# DUAL-PARTIAL MULTIMODAL FEDERATED LEARNING VIA UNCERTAINTY-GATED MODALITY–LABEL EVIDENCE GRAPHS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multimodal federated learning (MFL) is particularly challenging when modalities and labels are only partially observed. Missing modalities make cross-modal transfer necessary, whereas missing labels remove the supervision needed to verify that transfer. This *evidence dilemma* makes the targets that depend most on transfer the least directly verifiable. We formulate dual-partial MFL as progressively verified evidence transfer and propose Dual-Partial Evidence-Graph Federated Learning (DualP-EGFL). DualP-EGFL constructs signed modality–label evidence prototypes from verified observations, admits a cross-modal relation only after structural and target-side utility checks, and applies an accepted relation to an unknown target only when pair-specific out-of-distribution (OOD) gating and conservative Top-$K$ path aggregation provide sufficient support. Unsupported targets remain low-evidence rather than receiving fabricated pseudo-labels. Across the evaluated benchmarks, DualP-EGFL improves supported-hidden Macro-AUPRC by up to 7.65 percentage points over recent MFL baselines, with the largest gain in the 100-client MM-IMDb setting. In calibrated smaller-client settings, it also reduces the false-support rate relative to ungated graph transfer.

## 1 INTRODUCTION

Multimodal learning is increasingly important when complementary information is distributed across vision, language, audio, and other sensing modalities. In realistic deployments, these data often reside at different institutions or edge devices and cannot be centralized because of privacy, ownership, and communication constraints. Multimodal federated learning (MFL) enables collaborative learning without exchanging raw observations. Recent studies have extended MFL to heterogeneous settings with missing modalities, client-specific modality combinations, personalized models, and large multimodal architectures.

The central difficulty is not only that modalities and labels may be missing, but also that their missingness interacts. Different clients may observe different modality subsets and supervise different semantic categories. Thus, a modality–label pair is jointly verified only when both its input and label are available. Missing modalities make cross-modal transfer necessary, whereas missing labels remove the direct supervision needed to verify that transfer. Their conjunction creates an *evidence dilemma*: the targets that need transferred supervision most are precisely those for which transfer correctness is least observable. Sparse verified overlap can still create useful transfer paths, but modality-specific cues mean that representational similarity alone cannot establish target-side discrimination.

The evidence dilemma shifts the central question from how to connect heterogeneous clients to when transferred evidence is sufficiently supported to supervise an otherwise unknown target. It creates two coupled challenges:

- **C1: Reliable relation discovery under sparse verified overlap.** A cross-modal relation inferred from a small set of jointly observed modality–label examples may be compatible in representation space, statistically supported, uncertain, or stable across clients, yet fail to preserve the target modality's decision structure. Formally, high structural reliability $r_{\mathrm{struct}}(m' \to m, c)$ does not



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.

imply high predictive transfer utility $u(m' \to m, c)$. The first challenge is to identify transferable relations using only verified federated evidence.

- **C2: Selective supervision for unverifiable targets.** A relation with high utility on verified target observations need not apply to every sample with a missing label. For a target sample $x$, global utility $u(m' \to m, c)$ does not guarantee that $x$ lies within the relation's supported region. Moreover, aggregating many individually weak paths can create artificial support. The second challenge is to determine whether a verified relation applies to a particular unknown target while preventing unsupported evidence from becoming erroneous pseudo-supervision.

To resolve this dilemma, we propose Dual-Partial Evidence-Graph Federated Learning (DualP-EGFL), a reliability-aware framework for progressively verified cross-modal transfer. The framework constructs signed modality–label evidence prototypes from verified observations and admits a candidate relation only after checking structural support and predictive utility on verified target-modality data, addressing C1. For an unknown target, it calibrates pair-specific out-of-distribution scales, scores sample-level path reliability, and aggregates only the strongest supported paths, addressing C2. When sufficient federated evidence is unavailable, the target receives no fabricated pseudo-label and is instead constrained to remain low-evidence. Relation-level verification therefore determines whether knowledge can be transferred, whereas sample-level verification determines whether it should be used for a particular target.

We evaluate DualP-EGFL under heterogeneous multimodal federated settings with varying modality availability, label observability, and cross-client overlap. In the 100-client MM-IMDb experiment, it improves supported-hidden Macro-AUPRC by 7.65 percentage points over PePSY, the strongest recent MFL baseline in that setting. In calibrated smaller-client settings, it reduces the false-support rate relative to ungated graph transfer. Our contributions are threefold: (1) we formulate dual-partial MFL as reliable knowledge transfer under an evidence dilemma; (2) we introduce DualP-EGFL, which combines verified relation discovery with selective sample-level supervision; and (3) we evaluate the framework under joint modality and supervision incompleteness through a 100-client scalability study, support-reliability analysis, and component ablations.

## 2 Related Work

**Multimodal Federated Learning under Heterogeneity.** Multimodal federated learning (MFL) enables distributed clients to exchange complementary knowledge without sharing raw data. Existing methods address heterogeneous modalities through representation alignment and contrastive learning Yu et al. (2023), topology- or hypergraph-based collaboration Chen & Zhang (2024b); Qi & Li (2024), multimodal anchors and feature transfer Zhang et al. (2024c); Dai et al. (2024), and distillation or cross-modal consensus Chen & Zhang (2024a); Chen et al. (2024); Li et al. (2024). Recent work also considers modality incongruity, missing-data patterns, and foundation-model adaptation Saha et al. (2025); Phung et al. (2025); Nguyen et al. (2026); Luo et al. (2025). In parallel, federated multi-label and partial-label learning addresses heterogeneous or ambiguous supervision through label-correlation modeling and local label disambiguation Liu et al. (2024); Yan & Guo (2024); Di et al. (2024). FedMultimodal Feng et al. (2023) jointly benchmarks modality and label corruption, but the two forms of incompleteness are usually addressed separately. Existing collaboration mechanisms primarily determine *how* to exchange heterogeneous knowledge. They do not explicitly verify whether a selected cross-modal relation preserves the target decision boundary or remains reliable for a particular sample. We study this reliability problem when missing modalities and missing supervision occur jointly.

