13 \times 20 = 3 \times 10 \times 13 \times 20 = 7800.$$

This design isolates the effects of environmental severity and urban topology on detection, fairness, and privacy metrics.

Preprocessing. RGB frames are downsampled to 640×640 and normalized as in training; no random augmentations are applied at test time. Semantic masks are used only for bounding-box extraction.

E.2.1. Dataset Statistics

Table 3: Composition and partitioning of FACET and CARLA datasets

	FACET	**CARLA (Tune)**	**CARLA (Eval)**
Images	32 000	6 000	7 800
Person instances	50 000	6 515	8 163
Skin-tone levels	10	10	10
Urban layouts	N/A	3	3
Weather conditions	N/A	clear	13
Frames per (c,s,w)	N/A	200	20

Table 3 summarizes our real and simulated datasets. FACET provides 32000 images and 50000 person instances across ten MST levels, partitioned into four IID client shards of 8000 images each. The CARLA fine-tune set contains 6000 clear-weather frames (200 per town–skin-tone pair), and the evaluation set comprises 7800 frames collected over three towns under thirteen weather conditions, with twenty frames per (town, skin-tone, weather) tuple. This balanced design ensures comprehensive coverage for assessing `RESFL`'s performance under varied real-world and synthetic scenarios.

F. Implementation Details

All components of `RESFL` were implemented in Python (v3.9) using the PyTorch (v2.0) framework, and all experiments were executed on a single NVIDIA RTX 3070 GPU with 8 GB of VRAM. We adopted the

YOLOv8 object-detection architecture as our backbone, modifying its final detection head to emit nonnegative concentration vectors for evidential uncertainty estimation. Specifically, we replaced the standard softmax head with a softplus-plus-one layer to produce Dirichlet concentration parameters $\alpha_c = 1 + \ln(1 + e^{z_c})$. All code, including dataset wrappers, training scripts, and analysis notebooks, will be released upon publication to ensure full reproducibility.

Training proceeded in a standard federated loop over $T = 100$ communication rounds. In each round, the server distributed the current global model $\theta_G^{(t)}$ to $N = 4$ clients. Each client locally trained for one epoch (one full pass over its data) using stochastic gradient descent with momentum 0.9 and weight decay 1×10^{-4}. The initial learning rate was set to 1×10^{-3} and decayed by a factor of 0.1 at epochs 50 and 75. We fixed the batch size to 64 samples per step, and used a per-client training time of approximately 8 hours on the RTX 3070. All results are averaged over three independent random seeds to account for variability in data splits and optimizer initialization.

For the FACET dataset (Gustafson et al., 2023), we loaded 32,000 annotated images and partitioned them into four i.i.d. subsets of 8,000 images each, preserving the overall Monk Skin Tone distribution in each split. No raw image data were exchanged during training, clients only uploaded model gradients $\Delta\theta_i$ and a single Uncertainty Fairness Metric (UFM) scalar per round. For the CARLA simulator experiments (Dosovitskiy et al., 2017), we first fine-tuned the global model on 2,000 "neutral-weather" frames (200 per MST level), then evaluated on 2,600 held-out frames spanning 13 weather conditions (clear, fog, rain at five intensities each).

We set the adversarial gradient-reversal coefficient λ_{priv} to 0.1 and the uncertainty regularization weight λ_{fair} to 0.01. The server aggregation temperature β was chosen as 2.0 based on a preliminary grid search balancing fairness sensitivity against raw accuracy. All four clients performed synchronized local updates in each round, and the server aggregated via

$$\theta_G^{(t+1)} = \theta_G^{(t)} + \eta \sum_{i=1}^{N} \frac{\exp(-\beta\, \mathrm{UFM}_i)}{\sum_{j=1}^{N} \exp(-\beta\, \mathrm{UFM}_j)}\, \Delta\theta_i.$$

Hyperparameter values, dataset splits, and training protocols are summarized in Table 4.

