FedSEM: Mitigating Cross-Client Evidence Drift in Federated Multiple Instance Learning

Zhiqiang Kou[1*] **Beidi Wang**[2*] **Yuling Shi**[3] **Shengkun Zhu**[1]
Wei Tang[4] **Xiaodong Gu**[3] **Jiankai Zuo**[5] **Di Jiang**[1] **Yu Zhang**[6,✉]

[1] Hong Kong Polytechnic University [2] Tongji University
[3] Shanghai Jiao Tong University [4] Mohamed bin Zayed University of Artificial Intelligence
[5] Suzhou University of Science and Technology [6] Macquarie University
zhangyu.ansel@gmail.com

Abstract

Federated Multiple Instance Learning (MIFL) enables privacy-preserving weakly supervised learning across institutions, but existing methods mainly rely on parameter aggregation and overlook evidence-level inconsistency. Under bag-level supervision and heterogeneous client distributions, different clients may attend to inconsistent or misleading instances, causing cross-client critical evidence drift. To address this issue, we propose FedSEM, a Federated Shared Evidence Memory framework that aligns critical evidence beyond model aggregation. Each client compresses high-attention instances into compact prototypes with importance scores, while the server refines them into a stable shared evidence memory through redundancy removal, diversity selection, and score normalization. The memory is periodically broadcast to guide local attention learning with limited communication overhead and reduced privacy exposure. We further provide a non-convex convergence analysis and show that FedSEM only introduces a bounded memory-drift term. Experiments on four federated MIL benchmarks demonstrate that FedSEM achieves state-of-the-art performance under heterogeneous settings.

1 Introduction

Multiple Instance Learning (MIL) is an important weakly supervised learning paradigm in which each training sample is represented as a bag containing multiple instances, while labels are provided only at the bag level and instance-level annotations are unavailable [Dietterich et al., 1997, Ilse et al., 2018, Li et al., 2026a]. This formulation naturally matches many medical vision tasks, especially in computational pathology, where a whole-slide image is typically partitioned into a large number of patches and annotated only with a coarse slide-level label [Campanella et al., 2019, Lu et al., 2021]. As a result, MIL has been widely adopted for whole-slide image analysis, cancer diagnosis, lesion localization, and risk stratification [Campanella et al., 2019, Lu et al., 2021, Li et al., 2021a, 2026b]. The main advantage of MIL lies in its ability to circumvent the need for expensive and hard-to-obtain fine-grained instance annotations, yet its performance often depends on access to sufficiently large and diverse data sources [Zhang et al., 2025a, Campanella et al., 2019, Lu et al., 2021, Li and Wang, 2026]. In practice, data from a single medical institution is usually limited in both scale and distribution, and although multi-center collaboration can substantially improve model robustness and generalization, it is often constrained by data privacy concerns and data-sharing restrictions [Sheller et al., 2020, Rieke et al., 2020, Zhang et al., 2025b,c]. Federated Learning (FL) provides a practical solution to this dilemma by enabling multiple clients to perform collaborative model training without

*Equal Contribution.

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

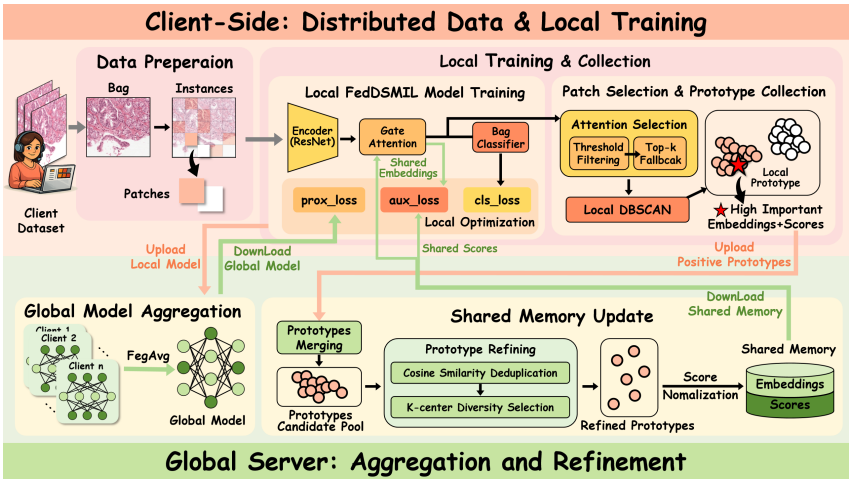

Figure 1: **Overall framework of FedSEM.** On the client side, each client trains a local FedDSMIL model and selects high-attention patch embeddings as representative evidence, which are further compressed into local prototypes. On the server side, FedSEM performs global model aggregation and updates a shared evidence memory by merging, deduplicating, diversifying, and normalizing the uploaded prototypes. The shared memory is then sent back to clients to guide local attention learning and improve cross-client evidence consistency.

exposing raw data [McMahan et al., 2017, Kairouz et al., 2021, Zhang et al., 2026a]. Therefore, Federated Multiple Instance Learning (MIFL) is gradually emerging as an important and promising research direction in medical weakly supervised learning [Lu et al., 2022].

Motivation. In MIFL, the final prediction is made at the patch (i.e., instance) level, while supervision is only available at the bag level; consequently, the model must first rely on attention mechanisms to identify candidate instances that are more likely to carry discriminative information from a large number of patches, and then leverage these high-value instances for subsequent patch-level learning [Ilse et al., 2018, Lu et al., 2021, Li et al., 2021a]. Therefore, although attention is not the final classification head, it determines which instances are regarded as valid evidence by the model; once the attention is biased, the downstream instance classifier will also be systematically misled. This issue is further amplified in the federated setting: due to the lack of explicit instance-level supervision and the significant data distribution heterogeneity across clients, each client can only infer important patches based on its local data, often resulting in inconsistent attention patterns that focus on different or even erroneous high-response regions, ultimately leading to cross-client instance-level evidence drift [Li et al., 2020, Karimireddy et al., 2020, Kairouz et al., 2021, Zhang et al., 2026b]. To mitigate this issue, parameter aggregation alone is insufficient, as the server can only average model weights but cannot explicitly align clients' judgments on "critical evidence." A potential solution is to introduce shared feature representations to correct the bias in evidence selection; however, directly uploading features of all high-attention patches would not only incur rapidly increasing communication costs with respect to the number of instances, but also introduce higher risks of privacy leakage [Zhu et al., 2019, Geiping et al., 2020].

To address the above issue, we propose FedSEM (Federated Shared Evidence Memory), which introduces a compact evidence-level communication pathway in addition to standard parameter aggregation. Each client selects high-attention instances as local critical evidence and compresses them into a few prototype centers with importance scores, rather than uploading raw patches or dense instance embeddings. The server refines these prototypes through redundancy removal, diversity selection, and score normalization, producing a stable shared evidence memory that is periodically broadcast to clients after a warm-up stage. This memory provides cross-client evidence anchors to guide local attention learning and mitigate evidence drift with limited communication overhead and reduced privacy exposure. Moreover, our convergence analysis shows that FedSEM maintains standard federated optimization behavior with only a bounded memory-drift term. Experiments on four federated MIL benchmarks verify its effectiveness across datasets and heterogeneity levels. The contributions of this paper are summarized as follows.

- *Method.* We reveal the problem of *cross-client critical evidence drift* in federated MIL and propose *FedSEM*, a shared evidence memory framework that explicitly aligns clients' critical instance-level evidence beyond parameter aggregation.
- *Theory.* In Theorem 1, we theoretically show that FedSEM maintains the convergence behavior of federated MIL, with only a bounded memory-drift term introduced by shared evidence memory updates.
- *Experiments.* Extensive experiments on four federated MIL benchmarks demonstrate that FedSEM achieves state-of-the-art performance under heterogeneous federated settings.

2 Methodology

Notations. We consider a federated multiple instance learning setting with K clients, where each client k holds a private dataset $\mathcal{D}_k = \{(\mathcal{B}_{k,i}, y_{k,i})\}_{i=1}^{n_k}$. Here, $\mathcal{B}_{k,i} = \{x_{k,i,j}\}_{j=1}^{m_{k,i}}$ denotes a bag containing multiple instances, and $y_{k,i}$ is the corresponding bag-level label. Instance-level labels are unavailable. At communication round t, the server sends the global MIL model θ^t to a subset of clients $\mathcal{S}^t$, and each client updates its local model θ_k^t using its private bags. The encoder maps each instance into an embedding $h_{k,i,j} \in \mathbb{R}^d$, and an attention module assigns an importance score $a_{k,i,j}$ to each instance for bag-level prediction. In standard federated MIL, only model parameters are exchanged between clients and the server. In FedSEM, we additionally maintain a shared evidence memory $\mathcal{M}^t = \{(m_r^t, \tilde{s}_r^t)\}_{r=1}^{M}$, where m_r^t is a shared evidence prototype and $\tilde{s}_r^t$ is its normalized importance score, to capture cross-client consensus over critical instance-level evidence.

Roadmap. FedSEM augments standard federated MIL with a compact shared evidence memory beyond model aggregation. Fig. 1 overviews the pipeline: Sec. 2.1 defines attention-based MIL and its weak evidence signal; Sec. 2.2 describes converting high-attention instances into scored prototypes; Sec. 2.3 presents server-side merging, deduplication, diversification, and score normalization; and Sec. 2.4 explains memory-guided local attention regularization. Appendix A lists notation, Appendix D.3 and Algorithm 1 detail the training pipeline, and Algorithm 2 gives memory-refinement pseudocode.

2.1 Federated Multiple Instance Learning

We consider a federated MIL setting with K clients. Each client k owns a private MIL dataset $\mathcal{D}_k = \{(\mathcal{B}_{k,i}, y_{k,i})\}_{i=1}^{n_k}$, where $\mathcal{B}_{k,i} = \{x_{k,i,j}\}_{j=1}^{m_{k,i}}$ denotes a bag containing $m_{k,i}$ instances, and $y_{k,i}$ is the corresponding bag-level label. Instance-level annotations are unavailable. Given an instance encoder g_θ, each instance is mapped into a latent embedding $h_{k,i,j} = g_\theta(x_{k,i,j})$. An attention module A_θ estimates the importance of each instance:

$$a_{k,i,j} = \frac{\exp(A_\theta(h_{k,i,j}))}{\sum_{r=1}^{m_{k,i}} \exp(A_\theta(h_{k,i,r}))}. \tag{1}$$

The bag representation is obtained by attention pooling $z_{k,i} = \sum_{j=1}^{m_{k,i}} a_{k,i,j} h_{k,i,j}$, and the bag-level prediction is produced by a classifier C_θ: $\hat{y}_{k,i} = C_\theta(z_{k,i})$. The standard MIL classification loss on client k is: $\mathcal{L}_{\text{MIL}}^k = \frac{1}{n_k} \sum_{i=1}^{n_k} \ell(\hat{y}_{k,i}, y_{k,i})$. In standard MIL, attention scores are only used for bag-level prediction. In FedSEM, we further treat them as weak evidence indicators for discovering locally important instances. This is crucial because, under client heterogeneity, different clients may focus on inconsistent or even misleading instances, leading to cross-client critical evidence drift.

2.2 Local Evidence Prototype Construction

To extract local discriminative evidence, each client first selects high-attention instances from positive bags $\mathcal{E}_k = \{h_{k,i,j} \mid y_{k,i} = 1,\ a_{k,i,j} \geq \tau\}$, where τ is the attention threshold. If too few instances satisfy this condition, we use a top-r fallback strategy and select the r highest-attention instances from each positive bag. This avoids empty or unstable evidence extraction during the early training stage. Directly uploading all selected instance embeddings would introduce substantial communication overhead and potential privacy risks. Therefore, FedSEM compresses the selected evidence set $\mathcal{E}_k$ into a small number of local prototypes. Specifically, client k clusters $\mathcal{E}_k$ into $\{\mathcal{C}_{k,q}\}_{q=1}^{Q_k}$, where each

cluster corresponds to one local evidence pattern. The prototype of cluster $\mathcal{C}_{k,q}$ is computed as the attention-weighted center:

$$p_{k,q} = \frac{\sum_{h_{k,i,j} \in \mathcal{C}_{k,q}} a_{k,i,j} h_{k,i,j}}{\sum_{h_{k,i,j} \in \mathcal{C}_{k,q}} a_{k,i,j}}. \tag{2}$$

The corresponding importance score is defined as the average attention score within the cluster:

$$s_{k,q} = \frac{1}{|\mathcal{C}_{k,q}|} \sum_{h_{k,i,j} \in \mathcal{C}_{k,q}} a_{k,i,j}. \tag{3}$$

Thus, each client uploads only a compact prototype set: $\mathcal{P}_k = \{(p_{k,q}, s_{k,q})\}_{q=1}^{Q_k}$. This design is communication-efficient: clients do not transmit raw patches, full bags, or all high-attention instance embeddings. Instead, they only upload a small number of compressed evidence prototypes and their scalar importance scores.

