UNCERTAINTY-AWARE EVIDENTIAL PROTOTYPE RECONSTRUCTION FOR ROBUST MULTIMODAL FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal federated learning must reconcile heterogeneous client data with different available modalities. Noisy supervision adds a further difficulty: disagreement may arise from mismatched modalities, corruption of one modality, or an incorrect label shared by otherwise consistent inputs. We introduce FedEPR, an evidential prototype reconstruction framework that uses these distinct reliability patterns to guide label correction. Modality-specific Dirichlet evidence provides an uncertainty proxy, while Dempster–Shafer fusion exposes cross-modal conflict. Globally calibrated prototypes anchor local semantics under non-IID data, and an ordered reconstruction rule selects a reliable modality, corrects confident consensus, or retains the observed label. Warm-up training stabilizes the anchors before reconstruction, and modality-specific evidence weights determine client aggregation. Across three bimodal benchmarks, FedEPR improves on the strongest compared baseline by 12.54 and 12.01 percentage points on AV-MNIST and UPMC Food-101 under 60% controlled structured noise. With naturally occurring noise, the corresponding improvements are 4.26 points on raw UPMC Food-101 and 3.78 points on non-IID WebVision. Additional noise processes, calibration sweeps, and aggregation diagnostics characterize where the reconstruction design is effective.

1 INTRODUCTION

Multimodal federated learning (MMFL) combines complementary signals across clients while keeping raw observations local. Its applications include biomedical analysis (Acosta et al., 2022; Chen & Pan, 2024) and autonomous driving (Zheng et al., 2023). Recent methods address missing modalities through prototype completion (Yu et al., 2026; Bao et al., 2026), balance heterogeneous gradients (Chen & Li, 2022), and exchange knowledge through contrastive ensembles (Yu et al., 2023). These advances make collaboration possible when clients differ in both their data distributions and their available sensors. Yet modality availability does not determine whether the available supervision is trustworthy. Web labels, automatic annotation, and weakly aligned inputs can introduce errors whose structure depends on relationships between modalities.

The difficulty is more specific than an incorrect class index. As illustrated in Figure 1, semantic mismatch occurs when the modalities support different concepts; asymmetric corruption leaves one modality informative while another is degraded; and consistent label error occurs when coherent modalities contradict their observed label. A confident prediction from one branch can be useful in the second case but misleading in the first. Agreement is also insufficient: two modalities may share an incorrect learned association. Reliable correction therefore requires a joint account of modal agreement, evidential support, and class geometry.

Federation further complicates this decision. Non-IID class distributions shift local feature centers, so a suspicious annotation may reflect either noise or limited local coverage. Unimodal clients cannot cross-check their inputs. At aggregation, a noisy branch can influence clients that do not share its failure mode. Existing noisy-label methods use sample selection, loss statistics, or pseudo-labeling (Li et al., 2020a; 2022; Xu et al., 2022; Wu et al., 2023); applying these signals independently to each modality leaves cross-modal disagreement implicit. The FedMultimodal benchmark documents the sensitivity of MMFL to imperfect data (Feng et al., 2023). Our focus is a reconstruction

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

mechanism that makes the different corruption signatures operational under both modality and statistical heterogeneity.

FedEPR combines evidential learning (Sensoy et al., 2018) with class prototypes and conflict-aware routing. Each modality supplies belief masses and an uncertainty proxy derived from its accumulated Dirichlet evidence. Dempster–Shafer (DS) composition (Dempster, 1968; Shafer, 1976) exposes a conflict factor that determines how reconstruction uses the modalities. This role differs from centralized evidence fusion (Han et al., 2021): conflict guides the choice of supervision, rather than merely determining a combined prediction. Global prototypes calibrate client-specific anchors, a warm-up phase precedes correction, and effective evidence weights the aggregation of each modality separately.

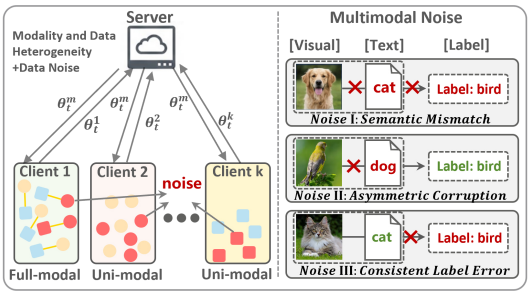

Figure 1: MMFL with heterogeneous clients and structured multimodal corruption.

Our main contributions are summarized as follows:

- We pioneer a systematic study of multimodal label noise in MMFL, formally defining three types of structured noise patterns and revealing their fundamental differences from unimodal noise.
- We propose an evidential learning-based multimodal federated framework that explicitly quantifies epistemic uncertainty through Dirichlet distributions, providing a principled foundation for noise identification.
- Extensive experiments on multiple benchmark datasets demonstrate that FedEPR significantly outperforms existing methods under various noise settings.

2 Related Work

2.1 Multimodal Federated Learning

MMFL addresses heterogeneous modality sets and distributions across clients (Lin et al., 2023; Xiong et al., 2022). FedMSplit adapts correlations across split multimodal networks (Chen & Zhang, 2022), CreamFL combines heterogeneous representations through contrastive learning (Yu et al., 2023), and modality-specific and modality-agnostic representations support cross-modal activity recognition (Yang et al., 2022). Hierarchical gradient blending addresses multimodal optimization under non-IID data by balancing gradient contributions (Chen & Li, 2022). PmcmFL uses prototype masking and contrasting (Bao et al., 2026); MPCPFL combines personalized aggregation with prototype-based missing-modality compensation (Yu et al., 2026). FedMobile reconstructs missing-modality features through cross-client knowledge distillation and evaluates contributions for aggregation (Liu et al., 2025).

These methods tackle complementary problems of representation sharing, incomplete inputs, and client-specific collaboration. The FedMultimodal benchmark further examines the behavior of federated learners under imperfect multimodal data (Feng et al., 2023). Modality availability and supervision reliability require different decisions. When an input is absent, the learner must determine how to train and aggregate the available representations. When both inputs are present, it must also determine whether their semantic evidence supports the observed label. A representation shared across clients can still be trained against an incorrect target, and agreement between modalities does not by itself validate that target. This distinction also changes the role of prototypes: beyond sharing class information or compensating for missing inputs, FedEPR uses globally calibrated class anchors to guide reconstruction decisions. Its combination of cross-modal conflict, evidential reliability, and prototype alignment connects representation sharing to the selection of training supervision under client heterogeneity. Aggregation then treats each modality separately, allowing clients with different modality sets to contribute to the branches they possess.

2.2 NOISY-LABEL LEARNING AND EVIDENTIAL FUSION

Centralized noisy-label learning includes co-teaching (Han et al., 2018), semi-supervised sample selection (Li et al., 2020a), contrastive learning (Li et al., 2022), and robust losses (Wang et al., 2019; Ma et al., 2020). Federated methods additionally account for client heterogeneity. FedCorr performs staged label correction (Xu et al., 2022); FedNoRo addresses class imbalance and heterogeneous noise (Wu et al., 2023); FedLN introduces corrections at different stages of federated training (Tsouvalas et al., 2024); and RHFL and FedDiv respectively use robust collaborative learning and collaborative filtering (Fang & Ye, 2022; Li et al., 2024). Our comparisons adapt noisy-label methods to modality branches while preserving their aggregation rules. Evidential deep learning parameterizes a Dirichlet distribution with nonnegative evidence (Sensoy et al., 2018). Trusted Multi-View Classification combines view-specific evidence using DS fusion (Han et al., 2021). FedEPR builds on this established representation and fusion rule. Its distinguishing mechanism uses conflict to route label reconstruction, with globally calibrated class anchors and modality-specific aggregation under federation. The uncertainty quantity is a reliability proxy; neither the Dirichlet parameterization nor DS fusion alone certifies prediction correctness.

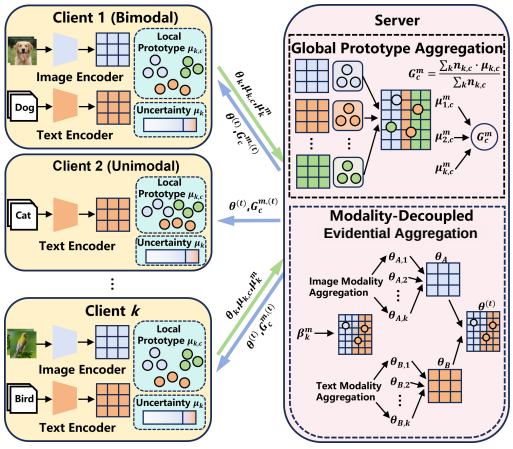

Figure 2: FedEPR server–client overview. Prototypes and modality models are aggregated separately.

3 PROPOSED METHOD

3.1 PROBLEM FORMULATION AND TRAINING CYCLE

A server coordinates K clients. Client k holds $\mathcal{D}_k = \{(\mathbf{x}_{k,i}, \tilde{y}_{k,i})\}_{i=1}^{N_k}$, where $\mathbf{x}_{k,i} = \{\mathbf{x}_{k,i}^m\}_{m\in\mathcal{M}_k}$ and $\mathcal{M}_k \subseteq \{A, B\}$. The observed label $\tilde{y}_{k,i}$ may differ from the latent clean class $y_{k,i} \in \{1, \dots, C\}$. The ideal clean-data objective is

$$\min_{\Theta} F(\Theta) = \sum_{k=1}^{K} \frac{N_k}{N} \mathbb{E}_{(\mathbf{x},y)\sim P_k} \ell(f_\Theta(\mathbf{x}), y), \qquad N = \sum_k N_k. \tag{1}$$

Training instead observes corruption that can depend on the modalities, $\tilde{y} \sim T(\cdot \mid \mathbf{x}^A, \mathbf{x}^B, y)$, together with possible input corruption. FedEPR constructs a target y_i' from the observed label, modality evidence, and calibrated prototypes. Semantic mismatch is disagreement between modal content; asymmetric corruption weakens one modality; consistent label error changes the annotation while the modalities remain coherent. They describe interpretable corruption patterns rather than an exhaustive noise model.

Each round broadcasts model parameters and modality-specific global prototypes. Clients compute evidence, calibrate their class prototypes, and reconstruct targets after warm-up. Local updates use an evidential objective, after which the server aggregates both class prototypes and modality-specific models. Figures 2 and 3 show the server–client overview and local sequence; Appendix D gives the full pseudocode.

3.2 EVIDENTIAL UNCERTAINTY AND PROTOTYPE CALIBRATION

For each available modality m, an encoder produces $\mathbf{z}_i^m \in \mathbb{R}^d$, and a Softplus evidential head outputs $\mathbf{e}_i^m \in \mathbb{R}_{\geq 0}^C$. Evidence parameterizes a Dirichlet distribution over class probabilities (Sensoy et al., 2018):

$$\boldsymbol{\alpha}_i^m = \mathbf{e}_i^m + \mathbf{1}, \qquad S_i^m = \sum_{c=1}^{C} \alpha_{i,c}^m, \qquad b_{i,c}^m = \frac{e_{i,c}^m}{S_i^m}, \qquad u_i^m = \frac{C}{S_i^m}. \tag{2}$$

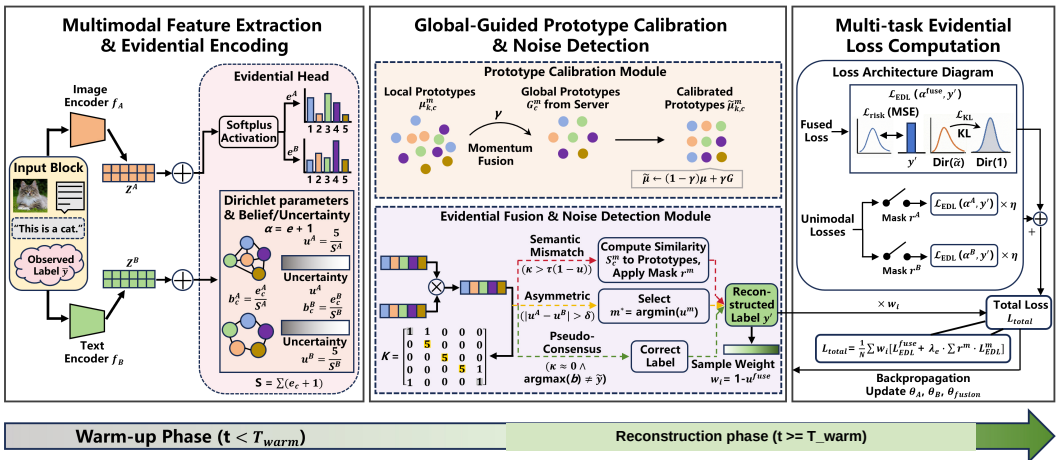

Figure 3: Local FedEPR pipeline. Modality-specific features produce Dirichlet evidence; calibrated class prototypes and modal conflict guide the reconstructed target. The objective includes a fused loss and masked modality losses. Warm-up uses observed labels for $t < T_{\text{warm}}$; reconstruction starts at $t \geq T_{\text{warm}}$. Equations (5)–(10) specify the selection and masking operations.

The masses satisfy $\sum_c b_{i,c}^m + u_i^m = 1$. Unlike relative softmax scores, u_i^m decreases with total evidence. We use it as an evidential reliability proxy; low uncertainty alone does not imply a correct class. The predictive probability is $\alpha_{i,c}^m / S_i^m = b_{i,c}^m + u_i^m / C$.

Class prototypes give these predictions a geometric reference. Local prototypes $\boldsymbol{\mu}_{k,c}^{m,(t)}$ are class means of the corresponding features. With $n_{k,c}^{(t)}$ samples assigned to class c, the server computes

$$\mathbf{G}_c^{m,(t)} = \frac{\sum_{k \in \mathcal{S}_m^{(t)}} n_{k,c}^{(t)} \boldsymbol{\mu}_{k,c}^{m,(t)}}{\sum_{k \in \mathcal{S}_m^{(t)}} n_{k,c}^{(t)}}, \qquad \tilde{\boldsymbol{\mu}}_{k,c}^{m,(t)} = (1-\gamma)\boldsymbol{\mu}_{k,c}^{m,(t)} + \gamma \mathbf{G}_c^{m,(t)}. \tag{3}$$

Here $\mathcal{S}_m^{(t)}$ contains the participating clients with modality m, and the class sum uses clients with nonzero class counts. The round index refers to the prototype state available for calibration. Warm-up uses observed assignments, while reconstruction supplies updated targets thereafter. Calibration anchors client-specific class centers to shared semantics without discarding local structure. The default $\gamma = 0.5$ is supported by the sweep in Appendix C.15; a convexity bound alone does not determine the optimal mixture.