Table 4: Implementation environment and hyperparameter settings

Category	**Setting**
Framework	PyTorch v2.0, Python 3.9
Hardware	NVIDIA RTX 3070 (8 GB VRAM)
Backbone	YOLOv8 with evidential head
Optimizer	SGD, momentum 0.9, weight decay 1×10^{-4}
Initial learning rate	1×10^{-3}, decayed by 0.1 at epochs 50, 75
Batch size	64
Communication rounds	100
Clients	4 (i.i.d. splits of FACET)
FACET split	32 k images $\rightarrow$ 4×8 k images
CARLA fine-tune / eval	6 k / 7.8 k frames (13 scenarios)
λ_{priv} (GRL)	0.1
λ_{fair} (uncertainty)	0.01
Aggregation temperature β	2.0
Random seeds	3
Per-client training time	8 h per 100 epochs

Table 5: FACET results under IID vs. Non-IID with **4** clients. Accuracy is mAP (higher is better). Fairness/Privacy Scores are averages of ($|1-\text{DI}|$, ΔEOP) and (MIA SR, AIA SR), respectively (lower is better).

	IID			Non-IID		
Algorithm	**Accuracy**	**Fairness Score**	**Privacy Score**	**Accuracy**	**Fairness Score**	**Privacy Score**
FedAvg	0.6378	**0.2261**	0.3886	0.4841	0.3892	0.4317
FedAvg-DP	0.2932	0.4048	**0.1960**	0.1757	0.4853	**0.2253**
FairFed	**0.7013**	0.2529	0.4832	**0.5080**	**0.2989**	0.5122
PUFFLE	0.4192	0.3348	0.2817	0.2726	0.3650	0.3140
Ours (`RESFL`	**0.6654**	**0.2123**	**0.1963**	**0.5384**	**0.2387**	**0.2131**

Table 6: FACET results under IID vs. Non-IID with **8** clients. Accuracy is mAP (higher is better). Fairness/Privacy Scores are averages of ($|1-\text{DI}|$, ΔEOP) and (MIA SR, AIA SR), respectively (lower is better).

	IID			Non-IID		
Algorithm	**Accuracy**	**Fairness Score**	**Privacy Score**	**Accuracy**	**Fairness Score**	**Privacy Score**
FedAvg	0.6217	**0.2395**	0.3973	0.3615	0.4279	0.4893
FedAvg-DP	0.2791	0.4179	**0.2067**	0.1327	0.5381	**0.2765**
FairFed	**0.6895**	0.2647	0.4216	**0.4284**	**0.3571**	0.5695
PUFFLE	0.3927	0.3529	0.2953	0.1983	0.4237	0.3719
Ours (`RESFL`	**0.6539**	**0.2197**	**0.2059**	**0.4627**	**0.2975**	**0.2635**

All experiments spanning privacy and fairness attacks, adversarial robustness tests, and ablation studies use the same training pipeline above. Our release will include detailed setup instructions, random seed logs, and pre-trained model checkpoints to facilitate both replication and future extension.

Compute resources. All experiments were run on a single workstation equipped with an NVIDIA RTX 3070 GPU (8 GB VRAM), an Intel Core i7-10700K CPU (8 cores, 16 threads) and 32 GB DDR4 RAM, with datasets and logs stored on a 1 TB NVMe SSD. Each 100-round federated training session required $\approx$8 hours of GPU time and $\approx$1 hour of CPU overhead per seed. CARLA fine-tuning and evaluation took $\approx$1.5 hours of GPU time per seed. Averaged over three random seeds, the total compute amounted to $\approx$27 GPU-hours and $\approx$12 CPU-hours, and consumed $\approx$50 GB of disk storage.

Table 7: Results on **Adult** and **TweetEval** with **4** clients (IID split, sensitive attribute = gender). Accuracy is overall classification accuracy ($\uparrow$). Fairness/Privacy Scores are lower-is-better ($\downarrow$).