2.3 Server-side Shared Evidence Memory Refinement

At communication round t, the server receives prototype sets from participating clients $\mathcal{S}^t$ and constructs a candidate evidence pool:

$$\mathcal{C}^t = \mathcal{M}^{t-1} \cup \bigcup_{k \in \mathcal{S}^t} \mathcal{P}_k^t, \tag{4}$$

where $\mathcal{M}^{t-1}$ denotes the shared evidence memory from the previous round. The server then refines the candidate pool through three operations. First, redundant prototypes are removed using cosine similarity. If two prototypes p_a and p_b are highly similar, they are merged by score-weighted averaging:

$$p_{\text{merge}} = \frac{s_a p_a + s_b p_b}{s_a + s_b}. \tag{5}$$

Second, the server performs diversity selection to preserve representative but diverse evidence modes. Instead of simply keeping the highest-score prototypes, the server selects M prototypes that best cover the candidate evidence space:

$$\mathcal{M}^t = \arg \min_{\mathcal{S} \subseteq \mathcal{C}^t, |\mathcal{S}| = M} \max_{p \in \mathcal{C}^t} \min_{m \in \mathcal{S}} d(p, m), \tag{6}$$

where $d(\cdot, \cdot)$ denotes the distance between two prototypes. This K-center selection prevents the memory from being dominated by a few frequent evidence patterns. Finally, the selected prototype scores are normalized: $\tilde{s}_r^t = \frac{\exp(s_r^t)}{\sum_{u=1}^{M} \exp(s_u^t)}$. The refined shared evidence memory is represented as: $\mathcal{M}^t = \{(m_r^t, \tilde{s}_r^t)\}_{r=1}^{M}$. The shared memory stores cross-client consensus over discriminative evidence patterns. It is not a raw feature bank; rather, it is a compact and refined prototype memory constructed from multiple clients.

2.4 Memory-guided Local Optimization

After the warm-up stage, the server periodically broadcasts the shared evidence memory $\mathcal{M}^t = \{(m_r^t, \tilde{s}_r^t)\}_{r=1}^{M}$ to clients. For each local instance embedding $h_{k,i,j}$, client k estimates its memory-induced evidence score by comparing it with the shared prototypes:

$$\begin{cases} \alpha_{k,i,j,r} = \dfrac{\exp(\text{sim}(h_{k,i,j}, m_r^t)/\gamma)}{\sum_{u=1}^{M} \exp(\text{sim}(h_{k,i,j}, m_u^t)/\gamma)}, \\ b_{k,i,j} = \sum_{r=1}^{M} \alpha_{k,i,j,r} \tilde{s}_r^t. \end{cases} \tag{7}$$

Here, $b_{k,i,j}$ reflects how likely an instance is to be critical evidence according to the shared memory. FedSEM uses this score to regularize the local attention distribution:

$$\mathcal{L}_{\text{mem}}^k = \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{j=1}^{m_{k,i}} (a_{k,i,j} - b_{k,i,j})^2. \tag{8}$$

The final local objective is:

$$\mathcal{L}_k = \underbrace{\frac{1}{n_k}\sum_{i=1}^{n_k}\ell(\hat{y}_{k,i}, y_{k,i})}_{\text{Bag-level MIL Loss}} + \lambda \underbrace{\frac{1}{n_k}\sum_{i=1}^{n_k}\sum_{j=1}^{m_{k,i}}(a_{k,i,j} - b_{k,i,j})^2}_{\text{Memory-guided Evidence Alignment}} + \mu \underbrace{\|\theta_k - \theta^t\|_2^2}_{\text{Proximal Regularization}}. \tag{9}$$

The proximal regularization term is inspired by FedProx [Li et al., 2020], which constrains local model drift under heterogeneous client distributions. After local optimization, the server aggregates client models by weighted averaging: $\theta^{t+1} = \sum_{k\in\mathcal{S}^t} \frac{n_k}{\sum_{u\in\mathcal{S}^t} n_u}\theta_k^{t+1}$. Thus, FedSEM updates global model parameters through FedAvg [McMahan et al., 2017] and updates the shared evidence memory through prototype refinement.

3 Convergence Analysis

In this section, we provide a non-convex convergence analysis for FedSEM. Although FedSEM introduces an additional shared evidence memory $\mathcal{M}^t$ beyond standard parameter aggregation, we show that, under mild assumptions, the global model parameters θ^t still converge to a first-order stationary point of the joint objective at the standard $\mathcal{O}(1/\sqrt{T})$ rate, and the additional cost induced by the memory update is controlled by a memory-drift term that vanishes once the memory stabilizes. For the analysis, we collect the per-client losses defined in Sec. 2 into the joint objective

$$F(\theta, \mathcal{M}) \;=\; \sum_{k=1}^{K} \frac{n_k}{n}\Big[\mathcal{L}_{\mathrm{MIL}}^k(\theta_k) + \lambda\,\mathcal{L}_{\mathrm{mem}}^k(\theta_k, \mathcal{M}) + \mu\|\theta_k - \theta\|_2^2\Big], \tag{10}$$

and write $F_k(\theta, \mathcal{M}) := \mathcal{L}_{\mathrm{MIL}}^k(\theta) + \lambda\mathcal{L}_{\mathrm{mem}}^k(\theta, \mathcal{M})$ for the per-client smooth part. At round t, each client $k \in \mathcal{S}^t$ performs E steps of local SGD, producing intermediate iterates $\{\theta_k^{t,e}\}_{e=0}^{E}$ with $\theta_k^{t,0} = \theta^t$ and $\theta_k^{t+1} := \theta_k^{t,E}$. The server then aggregates them by weighted averaging

$$\theta^{t+1} \;=\; \sum_{k\in\mathcal{S}^t} \alpha_k\,\theta_k^{t+1}, \qquad \alpha_k \;=\; \frac{n_k}{\sum_{u\in\mathcal{S}^t} n_u}, \tag{11}$$

matching the parameter-aggregation step at the end of Sec. 2.4, and we let $S = |\mathcal{S}^t|$.

Remark 1. *Assumption 4 [2],is the only FedSEM-specific assumption, and it is justified by the design choices in Sec. 2. During the warm-up stage the server only accumulates prototypes without broadcasting, so $\mathcal{M}^t$ is held fixed at clients during this period and $\Delta_{\mathcal{M}}^t = 0$. After the warm-up, the cosine deduplication and the K-center diversity selection in Eqs.* (5)–(6) *ensure that the symmetric difference between $\mathcal{M}^{t-1}$ and $\mathcal{M}^t$ is sparse and bounded, while the softmax score-normalization step in Sec. 2.3 keeps the importance scores $\tilde{s}_r^t$ uniformly bounded in $(0, 1)$ and hence renders $\mathcal{L}_{\mathrm{mem}}^k$ in Eq.* (8) *uniformly bounded as well. Together these mechanisms yield a finite per-round memory drift $\Delta_{\mathcal{M}}^t$ that decays as the memory stabilizes.*

3.1 Main Result

Theorem 1 (Convergence of FedSEM)**.** *Under Assumptions 1–4, choose the local learning rate $\eta \le \frac{1}{4LE}$. Then the iterates $\{\theta^t\}_{t=0}^{T}$ produced by FedSEM satisfy*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big\|\nabla_\theta F(\theta^t, \mathcal{M}^t)\big\|^2 \;\le\; \frac{4\big(F(\theta^0, \mathcal{M}^0) - F^\star\big)}{\eta ET} + \frac{2L\eta\sigma^2}{S} + 20L^2\eta^2E\,(\sigma^2{+}6EG^2) + \frac{4\,\Delta_{\mathcal{M}}}{\eta ET}, \tag{12}$$

where $F^\star := \inf_{\theta,\mathcal{M}} F(\theta, \mathcal{M})$. In particular, choosing $\eta = \frac{1}{L\sqrt{TE}}$ yields the rate

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla_\theta F(\theta^t, \mathcal{M}^t)\|^2 \;=\; \mathcal{O}\left(\frac{1}{\sqrt{TE}} + \frac{E}{T} + \frac{\Delta_{\mathcal{M}}}{\sqrt{T/E}}\right).$$

[2]Due to space limitations, all Assumptions are provided in Appendix C.

Hence, FedSEM converges to a first-order stationary point of F at the standard $\mathcal{O}(1/\sqrt{T})$ rate, provided that the cumulative memory drift $\Delta_{\mathcal{M}}$ is sublinear in T.

Proof. The proof proceeds in six steps. Throughout, $\mathbb{E}$ denotes expectation over the randomness of mini-batches and client sampling, conditioned on $(\theta^t, \mathcal{M}^t)$ unless otherwise specified. For brevity, write $\nabla F^t := \nabla_\theta F(\theta^t, \mathcal{M}^t)$ and $g_k^{t,e} := g_k(\theta_k^{t,e}; \xi_k^{t,e})$.

We first bound the client drift induced by local SGD. For any $k \in \mathcal{S}^t$ and $e \in \{0, \dots, E-1\}$, the local update satisfies $\theta_k^{t,e} - \theta^t = -\eta \sum_{s=0}^{e-1} g_k^{t,s}$. Combining $\|a_1 + \cdots + a_e\|^2 \le e \sum_{s=1}^{e} \|a_s\|^2$ with Assumptions 2 and 3 and the standard descent-lemma argument used in the analyses of FedAvg, one obtains

$$\mathbb{E}\big\|\theta_k^{t,e} - \theta^t\big\|^2 \;\le\; 5E\eta^2\big(\sigma^2 + 6EG^2\big) + 30E^2\eta^2\,\mathbb{E}\|\nabla F^t\|^2, \tag{13}$$

which is valid as long as $\eta \le \frac{1}{4LE}$. The proximal term $\mu\|\theta_k - \theta^t\|^2$ in Eq. (9) only tightens the constants in Eq. (13), since it acts as an additional contraction toward θ^t, and we therefore retain the cleaner form above without loss of generality.

We next derive a one-step descent inequality at fixed memory. By Assumption 1 applied to $F(\cdot, \mathcal{M}^t)$ and the global update $\theta^{t+1} - \theta^t = -\eta \sum_{k \in \mathcal{S}^t} \alpha_k \sum_{e=0}^{E-1} g_k^{t,e}$, we have

$$\mathbb{E}\big[F(\theta^{t+1}, \mathcal{M}^t)\big] \;\le\; F(\theta^t, \mathcal{M}^t) + \underbrace{\mathbb{E}\langle \nabla F^t, \theta^{t+1} - \theta^t\rangle}_{T_1} + \frac{L}{2}\,\underbrace{\mathbb{E}\|\theta^{t+1} - \theta^t\|^2}_{T_2}. \tag{14}$$

We then bound the linear term T_1. By the unbiasedness of $g_k^{t,e}$ and the identity $\nabla F^t = \sum_k \alpha_k \nabla_\theta F_k(\theta^t, \mathcal{M}^t)$,

$$T_1 = -\eta \sum_{e=0}^{E-1} \sum_{k \in \mathcal{S}^t} \alpha_k\, \mathbb{E}\langle \nabla F^t, \nabla_\theta F_k(\theta_k^{t,e}, \mathcal{M}^t)\rangle.$$

Decomposing $\nabla_\theta F_k(\theta_k^{t,e}, \mathcal{M}^t) = \nabla_\theta F_k(\theta^t, \mathcal{M}^t) + \big[\nabla_\theta F_k(\theta_k^{t,e}, \mathcal{M}^t) - \nabla_\theta F_k(\theta^t, \mathcal{M}^t)\big]$, the first piece contributes $-\eta E\|\nabla F^t\|^2$, while the second piece is controlled by Cauchy–Schwarz and Assumption 1, giving

$$T_1 \;\le\; -\eta E\,\mathbb{E}\|\nabla F^t\|^2 + \frac{\eta E}{2}\,\mathbb{E}\|\nabla F^t\|^2 + \frac{\eta L^2}{2} \sum_{e=0}^{E-1} \sum_{k} \alpha_k\, \mathbb{E}\|\theta_k^{t,e} - \theta^t\|^2.$$

The quadratic term T_2 is bounded analogously. Standard mini-batch averaging together with Assumption 2 gives

$$T_2 \;\le\; \eta^2 E^2\,\mathbb{E}\|\nabla F^t\|^2 + \frac{\eta^2 E \sigma^2}{S} + L^2\eta^2 \sum_{e=0}^{E-1} \sum_{k} \alpha_k\, \mathbb{E}\|\theta_k^{t,e} - \theta^t\|^2.$$

Plugging the bounds on T_1 and T_2 into Eq. (14) and absorbing the client-drift terms via Eq. (13), the choice $\eta \le \frac{1}{4LE}$ makes the coefficient of $\|\nabla F^t\|^2$ on the right-hand side equal to $-\eta E/4$, so that

$$\mathbb{E}[F(\theta^{t+1}, \mathcal{M}^t)] - F(\theta^t, \mathcal{M}^t) \le -\frac{\eta E}{4}\mathbb{E}\|\nabla F^t\|^2 + \frac{L\eta^2 E\sigma^2}{2S} + 5L^2\eta^3E^2(\sigma^2 + 6EG^2). \tag{15}$$

It remains to account for the memory update. The right-hand side of Eq. (15) is expressed in terms of $F(\theta^{t+1}, \mathcal{M}^t)$, whereas to telescope we need $F(\theta^{t+1}, \mathcal{M}^{t+1})$. Adding the memory-drift bound from Assumption 4 gives

$$\mathbb{E}[F(\theta^{t+1}, \mathcal{M}^{t+1})] - F(\theta^t, \mathcal{M}^t) \le -\frac{\eta E}{4}\mathbb{E}\|\nabla F^t\|^2 + \frac{L\eta^2 E\sigma^2}{2S} + 5L^2\eta^3E^2(\sigma^2 + 6EG^2) + \Delta_{\mathcal{M}}^t.$$

Summing over $t = 0, \dots, T-1$ and using $F(\theta^T, \mathcal{M}^T) \ge F^\star$ to telescope, we obtain

$$\frac{\eta E}{4} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla F^t\|^2 \le F(\theta^0, \mathcal{M}^0) - F^\star + \frac{L\eta^2 E\sigma^2 T}{2S} + 5L^2\eta^3E^2T(\sigma^2 + 6EG^2) + \Delta_{\mathcal{M}}.$$

Dividing both sides by $\eta ET/4$ yields Eq. (12), which completes the proof. □

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate FedSEM on four MIL benchmarks: *CAMELYON16* [Litjens et al., 2018], *PANDA* [Bulten et al., 2022], *Pascal VOC 2012* [Everingham et al., 2010], and *MNIST-Bags* [Zhang et al., 2022]. These datasets cover both medical and natural image MIL scenarios, where each sample is represented as a bag of instances and only bag-level labels are available.

