OPERATOR-SPACE FEDERATED LEARNING WITH DIFFUSION-BASED NEURAL OPERATORS FOR HETEROGENEOUS MULTIMODAL SENSING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated multimodal human sensing involves substantial heterogeneity across sensor modalities, spatial and temporal resolutions, local tokenizations, and client data distributions. We formulate federation over a shared semantic operator representation that supports aggregation across these heterogeneous local discretizations. A Spectral-Gated Linear Transformer constructs the semantic matrix $M = \phi(K)^\top V$, whose dimension is independent of the local token count, and applies spectral gating to retain dominant shared structure while suppressing modality-specific residual components. We further introduce Diffusion-Based Neural Operator Aggregation for server-side federation. Client semantic updates are represented in low-dimensional spectral coordinates, where a diffusion model learns the distribution used to form the global semantic update. The remaining model parameters are aggregated with FedAvg. A Subspace-Projected Differential Privacy mechanism protects the released semantic updates using a projection basis derived from the previously broadcast global semantic operator. Under the stated clipping and noise assumptions, each privatized semantic-update release satisfies (ϵ, δ)-differential privacy. Experiments on MM-Fi evaluate pose estimation, action recognition, and localization under heterogeneous client discretizations, non-IID data, missing modalities, and matched per-release privacy settings. The proposed operator-space formulation consistently improves performance over federated baselines across these settings. Additional evaluation on RELI11D demonstrates zero-adaptation cross-dataset out-of-distribution transfer with partially overlapping sensor stacks. Ablations further show complementary contributions from the local spectral representation and the learned server-side aggregation, with diffusion-based aggregation improving over simple averaging in the same spectral coordinates.

1 INTRODUCTION

Device-free localization (DFL) infers a person's position, pose, and activity from ambient sensor streams—Wi-Fi channel state information (1; 2; 3; 4), millimeter-wave radar (5; 6), and LiDAR or depth point clouds (7; 8)—without requiring users to wear any device. Fusing these modalities is attractive because each sensor compensates for the others' blind spots (9; 10; 11), but it raises a problem that is fundamentally not centralized: the raw streams reveal where a person is, what they are doing, and what their environment looks like, and they are too bandwidth-heavy to ship to a central server. Federated learning (FL) (12) is the natural framework, but standard FL does not match three properties of multi-modal sensing.

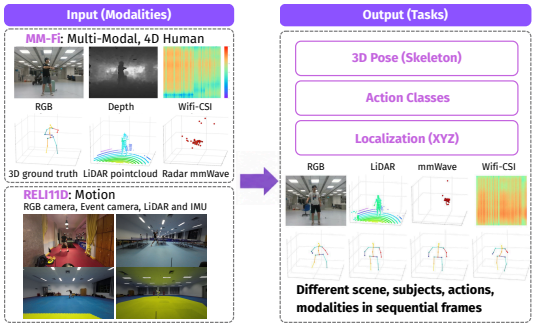


Figure 1: Overview of MM-Fi and RELI11D datasets, input modalities, and target tasks for multi-modal human sensing under heterogeneous scenes, subjects, actions, and sensors.

Clients have different sensors and discretizations. A Wi-Fi client and a LiDAR client produce representations with different token counts,

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

resolutions, and feature structures. Generic FL methods such as FedAvg (12), FedProx (13), and SCAFFOLD (14) aggregate updates in shared parameter coordinates, while multimodal sensing additionally requires a common semantic representation across heterogeneous local discretizations.

Client distributions drift. Different rooms, body sizes, and sensor placements shift the joint distribution of inputs and labels. Optimization-based FL methods (13; 14; 15; 16; 17) address heterogeneous optimization, while recent generative FL approaches (18; 19; 20) learn server-side distributions over client representations or updates. For multimodal sensing, aggregation additionally requires shared coordinates across heterogeneous local discretizations.

Privacy–utility trade-off. Device-free sensing exposes privacy-sensitive information about human appearance, activity, and location. Standard DP mechanisms (21; 22; 23; 24) clip updates and add isotropic Gaussian noise, which can strongly perturb dominant semantic directions. We therefore study subspace-structured privatization of the federated semantic update.