**Prototype-Based and Reliability-Aware Knowledge Transfer.** Prototype sharing provides a compact interface for heterogeneous federated models. FedProto and FedTGP exchange class-level statistics Tan et al. (2022); Zhang et al. (2024b), whereas FedMSplit, FedMEMA, and HAMFL use anchors or relational structures for multimodal collaboration Chen & Zhang (2022); Dai et al. (2024); Qi & Li (2024). Complementary work improves confidence reliability through federated calibration, posterior inference, or evidential prediction Chu et al. (2025); Zhang et al. (2024a); Pandey & Yu (2023). These techniques quantify geometric compatibility or predictive uncertainty, but neither property alone establishes reliable transfer. A structurally similar prototype may fail to predict verified target-modality labels, and a globally useful relation may be invalid outside its supported region. DualP-EGFL therefore treats structural similarity as a candidate criterion, verifies

target-side predictive utility, and conditions transferred supervision on sample-level support. This progressive criterion distinguishes the method from similarity-based transfer and population-level confidence calibration.

# 3 PRELIMINARIES

We first formalize multimodal federated learning with jointly incomplete modalities and label supervision. We then define the evidence and reliability concepts used to determine whether a cross-modal relation is plausible, useful on verified target data, and applicable to an individual sample.

## 3.1 PROBLEM SETTING

Consider $K$ clients indexed by $\mathcal{K} = \{1, \dots, K\}$, a global modality set $\mathcal{M} = \{1, \dots, M\}$, and a global label set $\mathcal{C} = \{1, \dots, C\}$. Client $i \in \mathcal{K}$ holds a private multi-label dataset

$$\mathcal{D}_i = \{(\mathbf{x}_{in}, \mathbf{q}_{in}, \mathbf{y}_{in}, \mathbf{o}_{in})\}_{n=1}^{N_i}, \tag{1}$$

where $\mathbf{x}_{in} = \{x_{in}^m\}_{m \in \mathcal{M}}$ denotes the multimodal input, $\mathbf{q}_{in} \in \{0,1\}^M$ records modality availability, $\mathbf{y}_{in} \in \{0,1\}^C$ contains binary semantic labels, and $\mathbf{o}_{in} \in \{0,1\}^C$ records label observability. An input $x_{in}^m$ or label $y_{inc}$ is accessed during training only when its corresponding indicator equals one. All raw inputs and labels remain local throughout federated training.

**Modality availability.** The indicator $q_{inm} = 1$ means that modality $m$ is available for sample $n$ at client $i$. The observed modality set is therefore

$$\mathcal{M}_{in} = \{m \in \mathcal{M} : q_{inm} = 1\}. \tag{2}$$

Because $\mathcal{M}_{in}$ may vary across clients and samples, this formulation captures both client-level modality heterogeneity and sample-level modality missingness.

**Label observability.** The indicator $o_{inc} = 1$ means that the binary label $y_{inc} \in \{0,1\}$ is observed and can provide direct supervision. In contrast, $o_{inc} = 0$ marks the label as *unknown*, indicating missing supervision rather than a verified negative label. We assign an observed label a sign $s \in \{+,-\}$, where $s = +$ if $y_{inc} = 1$ and $s = -$ if $y_{inc} = 0$. A modality–label observation is *verified* only when both the modality and its label are observed, that is, when $q_{inm} = 1$ and $o_{inc} = 1$. Only verified observations are used to construct or validate federated evidence.

The resulting setting is *dual-partial*: each client observes only part of the global modality space and receives supervision for only part of the global label space. Clients therefore need not share identical input or supervision spaces. Their partial modality–label overlap nevertheless creates potential transfer paths. Our objective is to exploit these paths without treating unverified relations as reliable supervision.

## 3.2 MULTIMODAL EVIDENCE REPRESENTATION

Cross-modal evidence is target-dependent. A modality representation that is informative for one label need not transfer to another. We therefore index each signed evidence unit by a modality, label, and observed sign,

$$v = (m, c, s), \qquad m \in \mathcal{M}, \quad c \in \mathcal{C}, \quad s \in \{+,-\}. \tag{3}$$

Verified observations associated with $v$ induce a prototype distribution summarized by an aligned mean $\boldsymbol{\mu}_v$ and diagonal variance $\boldsymbol{\sigma}_v^2$. The server also records the number of verified samples $N_v$, the number of contributing clients $K_v$, prototype uncertainty $\nu_v$, and cross-client stability $\mathrm{stab}_v$. These statistics characterize population-level evidence for a modality–label–sign relation. They do not expose or represent an individual raw sample.

Following evidential binary classification Pandey & Yu (2023), the predictive branch outputs non-negative positive and negative evidence, $e_{inc}^+ \geq 0$ and $e_{inc}^- \geq 0$, for each label $c$. These values parameterize a Beta distribution as

$$\alpha_{inc} = 1 + e_{inc}^+, \qquad \beta_{inc} = 1 + e_{inc}^-. \tag{4}$$

The corresponding predictive mean and evidential uncertainty are

$$p_{inc} = \frac{\alpha_{inc}}{\alpha_{inc} + \beta_{inc}}, \qquad \nu_{inc} = \frac{2}{\alpha_{inc} + \beta_{inc}}. \tag{5}$$

Thus, the model represents both its class probability and the amount of evidence supporting that prediction. In particular, low total evidence $e^+_{inc} + e^-_{inc}$ yields high uncertainty rather than an artificially confident label.

### 3.3 Reliable Cross-Modal Transfer

We distinguish three progressively stronger notions of transfer reliability. Consider a directed relation $m' \to m$ for label $c$, where evidence from source modality $m'$ may provide supervision for a sample represented in target modality $m$.

**Structural reliability.** Structural reliability $r_{\text{struct}}(m' \to m, c)$ measures whether signed source and target evidence is sufficiently supported and compatible in the aligned representation space. It combines verified-anchor availability with prototype support, geometric compatibility, uncertainty, and cross-client stability. Structural reliability therefore identifies plausible transfer candidates, but does not establish that source evidence preserves the target modality's decision boundary.

**Verified transfer utility.** For a structurally admissible relation, verified transfer utility $u(m' \to m, c)$ measures how well source evidence discriminates positive from negative examples among verified target-modality observations. It therefore tests whether a structurally plausible relation is useful for target-side prediction. This check is necessary because geometric compatibility does not imply preservation of the target decision structure.

**Sample-level applicability.** A relation with high verified utility may still be unreliable for a particular target sample. For an admissible transfer path $j$ to sample–label pair $(i, n, c)$, we denote its effective sample-level reliability by $g^{\text{path}}_{inc,j}$. This quantity discounts pair-level reliability when the target representation lies outside the verified region supporting the relation. Because a sample–label pair can receive evidence through multiple admissible paths, $G_{inc}$ denotes their conservatively aggregated reliability mass.

An unknown target is *federatively supported* only if $G_{inc} \geq \tau_g$, where $\tau_g$ is the minimum support threshold. Otherwise, it remains *unsupported*. This decision specifies whether transferred evidence may be used as supervision. It is not itself a pseudo-label, and an unsupported target is never interpreted as a negative instance.

