Reliable Federated Multi-View Learning via Conflict-Aware Evidence Calibration

Daoyuan Li
School of Automation
Guangdong University of Technology
Guangzhou, China
daoyuanli0914@aliyun.com

Zuyuan Yang*
School of Automation
Guangdong University of Technology
Guangzhou, China
zuyuangdut@aliyun.com

Hao Yang
School of Automation
Guangdong University of Technology
Guangzhou, China
yanghao15@mails.gdut.edu.cn

Jiawen Kang
School of Automation
Guangdong University of Technology
Guangzhou, China
kavinkang@gdut.edu.cn

Abstract

Federated Multi-View Learning (FedMVL) enables multiple clients with heterogeneous data views to collaboratively train global models without sharing private data. Existing FedMVL algorithms often overlook the reliability of local predictions, as both uncertainty quantification and calibration remain challenging in privacy-preserving federated settings. Heterogeneous sensing devices yield varying view qualities, but view-isolated training further exacerbates inherent bias of local model, inducing unreliable outputs. Without proper calibration, these artifacts cause spurious cross-view conflicts, ultimately undermining the accuracy and reliability of the aggregated global model. To address these issues, we propose a Reliable Federated Multi-View Learning with evidence calibration (FedRMVL-CAL), a FedMVL framework with adaptive evidence fusion. The local calibration scheme harmonizes uncertainty across views, while the fusion mechanism adaptively reconciles conflicting opinions, enhancing robustness in global inference. Theoretical analysis and extensive experiments demonstrate that FedRMVL-CAL achieves superior accuracy and reliability compared to existing approaches, ensuring trustworthy global predictions across heterogeneous views. The code is available at https://github.com/robotpear/NeurIps-2026-FedRMVL-CAL.

1 Introduction

Driven by the proliferation of multi-view data [1–3] and growing privacy concerns, Federated Multi-View Learning (FedMVL) has emerged as a crucial paradigm. It enables clients to collaboratively train models without sharing raw data, effectively mitigating privacy risks. Crucially, by capturing both consistent and complementary information across diverse views, FedMVL constructs more comprehensive global predictions than single-view methods [4–6], driving its widespread application in domains like recommendation systems [7] and medical analysis [8].

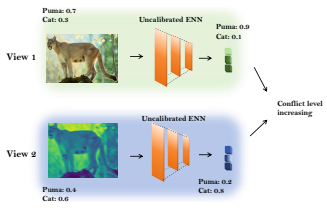


Figure 1: Conflicts amplified by uncalibrated predictions.

*Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

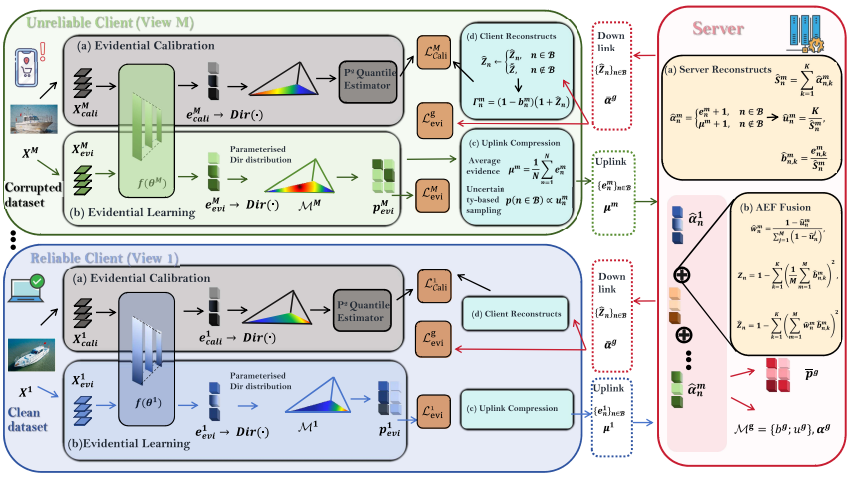


Figure 2: Overview of the FedRMVL-CAL framework. The tetrahedron represents the subjective opinion mass, and its base is the simplex of the Dirichlet distribution. In the local client training process, both parts (a) and (b) are performed simultaneously, using different sub-datasets.

The effectiveness of these methods hinges on the idealized assumption that the data quality of each client remains consistent and stable during training, and the resulting predictions are sufficiently reliable. Unfortunately, these assumptions are hard to satisfy in practical FedMVL scenarios. We identify that the cross-view conflicts undermining global prediction reliability comprise two components with fundamentally different roles: (1) **View-intrinsic heterogeneity**, which arises from genuine semantic differences across acquisition devices [9] and should be preserved as complementary information [10]. (2) **Uncalibrated conflict**, an artifact caused by uncalibrated local models producing overconfident or overconservative outputs [11]. The view-isolated training further exacerbates the bias. Treating these conflicts indiscriminately creates a dilemma: penalizing all divergence discards informative heterogeneity, while tolerating it allows calibration artifacts to corrupt global aggregation.

In this paper, we propose a novel FedMVL paradigm, Reliable Federated Multi-View Learning with evidence calibration (FedRMVL-CAL), as illustrated in Fig.2. At the client level, an evidential learning framework quantifies view-specific uncertainty and integrates the calibrated loss to correct model output bias without erasing inherent semantic heterogeneity. To minimize transmission overhead, an uncertainty-aware communication scheme selectively compresses and uploads informative evidence. At the central server, the Adaptive Evidence Fusion (AEF) strategy differentiates view-intrinsic heterogeneity from uncalibrated conflicts to achieve robust aggregation, which facilitates the broadcast of feedback to dynamically regulate local calibration pressure. This establishes a closed-loop feedback system that systematically eliminates persistent noise while faithfully preserving genuine complementary information. Specifically, the core contributions of this work are as follows:

- We identify and formalize the critical difference between view-intrinsic heterogeneity and uncalibrated-induced conflict in FedMVL, proposing the FedRMVL-CAL framework to treat these two sources asymmetrically through joint evidential learning and conflict-aware fusion.
- We present an evidence reliability calibration mechanism. By incorporating the broadcasted global conflict prior into the evidence reliability score, it adaptively regularizes local evidential learning, suppressing artificial overconfidence without destroying view-specific semantic representations.
- We propose the Adaptive Evidence Fusion (AEF) strategy, a conflict-robust aggregation rule with dual-modulation. It utilizes the uncertainty-aware deviation between uniform and weighted cross-view consensus to filter calibration-induced conflicts, ensuring reliable global Dirichlet inference under data heterogeneity.
- Implementation of an uncertainty-aware bidirectional communication scheme minimizes transmission overhead while maintaining fusion accuracy, supported by extensive theoretical analysis and empirical validation across diverse multi-view datasets.

2 Related works

Uncertainty estimation: Uncertainty estimation is crucial for reliable machine learning. Bayesian methods quantify uncertainty via posterior distributions over model parameters, as in Bayesian neural networks (BNNs)[12–14], while Monte Carlo dropout[15] provides a scalable approximation. Alternatively, the Dempster–Shafer (DS) theory represents uncertainty through basic probability assignments and supports multi-source evidence fusion[16–18]. Building on DS theory, Subjective Logic[19] models uncertainty using "opinions" consisting of belief and uncertainty masses, forming the foundation of reliable multi-view learning. Recent studies have extended these ideas to multi-view settings, where Han et al.[20] first incorporated uncertainty quantification, Liu et al.[21] proposed dynamic evidence decoupling, and Liang et al.[22] leveraged expert knowledge for improved reliability. However, existing multi-view uncertainty approaches often overlook conflicts introduced by calibration, resulting in suboptimal fusion performance—an issue that this work seeks to address.

Model calibration: Model calibration is essential for achieving reliable classification. Guo et al.[11] first revealed the severe overconfidence of modern deep networks and proposed temperature scaling as an effective post-hoc calibration method. Vaicenavicius et al.[23] further established a theoretical framework for evaluating calibration and introduced multidimensional reliability diagrams. A recent survey by Silva Filho et al.[24] systematized calibration concepts, metrics, and methods for both binary and multiclass tasks. Beyond post-hoc techniques, in-training strategies such as uncertainty-weighted optimization have been developed to improve calibration during learning (Lin et al.[25]). In federated settings, Peng et al.[26] proposed a federated calibration framework aligning local and global uncertainties across clients, while Meng et al.[27] enhanced global model reliability by calibrating cross-silo representations under heterogeneous data distributions.

Federated multi-view learning: Existing Federated Multi-View Learning (FedMVL) methods have demonstrated significant progress across various application domains and network architectures. Early efforts heavily focused on specific task formulations; for instance, Hu et al. proposed a fuzzy c-means approach for clustering [28], while Meng et al. designed a non-negative matrix factorization method for multi-label classification [29]. To address more complex data modalities, Chen et al. tackled the challenges posed by hybrid views [30], and Che et al. developed a secure framework tailored for private healthcare data analysis [8]. Additionally, to capture complex correlations, Li et al. leveraged hypergraphs to fuse information across views based on the self-expressive nature of the data [31].Recent work[32] has, for the first time, addressed reliability issues in the FedMVL environment. While recent SOTA methods have introduced evidential learning and reliability constraints to multi-view systems, they typically treat cross-view conflicts as a monolithic issue. Their inability to disentangle inherent view heterogeneity from artificial calibration noise severely constrains performance in complex federated settings. Our proposed framework directly addresses this limitation.

3 Methodology

3.1 Problem formulation

FedRMVL-CAL addresses heterogeneous data quality in FedMVL by accounting for view-wise uncertainty. Unreliable clients can degrade global reliability, while ignoring reliable views limits performance. The framework adaptively balances contributions to improve prediction reliability. We consider a K-class classification task in a vertical federated learning setting, where a multi-view feature space $\mathcal{X} = \{X^1, X^2, \ldots, X^M\}$ is partitioned across M clients. Each client m holds a distinct feature representation $X^m \in \mathbb{R}^{d_m}$, while all clients target a common label space $Y \in \mathbb{R}^K$.

3.2 Local Evidential Learning

Estimate uncertainty: To ensure reliable predictions under heterogeneous data quality, each client estimates view-specific uncertainty using Subjective Logic [19] and evidential deep learning [33]. For a K-class task, the m-th client formulates a subjective opinion $\mathcal{M}^m = \{b^m, u^m\}$, where $b^m \in \mathbb{R}^K$ and $u^m \in [0, 1]$ denote the belief masses and uncertainty, respectively. Given input X^m, the local Evidential Neural Network (ENN) with parameters θ^m outputs non-negative evidence $e^m = a(f(X^m; \theta^m))$, where $f(\cdot)$ is the penultimate layer and $a(\cdot)$ ensures non-negativity.