Our approach. We formulate multimodal federation over a shared semantic operator representation. A Spectral-Gated Linear Transformer (SGLT) constructs $\mathbf{M} = \phi(\mathbf{K})^\top \mathbf{V} \in \mathbb{R}^{d\times d}$, whose dimension is independent of the local token count, and filters its spectrum to retain dominant shared structure. The server projects semantic updates into low-dimensional spectral coordinates and aggregates them using Diffusion-Based Neural Operator Aggregation (Diff-NOA). Subspace-Projected Differential Privacy (SP-DP) privatizes each released semantic update using a basis derived from the previously broadcast global semantic operator, while the remaining parameters are aggregated with FedAvg.

Our contributions are:

- **Spectral-Gated Linear Transformer (SGLT).** A linear-attention encoder with spectral gating for heterogeneous multimodal representations whose semantic matrix has token-count-independent dimension.
- **Diffusion-Based Neural Operator Aggregation (Diff-NOA).** A server-side aggregator that learns the distribution of client semantic updates in shared low-dimensional spectral coordinates.
- **Subspace-Projected Differential Privacy (SP-DP).** A subspace-aware Gaussian mechanism providing (ϵ, δ)-DP for each privatized semantic-update release under the stated clipping and noise assumptions.

Experiments on MM-Fi (9) evaluate pose estimation, action recognition, and localization under discretization heterogeneity, non-IID data, missing modalities, and matched per-release privacy settings. Evaluation on RELI11D (8) further studies zero-adaptation cross-dataset OOD transfer with partially overlapping sensor stacks.

2 Related Work

Graph learning for heterogeneous signals. Unifying multi-modal sensors (LiDAR, Wi-Fi CSI, mmWave) is fundamentally about reconciling *discretization heterogeneity*. Voxelization and multi-view projection discard fine topology; point-based GNNs such as Point-GNN (25) and Graph Attention Convolution (26) build k-NN graphs to capture local geometry, but assume homogeneous node features and fail to align modalities of differing densities and noise. Graph Transformers (27; 28) model long-range dependencies but scale quadratically and amplify high-frequency sensor noise. Linear-attention variants (Performer (29), Nyströmformer (30), Linformer (31)) reduce cost but lack *spectral filtering* inductive biases—they treat all spectral components equally and cannot separate the shared low-rank semantic subspace (human pose) from modality-specific high-frequency artifacts (sensor noise), motivating an explicit spectral gate. Recent graph generative work further underscores the need to handle variable-size/discrete structures (continuous-time discrete-state graph diffusion (32), diffusion-free next-scale generation (33)).

Federated learning under non-IID drift. Optimization-based FL methods such as FedProx (13) and SCAFFOLD (14) regularize local optimization under client heterogeneity. Generative FL methods (18; 19; 20) learn server-side distributions for alignment or synthesis. Neural operators (34; 35) provide representations of operators across discretizations. Our formulation couples a shared semantic representation with learned aggregation of its spectral coordinates.