3.3 Conflict-aware label reconstruction

For a bimodal sample, DS composition combines belief and uncertainty (Dempster, 1968; Shafer, 1976):

$$\kappa_i = \sum_{c \neq d} b_{i,c}^A b_{i,d}^B, \qquad b_{i,c}^{\text{fuse}} = \frac{b_{i,c}^A b_{i,c}^B + b_{i,c}^A u_i^B + b_{i,c}^B u_i^A}{1 - \kappa_i}, \qquad u_i^{\text{fuse}} = \frac{u_i^A u_i^B}{1 - \kappa_i}. \tag{4}$$

Large κ_i indicates committed belief on different classes. Small conflict can reflect either agreement or weak evidence, so conflict must be interpreted together with uncertainty. DS is used here as an evidence-composition operator, not an assertion that the fused belief is the joint Bayesian posterior. Appendix A relates it to product fusion and gives conditional reliability arguments.

Two selection rules serve different branches. Prototype-guided selection combines the strongest class alignment with evidential reliability, whereas uncertainty-guided selection chooses the less uncertain modality:

$$\begin{aligned} s_{i,c}^m &= \cos(\mathbf{z}_i^m, \tilde{\boldsymbol{\mu}}_{k,c}^{m,(t)}), \qquad m_i^{\text{proto}} = \arg\max_{m \in \{A,B\}} \{\max_c s_{i,c}^m - \lambda_u u_i^m\}, \\ m_i^{\text{unc}} &= \arg\min_{m \in \{A,B\}} u_i^m, \qquad h_i = \mathbb{I}\{\kappa_i > \tau(1 - u_i^{\text{fuse}})\}. \end{aligned} \tag{5}$$

We use $\lambda_u = 0.5$. The high-conflict test scales the threshold with committed fused mass because attainable conflict decreases when a modality supplies little belief. It is computed once from the current evidence. Substituting u_i^{fuse} yields $\kappa_i^2 - (1+\tau)\kappa_i + \tau(1 - u_i^A u_i^B) < 0$, so no recursive threshold computation is needed.

The following cases are evaluated in order, with the first applicable branch taking precedence:

$$y_i' = \begin{cases} \arg\max_c b_{i,c}^{m_i^{\text{proto}}}, & h_i = 1, \\ \arg\max_c b_{i,c}^{m_i^{\text{unc}}}, & h_i = 0,\ |u_i^A - u_i^B| > \delta, \\ \arg\max_c b_{i,c}^{\text{fuse}}, & h_i = 0,\ |u_i^A - u_i^B| \le \delta,\ u_i^{\text{fuse}} < \epsilon, \\ & \quad \kappa_i < \tau,\ \arg\max_c b_{i,c}^{\text{fuse}} \neq \tilde{y}_i, \\ \tilde{y}_i, & \text{otherwise.} \end{cases} \tag{6}$$

High-conflict samples follow prototype-guided selection. The second branch trusts the less uncertain modality. The third changes a label when low-conflict, low-uncertainty consensus contradicts it. Otherwise the observed label is retained. Ordered evaluation also resolves the overlap $\tau(1 - u_i^{\text{fuse}}) < \kappa_i < \tau$ in favor of the conflict branch. For unimodal clients, reconstruction instead requires low uncertainty and weak alignment with the observed class:

$$y_i' = \begin{cases} \arg\max_c s_{i,c}^m, & s_{i,\tilde{y}_i}^m < \rho \text{ and } u_i^m < \epsilon, \\ \tilde{y}_i, & \text{otherwise.} \end{cases} \tag{7}$$

3.4 Evidential Training and Federated Aggregation

The fused masses determine the Dirichlet parameters used by the loss:

$$S_i^{\text{fuse}} = \frac{C}{u_i^{\text{fuse}}}, \qquad \alpha_{i,c}^{\text{fuse}} = 1 + b_{i,c}^{\text{fuse}} S_i^{\text{fuse}}. \tag{8}$$

For a one-hot target $\mathbf{y}_i'$, define the masked parameters $\tilde{\boldsymbol{\alpha}}_i = \mathbf{y}_i' + (\mathbf{1} - \mathbf{y}_i') \odot \boldsymbol{\alpha}_i$. The EDL loss combines the expected squared prediction error with regularization of non-target evidence:

$$\begin{aligned} \mathcal{L}_{\text{EDL}}(\boldsymbol{\alpha}_i, \mathbf{y}_i') = \sum_{c=1}^{C} \left[\left(y_{i,c}' - \frac{\alpha_{i,c}}{S_i} \right)^2 + \frac{\alpha_{i,c}(S_i - \alpha_{i,c})}{S_i^2 (S_i + 1)} \right] \\ + \lambda_{\text{KL}}\, \text{KL}[\text{Dir}(\mathbf{p} \mid \tilde{\boldsymbol{\alpha}}_i) \,\|\, \text{Dir}(\mathbf{p} \mid \mathbf{1})]\,. \end{aligned} \tag{9}$$

Thus $\mathcal{L}_i^m = \mathcal{L}_{\text{EDL}}(\boldsymbol{\alpha}_i^m, \mathbf{y}_i')$ and $\mathcal{L}_i^{\text{fuse}} = \mathcal{L}_{\text{EDL}}(\boldsymbol{\alpha}_i^{\text{fuse}}, \mathbf{y}_i')$. For unimodal clients, set $h_i = 0$ and use the available modality as the primary prediction. Auxiliary modality losses use

$$r_i^m = \begin{cases} \mathbb{I}\{m = m_i^{\text{proto}}\}, & h_i = 1, \\ \mathbb{I}\{m \in \mathcal{M}_k\}, & \text{otherwise,} \end{cases} \qquad \mathcal{L}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} w_i \left[\mathcal{L}_i^{\text{fuse}} + \lambda_e \sum_{m \in \mathcal{M}_k} r_i^m \mathcal{L}_i^m \right], \tag{10}$$

where $w_i = 1 - u_i^{\text{fuse}}$ for bimodal samples and $w_i = 1 - u_i^m$ for unimodal samples. During warm-up, targets remain $\tilde{y}_i$ and no conflict selection mask is applied. Reconstruction starts at round T_{warm}. The mask gates auxiliary losses only: the fused loss still updates both encoders on a high-conflict sample. Appendix C.16 quantifies this coupling without changing the main objective.

The server aggregates each modality only across participating clients that possess it. Client k uploads its local model, class prototypes and counts, and accumulated effective evidence:

$$\bar{e}_k^m = \sum_{i \in \mathcal{D}_k} (1 - u_{k,i}^m), \qquad \beta_k^m = \frac{\bar{e}_k^m}{\sum_{j \in \mathcal{S}_m} \bar{e}_j^m}, \qquad \theta_m^{(t+1)} = \sum_{k \in \mathcal{S}_m} \beta_k^m \theta_{m,k}^{(t+1)}. \tag{11}$$

The weights act as reliability-adjusted sample sizes. Inverse-variance weighting provides a conditional motivation, supported by the measured negative correlations between evidence, gradient variance, and residual error in Appendix C.18. Confidence can nevertheless favor a consistently wrong client. Raw inputs and per-sample evidence remain local; shared prototypes and model updates do not by themselves provide a formal privacy guarantee.

3.5 WHY RECONSTRUCTION CAN IMPROVE SUPERVISION

The reconstruction rule balances correction of noisy labels against damage to clean ones. Let $q_i = \Pr(\tilde{y}_i \neq y_i)$, let a_i be the probability of making a clean annotation wrong, and let b_i be the probability that an initially wrong annotation remains wrong after reconstruction. The residual error is

$$\pi_i = (1 - q_i)a_i + q_i b_i. \tag{12}$$

For $0 < q_i < 1$, reconstruction improves label quality when $a_i < q_i(1 - b_i)/(1 - q_i)$. This condition explains the different branches: disagreement prompts a reliability comparison, a large uncertainty gap identifies an informative source, and consensus correction requires low uncertainty. The fallback protects examples without sufficiently strong signals. The condition also explains why maximizing the number of changed labels is not the objective.

Prototype calibration affects this decision through its semantic reference. Let a and b denote the local and global prototype errors relative to a clean class center, with $A = \mathbb{E}\|a\|^2$, $B = \mathbb{E}\|b\|^2$, and $D = \mathbb{E}\langle a, b\rangle$. The calibrated error is

$$\mathbb{E}\|(1 - \gamma)a + \gamma b\|^2 = (1 - \gamma)^2 A + 2\gamma(1 - \gamma)D + \gamma^2 B. \tag{13}$$

When $A + B - 2D > 0$, the optimal mixture is the projection of $(A - D)/(A + B - 2D)$ onto $[0, 1]$. An interior mixture can therefore improve on either anchor when their errors are not perfectly aligned. The pure-global endpoint must be evaluated empirically; minimizing a loose convexity bound cannot establish its optimality. Likewise, evidence-weighted aggregation minimizes a conditional variance bound when client update variance is inversely related to effective evidence. Its benefit depends on informative reliability estimates. Appendix A states the separation and variance conditions, derives the exact prototype mixing error, and gives a stationarity bound that retains reconstruction and aggregation bias.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate three bimodal benchmarks. AV-MNIST combines digit images and spoken-digit spectrograms with ten classes (Vielzeuf et al., 2018; LeCun et al., 1998; Jackson et al., 2018). UPMC Food-101 pairs food images with recipe text across 101 classes (Wang et al., 2015). Controlled injection uses the filtered set of approximately 73,500 samples; natural-noise evaluation uses approximately 90,840 raw web-collected pairs. CMU-MOSEI provides approximately 23,500 video clips with seven-class sentiment labels; we use language and acoustic features (Bagher Zadeh et al., 2018).

Federated and Noise Protocol. The default federation has ten clients, Dirichlet class partitioning with $\alpha = 0.5$ (Hsu et al., 2019), and 30% unimodal clients. Structured corruption affects 20%, 40%, or 60% of training samples, with a 40/30/30 mixture of semantic mismatch, asymmetric corruption, and consistent label error. Client rates vary within ten percentage points of the nominal rate. Mismatch replaces one input with a different-class sample, asymmetric corruption adds Gaussian noise to one modality, and consistent error flips the label while leaving inputs unchanged. These tests isolate the designed corruption patterns; additional processes and naturally noisy data test a broader range of conditions.

Implementation. Training uses 100 rounds, 20 warm-up rounds, two local epochs per round, batch size 32, SGD with learning rate 0.01 and momentum 0.9, and 128-dimensional embeddings. AV-MNIST uses three-layer CNNs, Food-101 uses ResNet-18 (He et al., 2016) and a two-layer LSTM, and CMU-MOSEI uses an acoustic CNN and frozen BERT (Devlin et al., 2019) with a trainable projection. Defaults are $\lambda_{\mathrm{KL}} = 0.1$, $\lambda_e = \lambda_u = \gamma = \tau = \rho = 0.5$, $\delta = 0.3$, and $\epsilon = 0.2$. Appendix B details the protocol.

Baselines. Comparisons include FedAvg (McMahan et al., 2017), FedProx (Li et al., 2020b), FedMSplit, CreamFL, PmcmFL, MPCPFL, and FedMobile. All methods share partitions, corrupted samples, seeds, rounds, local epochs, and validation protocols, with comparable search ranges and the same number of tuning trials. All experiments use three random seeds; we report means and standard deviations. Comparison tables mark the top three results with the colors defined in Table 1, accounting for each metric's direction. Additional adapted noisy-label FL baselines appear in Appendix C.8.

Table 1: Test accuracy (%) under structured noise, with Dirichlet $\alpha = 0.5$ and 30% unimodal clients. Entries show mean$_{\pm \text{std}}$ over three seeds. First, second, and third mark column ranks; the best mean is bold.

	AV-MNIST			UPMC Food-101			CMU-MOSEI		
Noise rate	20%	40%	60%	20%	40%	60%	20%	40%	60%
FedAvg	$89.34_{\pm 0.28}$	$81.67_{\pm 0.41}$	$70.42_{\pm 0.58}$	$74.23_{\pm 0.32}$	$64.85_{\pm 0.47}$	$53.62_{\pm 0.69}$	$42.18_{\pm 0.45}$	$35.24_{\pm 0.52}$	$28.37_{\pm 0.64}$
FedProx	$90.18_{\pm 0.25}$	$82.45_{\pm 0.37}$	$71.83_{\pm 0.52}$	$75.47_{\pm 0.29}$	$65.93_{\pm 0.43}$	$54.86_{\pm 0.61}$	$43.05_{\pm 0.41}$	$36.17_{\pm 0.48}$	$29.24_{\pm 0.59}$
FedMSplit	$91.72_{\pm 0.22}$	$84.53_{\pm 0.33}$	$74.26_{\pm 0.47}$	$77.38_{\pm 0.26}$	$68.52_{\pm 0.38}$	$58.17_{\pm 0.54}$	$45.34_{\pm 0.38}$	$38.46_{\pm 0.44}$	$31.58_{\pm 0.55}$
CreamFL	$92.45_{\pm 0.20}$	$86.18_{\pm 0.30}$	$78.35_{\pm 0.43}$	$78.65_{\pm 0.24}$	$70.74_{\pm 0.35}$	$61.43_{\pm 0.49}$	$46.82_{\pm 0.35}$	$40.15_{\pm 0.41}$	$33.27_{\pm 0.51}$
PmcmFL	$93.67_{\pm 0.18}$	$87.24_{\pm 0.27}$	$77.18_{\pm 0.41}$	$80.27_{\pm 0.22}$	$73.28_{\pm 0.32}$	$64.52_{\pm 0.46}$	$48.15_{\pm 0.32}$	$41.37_{\pm 0.38}$	$34.28_{\pm 0.48}$
MPCPFL	$94.38_{\pm 0.16}$	$88.56_{\pm 0.24}$	$77.92_{\pm 0.39}$	$81.92_{\pm 0.20}$	$72.65_{\pm 0.30}$	$66.83_{\pm 0.43}$	$48.73_{\pm 0.30}$	$42.28_{\pm 0.36}$	$35.12_{\pm 0.45}$
FedMobile	$93.85_{\pm 0.17}$	$89.17_{\pm 0.23}$	$79.64_{\pm 0.37}$	$80.73_{\pm 0.21}$	$74.56_{\pm 0.29}$	$68.25_{\pm 0.41}$	$47.86_{\pm 0.31}$	$42.63_{\pm 0.35}$	$35.74_{\pm 0.44}$
FedEPR	$\mathbf{98.23}_{\pm 0.12}$	$\mathbf{95.47}_{\pm 0.17}$	$\mathbf{92.18}_{\pm 0.24}$	$\mathbf{89.17}_{\pm 0.15}$	$\mathbf{85.34}_{\pm 0.21}$	$\mathbf{80.26}_{\pm 0.32}$	$\mathbf{53.28}_{\pm 0.26}$	$\mathbf{46.15}_{\pm 0.31}$	$\mathbf{40.37}_{\pm 0.39}$

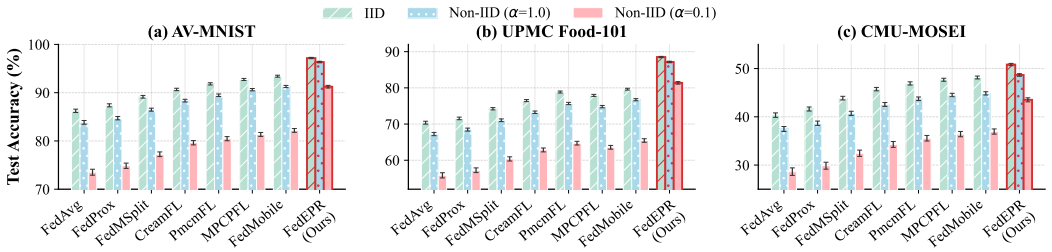

Figure 4: Accuracy under IID, moderately non-IID ($\alpha = 1.0$), and severely non-IID ($\alpha = 0.1$) partitions at 40% structured corruption. Panels show AV-MNIST, Food-101, and CMU-MOSEI; error bars show standard deviations over three seeds.