Baselines. We compare FedSEM with representative federated and MIL baselines, including FedAvg [McMahan et al., 2017], FedProx [Li et al., 2020], FedRDN [Yan et al., 2025], MOON [Li et al., 2021b], MIL-Dropout [Zhu et al., 2025], and DSMIL [Li et al., 2021a]. All methods are evaluated using bag-level classification accuracy.

Table 1: Client-side bag-level ACC(%) on four federated MIL benchmarks under different heterogeneity levels. The best results are highlighted in **bold**, and the second-best results are underlined.

Method	CAMELYON16			PANDA			Pascal VOC 2012			MNIST-Bags		
β	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
FedSEM (Ours)	**66.67**	**77.78**	68.52	**75.20**	**77.46**	**77.51**	65.29	**71.14**	**70.84**	**90.00**	**94.00**	**90.00**
FedAvg	62.96	68.52	62.96	74.68	76.94	76.47	**65.81**	65.25	68.11	88.00	90.00	88.00
FedProx	62.96	68.52	68.52	75.06	76.94	77.37	64.99	64.60	70.28	86.00	90.00	**90.00**
FedRDN	61.11	61.11	61.11	72.75	72.75	72.75	64.60	64.90	64.64	50.00	52.00	52.00
MOON	59.26	66.67	64.81	72.75	76.47	72.75	64.64	66.77	64.60	88.00	90.00	82.00
MIL-Dropout	61.11	66.67	59.26	73.83	77.23	75.67	64.64	66.81	65.08	74.00	90.00	74.00
DSMIL	**66.67**	59.26	**70.37**	72.80	72.94	72.75	64.60	64.60	65.77	74.00	70.00	74.00

4.2 Main Results

Table 1 reports the client-side bag-level accuracy of FedSEM and competing methods on four federated MIL benchmarks under different heterogeneity levels. From the table, we can draw the following conclusions:

- **Finding 1.** FedSEM obtains the best or tied-best results in 10 out of 12 settings and ranks second in the remaining two settings, showing stable superiority across different datasets and heterogeneity levels.
- **Finding 2.** Compared with FedAvg and FedProx, FedSEM achieves clear improvements on CAMELYON16 and PANDA. This indicates that standard parameter aggregation alone is insufficient for federated MIL, while the proposed evidence-level memory helps align critical instance patterns across clients.
- **Finding 3.** Besides pathology datasets, FedSEM also performs strongly on Pascal VOC 2012 and MNIST-Bags. This suggests that the proposed prototype-based shared memory is not tailored to a single medical dataset, but can benefit general MIL tasks with diverse instance patterns.

4.3 Ablation Study

Table 2 reports the ablation results of FedSEM. ***Panel A*** evaluates the effect of different local optimization components, where FedSEM-a denotes the original FedAvg-style aggregation strategy, FedSEM-b adds the proximal regularization term, and FedSEM is our full method. Compared with FedSEM-a, FedSEM-b brings moderate improvements in several settings, indicating that constraining local model drift is useful under heterogeneous client distributions. However, proximal regularization alone only operates at the parameter level and cannot explicitly correct inconsistent evidence selection

Table 2: Ablation study on four federated MIL benchmarks. Panel A evaluates the local optimization objective. Panel B reports the communication–performance profile of server-side prototype filtering, where each Raw/Refined entry is communication cost / ACC(%). Positive ΔACC indicates lower communication cost with better accuracy, while negative ΔACC indicates a communication–accuracy trade-off.

Panel A: Local optimization objective

CAMELYON16

Method	$\beta = 0.25$	$\beta = 0.5$	$\beta = 0.75$
FedSEM-a	62.96	68.52	62.96
FedSEM-b	62.96 —	68.52 —	68.52↑**5.56**
FedSEM(Ours)	**66.67**↑**3.71**	**77.78**↑**9.26**	68.52 —

PANDA

Method	$\beta = 0.25$	$\beta = 0.5$	$\beta = 0.75$
FedSEM-a	74.68	76.94	76.47
FedSEM-b	75.06↑**0.38**	76.94 —	77.37↑**0.90**
FedSEM(Ours)	75.20↑**0.14**	**77.46**↑**0.52**	**77.51**↑**0.14**

Pascal VOC 2012

Method	$\beta = 0.25$	$\beta = 0.5$	$\beta = 0.75$
FedSEM-a	65.81	65.25	68.11
FedSEM-b	64.99	64.60	70.28↑**2.17**
FedSEM(Ours)	65.29↑**0.30**	**71.14**↑**6.54**	**70.84**↑**0.56**

MNIST-Bags

Method	$\beta = 0.25$	$\beta = 0.5$	$\beta = 0.75$
FedSEM-a	88.00	90.00	88.00
FedSEM-b	86.00	90.00 —	**90.00**↑**2.00**
FedSEM(Ours)	**90.00**↑**4.00**	**94.00**↑**4.00**	**90.00** —

Panel B: Server-side prototype filtering

CAMELYON16

β	Raw	Refined	ΔCost	ΔACC	Profile
0.25	13.24/61.11	4.88/**66.67**	↓**8.37**	↑**5.56**	WIN
0.5	18.40/70.37	6.46/**77.78**	↓**11.94**	↑**7.41**	WIN
0.75	18.40/**70.37**	6.46/68.52	↓**11.94**	↓**1.85**	TRADE

PANDA

β	Raw	Refined	ΔCost	ΔACC	Profile
0.25	17.18/**77.04**	5.01/75.20	↓**12.17**	↓**1.84**	TRADE
0.5	17.18/77.27	4.83/**77.46**	↓**12.35**	↑**0.19**	WIN
0.75	17.18/77.46	5.01/**77.51**	↓**12.17**	↑**0.05**	WIN

Pascal VOC 2012

β	Raw	Refined	ΔCost	ΔACC	Profile
0.25	14.85/**67.11**	8.63/65.29	↓**6.22**	↓**1.82**	TRADE
0.5	17.12/65.29	5.34/**71.14**	↓**11.78**	↑**5.85**	WIN
0.75	17.14/69.84	8.30/**70.84**	↓**8.85**	↑**1.00**	WIN

MNIST-Bags

β	Raw	Refined	ΔCost	ΔACC	Profile
0.25	23.33/88.00	9.77/**90.00**	↓**13.56**	↑**2.00**	WIN
0.5	24.54/80.00	13.46/**94.00**	↓**11.08**	↑**14.00**	WIN
0.75	24.54/88.00	13.46/**90.00**	↓**11.08**	↑**2.00**	WIN

across clients. By further introducing memory-guided evidence alignment, the full FedSEM achieves the best overall performance, improving the average accuracy from 73.98% and 74.60% to 77.03%. This verifies that the shared evidence memory provides effective cross-client evidence guidance for local MIL training.

Panel B evaluates the effect of server-side prototype refinement. Compared with directly using raw uploaded prototypes, the refined memory consistently reduces communication cost by removing redundant prototypes, preserving diverse evidence patterns, and normalizing importance scores. Meanwhile, the refined memory maintains or improves accuracy in most settings, leading to a better accuracy–communication trade-off. For instance, on MNIST-Bags under $\beta = 0.5$, refinement reduces the cost from 24.54 to 13.46 and improves accuracy from 80.00% to 94.00%. These results show that FedSEM benefits not only from evidence sharing, but also from carefully refining the shared memory before broadcasting it back to clients.

4.4 Parameter Sensitivity Analysis.

We analyze the sensitivity of FedSEM to six key parameters in Fig. 2, including λ, μ, β, the number of clients, τ, and Top-K. Overall, FedSEM is relatively stable under different parameter choices, while moderate values usually lead to better performance.

For the optimization-related parameters, λ controls the strength of memory-guided evidence alignment, and μ controls the proximal regularization strength. The results show that a moderate λ and a

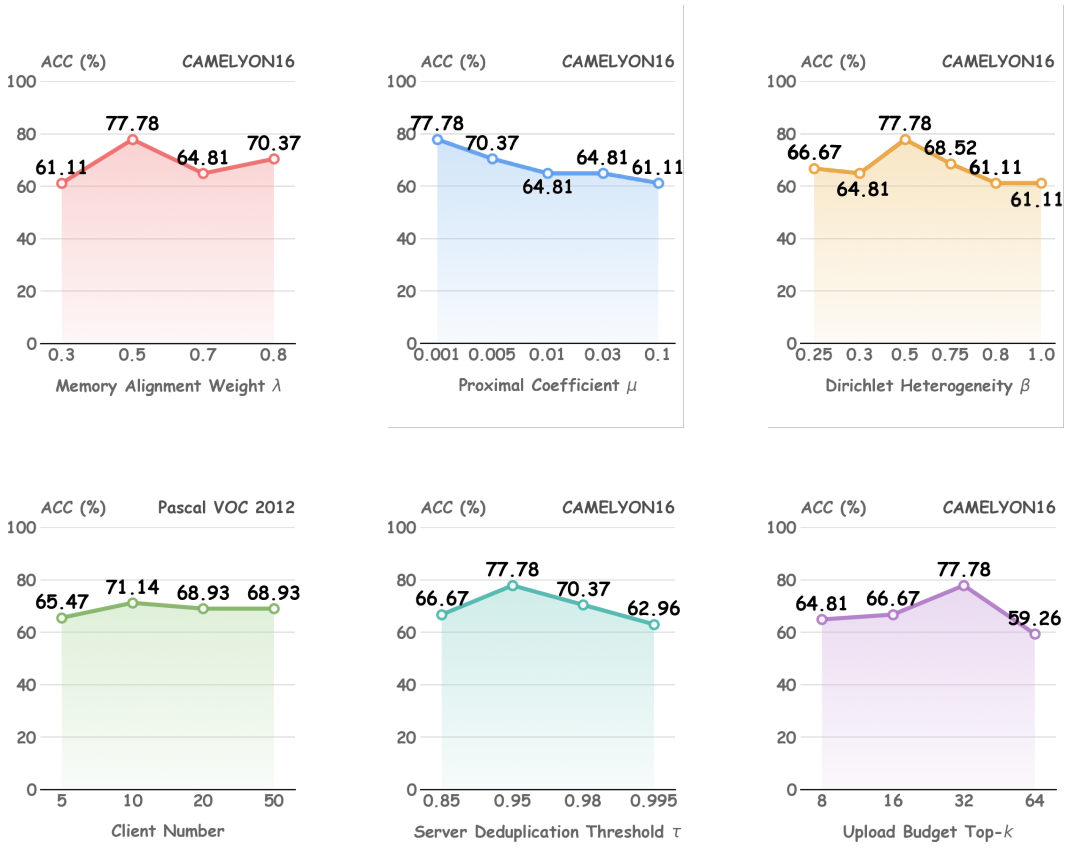

Figure 2: Parameter sensitivity analysis of FedSEM with respect to the memory alignment weight λ, the proximal coefficient μ, the Dirichlet heterogeneity parameter β, the number of clients, the server deduplication threshold τ, and the upload budget Top-K.

small μ are preferable, since overly weak alignment cannot fully exploit the shared memory, while overly strong regularization may restrict local optimization.

For the federated data setting, β reflects the degree of client heterogeneity, and the client number controls the scale of federation. FedSEM maintains stable performance under different heterogeneity levels, but too many clients may reduce the amount of local data per client and make evidence discovery more difficult.

For memory construction, τ controls server-side prototype deduplication, and Top-K determines the upload budget of high-attention instances. A proper τ preserves diverse but non-redundant evidence, while a moderate Top-K provides sufficient evidence without introducing excessive noisy prototypes.

5 Conclusion

In this paper, we proposed FedSEM, a federated shared evidence memory framework for federated multiple instance learning. FedSEM addresses the cross-client critical evidence drift problem by allowing clients to upload compact evidence prototypes instead of raw instances, and by constructing a shared evidence memory on the server to guide local attention learning. Together with memory-guided evidence alignment and proximal regularization, FedSEM improves evidence consistency across heterogeneous clients while maintaining communication efficiency. Extensive experiments on multiple benchmark datasets demonstrate the effectiveness of FedSEM.

References

Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1–2):31–71, 1997. doi: 10.1016/S0004-3702(96)00034-3. URL https://doi.org/10.1016/S0004-3702(96)00034-3.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR, 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

Qi Li, Bo Yin, Weiqi Huang, Ruhao Liu, Bojun Zou, Runpeng Yu, Jingwen Ye, Weihao Yu, and Xinchao Wang. Vision-language-action safety: Threats, challenges, evaluations, and mechanisms. *arXiv preprint arXiv:2604.23775*, 2026a. doi: 10.48550/arXiv.2604.23775. URL https://arxiv.org/abs/2604.23775.

Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019. doi: 10.1038/s41591-019-0508-1. URL https://doi.org/10.1038/s41591-019-0508-1.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021. doi: 10.1038/s41551-020-00682-w. URL https://doi.org/10.1038/s41551-020-00682-w.

Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2021a. doi: 10.1109/CVPR46437.2021.01409. URL https://openaccess.thecvf.com/content/CVPR2021/html/Li_Dual-Stream_Multiple_Instance_Learning_Network_for_Whole_Slide_Image_Classification_CVPR_2021_paper.html.

Qi Li, Xingyi Yang, and Xinchao Wang. Badwam: When world-action models dream right but act wrong. *arXiv preprint arXiv:2607.15207*, 2026b. doi: 10.48550/arXiv.2607.15207. URL https://arxiv.org/abs/2607.15207.

Xueyi Zhang, Peiyin Zhu, Yuan Liao, Xiyu Wang, Mingrui Lao, Siqi Cai, Yanming Guo, and Haizhou Li. Trustclip: Learning from noisy labels via semantic label verification and trust-aligned gradient projection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 4388–4397, 2025a. doi: 10.1145/3746027.3755415. URL https://doi.org/10.1145/3746027.3755415.

Qi Li and Xinchao Wang. Jeva: System one models from first principles. *arXiv preprint arXiv:2610.18773*, 2026. doi: 10.48550/arXiv.2610.18773. URL https://arxiv.org/abs/2610.18773.

Micah J. Sheller, Brandon Edwards, G. Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R. Colen, and Spyridon Bakas. Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10(1):12598, 2020. doi: 10.1038/s41598-020-69250-1. URL https://doi.org/10.1038/s41598-020-69250-1.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus Maier-Hein, Sébastien Ourselin, Micah Sheller, Ronald M. Summers, Andrew Trask, Daguang Xu, Maximilian Baust, and M. Jorge Cardoso. The future of digital health with federated learning. *npj Digital Medicine*, 3(1):119, 2020. doi: 10.1038/s41746-020-00323-1. URL https://doi.org/10.1038/s41746-020-00323-1.

Xueyi Zhang, Peiyin Zhu, Chengwei Zhang, Zhiyuan Yan, Jikang Cheng, Mingrui Lao, Siqi Cai, and Yanming Guo. Generalization-preserved learning: Closing the backdoor to catastrophic forgetting in continual deepfake detection. In *2025 IEEE/CVF International Conference on Computer*

Vision (ICCV), pages 3798–3808. IEEE, 2025b. doi: 10.1109/ICCV51701.2025.00362. URL https://doi.org/10.1109/ICCV51701.2025.00362.

Xueyi Zhang, Peiyin Zhu, Jinping Sui, Xiaoda Yang, Jiahe Tian, Mingrui Lao, Siqi Cai, Yanming Guo, and Jun Tang. Choose your expert: Uncertainty-guided expert selection for continual deepfake detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 11502–11511, 2025c. doi: 10.1145/3746027.3755398. URL https://doi.org/10.1145/3746027.3755398.

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*, pages 1273–1282. PMLR, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, Hossein Esfandiari, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2): 1–210, 2021. doi: 10.1561/2200000083. URL https://doi.org/10.1561/2200000083.

Xueyi Zhang, Ruicong Wang, Jialu Sun, Siqi Cai, and Haizhou Li. Listening to the brain: Multi-band seeg auditory reconstruction via dynamic spatio-temporal hypergraphs. *Advances in Neural Information Processing Systems*, 38, 2026a. doi: 10.52202/085713-4302. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/bbc6396ae4b4dbf78ce328aa8beafd27-Abstract-Datasets_and_Benchmarks_Track.html.

Ming Y. Lu, Richard J. Chen, Dehan Kong, Jana Lipkova, Rajendra Singh, Drew F. K. Williamson, Tiffany Y. Chen, and Faisal Mahmood. Federated learning for computational pathology on gigapixel whole slide images. *Medical Image Analysis*, 76:102298, 2022. doi: 10.1016/j.media.2021.102298. URL https://doi.org/10.1016/j.media.2021.102298.

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, pages 429–450, 2020. URL https://proceedings.mlsys.org/paper_files/paper/2020/hash/1f5fe83998a09396ebe6477d9475ba0c-Abstract.html.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5132–5143. PMLR, 2020. URL https://proceedings.mlr.press/v119/karimireddy20a.html.

Xueyi Zhang, Chengwei Zhang, Zheng Li, Xiyu Wang, Siqi Cai, Mingrui Lao, Yanming Guo, and Huiping Zhuang. Rep deep & machine learning: Exemplar-free continual video action recognition via slow-fast collaborative learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 36075–36083, 2026b. doi: 10.1609/aaai.v40i42.40924. URL https://doi.org/10.1609/aaai.v40i42.40924.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper/9617-deep-leakage-from-gradients.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients – how easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems*, volume 33, pages 16937–16947, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c4ede56bbd98819ae6112b20ac6bf145-Abstract.html.

Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob Van de Loo, Rob Vogels, et al. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: The CAMELYON dataset. *GigaScience*, 7(6):giy065, 2018. doi: 10.1093/gigascience/giy065. URL https://doi.org/10.1093/gigascience/giy065.

Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F. Steiner, Hester Van Boven, Robert Vink, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: The PANDA challenge. *Nature Medicine*, 28(1):154–163, 2022. doi: 10.1038/s41591-021-01620-2. URL https://doi.org/10.1038/s41591-021-01620-2.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. doi: 10.1007/s11263-009-0275-4. URL https://doi.org/10.1007/s11263-009-0275-4.

Weijia Zhang, Xuanhui Zhang, Hanwen Deng, and Min-Ling Zhang. Multi-instance causal representation learning for instance label prediction and out-of-distribution generalization. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-2532. URL https://papers.nips.cc/paper_files/paper/2022/hash/e261e92e1cfb820da930ad8c38d0aead-Abstract-Conference.html.

Yunlu Yan, Huazhu Fu, Yuexiang Li, Jinheng Xie, Jun Ma, Guang Yang, and Lei Zhu. A simple data augmentation for feature distribution skewed federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25749–25758, 2025. doi: 10.1109/CVPR52734.2025.02398. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yan_A_Simple_Data_Augmentation_for_Feature_Distribution_Skewed_Federated_Learning_CVPR_2025_paper.html.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10713–10722, 2021b. doi: 10.1109/CVPR46437.2021.01057. URL https://openaccess.thecvf.com/content/CVPR2021/html/Li_Model-Contrastive_Federated_Learning_CVPR_2021_paper.html.

Wenhui Zhu, Peijie Qiu, Xiwen Chen, Zhangsihao Yang, Aristeidis Sotiras, Abolfazl Razi, and Yalin Wang. How effective can dropout be in multiple instance learning? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 80090–80106. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhu25q.html.

Bao Li, Zhenyu Liu, Lizhi Shao, Bensheng Qiu, Hong Bu, and Jie Tian. Point transformer with federated learning for predicting breast cancer HER2 status from hematoxylin and eosin-stained whole slide images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(4):3000–3008, 2024. doi: 10.1609/aaai.v38i4.28082. URL https://doi.org/10.1609/aaai.v38i4.28082.

Farchan Hakim Raswa, Chun-Shien Lu, and Jia-Ching Wang. HistoFS: Non-IID histopathologic whole slide image classification via federated style transfer with RoI-preserving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 30251–30260, 2025. doi: 10.1109/CVPR52734.2025.02816. URL https://openaccess.thecvf.com/content/CVPR2025/html/Raswa_HistoFS_Non-IID_Histopathologic_Whole_Slide_Image_Classification_via_Federated_Style_CVPR_2025_paper.html.

Zhiqiang Kou, Junxiang Wu, Wenke Huang, Wenwen He, Ming-Kun Xie, Changwei Wang, Yuheng Jia, Di Jiang, Yang Liu, Xin Geng, and Qiang Yang. FedHarmony: Harmonizing heterogeneous

label correlations in federated multi-label learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10324–10334, 2026a. doi: 10.48550/arXiv.2604.28024. URL https://arxiv.org/abs/2604.28024.

Junxiang Wu, Zhiqiang Kou, Hongwei Zeng, Wenke Huang, Biao Liu, Hanlin Gu, Yuheng Jia, Di Jiang, Yang Liu, and Xin Geng. Trustworthy federated label distribution learning under annotation quality disparity, 2026. URL https://arxiv.org/abs/2605.04827.

Jiashuai Liu, Anyu Mao, Yi Niu, Xianli Zhang, Tieliang Gong, Chen Li, and Zeyu Gao. PAMIL: Prototype attention-based multiple instance learning for whole slide image classification. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15004 of *Lecture Notes in Computer Science*, pages 362–372. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72083-3_34. URL https://doi.org/10.1007/978-3-031-72083-3_34.

Susu Sun, Dominique van Midden, Geert Litjens, and Christian F. Baumgartner. Prototype-based multiple instance learning for gigapixel whole slide image classification. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15973 of *Lecture Notes in Computer Science*, pages 507–517. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-05185-1_49. URL https://doi.org/10.1007/978-3-032-05185-1_49.

Zhiqiang Kou, Jing Wang, Jiawei Tang, Yuheng Jia, Boyu Shi, and Xin Geng. Exploiting multi-label correlation in label distribution learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 4326–4334, 2024. doi: 10.24963/ijcai.2024/478. URL https://www.ijcai.org/proceedings/2024/478.

Zhiqiang Kou, Yucheng Xie, Hailin Wang, Junyang Chen, Jing Wang, Ming-Kun Xie, Shuo Chen, Yuheng Jia, Tongliang Liu, and Xin Geng. RankMatch: A novel approach to semi-supervised label distribution learning leveraging rank correlation between labels. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. doi: 10.52202/085713-5282. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/e7cc7f2467171d81b7224eae1373c493-Abstract-Conference.html.

Zhiqiang Kou, Jing Wang, Yuheng Jia, Biao Liu, and Xin Geng. Instance-dependent inaccurate label distribution learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1): 1425–1437, 2025b. doi: 10.1109/TNNLS.2023.3329870. URL https://doi.org/10.1109/TNNLS.2023.3329870.

Zhiqiang Kou, Si Qin, Hailin Wang, Jing Wang, Mingkun Xie, Shuo Chen, Yuheng Jia, Tongliang Liu, Masashi Sugiyama, and Xin Geng. Label distribution learning with biased annotations assisted by multi-label learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 5545–5553, 2025c. doi: 10.24963/ijcai.2025/617. URL https://www.ijcai.org/proceedings/2025/617.

Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. FedProto: Federated prototype learning across heterogeneous clients. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):8432–8440, 2022. doi: 10.1609/aaai.v36i8.20819. URL https://doi.org/10.1609/aaai.v36i8.20819.

Haolong Jin, Shenglin Liu, Cong Cong, Qingmin Feng, Yongzhi Liu, Lina Huang, and Yingzi Hu. FedWSIDD: Federated whole slide image classification via dataset distillation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15973 of *Lecture Notes in Computer Science*, pages 178–188. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-05185-1_18. URL https://doi.org/10.1007/978-3-032-05185-1_18.

Zhiqiang Kou, Junyang Chen, Xin-Qiang Cai, Xiaobo Xia, Ming-Kun Xie, Dong-Dong Wu, Biao Liu, Yuheng Jia, Xin Geng, Masashi Sugiyama, and Tat-Seng Chua. Positive-unlabeled reinforcement learning distillation for on-premise small models, 2026b. URL https://arxiv.org/abs/2601.20687.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, and the CAMELYON16 Consortium. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017. doi: 10.1001/jama.2017.14585. URL https://doi.org/10.1001/jama.2017.14585.

Jasper R. R. Uijlings, Koen E. A. van de Sande, Theo Gevers, and Arnold W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013. doi: 10.1007/s11263-013-0620-5. URL https://doi.org/10.1007/s11263-013-0620-5.

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. URL https://doi.org/10.1109/5.726791.

Appendix Table of Contents

A Notation **16**

B Related Work **16**

B.1 Federated Multiple Instance Learning for Whole-Slide Images 16

B.2 Evidence Modeling and Prototype Communication 17

C Details of the Assumption **18**

D Additional Method Details **18**

D.1 Federated Setting 18

D.2 Evaluation Metric 18

D.3 FedSEM Training Pipeline 18

D.4 Local Prototype Construction 19

D.5 Server-side Memory Refinement 19

E Dataset Details **20**

E.1 CAMELYON16 20

E.2 PANDA 20

E.3 Pascal VOC 2012 21

E.4 MNIST-Bags 21

F Experimental Details **21**

F.1 Federated Configuration 21

F.2 Optimization and Architecture 22

G Parameter Analysis **22**

G.1 Memory Alignment Weight λ 22

G.2 Proximal Coefficient μ 22

G.3 Number of Clients 23

G.4 Dirichlet Heterogeneity Parameter β 23

G.5 Upload Budget Top-k 24

G.6 Server Deduplication Threshold τ 25

H Feature-attention Heatmaps **25**

I Additional Discussion **26**

I.1 Limitations 26

I.2 Future Directions 26

A Notation

Table 3 summarizes the main notation used throughout the paper. Superscript t denotes the communication round, while subscripts k, i, j, q, and r index clients, bags, instances, local clusters, and shared-memory entries, respectively.

Table 3: Summary of the main notations.