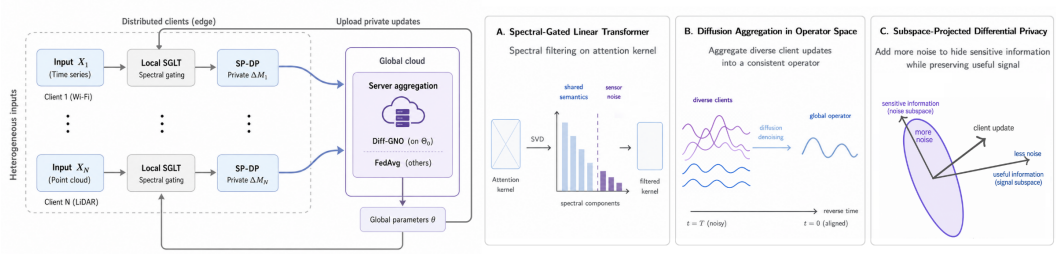


Figure 2: Architecture and mechanisms of our method. **Left:** end-to-end pipeline. Heterogeneous clients (Wi-Fi time series, LiDAR point clouds, etc.) apply SGLT to produce kernel updates $\Delta\mathbf{M}_k$, privatize them via SP-DP, and upload to the server, which aggregates with Diff-NOA on $\mathbf{\Theta}_0 = \mathrm{vec}(\Delta\tilde{\mathbf{M}})$ and FedAvg on the rest, then broadcasts θ. **Right:** *(A) SGLT* retains low-rank shared semantics (blue bars) and suppresses high-frequency sensor noise (purple bars) on the attention kernel. *(B) Diff-NOA* learns and samples a consensus update in the shared spectral coordinates via reverse-time diffusion. *(C) SP-DP* adds more noise in the null (sensitive) subspace and less in the signal (useful) subspace, preserving utility under (ϵ, δ)-DP.

Privacy–utility trade-offs. DP-SGD (21) clips gradients and injects *isotropic* Gaussian noise. In geometric deep learning this is harmful: the structural utility of a graph (e.g., localization geometry) concentrates in the *dominant eigenspace*, so uniform perturbation yields a poor privacy-utility frontier. Topology-aware mechanisms (36) adjust noise by edge existence or node degree but target structural anonymity (hiding links) rather than the *semantic utility* of node features. Missing is a mechanism that uses *subspace projection* to confine noise to the signal's null space, masking sensitive updates without destroying the low-rank geometry essential for localization and pose estimation.

Structure-aware federated learning. Graph federated learning models inter-client relations explicitly through edges or learned similarities (36; 37; 38; 39), with FedGraphNN (40) and OpenFGL (41) providing representative benchmarks. We include graph-aware FL methods as structure-aware federation baselines. Our federated semantic block has fixed dimension $\mathbf{M} \in \mathbb{R}^{d\times d}$ across local token counts and is aggregated through shared spectral coordinates.

3 Method

Overview. Let $\mathcal{C} = \{1, \ldots, N\}$ denote the clients, with local datasets $\mathcal{D}_k = \{(\mathbf{X}_k^{(i)}, \mathbf{Y}_k^{(i)})\}_{i=1}^{n_k}$. Clients may differ in sensor modality, resolution, and token count. A local encoder maps $\mathbf{X}_k$ to $\mathbf{H}_k \in \mathbb{R}^{L_k\times d}$, while linear attention contracts the token dimension into the semantic matrix $\mathbf{M}_k = \phi(\mathbf{K}_k)^\top \mathbf{V}_k \in \mathbb{R}^{d\times d}$. Thus, the federated semantic representation has the same dimension across different L_k.

We optimize the standard federated objective

$$\theta^* = \arg\min_\theta \; \mathbb{E}_{k\sim\mathcal{P}} \left[\mathcal{L}_k(\theta; \mathcal{D}_k)\right]. \tag{1}$$

Our framework comprises SGLT for local spectral representation, SP-DP for privatizing semantic-update releases, and Diff-NOA for server-side aggregation of their spectral coordinates.

3.1 Spectral-Gated Linear Transformer (SGLT)

To reconcile multi-modal heterogeneity, we define a local encoder Φ_k that maps discrete inputs $\mathbf{X}_k$ to a latent sequence $\mathbf{H}_k \in \mathbb{R}^{L\times d}$. Standard self-attention suffers from quadratic complexity and overfits to high-frequency sensor noise. We propose *SGLT*, which incorporates a low-rank spectral filtering mechanism.

Linearized Attention via Kernel Approximation. Let $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{L\times d}$ be the query, key, and value matrices. We utilize a linearized attention mechanism via a feature map:

$$\mathrm{SGLT}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \phi(\mathbf{Q}) \left(\phi(\mathbf{K})^\top \mathbf{V}\right). \tag{2}$$

The term $\mathbf{M} = \phi(\mathbf{K})^\top \mathbf{V} \in \mathbb{R}^{d\times d}$ encapsulates the global semantic structure. In multi-modal settings, modality-specific noise (e.g., Wi-Fi multipath, LiDAR outliers) manifests as high-frequency perturbations in $\mathbf{M}$.