	Adult			TweetEval		
Algorithm	**Accuracy** $\uparrow$	**Fairness Score** $\downarrow$	**Privacy Score** $\downarrow$	**Accuracy** $\uparrow$	**Fairness** $\downarrow$	**Privacy** $\downarrow$
FedAvg	**0.8527**	0.3185	0.3721	**0.5258**	0.0439	0.3426
FedAvg-DP	0.7063	0.3018	**0.2217**	0.3724	0.0415	**0.1950**
FairFed	0.8449	**0.2564**	0.3965	**0.5310**	**0.0360**	0.4079
PUFFLE	0.8294	0.2951	0.2853	0.4959	0.0441	0.2851
Ours (RESFL)	**0.8481**	**0.2317**	**0.2389**	0.5067	**0.0334**	**0.2353**

G. FACET EVALUATION RESULTS

We evaluate on FACET using two federation sizes (4 and 8 clients) under *IID* and *Non-IID* partitions. For *IID*, we perform stratified sampling that preserves the joint distribution of task labels and sensitive groups within each client; each client thus receives an equal-sized subset with approximately identical class and group proportions. For *Non-IID*, we induce heterogeneity via Dirichlet allocation over class–group pairs with concentration $\alpha = 0.5$. The total number of samples and the per-client sizes are matched across IID/Non-IID, and all methods use the same local training budget and optimizer settings; full hyperparameters are listed in Appendix F.

IID vs. Non-IID with 4 clients (Table 5. Under IID, `RESFL` attains the best combined fairness–privacy profile while preserving competitive mAP. Compared to FedAvg, `RESFL` reduces the fairness score (lower is better) from 0.2261 to 0.2123 and the privacy score from 0.3886 to 0.1963, with a modest accuracy gain over most baselines. FedAvg-DP ($\epsilon = 0.1$) achieves the strongest privacy (0.1960) but at a steep accuracy cost (0.2932), illustrating the classic privacy–utility tension. FairFed delivers the highest mAP (0.7013) but with weaker privacy (0.4832). Under Non-IID, all methods degrade—as expected—yet `RESFL` retains the lowest fairness (0.2387) and near-best privacy (0.2131) with the top mAP (0.5384), indicating robustness to client heterogeneity.

Scaling to 8 clients (Table 6. The trends persist when increasing the client count: under IID, `RESFL` again achieves strong accuracy (0.6539) with the best fairness (0.2197) and near-best privacy (0.2059). In Non-IID, the gap between data-size–agnostic and DP-based methods widens; FedAvg-DP preserves privacy (0.2765) but collapses in mAP (0.1327). FairFed remains accuracy-leaning (0.4284) yet with weaker privacy (0.5695). `RESFL` continues to balance all three criteria (mAP 0.4627; fairness 0.2975; privacy 0.2635), suggesting that uncertainty-guided aggregation can mitigate distributional skew without over-penalizing utility. All scores are averaged over multiple runs; full seed-wise statistics and confidence intervals are provided in the supplementary analysis.

H. CARLA EVALUATION RESULTS

Table 8 shows accuracy (mAP), fairness ($|1-\mathrm{DI}|$, $\Delta\mathrm{EOP}$), privacy-attack success rates (MIA SR, AIA SR), and robustness metrics (BA AD, DPA EODD) for all FL algorithms under varying cloud intensities.Table 9 reports the same set of metrics across rain intensity levels. Table 10 presents these metrics under fog conditions at increasing severity.

I. ADDITIONAL EVALUATION

To assess the domain-agnostic capability of `RESFL`, we conducted experiments on two distinct modalities: tabular data (Adult (Basile et al., 2019) income prediction) and textual data (TweetEval (Barbieri et al., 2020) sentiment classification). For Adult, we used a lightweight TabularNet architecture with three fully-connected layers and ReLU activations, trained to predict whether income exceeds $50K based on census features and set 'race' as sensitive attribute). For TweetEval, we fine-tuned DistilBERT with a classification head on the sentiment analysis task. In both cases, we created a federation of four clients with an IID split, ensuring that the class distribution and the sensitive attribute (gender) proportions were approximately balanced across clients. We used the same training protocol as in the main experiments, with local SGD updates, a fixed number of local epochs, and the same optimizer hyperparameters. Fairness was measured using the average of demographic parity gap ($|1 - \mathrm{DI}|$) and equality-of-opportunity gap ($\Delta\mathrm{EOP}$), while privacy leakage was quantified by the success rates of Membership Inference (MIA) and Attribute Inference (AIA) attacks, following the methodology in Section A.4.