4.2 Controlled and Naturally Occurring Noise

Table 1 shows consistent gains across the controlled settings. At 60% corruption, FedEPR reaches 92.18% on AV-MNIST and 80.26% on Food-101, improving on FedMobile by 12.54 and 12.01 percentage points. On CMU-MOSEI, the corresponding gain is 4.63 points. The improvement is larger at severe corruption on the image-based tasks, while sentiment remains more difficult. These measurements describe the structured generator, whose cases overlap with the reconstruction rules.

Figure 4 varies the class distribution. FedEPR reaches 91.24% on AV-MNIST and 81.38% on Food-101 at $\alpha = 0.1$. Global calibration supplies anchors when local class centers are poorly estimated, while reconstruction uses modal reliability to limit misleading supervision. This complementary role is tested by removing calibration and evidential learning separately below.

Naturally noisy evaluation gives a distinct estimate of performance. On raw UPMC data, FedEPR improves from the strongest baseline's 84.21% to 88.47%; on the low-agreement CMU subset it improves from 39.83% to 44.21%. No synthetic corruption is added. A separate audit of 1,500 UPMC training triplets evaluates correction and detection without exposing audit labels to training. WebVision Google500 provides an additional naturally noisy image–text setting: FedEPR reaches 68.61% IID and 64.52% non-IID top-1 accuracy, improving on FedMobile by 3.23 and 3.78 points. Full protocols and results are in Appendices C.3 and C.12. These smaller gains distinguish natural-noise performance from the controlled stress tests.

Table 2: Natural-noise accuracy (%, mean ± std., three seeds). UPMC uses raw web pairs; CMU uses low-agreement labels.

Method	UPMC Raw	CMU Low-agr.
FedAvg	78.64 ± 0.36	34.18 ± 0.48
FedProx	79.42 ± 0.33	35.02 ± 0.45
CreamFL	82.73 ± 0.28	38.24 ± 0.39
MPCPFL	83.56 ± 0.25	39.06 ± 0.37
FedMobile	84.21 ± 0.24	39.83 ± 0.35
FedEPR	$\mathbf{88.47} \pm 0.19$	$\mathbf{44.21} \pm 0.31$

The extra noise processes include instance-dependent flips, adjacent-class flips, realistic input degradation, and simultaneous degradation of both modalities. FedEPR remains ahead of the compared

Table 3: Component ablations at 40% structured noise and $\alpha = 0.5$. Mean $\pm$ standard deviation over three runs. EL: evidential learning; GPC: global prototype calibration; DR: dynamic reconstruction.

Variant	AV-MNIST	Food-101	CMU-MOSEI
FedAvg	81.67 ± 0.41	64.85 ± 0.47	35.24 ± 0.52
w/o EL	88.35 ± 0.28	76.48 ± 0.34	40.72 ± 0.41
w/o GPC	90.26 ± 0.24	78.65 ± 0.30	42.18 ± 0.38
w/o DR	91.82 ± 0.22	80.47 ± 0.28	43.56 ± 0.35
w/o EL & GPC	85.74 ± 0.33	72.36 ± 0.39	38.45 ± 0.46
FedEPR	$\mathbf{95.47} \pm 0.17$	$\mathbf{85.34} \pm 0.21$	$\mathbf{46.15} \pm 0.31$

methods, although the processes still partly overlap with the corruption structures used for design. At 500 clients with about ten active clients per round, its AV-MNIST accuracy is 78.26% against FedAvg's 68.63%. This is a joint stress test of client sparsity and participation, rather than an isolated computational scaling experiment (Appendix C).

Under instance-dependent flips, the gains over FedMobile are 3.39 points on AV-MNIST and 6.71 on Food-101; with both modalities degraded, they are 3.52 and 6.39 points. AV-MNIST correction precision falls from 86.74% under single-modality degradation to 75.32% when both inputs degrade, showing the value of a reliable counterpart.

The UPMC audit uses two independent annotators and a third to resolve disagreements; agreement is approximately 0.82 for clean/noisy status and 0.74 for corruption type. FedEPR achieves 88.6% detection AUROC and 81.7% correction precision versus 72.6% and 62.8% for softmax confidence. Precision is 84.5% for semantic mismatch, 82.8% for asymmetric corruption, and 76.9% for consistent label error, where coherent inputs offer less evidence of annotation error. Audit labels are excluded from training, model selection, and reconstruction.

The missing-modality sweep tests the availability of a second source for cross-checking. At 70% unimodal clients and 40% structured corruption, FedEPR reaches 89.62%, 76.48%, and 39.56% on AV-MNIST, Food-101, and CMU-MOSEI, ahead of MPCPFL by 5.88, 8.66, and 4.32 points. Unimodal clients use Equation (7); shared anchors and modality-specific aggregation complement the checks available on bimodal clients. Table 14 gives the full sweep.

4.3 Component and Design Analysis

Removing evidential learning produces the largest single-component drop in Table 3: 7.12 points on AV-MNIST, 8.86 on Food-101, and 5.43 on CMU-MOSEI. Removing global prototype calibration lowers Food-101 accuracy by 6.69 points, and removing dynamic reconstruction lowers it by 4.87 points. The components address different sources of error: evidence supplies reliability estimates, prototypes stabilize semantic comparisons, and reconstruction changes the supervision used for training.

The fusion comparison in Table 4 separates evidential representation from routing. At 40% corruption, AV-MNIST accuracy rises from 88.35% with softmax averaging to 92.36% with evidential averaging, 93.71% with DS without routing, and 95.47% with FedEPR. Detection AUROC reaches 94.83%, and calibration error decreases to 2.86%. These results support the combined reconstruction signals rather than attributing all gains to DS fusion alone.

Figure 5 shows the representation change. At round 100, AV-MNIST prototype-to-centroid distance is 0.298 for FedAvg, 0.172 without global calibration, and 0.053 for FedEPR. Clean centroids are used only for this diagnostic. From rounds 20 to 100, detection precision rises from 78.2% to 91.5% while prototype distance falls from 0.156 to 0.053, linking improved alignment to more reliable correction (Appendix C).

Figure 6 shows the interaction of the conflict threshold and KL regularization. The best tested pair is $\tau = 0.5$, $\lambda_{\mathrm{KL}} = 0.1$ on both datasets. The separate prototype sweep finds $\gamma = 0.5$ best on all three benchmarks; using only global prototypes ($\gamma = 1$) lowers accuracy by 1.09, 1.52, and 1.19 points. Thus a mixture is useful under the tested non-IID setting. A fixed conflict threshold yields 94.86% and 84.73% on AV-MNIST and Food-101, compared with 95.47% and 85.34% for the adaptive rule.

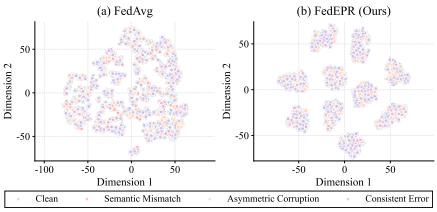

Figure 5: t-SNE on AV-MNIST at 40% corruption. FedEPR learns more compact clusters than FedAvg. Colors denote clean samples and the three corruption types.

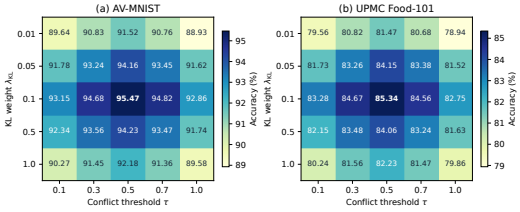

Figure 6: Sensitivity to conflict threshold τ and KL weight λ_{KL} at 40% corruption and $\alpha = 0.5$. Cells show mean accuracy over three seeds.

Table 4: Uncertainty and fusion analysis at 40% structured corruption and $\alpha = 0.5$ (mean$_{\pm\mathrm{std}}$, three seeds). Accuracy and AUROC are higher-is-better; ECE is lower-is-better. All metrics are percentages.

	AV-MNIST			UPMC Food-101			CMU-MOSEI		
Variant	Acc.↑	AUROC↑	ECE↓	Acc.↑	AUROC↑	ECE↓	Acc.↑	AUROC↑	ECE↓
Softmax Avg.	$88.35_{\pm 0.28}$	$82.41_{\pm 0.36}$	$8.74_{\pm 0.18}$	$76.48_{\pm 0.34}$	$75.63_{\pm 0.42}$	$10.92_{\pm 0.24}$	$40.72_{\pm 0.41}$	$70.86_{\pm 0.51}$	$13.47_{\pm 0.31}$
Softmax PoE	$89.62_{\pm 0.25}$	$83.17_{\pm 0.34}$	$8.21_{\pm 0.16}$	$77.65_{\pm 0.31}$	$76.42_{\pm 0.39}$	$10.35_{\pm 0.22}$	$41.26_{\pm 0.39}$	$71.58_{\pm 0.48}$	$12.96_{\pm 0.28}$
Evidential Avg.	$92.36_{\pm 0.21}$	$89.74_{\pm 0.27}$	$5.36_{\pm 0.12}$	$80.71_{\pm 0.28}$	$82.58_{\pm 0.33}$	$7.42_{\pm 0.18}$	$42.88_{\pm 0.36}$	$76.94_{\pm 0.43}$	$10.81_{\pm 0.25}$
Evidential PoE	$93.04_{\pm 0.20}$	$90.38_{\pm 0.25}$	$5.12_{\pm 0.11}$	$81.56_{\pm 0.26}$	$83.19_{\pm 0.31}$	$7.18_{\pm 0.17}$	$43.21_{\pm 0.34}$	$77.36_{\pm 0.41}$	$10.52_{\pm 0.24}$
DS w/o Routing	$93.71_{\pm 0.19}$	$91.46_{\pm 0.24}$	$4.61_{\pm 0.10}$	$82.64_{\pm 0.24}$	$84.77_{\pm 0.29}$	$6.53_{\pm 0.15}$	$44.07_{\pm 0.33}$	$78.82_{\pm 0.39}$	$9.74_{\pm 0.22}$
FedEPR	$\mathbf{95.47}_{\pm 0.17}$	$\mathbf{94.83}_{\pm 0.20}$	$\mathbf{2.86}_{\pm 0.08}$	$\mathbf{85.34}_{\pm 0.21}$	$\mathbf{88.91}_{\pm 0.25}$	$\mathbf{4.37}_{\pm 0.12}$	$\mathbf{46.15}_{\pm 0.31}$	$\mathbf{82.46}_{\pm 0.35}$	$\mathbf{7.28}_{\pm 0.19}$

Removing the fused loss for high-conflict samples changes accuracy by 0.21 points on AV-MNIST and 0.19 on Food-101. The original objective still updates non-selected modalities toward reconstructed targets. Across 17 client–modality pairs, accumulated evidence has Spearman correlations of -0.91 with gradient variance and -0.87 with residual label error; mean evidence correlates with variance at -0.83. These empirical diagnostics and their scope are reported in Appendix C.

4.4 Efficiency and Privacy Scope

Table 5: Per-round overhead on AV-MNIST. Ratio refers to communication.

Method	MB	Seconds	Ratio
FedAvg	42.6	12.4	1.00×
FedProx	42.6	13.1	1.00×
FedMSplit	48.3	18.7	1.13×
CreamFL	54.2	24.5	1.27×
PmcmFL	43.8	15.6	1.03×
MPCPFL	44.5	16.8	1.04×
FedMobile	51.7	28.3	1.21×
FedEPR	43.4	14.8	1.02×

On AV-MNIST, FedEPR exchanges 43.4 MB per round versus FedAvg's 42.6 MB, a communication ratio of 1.02. Round time is 14.8 versus 12.4 seconds, approximately 1.19 times as large. Additional uploads consist of class prototypes, counts, and one accumulated evidence statistic per available modality; individual uncertainty values and raw inputs stay on clients. Table 5 compares all baselines; Appendix B.1 gives complexity estimates. Gaussian perturbation of the additional statistics gives 93.26% accuracy at a reported budget of $\varepsilon_{\mathrm{DP}} = 5$, compared with 95.47% without perturbation. This experiment protects the extra statistics only; model updates are outside its accountant (Appendix C.13). At $\varepsilon_{\mathrm{DP}} = 1$, accuracy falls to 89.41%, below the 90.26% of the variant without global calibration.

5 Conclusion

FedEPR reconstructs supervision in bimodal federated learning using evidential reliability, modal conflict, and calibrated prototypes. Its ordered rule separates conflict, unequal reliability, and disagreement with the observed label. Gains are larger under controlled corruption and smaller but consistent on naturally noisy UPMC and WebVision. Diagnostics support the design while identifying limitations: simultaneous modality degradation, confidently incorrect clients, fused-loss gradient coupling, and extension beyond two modalities.