Notation	Description
K	Total number of federated clients.
k	Client index.
t, T	Communication-round index and total number of communication rounds.
E, T_w	Number of local optimization epochs/steps per round and memory warm-up rounds.
$\mathcal{S}^t, S$	Participating client set at round t and its size $\lvert\mathcal{S}^t\rvert$.
$\mathcal{D}_k$	Local MIL dataset of client k.
n_k, n	Number of bags on client k and total number of bags.
$\mathcal{B}_{k,i}$	The i-th bag on client k.
$y_{k,i}$	Bag-level label of $\mathcal{B}_{k,i}$.
$x_{k,i,j}$	The j-th instance in bag $\mathcal{B}_{k,i}$.
$m_{k,i}$	Number of instances in bag $\mathcal{B}_{k,i}$.
θ^t	Global model parameters at round t.
$\theta_k^t, \theta_k^{t,e}$	Local model of client k and its intermediate local iterate.
$g_\theta, A_\theta, C_\theta$	Instance encoder, attention module, and bag-level classifier.
$h_{k,i,j} \in \mathbb{R}^d, d$	Instance embedding and embedding dimension.
$a_{k,i,j}$	Attention score, used as a local evidence-importance signal.
$z_{k,i}, \hat{y}_{k,i}$	Attention-pooled bag representation and bag-level prediction.
$\ell(\cdot,\cdot)$	Bag-level classification loss.
$\mathcal{L}_{\mathrm{MIL}}^k$	Standard MIL classification loss on client k.
$\mathcal{E}_k$	Selected high-attention evidence set on client k.
τ_{att}	Attention threshold for local evidence selection; written as τ in Sec. 2.2.
k_{up} or Top-k	Upload/fallback budget for high-attention instances.
$\mathcal{C}_{k,q}, Q_k$	The q-th local evidence cluster and number of clusters on client k.
$p_{k,q}, s_{k,q}$	Local evidence prototype and its importance score.
$\mathcal{P}_k^t$	Prototype set uploaded by client k at round t.
$\mathcal{M}^t, M$	Shared evidence memory at round t and memory size.
$m_r^t, \tilde{s}_r^t$	The r-th shared evidence prototype and normalized importance score.
$\mathcal{C}^t$	Server-side candidate evidence pool before memory refinement.
p_{merge}	Score-weighted merged prototype after redundancy removal.
τ_{dedup}	Cosine-similarity threshold for server-side prototype deduplication.
$d(\cdot,\cdot), \mathrm{sim}(\cdot,\cdot)$	Prototype distance and embedding-prototype similarity.
γ	Temperature for memory-based soft assignment.
$\alpha_{k,i,j,r}$	Soft assignment of instance $h_{k,i,j}$ to memory prototype m_r^t.
$b_{k,i,j}$	Memory-induced evidence score for instance $h_{k,i,j}$.
$\mathcal{L}_{\mathrm{mem}}^k$	Memory-guided evidence alignment loss.
λ, μ	Memory-alignment weight and proximal regularization coefficient.
$\mathcal{L}_k$	Final local FedSEM objective minimized by client k.
α_k	FedAvg aggregation weight of client k.
$F(\theta, \mathcal{M}), F_k(\theta, \mathcal{M})$	Global objective and per-client smooth objective in the convergence analysis.
η, L	Local learning rate and smoothness constant in the convergence analysis.
$g_k(\theta;\xi_k), \xi_k$	Stochastic gradient and mini-batch on client k.
σ^2, G^2	Stochastic-gradient variance bound and client-heterogeneity bound.
$\Delta_{\mathcal{M}}^t, \Delta_{\mathcal{M}}$	Per-round and cumulative memory-drift bounds.
$F^\star$	Infimum of the global objective.
β	Dirichlet concentration parameter controlling label-distribution heterogeneity.
ACC	Bag-level classification accuracy.

B Related Work

Federated multiple instance learning for whole-slide images. Whole-slide image (WSI) analysis provides a natural setting for federated multiple instance learning: each slide contains many patches,

supervision is usually available at the slide level, and data are distributed across institutions. Lu et al. [2022] demonstrated federated WSI classification and survival prediction with attention-based MIL and differential privacy. Subsequent methods address different sources of variation across participating institutions. PointTransformerFL [Li et al., 2024] combines point-transformer representations with strategies for label imbalance and feature sampling in federated HER2 prediction. HistoFS [Raswa et al., 2025] addresses non-IID histopathological appearance through federated style transfer that preserves regions of interest. These studies establish the feasibility of collaborative WSI learning and show the importance of accounting for variation in both slide representations and local data distributions.

The heterogeneity encountered in federated learning extends beyond visual appearance. FedAvg [McMahan et al., 2017] provides the basic model-aggregation framework, while FedProx [Li et al., 2020] regularizes local optimization and FedRDN [Yan et al., 2025] addresses feature-distribution skew through data augmentation. At the level of output structure, FedHarmony [Kou et al., 2026a] studies heterogeneous label correlations in federated multi-label learning. Annotation quality introduces another source of client variation: FedQual [Wu et al., 2026] combines quality-adaptive local training with reliability-aware aggregation for federated label distribution learning. These approaches concern optimization, feature distributions, label dependencies, or supervision quality. FedSEM considers the additional problem that clients trained with bag labels can assign importance to different instance patterns, even when their model parameters are aggregated.

Evidence modeling and prototype communication. Attention-based MIL and related WSI models infer bag predictions from learned instance representations and importance estimates [Ilse et al., 2018, Lu et al., 2021, Li et al., 2021a]. Prototype-based approaches make the representation of this evidence more explicit. PAMIL [Liu et al., 2024] combines prototype and instance attention to support instance-level interpretation, whereas ProtoMIL [Sun et al., 2025] represents predictions through interpretable concepts and supports intervention on those concepts. Their emphasis is on evidence representation and interpretation within a centralized model; federated learning additionally requires deciding which evidence should be communicated and how it should guide other clients.

Structural information can also be exploited in the supervision itself. In MIL, DSMIL [Li et al., 2021a] connects instance predictions with bag-level aggregation, while PointTransformerFL [Li et al., 2024] models dependencies among slide features. In the output space, Kou et al. [2024] introduce an auxiliary multi-label learning component to capture label correlations for label distribution learning, avoiding a direct low-rank assumption on the label distribution matrix. RankMatch [Kou et al., 2025a] uses label-ranking relationships to construct and exploit pseudo-labels in semi-supervised label distribution learning. These methods illustrate how the structure of an auxiliary learning signal matters beyond its numerical fit. Their label distributions and ranking targets differ from FedSEM's bag-only supervision, where the shared signal consists of instance prototypes and attention-derived importance scores.

The reliability of such signals is a separate concern. MIL-Dropout [Zhu et al., 2025] investigates regularization under bag-level supervision, and causal MIL [Zhang et al., 2022] studies instance-label prediction and out-of-distribution generalization. Related issues arise when the observed supervision is itself inaccurate. Instance-dependent inaccurate label distribution learning [Kou et al., 2025b] explicitly considers annotation errors that depend on the input instance. Biased-annotation LDL [Kou et al., 2025c] uses an auxiliary multi-label representation to recover label information from biased distributions. Although these annotation models differ from attention estimated without instance labels, they motivate distinguishing the amount of shared information from its reliability. For FedSEM, high attention and prototype agreement are useful selection signals, but neither establishes that an instance provides causal evidence for the bag label.

Communication methods differ in the objects they transfer. FedProto [Tan et al., 2022] exchanges class prototypes to support learning across heterogeneous clients, while FedWSIDD [Jin et al., 2025] distills WSI data into synthetic slides for federated training with heterogeneous local models. Beyond visual learning, positive-unlabeled reinforcement learning distillation [Kou et al., 2026b] uses teacher-generated response anchors to construct preference supervision for an on-premise language-model student. This is a different task and supervision regime, but it illustrates another way of transferring selected learning signals for local training. FedSEM communicates prototypes of high-attention instances from positive bags, refines the current uploads into a shared memory, and regularizes local attention on that memory. Its focus is therefore the communication and reuse of

weakly supervised instance evidence across clients, rather than class-mean alignment, synthetic-slide training, or language-model preference distillation.

C Details of the Assumption

Assumption 1 (L-smoothness)**.** *For every client k and every fixed memory $\mathcal{M}$, the function $F_k(\theta, \mathcal{M})$ is L-smooth with respect to θ, i.e., $\|\nabla_\theta F_k(\theta, \mathcal{M}) - \nabla_\theta F_k(\theta', \mathcal{M})\| \le L\|\theta - \theta'\|$ for all θ, θ'.*

Assumption 2 (Bounded stochastic gradient variance)**.** *For every client k, the stochastic gradient $g_k(\theta; \xi_k)$ computed from a mini-batch $\xi_k \sim \mathcal{D}_k$ is unbiased and has bounded variance: $\mathbb{E}_{\xi_k}[g_k(\theta; \xi_k)] = \nabla_\theta F_k(\theta, \mathcal{M})$ and $\mathbb{E}_{\xi_k}\|g_k(\theta; \xi_k) - \nabla_\theta F_k(\theta, \mathcal{M})\|^2 \le \sigma^2$.*

Assumption 3 (Bounded client heterogeneity)**.** *There exists $G \ge 0$ such that for every θ and every $\mathcal{M}$,*

$$\frac{1}{K}\sum_{k=1}^{K}\left\|\nabla_\theta F_k(\theta, \mathcal{M}) - \nabla_\theta F(\theta, \mathcal{M})\right\|^2 \;\le\; G^2.$$

Assumption 4 (Bounded memory drift)**.** *There exist non-negative scalars $\{\Delta_{\mathcal{M}}^t\}_{t\ge 0}$ such that for any θ,*

$$\left|F(\theta, \mathcal{M}^{t+1}) - F(\theta, \mathcal{M}^t)\right| \;\le\; \Delta_{\mathcal{M}}^t, \qquad \Delta_{\mathcal{M}} \;:=\; \sum_{t=0}^{T-1}\Delta_{\mathcal{M}}^t < \infty.$$

D Additional Method Details

D.1 Federated Setting.

We simulate a federated learning environment with $K = 10$ clients. The training data are partitioned across clients using a Dirichlet distribution over class labels, which induces non-IID data distributions. The concentration parameter β controls the degree of heterogeneity, and we consider $\beta \in \{0.25, 0.5, 0.75\}$ in our experiments, where smaller β corresponds to more skewed data distributions. During training, each client performs 5 local epochs before uploading model updates to the server. The total number of communication rounds is set to 100, and the server aggregates local models via weighted averaging in each round.

D.2 Evaluation Metric.

We evaluate all methods using bag-level classification accuracy (ACC). ACC is computed as the proportion of correctly classified test bags across clients. Since instance-level annotations are unavailable, all evaluations are conducted at the bag level, which is consistent with the weakly supervised MIL setting.

D.3 FedSEM Training Pipeline

Algorithm 1 summarizes the full training procedure of FedSEM, and Algorithm 2 details the server-side shared-memory refinement step. At each communication round, the server samples online clients and broadcasts the current global model. Each selected client performs local MIL optimization and extracts high-attention evidence instances from positive bags. The selected embeddings are compressed into prototype summaries before upload. During the warm-up period, the server only accumulates and refines uploaded prototypes. After the warm-up period, the refined shared evidence memory is broadcast to clients and used as evidence anchors for memory-guided local optimization.

Algorithm 1: FedSEM: Federated Shared Evidence Memory

Require: Client datasets $\{\mathcal{D}_k\}_{k=1}^K$; initial model θ^0; communication rounds T; local epochs E; warm-up rounds T_w; client sampling rule $\pi(\cdot)$; memory size M

Ensure: Trained global model θ^T and shared evidence memory $\mathcal{M}^T$

1: Initialize shared evidence memory $\mathcal{M}^0 \leftarrow \emptyset$

2: **for** $t = 0, \ldots, T-1$ **do**

Sample participating clients $\mathcal{S}^t \sim \pi(K)$
Broadcast θ^t to all $k \in \mathcal{S}^t$
for all clients $k \in \mathcal{S}^t$ **in parallel do**
  $\theta_k^{t,0} \leftarrow \theta^t, \mathcal{U}_k^t \leftarrow \emptyset$
  if $t < T_w$ **then**
    $\mathcal{M}_k^t \leftarrow \emptyset$
  else
    $\mathcal{M}_k^t \leftarrow \mathcal{M}^t$
  end if
  for $e = 0, \ldots, E-1$ **do**
    for all bags $(\mathcal{B}_{k,i}, y_{k,i}) \in \mathcal{D}_k$ **do**
      Compute instance embeddings $\{h_{k,i,j}\}_{j=1}^{m_{k,i}}$ and attention scores $\{a_{k,i,j}\}_{j=1}^{m_{k,i}}$
      Update $\theta_k^{t,e}$ by minimizing Eq. (9) with memory $\mathcal{M}_k^t$
      if $y_{k,i} = 1$ **then**
        $\mathcal{E}_{k,i}^t \leftarrow \{(h_{k,i,j}, a_{k,i,j}) : a_{k,i,j} \geq \tau_{\text{att}}\}$
        if $\mathcal{E}_{k,i}^t = \emptyset$ **then**
          $\mathcal{E}_{k,i}^t \leftarrow$ top-k_{up} instances in $\mathcal{B}_{k,i}$ ranked by $a_{k,i,j}$
        end if
        $\mathcal{U}_k^t \leftarrow \mathcal{U}_k^t \cup \mathcal{E}_{k,i}^t$
      end if
    end for
  end for
  $\mathcal{P}_k^t \leftarrow$ ClusterToPrototypes$(\mathcal{U}_k^t)$
  Upload $(\theta_k^{t,E}, \mathcal{P}_k^t, n_k)$ to the server
end for
$\theta^{t+1} \leftarrow \sum_{k \in \mathcal{S}^t} \frac{n_k}{\sum_{u \in \mathcal{S}^t} n_u} \theta_k^{t,E}$
$\mathcal{M}^{t+1} \leftarrow$ RefineMemory$(\mathcal{M}^t, \{\mathcal{P}_k^t\}_{k \in \mathcal{S}^t}, M)$
end for
return $\theta^T, \mathcal{M}^T$

D.4 Local Prototype Construction

For each positive bag, FedSEM first ranks instances according to the attention scores produced by the local MIL model. Instances whose attention scores are larger than the threshold are selected as local evidence candidates. When no instance satisfies the threshold, FedSEM falls back to the top-k highest-attention instances in the bag. This fallback is important in early rounds, because the attention module may still be under-trained and overly flat.