Table 7 reports the results. On Adult income prediction, RESFL achieves competitive accuracy (0.8481) while attaining the lowest fairness disparity (0.2317) and strong privacy protection (0.2389), outperforming FedAvg and FairFed in terms of fairness without sacrificing utility. FedAvg-DP improves privacy but incurs a large accuracy drop (0.7063), highlighting the advantage of our adversarial privacy disentanglement which preserves utility. On TweetEval, RESFL also achieves a favorable fairness score (0.0334) with improved privacy compared to FedAvg, indicating that the uncertainty-guided aggregation generalizes to textual tasks. These results collectively demonstrate that RESFL is not restricted to vision-based detection but applies broadly to heterogeneous data modalities, reinforcing its claim as a domain-agnostic framework for responsible federated learning.

J. Broader Impacts

RESFL aims to make federated learning safer and more equitable across domains such as autonomous driving, healthcare, and edge sensing by improving group fairness and reducing sensitive-attribute leakage without sharing raw data. Its uncertainty-guided aggregation can help models remain reliable under distribution shift and adverse conditions, potentially improving real-world safety and user trust. At the same time, risks remain: stakeholders could tout fairness or privacy benefits without adequate validation, obscure data quality issues behind adversarial masking, or impose extra compute/communication costs that burden smaller clients. These concerns call for transparent reporting of hyperparameters and metrics, independent audits of fairness–privacy–utility trade-offs, and safeguards against gaming self-reported signals. Overall, RESFL offers a practical step toward responsible FL while highlighting the need for oversight, reproducibility artifacts, and domain-specific governance in deployments.

LLM Usage

We used an LLM (GPT-5 Thinking) only to aid writing polish and literature discovery. For writing, it suggested alternative phrasings, grammar/clarity edits, and minor LaTeX fixes; all technical content, claims, math, algorithmic choices, figures, tables, and results were authored and verified by the authors. For retrieval, it helped brainstorm search queries and surface candidate related-work papers; we independently checked every citation and read primary sources before inclusion. The LLM did not generate datasets, code, experiments, proofs, or results, nor did it design evaluations. We reviewed and edited any suggested text to ensure originality and accuracy, and we did not include verbatim model output beyond trivial boilerplate. No sensitive or proprietary data were shared in prompts. This usage is also disclosed in the submission form.

Table 8: Performance comparison of federated learning algorithms under **Cloud** in CARLA simulation: The table reports accuracy (mAP), fairness ($|1 - \text{DI}|$, ΔEOP), privacy risks (MIA, AIA), and robustness (BA AD, DPA EODD) across cloud intensity levels.