Subjective Logic links this evidence to a Dirichlet distribution p^m parameterized by $\alpha_k^m = e_k^m + 1$. The belief, uncertainty, and total Dirichlet strength S^m are derived as:

$$b_k^m = \frac{e_k^m}{S^m}, \quad u^m = \frac{K}{S^m}, \quad S^m = \sum_{k=1}^{K} \alpha_k^m \tag{1}$$

subject to the constraint $\sum_{k=1}^{K} b_k^m + u^m = 1$. Finally, the expected predictive probability for class k is obtained by $p_k^m = \frac{\alpha_k^m}{S^m}$. Unlike classical point estimates, ENN models capture uncertainty via a Dirichlet distribution $\mathrm{Dir}(p^m \mid \alpha^m)$ over class probabilities $p^m \in \mathbb{R}^K$, parameterized by $\alpha^m \in \mathbb{R}^K$ [33]. This serves as a conjugate prior over the categorical likelihood $p(y \mid p^m)$, with one-hot label $y \in \mathbb{R}^K$. The evidential cross-entropy loss is:

$$\mathcal{L}(\alpha^m) = -\sum_{k=1}^{K} y_k \mathbb{E}_{p_k^m \sim \mathrm{Dir}(\alpha^m)}[\log p_k^m] = \sum_{k=1}^{K} y_k [\psi(S^m) - \psi(\alpha_k^m)] \tag{2}$$

where $\psi(\cdot)$ is the digamma function. To explicitly suppress evidence for incorrect classes, we introduce a Kullback-Leibler (KL) divergence term and the total evidential loss is then formulated as:

$$\mathcal{L}_{\mathrm{evi}}(\alpha^m) = \mathcal{L}(\alpha^m) + \lambda_k \mathrm{KL}[\mathrm{Dir}(p^m \mid \tilde{\alpha}^m) \| \mathrm{Dir}(p^m \mid 1)] \tag{3}$$

where the adjusted parameter $\tilde{\alpha}_k^m = y_k + (1 - y_k)\alpha_k^m$ prevents penalizing the ground-truth class, and λ_k is an annealing trade-off coefficient.

Conflict decoupling and calibration: To illustrate why view-isolated training can cause additional conflicts in FL, consider a simple case with $K = 3$ classes where the same sample is independently observed by two clients. Unlike the centralized paradigm, federated clients lack visibility into the representations and uncertainties of other views. As a result, view $m = 1, 2$ outputs $b^1 = (0.82, 0.10, 0.08)$ and $b^2 = (0.12, 0.85, 0.03)$ with uncertainties $u^1 = 0.05$ and $u^2 = 0.06$. View 1 is highly confident in class 1, while view 2 is equally confident in class 2, both showing very low uncertainty. Assuming the true label is class 2, view 1 is severely overconfident despite making an incorrect prediction. Although each view seems reliable locally, their predictions are globally contradictory. Aggregating such uncalibrated evidence tends to induce spurious cross-view conflicts, highlighting the necessity of calibration to align local confidence and achieve coherent multi-view fusion. To mitigate this, we introduce a calibration strategy via variant pinball loss. The local dataset X^m is split into X_{evi}^m for ENN training and X_{cali}^m for calibration (e.g., 10% of the training samples). Given an instance from X_{cali}^m with ground-truth class y, its evidence reliability score is formulated as:

$$\Gamma^m = (1 - b_y^m)(1 + \hat{Z}). \tag{4}$$

Here, b_y^m is the belief mass on the true class y, and $\hat{Z} \in [0, 1)$ is the calibration-weighted conflict prior from the preceding server round, which gauges the deviation of local evidence from a reliable prediction. Consequently, a lower score indicates higher reliability: $\Gamma^m \to 0$ corresponds to highly confident and accurate evidence, whereas higher scores (e.g., $\Gamma^m \geq 1$) signify unreliable evidence, with the theoretical maximum approaching 2 under extreme historical conflict.

The dynamic threshold c tracks the τ-quantile ($\tau = 0.5$ in experiments) of the evidence reliability scores online via the P^2 algorithm [34] (details in Appendix A.1). To calibrate the predictions, we employ an asymmetric variant pinball loss:

$$\mathcal{L}_{\mathrm{cali}} = (\Gamma^m - c)\Big(\tau \mathbb{I}[\hat{y} = y, \Gamma^m \geq c] - (1 - \tau)\mathbb{I}[\Gamma^m < c]\Big), \tag{5}$$

where $\hat{y} = \arg\max b_k^m$ and $\mathbb{I}[\cdot]$ is the indicator function. The formulation balances model confidence in a principled manner. When $\Gamma^m < c$, the local ENN assigns excessive belief to the true class relative to c, indicating overconfidence induced by view-isolated training; the loss penalises such instances by driving Γ^m toward c from below, which reduces b_y^m and increases local uncertainty u^m. Conversely, when $\Gamma^m \geq c$ and the prediction is correct, the model assigns insufficient belief to the true class; the loss drives b_y^m upward to align local confidence with empirical accuracy. Incorrect predictions are excluded from this upward correction to avoid reinforcing erroneous evidence.

3.3 Server Opinion Fusion

To reconcile conflicting multi-view evidence, we propose an Adaptive Evidence Fusion (AEF) framework at the serve side. AEF leverages client-specific calibrated uncertainty to adaptively down-weight unreliable sources, mitigating conflicts while preserving informative contributions to ensure stability and fidelity in the fused evidence.

Definition 1 (Adaptive Evidence Fusion Rule)**.** Let the subjective opinion masses over K classes from client m be denoted by $\mathcal{M}^m = \{b^m, u^m\}$, with total strength $S^m = \sum_{k=1}^{K} \alpha_k^m$. The uncertainty-aware view weight w^m and the two conflict measures are defined as:

$$w^m = \frac{1-u^m}{\sum_{j=1}^{M}(1-u^j)}, \qquad Z = 1 - \sum_{k=1}^{K}\left(\frac{1}{M}\sum_{m=1}^{M} b_k^m\right)^2, \qquad \hat{Z} = 1 - \sum_{k=1}^{K}\left(\sum_{m=1}^{M} w^m b_k^m\right)^2 \tag{6}$$

where Z measures cross-view conflict under uniform aggregation and the $\hat{Z}$ is the calibration-weighted conflict. The fused Dirichlet parameter for class k with dual conflict modulation is:

$$\alpha_k^g = \frac{1-Z}{1+\hat{Z}} \sum_{m=1}^{M} w^m \alpha_k^m. \tag{7}$$

The fused total strength, uncertainty, and global predictive probability for class k are derived inline as $S^g = \sum_{k=1}^{K} \alpha_k^g$, $u^g = \frac{K}{S^g}$, and $p_k^g = \frac{\alpha_k^g}{S^g}$. The numerator $(1 - Z)$ penalizes conflict under uniform aggregation. It provides a baseline measure of inherent data heterogeneity. Conversely, the denominator $(1 + \hat{Z})$ imposes an adaptive penalty based on the calibration-weighted conflict. During standard optimization, local calibration improves and confident views converge. The progressively reduces the fusion penalty and facilitates high-confidence consensus. Conversely, severe view-isolation or data corruption can cause strict contradictions among highly confident views. Under these extreme conditions, the denominator amplifies the conservatism of the global update beyond the baseline. It significantly discounts the aggregated evidence to prevent highly confident misclassifications during severe cross-view divergence.

3.4 Evidence Compression and Communication

Transmitting the full evidence matrix $\alpha^m \in \mathbb{R}^{N\times K}$ incurs a prohibitive $\mathcal{O}(NK)$ communication cost per client per round, which scales poorly for large datasets. To alleviate this, we propose a bidirectional uncertainty-aware communication scheme that prioritizes informative samples while summarizing the rest.

Uplink Compression and Server Fusion: Instead of full transmission, clients selectively upload evidence based on local uncertainty. Since low-uncertainty samples reflect high confidence with concentrated evidence distributions, approximating them with a mean summary introduces minimal perturbation to the global fusion. Specifically, each client calculates the mean evidence $\mu^m = \frac{1}{N}\sum_{n=1}^{N} e_n^m \in \mathbb{R}^K$ and selects a subset $\mathcal{B}$ of size $\beta = \min(\beta_{\max}, N)$. The sampling probability is directly proportional to the local uncertainty: $p(n \in \mathcal{B}) \propto u_n^m = \frac{K}{S_n^m}$. The client then transmits only the exact evidence $\{e_n^m\}_{n\in\mathcal{B}}$ and the mean summary μ^m.

Upon receiving this condensed information, the server reconstructs the local Dirichlet parameters as:

$$\hat{\alpha}_n^m = \begin{cases} e_n^m + 1, & n \in \mathcal{B}, \\ \mu^m + 1, & n \notin \mathcal{B}, \end{cases} \tag{8}$$

and performs AEF using $\{\hat{\alpha}^m\}_{m=1}^{M}$ to yield the global parameters α_n^g, the per-sample intrinsic conflict Z_n, and the calibration-weighted residual conflict $\hat{Z}_n$, along with their global mean $\bar{\hat{Z}} = \frac{1}{N}\sum_{n=1}^{N} \hat{Z}_n$.

Cold-start Warm-up: Since uncertainty rankings may be unstable during early training, clients transmit the full evidence matrix for the first five communication rounds. Afterwards, they switch to the uncertainty-proportional sampling rule described above. This transition avoids premature reliance on unstable uncertainty estimates, with full-transmission overhead restricted to the warm-up phase

Downlink Broadcast and Client Reconstruction: To compress the downlink payload, the server broadcasts the exact $\{\hat{Z}_n\}_{n\in\mathcal{B}}$, the global mean $\bar{\hat{Z}}$, and the global mean Dirichlet parameter $\bar{\alpha}^g$. Each client stores the conflict as the prior for the subsequent round:

$$\hat{Z}_n \leftarrow \begin{cases} \hat{Z}_n, & n \in \mathcal{B}, \\ \bar{\hat{Z}}, & n \notin \mathcal{B}. \end{cases} \tag{9}$$

The stored prior $\hat{Z}_n$ then enters the AEF denominator in the next round and simultaneously regulates the per-sample calibration strength via the evidence reliability score $\Gamma^m_{n,y_n} = (1 - b^m_{n,y_n})(1 + \hat{Z}_n)$. Since $\hat{Z}_n$ tracks calibration-weighted conflict, its decay across rounds directly reflects calibration improvement, ensuring that fusion conservatism and calibration pressure are released in tandem as views become more consistent and reliable. Simultaneously, the broadcasted $\bar{\alpha}^g$ serves as a lightweight target for the global alignment loss $\mathcal{L}_{\text{evi}}(\alpha^g)$. Ultimately, the total per-round communication cost is reduced from $\mathcal{O}(MNK)$ to $\mathcal{O}(M\beta K + MK + \beta)$, achieving an approximately N/β-fold compression ratio when $\beta \ll N$.

3.5 General objectives and theoretical analyses

To jointly model cross-view consistency and view-specific complementarity, the client loss is defined as

$$\mathcal{L}^m_{\text{client}} = \mathcal{L}_{\text{evi}}(\alpha^m) + \mathcal{L}_{\text{cali}} + \lambda_g \mathcal{L}_{\text{evi}}(\alpha^g), \tag{10}$$

where λ_g balances local and global evidence alignment. The pseudocode for the algorithm is provided in AppendixA.2. The following propositions summarize the key theoretical properties of our framework.

Proposition 1 (Monotonic Uncertainty Regulation)**.** The u^g increases with the conflict prior $\hat{Z}$, and decreases as reliable local evidence S^m increases.

Proof. Since $u^g = \frac{K}{S^g}$, analyzing the monotonicity of u^g is strictly equivalent to analyzing the inverse monotonicity of S^g.

For the historical conflict $\hat{Z}$, it solely regulates the denominator of S^g. The partial derivative is strictly negative:

$$\frac{\partial S^g}{\partial \hat{Z}} = -\frac{1 - Z}{(1 + \hat{Z})^2} \sum_{m=1}^{M} w^m S^m < 0 \implies \frac{\partial u^g}{\partial \hat{Z}} > 0.$$

Thus, global uncertainty u^g strictly inflates with accumulated historical conflict.