## 4 Method

### 4.1 Overview

DualP-EGFL addresses the two reliability gaps in Section 3.3 through two consecutive checks. Edge-level verification first determines whether a cross-modal relation is structurally plausible and useful on verified target-modality data, addressing C1. Sample-level verification then determines whether an accepted relation applies to the current sample and whether its evidence is sufficient for supervision, addressing C2. The framework uses only observed modalities and verified training labels to construct and calibrate the graph. Unknown labels and test labels are never used.

For each available modality $m$, client $i$ forms

$$\mathbf{z}^m_{in,\text{raw}} = P_m(E_m(x^m_{in})), \mathbf{z}^m_{in,\text{graph}} = R_m \mathbf{z}^m_{in,\text{raw}}, \mathbf{z}^m_{in,\text{personal}} = \mathbf{z}^m_{in,\text{raw}} + A_{i,m}(\mathbf{z}^m_{in,\text{raw}}), \tag{6}$$

where $R_m$ aligns modality $m$ with the shared graph space and $A_{i,m}$ is a private adapter. The authoritative branch $\mathbf{z}_{\text{graph}}$ is used only for graph construction and verification, whereas the personalized branch $\mathbf{z}_{\text{personal}}$ produces local predictions. For sign $s \in \{+, -\}$, modality-specific evidence is

$$e^{m,s}_{inc} = \text{softplus}\Big(\mathbf{w}^{s\top}_c \mathbf{z}^m_{in,\text{personal}} + b^s_{i,c}\Big). \tag{7}$$

A shared uncertainty-aware fusion operator combines available modality evidence into the Beta parameters $(\alpha_{inc}, \beta_{inc})$ defined in Section 3.2.

### 4.2 EDGE-LEVEL TRANSFER VERIFICATION

**Structurally reliable evidence graph.** To address C1, the server first constructs candidate relations from verified signed prototypes. A client contributes to node $v = (m, c, s)$ only when $q_{inm} = o_{inc} = 1$ and the observed label has sign $s$. Clients upload sufficient statistics, from which the server estimates the aligned prototype $(\boldsymbol{\mu}_v, \boldsymbol{\sigma}_v^2)$, support counts $(N_v, K_v)$, uncertainty $\nu_v$, and cross-client stability $\mathrm{stab}_v$. Orthogonal transformations $\{R_m\}$ are updated from verified paired anchors using generalized Procrustes alignment.

For a signed edge $a = (m', c, s) \to b = (m, c, s)$, structural reliability is

$$r_{\mathrm{struct}}(a, b) = I_{\mathrm{anchor}} I_{\mathrm{eligible}} \exp\left(-\frac{\widetilde{D}_{\mathrm{Mah}}(a, b)}{2}\right) \rho_{\mathrm{count}} \rho_{\mathrm{unc}} \rho_{\mathrm{stab}}, \tag{8}$$

where the indicator terms enforce anchor and support requirements, $\widetilde{D}_{\mathrm{Mah}}$ is the normalized diagonal Mahalanobis distance, and the three $\rho$ terms measure sample support, evidential certainty, and cross-client stability. Because binary transfer requires both sides of the decision structure, we define

$$r_{\mathrm{pair}}^{\mathrm{struct}}(m', m, c) = \min_{s \in \{+,-\}} r_{\mathrm{struct}}((m', c, s), (m, c, s)), \tag{9}$$

and discard pairs below the structural threshold $\tau_e$.

**Verified transfer utility.** Structural compatibility alone does not guarantee that source evidence preserves the target decision boundary. We therefore evaluate each surviving pair on verified target-modality samples. For $\mathbf{z}_{in,\mathrm{graph}}^m$, let $d_{inc}^+$ and $d_{inc}^-$ be its variance-normalized distances to the source prototypes $(m', c, +)$ and $(m', c, -)$. The source-induced prediction is

$$p_{inc}^{\mathrm{tr}}(m' \to m, c) = \sigma\left(\frac{d_{inc}^- - d_{inc}^+}{\tau_d}\right). \tag{10}$$

Using only verified target samples and requiring both classes, the server computes

$$u(m', m, c) = \frac{1}{2}\mathrm{BA}(m', m, c) + \frac{1}{2}\,\mathrm{clip}\left(1 - \frac{\mathrm{Brier}_{\mathrm{bal}}(m', m, c)}{0.25}, 0, 1\right). \tag{11}$$

The final pair reliability, denoted by $r_{\mathrm{pair}}(m', m, c)$, is the thresholded structural score multiplied by a utility factor $\phi(u)$. A measured zero utility removes the pair, utility below $\tau_{\mathrm{transfer}}$ down-weights it, and insufficient calibration data temporarily retain the structural score. An edge therefore becomes transferable only after passing both representation-level and target-prediction tests.

### 4.3 SAMPLE-LEVEL TRANSFER APPLICABILITY

**Pair-specific applicability and conservative support.** To address C2, DualP-EGFL calibrates a separate OOD scale for each active pair. From verified target samples, the server sets

$$\tau_{\mathrm{ood}}(m', m, c) = \mathrm{clip}\big(Q_{0.90}\{\min(d_{inc}^+, d_{inc}^-)\}, \tau_{\mathrm{ood}}^{\min}, \tau_{\mathrm{ood}}^{\max}\big). \tag{12}$$

For path $j = (m' \to m, c)$, its reliability on the current sample is

$$g_{inc,j}^{\mathrm{path}} = r_{\mathrm{pair}}(m', m, c) \exp\left(-\frac{\min(d_{inc}^+, d_{inc}^-)}{\tau_{\mathrm{ood}}(m', m, c)}\right). \tag{13}$$

Rather than summing all available paths, we retain the $K_{\mathrm{path}}$ largest values. For their index set $\mathcal{I}_{inc}$,

$$G_{inc} = \sum_{j \in \mathcal{I}_{inc}} g_{inc,j}^{\mathrm{path}}, \qquad \mu_{inc}^T = \frac{\sum_{j \in \mathcal{I}_{inc}} g_{inc,j}^{\mathrm{path}} p_{inc,j}^{\mathrm{tr}}}{G_{inc} + \epsilon}. \tag{14}$$

An unknown label is federatively supported only when $G_{inc} \geq \tau_g$ and at least one federated evidence source exists for label $c$. Top-$K$ aggregation prevents many individually weak paths from producing artificial support.