AI USE STATEMENT

Generative AI tools assisted with language polishing, literature search and verification, and revision of mathematical derivations and proofs. The authors take responsibility for the final content of the manuscript.

ETHICS STATEMENT

The experiments use established research datasets. A human audit of 1,500 UPMC Food-101 training triplets assesses annotation and alignment errors; these labels are kept separate from training and from the test set. Raw observations remain local in the federated protocol, but model updates and class prototypes can disclose information. The statistical perturbation experiment does not establish end-to-end differential privacy. Evidence-weighted aggregation can also favor confidently incorrect clients, which is relevant when the method is considered for sensitive applications.

REPRODUCIBILITY STATEMENT

Section 3 specifies reconstruction, losses, and aggregation. Appendix B describes data preparation and the shared comparison protocol; Appendix A states the conditions used in the analysis. Appendix C provides the complete result tables and diagnostic settings, and Appendix D provides round-level pseudocode.

REFERENCES

Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. Multimodal biomedical AI. *Nature Medicine*, 28(9):1773–1784, 2022. doi: 10.1038/s41591-022-01981-2.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, 2018. doi: 10.18653/v1/P18-1208. URL https://aclanthology.org/P18-1208/.

Guangyin Bao, Qi Zhang, Duoqian Miao, Zixuan Gong, Chaochao Chen, Liang Hu, and Longbing Cao. Incomplete Multimodal Federated Learning via Masking and Contrasting Prototypes. *IEEE Transactions on Neural Networks and Learning Systems*, 37(9):4050–4062, 2026. doi: 10.1109/TNNLS.2026.3658522.

Jiayi Chen and Aidong Zhang. FedMSplit: Correlation-Adaptive Federated Multi-Task Learning Across Multimodal Split Networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 87–96, 2022. doi: 10.1145/3534678.3539384.

Jieying Chen and Rong Pan. Medical report generation based on multimodal federated learning. *Computerized Medical Imaging and Graphics*, 113:102342, 2024. doi: 10.1016/j.compmedimag.2024.102342.

Sijia Chen and Baochun Li. Towards optimal multi-modal federated learning on non-iid data with hierarchical gradient blending. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pp. 1469–1478. IEEE, 2022. doi: 10.1109/INFOCOM48880.2022.9796724.

Arthur P Dempster. A generalization of Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232, 1968. doi: 10.1111/j.2517-6161.1968.tb00722.x.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Xiuwen Fang and Mang Ye. Robust federated learning with noisy and heterogeneous clients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10072–10081, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Fang_Robust_Federated_Learning_With_Noisy_and_Heterogeneous_Clients_CVPR_2022_paper.html.

Tiantian Feng, Digbalay Bose, Tuo Zhang, Rajat Hebbar, Anil Ramakrishna, Rahul Gupta, Mi Zhang, Salman Avestimehr, and Shrikanth Narayanan. FedMultimodal: A Benchmark for Multimodal Federated Learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4035–4045. ACM, 2023. doi: 10.1145/3580305.3599825.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/a19744e268754fb0148b017647355b7b-Abstract.html.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted Multi-View Classification. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=OOsR8BzCnl5.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019. URL https://arxiv.org/abs/1909.06335.

Zohar Jackson, César Souza, Jason Flaks, Yuxin Pan, Hereman Nicolas, and Adhish Thite. Jakobovski/free-spoken-digit-dataset: v1.0.8, 2018. URL https://zenodo.org/records/1342401. Version 1.0.8.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Jichang Li, Guanbin Li, Hui Cheng, Zicheng Liao, and Yizhou Yu. FedDiv: Collaborative noise filtering for federated learning with noisy labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3118–3126, 2024. doi: 10.1609/aaai.v38i4.28095.

Junnan Li, Richard Socher, and Steven C. H. Hoi. DivideMix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020a. URL https://openreview.net/forum?id=HJgExaVtwr.

Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. Selective-supervised contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 316–325, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Li_Selective-Supervised_Contrastive_Learning_With_Noisy_Labels_CVPR_2022_paper.html.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pp. 429–450, 2020b. URL https://proceedings.mlsys.org/paper/2020/hash/1f5fe83998a09396ebe6477d9475ba0c-Abstract.html.

Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. WebVision Database: Visual Learning and Understanding from Web Data. *arXiv preprint arXiv:1708.02862*, 2017. URL https://arxiv.org/abs/1708.02862.

Yi-Ming Lin, Yuan Gao, Mao-Guo Gong, Si-Jia Zhang, Yuan-Qiao Zhang, and Zhi-Yuan Li. Federated learning on multimodal data: A comprehensive survey. *Machine Intelligence Research*, 20(4): 539–553, 2023. doi: 10.1007/s11633-022-1398-0.

Yi Liu, Cong Wang, and Xingliang Yuan. FedMobile: Enabling Knowledge Contribution-aware Multi-modal Federated Learning with Incomplete Modalities. In *Proceedings of the ACM on Web Conference 2025*, pp. 2775–2786, 2025. doi: 10.1145/3696410.3714623.

Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6543–6553. PMLR, 2020. URL https://proceedings.mlr.press/v119/ma20c.html.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Yulei Qin, Xingyu Chen, Yunhang Shen, Chaoyou Fu, Yun Gu, Ke Li, Xing Sun, and Rongrong Ji. CAPro: Webly Supervised Learning with Cross-modality Aligned Prototypes. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53565–53592, 2023. doi: 10.52202/075280-2331. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a7e0d77325db843fd5baf1298163e89a-Abstract-Conference.html.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/a981f2b708044d6fb4a71a1463242520-Abstract.html.

Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976. doi: 10.1515/9780691214696.

Vasileios Tsouvalas, Aaqib Saeed, Tanir Ozcelebi, and Nirvana Meratnia. Labeling Chaos to Learning Harmony: Federated Learning with Noisy Labels. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–26, 2024. doi: 10.1145/3626242.

Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. CentralNet: A Multilayer Approach for Multimodal Fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 575–589, 2018. URL https://openaccess.thecvf.com/content_ECCVW_2018/papers/11134/Vielzeuf_CentralNet_a_Multilayer_Approach_for_Multimodal_Fusion_ECCVW_2018_paper.pdf.

Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1–6. IEEE, 2015. doi: 10.1109/ICMEW.2015.7169757.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 322–330, 2019. doi: 10.1109/ICCV.2019.00041.

Nannan Wu, Li Yu, Xuefeng Jiang, Kwang-Ting Cheng, and Zengqiang Yan. FedNoRo: Towards noise-robust federated learning by addressing class imbalance and label noise heterogeneity. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 4424–4432, 2023. doi: 10.24963/ijcai.2023/492. URL https://www.ijcai.org/proceedings/2023/492.

Baochen Xiong, Xiaoshan Yang, Fan Qi, and Changsheng Xu. A unified framework for multi-modal federated learning. *Neurocomputing*, 480:110–118, 2022. doi: 10.1016/j.neucom.2022.01.063.

Jingyi Xu, Zihan Chen, Tony Q. S. Quek, and Kai Fong Ernest Chong. FedCorr: Multi-stage federated learning for label noise correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10184–10193, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Xu_FedCorr_Multi-Stage_Federated_Learning_for_Label_Noise_Correction_CVPR_2022_paper.html.

Xiaoshan Yang, Baochen Xiong, Yi Huang, and Changsheng Xu. Cross-modal federated human activity recognition via modality-agnostic and modality-specific representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3063–3071, 2022. doi: 10.1609/aaai.v36i3.20213.

Qiwei Yu, Hongjian Shi, Zhuolong Jiang, Yuhan Qiu, Ruhui Ma, Shiliang Liao, Xing Zhou, and Haibing Guan. MPCPFL: Prototype-Based Multimodal Personalized Federated Learning With Missing Modalities in Medical Domain. *IEEE Transactions on Computational Social Systems*, 13 (4):5359–5373, 2026. doi: 10.1109/TCSS.2025.3638431.

Qiying Yu, Yang Liu, Yimu Wang, Ke Xu, and Jingjing Liu. Multimodal federated learning via contrastive representation ensemble. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2302.08888.

Tianyue Zheng, Ang Li, Zhe Chen, Hongbo Wang, and Jun Luo. AutoFed: Heterogeneity-Aware Federated Multimodal Learning for Robust Autonomous Driving. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, pp. 1–15, 2023. doi: 10.1145/3570361.3592517.

APPENDIX

In this appendix, we provide additional theoretical, empirical, and implementation evidence for FedEPR. Table 6 summarizes the organization of the appendix and the main contents of each section.

Table 6: Roadmap of the appendix.

Appendix	**Main Contents**
Appendix A	• Signal consistency under structured noise (A.1). • Evidential uncertainty (A.2), conflict and the adaptive threshold (A.3), and relation to product fusion (A.4). • Prototype calibration (A.5) and prototype-guided routing (A.6). • Label reconstruction (A.7), evidence-weighted aggregation (A.8), and convergence (A.9).
Appendix B	• Computational and communication complexity (B.1), data partitions, and corruption protocols (B.2). • Architectures and training settings (B.3), baseline adaptation, and model selection (B.4).
Appendix C	• Complete controlled-noise results (C.1), fusion ablations (C.2), and natural noise and human audit (C.3). • Warm-up (C.4), noise composition (C.5), missing modalities (C.6), federation size, and per-class results (C.7). • Noisy-label baselines (C.8), detection and representations (C.9), and prototype stabilization and progressive detection (C.10). • Additional corruption processes (C.11), WebVision (C.12), statistics perturbation (C.13), and partial participation (C.14). • Prototype mixing (C.15), fused-loss coupling (C.16), conflict thresholds (C.17), and evidence correlations (C.18).
Appendix D	• Complete round-level training procedure: warm-up, evidential estimation, conflict-aware reconstruction, local updates, prototype aggregation, and modality-decoupled evidence-weighted aggregation.
Appendix E	• Limitations and broader scope: additional modalities, prototype shift, confidently incorrect clients, and privacy of released statistics.

A THEORETICAL ANALYSIS

The analysis explains the signals used for reconstruction and the conditions under which they are useful. DS composition is an evidence operator, and u is a reliability proxy. The statements below do not equate fused belief with the joint posterior. Table 7 summarizes the main notation.

Table 7: Notation used by FedEPR.

Symbol	Meaning
K, N_k, C, d	Clients, local samples, classes, embedding dimension
$\mathcal{M}_k, \mathcal{S}_m^{(t)}$	Available modalities; participating clients with modality m
$y_i, \tilde{y}_i, y_i'$	Latent clean, observed, and reconstructed labels
$\mathbf{z}_i^m, \mathbf{e}_i^m, \boldsymbol{\alpha}_i^m$	Features, nonnegative evidence, Dirichlet parameters
$S_i^m, b_{i,c}^m, u_i^m$	Dirichlet strength, class belief, uncertainty proxy
κ_i, h_i	Cross-modal conflict and high-conflict indicator
$\boldsymbol{\mu}_{k,c}^{m,(t)}, \mathbf{G}_c^{m,(t)}$	Local and global class prototypes
$\tilde{\boldsymbol{\mu}}_{k,c}^{m,(t)}, \gamma$	Calibrated prototype and global mixing weight
$m_i^{\text{proto}}, m_i^{\text{unc}}$	Prototype-guided and minimum-uncertainty modality choices
$\tau, \delta, \epsilon, \rho$	Conflict, uncertainty-gap, uncertainty, similarity thresholds
w_i, r_i^m	Sample reliability weight and auxiliary modality-loss mask
$\bar{e}_k^m, \beta_k^m$	Accumulated effective evidence and aggregation weight

A.1 Signal consistency for structured multimodal noise

Let $c_m = \arg\max_c b_c^m$. A reliable modality for class y satisfies

$$b_y^m \geq 1 - u^m - \varepsilon_b, \quad u^m \leq \varepsilon_u, \quad \cos(\mathbf{z}^m, \tilde{\boldsymbol{\mu}}_{k,y}^m) - \max_{c \neq y} \cos(\mathbf{z}^m, \tilde{\boldsymbol{\mu}}_{k,c}^m) \geq \Delta_\mu > 0. \tag{14}$$

The first two requirements give strong belief and low uncertainty; the third supplies geometric separation. These are sufficient signal conditions, not assumptions that all real-world corruptions satisfy.

Proposition A.1 (Consistency of the three signal patterns)**.** *Suppose reliable modalities satisfy Equation (14), with $\varepsilon_b < 1/2$ and $1 - \varepsilon_u > 2\varepsilon_b$. If reliable modalities support different classes, then $\kappa \geq (1 - \varepsilon_u - \varepsilon_b)^2$. If a reliable modality and a corrupted modality have uncertainty separation $u^B - u^A \geq \Delta_u$, then $|u^A - u^B| \geq \Delta_u$. If both reliable modalities support the same class $y \neq \tilde{y}$, then*

$$\kappa \leq 2\varepsilon_b, \qquad u^{\text{fuse}} \leq \frac{\varepsilon_u^2}{1 - 2\varepsilon_b}, \qquad \arg\max_c b_c^{\text{fuse}} = y \neq \tilde{y}. \tag{15}$$

Proof. For different classes, the conflict sum contains $b_{c_A}^A b_{c_B}^B$, which gives the lower bound. The uncertainty-gap conclusion follows from its stated separation condition. For a common class y, write $a = 1 - u^A$, $b = 1 - u^B$. The total belief outside y is at most ε_b in either modality, so

$$\kappa = ab - \sum_c b_c^A b_c^B \leq ab - (a - \varepsilon_b)(b - \varepsilon_b) \leq 2\varepsilon_b.$$

The fused-uncertainty bound follows directly. Since $a, b > 2\varepsilon_b$, the belief of y exceeds that of every other class in each modality. Every term in its fused numerator is therefore at least the corresponding term for any competing class, with a strict advantage, so y is the fused maximizer. □

This result interprets the routing signatures under explicit separation. It does not prove that a detector can identify every corruption from observed data, nor does low conflict rule out correlated mistakes.