For reliable local evidence S^m, an increase in S^m strictly decreases local uncertainty $u^m = \frac{K}{S^m}$, which monotonically increases its view weight w^m. For a reliable client, this increased weight concentrates the consensus, strictly lowering the instantaneous geometric conflict Z. Consequently, both the scaling factor $(1 - Z)$ and the weighted evidence $\sum_{m=1}^{M} w^m S^m$ strictly increase. The guarantees S^g strictly increases, thereby strictly reducing u^g. □

Proposition 2 (Transmission Approximation Bound)**.** Let α^g and $\hat{\alpha}^g$ denote the raw and server-reconstructed global Dirichlet parameters under AEF, respectively. For any unobserved instance, the server substitutes $\hat{\alpha}^m_k = \mu^m_k + 1$ for the unobserved evidence. The resulting reconstruction error satisfies:

$$\left\|\alpha^g - \hat{\alpha}^g\right\|_1 \le \frac{1 - Z}{1 + \hat{Z}} \sum_{m=1}^{M} w^m \left\|e^m - \mu^m\right\|_1$$

and vanishes as all instances are transmitted or as local evidence concentrates ($e^m \to \mu^m$).

Proof. Since the AEF formula is linear in $\{\alpha^m_k\}$, the error in the global parameter for class k is:

$$\alpha^g_k - \hat{\alpha}^g_k = \frac{1 - Z}{1 + \hat{Z}} \sum_{m=1}^{M} w^m \left(\alpha^m_k - \hat{\alpha}^m_k\right).$$

For an unobserved instance, the substitution gives $\alpha_k^m - \hat{\alpha}_k^m = e_k^m - \mu_k^m$, where $e^m = \alpha^m - 1$ denotes the raw evidence vector. Summing over k and applying the triangle inequality:

$$\left\|\alpha^g - \hat{\alpha}^g\right\|_1 \leq \frac{1-Z}{1+\hat{Z}} \sum_{m=1}^{M} w^m \left\|e^m - \mu^m\right\|_1.$$

Unobserved instances are selected to have low u^m, implying large $S^m \gg K$ and evidence e^m concentrated near μ^m, so $\|e^m - \mu^m\|_1$ is small by the sampling design. The bound collapses to zero identically when all instances are transmitted. □

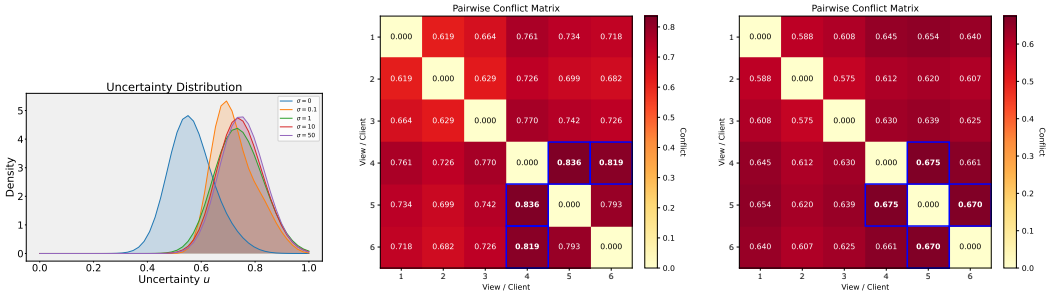
Proposition 3 (Calibration–Fusion Synergy)**.** Suppose a client m applies a calibration increment $\delta > 0$ to the target belief b_y^m of a correctly classified instance, with non-target beliefs decaying proportionally to maintain the simplex constraint. Then, both the Z and the $\hat{Z}$ decrease, leading to a monotonic increase in the fused S^g and an attenuation of the global uncertainty u^g.

Proof. Let $\bar{b}_k = \frac{1}{M}\sum_{j=1}^{M} b_k^j$. An increment δ on the target belief b_y^m yields $\frac{\partial \bar{b}_y}{\partial \delta} = \frac{1}{M}$ and proportionally decays non-target beliefs as $\frac{\partial \bar{b}_{k\neq y}}{\partial \delta} = -\frac{b_k^m}{M(1-b_y^m)}$. The variation of Z is derived as:

$$\frac{\partial Z}{\partial \delta} = -2\sum_{k=1}^{K} \bar{b}_k \frac{\partial \bar{b}_k}{\partial \delta} = -\frac{2}{M}\left(\bar{b}_y - \sum_{k\neq y} \bar{b}_k \frac{b_k^m}{1-b_y^m}\right) < 0. \tag{11}$$

The inequality holds since $\bar{b}_y$ strictly dominates the convex combination of non-target beliefs under correct classification. Analogously, this evidence alignment ensures $\frac{\partial \hat{Z}}{\partial \delta} < 0$. Because the AEF rule yields $\frac{\partial S^g}{\partial Z}, \frac{\partial S^g}{\partial \hat{Z}} < 0$, the dual reduction of conflicts strictly increases S^g, monotonically attenuating the global uncertainty $u^g = \frac{K}{S^g}$. □

The detailed proofs provided in Appendix B.

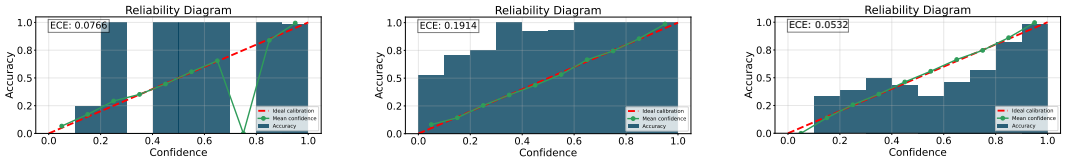


(a) Density of uncertainty in Cal-20 (b) Uncalibrated (c) With calibration

Figure 3: Attributes of reliable learning

Table 1: Classification performance (mean result of 5 experiments±std%) in ACC

Categories	Method	HW	Cal-20	AwA	PIE	MNIST2	NUSWIDE
Centralized	DUA-Nets(2020)	96.1±0.3	86.2±0.2	28.2±0.0	90.6±0.1	97.2±0.0	27.7±0.0
	TMC(2023)	98.5±0.9	92.2±0.5	35.1±1.2	96.5±1.2	98.3±0.9	33.4±0.6
	CCML(2024)	98.8±0.1	92.9±0.1	38.5±0.1	94.6±1.8	99.0±0.0	38.9±0.1
	ECML(2024)	99.1±0.0	93.9±0.3	39.2±0.8	97.5±0.2	99.2±0.1	33.8±0.5
	TMCEK(2025)	98.8±0.2	94.1±0.2	37.3±0.2	97.1±1.8	99.0±0.0	35.6±0.2
	TEF(2025)	99.2±0.0	93.4±0.6	41.2±1.1	98.6±0.0	99.5±0.0	38.7±0.0
FL	FedAvg(2016)	94.2±0.0	84.1±0.0	23.1±0.2	86.3±0.0	98.2±0.0	27.8±0.0
	FedProx(2018)	96.0±0.0	87.8±0.0	27.4±0.2	89.1±0.0	98.6±0.0	28.3±0.0
	β-preBayes(2024)	95.1±0.2	86.1±0.2	29.3±1.1	93.1±2.0	98.1±0.5	30.1±0.8
	RHFL+(2025)	96.4±0.0	88.7±0.1	31.8±0.4	92.2±0.0	98.9±0.1	31.9±0.8
FedMVL	FedRMVL(2026)	98.7±0.0	91.6±1.3	36.1±0.2	94.5±0.1	98.5±0.0	34.1±0.6
	FedRMVL-CAL(Ours)	99.1±0.0	93.1±0.2	40.1±0.2	97.5±0.1	99.5±0.0	37.6±0.1



(a) CCML (b) TMCEK (c) FedRMVL-CAL

Figure 4: Comparison of prediction reliability in Cal-20

Table 2: Ablation Study on Six Datasets

Method	Cal-20		HW		Awa		MNIST2		PIE		NUSWIDE	
	ACC↑	ECE↓	ACC↑	ECE↓	ACC↑	ECE↓	ACC↑	ECE↓	ACC↑	ECE↓	ACC↑	ECE↓
FedRMVL-CAL	0.9308	0.0532	0.9910	0.0037	0.4008	0.2753	0.9946	0.0020	0.9751	0.0140	0.3757	0.2552
w/o $\mathcal{L}_{\text{cali}}$	0.9203	0.0626	0.9875	0.0110	0.3917	0.4213	0.9930	0.0040	0.9708	0.0157	0.3797	0.3213
w/o $\mathcal{L}_{\text{KL}}$	0.8637	0.0847	0.9850	0.0165	0.3693	0.5002	0.9921	0.0045	0.9692	0.0172	0.3612	0.3165
w/o $\mathcal{L}_{\text{cali}}$ + $\mathcal{L}_{\text{KL}}$	0.8532	0.0828	0.9725	0.0112	0.3601	0.4828	0.9807	0.0105	0.9569	0.0212	0.3538	0.3986

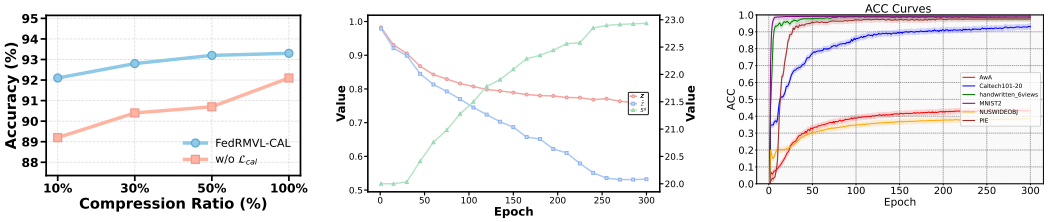
4 Experiments

4.1 Experiment settings

We conduct comprehensive experiments across six diverse datasets: **HW**, **Cal-20**, **AwA**, **PIE**, **MNIST2**, and **NUSWIDE**. We establish comparisons against a broad spectrum of baselines, including centralized reliable multi-view methods (DUA-Nets [35], TMC [20], CCML [21], ECML [36], TMCEK [22], TEF [37]), standard single-view FL frameworks (FedAvg [38], FedProx [39], β-preBayes [40], RHFL+ [41]), and the recent FedMVL approach (FedRMVL [32]). Under the federated setting, each client holds data from a specific view. Due to the dimension difference among these view datasets, communication is strictly limited to network parameters that share isomorphic structures(details in Appendix C.5). The proposed FedRMVL-CAL model is instantiated as a three-layer MLP trained via the Adam optimizer. A consistent training budget of 300 iterations/communication rounds is applied to all methods to guarantee sufficient and stable convergence. The hardware environment comprises an Intel i9-14900KF CPU and an NVIDIA RTX 4090 GPU. Further details regarding the experimental setup and extended results can be found in Appendix C.

4.2 Classification results

As shown in Tab. 1, FedRMVL-CAL consistently outperforms all evaluated single-view FL baselines and the FedRMVL across the six datasets. Notably, it achieves a substantial performance leap on challenging datasets such as AwA, while effectively narrowing the performance gap with state-of-the-art centralized multi-view methods and attaining on-par accuracy on datasets like HW and MNIST2. This empirical advantage stems from our framework, which first quantifies view-specific data uncertainty at the client level and subsequently harmonizes these local estimates via the e calibration mechanism. By principledly handling uncertainty, the AEF at the server side can robustly integrate reliable information while reconciling cross-view conflicts, ultimately yielding more accurate and dependable global predictions.