**Detached evidential teacher.** For a supported unknown label, reliability determines teacher concentration:

$$\kappa_{inc} = \kappa_{\max} \min(1, G_{inc}), \qquad \alpha_{inc}^T = 1 + \kappa_{inc}\mu_{inc}^T, \qquad \beta_{inc}^T = 1 + \kappa_{inc}(1 - \mu_{inc}^T). \tag{15}$$

The teacher is detached before optimization. Unsupported labels receive no transferred target and remain subject only to a low-evidence constraint. They are never interpreted as negative labels.

### 4.4 Federated Optimization

For client $i$, let $\mathcal{S}_i$ and $\mathcal{U}_i$ denote its supported and unsupported unknown entries. The local objective is

$$\mathcal{L}^{(i)} = \mathcal{L}_{\text{obs}}^{(i)} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}^{(i)} + \lambda_{\text{trans}}\mathcal{L}_{\text{trans}}^{(i)} + \lambda_{\text{unc}}\mathcal{L}_{\text{unsupported}}^{(i)}, \tag{16}$$

where $\mathcal{L}_{\text{obs}}^{(i)}$ is the masked evidential loss on observed labels and $\mathcal{L}_{\text{align}}^{(i)}$ aligns authoritative representations with detached verified prototypes. The two unknown-label losses are

$$\mathcal{L}_{\text{trans}}^{(i)} = \frac{1}{\max(1, |\mathcal{S}_i|)} \sum_{(n,c)\in\mathcal{S}_i} w_{inc}^T D_{\text{KL}}\big(\text{Beta}(\alpha_{inc}^T, \beta_{inc}^T) \parallel \text{Beta}(\alpha_{inc}, \beta_{inc})\big), \tag{17}$$

$$\mathcal{L}_{\text{unsupported}}^{(i)} = \frac{1}{\max(1, |\mathcal{U}_i|)} \sum_{(n,c)\in\mathcal{U}_i} \left[\log(1 + \text{ReLU}(\alpha_{inc} + \beta_{inc} - s_0))\right]^2. \tag{18}$$

Here $w_{inc}^T = \min(1, G_{inc})$. The transfer and unsupported losses update only private personalization parameters. Shared encoders and predictors are trained with observed supervision and verified alignment. This gradient separation prevents graph-generated targets from directly rewriting the representation used to validate the graph.

Each communication round alternates local optimization with shared-parameter aggregation. On graph-refresh rounds, clients additionally upload verified counts and moments. The server then updates alignments, prototypes, edge utilities, and pair-specific OOD thresholds. Raw inputs and unknown labels remain local. Constructing the structural graph costs $\mathcal{O}(M^2Cd)$, while calibrating $P$ active pairs over $N_{\text{cal}}$ verified target samples costs $\mathcal{O}(PN_{\text{cal}}d)$. These information restrictions reduce exchanged data but do not constitute a formal privacy guarantee.

## 5 Experiments

### 5.1 Experimental Setup

**Datasets and protocol.** We evaluate DualP-EGFL on MM-IMDb Arevalo et al. (2017), MELD Poria et al. (2019), and a controlled synthetic benchmark. MM-IMDb uses eight genre labels in a six-client main setting and a 100-client scalability setting. MELD contains four emotion labels distributed across six clients. The synthetic benchmark contains nine clients, three modalities, and six labels, with misleading cross-modal signals and out-of-distribution samples injected. All benchmarks follow the dual-partial protocol with client-level and sample-level modality and label missingness. Unknown labels are masked during training and revealed only for evaluation. Detailed preprocessing, partitioning, and missingness configurations are provided in Appendix A.1.

**Baselines and implementation.** We compare against local training, FedAvg, FedProx, CreamFL Yu et al. (2023), PePSY Nguyen et al. (2026), and HAMFed Qi & Li (2024). We also include two graph controls: *Prototype-Ungated*, which removes relation gating, and *Full-Graph*, which removes edge and sample filtering. All methods use identical data splits, feature backends, model capacity, and communication schedules. MM-IMDb and the synthetic benchmark are trained for 16 rounds, whereas MELD is trained for 12 rounds. Each round uses one local epoch, and the 100-client experiment uses full participation. Results use seed 7 for the real datasets and seed 1 for the synthetic benchmark. We therefore report observed results without statistical-significance claims.

Table 1: Macro-AUPRC (%) under joint modality and label incompleteness. "Supp.-H" denotes supported hidden entries. All methods within each dataset use the same split and seed. Best results are bold.

| Dataset | Method | Observed ↑ | Hidden ↑ | Supp.-H ↑ | All ↑ |
|---|---|---|---|---|---|
| Synthetic | FedAvg | 99.91 | 94.18 | 93.86 | 96.07 |
| | PePSY | 99.82 | 95.99 | 95.09 | 97.11 |
| | **DualP-EGFL** | **99.97** | **97.38** | **97.01** | **98.28** |
| MELD | FedAvg | 29.60 | 24.91 | 28.19 | 28.23 |
| | CreamFL | 30.26 | 24.40 | 26.81 | 27.58 |
| | PePSY | **30.61** | 25.13 | 28.03 | 28.03 |
| | HAMFed | 28.78 | 24.31 | 27.48 | 27.45 |
| | Prototype-Ungated | 29.66 | 25.49 | 28.16 | 27.54 |
| | Full-Graph | 29.36 | 24.06 | 26.56 | 26.44 |
| | **DualP-EGFL** | 30.26 | **25.78** | **28.36** | **28.50** |
| MM-IMDb | Local | **59.26** | 47.06 | 57.84 | 50.47 |
| | FedAvg | 56.08 | 48.31 | 58.09 | 49.53 |
| | Prototype-Ungated | 58.25 | 49.68 | 59.91 | 51.41 |
| | Full-Graph | 57.16 | 48.91 | 59.42 | 50.58 |
| | **DualP-EGFL** | 59.00 | **52.02** | **63.04** | **53.69** |

**Metrics.** Our primary metric is Macro-AUPRC on *supported hidden* entries, which measures recovery of unknown labels with available federated evidence. We additionally report Macro-AUPRC on observed, hidden, and all entries, together with support-detection AUROC, transfer coverage, and false-support rate (FSR).

### 5.2 Main Results

**Hidden-label prediction.** Table 1 shows that DualP-EGFL provides its clearest benefit on labels lacking local supervision. On MM-IMDb, it reaches $63.04\%$ supported-hidden Macro-AUPRC, improving over Prototype-Ungated by $3.14$ percentage points. It also improves Macro-AUPRC on all hidden entries by $2.33$ points and on all entries by $2.28$ points, while observed-label performance remains within $0.26$ points of the best local result. On the synthetic benchmark, DualP-EGFL improves supported-hidden and all-entry Macro-AUPRC over PePSY by $1.92$ and $1.18$ points, respectively. These results support the intended role of federated evidence, namely, improving targets without direct local supervision rather than merely fitting observed labels.