The selected candidates collected from all positive bags on a client are then compressed by density-based clustering in the embedding space. In our implementation, we use cosine distance and cluster the selected embeddings into evidence modes. Each cluster is represented by an attention-weighted prototype, while isolated samples are retained as individual prototypes instead of being discarded. This design avoids losing rare but potentially discriminative evidence patterns and substantially reduces the number of uploaded instance embeddings compared with direct all-feature upload.

D.5 Server-side Memory Refinement

The server receives local prototype sets from online clients and merges them with the previous memory. The refinement step follows three principles. First, highly similar prototypes are greedily deduplicated by cosine similarity, prioritizing prototypes with higher importance scores. Second, when the candidate set is larger than the memory budget, FedSEM applies farthest-first K-center selection so that the retained prototypes cover diverse evidence modes rather than only the most frequent or highest-score modes. Third, prototype scores are normalized before being stored in the memory, which reduces scale differences induced by heterogeneous local attention distributions.

The raw-memory ablation in Table 2 removes the server-side refinement and directly forwards uploaded prototypes. The full FedSEM variant consistently benefits from refinement, indicating that

the memory quality is not determined only by whether evidence is shared, but also by how redundant and noisy evidence is filtered before feedback.

Algorithm 2: Server-side shared evidence memory refinement

Require: Previous memory $\mathcal{M}^t$; uploaded prototype sets $\{\mathcal{P}_k^t\}_{k\in\mathcal{S}^t}$; memory size M; deduplication threshold τ_{dedup}
Ensure: Refined memory $\mathcal{M}^{t+1}$

$\mathcal{C}^t \leftarrow \mathcal{M}^t \cup \bigcup_{k\in\mathcal{S}^t} \mathcal{P}_k^t$
for all classes c represented in $\mathcal{C}^t$ **do**
  $\mathcal{C}_c^t \leftarrow \{(p,s) \in \mathcal{C}^t : \text{class}(p) = c\}$
  Sort $\mathcal{C}_c^t$ by score s in descending order
  $\mathcal{R}_c^t \leftarrow \emptyset$
  for all $(p,s) \in \mathcal{C}_c^t$ **do**
    if $\mathcal{R}_c^t = \emptyset$ **or** $\max_{(q,\cdot)\in\mathcal{R}_c^t} \cos(p,q) < \tau_{\text{dedup}}$ **then**
      $\mathcal{R}_c^t \leftarrow \mathcal{R}_c^t \cup \{(p,s)\}$
    end if
  end for
  if $|\mathcal{R}_c^t| > M$ **then**
    $\mathcal{R}_c^t \leftarrow \text{KCENTERSELECT}(\mathcal{R}_c^t, M)$
  end if
  Normalize scores in $\mathcal{R}_c^t$ into $[0,1]$
end for
$\mathcal{M}^{t+1} \leftarrow \bigcup_c \mathcal{R}_c^t$
return $\mathcal{M}^{t+1}$

E Dataset Details

We evaluate FedSEM on two pathology WSI datasets and two general MIL benchmarks. All four datasets are converted into a bag-level classification protocol, where each slide or image is a bag and its patches or regions are instances. Only bag-level labels are used for training. Instance-level annotations, when available in the original data source, are not used as supervision.

E.1 CAMELYON16

CAMELYON16 is a breast cancer metastasis detection benchmark built from hematoxylin-and-eosin stained lymph-node whole-slide images [Bejnordi et al., 2017]. Following the MIL formulation used in computational pathology, each WSI is treated as one bag, and the bag label indicates whether the slide contains tumor tissue. We convert the reference labels into a binary target, where tumor slides are positive and normal slides are negative. Each WSI is tiled into 224×224 patches at the selected slide level, and low-saturation background patches are removed by a tissue filter. During training, a fixed maximum number of patches is sampled from each slide to control memory and compute cost.

For federated simulation, slide-level training samples are partitioned across clients with a Dirichlet distribution over binary labels. This produces client-specific class imbalance and reflects cross-institution heterogeneity. The validation split is created before client partitioning with stratified sampling, so that evaluation is conducted on held-out bags rather than on client training bags.

E.2 PANDA

PANDA is a large-scale prostate pathology WSI benchmark for prostate cancer grade assessment [Bulten et al., 2022]. We use it as a second medical MIL benchmark with different tissue morphology, stain variation, and disease evidence patterns from CAMELYON16. Each prostate WSI is treated as one bag. The original slide-level diagnostic labels are converted into a binary classification task using the provided binary reference file: cancer-positive or high-grade slides are treated as positive bags, and benign or non-target slides are treated as negative bags.

The WSI preprocessing protocol follows the same pipeline as CAMELYON16. Slides are tiled into 224×224 patches, background-dominated patches are removed using a saturation threshold, and a

maximum of 64 patches per bag is sampled during training. The same Dirichlet partitioning protocol is used to simulate non-IID federated clients. This dataset is particularly useful for testing whether the shared evidence memory can remain stable under large visual variability and client-level label skew.

E.3 Pascal VOC 2012

Pascal VOC 2012 is a natural-image recognition benchmark with 20 object categories [Everingham et al., 2010]. To adapt it to MIL, each image is treated as one bag, and object proposal regions are treated as instances. We construct a binary one-vs-rest task from the trainval split, using *person* as the target positive category. The manifest records the image id, image path, binary bag label, and the list of object categories in the image. Images containing at least one *person* instance are labeled as positive bags, while images that do not contain the *person* category are labeled as negative bags. Difficult objects are ignored when constructing the manifest unless explicitly included.

Instances are generated by Selective Search [Uijlings et al., 2013]. Each proposal is cropped from the original image and resized to 224×224. We keep at most 50 proposals per image and discard boxes whose shorter side is smaller than 20 pixels. This creates a standard weakly supervised MIL setting: the bag label tells whether the target object category appears in the image, but the learner does not receive the proposal-level object label or bounding-box supervision.

E.4 MNIST-Bags

MNIST-Bags is a synthetic MIL benchmark derived from MNIST digit images [LeCun et al., 1998, Zhang et al., 2022]. Each bag contains a variable number of digit instances. A bag is labeled positive if it contains the target digit and negative otherwise. In our experiments, the target digit is set to "9"; therefore, bags containing at least one digit "9" instance are treated as positive bags, while bags without digit "9" are treated as negative bags. The mean bag length is 10, the bag-length variance is 2, and 250 training bags are generated. The single-channel MNIST digit images are repeated across three channels so that the same ResNet-based encoder can be used as in the other datasets.

Unlike the WSI and VOC settings, MNIST-Bags has a clean instance-generation mechanism and a controlled target concept. It therefore serves as a sanity-check benchmark for whether FedSEM can align evidence patterns when the discriminative instance is visually compact and semantically unambiguous. The federated partition is still performed at the bag level with a Dirichlet distribution over positive and negative bag labels.

Table 4: Dataset characteristics used in our federated MIL experiments.

Dataset	Domain	#bag	Train / Val	#clients	Instance type	Max #ins/bag	#dim	#class
CAMELYON16	Pathology WSI	270	216 / 54	100 (10 active)	Tissue patch	64	512	2
PANDA	Pathology WSI	10,606	8,485 / 2,121	100	Tissue patch	64	512	2
Pascal VOC 2012	Natural image	11,540	9,232 / 2,308	100	Selective Search region	50	512	2
MNIST-Bags	Digits	250	200 / 50	10	Digit image	mean 10, var 2	512	2

F Experimental Details

F.1 Federated Configuration

Unless otherwise specified, all methods are trained for 100 communication rounds with 5 local epochs per communication round. The default number of clients is 10 in the main experiments, and a client fraction of 0.1 is used when simulating larger federations. For CAMELYON16, we select an active subset of clients before partitioning to avoid extremely small client datasets in the small-slide regime. The Dirichlet concentration parameter β controls label-distribution heterogeneity; smaller β produces stronger client skew.

All reported results use bag-level accuracy as the primary metric, consistent with the weakly supervised MIL setting. During training and validation, no instance-level labels are used. For binary tasks, model outputs are thresholded at 0.5 to compute accuracy.

F.2 Optimization and Architecture

All datasets use a ResNet-18 instance encoder followed by an attention-based MIL aggregation head. The local optimizer uses a learning rate of 10^{-4}, weight decay of 10^{-4}, and gradient clipping with norm 10. For FedSEM, the default shared-memory upload mode is clustered prototype upload. The attention threshold is 0.8, the upload budget is top-32 instances per bag, the shared memory pool size is 256, the prototype clustering radius is 0.5 under cosine distance, and the minimum cluster size is 2. The first 5 communication rounds are used as memory warm-up rounds.

Table 5: Default implementation hyperparameters used by FedSEM. These defaults are fixed unless the corresponding parameter is explicitly varied in Appendix G.

Group	Hyperparameter	Default value
Federated training	Communication rounds	100
Federated training	Local epochs per round	5
Federated training	Client participation ratio	0.1
Optimization	Local learning rate	1×10^{-4}
Optimization	Weight decay	1×10^{-4}
Optimization	Gradient clipping norm	10
Model	Backbone	ResNet-18
MIL preprocessing	WSI patch size / VOC proposal size	224×224
FedSEM memory	Warm-up rounds	5
FedSEM memory	Attention threshold	0.8
FedSEM memory	Upload budget top-k	32
FedSEM memory	Shared pool size	256
FedSEM memory	Prototype clustering radius	0.5
FedSEM memory	Prototype clustering minimum samples	2
FedSEM memory	Server deduplication threshold τ	0.95

G Parameter Analysis

This section reports the parameter analysis experiments for FedSEM. Each experiment varies one parameter while keeping the remaining settings fixed to the default configuration in Appendix F.

G.1 Memory Alignment Weight λ

The coefficient λ controls the strength of the memory-guided evidence alignment term in Eq. (9). Small values weaken cross-client evidence alignment, while overly large values may force local attention to overfit the shared memory and reduce the effect of bag-level supervision. Table 6 shows that the default range around $\lambda = 0.5$–0.7 provides strong performance.

Table 6: Parameter analysis of λ under bag-level ACC(%).

λ	CAMELYON16	PANDA	Pascal VOC 2012	MNIST-Bags
0.3	61.11	76.71	65.90	82.00
0.5	77.78	76.28	69.24	94.00
0.7	64.81	75.43	71.14	92.00
0.8	70.37	77.46	68.59	84.00

G.2 Proximal Coefficient μ

The coefficient μ controls the FedProx-style local proximal regularization. It constrains client drift under heterogeneous local distributions, but too strong a constraint can also limit local adaptation. Table 7 indicates that a moderate proximal coefficient is generally preferable, while large μ values may degrade performance on several datasets.

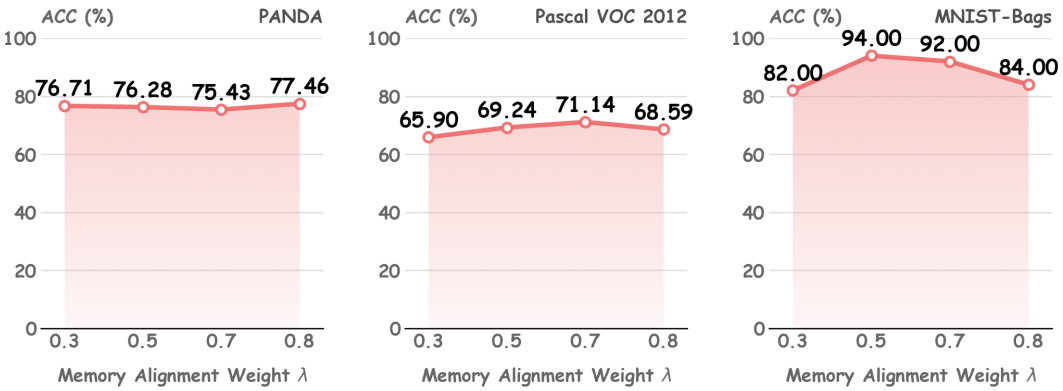

Figure 3: Additional λ sensitivity panels not shown in the main text.

Table 7: Parameter analysis of μ under bag-level ACC(%).