(a) Effect of compression ratio (b) Conflict trends (c) ACC variation curves

Figure 5: Training dynamic

4.3 Attributes of reliable learning

Uncertainty estimation: To verify the performance of the proposed algorithm in quantifying data uncertainty, we add noise to the test data. Fig.3a illustrates that the proposed method effectively

captures data uncertainty and reflects its variation under different noise intensities. When no noise is added, the uncertainty values are concentrated at relatively low levels, indicating confident predictions on clean data. As the noise level increases ($\sigma = 0.1, 1, 10, 50$), the uncertainty distribution gradually shifts toward higher values and becomes more dispersed, suggesting that the model can sensitively perceive and quantify the degradation in data quality. The consistent and monotonic trend across noise scales confirms that the evidential learning framework with calibration mechanism produces uncertainty estimates that faithfully represent the reliability of the participation views (clients).

Impact of evidence reliability calibration: The pairwise matrices in Fig.3b and 3c evaluate the calibration-weighted conflict (Pairwise $\hat{Z}$), mirroring the AEF rule in server side: $1 - \sum_k (w^i b_k^i + w^j b_k^j)^2$. This metric vividly captures the synergy between local calibration and global fusion. Before calibration, view-isolated overconfidence forces nearly equal fusion weights, causing severe cross-view collisions. After calibration, extreme overconfident outputs are attenuated into higher uncertainty, allowing the AEF mechanism to adaptively down-weight unreliable views. Consequently, the conflict matrix becomes substantially more homogeneous. This process resolves spurious artifacts without collapsing genuine cross-view differences, ensuring coherent and robust multi-view aggregation.

Prediction reliability analysis: Fig. 4 compares the prediction reliability of various multi-view methods. As shown in the reliability diagrams, FedRMVL-CAL most closely aligns with the perfect calibration diagonal and achieves the lowest Expected Calibration Error (ECE). Furthermore, the confidence histograms indicate that our model does not artificially lower overall confidence; instead, it maintains prevalent high-confidence predictions that accurately match empirical accuracies. Taken together, these results demonstrate that incorporating evidence reliability calibration into the FedMVL pipeline leads to highly trustworthy predictions.

4.4 Ablation experiments

Tab.2 presents the ablation results, evaluating the impact of the two regularization terms. The analysis highlights the impact of regularization on predictive reliability in FedRMVL-CAL. The ablation results demonstrate that removing either the $\mathcal{L}_{\text{cali}}$ or the $\mathcal{L}_{\text{KL}}$ consistently degrades performance across all datasets, reflected in lower accuracy and higher calibration error. The complete FedRMVL-CAL framework achieves the most favorable balance, yielding the highest accuracy together with the lowest ECE, which indicates that both components contribute complementary benefits. Variants lacking one or both terms converge to weaker solutions, with particularly pronounced declines on more challenging datasets, highlighting the necessity of jointly modeling calibration and regularization to ensure stable and reliable learning.

4.5 Training dynamic analysis

Communication compression analysis: Fig.5a illustrates the effect of varying transmission compression ratios on classification accuracy. Under extremely low transmission ratios, the complete FedRMVL-CAL model maintains a high and stable accuracy, whereas the model variant lacking the calibration module exhibits significant performance degradation and fluctuation. This demonstrates that the uncertainty-aware communication scheme precisely filters and transmits the most informative local evidence, effectively ensuring the robustness of global aggregation performance while substantially reducing communication overhead.

Conflict trends analysis: Fig.5b reflects the dynamic behavioral differences between the global Z and the global calibration-weighted conflict $\hat{Z}$ during training. As iterations proceed, both metrics exhibit a significant decline and eventually converge. The reduction in Z indicates that local calibration operations effectively align cross-view evidence and mitigate prediction discrepancies in the current round. Concurrently, the continuous decay of $\hat{Z}$ directly quantifies the improvement in global calibration quality. Their synergistic downward trajectory verifies that the closed-loop feedback system, upon eliminating artificial conflicts induced by uncalibrated outputs, can dynamically and simultaneously release fusion conservatism and calibration pressure, thereby preserving genuine complementary view information.

Convergence analysis: Fig.5c shows the accuracy curves of FedRMVL-CAL on six multi-view datasets.The curves exhibit a rapid increase during the initial training epochs, reflecting efficient early-stage learning across the evaluated datasets. After this sharp rise, the trajectories gradually

flatten, indicating that the model approaches a stable performance plateau. The convergence behavior is consistent and smooth, with no significant oscillations, suggesting that FedRMVL-CAL achieves reliable optimization and maintains stable performance once the training process matures.

5 Conclusion and discussion

We presented FedRMVL-CAL, a novel paradigm that uniquely disentangles view-intrinsic heterogeneity from calibration-induced artifacts. By integrating calibration loss with adaptive evidence fusion, FedRMVL-CAL establishes a dynamic feedback loop that harmonizes cross-view uncertainty while preserving essential complementary semantics. Theoretical and empirical results demonstrate that our framework achieves SOTA accuracy and superior reliability with high communication efficiency. Overall, FedRMVL-CAL provides a robust, theoretically grounded solution for reliable multi-view classification in federated environments. Limitations are discussed in Appendix D.

6 Acknowledgements

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515010589, in part by the National Natural Science Foundation of China under Grants 62273106, 12326604, and 82305172. The authors would like to extend special thanks to Dr. Zhihong Li for his assistance with the paper.

References

[1] IDC, "Worldwide idc global datasphere forecast, 2024–2028: Ai everywhere, but upsurge in data will take time." `https://www.idc.com/research/viewtoc.jsp?containerId=US52076424`. Accessed: May 2024.

[2] X. Chen, J. Xu, Y. Ren, X. Pu, C. Zhu, X. Zhu, Z. Hao, and L. He, "Federated deep multi-view clustering with global self-supervision," in *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, (New York, NY, USA), p. 3498–3506, Association for Computing Machinery, 2023.

[3] H. Chen, Y. Zhang, D. Krompass, J. Gu, and V. Tresp, "Feddat: An approach for foundation model finetuning in multi-modal heterogeneous federated learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 11285–11293, Mar. 2024.

[4] S. Huang, W. Shi, Z. Xu, I. W. Tsang, and J. Lv, "Efficient federated multi-view learning," *Pattern Recognition*, vol. 131, p. 108817, 2022.

[5] R. Patole, N. Singh, M. Adhikari, and A. K. Singh, "Multi-view ensemble federated learning for efficient prediction of consumer electronics applications in fog networks," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 4597–4604, 2024.

[6] Y. Yan, S. Wang, F. Sun, and X. Tong, "Personalized federated learning with multi-view geometry structure," *IEEE Internet of Things Journal*, pp. 1–1, 2024.

[7] A. Flanagan, W. Oyomno, A. Grigorievskiy, K. E. Tan, S. A. Khan, and M. Ammad-Ud-Din, "Federated multi-view matrix factorization for personalized recommendations," in *Joint European conference on machine learning and knowledge discovery in databases*, pp. 324–347, Springer, 2020.

[8] S. Che, Z. Kong, H. Peng, L. Sun, A. Leow, Y. Chen, and L. He, "Federated multi-view learning for private medical data integration and analysis," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 4, pp. 1–23, 2022.

[9] Q. Wang, C. Yin, H. Song, T. Shen, and Y. Gu, "Utfnet: Uncertainty-guided trustworthy fusion network for rgb-thermal semantic segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[10] R. Halabi, R. Selvarajan, Z. Lin, C. Herd, X. Li, J. Kabrit, M. Tummalacherla, E. Chaibub Neto, and A. Pratap, "Comparative assessment of multimodal sensor data quality collected using android and ios smartphones in real-world settings," *Sensors*, vol. 24, no. 19, 2024.

[11] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, PMLR, 06–11 Aug 2017.

[12] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," *Advances in neural information processing systems*, vol. 30, 2017.

[13] M. Clyde and E. I. George, "Model Uncertainty," *Statistical Science*, vol. 19, no. 1, pp. 81 – 94, 2004.

[14] M. Breuer and H. H. Schütt, "Accounting for uncertainty: an application of bayesian methods to accruals models," *Review of Accounting Studies*, vol. 28, no. 2, pp. 726–768, 2023.

[15] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, pp. 1050–1059, PMLR, 2016.

[16] A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," in *Classic works of the Dempster-Shafer theory of belief functions*, pp. 57–72, Springer, 2008.

[17] A. P. Dempster, "A generalization of bayesian inference," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 30, no. 2, pp. 205–232, 1968.

[18] K. Sentz and S. Ferson, "Combination of evidence in dempster-shafer theory," *SANDIA REPORT*, 2002.

[19] A. Jøsang, R. Ismail, and C. Boyd, "A survey of trust and reputation systems for online service provision," *Decision support systems*, vol. 43, no. 2, pp. 618–644, 2007.

[20] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification with dynamic evidential fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 2551–2566, 2023.

[21] Y. Liu, L. Liu, C. Xu, X. Song, Z. Guan, and W. Zhao, "Dynamic evidence decoupling for trusted multi-view learning," in *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, (New York, NY, USA), p. 7269–7277, Association for Computing Machinery, 2024.

[22] X. Liang, S. Wang, Y. Qian, Q. Guo, L. Du, B. Jiang, T. Luo, and F. Li, "Trusted multi-view classification with expert knowledge constraints," in *Proceedings of the 42nd International Conference on Machine Learning* (A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, and J. Zhu, eds.), vol. 267 of *Proceedings of Machine Learning Research*, pp. 37409–37426, PMLR, 13–19 Jul 2025.

[23] J. Vaicenavicius, D. Widmann, C. Andersson, F. Lindsten, J. Roll, and T. Schön, "Evaluating model calibration in classification," in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* (K. Chaudhuri and M. Sugiyama, eds.), vol. 89 of *Proceedings of Machine Learning Research*, pp. 3459–3467, PMLR, 16–18 Apr 2019.

[24] T. Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach, "Classifier calibration: a survey on how to assess and improve predicted class probabilities," *Machine Learning*, vol. 112, pp. 3211–3260, 2023.

[25] J. Lin, L. Tao, M. Dong, and C. Xu, "Uncertainty weighted gradients for model calibration," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15497–15507, 2025.

[26] H. Peng, H. Yu, X. Tang, and X. Li, "FedCal: Achieving local and global calibration in federated learning via aggregated parameterized scaler," in *Proceedings of the 41st International Conference on Machine Learning* (R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, eds.), vol. 235 of *Proceedings of Machine Learning Research*, pp. 40331–40346, PMLR, 21–27 Jul 2024.

[27] L. Meng, Z. Qi, L. Wu, X. Du, Z. Li, L. Cui, and X. Meng, "Improving global generalization and local personalization for federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 1, pp. 76–87, 2025.

[28] X. Hu, J. Qin, Y. Shen, W. Pedrycz, X. Liu, and J. Liu, "An efficient federated multiview fuzzy c-means clustering method," *IEEE Transactions on Fuzzy Systems*, vol. 32, no. 4, pp. 1886–1899, 2023.

[29] H. Meng, Y. Deng, Q. Zhong, Y. Wang, Z. Yang, and G. Lyu, "Federated multi-view multi-label classification," *IEEE Transactions on Big Data*, pp. 1–13, 2024.

[30] X. Chen, Y. Ren, J. Xu, F. Lin, X. Pu, and Y. Yang, "Bridging gaps: Federated multi-view clustering in heterogeneous hybrid views," *Advances in Neural Information Processing Systems*, vol. 37, pp. 37020–37049, 2024.