**Performance on MELD.** The MELD setting is more difficult and yields smaller margins. DualP-EGFL obtains the best hidden, supported-hidden, and all-entry Macro-AUPRC, exceeding the strongest baseline by $0.29$, $0.17$, and $0.26$ percentage points, respectively. It does not outperform PePSY on observed labels, where the difference is $0.35$ points. This pattern is consistent with the method's objective, but the small single-seed margin remains preliminary until multi-seed uncertainty estimates are available.

### 5.3 Scaling to 100 Clients

**Large-scale hidden-label recovery.** Table 2 shows that the advantage of DualP-EGFL increases when the same modality–label evidence is distributed across 100 partial clients. Relative to the strongest recent MFL baseline for each metric, DualP-EGFL improves hidden, supported-hidden, and all-entry Macro-AUPRC by $6.08$, $7.65$, and $5.46$ percentage points, respectively. Supported-hidden Macro-AUPRC increases from PePSY's $44.69\%$ to $52.34\%$, a $7.65$-point or $17.11\%$ relative improvement. DualP-EGFL remains $0.90$ points below HAMFed on observed entries. This pattern indicates that its advantage comes from recovering locally missing supervision rather than uniformly improving already observed labels.

Table 2: Macro-AUPRC (%) on the 100-client MM-IMDb experiment (seed 7). All methods use the same dual-partial split and full client participation. Best results are bold and the best recent MFL baseline is underlined.

| Method | Observed ↑ | Hidden ↑ | Supp.-H ↑ | All ↑ |
|---|---|---|---|---|
| Local | 70.03 | 47.89 | 46.77 | 54.58 |
| FedAvg | 62.80 | 42.80 | 41.18 | 47.96 |
| FedProx | 34.31 | 24.98 | 26.68 | 27.06 |
| CreamFL | 74.57 | 46.59 | 44.60 | <u>53.37</u> |
| PePSY | 74.07 | <u>46.66</u> | <u>44.69</u> | 53.31 |
| HAMFed | **74.59** | 46.51 | 44.50 | 53.30 |
| Prototype-Ungated | 69.19 | 46.42 | 45.87 | 52.01 |
| Full-Graph | 69.68 | 46.70 | 46.08 | 52.27 |
| **DualP-EGFL** | 73.69 | **52.74** | **52.34** | **58.83** |

Table 3: Support quality on the real datasets (%). Coverage is the fraction of hidden entries accepted for transfer, and FSR is the fraction of unsupported entries accepted incorrectly.

| Dataset | Method | AUROC ↑ | Coverage | FSR ↓ |
|---|---|---|---|---|
| MM-IMDb (6) | Prototype-Ungated | 48.77 | 100.00 | 100.00 |
| | Full-Graph | 48.35 | 100.00 | 100.00 |
| | DualP-EGFL | **78.14** | 75.41 | **55.01** |
| MELD | Prototype-Ungated | 66.17 | 100.00 | 100.00 |
| | Full-Graph | 60.55 | 94.39 | 97.08 |
| | DualP-EGFL | 60.79 | 3.76 | **0.68** |

### 5.4 Transfer Reliability

**Selective support controls negative transfer.** Table 3 reveals a failure hidden by aggregate prediction metrics. Both ungated graph variants accept nearly every hidden entry, including unsupported ones. In the calibrated six-client MM-IMDb setting, DualP-EGFL reduces FSR from $100\%$ to $55.01\%$ and improves support AUROC to $78.14\%$. On MELD, the strict gate reduces FSR to $0.68\%$ but accepts only $3.76\%$ of hidden entries. This result demonstrates conservative rejection rather than broad recovery and exposes a coverage–reliability trade-off when verified cross-modal overlap is scarce. Conversely, the 100-client configuration accepts $99.99\%$ of hidden entries and has a $100\%$ FSR despite its strong prediction results. Predictive scalability therefore does not yet imply reliable support calibration at this scale.

### 5.5 Ablation Study

**Both verification levels contribute.** Table 4 links the ablations to the two challenges in Section 4.1. Removing anchor, uncertainty, or stability information from edge-level verification decreases supported-hidden Macro-AUPRC by 2.25–2.68 points. Disabling the sample-level OOD gate causes a 2.34-point drop, whereas removing graph supervision or the low-evidence constraint loses 2.36 and 2.63 points, respectively. These similar degradations indicate that structural edge verification and sample-specific applicability are complementary.

**Current evaluation scope.** Figure 1 extends the component ablation with two 10-seed robustness sweeps on the balanced MM-IMDb protocol. Increasing dual-partial severity reduces supported-hidden and all-hidden Macro-AUPRC, transfer coverage, and false-support rate, indicating that scarcer verified evidence limits both recovery and accepted support. In contrast, increasing the sample-support threshold $\tau_g$ leaves Macro-AUPRC nearly unchanged while sharply reducing coverage and false-support rate. These results make the coverage–reliability trade-off explicit. Multi-seed uncertainty estimates for all baselines and broader cross-dataset severity and threshold studies remain outside the current evaluation scope.

Table 4: Component ablation on the balanced MM-IMDb setting (seed 7). $\Delta$ is the change in supported-hidden Macro-AUPRC relative to the full method.

| Variant | Supp.-H ↑ | $\Delta$ |
|---|---|---|
| DualP-EGFL | **58.88** | – |
| w/o paired anchors | 56.63 | −2.25 |
| w/o uncertainty-aware edges | 56.20 | −2.68 |
| w/o stability factor | 56.53 | −2.34 |
| w/o sample OOD gate | 56.53 | −2.34 |
| w/o graph distillation | 56.51 | −2.36 |
| w/o unsupported regularization | 56.25 | −2.63 |