A.2 Evidential uncertainty as a reliability proxy

For $S^m = \sum_c (e_c^m + 1)$, the uncertainty proxy satisfies

$$\frac{\partial u^m}{\partial e_c^m} = -\frac{C}{(S^m)^2} < 0. \tag{16}$$

This monotonicity ranks the total support supplied by the evidential head. To connect it with detection, let U_1 and U_0 be the scores of an unreliable and a reliable prediction. Suppose $\mathbb{E}[U_1 - U_0] \geq \Delta_u > 0$ and the centered difference is sub-Gaussian with variance proxy σ_u^2. The one-sided tail bound gives

$$\Pr(U_1 > U_0) \geq 1 - \exp\left(-\frac{\Delta_u^2}{2\sigma_u^2}\right). \tag{17}$$

AUROC, including half weight for ties, is at least this strict-ranking probability. The bound is conditional on separation; the empirical results in Table 17 evaluate whether the ordering is useful in the tested setting.

A.3 Conflict and the adaptive threshold

The conflict factor can be computed without enumerating all class pairs:

$$\kappa = (1 - u^A)(1 - u^B) - \langle \mathbf{b}^A, \mathbf{b}^B \rangle. \tag{18}$$

The first term is total committed belief interaction; the inner product removes class-matched agreement. In particular, $0 \leq \kappa \leq (1 - u^A)(1 - u^B)$. If modal maximizers differ, $\kappa \geq b_{c_A}^A b_{c_B}^B$. For fixed uncertainties and feasible conflict,

$$u^{\text{fuse}} = \frac{u^A u^B}{1 - \kappa}, \qquad \frac{\partial u^{\text{fuse}}}{\partial \kappa} = \frac{u^A u^B}{(1 - \kappa)^2} > 0. \tag{19}$$

The adaptive test $\kappa > \tau(1 - u^{\text{fuse}})$ is an algebraic test on the current masses. For $\kappa < 1$, it is equivalent to

$$\kappa^2 - (1 + \tau)\kappa + \tau(1 - u^A u^B) < 0. \tag{20}$$

For $0 < \tau \leq 1$, the larger root is at least one. Hence on the feasible domain the smaller root gives an explicit threshold:

$$\kappa > \frac{1 + \tau - \sqrt{(1 - \tau)^2 + 4\tau u^A u^B}}{2}. \tag{21}$$

When $u^A u^B$ is positive, the threshold is below τ. High conflict is evaluated before the uncertainty-gap and consensus cases, including the overlap band $\tau(1 - u^{\text{fuse}}) < \kappa < \tau$. The fraction in this band and the fixed-threshold comparison are reported in Appendix C.17.

A.4 Relation to product fusion

When $u^m < 1$, define normalized committed belief $\hat{p}_c^m = b_c^m/(1 - u^m)$. Product fusion normalizes $\hat{p}_c^A \hat{p}_c^B$. In the zero-uncertainty limit, provided the common class support is nonzero,

$$b_c^{\text{fuse}} = \frac{b_c^A b_c^B}{\sum_j b_j^A b_j^B} = p_c^{\text{PoE}}. \tag{22}$$

With nonzero uncertainty, DS additionally retains

$$b_c^{\text{fuse}} = \underbrace{\frac{b_c^A b_c^B}{1 - \kappa}}_{\text{matched belief}} + \underbrace{\frac{b_c^A u^B + b_c^B u^A}{1 - \kappa}}_{\text{belief with uncertain counterpart}}. \tag{23}$$

These terms let one modality contribute when its counterpart is uncertain. They also expose a conflict variable used in reconstruction. Neither comparison makes a conditional-independence claim about the data-generating process.

A.5 Prototype calibration and the mixing weight

Write a local prototype as $\boldsymbol{\mu}_{k,c}^m = \boldsymbol{\mu}_c^{m,\star} + \Delta_{k,c}^m + \varepsilon_{k,c}^m$, where Δ is systematic drift and ε is finite-sample estimation noise. For fixed normalized weights $\omega_{k,c}^m$, mean-zero errors uncorrelated across clients yield

$$\mathbb{E}\,\|\mathbf{G}_c^m - \boldsymbol{\mu}_c^{m,\star}\|^2 = \left\|\sum_k \omega_{k,c}^m \Delta_{k,c}^m\right\|^2 + \sum_k (\omega_{k,c}^m)^2 \mathbb{E}\|\varepsilon_{k,c}^m\|^2. \tag{24}$$

The weights are proportional to class counts. Non-aligned drifts can partially cancel, and averaging can reduce estimation variance under the stated error conditions. Global prototypes may still retain systematic bias.

Let $a = \boldsymbol{\mu}_{k,c}^m - \boldsymbol{\mu}_c^{m,\star}$ and $b = \mathbf{G}_c^m - \boldsymbol{\mu}_c^{m,\star}$, and define $A = \mathbb{E}\|a\|^2$, $B = \mathbb{E}\|b\|^2$, $D = \mathbb{E}\langle a, b\rangle$. The exact calibration error is

$$\mathbb{E}\|(1 - \gamma)a + \gamma b\|^2 = (1 - \gamma)^2 A + 2\gamma(1 - \gamma)D + \gamma^2 B. \tag{25}$$

If $A + B - 2D > 0$, the unconstrained minimizer is $(A - D)/(A + B - 2D)$; the constrained minimizer is its projection onto $[0, 1]$. Thus an interior mixture can minimize the error. Convexity also gives

$$\mathbb{E}\|(1 - \gamma)a + \gamma b\|^2 \leq (1 - \gamma)A + \gamma B, \tag{26}$$

but minimizing this loose linear bound is not equivalent to minimizing Equation (25). The empirical sweep includes the pure-global endpoint rather than inferring it from the bound.

A.6 Reliability of prototype-guided routing

Let $R_i^m = \max_c s_{i,c}^m - \lambda_u u_i^m$ and let $m^\dagger$ denote the reliable modality. Assume its ideal score has margin $R_i^{m^\dagger} - \max_{m \neq m^\dagger} R_i^m \geq \Delta_R > 0$. If the estimated scores are $\widehat{R}_i^m = R_i^m + \xi_i^m$ and each difference $\xi_i^m - \xi_i^{m^\dagger}$ is centered sub-Gaussian with proxy $\sigma_{\Delta R}^2$, then

$$\Pr(\widehat{m}_i \neq m^\dagger) \leq (|\mathcal{M}_k| - 1) \exp\left(-\frac{\Delta_R^2}{2\sigma_{\Delta R}^2}\right). \tag{27}$$

This follows by a union bound over competing modalities. If individual errors are independent with proxy σ_R^2, one may take $\sigma_{\Delta R}^2 = 2\sigma_R^2$. Specifying the proxy for the difference is necessary for the constant in the exponent. Reconstruction error obeys

$$\Pr(y_i' \neq y_i \mid \text{conflict}) \leq \Pr(\widehat{m}_i \neq m^\dagger \mid \text{conflict}) + \Pr(\arg\max_c b_{i,c}^{m^\dagger} \neq y_i \mid \text{conflict}). \tag{28}$$

Routing quality and prediction quality of the selected modality both matter. Prototype alignment and uncertainty can improve the first only when they separate reliable and unreliable sources.

A.7 Effect of dynamic label reconstruction

Let $q_i = \Pr(\tilde{y}_i \neq y_i)$, $a_i = \Pr(y_i' \neq y_i \mid \tilde{y}_i = y_i)$, and $b_i = \Pr(y_i' \neq y_i \mid \tilde{y}_i \neq y_i)$. The residual error is

$$\pi_i = (1 - q_i)a_i + q_i b_i. \tag{29}$$

For $0 < q_i < 1$, improvement $\pi_i < q_i$ is equivalent to $a_i < q_i(1 - b_i)/(1 - q_i)$. Correcting noisy labels must therefore compensate for the risk of changing clean ones. This motivates retaining observed labels when the reconstruction signals are weak.

For $g_i(\theta, y) = \nabla_\theta \ell(f_\theta(\mathbf{x}_i), y)$, suppose $\|g_i(\theta, c) - g_i(\theta, y_i)\| \leq G_y \mathbb{I}\{c \neq y_i\}$. Then

$$\|\mathbb{E}[g_i(\theta, y_i') - g_i(\theta, y_i)]\| \leq G_y \pi_i, \qquad \|\zeta_k(\theta)\| \leq G_y \pi_k, \tag{30}$$

where $\pi_k = \mathbb{E}_{i \in \mathcal{D}_k} \pi_i$. These bounds compare the same prediction and weighting mechanism under two targets. They do not remove bias from sample-dependent weighting or heterogeneous client aggregation.

A.8 Evidence-weighted aggregation

Consider modality updates $g_k^m = \nabla F^m(\theta) + \epsilon_k^m + \zeta_k^m$, where the stochastic errors are mean-zero and uncorrelated across clients, conditional on the aggregation weights. Suppose $\mathbb{E}\|\epsilon_k^m\|^2 \leq \sigma_m^2 / \bar{e}_k^m$ for positive accumulated evidence. Minimizing the variance upper bound $\sum_k (w_k^m)^2 \sigma_m^2 / \bar{e}_k^m$ under $\sum_k w_k^m = 1$ gives

$$w_k^m = \frac{\bar{e}_k^m}{\sum_j \bar{e}_j^m}. \tag{31}$$

Indeed, differentiating the Lagrangian gives $2w_k^m \sigma_m^2 / \bar{e}_k^m + \lambda = 0$, so the weights are proportional to evidence. This minimizes a conditional stochastic-variance bound; it does not eliminate the residual biases ζ_k^m. Appendix C.18 measures evidence–variance correlations in one protocol rather than treating inverse proportionality as universal.

Under Equation (30), the residual reconstruction bias satisfies

$$\left\| \sum_k \beta_k^m \zeta_k^m \right\| \leq G_y \sum_k \beta_k^m \pi_k^m. \tag{32}$$

Let averages be uniform over the participating modality clients. Then

$$\sum_k \beta_k^m \pi_k^m = \mathbb{E}_k \pi_k^m + \frac{\mathrm{Cov}_k(\bar{e}_k^m, \pi_k^m)}{\mathbb{E}_k \bar{e}_k^m}. \tag{33}$$

A nonpositive covariance reduces this bound relative to uniform client weighting. The comparison is not automatically a comparison with sample-size-weighted FedAvg; it depends on client sizes and error structure. Confidently incorrect clients can violate the desired covariance condition.

A.9 Convergence with reconstruction and aggregation bias

We make explicit the terms needed when analyzing a clean objective under evidence-dependent aggregation. Assume F is L-smooth and bounded below by $F_\star$. Write the round update as

$$\theta_{t+1} = \theta_t - a\big(\nabla F(\theta_t) + b_t + \xi_t\big), \qquad a = \eta Q, \tag{34}$$

where Q is the number of local steps, $\mathbb{E}_t \xi_t = 0$, and $\mathbb{E}_t \|\xi_t\|^2 \leq V_t$. The bias b_t contains residual reconstruction, local drift, and any change from the clean objective's weights to evidence-based client or sample weights. Conditional expectations are with respect to the current training history.

Theorem A.2 (Stationarity under a biased aggregate update)**.** *For Equation (34), if $a \le 1/(4L)$, then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\theta_t)\|^2 \le \frac{2(F(\theta_0)-F_\star)}{aT} + \frac{5}{2T}\sum_{t=0}^{T-1}\mathbb{E}\|b_t\|^2 + \frac{La}{T}\sum_{t=0}^{T-1}\mathbb{E}V_t. \tag{35}$$

Proof. Smoothness and conditional centering give

$$\mathbb{E}_t F(\theta_{t+1}) \le F(\theta_t) - a\|g_t\|^2 - a\langle g_t, b_t\rangle + \frac{La^2}{2}\big(\|g_t+b_t\|^2 + V_t\big), \quad g_t = \nabla F(\theta_t).$$

Use $|\langle g_t, b_t\rangle| \le \|g_t\|^2/4 + \|b_t\|^2$ and $\|g_t+b_t\|^2 \le 2\|g_t\|^2 + 2\|b_t\|^2$. Since $La \le 1/4$, the coefficient of $\|g_t\|^2$ is at most $-a/2$, and that of $\|b_t\|^2$ is at most $5a/4$. Sum over rounds, use the lower bound on F, and divide by $aT/2$. □

To interpret the bias, define $\bar{\pi}^{\beta}_{m,t} = \sum_k \beta^m_{k,t}\pi^m_{k,t}$. Let D_t bound bias from weighting relative to the clean objective, and let H^{loc}_t bound local-update drift. The block-wise label bound implies

$$\|b_t\|^2 \le 3G_y^2 \sum_m (\bar{\pi}^{\beta}_{m,t})^2 + 3D_t^2 + 3(H^{\mathrm{loc}}_t)^2. \tag{36}$$

When V_t is bounded, a scales as $T^{-1/2}$, and the bias terms vanish, Equation (35) has the usual $T^{-1/2}$ stationarity behavior. Otherwise the average squared bias determines a neighborhood. The benefit of reconstruction is to reduce its residual-label contribution. It is not a guarantee of convergence to the clean objective merely from smoothness and lower label error, because evidence-dependent weighting can change that objective.

B Complexity, Implementation, and Experimental Protocol

B.1 Computational and communication complexity

Let N denote samples per client, C classes, d feature dimension, M available modalities, and $|\Theta|$ model size. Evidence transformation costs $\mathcal{O}(MNC)$. Class-mean construction and calibration cost $\mathcal{O}(MNd + MCd)$. Equation (18) computes each bimodal conflict in $\mathcal{O}(C)$ rather than $\mathcal{O}(C^2)$. Comparing features with all class prototypes costs $\mathcal{O}(MNCd)$, the dominant extra local operation. The cost remains linear in local dataset size for fixed C, d, M.

The server aggregates prototypes in $\mathcal{O}(KMCd)$ and models in $\mathcal{O}(K|\Theta|)$. Extra client uploads consist of $\mathcal{O}(MCd)$ prototype values, class counts, and $\mathcal{O}(M)$ accumulated evidence scalars. These are aggregate statistics rather than per-example uncertainty. Table 5 separates the measured communication ratio from training time: $43.4/42.6 \simeq 1.02$ for communication, while $14.8/12.4 \simeq 1.19$ for time.

B.2 Data partitions and corruption

Only the training split is repartitioned across clients; standard evaluation splits remain separate. Dirichlet allocation of class proportions produces statistical heterogeneity, and client-level modality masks introduce modality heterogeneity. The default is $K = 10$, $\alpha = 0.5$, and three unimodal clients. The default corruption mixture comprises 40% semantic mismatch, 30% asymmetric corruption, and 30% consistent label error among the affected examples. Nominal total rates are 20%, 40%, and 60%; client-wise rates are sampled within ten percentage points of the nominal rate.