[31] D. Li, Z. Yang, and S. Xie, "Fedmsgl: A self-expressive hypergraph based federated multi-view learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 18244–18252, 2025.

[32] D. Li, Z. Yang, J. Kang, Z. Xiong, D. Niyato, and S. Xie, "Reliable federated multi-view learning for heterogeneous information fusion in mobile edge computing," *IEEE Transactions on Mobile Computing*, vol. 25, no. 5, pp. 7062–7076, 2026.

[33] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Advances in Neural Information Processing Systems*, vol. 31, Curran Associates, Inc., 2018.

[34] R. Jain and I. Chlamtac, "The p2 algorithm for dynamic calculation of quantiles and histograms without storing observations," *Commun. ACM*, vol. 28, p. 1076–1085, Oct. 1985.

[35] Y. Geng, Z. Han, C. Zhang, and Q. Hu, "Uncertainty-aware multi-view representation learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 7545–7553, 2021.

[36] C. Xu, J. Si, Z. Guan, W. Zhao, Y. Wu, and X. Gao, "Reliable conflictive multi-view learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, pp. 16129–16137, 2024.

[37] X. Liang, P. Fu, Y. Qian, Q. Guo, and G. Liu, "Trusted multi-view classification via evolutionary multi-view fusion," in *International Conference on Learning Representations*, 2025.

[38] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*, pp. 1273–1282, PMLR, 2017.

[39] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proceedings of Machine Learning and Systems*, vol. 2, pp. 429–450, 2020.

[40] M. Hasan, G. Zhang, K. Guo, X. Chen, and P. Poupart, "Calibrated one round federated learning with bayesian inference in the predictive space," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 12313–12321, Mar. 2024.

[41] X. Fang and M. Ye, "Noise-robust federated learning with model heterogeneous clients," *IEEE Transactions on Mobile Computing*, vol. 24, no. 5, pp. 4053–4071, 2025.

A Additional technical details

A.1 Quantile estimation based on the p2 Algorithm

The P2 algorithm (Piecewise-Parabolic Quantile Estimation), introduced by Jain and Chlamtac in 1985 [33], constitutes a single-pass, constant-memory method for dynamic quantile estimation that maintains only five markers to track data distributions without storing all observations. For a target quantile p (where in this calibration framework $p = \tau = 1 - \eta$), the algorithm preserves five markers with distinct statistical interpretations. The first marker q_0 represents the current minimum observation, while q_4 denotes the maximum. The intermediate markers q_1, q_2, and q_3 track the $p/2$, p, and $(1 + p)/2$ quantiles respectively. For instance, when $p = 0.9$, these correspond to the 0.45, 0.9, and 0.95 quantiles. Each marker maintains three attributes: its height q_i (the actual numerical value), its position n_i (the count of observations less than or equal to q_i), and its desired position n'_i (the ideal location in the sorted sequence). The desired positions evolve according to $n'_0 = 0$, $n'_1 = (n-1) \cdot p/2$, $n'_2 = (n-1) \cdot p$, $n'_3 = (n-1) \cdot (1+p)/2$, and $n'_4 = n - 1$, where n represents the total number of processed observations.

The algorithmic procedure comprises two fundamental phases. During initialization, the first five observations are collected and sorted to establish the initial marker heights q_0 through q_4, with initial positions set as $n_i = i$ for $i = 0, 1, 2, 3, 4$. The online update phase executes for each subsequent observation x through a sequence of operations. First, the algorithm identifies the insertion interval k such that $q_k \leq x < q_{k+1}$. Second, for all markers with index $i \geq k+1$, the positions are incremented virtually as $n_i \leftarrow n_i + 1$. Third, if x constitutes a new extremum, the minimum or maximum marker is updated accordingly. Fourth, the desired positions are incremented by constant values dn'_i specific to each marker, where $dn'_2 = p$ for the target quantile marker. Fifth, for non-extreme markers q_1, q_2, and q_3, the algorithm examines the deviation $d = n'_i - n_i$ between desired and actual positions.

When $|d| \geq 1$ and marker movement does not violate ordering constraints, the P2 algorithm employs piecewise-parabolic interpolation to adjust marker heights. The adjustment formula is given by:

$$q'_i = q_i + \frac{d}{n_{i+1} - n_{i-1}} \left[(n_i - n_{i-1} + d) \frac{q_{i+1} - q_i}{n_{i+1} - n_i} + (n_{i+1} - n_i - d) \frac{q_i - q_{i-1}}{n_i - n_{i-1}} \right] \tag{12}$$

If the parabolic interpolation yields a value outside the permissible range, the algorithm falls back to linear interpolation:

$$q'_i = q_i + d \cdot \frac{q_{i+d} - q_i}{n_{i+d} - n_i} \tag{13}$$

The adoption of P2 algorithm in our quantile calibration framework addresses three critical challenges inherent to the proposed methodology. The calibration set X^m_{cali} may contain numerous samples, making storage of all evidence reliability scores Γ^m_n impractical. Furthermore, model parameters evolve during training, causing the distribution of Γ^m_n to drift over time. Additionally, the calibration process demands real-time computation of threshold c without awaiting complete processing of all samples. P2 resolves these issues through its constant memory footprint of five markers, single-pass processing capability, and inherent adaptation to distribution shifts.

The mapping between P2 algorithmic components and the calibration framework proceeds as follows. The data stream corresponds to the sequence of evidence reliability scores $\{\Gamma^m_1, \Gamma^m_2, \ldots, \Gamma^m_N\}$. The target quantile level is $p = \tau = 1 - \eta$, with the dynamic threshold c represented by the central marker q_2. The variant pinball loss function evaluates residuals as $\rho(r^m_n) = r^m_n(\tau - \mathbb{I}\{r^m_n < 0\})$, where $r^m_n = \Gamma^m - c$ represents the difference between sample-specific reliability scores and the P2-estimated quantile threshold.

The computational workflow integrates P2 estimation into the calibration procedure through a multi-stage process. During initialization, the algorithm processes the first five samples from the calibration set to establish the marker positions, setting q_0 to the minimum observed score, q_4 to the maximum, and initializing q_2 as a preliminary p-quantile estimate. For each subsequent calibration sample, the algorithm executes the complete online update sequence: determining the insertion interval, incrementing marker positions, adjusting desired positions by the constant increments $dns_i = [0, \tau/2, \tau, (1+\tau)/2, 1]$, and correcting non-extreme markers q_1, q_2, and q_3 based on the deviation $d = n'_i - n_i$. The marker correction employs the parabolic interpolation formula when feasible, reverting to linear interpolation if necessary, and increments the actual position n_i by

sign(d) upon successful adjustment. The resulting q_2 value serves as the dynamic threshold c for loss computation. Within each training iteration, before backpropagation, the current q_2 value is retrieved as c. The residual $r_n^m = \Gamma^m - c$ is computed for each sample, and the pinball objective is evaluated. The negative-class regularization term subsequently enforces constraints on incorrect classes using the same dynamically computed c.

The asymmetric variant pinball loss exhibits quantile-specific penalization behavior that synergizes with P2's dynamic threshold estimation. When $r_n^m < 0$ (indicating $\Gamma^m < c$), the loss applies weight τ, imposing substantial penalty on samples with scores below the threshold and driving the model to increase confidence for these instances. Conversely, when $r_n^m \geq 0$ (indicating $\Gamma^m \geq c$), the weight becomes $\tau - 1$, applying minimal penalty to samples with scores exceeding the threshold and permitting them to maintain higher uncertainty. By continuously estimating the true τ-quantile of the evolving reliability score distribution, the P2 algorithm ensures that the penalization strength adapts automatically to distribution changes, thereby maintaining effective calibration throughout the training process.

A.2 Federated update process

Algorithm 1 Reliable Federated Multi-View Learning (FedRMVL-CAL)

Initialize: M-view datasets $X = \{X^1, X^2, \ldots, X^M\}$, common label space Y; Trade-off hyperparameters λ_g, λ_k; Initial conflict prior $\hat{Z}_n$ and global mean Dirichlet parameter $\bar{\alpha}^g$.
Output: Global predictive probability $p_k^g = \alpha_k^g / S^g$.
repeat
  Local Client Process (Perform parallel for client $m = 1$ to M):
  Split local dataset X^m into evidential subset X_{evi}^m and calibration subset X_{cali}^m.
  Evidence Reliability Calibration:
    Retrieve stored conflict prior $\hat{Z}_n$(Eq.9) and $\bar{\alpha}^g$ from server.
    Calculate evidence reliability score Γ^m (Eq.4).
    Compute calibration loss $\mathcal{L}_{cal}$ (Eq. 5).
  Evidential Learning:
    Generate non-negative evidence $e^m = a(f(X^m; \theta^m))$.
    Derive belief b_k^m, uncertainty u^m, and Dirichlet strength S^m (Eq.1).
    Compute local evidential loss $\mathcal{L}_{evi}(\alpha^m)$ and global alignment loss $\mathcal{L}_{evi}(\alpha^g)$ (Eq.3).
  Uplink Compression:
    Calculate mean evidence $\mu^m = \frac{1}{N}\sum_{n=1}^{N} e_n^m$.
    Select subset $\mathcal{B}$ of size β, where sampling probability $p(n \in \mathcal{B}) \propto u_n^m$.
    Send exact evidence $\{e_n^m\}_{n\in\mathcal{B}}$ and mean summary μ^m to server.
  Central Server Process:
  Reconstruction and Fusion:
    Reconstruct local Dirichlet parameters $\hat{\alpha}_n^m$ for all samples (Eq.8).
    Perform AEF to yield global parameters α_k^g (Eq.7).
    Calculate per-sample the conflict Z_n and $\hat{Z}_n$ (Eq. 6).
  Downlink Broadcast:
    Compute global means $\bar{\hat{Z}} = \frac{1}{N}\sum \hat{Z}_n$ and $\bar{\alpha}^g = \frac{1}{N}\sum \alpha_n^g$.
    Broadcast $\{\hat{Z}_n\}_{n\in\mathcal{B}}$, $\bar{\hat{Z}}$, and $\bar{\alpha}^g$ to all clients.
until Convergence

B Detailed Proofs of Theoretical Properties

In this section, we provide the comprehensive mathematical derivations for the three propositions presented in Section 3.5. We consistently denote the fused Dirichlet parameters as α_k^g, the fused total strength as S^g, and the global uncertainty as u^g. The Adaptive Evidence Fusion (AEF) rule is defined as:

$$\alpha_k^g = \frac{1-Z}{1+\hat{Z}} \sum_{m=1}^{M} w^m \alpha_k^m, \quad S^g = \sum_{k=1}^{K} \alpha_k^g = \frac{1-Z}{1+\hat{Z}} \sum_{m=1}^{M} w^m S^m, \tag{14}$$

where $w^m = (1-u^m)/\sum_j (1-u^j)$ is the uncertainty-aware view weight, Z represents the baseline uniform-aggregation conflict, and $\hat{Z}$ represents the calibration-weighted conflict prior.

B.1 Proof of Proposition 1: Monotonic Uncertainty Regulation

Proposition 1. The global uncertainty u^g increases with the historical conflict prior $\hat{Z}$, and decreases as reliable local evidence S^m increases.