μ	CAMELYON16	PANDA	Pascal VOC 2012	MNIST-Bags
0.001	77.78	76.28	71.14	90.00
0.005	70.37	75.43	66.46	94.00
0.01	64.81	76.52	64.69	88.00
0.03	64.81	76.10	66.72	84.00
0.1	61.11	72.75	64.64	84.00

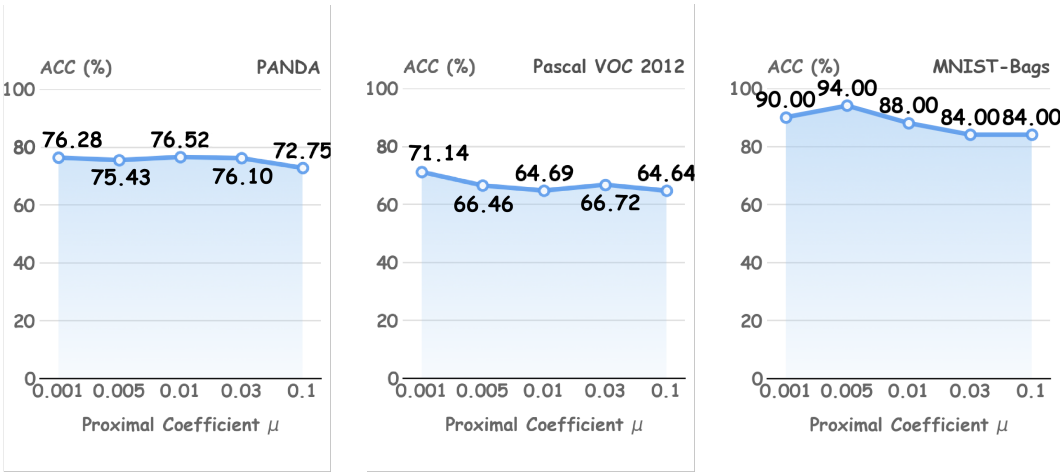

Figure 4: Additional μ sensitivity panels not shown in the main text.

G.3 Number of Clients

We vary the number of federated clients to study scalability under different federation sizes. Increasing the number of clients usually reduces the amount of data available to each client and increases the severity of client drift, especially under label-skewed partitioning. The current results in Table 8 show that FedSEM remains competitive when increasing the client count.

G.4 Dirichlet Heterogeneity Parameter β

The Dirichlet concentration parameter β determines the label-distribution skew across clients. Smaller β values produce stronger non-IID partitions, while larger values move the partition toward a

Table 8: Parameter analysis of the number of clients under bag-level ACC.

Client number	CAMELYON16	PANDA	Pascal VOC 2012	MNIST-Bags
5	62.96	76.19	65.47	92.00
10	77.78	76.28	71.14	94.00
20	62.96	76.10	68.93	84.00
50	61.11	76.52	68.93	88.00

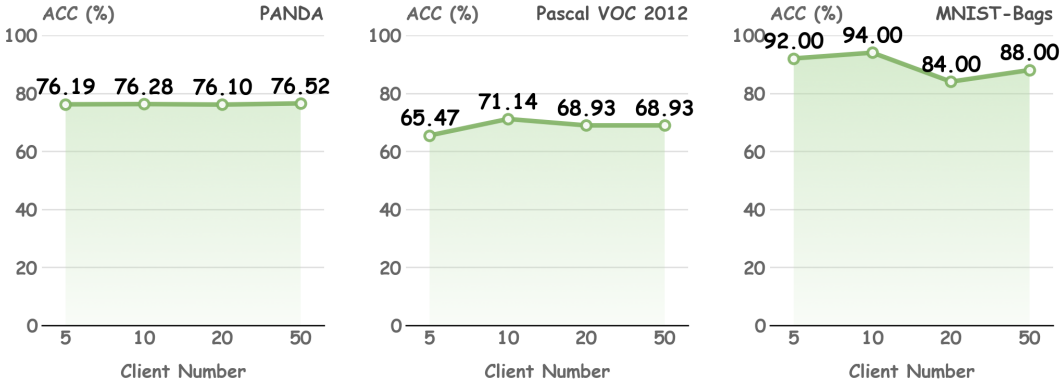

Figure 5: Additional client-number sensitivity panels not shown in the main text.

more balanced distribution. Table 9 complements the main results by isolating the effect of data heterogeneity on FedSEM.

Table 9: Parameter analysis of Dirichlet heterogeneity β under bag-level ACC(%).

β	CAMELYON16	PANDA	Pascal VOC 2012	MNIST-Bags
0.25	66.67	75.20	65.29	86.00
0.3	64.81	73.08	66.81	84.00
0.5	77.78	77.46	71.14	94.00
0.75	68.52	77.51	70.84	90.00
0.8	61.11	75.77	65.12	90.00
1.0	61.11	76.42	70.10	84.00

G.5 Upload Budget Top-k

The top-k parameter controls the maximum number of high-attention instances selected per bag before client-side clustering. It is both a performance parameter and a communication parameter. Small values may fail to cover diverse evidence patterns, whereas large values increase communication and may introduce additional noisy candidates. As shown in Table 10, intermediate values such as $k = 32$ provide a strong trade-off in the runs.

Table 10: Parameter analysis of upload budget top-k under bag-level ACC(%).

Top-k	CAMELYON16	PANDA	Pascal VOC 2012	MNIST-Bags
8	64.81	78.17	64.60	84.00
16	66.67	78.03	65.03	90.00
32	77.78	76.28	71.14	94.00
64	59.26	76.85	64.64	90.00

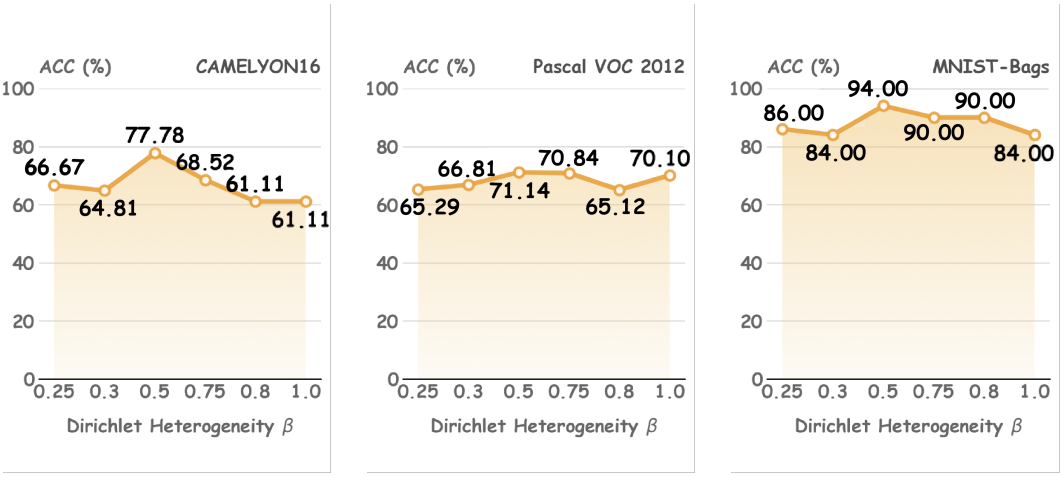

Figure 6: Additional β sensitivity panels not shown in the main text.

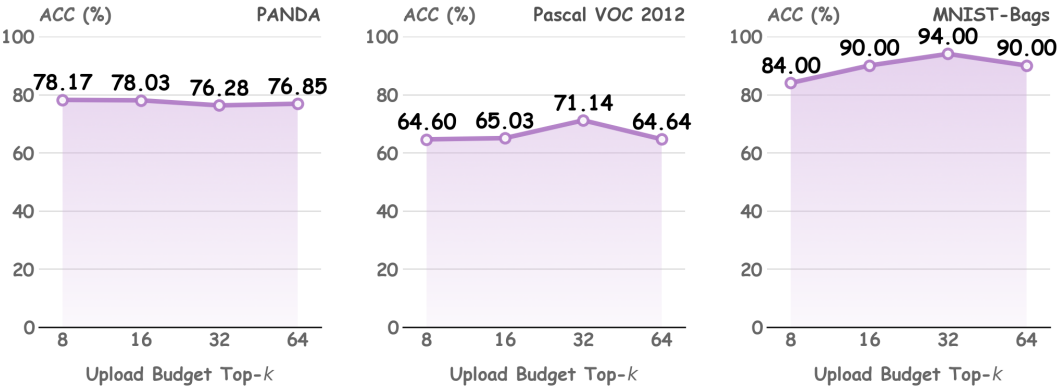

Figure 7: Additional Top-k sensitivity panels not shown in the main text.

G.6 Server Deduplication Threshold τ

The threshold τ controls server-side cosine deduplication. A smaller value removes more prototypes as duplicates, leading to a more compact but potentially less diverse memory. A larger value preserves more fine-grained prototypes, but may keep redundant evidence anchors. Table 11 suggests that $\tau = 0.95$ is a stable default among the tested values.

Table 11: Parameter analysis of server deduplication threshold τ under bag-level ACC(%).

τ	CAMELYON16	PANDA	Pascal VOC 2012	MNIST-Bags
0.85	66.67	76.05	68.02	86.00
0.95	77.78	77.46	71.14	94.00
0.98	70.37	76.47	65.55	90.00
0.995	62.96	76.28	65.38	90.00

H Feature-attention Heatmaps

Feature-attention heatmaps Fig 9 qualitatively evaluate whether FedSEM focuses on discriminative evidence regions. For WSI datasets, attention scores are mapped back to patch coordinates on the

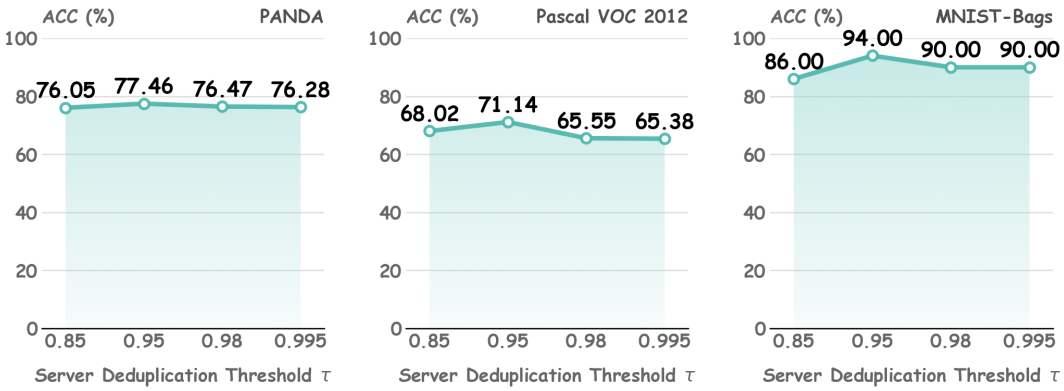

Figure 8: Additional τ sensitivity panels not shown in the main text.

slide thumbnail. For Pascal VOC 2012, proposal-level attention scores are projected onto image regions. For MNIST-Bags, instance attention scores are displayed over the digit instances in each bag. These visualizations are intended to compare FedSEM with parameter-only federated baselines and show whether the shared evidence memory reduces attention drift across clients.

I Additional Discussion

I.1 Limitations

FedSEM shares attention-derived importance, but can retain systematic label-correlated shortcuts, as the pseudo-patch test in Appendix J illustrates. Warm-up delays sharing, and deduplication reduces repeated prototypes; neither guarantees the removal of incorrect evidence. The current evaluation focuses on simulated federated partitions; real multi-institution deployment may introduce additional shifts in staining, scanning devices, labeling policies, and patient populations. Prototypes and model updates are visible to the server and broadcast prototypes to participating clients. They may reveal membership, class prevalence, or reconstructible feature information. No differential-privacy guarantee, secure prototype aggregation, or privacy-attack evaluation is established here. Fewer transmitted embeddings do not prove lower leakage. Most original results are single-run; repeated experiments cover MNIST-Bags only. Best-validation checkpoint selection can be optimistic, and simulations do not validate clinical deployment.

Broader impact. Collaborative MIL may improve access to diverse training data without pooling raw patient images. However, unrecognized shortcuts, uneven institutional performance, and leakage through representations could harm patients or underrepresented groups. Clinical use would require external validation, privacy assessment, and human oversight.

I.2 Future Directions

Future work will extend FedSEM in three directions. First, we plan to add uncertainty-aware memory updates so that low-confidence or unstable prototypes are down-weighted automatically. Second, we will study stronger privacy mechanisms for prototype communication, such as secure aggregation or differentially private perturbation of uploaded prototypes. Third, we will evaluate FedSEM on real cross-center WSI cohorts where client identity corresponds to actual institutions rather than simulated Dirichlet partitions.