Semantic mismatch replaces one modality with a randomly chosen sample from another class while preserving the other modality and the label. Asymmetric corruption adds Gaussian noise with $\sigma = 0.5$ to one randomly chosen modality, leaving the other intact. Consistent label error replaces the class with a random incorrect class without changing either input. Thus the total corruption rate includes input corruption as well as annotation changes; it is not a purely symmetric label-flip rate.

The controlled Food-101 experiment uses a filtered subset of approximately 73,500 samples. The natural-noise experiment retains approximately 90,840 raw web-collected pairs. In the audit, 1,500 training triplets are randomly selected and independently annotated by two annotators for clean/noisy

status and noise type. A third annotator resolves disagreements. Cohen's agreement is approximately 0.82 for clean/noisy status and 0.74 for type. Audit labels are hidden from all training methods and used only for detection and correction analyses; audited examples do not overlap the standard clean Food-101 test split.

B.3 Architectures and training schedule

AV-MNIST uses three-layer image and spectrogram CNNs with 32, 64, and 128 channels and 3×3 kernels. UPMC Food-101 uses ResNet-18 for images and a two-layer text LSTM with hidden size 256. CMU-MOSEI uses a one-dimensional acoustic CNN and frozen pretrained BERT followed by a trainable language projection. All modality branches project to 128-dimensional features before the evidential head. WebVision uses the image and text encoders described in Appendix C.12.

Training lasts 100 communication rounds. In the first 20 rounds, local evidential classification uses observed labels and initializes prototypes. Reconstruction is then enabled. Each client performs two local epochs per round with batch size 32. The optimizer is SGD with learning rate 0.01 and momentum 0.9. Table 8 gives the fixed coefficients, including distinct notation for the conflict and uncertainty-based modality choices.

Table 8: Default hyperparameters. Sensitivity experiments vary one stated setting while retaining the rest.

Quantity	Symbol	Value
Communication rounds	T	100
Warm-up rounds	T_{warm}	20
Local epochs per round	E	2
Batch size		32
Embedding dimension	d	128
Learning rate / momentum	η_{lr} / momentum	0.01 / 0.9
KL coefficient	λ_{KL}	0.1
Auxiliary modality-loss weight	λ_e	0.5
Routing uncertainty penalty	λ_u	0.5
Global prototype mixing weight	γ	0.5
Conflict threshold	τ	0.5
Uncertainty-gap threshold	δ	0.3
Consensus/correction uncertainty threshold	ϵ	0.2
Observed-class prototype similarity threshold	ρ	0.5
Dirichlet partition parameter	α	0.5
Unimodal-client fraction		30%
Repeated runs		3

B.4 Comparison and model selection

Comparisons start from official method implementations and use identical dataset partitions, corrupted examples, modality masks, seeds, local epochs, communication rounds, and validation protocols. Method-specific parameters are selected independently on validation data. Shared parameters use comparable search ranges, with the same number of tuning trials. Test results are not used for selection. WebVision instead selects models using training data and reports on the human-verified validation split, as detailed in its separate protocol.

For the additional noisy-label baselines, each modality branch applies the original method's sample detection or correction. Available branch probabilities are averaged at prediction time, while each baseline's original aggregation rule is retained. This gives access to both modalities without introducing FedEPR's conflict routing. Standard deviations refer to the three reported runs; where an auxiliary experiment reports only means, no unreported uncertainty estimates are inferred.

C Extended Results and Additional Experiments

C.1 Complete controlled-noise statistics

Table 9 gives all three-run means and standard deviations for the controlled-noise results summarized in the main text. The 40/30/30 corruption mixture, client-wise rate variability, and modality masks are shared across methods.

Table 9: Test accuracy (%, mean ± standard deviation over three runs) under 20%, 40%, and 60% structured noise, Dirichlet $\alpha = 0.5$, and 30% unimodal clients. These are the full statistics underlying Table 1.

Dataset	Method	20%	40%	60%
AV-MNIST	FedAvg	89.34 ± 0.28	81.67 ± 0.41	70.42 ± 0.58
	FedProx	90.18 ± 0.25	82.45 ± 0.37	71.83 ± 0.52
	FedMSplit	91.72 ± 0.22	84.53 ± 0.33	74.26 ± 0.47
	CreamFL	92.45 ± 0.20	86.18 ± 0.30	78.35 ± 0.43
	PmcmFL	93.67 ± 0.18	87.24 ± 0.27	77.18 ± 0.41
	MPCPFL	94.38 ± 0.16	88.56 ± 0.24	77.92 ± 0.39
	FedMobile	93.85 ± 0.17	89.17 ± 0.23	79.64 ± 0.37
	FedEPR (Ours)	**98.23** ± 0.12	**95.47** ± 0.17	**92.18** ± 0.24
Food-101	FedAvg	74.23 ± 0.32	64.85 ± 0.47	53.62 ± 0.69
	FedProx	75.47 ± 0.29	65.93 ± 0.43	54.86 ± 0.61
	FedMSplit	77.38 ± 0.26	68.52 ± 0.38	58.17 ± 0.54
	CreamFL	78.65 ± 0.24	70.74 ± 0.35	61.43 ± 0.49
	PmcmFL	80.27 ± 0.22	73.28 ± 0.32	64.52 ± 0.46
	MPCPFL	81.92 ± 0.20	72.65 ± 0.30	66.83 ± 0.43
	FedMobile	80.73 ± 0.21	74.56 ± 0.29	68.25 ± 0.41
	FedEPR (Ours)	**89.17** ± 0.15	**85.34** ± 0.21	**80.26** ± 0.32
CMU-MOSEI	FedAvg	42.18 ± 0.45	35.24 ± 0.52	28.37 ± 0.64
	FedProx	43.05 ± 0.41	36.17 ± 0.48	29.24 ± 0.59
	FedMSplit	45.34 ± 0.38	38.46 ± 0.44	31.58 ± 0.55
	CreamFL	46.82 ± 0.35	40.15 ± 0.41	33.27 ± 0.51
	PmcmFL	48.15 ± 0.32	41.37 ± 0.38	34.28 ± 0.48
	MPCPFL	48.73 ± 0.30	42.28 ± 0.36	35.12 ± 0.45
	FedMobile	47.86 ± 0.31	42.63 ± 0.35	35.74 ± 0.44
	FedEPR (Ours)	**53.28** ± 0.26	**46.15** ± 0.31	**40.37** ± 0.39

C.2 Uncertainty and fusion ablations

Table 10 compares softmax and evidential representations, averaging and product-of-experts (PoE) fusion, DS without routing, and the complete method. Accuracy, noisy-label detection AUROC, and calibration error improve across these ablations. On Food-101, DS without routing obtains 82.64% accuracy, 84.77% AUROC, and 6.53% ECE, while FedEPR obtains 85.34%, 88.91%, and 4.37%. The additional gain tests the routing mechanism beyond a change of fusion operator.

C.3 Naturally occurring noise and the human audit

The raw UPMC experiment preserves original web-collected image–text–label triplets, including natural annotation and alignment errors. The CMU experiment evaluates low-agreement sentiment supervision. Table 2 reports accuracy without injected corruption. FedEPR improves on FedMobile by 4.26 points on UPMC Raw and 4.38 points on CMU Low-agreement. These settings complement the controlled experiments but do not exhaust natural annotation artifacts.

The human audit uses the 1,500 training triplets and independent annotation protocol in Appendix B. It provides detection and correction labels only, rather than an additional classification test split. Table 11 reports AUROC, AUPRC, ECE, and correction precision, recall, and F1. Correction precision counts corrected labels matching the audit. FedEPR reaches 88.6% AUROC and 81.7% correction precision. The type breakdown in Table 12 reports 63.1% clean triplets, 13.6% mismatch, 9.8%

Table 10: Uncertainty and fusion analysis at 40% structured noise and $\alpha = 0.5$. Values are mean $\pm$ standard deviation over three runs. AUROC measures noisy-label detection; ECE is expected calibration error. All metrics are percentages.

Dataset	Method	Accuracy ↑	AUROC ↑	ECE ↓
AV-MNIST	Softmax Avg.	88.35 ± 0.28	82.41 ± 0.36	8.74 ± 0.18
	Softmax PoE	89.62 ± 0.25	83.17 ± 0.34	8.21 ± 0.16
	Evidential Avg.	92.36 ± 0.21	89.74 ± 0.27	5.36 ± 0.12
	Evidential PoE	93.04 ± 0.20	90.38 ± 0.25	5.12 ± 0.11
	DS w/o Routing	93.71 ± 0.19	91.46 ± 0.24	4.61 ± 0.10
	FedEPR	$\mathbf{95.47} \pm 0.17$	$\mathbf{94.83} \pm 0.20$	$\mathbf{2.86} \pm 0.08$
Food-101	Softmax Avg.	76.48 ± 0.34	75.63 ± 0.42	10.92 ± 0.24
	Softmax PoE	77.65 ± 0.31	76.42 ± 0.39	10.35 ± 0.22
	Evidential Avg.	80.71 ± 0.28	82.58 ± 0.33	7.42 ± 0.18
	Evidential PoE	81.56 ± 0.26	83.19 ± 0.31	7.18 ± 0.17
	DS w/o Routing	82.64 ± 0.24	84.77 ± 0.29	6.53 ± 0.15
	FedEPR	$\mathbf{85.34} \pm 0.21$	$\mathbf{88.91} \pm 0.25$	$\mathbf{4.37} \pm 0.12$
CMU-MOSEI	Softmax Avg.	40.72 ± 0.41	70.86 ± 0.51	13.47 ± 0.31
	Softmax PoE	41.26 ± 0.39	71.58 ± 0.48	12.96 ± 0.28
	Evidential Avg.	42.88 ± 0.36	76.94 ± 0.43	10.81 ± 0.25
	Evidential PoE	43.21 ± 0.34	77.36 ± 0.41	10.52 ± 0.24
	DS w/o Routing	44.07 ± 0.33	78.82 ± 0.39	9.74 ± 0.22
	FedEPR	$\mathbf{46.15} \pm 0.31$	$\mathbf{82.46} \pm 0.35$	$\mathbf{7.28} \pm 0.19$

asymmetric corruption, and 13.5% consistent label error. Correction is most precise for mismatch and least precise for consistent errors, where agreement alone offers little protection against a wrong label.

Table 11: Detection (AUROC, AUPRC, ECE) and correction (precision, recall, F1) on 1,500 human-audited UPMC training triplets. Values are percentages with the reported standard deviations. Audit labels are used only for analysis.

Method	AUROC	AUPRC	ECE	Prec.	Rec.	F1
Softmax Confidence	72.6 ± 1.8	61.4 ± 2.2	11.84 ± 0.31	62.8 ± 2.4	43.6 ± 2.7	51.5 ± 2.5
Evidential w/o DS	80.7 ± 1.4	72.3 ± 1.8	7.36 ± 0.22	71.4 ± 1.9	55.2 ± 2.2	62.3 ± 2.0
DS w/o Routing	84.9 ± 1.2	77.5 ± 1.5	5.92 ± 0.18	76.8 ± 1.6	60.7 ± 1.8	67.8 ± 1.6
FedEPR	$\mathbf{88.6} \pm 1.0$	$\mathbf{82.4} \pm 1.2$	$\mathbf{4.51} \pm 0.14$	$\mathbf{81.7} \pm 1.3$	$\mathbf{68.9} \pm 1.5$	$\mathbf{74.8} \pm 1.3$

Table 12: Observed noise composition and correction precision (%) on the human-audited UPMC subset.

Metric	Clean	Mismatch	Asym. corr.	Label error
Composition (%)	63.1	13.6	9.8	13.5
Correction Prec. (%)	-	84.5	82.8	76.9

C.4 Warm-up duration

Warm-up stabilizes representation and prototype estimates before reconstruction. Figure 7 varies its duration from zero to 40 rounds while holding total training at 100 rounds. The sweep peaks at 20 rounds on AV-MNIST, 30 on Food-101, and ten on CMU-MOSEI. Activating reconstruction immediately reduces accuracy on every dataset; extending warm-up too far leaves fewer reconstruction rounds and can increase memorization of noisy supervision. The common default in the main protocol is 20 rounds.

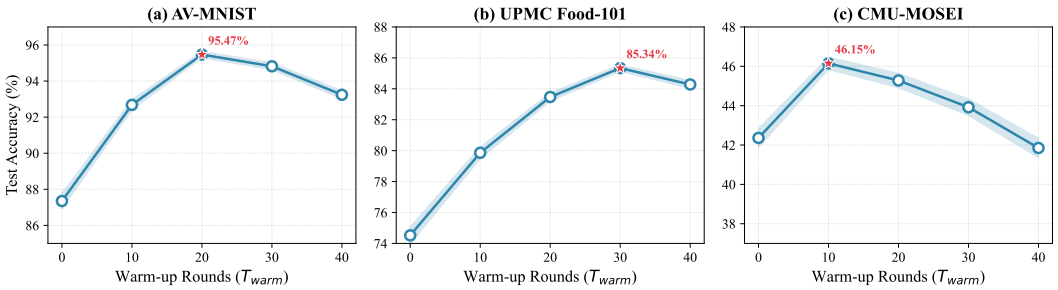

Figure 7: Accuracy versus warm-up rounds at 40% structured noise and $\alpha = 0.5$, with 100 total rounds. The stars mark the maxima of this warm-up sweep.

C.5 Noise-type composition

Table 13 changes the mixture while retaining 40% total corruption. FedEPR performs best in the asymmetric-corruption-dominant setting, where a reliable modality can guide reconstruction, and obtains lower accuracy under consistent-label-error-dominant corruption. In the latter case, cross-modal agreement is compatible with the wrong observed annotation, making evidence calibration and class geometry particularly relevant. The table preserves the independently reported composition-sweep values.

Table 13: Accuracy (%) under different proportions of semantic mismatch, asymmetric corruption, and consistent label error at 40% total corruption. The three proportions in each row are percentages in that order.