Proof. By definition, the global uncertainty is given by $u^g = K/S^g$. Analyzing the monotonicity of u^g is strictly equivalent to examining the inverse monotonicity of the fused Dirichlet strength S^g. We first evaluate the dependency of the global uncertainty on the historical conflict prior $\hat{Z}$. By taking the partial derivative of S^g with respect to $\hat{Z}$ while treating Z, w^m, and S^m as constants for the current iteration, we obtain:

$$\frac{\partial S^g}{\partial \hat{Z}} = \left((1-Z) \sum_{m=1}^{M} w^m S^m \right) \cdot \frac{\partial}{\partial \hat{Z}} (1+\hat{Z})^{-1} = -\frac{1-Z}{(1+\hat{Z})^2} \sum_{m=1}^{M} w^m S^m. \tag{15}$$

According to the definitions in the methodology, the instantaneous conflict Z is strictly bounded within $[0, 1)$, the weights satisfy $w^m > 0$, and the local strengths satisfy $S^m \geq K > 0$. Consequently, the entire expression is strictly negative, yielding $\frac{\partial S^g}{\partial \hat{Z}} < 0$. Applying the chain rule to the global uncertainty u^g gives:

$$\frac{\partial u^g}{\partial \hat{Z}} = \frac{\partial (K/S^g)}{\partial S^g} \cdot \frac{\partial S^g}{\partial \hat{Z}} = -\frac{K}{(S^g)^2} \cdot \left(-\frac{1-Z}{(1+\hat{Z})^2} \sum_{m=1}^{M} w^m S^m \right) > 0. \tag{16}$$

This positive derivative establishes that the global uncertainty u^g strictly inflates with the accumulation of the historical conflict prior $\hat{Z}$.

Furthermore, we examine the effect of increasing the local evidence strength S^m from a reliable client. A strict increase in S^m intrinsically leads to a strict decrease in the local uncertainty $u^m = K/S^m$. Based on the weight normalization function $w^m = (1-u^m)/\sum_j (1-u^j)$, a decrease in u^m strictly monotonically increases its view weight w^m. For a reliable client, whose predicted belief mass correctly aligns with the true class, this increased weight w^m further concentrates the global consensus. Such concentration explicitly lowers the cross-view geometric conflict, meaning Z strictly decreases, which in turn causes the scaling factor $(1-Z)$ to strictly increase. Concurrently, the weighted evidence sum $\sum_{m=1}^{M} w^m S^m$ strictly increases because the increment in S^m outpaces the change in the weight denominator. Given that both $(1-Z)$ and the weighted sum strictly increase while the historical prior $\hat{Z}$ remains unchanged in the current forward pass, the global Dirichlet strength S^g strictly increases. As a result, the global uncertainty $u^g = K/S^g$ strictly decreases. □

B.2 Proof of Proposition 2: Transmission Approximation Bound

Proposition 2. Let α^g and $\hat{\alpha}^g$ denote the raw and server-reconstructed global Dirichlet parameters under AEF, respectively. For any unobserved instance, the server substitutes $\hat{\alpha}_k^m = \mu_k^m + 1$ for the unobserved evidence. The resulting reconstruction error satisfies:

$$\|\alpha^g - \hat{\alpha}^g\|_1 \leq \frac{1-Z}{1+\hat{Z}} \sum_{m=1}^{M} w^m \|e^m - \mu^m\|_1,$$

and vanishes as all instances are transmitted or as local evidence concentrates ($e^m \to \mu^m$).

Proof. Let the error in the reconstructed global parameter for class k be denoted as $\Delta\alpha_k^g = \alpha_k^g - \hat{\alpha}_k^g$. Given the linearity of the AEF formulation with respect to the client parameters, we can express this error as:

$$\Delta\alpha_k^g = \frac{1-Z}{1+\hat{Z}} \sum_{m=1}^{M} w^m (\alpha_k^m - \hat{\alpha}_k^m). \tag{17}$$

For an unobserved instance $n \notin \mathcal{B}$ that is compressed and not explicitly transmitted, the server substitutes the local Dirichlet parameter using the mean evidence summary, yielding $\hat{\alpha}_k^m = \mu_k^m + 1$. Recognizing that the raw parameter is derived from the exact evidence as $\alpha_k^m = e_k^m + 1$, the substitution error for class k at client m simplifies to:

$$\alpha_k^m - \hat{\alpha}_k^m = (e_k^m + 1) - (\mu_k^m + 1) = e_k^m - \mu_k^m. \tag{18}$$

To quantify the total deviation across the entire probability simplex, we compute the L_1 norm of the global parameter error, which is the sum of the absolute deviations over all K classes:

$$\|\alpha^g - \hat{\alpha}^g\|_1 = \sum_{k=1}^{K} |\Delta\alpha_k^g| = \sum_{k=1}^{K} \left| \frac{1-Z}{1+\hat{Z}} \sum_{m=1}^{M} w^m (e_k^m - \mu_k^m) \right|. \tag{19}$$

By applying the triangle inequality to the absolute value of the summation, we can move the absolute value operator inside the client summation. Since the factor $\frac{1-Z}{1+\hat{Z}}$ and the weights w^m are non-negative scalars, they can be factored out of the absolute value:

$$\|\alpha^g - \hat{\alpha}^g\|_1 \leq \frac{1-Z}{1+\hat{Z}} \sum_{k=1}^{K} \sum_{m=1}^{M} w^m |e_k^m - \mu_k^m|. \tag{20}$$

Since the summations over k and m are finite, we are permitted to safely interchange the order of summation:

$$\|\alpha^g - \hat{\alpha}^g\|_1 \leq \frac{1-Z}{1+\hat{Z}} \sum_{m=1}^{M} w^m \left(\sum_{k=1}^{K} |e_k^m - \mu_k^m| \right). \tag{21}$$

Recognizing that the inner summation $\sum_{k=1}^{K} |e_k^m - \mu_k^m|$ is exactly the definition of the L_1 norm for the evidence vector deviation $\|e^m - \mu^m\|_1$, we arrive at the final bound:

$$\|\alpha^g - \hat{\alpha}^g\|_1 \leq \frac{1-Z}{1+\hat{Z}} \sum_{m=1}^{M} w^m \|e^m - \mu^m\|_1. \tag{22}$$

Based on the uncertainty-aware sampling design, unobserved instances are specifically selected to have low local uncertainty u^m, which mathematically implies a massive total strength $S^m \gg K$. In this highly confident regime, the evidence distribution e^m becomes extremely concentrated and sharply peaked around its local mean μ^m, causing the residual $\|e^m - \mu^m\|_1$ to be practically negligible. Furthermore, the bound collapses to zero identically when the transmission ratio β equals the dataset size N, confirming that the compression scheme approaches lossless fusion as transmission bandwidth increases. □

B.3 Proof of Proposition 3: Calibration–Fusion Synergy

Proposition 3. Suppose a client m applies a calibration increment $\delta > 0$ to the target belief b_y^m of a correctly classified instance, with non-target beliefs decaying proportionally to maintain the simplex constraint. Then, both the baseline conflict Z and the weighted conflict prior $\hat{Z}$ strictly decrease, leading to a monotonic increase in the fused Dirichlet strength S^g and a strict attenuation of the global uncertainty u^g.

Proof. Let b_y^m represent the belief mass assigned to the ground-truth target class y by client m. Under the evidence reliability calibration, the target belief experiences a positive continuous increment parameterized by $\delta > 0$. To strictly preserve the probability simplex constraint $\sum_k b_k^m = 1 - u^m$, the non-target belief components b_k^m (where $k \neq y$) must collectively decay. Assuming a proportional decay relative to their existing mass, the continuous variation rate of the beliefs can be defined as $\frac{\partial b_y^m}{\partial \delta} = 1$ and $\frac{\partial b_k^m}{\partial \delta} = -\frac{b_k^m}{1-b_y^m}$. Consequently, the uniform arithmetic mean belief $\bar{b}_k = \frac{1}{M}\sum_{j=1}^{M} b_k^j$ changes with respect to δ at the rates:

$$\frac{\partial \bar{b}_y}{\partial \delta} = \frac{1}{M}, \quad \text{and} \quad \frac{\partial \bar{b}_k}{\partial \delta} = -\frac{1}{M}\frac{b_k^m}{1-b_y^m} \quad \text{for } k \neq y. \tag{23}$$

We now evaluate the impact of this calibration mapping on the baseline uniform-aggregation geometric conflict, defined as $Z = 1 - \sum_{k=1}^{K} \bar{b}_k^2$. Taking the partial derivative of Z with respect to the calibration increment δ yields:

$$\frac{\partial Z}{\partial \delta} = \frac{\partial}{\partial \delta}\left(1 - \sum_{k=1}^{K} \bar{b}_k^2\right) = -2 \sum_{k=1}^{K} \bar{b}_k \frac{\partial \bar{b}_k}{\partial \delta}. \tag{24}$$

By isolating the target class y from the non-target classes and substituting the derived rates of change, we obtain:

$$\frac{\partial Z}{\partial \delta} = -2\left(\bar{b}_y\left(\frac{1}{M}\right) + \sum_{k \neq y} \bar{b}_k\left(-\frac{1}{M}\frac{b_k^m}{1 - b_y^m}\right)\right) = -\frac{2}{M}\left(\bar{b}_y - \sum_{k \neq y} \bar{b}_k \frac{b_k^m}{1 - b_y^m}\right). \tag{25}$$

To determine the sign of this derivative, we analyze the term $\sum_{k \neq y} \bar{b}_k \frac{b_k^m}{1 - b_y^m}$. Notice that the scaling factors $\frac{b_k^m}{1 - b_y^m}$ are strictly non-negative and sum exactly to 1 across all non-target classes. Therefore, this summation represents a strict convex combination of the non-target mean beliefs $\{\bar{b}_k\}_{k \neq y}$. Since the instance is correctly classified globally, the target mean belief inherently strictly dominates any individual non-target mean belief, meaning $\bar{b}_y > \bar{b}_k$ for all $k \neq y$. The convex combination of a set of elements strictly less than $\bar{b}_y$ is mathematically guaranteed to be strictly less than $\bar{b}_y$. Thus, $\bar{b}_y - \sum_{k \neq y} \bar{b}_k \frac{b_k^m}{1 - b_y^m} > 0$, confirming that $\frac{\partial Z}{\partial \delta} < 0$. This proves that the calibration-induced alignment of local evidence towards the true class strictly monotonically reduces the baseline conflict Z.

A structurally identical logic applies to the conflict prior $\hat{Z} = 1 - \sum_{k=1}^{K} \bar{b}_{w,k}^2$, where $\bar{b}_{w,k} = \sum_{j=1}^{M} w^j b_k^j$. The calibration increment on b_y^m intrinsically reduces the local uncertainty u^m, which synchronously increases the adaptive weight w^m. The combination of an increased target belief and an increased weight ensures that the precision-weighted mean vector is driven towards the target class consensus at an even faster rate than the uniform mean, unequivocally resulting in a strict reduction of the weighted conflict, $\frac{\partial \hat{Z}}{\partial \delta} < 0$.