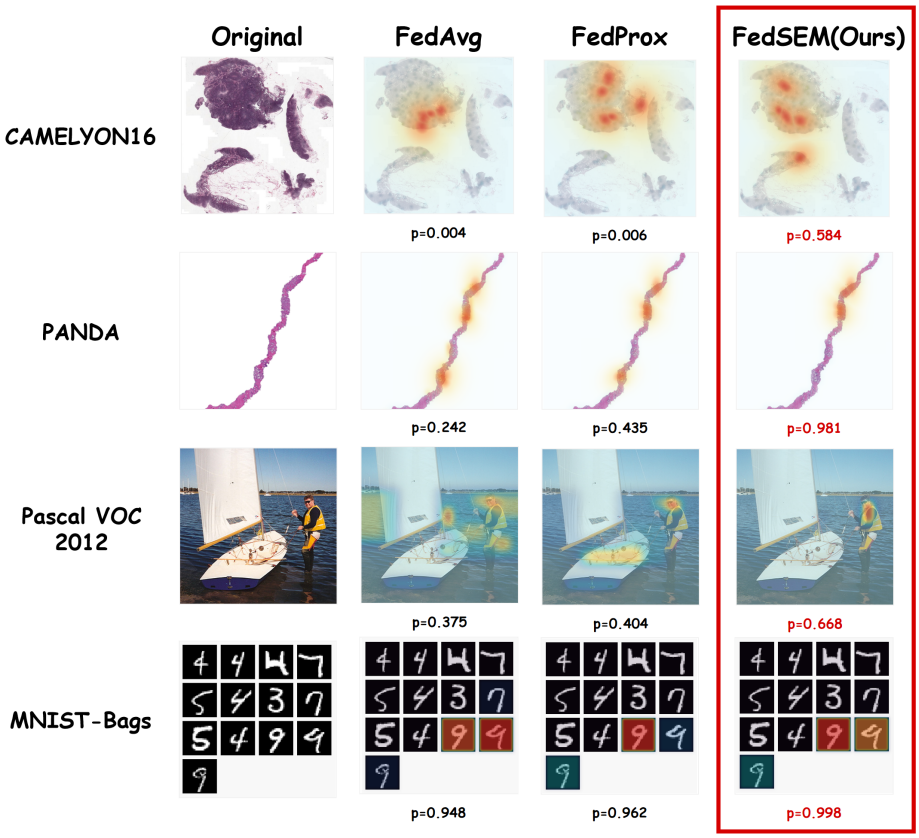

Figure 9: Qualitative feature-attention heatmaps, where p denotes the model-predicted probability that the input bag belongs to the positive class

J Additional Diagnostics and Reproducibility

J.1 Repeated-run protocol and client distributions

We reaggregate 84 archived histories: 63 MNIST-Bags benchmark runs, 12 pseudo-evidence runs, and 9 memory-refresh runs. Each history contains 100 rounds. The benchmark/refresh histories for seeds 43 and 44 have identical non-timing recorded trajectories in each corresponding pair; they are not treated as independent additional repetitions. Results below are the arithmetic mean and sample standard deviation ($n = 3$, denominator $n - 1$) over seeds 42, 43, and 44. For each run, we select the earliest maximum-validation-ACC round and read ACC, AUC, and balanced ACC from that same checkpoint. We do not separately maximize each metric. All 84 selected rounds agree with the saved best-ACC metadata. The histories identify run seeds but do not record the complete bag-construction and split-seed settings; their contributions to variability cannot be separated. The reported standard deviation is not a confidence interval or a significance test. Validation uses 50 bags; repeated tuning and checkpoint selection on this small set can yield optimistic estimates.

FedSEM is competitive with FedProx, with overlapping empirical variability. At $\beta = 0.25$, FedProx has the higher mean; at 0.5 and 0.75 FedSEM has the higher mean. These three runs do not establish a general ranking.Finite realized partitions need not become monotonically more balanced as β increases.

J.2 Controlled pseudo-evidence stress test

The MNIST-Bags diagnostic adds a fixed label-correlated pseudo patch to positive training bags. Clean validation contains no injected patch. A separate probe injects the pseudo patch into positive validation bags (25 probe bags) and uses known digit/pseudo identities only for evaluation. This

Table 12: MNIST-Bags validation ACC (%), mean ± sample SD over three seeds. These repeated-run results are distinct from the historical single-run main table.

Method	$\beta = 0.25$	$\beta = 0.5$	$\beta = 0.75$
DSMIL	82.00 ± 7.21	82.67 ± 11.37	85.33 ± 9.87
FedAvg	88.00 ± 4.00	92.00 ± 3.46	89.33 ± 2.31
FedProx	91.33 ± 4.62	93.33 ± 3.06	91.33 ± 4.16
FedRDN	52.00 ± 2.00	60.00 ± 12.17	60.67 ± 11.72
MIL-Dropout	70.67 ± 7.57	79.33 ± 7.02	80.00 ± 10.39
MOON	87.33 ± 5.03	88.00 ± 3.46	68.00 ± 13.11
FedSEM	90.67 ± 4.16	94.00 ± 4.00	92.00 ± 4.00

construction tests a systematic shortcut under synthetic conditions, not the reliability of real pathology annotations. Table 13 uses the checkpoint rule above. Table 14 instead uses the final round (100) for every method, so its attention and retention statistics refer to a common training time. The two tables must not be interpreted as diagnostics of the same selected checkpoint.

Table 13: Clean-validation performance under pseudo-patch training, three seeds. AUC and balanced ACC use the best-ACC checkpoint; the last column reports round-100 ACC separately.

Method	Best ACC (%)	AUC (%)	Bal. ACC (%)	Final ACC (%)
FedAvg	58.00 ± 2.00	61.76 ± 4.74	58.00 ± 2.00	50.00 ± 0.00
FedProx	57.33 ± 7.02	63.20 ± 10.59	57.33 ± 7.02	50.00 ± 0.00
FedSEM-Raw	56.67 ± 8.33	58.40 ± 8.07	56.67 ± 8.33	50.00 ± 0.00
FedSEM-Filtered	56.00 ± 7.21	60.32 ± 11.35	56.00 ± 7.21	50.00 ± 0.00

Table 14: Final-round pseudo-patch probe and memory retention, three seeds. Mass is the summed attention on the indicated instances, averaged across probe bags. Top-1 is the fraction whose maximum-attention instance is the pseudo patch. Retention is retained/uploaded pseudo mass in prototype diagnostic flags; it is not the fraction of causally valid prototypes. A dash means no memory.

Method	Pseudo mass	Target mass	Pseudo top-1	Retention
FedAvg	0.4016 ± 0.4723	0.1195 ± 0.1000	0.6667 ± 0.5774	–
FedProx	0.9691 ± 0.0348	0.0061 ± 0.0072	1.0000 ± 0.0000	–
FedSEM-Raw	0.9813 ± 0.0172	0.0034 ± 0.0030	1.0000 ± 0.0000	1.0000 ± 0.0000
FedSEM-Filtered	0.9806 ± 0.0292	0.0033 ± 0.0049	1.0000 ± 0.0000	0.1120 ± 0.0016

For FedSEM-Filtered, the selected rounds are 5, 1, and 1 for seeds 42, 43, and 44: all precede the first possible memory-guided update in round 7. For FedSEM-Raw, they are 5, 1, and 37. Thus the best-checkpoint table, especially its filtered row, does not directly evaluate clean performance after memory guidance has been enabled. Both memory variants end at 50% ACC for all three seeds. FedAvg and FedProx have slightly higher best-checkpoint mean ACC than the memory variants. Filtering reduces pseudo-prototype retention from 1 to approximately 0.112, yet final-round attention still concentrates on the injected shortcut. Therefore duplicate removal can compress repeated pseudo evidence without identifying or removing the underlying confounder. High attention and cross-client agreement are insufficient to establish semantic correctness or causal relevance.

J.3 Memory refresh and feature-drift diagnostics

We compare default per-round refresh, decay retention (old scores multiplied by 0.9 before pooling and filtering), and periodic refresh (every fifth co-training update). The periodic histories record 19 updates and 76 skips over 95 co-training rounds, giving an average stale gap of 2 rounds. For classes

present in both consecutive memories, the logged diagnostic is

$$d_t = \frac{1}{|\mathcal{C}_t|} \sum_{c \in \mathcal{C}_t} \left[1 - \frac{1}{M_c^t} \sum_{r=1}^{M_c^t} \max_{q \le M_c^{t-1}} \cos(m_{c,r}^t, m_{c,q}^{t-1}) \right]. \tag{16}$$

This directed nearest-prototype cosine discrepancy ignores unmatched classes and scores. The first undefined comparison is excluded; skipped periodic updates contribute logged zero drift. It does not bound objective drift in Theorem 1. Mean size and stale gap average all logged co-training rounds.

Table 15: Memory strategies on MNIST-Bags, three seeds. Performance uses the best-ACC checkpoint; drift/size summarize the full trajectory. Hours sum logged round times, without assuming matched hardware or scheduling.

Variant	ACC (%)	AUC (%)	Mean drift	Mean size	Hours
Refresh	94.00 ± 4.00	94.93 ± 3.37	0.1293 ± 0.0430	3.50 ± 0.62	5.29 ± 0.08
Decay retention	90.67 ± 6.11	89.97 ± 5.69	0.0058 ± 0.0005	128.01 ± 37.16	1.70 ± 0.02
Periodic refresh	92.67 ± 3.06	92.27 ± 4.42	0.0295 ± 0.0035	3.75 ± 0.67	1.59 ± 0.05

Per-round refresh has the highest mean accuracy among these variants. Decay retention lowers measured drift but retains substantially more entries. Periodic refresh lowers average drift partly by construction because unchanged rounds have zero drift. Lower feature drift alone therefore need not imply a more useful or more current memory. The results do not verify summable objective drift or prove an optimal refresh schedule.

J.4 Communication and computation accounting

Let B_θ be the serialized model-state tensor bytes, S_t the number of participating clients, Q_k^t client k's uploaded prototype count, and $M_t = \sum_c M_c^t$ the shared count. With float32 embeddings of dimension d and one float32 score, each prototype requires $4(d+1)$ bytes before metadata. For point-to-point transfers to participating clients, round t has

$$B_{\text{up}}^t = S_t B_\theta + 4(d+1) \sum_{k \in \mathcal{S}^t} Q_k^t, \tag{17}$$

$$B_{\text{down}}^t = S_t B_\theta + 4(d+1) S_t M_t. \tag{18}$$

If memory is broadcast to all clients, replace the second S_t by K; network multicast has different physical accounting. For ResNet-18 features ($d = 512$), a prototype takes 2052 bytes; a full binary memory of 256 entries takes 525312 bytes (0.525 MB), or 5.253 MB for ten point-to-point recipients. These are analytic payload estimates, not measured network traffic. Headers, serialization overhead, diagnostic flags, and retransmissions are excluded. Model-state bytes should include buffers, not only trainable parameters. Server refinement primarily reduces subsequent prototype *downlink*; it does not retroactively reduce the current upload, whose reduction comes from local selection and clustering.

Client compression may have quadratic pairwise-distance cost in the number of locally selected embeddings. Greedy server deduplication has worst-case quadratic similarity cost in uploaded count, and farthest-first selection costs $\mathcal{O}(MQd)$ for Q candidates. The memory loss adds head computation on the shared embeddings at each local training step, without re-encoding raw images. Table 15 provides logged runtime, but the archive does not identify consistent hardware, peak memory, or the full development budget. The runtime differences are not a controlled efficiency comparison.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes].

   Justification: The abstract and introduction clearly state the main problem, namely cross-client critical evidence drift in federated MIL, and summarize the proposed FedSEM framework, theoretical convergence result, and empirical evaluation. The claims are consistent with the method design in Sec. 2, the convergence analysis in Sec. 3, and the experimental results in Sec. 4.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes].

   Justification: The paper includes a dedicated limitation discussion in Appendix I.1. It discusses potential noisy memory entries caused by biased early attention, the gap between simulated federated partitions and real multi-institution deployment, and the privacy considerations of transmitting learned prototypes.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes].

   Justification: The paper provides the main convergence theorem in Sec. 3 and gives the proof following the theorem. The assumptions used in the theorem are provided in Appendix C, including smoothness, bounded variance, bounded client heterogeneity, and bounded memory drift.

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

4. **Experimental result reproducibility**

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

   Answer: [Yes].

   Justification: The paper describes the evaluated datasets, baselines, metrics, federated configuration, model architecture, optimization settings, and FedSEM-specific hyperparameters in Sec. 4 and Appendix E–F. The training pipeline and server-side memory refinement procedure are also provided in Algorithm 1 and Algorithm 2.

   Guidelines:

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

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [No].

   Justification: The paper uses publicly available benchmark datasets and provides detailed experimental settings.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes].

   Justification: The paper specifies dataset construction and preprocessing in Appendix E, federated configuration in Appendix F.1, and optimization/model hyperparameters in Appendix F.2. The evaluation metric is also described in Appendix D.2.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No].

   Justification: The current version reports deterministic accuracy values across datasets and heterogeneity settings but does not include error bars, confidence intervals, or statistical significance tests. This is mainly due to the high computational cost of repeated federated MIL experiments on large WSI-style datasets.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [No].

   Justification: The paper reports the model architecture, optimization hyperparameters, and communication settings, but it does not currently disclose the GPU/CPU type, memory, runtime per experiment, or total compute budget.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes].

   Justification: The work studies privacy-preserving federated MIL using existing benchmark datasets and does not involve deceptive, harmful, or unsafe data collection practices. The paper also discusses privacy-related limitations and notes that prototype communication should be combined with standard privacy-preserving safeguards in sensitive medical settings.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes].

    Justification: Appendix I.1 discusses collaborative-learning benefits and risks of leakage, shortcuts, unequal performance, and unvalidated clinical deployment.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A].

    Justification: The paper does not release a high-risk pretrained model, image generator, scraped dataset, or other asset with substantial misuse potential. The method is evaluated on existing MIL benchmarks and does not introduce a new public high-risk dataset or generative model.

    Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [No].

Justification: The paper cites the original sources of the datasets and related methods, including CAMELYON16, PANDA, Pascal VOC 2012, MNIST, and MIL baselines. However, the current version does not explicitly mention the license or terms of use for each existing dataset or asset.

Guidelines:

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

13. **New assets**

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

Answer: [N/A].

Justification: The paper does not introduce or release a new dataset, benchmark, pretrained model, or other standalone asset. The contribution is a learning framework evaluated on existing datasets.

Guidelines:

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

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

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

Answer: [N/A].

Justification: The paper does not involve crowdsourcing, user studies, or newly recruited human participants. All experiments are conducted on existing benchmark datasets.

Guidelines:

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

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

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

    Answer: [N/A].

    Justification: The paper does not conduct new human-subject research or collect new human data. It uses existing benchmark datasets for federated MIL experiments.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A].

    Justification: The core method does not use LLMs as an important, original, or non-standard component. LLMs were used to assist with language polishing, data organization, and manuscript formatting during the research and paper preparation process.

    Guidelines:

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