Dataset	Composition	FedAvg	MPCPFL	FedEPR
AV-MNIST	SM-dominant (70/15/15)	78.42	86.73	94.86
	AC-dominant (15/70/15)	83.56	89.82	96.23
	CLE-dominant (15/15/70)	79.83	87.24	93.45
	Balanced (40/30/30)	81.67	88.56	95.47
	Uniform (33/33/34)	80.94	88.15	95.12
Food-101	SM-dominant (70/15/15)	61.35	72.48	84.52
	AC-dominant (15/70/15)	66.74	76.35	86.18
	CLE-dominant (15/15/70)	62.58	73.62	82.76
	Balanced (40/30/30)	64.85	74.82	85.34
	Uniform (33/33/34)	63.92	74.26	84.87
CMU-MOSEI	SM-dominant (70/15/15)	32.18	40.56	44.83
	AC-dominant (15/70/15)	36.85	43.72	47.56
	CLE-dominant (15/15/70)	33.47	41.28	44.12
	Balanced (40/30/30)	35.24	42.28	46.15
	Uniform (33/33/34)	34.56	41.85	45.72

C.6 Modality availability

Table 14 increases the proportion of unimodal clients from zero to 70%. FedEPR retains 89.62% AV-MNIST accuracy and 76.48% Food-101 accuracy at 70%. Cross-modal verification becomes less available as the fraction rises, so unimodal prototype consistency and modality-specific aggregation become more important. The reported margins vary across datasets; for AV-MNIST the margin over MPCPFL grows from 4.45 to 5.88 points in this sweep.

C.7 Federation size and per-class performance

Figure 8 contains the original 10–100-client sweep. Its results characterize the interaction between federation size and data fragmentation. The separate 10–500-client partial-participation experiment is specified in Appendix C.14; the two protocols should not be interpreted as the same sweep.

Table 14: Accuracy (%) with different percentages of unimodal clients at 40% structured noise. Baseline values are those reported in this missing-modality sweep.

Dataset	Method	0%	30%	50%	70%
AV-MNIST	FedAvg	84.26	81.67	78.35	73.48
	PmcmFL	91.45	89.47	86.82	82.35
	MPCPFL	92.83	90.62	87.56	83.74
	FedEPR	97.28	95.47	93.15	89.62
	Δ vs. MPCPFL	+4.45	+4.85	+5.59	+5.88
Food-101	FedAvg	68.72	64.85	60.56	54.83
	PmcmFL	78.36	75.68	71.94	66.47
	MPCPFL	79.85	74.82	72.68	67.82
	FedEPR	88.65	85.34	81.72	76.48
	Δ vs. MPCPFL	+8.80	+10.52	+9.04	+8.66
CMU-MOSEI	FedAvg	38.56	35.24	31.82	27.45
	PmcmFL	44.82	41.37	38.26	34.18
	MPCPFL	45.67	42.28	39.15	35.24
	FedEPR	49.83	46.15	43.28	39.56
	Δ vs. MPCPFL	+4.16	+3.87	+4.13	+4.32

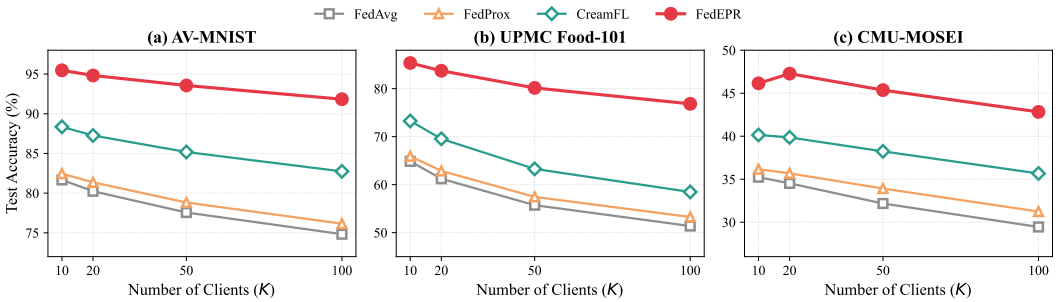

Figure 8: Accuracy as the original scalability experiment increases the federation from ten to 100 clients at 40% structured noise. Panels show AV-MNIST, UPMC Food-101, and CMU-MOSEI.

Table 15 gives the reported AV-MNIST per-class diagnostic. Digits 3, 5, and 8 improve by more than 16 points over FedAvg. These are descriptive class-level comparisons; aggregate figures from different diagnostic runs are not substituted for the main results.

Table 15: Per-class AV-MNIST accuracy (%) at 40% structured noise. The final column reproduces the aggregate reported for this per-class diagnostic.

Method	0	1	2	3	4	5	6	7	8	9	Reported avg.
FedAvg	85.2	89.4	78.3	76.5	80.7	74.8	83.6	82.4	77.2	88.6	81.67
MPCPFL	92.4	94.7	87.6	85.3	88.5	84.2	90.8	89.6	86.4	93.1	89.26
FedEPR	97.3	98.2	94.5	92.8	95.4	91.6	96.2	95.8	93.7	98.2	95.37
Gain vs. FedAvg	+12.1	+8.8	+16.2	+16.3	+14.7	+16.8	+12.6	+13.4	+16.5	+9.6	+13.70

C.8 Adapted noisy-label federated baselines

For each modality, a baseline produces $p_i^m = h^m(f^m(\mathbf{x}_i^m))$. Its adapted multimodal output is

$$p_i^{\text{ada}} = \frac{1}{|\mathcal{M}_k|} \sum_{m \in \mathcal{M}_k} p_i^m. \tag{37}$$

This reduces to the available prediction for a unimodal client. FedCorr, FedNoRo, and FedLN retain their original aggregation rules and apply correction or detection to each branch separately. Table 16 compares these adaptations with FedEPR. At 60% corruption on AV-MNIST, their accuracies

range from 76.56% to 77.83%, versus 92.18% for FedEPR. This comparison evaluates branch-wise adaptations; it does not supply those methods with the proposed cross-modal routing rule.

Table 16: Accuracy (%, mean ± standard deviation over three runs) for noisy-label FL methods adapted to multimodal inputs at $\alpha = 0.5$. The dagger denotes branch-wise correction followed by average prediction fusion.

Dataset	Method	20%	40%	60%
AV-MNIST	FedAvg	89.34 ± 0.28	81.67 ± 0.41	70.42 ± 0.58
	FedCorr†	92.24 ± 0.23	84.72 ± 0.36	77.83 ± 0.44
	FedNoRo†	91.87 ± 0.24	86.35 ± 0.32	76.56 ± 0.47
	FedLN†	92.45 ± 0.21	86.18 ± 0.34	77.45 ± 0.45
	FedEPR	$\mathbf{98.23} \pm 0.12$	$\mathbf{95.47} \pm 0.17$	$\mathbf{92.18} \pm 0.24$
Food-101	FedAvg	74.23 ± 0.32	64.85 ± 0.47	53.62 ± 0.69
	FedCorr†	77.82 ± 0.27	69.83 ± 0.39	60.74 ± 0.53
	FedNoRo†	78.56 ± 0.25	69.47 ± 0.41	60.28 ± 0.56
	FedLN†	77.35 ± 0.29	70.15 ± 0.37	59.86 ± 0.55
	FedEPR	$\mathbf{89.17} \pm 0.15$	$\mathbf{85.34} \pm 0.21$	$\mathbf{80.26} \pm 0.32$
CMU-MOSEI	FedAvg	42.18 ± 0.45	35.24 ± 0.52	28.37 ± 0.64
	FedCorr†	44.83 ± 0.40	38.56 ± 0.47	31.85 ± 0.55
	FedNoRo†	44.12 ± 0.43	38.87 ± 0.45	32.68 ± 0.52
	FedLN†	44.56 ± 0.41	39.24 ± 0.44	31.74 ± 0.57
	FedEPR	$\mathbf{53.28} \pm 0.26$	$\mathbf{46.15} \pm 0.31$	$\mathbf{40.37} \pm 0.39$

C.9 Detection and Representation Diagnostics

Table 17 compares evidential and softmax detection using the same thresholding protocol. On Food-101, evidential precision and recall are 87.7% and 84.0%, compared with 71.3% and 62.2% for softmax. The comparison supports uncertainty as a ranking signal under the evaluated corruption process.

Table 17: Noise-detection precision and recall (%) at 40% structured noise, averaged over the three corruption types. The softmax comparison uses the same thresholding protocol.

Dataset	EDL prec.	EDL rec.	Softmax prec.	Softmax rec.
AV-MNIST	91.1 ± 1.1	87.9 ± 1.4	77.4 ± 2.3	68.2 ± 3.1
UPMC Food-101	87.7 ± 1.3	84.0 ± 1.6	71.3 ± 2.6	62.2 ± 3.4
CMU-MOSEI	81.9 ± 1.6	77.8 ± 1.9	66.5 ± 2.8	57.7 ± 3.7

Figure 5 compares learned representations, and Figure 9 visualizes conflict, the uncertainty gap, and fused uncertainty. Mismatch moves samples toward larger conflict, while asymmetric corruption tends to increase the uncertainty gap. Consistent label errors are more difficult because modal agreement can remain strong. The visible separation weakens under severe non-IID data.

C.10 Prototype Stabilization and Progressive Detection

Table 18 measures cosine distance to clean held-out class centroids. These clean centroids are used for analysis, not for training. At round 100, the AV-MNIST distance is 0.053 with FedEPR, compared with 0.172 without global calibration and 0.298 with FedAvg. The improvement indicates that the learned anchors move closer to the diagnostic class centers.

Table 19 follows AV-MNIST detection through training. Overall precision rises from 78.2% at round 20 to 91.5% at round 100; recall rises from 74.6% to 87.9%. Prototype distance falls from 0.156 to 0.053. These jointly improving signals are consistent with the intended warm-up and reconstruction cycle, without treating the correlation as proof of causation.

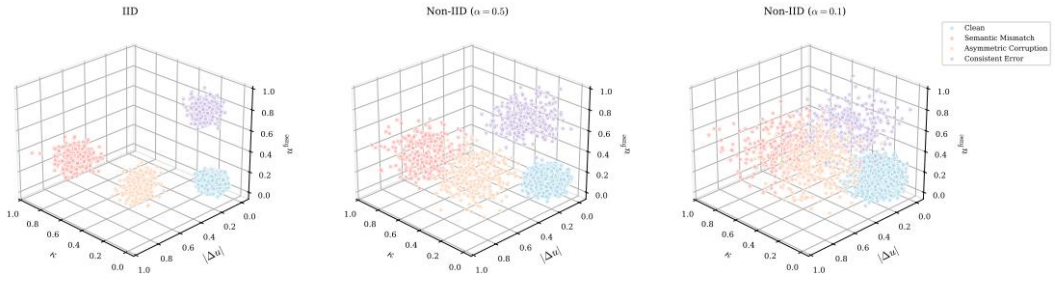

Figure 9: Conflict–uncertainty visualization under IID, non-IID ($\alpha = 0.5$), and severe non-IID ($\alpha = 0.1$) partitions. Axes are modal conflict κ, uncertainty gap $|u^A - u^B|$, and fused uncertainty u^{fuse}.

Table 18: Cosine distance to clean held-out class centroids at 40% structured corruption (mean $\pm$ standard deviation over three seeds). Lower is better.

Round	Dataset	FedAvg	w/o GPC	FedEPR
20	AV-MNIST	0.318 ± 0.021	0.247 ± 0.017	$\mathbf{0.156} \pm 0.012$
	UPMC Food-101	0.462 ± 0.028	0.384 ± 0.024	$\mathbf{0.278} \pm 0.018$
	CMU-MOSEI	0.508 ± 0.033	0.436 ± 0.028	$\mathbf{0.334} \pm 0.022$
100	AV-MNIST	0.298 ± 0.019	0.172 ± 0.014	$\mathbf{0.053} \pm 0.006$
	UPMC Food-101	0.435 ± 0.026	0.294 ± 0.019	$\mathbf{0.098} \pm 0.009$
	CMU-MOSEI	0.487 ± 0.031	0.358 ± 0.024	$\mathbf{0.148} \pm 0.014$

Table 19: Progressive AV-MNIST detection at 40% structured noise. Per-type and overall detection metrics are percentages; distance is prototype-to-clean-centroid cosine distance.

Round	Stage	SM prec.	AC prec.	CLE prec.	Prec.	Rec.	F1	Distance
20	Warm-up	82.4 ± 1.8	79.6 ± 2.1	72.5 ± 2.6	78.2 ± 2.0	74.6 ± 2.1	76.4 ± 1.9	0.156 ± 0.012
30	Recon.	88.7 ± 1.3	85.2 ± 1.6	80.4 ± 2.0	84.8 ± 1.5	81.3 ± 1.6	83.0 ± 1.5	0.118 ± 0.010
50		92.6 ± 0.9	90.3 ± 1.1	85.8 ± 1.5	89.6 ± 1.1	86.2 ± 1.2	87.9 ± 1.1	0.087 ± 0.008
75		93.8 ± 0.8	91.4 ± 1.0	86.7 ± 1.4	90.7 ± 1.0	87.4 ± 1.1	89.0 ± 1.0	0.064 ± 0.007
100		94.3 ± 0.8	91.8 ± 1.0	87.2 ± 1.4	91.5 ± 0.9	87.9 ± 1.0	89.7 ± 0.9	0.053 ± 0.006

C.11 Noise processes beyond the design generator

We evaluate four processes not used for method design or tuning, using AV-MNIST and UPMC Food-101 at 40% corruption and non-IID $\alpha = 0.5$. Each corrupted split is generated once per seed and shared by every method; selection uses validation data only. The independent classifier used by the first two generators is unavailable to the federated learners.

For instance-dependent flips (P1), a centralized multimodal classifier trained before injection supplies margin $q_i = p(y_i) - \max_{c \neq y_i} p(c)$. Forty percent of training samples are selected without replacement with weights proportional to $1 - q_i$ and relabeled to the classifier's second-ranked class; modalities are unchanged. For adjacent-class flips (P2), the same independent classifier supplies a confusion matrix C. Each selected label c is replaced by $\text{adj}(c) = \arg\max_{c' \neq c} C_{c,c'}$.

Realistic input degradation (P3) uses image motion blur with kernel size five, text token dropout at rate 0.3 for Food-101, and audio reverberation with $\text{RT60} = 0.4$ seconds plus ambient noise at 15 dB SNR. P4 independently applies the corresponding degradations to both modalities. P1 and P2 still contain coherent modalities with incorrect labels; P3 and P4 evaluate input degradation rather than new label-transition families. Thus the tests broaden the protocol without eliminating overlap with the structural cases used by FedEPR.

FedEPR leads the compared methods in all eight settings. P4 lowers correction precision to 75.32%, reflecting the difficulty of choosing reliable supervision when both inputs are degraded. The reported label-change fractions should be interpreted together with correction precision: more changes do not necessarily mean better correction.

Table 20: Accuracy (%, mean ± standard deviation over three seeds) under the four additional corruption processes at 40% corruption and $\alpha = 0.5$.