Finally, we translate the reduction in these conflict terms into the dynamic behavior of the global parameters. The fused Dirichlet strength is defined as $S^g = \frac{1-Z}{1+\hat{Z}} \sum_{m=1}^{M} w^m S^m$. The partial derivatives of S^g with respect to the conflict variables are:

$$\frac{\partial S^g}{\partial Z} = -\frac{1}{1+\hat{Z}} \sum_{m=1}^{M} w^m S^m < 0, \quad \text{and} \quad \frac{\partial S^g}{\partial \hat{Z}} = -\frac{1-Z}{(1+\hat{Z})^2} \sum_{m=1}^{M} w^m S^m < 0. \tag{26}$$

Because both Z and $\hat{Z}$ strictly decrease during the calibration update, and both possess strictly negative partial derivatives regarding S^g, their simultaneous reduction strictly monotonically amplifies the global Dirichlet strength S^g. Consequently, the global uncertainty, explicitly defined as the inverse relation $u^g = K/S^g$, strictly attenuates. This end-to-end continuous differentiation firmly establishes that integrating evidence reliability calibration on the client side fundamentally regulates and systematically enhances the certainty and stability of the global evidence fusion mechanism. □

B.4 Convergency analysis

Proposition 4 (Convergence under Partial Transmission)**.** Suppose $\mathcal{L}_{\text{client}}^m$ is Lipschitz continuous with bounded subgradients and stochastic gradients have bounded variance σ^2. Running SGD with step size $\eta_{\text{lr}} = \frac{1}{\sqrt{T}}$ yields:

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left\|\nabla \mathcal{L}_{\text{client}}^m\right\|^2 \;\leq\; \mathcal{O}\left(\frac{L\sigma^2}{\sqrt{T}}\right) + \lambda_g L' \cdot \frac{N-\beta}{N} \cdot \bar{\delta} \tag{27}$$

where L' is the Lipschitz constant of $\nabla_{\hat{\boldsymbol{\alpha}}^g} \mathcal{L}_{\text{evi}}(\boldsymbol{\alpha}^g)$ and $\bar{\delta} = \mathbb{E}_{n \notin \mathcal{B}}\left[\frac{1}{M}\sum_m \|\boldsymbol{e}_n^m - \boldsymbol{\mu}^m\|_1\right]$. The bias term vanishes when $\beta = N$ or as $\bar{\delta} \to 0$.

Proof. Partial transmission introduces gradient bias only through the global evidence term $\lambda_g \mathcal{L}_{\text{evi}}(\boldsymbol{\alpha}^g)$ in Eq.(10), since $\mathcal{L}_{\text{evi}}(\boldsymbol{\alpha}^m)$ and $\mathcal{L}_{\text{cali}}$ are computed exactly on local data. For $n \notin \mathcal{B}$,

the AEF formula is linear in $\{\alpha_{n,k}^m\}$, so the reconstruction error satisfies:

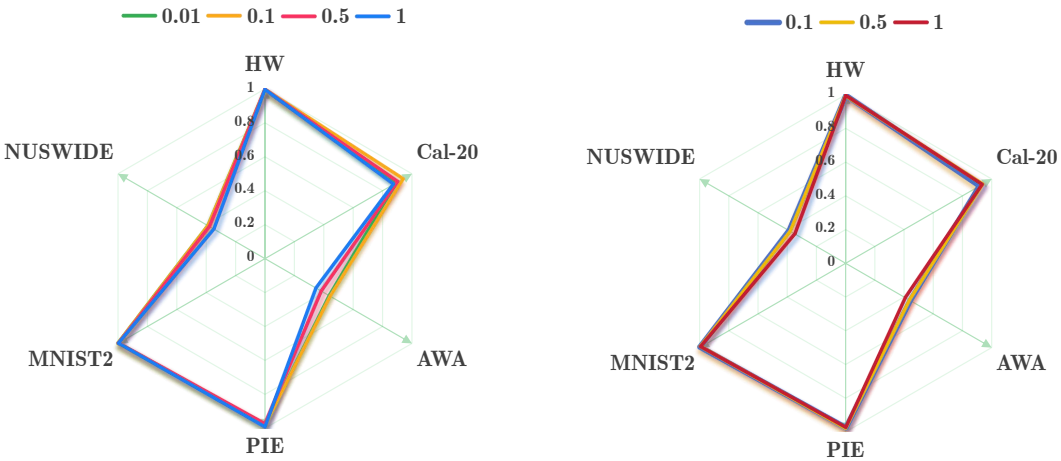
$$\left\|\hat{\boldsymbol{\alpha}}_n^g - \boldsymbol{\alpha}_n^g\right\|_1 = \frac{1 - Z_n}{1 + \hat{Z}_n} \sum_m w_n^m \left\|\boldsymbol{e}_n^m - \boldsymbol{\mu}^m\right\|_1 \;\leq\; \bar{\delta} \tag{28}$$

by definition of $\bar{\delta}$ and since $\frac{1-Z_n}{1+\hat{Z}_n} \leq 1$. By L'-smoothness of $\mathcal{L}_{\text{evi}}$ in $\hat{\boldsymbol{\alpha}}^g$:

$$\left\|\nabla \mathcal{L}_{\text{evi}}(\hat{\boldsymbol{\alpha}}^g) - \nabla \mathcal{L}_{\text{evi}}(\boldsymbol{\alpha}^g)\right\| \;\leq\; L' \sum_{n \notin \mathcal{B}} \left\|\hat{\boldsymbol{\alpha}}_n^g - \boldsymbol{\alpha}_n^g\right\|_1 \;\leq\; L'(N - \beta)\,\bar{\delta} \tag{29}$$

Incorporating this bias into the standard SGD convergence analysis under L-smoothness and variance σ^2, and scaling by $\frac{\lambda_g}{N}$, yields the stated bound. When $\beta = N$, all samples are transmitted exactly and the bias vanishes. When evidence concentrates ($\boldsymbol{e}_n^m \to \boldsymbol{\mu}^m$ for $n \notin \mathcal{B}$), $\bar{\delta} \to 0$ and the bias disappears adaptively, confirming that accurate mean estimation suffices for convergence under partial transmission. □

C Experimental setup and additional experimental results



(a) λ_g (b) λ_k

Figure 6: The parameter sensitivity analysis

C.1 Parameter sensitivity analysis

We conduct a sensitivity analysis to evaluate the impact of the key hyperparameters, λ_g and λ_k, on the performance of our model. As illustrated in Fig. 6, we varied λ_g and λ_k across a wide range of values on six diverse datasets. The results demonstrate that the accuracy of the model exhibits remarkable stability, showing only marginal variations under different hyperparameter settings for both λ_g and λ_k. This pronounced robustness indicates that our method is not sensitive to the precise choice of these parameters, which alleviates the burden of extensive hyperparameter tuning and underscores the practical utility of our approach.

C.2 Detailed experimental Setup

The project code runs on the PyTorch framework using Python version 3.11.11, with an Intel Core i9-14900K CPU and an NVIDIA RTX 4090 (24GB) GPU. The random seed is fixed at 52. The Adam optimizer is employed with a learning rate of 1e-2 and a weight decay of 1e-4.The data is partitioned as follows: 80% of the samples are used for training, and 20% are used for testing. 10% of the samples in the training set are used for non-consistency calibration.

C.3 Dataset introduction and configuration

Multi-view datasets, which provide complementary representations of samples through heterogeneous feature spaces, are essential for evaluating multi-view learning algorithms. Below, we formally introduce the six benchmark datasets employed in this study.

Animals with Attributes (AwA)[1] is a widely recognized multi-view dataset consisting of 30,475 images across 50 animal categories. Each sample is represented by six distinct feature views, including diverse visual descriptors such as color histograms, SIFT, and VGG features, with dimensionalities of 2688, 2000, 252, 2000, 2000, and 2000, respectively. This dataset is particularly valuable for attribute-based classification and cross-view retrieval tasks.

MNIST2[2] is a derived benchmark constructed from the original MNIST handwritten digit corpus, encompassing 10 classes (digits 0-9) with 20,000 total samples. It comprises three heterogeneous feature views extracted via ISOMAP (ISO), Linear Discriminant Analysis (LDA), and Neighborhood Preserving Embedding (NPE), with dimensions of 342, 1024, and 64. This dataset serves as a standard testbed for low-dimensional multi-view representation learning.

NUS-WIDEOBJ[3], a subset of the NUS-WIDE dataset, is a large-scale multi-view object recognition benchmark containing 30,000 images from 31 concept categories. Each image is characterized by five complementary visual descriptors with dimensions of 65, 226, 145, 74, and 129, capturing diverse aspects of visual appearance. The dataset's multi-label nature and view heterogeneity make it suitable for evaluating robust multi-view clustering and classification methods.

CMU PIE (Pose, Illumination, and Expression)[4] is a multi-view face recognition dataset derived from the CMU Multi-PIE database, featuring 680 facial images of 68 subjects. Three feature views are extracted: intensity values (1024-dimensional), Local Binary Patterns (LBP), and Gabor wavelet responses, each encoding different facial characteristics under varying pose, illumination, and expression conditions. This dataset is extensively used for assessing multi-view subspace learning algorithms in biometric applications.

handwritten[5] is a specialized multi-view dataset for handwritten numeral recognition, comprising samples from 10 digit classes. Six statistical feature sets are extracted from each image, including Fourier coefficients (FOU), profile correlations (FAC), Karhunen-Loève coefficients (KAR), pixel averages (PIX), Zernike moments (ZER), and morphological features (MOR), with dimensionalities ranging from low to moderate. This dataset provides a clean, well-structured benchmark for evaluating view fusion strategies.

Caltech101-20[6] is a curated subset of the Caltech101 dataset, containing 2,386 images from 20 object categories. Six widely-used visual descriptors are extracted: Gabor filters (48D), CENTRIST (40D), Histogram of Oriented Gradients (254D), Wavelet moments (1984D), GIST (512D), and Local Binary Patterns (928D). The high dimensionality and significant heterogeneity across views make this dataset particularly challenging and representative for testing multi-view dimensionality reduction and classification algorithms.

Collectively, these datasets span diverse domains from object recognition and face verification to handwritten digit classification and exhibit varying numbers of views, feature dimensions, and sample sizes, providing a comprehensive evaluation suite for multi-view learning methodologies.

The data allocation and configuration used by each client during training are shown in 3.

C.4 Compared methods introduction

We compare against centralized multi-view methods and single-view FL baselines. Below are detailed introductions to these comparison methods.