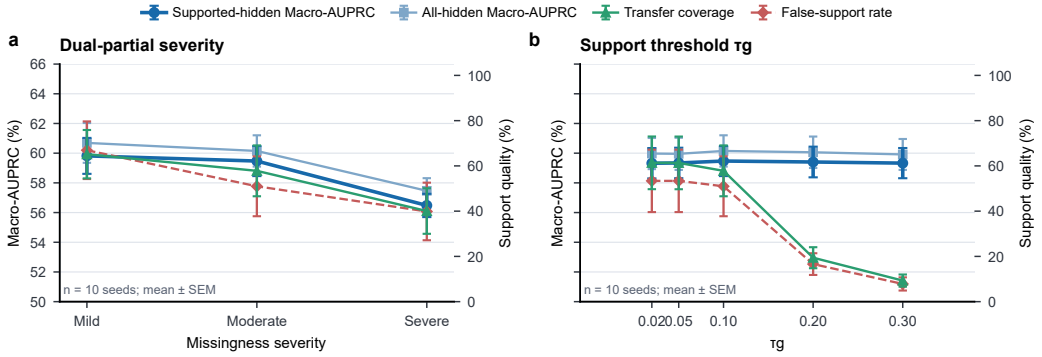



Figure 1: Multi-seed robustness analyses on the balanced MM-IMDb protocol. (a) Increasing dual-partial severity reduces supported-hidden and all-hidden Macro-AUPRC and decreases transfer coverage, while false-support rate also declines as the available verified evidence becomes more limited. (b) Increasing the sample-support threshold $\tau_g$ leaves Macro-AUPRC nearly unchanged but sharply reduces both transfer coverage and false-support rate, exposing the coverage–reliability trade-off. Curves show means and error bars show $\pm$ one SEM over 10 seeds; Macro-AUPRC uses the left axis and support-quality metrics use the right axis.

## 6 Conclusion

This work addresses reliable knowledge transfer in multimodal federated learning with jointly missing modalities and labels. DualP-EGFL verifies cross-modal transfer at both relation and sample levels while avoiding unsupported supervision. Experiments on synthetic, MELD, and MM-IMDb benchmarks show improved hidden-label prediction, and the ablations support complementary roles for edge-level verification and sample-level reliability control. The current evaluation is limited to controlled missingness and predominantly single-seed comparisons. The 100-client setting also reveals challenges in false-support control at scale. Future work should examine natural missingness, multi-seed uncertainty, and scale-adaptive calibration to balance transfer coverage and reliability.

### AI USE STATEMENT

In this work, we used generative AI tools for [generating synthetic datasets, assisting with implementation, and translation]. We did not use generative AI tools for [other tasks requiring disclosure], and [the remaining disclosure categories] are not applicable. We also used generative AI tools for [suggesting experimental parameters and identifying research topics or gaps]. We reviewed all AI-assisted work. [Complete this disclosure with the verification procedure before submission.] We take responsibility for the final content, including all claims and artifacts produced with AI assistance.

### REPRODUCIBILITY STATEMENT

The complete implementation, configuration files, and evaluation scripts for reproducing the reported experiments will be released at `https://anonymous.4open.science/r/DualP-MMFL-578F`. The release will use the dual-partial protocols, fixed data splits, feature backends, communication schedules, and random seeds described in the paper. The theoretical claims are documented in Appendix A.2. Proposition 1 gives a finite-sample concentration result for verified transfer utility, Proposition 2 establishes immunity to weak-path accumulation, Theorem 1 gives a sufficient support-separation condition, and Propositions 3–4 characterize the reliability-controlled evidential teacher and the supported and unsupported label objectives. Their assumptions and complete proofs are provided with the statements.

The experiments use three explicitly described data sources. MM-IMDb is evaluated as eight-label genre prediction from poster images and plot text in both six-client and 100-client settings. MELD uses MobileBERT text features and MFCC audio features for four emotion labels across six clients. The controlled synthetic benchmark contains nine clients, three modalities, and six labels, with misleading cross-modal signals and out-of-distribution samples injected to test transfer reliability. For every benchmark, modality and label availability is generated by the dual-partial protocol. Unknown labels are masked during training and revealed only for evaluation. The supplementary release will include the preprocessing, partitioning, masking, and evaluation commands needed to reconstruct these settings.

### ACKNOWLEDGMENTS

Acknowledgments and funding information will be added in the final version.

## REFERENCES

John Arevalo, Thamar Solorio, Manuel Montes-y Gomez, and Fabio A. Gonzalez. Gated multimodal units for information fusion. In *International Conference on Learning Representations Workshop*, 2017.

Haokun Chen, Yao Zhang, Denis Krompass, Jindong Gu, and Volker Tresp. Feddat: An approach for foundation model finetuning in multi-modal heterogeneous federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 11285–11293, 2024.

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.

Jiayi Chen and Aidong Zhang. On disentanglement of asymmetrical knowledge transfer for modality-task agnostic federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11311–11319, 2024a.

Jiayi Chen and Aidong Zhang. Fedmbridge: Bridgeable multimodal federated learning. In *ICML*, pp. 7667–7686, 2024b.

Yun-Wei Chu, Dong-Jun Han, Seyyedali Hosseinalipour, and Christopher Brinton. Unlocking the potential of model calibration in federated learning. In *International Conference on Learning Representations*, volume 2025, pp. 99316–99343, 2025.

Qian Dai, Dong Wei, Hong Liu, Jinghan Sun, Liansheng Wang, and Yefeng Zheng. Federated modality-specific encoders and multimodal anchors for personalized brain tumor segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1445–1453, 2024.

Zonglin Di, Zhaowei Zhu, Xiaoxiao Li, and Yang Liu. Federated learning with local openset noisy labels. In *European conference on computer vision*, pp. 38–56. Springer, 2024.

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, 2023.

Shuai Li, Fan Qi, Zixin Zhang, and Changsheng Xu. Cross-modal meta consensus for heterogeneous federated learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 975–984, 2024.

I-Jieh Liu, Ci-Siang Lin, Fu-En Yang, and Yu-Chiang Frank Wang. Language-guided transformer for federated multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13882–13890, 2024.

Jun Luo, Chen Chen, and Shandong Wu. Mixture of experts made personalized: Federated prompt learning for vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 58061–58080, 2025.

Duong Nguyen, Nghia Hoang, Thanh Trung Huynh, Quoc Viet Hung Nguyen, and Phi Le Nguyen. Learning reconfigurable representations for multimodal federated learning with missing data. *Advances in Neural Information Processing Systems*, 38:28808–28845, 2026.

Deep Shankar Pandey and Qi Yu. Learn to accumulate evidence from all training samples: theory and practice. In *International Conference on Machine Learning*, pp. 26963–26989. PMLR, 2023.

Thu Hang Phung, Duong M Nguyen, Thanh Trung Huynh, Quoc Viet Hung Nguyen, Trong Nghia Hoang, and Phi Le Nguyen. Federated prompt-tuning with heterogeneous and incomplete multimodal client data. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3936–3946. IEEE, 2025.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, 2019.

Fan Qi and Shuai Li. Adaptive hyper-graph aggregation for modality-agnostic federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12312–12321, 2024.

Pramit Saha, Divyanshu Mishra, Felix Wagner, Konstantinos Kamnitsas, and J Alison Noble. Incongruent multimodal federated learning for medical vision and language-based multi-label disease detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 28331–28339, 2025.

Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 8432–8440, 2022.

Yan Yan and Yuhong Guo. Federated partial label learning with local-adaptive augmentation and regularization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 16272–16280, 2024.

Qiying Yu, Yang Liu, Yimu Wang, Ke Xu, and Jingjing Liu. Multimodal federated learning via contrastive representation ensemble. *arXiv preprint arXiv:2302.08888*, 2023.

Hao Zhang, Chenglin Li, Nuowen Kan, Ziyang Zheng, Wenrui Dai, Junni Zou, and Hongkai Xiong. Improving generalization in federated learning with model-data mutual information regularization: A posterior inference approach. *Advances in Neural Information Processing Systems*, 37: 136646–136678, 2024a.

Jianqing Zhang, Yang Liu, Yang Hua, and Jian Cao. Fedtgp: Trainable global prototypes with adaptive-margin-enhanced contrastive learning for data and model heterogeneity in federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 16768–16776, 2024b.

Zixin Zhang, Fan Qi, and Changsheng Xu. Enhancing storage and computational efficiency in federated multimodal learning for large-scale models. In *ICML*, pp. 59685–59699, 2024c.

# A Appendix

## A.1 Experimental Setup and Dual-Partial Protocol

All experiments use the dual-partial protocol, which independently masks modality inputs and label supervision at the client and sample levels. Table 5 summarizes the missingness configurations. A missing label denotes unavailable supervision rather than a verified negative label. During training, unknown labels are masked from the loss and revealed only for evaluation.

Table 5: Dual-partial missingness configurations used in the experiments. "C-M" and "C-L" denote client-level modality and label missingness, respectively; "S-M" and "S-L" denote the corresponding sample-level rates.

| Setting | C-M | C-L | S-M | S-L |
|---|---|---|---|---|
| MM-IMDb (6 clients) | 0.55 | 0.70 | 0.18 | 0.55 |
| MM-IMDb (100 clients) | 0.45 | 0.65 | 0.18 | 0.55 |
| MELD (6 clients) | 0.20 | 0.28 | 0.08 | 0.10 |
| Synthetic (9 clients) | 0.55 | 0.70 | 0.18 | 0.55 |

MM-IMDb is evaluated as eight-label genre prediction from poster images and plot text. The six-client setting is used for the main comparison. The 100-client setting assigns 120 training and 50 test samples to each client, yielding 40,000 test label entries. MELD uses MobileBERT text features and MFCC audio features for four-way emotion prediction across six clients. The synthetic benchmark contains nine clients, three modalities, and six labels. It includes misleading cross-modal signals and out-of-distribution samples to test selective transfer. Its missingness rates match those of the six-client MM-IMDb setting.

## A.2 Theoretical Properties of Reliable Transfer

This section establishes properties of the two reliability checks used by DualP-EGFL. We first give a finite-sample guarantee for target-side utility calibration. We then analyze conservative path aggregation, the evidential teacher, and the losses applied to unknown labels. These results characterize when the reliability controls behave as intended. They do not claim global convergence of the non-convex federated optimization problem.

### A.2.1 Concentration of Verified Transfer Utility

Consider a fixed modality–label pair $(m', m, c)$ and its source-induced predictor in equation 10. Let the predictor be fixed independently of a calibration set containing $n_+$ positive and $n_-$ negative verified target-modality examples, drawn independently within each class. Denote $n_{\min} = \min(n_+, n_-)$. Let $\widehat{\mathrm{BA}}$ and $\widehat{\mathrm{Brier}}_{\mathrm{bal}}$ denote empirical balanced accuracy and balanced Brier score. Let $\mathrm{BA}^\star$ and $\mathrm{Brier}^\star_{\mathrm{bal}}$ denote the corresponding population quantities. Define

$$h(b) = \mathrm{clip}(1 - 4b, 0, 1), \qquad u^\star = \tfrac{1}{2}\mathrm{BA}^\star + \tfrac{1}{2}h(\mathrm{Brier}^\star_{\mathrm{bal}}), \tag{19}$$

and let $\widehat{u}$ be the empirical utility in equation 11.

**Proposition 1 (finite-sample utility calibration).** For any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$|\widehat{u} - u^\star| \le \frac{5}{2}\sqrt{\frac{\log(8/\delta)}{2n_{\min}}}. \tag{20}$$

*Proof.* Balanced accuracy and balanced Brier score are each averages of two class-conditional empirical means. Their four underlying random variables, two correctness indicators and two squared errors, lie in $[0,1]$. By Hoeffding's inequality and a union bound, with probability at least $1 - 8\exp(-2n_{\min}t^2)$, all four empirical means differ from their expectations by at most $t$. On this event,

$$\left|\widehat{\mathrm{BA}} - \mathrm{BA}^\star\right| \le t, \qquad \left|\widehat{\mathrm{Brier}}_{\mathrm{bal}} - \mathrm{Brier}^\star_{\mathrm{bal}}\right| \le t. \tag{21}$$

The clipped function $h$ is 4-Lipschitz. Therefore,

$$|\widehat{u} - u^\star| \le \tfrac{1}{2}|\widehat{\mathrm{BA}} - \mathrm{BA}^\star| + \tfrac{1}{2}|h(\widehat{\mathrm{Brier}}_{\mathrm{bal}}) - h(\mathrm{Brier}^\star_{\mathrm{bal}})| \tag{22}$$

$$\le \tfrac{1}{2}t + 2t = \tfrac{5}{2}t. \tag{23}$$

Setting $t = \sqrt{\log(8/\delta)/(2n_{\min})}$ proves equation 20. □

Proposition 1 implies that an edge whose population utility is separated from $\tau_{\mathrm{transfer}}$ by more than the right-hand side of equation 20 will, with probability at least $1-\delta$, fall on the correct side of the empirical utility threshold. The independence condition can be enforced with a held-out or cross-fitted verified calibration subset. Without it, this bound does not account for adaptive reuse of the same samples to construct and evaluate a prototype.

### A.2.2 Conservative Sample-Level Support

For a fixed unknown entry $(i,n,c)$, abbreviate the candidate path reliabilities in equation 13 as $g_1,\dots,g_P \ge 0$. Let $K = K_{\mathrm{path}}$, and let $G$ be the sum of the $\min(K,P)$ largest values, as in equation 14.

**Proposition 2 (immunity to weak-path accumulation).** If $g_j \le \eta$ for every candidate path, then

$$G \le \min(K,P)\eta \le K\eta. \tag{24}$$

Consequently, if $K\eta < \tau_g$, the entry cannot be declared supported, regardless of the total number $P$ of weak candidate paths.