Algorithm	Cloud Intensity (%)	Utility	Fairness		Privacy Attacks		Robustness Attack	
		Overall mAP	$\|1 - \text{DI}\|$	ΔEOP	MIA SR	AIA SR	BA AD	DPA EODD
FedAvg	0	0.3952	0.2356	0.2446	0.3915	0.4235	0.1531	0.0738
	25	0.4005	0.2462	0.2460	0.3980	0.4085	0.1053	0.0821
	50	0.3850	0.2394	0.2501	0.4052	0.3520	0.0975	0.0769
	75	0.3662	0.2535	0.2552	0.4105	0.3401	0.0908	0.0952
	100	0.3387	0.2587	0.2604	0.4203	0.3328	0.0803	0.0893
FedAvg-DP	0	0.2741	0.3557	0.3789	0.2327	0.2494	0.1834	0.1842
	25	0.2538	0.3681	0.3802	0.2382	0.2023	0.1254	0.1950
	50	0.2520	0.3705	0.3828	0.2453	0.2501	0.1107	0.1885
	75	0.2205	0.3852	0.3871	0.2551	0.2387	0.0958	0.1782
	100	0.1890	0.3905	0.3922	0.2658	0.2204	0.0852	0.1707
FairFed	0	0.5013	0.2759	0.2593	0.3930	0.4384	0.2132	0.0638
	25	0.4782	0.2781	0.2622	0.3984	0.4420	0.1759	0.0704
	50	0.4845	0.2803	0.2650	0.4057	0.4483	0.1602	0.0807
	75	0.4281	0.3045	0.2689	0.4120	0.4557	0.1453	0.0945
	100	0.3820	0.3190	0.2725	0.4201	0.4635	0.1307	0.0856
PUFFLE	0	0.3526	0.3016	0.3882	0.2636	0.2863	0.1352	0.1673
	25	0.3502	0.3050	0.3905	0.2707	0.2921	0.1508	0.1785
	50	0.3450	0.3285	0.3942	0.2751	0.2985	0.1357	0.1614
	75	0.3389	0.3422	0.3987	0.2825	0.3054	0.1203	0.1831
	100	0.3128	0.3565	0.4034	0.2901	0.3132	0.1052	0.1909
Ours (RESFL	0	**0.4621**	0.2332	0.2434	0.1939	0.1420	0.2726	0.0807
	25	0.4600	0.2555	0.2452	0.1985	0.1482	0.1658	0.0912
	50	0.4557	0.2789	0.2489	0.2057	0.1573	0.1504	0.0783
	75	0.4008	0.2925	0.2523	0.2121	0.1658	0.1357	0.1025
	100	0.3851	0.3070	0.2575	0.2205	0.1727	0.1202	0.1093

Table 9: Performance comparison of federated learning algorithms under **Rain** in CARLA simulation: The result presents accuracy (mAP), fairness ($|1 - \text{DI}|$, ΔEOP), privacy risks (MIA, AIA), and robustness (BA AD, DPA EODD) across rain intensity levels.

Algorithm	Rain Intensity (%)	Utility	Fairness		Privacy Attacks		Robustness Attack	
		Overall mAP	$\|1 - \text{DI}\|$	ΔEOP	MIA SR	AIA SR	BA AD	DPA EODD
FedAvg	0	0.3852	0.2356	0.2446	0.3915	0.4235	0.1531	0.0738
	25	0.3801	0.2389	0.2485	0.3998	0.4302	0.1307	0.0814
	50	0.3705	0.2441	0.2540	0.4072	0.4375	0.1185	0.0912
	75	0.3583	0.2515	0.2628	0.4150	0.4451	0.1023	0.1028
	100	0.3120	0.2580	0.2702	0.4228	0.4527	0.0790	0.1305
FedAvg-DP	0	0.2741	0.3557	0.3789	0.2327	0.2494	0.1834	0.1842
	25	0.2705	0.3583	0.3821	0.2425	0.2459	0.1264	0.1953
	50	0.2672	0.3608	0.3854	0.2501	0.2653	0.1109	0.2012
	75	0.2621	0.3655	0.3902	0.2585	0.2321	0.0987	0.2250
	100	0.2289	0.3708	0.3950	0.2683	0.2203	0.0552	0.2904
FairFed	0	0.5013	0.2759	0.2593	0.3930	0.4384	0.2132	0.0638
	25	0.4950	0.2782	0.2625	0.4008	0.4453	0.1752	0.0725
	50	0.4820	0.2850	0.2703	0.4125	0.4550	0.1598	0.0914
	75	0.4652	0.2980	0.2810	0.4250	0.4705	0.1257	0.1042
	100	0.4380	0.3125	0.2947	0.4401	0.4852	0.1004	0.1501
PUFFLE	0	0.3526	0.3016	0.3882	0.2636	0.2863	0.1352	0.1673
	25	0.3500	0.3060	0.3905	0.2703	0.2908	0.1527	0.1785
	50	0.3452	0.3105	0.3940	0.2785	0.2983	0.1358	0.1895
	75	0.3385	0.3157	0.3987	0.2850	0.3050	0.1003	0.2259
	100	0.3023	0.3202	0.4043	0.2951	0.3128	0.0859	0.2881
Ours (RESFL	0	**0.4621**	0.2332	0.2434	0.1939	0.1420	0.2726	0.0807
	25	0.4605	0.2357	0.2467	0.1984	0.1471	0.1589	0.0925
	50	0.4560	0.2389	0.2503	0.2052	0.1552	0.1403	0.1082
	75	0.4508	0.2425	0.2545	0.2121	0.1658	0.1204	0.1301
	100	0.4151	0.2470	0.2598	0.2205	0.1727	0.0753	0.1803