Centralized Multi-view Methods:

[1] https://cvml.ist.ac.at/AwA/

[2] http://yann.lecun.com/exdb/mnist/

[3] https://lms.comp.nus.edu.sg/wp-content/uploads/2019/

[4] https://www.cs.cmu.edu/afs/cs/project/PIE/

[5] https://archive.ics.uci.edu/dataset/72/multiple+features

[6] http://www.vision.caltech.edu/archive.html

Table 3: Dataset configurations

Dataset	Test Batch Size	Train Batch Size	Dimensions (Views)
AwA	6059	20000	[2688, 2000, 252, 2000, 2000, 2000]
MNIST2	12000	20000	[342, 1024, 64]
NUSWIDEOBJ	6000	10000	[65, 226, 145, 74, 129]
PIE	325	400	[1024, 1024, 1024]
handwritten_6views	600	600	[240, 76, 216, 47, 64, 6]
Caltech101-20	477	600	[48, 40, 254, 1984, 512, 928]

- **DUA-Nets (Dynamic Uncertainty-Aware Networks)**: Proposed by Geng et al.[35], DUA-Nets is an uncertainty-aware multi-view fusion method that employs reversal networks to integrate intrinsic information from different views into a unified representation. It dynamically assesses the reliability of each view and incorporates uncertainty quantification to improve robustness in multi-view decision-making.
- **TMC (Trusted Multi-view Classification)**: Introduced by Han et al.[20], TMC is a pioneering uncertainty-aware multi-view classification framework that enhances decision reliability by estimating classification probabilities and uncertainty through evidential deep neural networks. It transforms traditional classification networks into evidential neural networks by replacing the final softmax function with a ReLU activation, enabling the modeling of Dirichlet distributions. The method employs Dempster's combination rule to fuse multi-view opinions, providing both class probabilities and uncertainty estimates for trustworthy predictions.
- **CCML (Consistent and Complementary-aware trusted Multi-view Learning)**: Recently proposed by Liu et al.[21], CCML addresses the semantic vagueness phenomenon in real-world multi-view data by dynamically decoupling evidence into consistent and complementary components. Consistent evidence captures shared information across all views, while complementary evidence aggregates view-specific variations. This decoupling strategy allows CCML to handle cases where individual views cannot effectively differentiate all categories, significantly improving both accuracy and reliability.
- **ECML (Evidential Conflictive Multi-view Learning)**: Developed by Xu et al.[36], ECML tackles the challenge of conflictive multi-view data where low-quality instances exhibit contradictory information across views. Unlike methods that simply eliminate conflictive instances, ECML constructs view-specific opinions using evidential theory and proposes a conflictive opinion aggregation strategy that models both common and view-specific reliabilities, enabling reliable decision-making for conflictive data.
- **TMCEK (Trusted Multi-view Classification with Evidential Kernel)**: Proposed by Liang et al.[22], TMCEK addresses fundamental limitations in existing Dempster-Shafer theory-based multi-view classification methods. While prior approaches focus on providing decision confidence without explanatory capability and rely on first-order statistical magnitudes of belief masses that inadequately capture intrinsic uncertainty, TMCEK integrates expert knowledge to enhance feature-level interpretability and introduces a distribution-aware subjective opinion mechanism for more reliable confidence estimation. The theoretical superiority of its uncertainty measure over conventional approaches is rigorously established.
- **TEF (Trusted Multi-view Classification via Evolutionary Multi-view Fusion):** Developed by Liang et al.[37], TEF addresses the limitations of late fusion strategies in Dempster-Shafer theory-based methods, which restrict information interaction among views and lead to suboptimal data utilization. To overcome the vulnerability of conventional pseudo-view generation to noisy views and imbalanced learning, TEF employs an evolutionary multi-view architecture search method to adaptively select views and fusion operators, constructing a high-quality fusion architecture. Furthermore, it enhances individual views by concatenating the decision output of this fusion architecture with each respective view, effectively mitigating imbalanced multi-view learning issues, particularly on complex datasets with numerous views.

Single-view federated learning methods

- **FedAvg (Federated Averaging)**: As the foundational algorithm in federated learning proposed by McMahan et al.[38], FedAvg serves as the primary baseline. It performs local stochastic gradient descent (SGD) on each client for multiple epochs and aggregates the model updates by weighted averaging based on each client's data size. FedAvg assumes independent and identically distributed (IID) data and homogeneous client capabilities.
- **FedProx (Federated Proximal)**: Introduced by Li et al.[39], FedProx addresses statistical and systems heterogeneity in federated networks by adding a proximal term to the local objective function. This term constrains local updates to remain close to the global model, enabling more stable convergence with non-IID data and heterogeneous client participation. FedProx generalizes FedAvg and provides theoretical convergence guarantees under these challenging conditions.
- **β-preBayes (Beta-preBayes)**: This Bayesian federated learning approach incorporates Beta distributions for uncertainty quantification in the global aggregation process[40]. It extends conventional federated averaging by modeling client contributions probabilistically, allowing for robust aggregation that accounts for client reliability and data quality. The method uses Beta priors to weight client updates based on their historical performance and uncertainty estimates.
- **RHFL+ (Robust Heterogeneous Federated Learning)**: RHFL+[41]is an advanced federated learning framework designed for robust training in heterogeneous environments. It employs confidence-based clustering to group clients with similar data distributions and implements adaptive aggregation strategies that account for both client reliability and model uncertainty. The method enhances robustness against malicious or low-quality client updates while preserving privacy guarantees in heterogeneous federated settings.

Multi-view federated learning method

- **FedRMVL (Federated Reliable Multi-view Classification)**: FedRMVL[32] is a vertical federated multi-view learning framework tailored for reliable classification in mobile edge environments. It incorporates Subjective Logic for lightweight uncertainty quantification at local edge devices and utilizes the Dempster-Shafer combination rule to achieve adaptive and reliable multi-view opinion fusion at the central server. To address feature dimension heterogeneity among diverse data views, the framework introduces a partial parameter-sharing strategy during federated optimization. FedRMVL enhances global model performance by capturing both consistency and complementarity across heterogeneous views, with its effectiveness and robustness validated through rigorous theoretical analysis and extensive experiments on real-world datasets.

C.5 Detailed experimental setup

Network Architecture and Parameter Sharing: To accommodate the inherent heterogeneity in feature dimensions across different views (clients), all evaluated models are instantiated using a three-layer Multi-Layer Perceptron (MLP). Specifically, the network is decoupled into a view-specific local feature extractor (the first linear layer, mapping from the unique input dimension d_m to a common hidden dimension) and a shared isomorphic classifier (the subsequent two layers). During the communication rounds, only the parameters of the shared isomorphic classifier are transmitted and aggregated. The view-specific feature extractors are kept strictly local to preserve specific view representations and accommodate dimensional discrepancies.

Adaptation of Single-View FL Baselines to FedMVL: Since standard single-view FL baselines are originally designed for homogeneous architectures, we rigorously adapted them to our heterogeneous multi-view setting to ensure fair comparisons:

- *Weight-space aggregation baselines (FedAvg, FedProx):* These algorithms are executed by aggregating only the aforementioned isomorphic classifier layers. The local feature extractors are updated exclusively via local backpropagation.
- *Predictive-space and heterogeneous baselines (β-preBayes, RHFL+):* Since these methods inherently support model heterogeneity by exchanging information in the predictive/logit space or utilizing knowledge distillation, they are seamlessly applied to our multi-view scenario without architectural constraints.

- *Testing Protocol:* For single-view FL baselines, the global predictive performance is evaluated by uniformly ensembling the output logits (or probabilities) from all participating view-specific models on the test set.

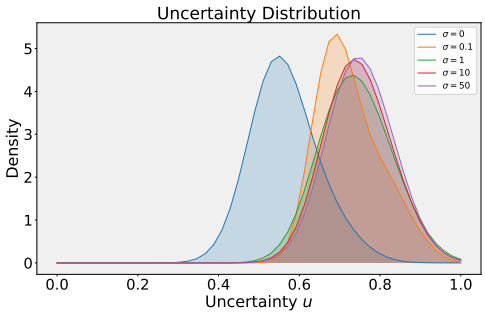
Label Availability: In our experimental setup, we assume that all clients share a common label space and have access to the ground-truth labels for local supervised optimization. This ensures that each local view can independently quantify its view-specific uncertainty and compute the necessary calibration loss before global evidence fusion.

C.6 Extend results of uncertainty estimation

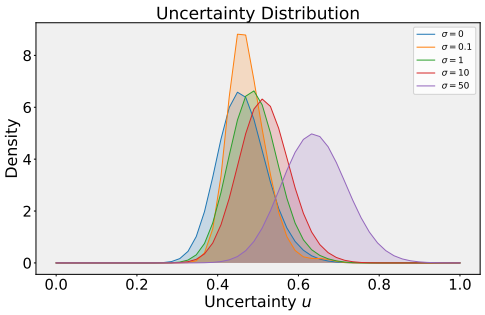
The uncertainty distributions presented in Fig.7 corroborate the proposed algorithm's efficacy in quantifying data reliability, exhibiting a consistent monotonic trend where probability density curves shift towards higher uncertainty intervals and show increased dispersion as noise intensity escalates from $\sigma = 0$ to $\sigma = 50$. While datasets such as MNIST2 and HW display a smooth, progressive transition in uncertainty reflecting the gradual degradation of data quality, others like AWA and NUS reveal a distinct stratification with a sharp demarcation between low-noise and high-noise scenarios, indicating a threshold-based sensitivity to severe corruption. Furthermore, the relatively high baseline uncertainty observed in CAL and PIE implies that the model is acutely calibrated to inherent data complexity even in the absence of significant noise, collectively confirming that the evidential learning framework successfully maps varying degrees of data degradation to reliable uncertainty estimates across all participation views.

D Limitation

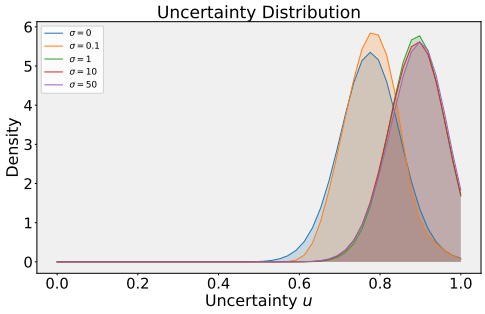
The proposed algorithm addresses the reliability degradation problem in federated multi-view learning setting, which arises from inconsistent data quality across views in real-world applications. However, the current experimental setup assumes an simplified scenario. In practice, it is common for both single-view and multi-view data to coexist on the same client, or for only a subset of views to exhibit degraded quality. Refining the problem setup to account for these more realistic conditions could improve the generalization ability of the algorithm, particularly in safety-critical domains such as finance, healthcare, and autonomous trajectory planning, where robustness and reliability are essential. This will also be the main task of our follow-up work.



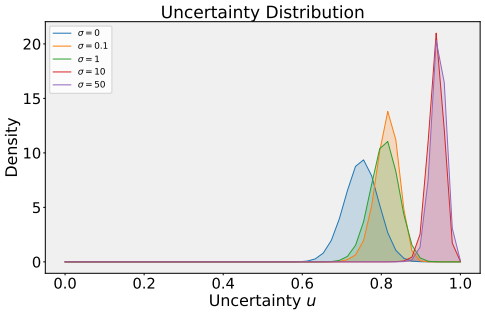
(a) CAL



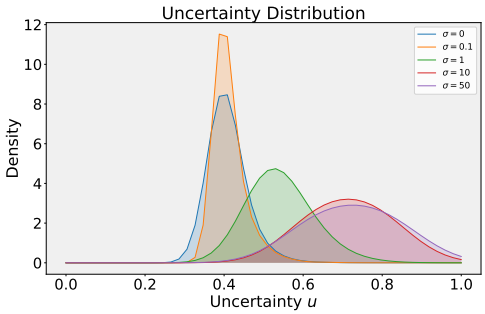
(b) HW



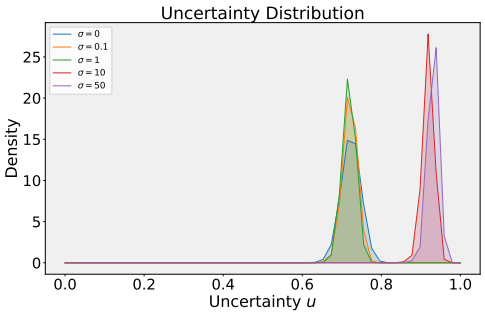
(c) PIE



(d) AWA



(e) MNIST2



(f) NUS

Figure 7: The uncertainty distribution changes in 6 datasets

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See the abstract and section 1.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See the appendix D.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: See the section 3.5.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See the appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: See the supplemental material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See the appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See the section 4 and Table. 1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See the appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All codes are anonymity.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See the section 5

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The authors has cited the original paper that produced the code package or dataset.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.