*Proof.* The retained index set contains at most $K$ paths, each bounded above by $\eta$. Summing these terms yields equation 24. The support condition $G \ge \tau_g$ then cannot hold when $K\eta < \tau_g$. □

Unlike an unrestricted sum, whose upper bound $P\eta$ grows with graph density, the bound in Proposition 2 depends only on $K$. Increasing the number of individually weak relations therefore cannot create support by itself.

**Theorem 1 (sufficient support-separation condition).** Write each path score as

$$g_j = r_j \exp(-d_j/\tau_j), \tag{25}$$

where $r_j$ is its verified pair reliability, $d_j = \min(d^+_{inc}, d^-_{inc})$, and $\tau_j$ is its pair-specific OOD scale. Assume at least $K$ candidate paths exist. Suppose an in-support entry has at least one federated evidence source for label $c$ and at least $K$ paths satisfying

$$r_j \ge r_{\mathrm{in}}, \qquad d_j \le d_{\mathrm{in}}, \qquad \tau_j \ge \underline{\tau}, \tag{26}$$

whereas every path of an out-of-support entry satisfies

$$r_j \le r_{\mathrm{out}}, \qquad d_j \ge d_{\mathrm{out}}, \qquad \tau_j \le \overline{\tau}. \tag{27}$$

If

$$K r_{\mathrm{out}} \exp(-d_{\mathrm{out}}/\overline{\tau}) < \tau_g \le K r_{\mathrm{in}} \exp(-d_{\mathrm{in}}/\underline{\tau}), \tag{28}$$

then the in-support entry is accepted and the out-of-support entry is rejected by the sample-level support rule.

*Proof.* Each of the $K$ stated in-support paths obeys

$$g_j \geq r_{\text{in}} \exp(-d_{\text{in}}/\underline{\tau}). \tag{29}$$

Because the Top-$K$ operator selects the largest scores, its sum is at least the sum of these $K$ lower bounds. Conversely, every out-of-support path obeys

$$g_j \leq r_{\text{out}} \exp(-d_{\text{out}}/\overline{\tau}), \tag{30}$$

so Proposition 2 upper-bounds its retained mass by $K$ times this quantity. Equation 28 places $\tau_g$ between the two bounds. The assumed federated evidence source completes the support requirement for the in-support entry, while the out-of-support entry already fails the reliability-mass requirement. □

Theorem 1 makes the roles of the two verification levels explicit. Edge-level verification enlarges the reliability gap $r_{\text{in}} - r_{\text{out}}$, whereas the sample-level OOD gate enlarges the effective distance gap. Either source of separation can compensate for the other, but both contribute to a feasible threshold interval in equation 28.

### A.2.3 Conservativeness of the Evidential Teacher

**Proposition 3 (reliability-controlled teacher strength).** Let $\mu = \mu_{inc}^T \in [0, 1]$ and $\kappa = \kappa_{inc}$ in equation 15. The concentration $S^T$, predictive mean $\bar{\mu}^T$, and evidential uncertainty $\nu^T$ of the teacher satisfy

$$S^T = \alpha_{inc}^T + \beta_{inc}^T = 2 + \kappa, \tag{31}$$

$$\bar{\mu}^T = \frac{\alpha_{inc}^T}{S^T} = \frac{1}{2} + \frac{\kappa}{\kappa + 2}\left(\mu - \frac{1}{2}\right), \tag{32}$$

$$\nu^T = \frac{2}{S^T} = \frac{2}{2 + \kappa}. \tag{33}$$

Hence $|\bar{\mu}^T - 1/2| \leq |\mu - 1/2|$: the Beta teacher is never more extreme than its aggregated path prediction. For fixed $\mu$, increasing $G_{inc}$ up to one monotonically increases concentration, decreases uncertainty, and moves the teacher mean from $1/2$ toward $\mu$. All three effects saturate for $G_{inc} \geq 1$.

*Proof.* Substituting the two parameters from equation 15 gives $S^T = 2 + \kappa$ and the identities in equation 33. Since $\kappa/(\kappa + 2) \in [0, 1)$ and is increasing in $\kappa$, the shrinkage and monotonicity statements follow. Finally, $\kappa = \kappa_{\max} \min(1, G_{inc})$ is non-decreasing in $G_{inc}$ and constant after $G_{inc} = 1$. □

### A.2.4 Behavior of the Unknown-Label Objectives

**Proposition 4 (faithful supported and unsupported targets).** Consider one unknown entry and optimize its evidential Beta parameters over $\alpha, \beta \geq 1$, with the graph quantities detached and $s_0 \geq 2$. Then:

1. For a supported entry with $w^T > 0$, the transfer term is minimized if and only if $(\alpha, \beta) = (\alpha^T, \beta^T)$.
2. For an unsupported entry, the penalty

$$\ell_{\text{unsup}}(\alpha, \beta) = [\log(1 + \text{ReLU}(\alpha + \beta - s_0))]^2 \tag{34}$$

is minimized exactly on the nonempty set of feasible $(\alpha, \beta)$ satisfying $\alpha + \beta \leq s_0$. It is invariant to exchanging $\alpha$ and $\beta$ and therefore imposes no positive or negative pseudo-label.

*Proof.* For the first statement, KL divergence is non-negative and equals zero if and only if its two distributions agree almost everywhere. The Beta family is identifiable, and the teacher parameters satisfy $\alpha^T, \beta^T \geq 1$, so equality holds at the unique feasible point where the two shape parameters match. Multiplication by $w^T > 0$ does not change the minimizer. For the second statement, $s_0 \geq 2$ makes the zero-penalty set feasible because $\alpha + \beta \geq 2$. Moreover, $\ell_{\text{unsup}} \geq 0$ and equals zero exactly when $\text{ReLU}(\alpha + \beta - s_0) = 0$, or $\alpha + \beta \leq s_0$. When $S = \alpha + \beta > s_0$,

$$\frac{\partial \ell_{\text{unsup}}}{\partial S} = \frac{2 \log(1 + S - s_0)}{1 + S - s_0} > 0, \tag{35}$$

so the penalty pushes excessive concentration downward. Because it depends only on $\alpha + \beta$, swapping the positive and negative evidence leaves it unchanged. □

Proposition 4 formalizes the asymmetric treatment of unknown labels. Supported entries receive a reliability-scaled distributional target, whereas unsupported entries are constrained only in evidence magnitude. These entrywise statements do not imply that a shared neural parameterization can attain every individual optimum simultaneously. They characterize the supervision encoded by the objective itself.