Table 10: Performance comparison of federated learning algorithms under **Fog** in CARLA simulation: The result presents accuracy (mAP), fairness ($|1 - \text{DI}|$, ΔEOP), privacy risks (MIA, AIA), and robustness (BA AD, DPA EODD) across fog intensity levels.

Algorithm	Fog Intensity (%)	Utility	Fairness		Privacy Attacks		Robustness Attack	
		Overall mAP	$\|1 - \text{DI}\|$	ΔEOP	MIA SR	AIA SR	BA AD	DPA EODD
FedAvg	0	0.3952	0.2356	0.2446	0.3915	0.4235	0.1531	0.0738
	25	0.3650	0.2402	0.2605	0.4251	0.4357	0.1483	0.0851
	50	0.3157	0.2650	0.2872	0.4175	0.4901	0.0891	0.1002
	75	0.1304	0.3853	0.4157	0.4805	0.5502	0.0908	0.1657
	100	0.0001	0.5202	0.5358	0.6153	0.6950	0.0000	0.3203
FedAvg-DP	0	0.2741	0.3557	0.3789	0.2327	0.2494	0.1834	0.1842
	25	0.2500	0.3723	0.4001	0.2801	0.2703	0.1602	0.2015
	50	0.2058	0.3905	0.4502	0.3156	0.3457	0.0558	0.2301
	75	0.0953	0.5058	0.5204	0.4123	0.4605	0.0053	0.3879
	100	0.0000	0.6852	0.7285	0.5207	0.5784	0.0000	0.4907
FairFed	0	0.5013	0.2759	0.2593	0.3930	0.4384	0.2132	0.0638
	25	0.4950	0.2805	0.2681	0.4052	0.4552	0.2085	0.0709
	50	0.4608	0.3107	0.2978	0.4451	0.4703	0.1505	0.0953
	75	0.1952	0.4503	0.3859	0.5085	0.5598	0.1104	0.2156
	100	0.0753	0.5801	0.4902	0.5802	0.6350	0.0753	0.2851
PUFFLE	0	0.3526	0.3016	0.3882	0.2636	0.2863	0.1352	0.1673
	25	0.3458	0.3152	0.4027	0.3104	0.3057	0.1205	0.1854
	50	0.2801	0.3682	0.4558	0.3405	0.3708	0.1104	0.2128
	75	0.0957	0.4827	0.5782	0.4256	0.5083	0.0552	0.3304
	100	0.0125	0.6250	0.7208	0.5507	0.6005	0.0125	0.4708
Ours (RESFL	0	**0.4621**	0.2332	0.2434	0.1939	0.1420	0.2726	0.0807
	25	0.4503	0.2452	0.2583	0.2054	0.1658	0.1859	0.0910
	50	0.4051	0.2780	0.2872	0.2601	0.2153	0.1201	0.1208
	75	0.3107	0.3505	0.4058	0.3702	0.3557	0.1108	0.2005
	100	0.1652	0.4850	0.5152	0.4450	0.4308	0.0552	0.2993