Process	Dataset	FedAvg	MPCPFL	FedMobile	FedEPR
P1	AV-MNIST	81.23 ± 0.44	88.14 ± 0.27	88.76 ± 0.25	$\mathbf{92.15} \pm 0.22$
	Food-101	64.47 ± 0.51	72.18 ± 0.33	74.12 ± 0.31	$\mathbf{80.83} \pm 0.28$
P2	AV-MNIST	80.82 ± 0.46	87.63 ± 0.28	88.45 ± 0.26	$\mathbf{93.72} \pm 0.20$
	Food-101	63.91 ± 0.53	71.74 ± 0.34	73.58 ± 0.32	$\mathbf{82.91} \pm 0.26$
P3	AV-MNIST	81.38 ± 0.43	88.05 ± 0.26	88.82 ± 0.24	$\mathbf{94.38} \pm 0.19$
	Food-101	64.27 ± 0.50	72.31 ± 0.32	73.94 ± 0.30	$\mathbf{83.67} \pm 0.25$
P4	AV-MNIST	79.45 ± 0.52	86.18 ± 0.31	87.34 ± 0.29	$\mathbf{90.86} \pm 0.26$
	Food-101	62.14 ± 0.57	70.42 ± 0.37	72.53 ± 0.35	$\mathbf{78.92} \pm 0.31$

Table 21: Detection and correction (%) on AV-MNIST for the processes in Table 20. Labels changed is the fraction of training labels modified by reconstruction.

Process	Detection AUROC	Correction precision	Labels changed
P1	82.36 ± 1.7	78.54 ± 2.1	22.4
P2	89.72 ± 1.2	83.21 ± 1.7	30.7
P3	92.85 ± 0.9	86.74 ± 1.3	34.2
P4	85.63 ± 1.5	75.32 ± 2.0	37.5

C.12 Naturally noisy WebVision Google500

WebVision Google500 is the subset used by CAPro (Qin et al., 2023), containing approximately 0.48 million Google images from 500 randomly selected WebVision classes (Li et al., 2017). Each sample pairs the image with available title, description, and tags, retaining its original query-derived class. Missing metadata remains missing, and no synthetic corruption is injected.

Training data are distributed across ten clients using IID or Dirichlet non-IID partitions with $\alpha = 0.5$ and 30% unimodal clients. All methods use identical data, partitions, modality masks, seeds, ResNet-18 image and two-layer LSTM text encoders, and 100 communication rounds. Model selection uses training data only. Evaluation uses the human-verified WebVision500 validation split and averages three seeds. For scores $p_{i,c}$ and top-five class set $T_5(i)$,

$$\text{Acc@1} = \frac{1}{N}\sum_i \mathbb{I}\{\arg\max_c p_{i,c} = y_i\}, \qquad \text{Acc@5} = \frac{1}{N}\sum_i \mathbb{I}\{y_i \in T_5(i)\}. \tag{38}$$

Table 22: WebVision Google500 accuracy (%, mean ± standard deviation, three seeds) with natural web labels and no synthetic corruption.

Method	IID top-1	IID top-5	Non-IID top-1	Non-IID top-5
FedAvg	60.74 ± 0.57	82.15 ± 0.44	55.83 ± 0.65	78.42 ± 0.54
FedProx	61.63 ± 0.52	82.87 ± 0.47	56.72 ± 0.63	79.05 ± 0.58
MPCPFL	64.52 ± 0.54	85.23 ± 0.38	59.87 ± 0.56	82.18 ± 0.45
FedMobile	65.38 ± 0.41	85.76 ± 0.36	60.74 ± 0.48	82.63 ± 0.40
FedEPR	$\mathbf{68.61} \pm 0.46$	$\mathbf{88.14} \pm 0.33$	$\mathbf{64.52} \pm 0.53$	$\mathbf{85.37} \pm 0.43$

Compared with FedMobile, FedEPR improves top-1/top-5 accuracy by 3.23/2.38 points under IID and 3.78/2.74 points under non-IID partitioning. The evaluation supplies a second web-collected image–text setting beyond UPMC, at a larger sample and class scale.

C.13 Perturbation of prototype and evidence statistics

FedEPR preserves raw-data locality, but prototypes can reveal class-conditional feature information, and model updates can also leak information. Secure aggregation is compatible with the sums used by prototypes and evidence weights. We assess Gaussian perturbation of both additional uploads on AV-MNIST at 40% structured noise and $\alpha = 0.5$, averaging three seeds.

The stated accounting uses record-level adjacency, an ℓ_2 clipping bound of 1.0 on normalized feature contributions, $\delta_{\mathrm{DP}} = 10^{-5}$, and a Rényi accountant with optimal order selection across 100 rounds. Table 23 reports the budgets for these additional statistics. Model updates are not protected by this mechanism, so these budgets do not describe an end-to-end private training algorithm.

Table 23: Statistics-perturbation utility on AV-MNIST. Accuracy is mean ± standard deviation (%); drops and gains are percentage points. FedAvg and w/o GPC accuracies are 81.67 and 90.26.

Protection	$\varepsilon_{\mathrm{DP}}$	Accuracy	Drop	vs. FedAvg	vs. w/o GPC
No perturbation	∞	95.47 ± 0.17	0.00	13.80	5.21
	10	94.83 ± 0.19	0.64	13.16	4.57
Gaussian	5	93.26 ± 0.23	2.21	11.59	3.00
	1	89.41 ± 0.36	6.06	7.74	−0.85

At $\varepsilon_{\mathrm{DP}} = 5$, accuracy remains three points above the variant without global calibration. At $\varepsilon_{\mathrm{DP}} = 1$, the perturbed method falls 0.85 points below it. Strong perturbation can therefore remove the benefit of shared prototype information. Protecting model updates and accounting jointly for all released quantities would be necessary for an end-to-end guarantee.

C.14 Partial participation with up to 500 clients

The AV-MNIST stress test uses 40% structured noise, $\alpha = 0.5$, and approximately ten participating clients per round. Increasing the federation simultaneously decreases local sample counts and participation rates. At 500 clients, FedEPR retains a 9.63-point advantage over FedAvg, down from 13.80 points at ten clients (Table 24). This measures a deployment regime with class-sparse clients, not invariance to federation size.

Table 24: AV-MNIST partial-participation stress test. Accuracy (%) is averaged over three seeds; gains are percentage points.

Clients	Participation	FedAvg	FedEPR	Gain
10	100%	81.67	95.47	13.80
100	10%	76.35	88.72	12.37
200	5%	73.18	84.47	11.29
500	2%	68.63	78.26	9.63

C.15 Prototype-calibration sweep

Table 25 includes the pure-local and pure-global endpoints as well as three mixtures. Every setting uses 40% structured noise and $\alpha = 0.5$. The best tested weight is $\gamma = 0.5$ across datasets. Pure-global calibration lowers accuracy by 1.09 points on AV-MNIST, 1.52 on Food-101, and 1.19 on CMU-MOSEI. The exact error decomposition in Appendix A.5 explains why correlated local and global prototype errors can favor an interior mixture.

Table 25: Prototype mixing at 40% structured noise and $\alpha = 0.5$. Accuracy (%, mean ± standard deviation over three seeds).

γ	AV-MNIST	Food-101	CMU-MOSEI
0.00	90.26 ± 0.24	78.65 ± 0.30	42.18 ± 0.38
0.25	93.14 ± 0.21	82.43 ± 0.27	44.52 ± 0.35
0.50	$\mathbf{95.47} \pm 0.17$	$\mathbf{85.34} \pm 0.21$	$\mathbf{46.15} \pm 0.31$
0.75	95.31 ± 0.18	85.17 ± 0.22	45.89 ± 0.33
1.00	94.38 ± 0.20	83.82 ± 0.26	44.96 ± 0.35

C.16 FUSED-LOSS GRADIENT COUPLING

The original objective masks only the auxiliary modality losses. Even if the reconstructed label follows one modality on a high-conflict sample, $\mathcal{L}_i^{\text{fuse}}$ backpropagates through both encoders. We evaluate a diagnostic that omits this term only when $h_i = 1$. The original objective is retained for all main results.

Table 26: Fused-loss diagnostic at 40% structured noise and $\alpha = 0.5$. Accuracies are the reported means over three seeds. The middle column gives the reported mean sample weight on high-conflict examples.

Objective	Mean w_i	AV-MNIST	Food-101
Original	0.83	95.47	85.34
High-conflict fused-loss gating	0.84	95.68	85.53

The changes are 0.21 and 0.19 points. They are small in these settings, but the original objective still permits updates to a non-selected modality. In particular, the observed mean weight of 0.83 is not close to zero and should not be described as removing gradient leakage.

C.17 ADAPTIVE VERSUS FIXED CONFLICT THRESHOLD

The first-applicable-branch rule in Equation (6) assigns the overlap band $\tau(1 - u^{\text{fuse}}) < \kappa < \tau$ to prototype-guided conflict handling. This band contains 4.7% of bimodal AV-MNIST samples at 40% structured noise. The fixed-threshold diagnostic replaces the adaptive test by $\kappa > \tau$; accuracies are 94.86% on AV-MNIST and 84.73% on Food-101, against 95.47% and 85.34% for the adaptive rule. Both gains are 0.61 points. The comparison supports the selected rule in the tested configuration rather than a universal optimal threshold.

C.18 EVIDENCE, GRADIENT VARIANCE, AND RESIDUAL ERROR

At round 100 of synthetic AV-MNIST training with 40% noise, we examine 17 client–modality pairs. Clean labels are used only to measure residual error. Per-example gradients use the reconstructed targets:

$$g_{k,i}^m = \nabla_{\theta_m} \mathcal{L}_{\text{EDL}}(\boldsymbol{\alpha}_i^m, y_i'), \quad \bar{g}_k^m = \frac{1}{N_k} \sum_{i \in \mathcal{D}_k} g_{k,i}^m, \quad V_k^m = \frac{1}{N_k} \sum_{i \in \mathcal{D}_k} \|g_{k,i}^m - \bar{g}_k^m\|^2. \tag{39}$$

Table 27: Spearman correlations across 17 client–modality pairs at round 100 on AV-MNIST, 40% structured noise.

Evidence statistic	Paired measurement	Spearman ρ	Reported p
Accumulated $\bar{e}_k^m$	Gradient variance V_k^m	-0.91	< 0.001
	Residual label error	-0.87	< 0.001
Mean $\bar{e}_k^m / N_k$	Gradient variance V_k^m	-0.83	< 0.01

The mean-evidence result reduces the influence of client sample count on the correlation. These measurements motivate the evidence weights in this protocol. They do not establish exact inverse proportionality, independence between pairs from the same client, or protection against confidently incorrect clients.

D ROUND-LEVEL ALGORITHM

Figure 2 gives the server–client overview. Algorithm 1 makes the warm-up boundary and ordered correction explicit. Models and prototypes are exchanged at round boundaries; raw observations and per-example uncertainty remain on clients.

Algorithm 1 FedEPR: evidential prototype reconstruction

Require: Client data $\{\mathcal{D}_k\}$, modality sets $\{\mathcal{M}_k\}$, rounds T, warm-up T_{warm}, local epochs E, coefficients in Table 8
Ensure: Global modality models and class prototypes

Initialize model parameters and prototype states
for $t = 0, \ldots, T-1$ **do**
  Select participating clients and broadcast current models and prototypes
  for participating client k in parallel **do**
    Calibrate available local class prototypes using Equation (3)
    for $e = 1, \ldots, E$ **do**
      for minibatch $\mathcal{B} \subset \mathcal{D}_k$ **do**
        Compute features, evidence, beliefs, and uncertainty for available modalities
        if $|\mathcal{M}_k| = 2$ **then**
          Compute DS fusion and fused Dirichlet parameters, Equations (4), (8)
        else
          Use the available modality as the primary prediction
        end if
        if $t < T_{\text{warm}}$ **then**
          Set $y_i' = \tilde{y}_i$ and activate available auxiliary losses
        else if $|\mathcal{M}_k| = 2$ **then**
          Compute m_i^{proto}, m_i^{unc}, and h_i
          Apply conflict, uncertainty-gap, consensus, then fallback cases in order
          Set auxiliary masks r_i^m using Equation (10)
        else
          Apply unimodal reconstruction, Equation (7)
        end if
        Optimize the original evidential objective, Equation (10)
      end for
    end for
    Recompute class prototypes and counts; accumulate $\bar{e}_k^m = \sum_i (1 - u_{k,i}^m)$
    Upload local models, class statistics, and effective evidence
  end for
  Aggregate each modality over clients that possess it, Equation (11)
  Aggregate class prototypes and advance the prototype state to round $t+1$
end for

E LIMITATIONS AND BROADER SCOPE

The method and experiments concern bimodal inputs. With more than two available modalities, pairwise conflicts form a set $\{\kappa_i^{mn} : m, n \in \mathcal{M}_k, m < n\}$ of size $\binom{|\mathcal{M}_k|}{2}$. A two-against-one split can arise, and a majority of modalities can share correlated corruption. The class-support count $v_{i,c} = \sum_{m \in \mathcal{M}_k} \mathbb{I}\{\arg\max_j b_{i,j}^m = c\}$ does not certify correctness. Sequential DS fusion can combine evidence, but does not specify which coalition should determine a reconstructed target. Extending the reconstruction rule requires a higher-order account of support, dependence, and modality reliability.

Prototype geometry must remain informative. Severe class sparsity or concept shift can weaken both local and global anchors; the 500-client stress test shows decreasing accuracy and a smaller advantage over FedAvg. The primary controlled generator also follows the three structures used by the method. Additional processes and naturally noisy UPMC and WebVision data expand the evidence, but do not establish robustness to arbitrary annotation or sensor failures. The human audit is limited to 1,500 UPMC training triplets.

Evidence measures support, not correctness. A client that confidently fits corrupted labels can receive excessive aggregation weight. The measured negative correlations with residual error and gradient variance apply to the evaluated protocol. The original fused objective also updates both modalities on high-conflict samples; the auxiliary mask does not remove this pathway. The gated diagnostic measures a small difference on two datasets but leaves the mechanism as an observed limitation.

Keeping raw data on clients is a data-locality property. Model parameters, prototypes, and counts can still reveal information. The perturbation experiment protects the additional statistics only, and the method does not provide end-to-end differential privacy or robustness against adversarial evidence manipulation. These distinctions are relevant to applications involving sensitive multimodal observations.