Spectral Gating Mechanism. We compute the SVD

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{W}^\top. \tag{3}$$

During training, we use a differentiable relaxation

$$g(\sigma_i) \;=\; \mathrm{sigmoid}\!\left(\frac{\sigma_i - \tau}{\beta}\right), g(\mathbf{\Sigma}) := \mathrm{diag}(g(\sigma_1), \ldots, g(\sigma_d)), \tag{4}$$

and reconstruct

$$\hat{\mathbf{M}} \;=\; \mathbf{U}\big(g(\mathbf{\Sigma}) \odot \mathbf{\Sigma}\big)\mathbf{W}^\top. \tag{5}$$

For analysis and ablations, we also consider the hard threshold gate $g(\sigma_i) = \mathbb{I}(\sigma_i > \tau)$, which recovers a truncated SVD.

Output. We apply the filtered semantic matrix $\hat{\mathbf{M}}$ with the standard linear-attention normalization:

$$\mathbf{H}' \;=\; \frac{\phi(\mathbf{Q})\,\hat{\mathbf{M}}}{\phi(\mathbf{Q})\,(\phi(\mathbf{K})^\top \mathbf{1}) + \varepsilon}, \tag{6}$$

where $\mathbf{1} \in \mathbb{R}^L$ is an all-ones vector and $\varepsilon > 0$ ensures stability.

3.2 Subspace-Projected Differential Privacy (SP-DP)

To address the privacy-utility trade-off while ensuring rigorous guarantees, we propose *Safe SP-DP*. Unlike standard DP mechanisms that perturb all parameters isotropically, we decouple the subspace estimation from the private gradient update to avoid privacy leakage through the projection basis.

Safe Subspace Projection. Instead of using the client's private local sensitivity basis, we leverage the *global* semantic operator from the previous round, θ_M^t, to define the projection manifold. Let $\theta_M^t = \mathbf{U}_g\mathbf{\Sigma}_g\mathbf{W}_g^\top$ be the SVD of the current global semantic operator. We define the public signal projector $\mathbf{P}_\mathcal{S} = \mathbf{U}_{g,1:r}\mathbf{U}_{g,1:r}^\top$ based on the top-r dominant eigenspace.

The client's update $\Delta\mathbf{M}_k$ is privatized in three steps. First, let $\Delta\mathbf{m} := \mathrm{vec}(\Delta\mathbf{M}_k) \in \mathbb{R}^{d^2}$. We bound the update sensitivity via clipping:

$$\Delta\bar{\mathbf{m}} \;=\; \Delta\mathbf{m} \cdot \min\left\{1, \frac{C}{\|\Delta\mathbf{m}\|_2}\right\}, \tag{7}$$

where C is the clipping threshold. Second, we inject anisotropic Gaussian noise projected onto the global signal/null subspaces:

$$\Delta\tilde{\mathbf{m}} = \Delta\bar{\mathbf{m}} + (\mathbf{P}_\mathcal{S} \otimes \mathbf{I})\mathbf{z}_{\mathrm{sig}} + (\mathbf{I} - \mathbf{P}_\mathcal{S} \otimes \mathbf{I})\mathbf{z}_{\mathrm{null}}, \tag{8}$$

where $\mathbf{z}_{\mathrm{sig}} \sim \mathcal{N}(0, \sigma_{\mathrm{sig}}^2\mathbf{I})$ and $\mathbf{z}_{\mathrm{null}} \sim \mathcal{N}(0, \sigma_{\mathrm{null}}^2\mathbf{I})$ are independent noise vectors. Finally, to ensure (ϵ, δ)-differential privacy, we calibrate the base noise scale σ_{sig} according to the Gaussian mechanism:

$$\sigma_{\mathrm{sig}} \;\geq\; \frac{C\sqrt{2\ln(1.25/\delta)}}{\epsilon}. \tag{9}$$

Since $\mathbf{P}_\mathcal{S}$ is derived from the previously broadcast global semantic operator and $\sigma_{\mathrm{null}} \geq \sigma_{\mathrm{sig}}$, each release of the privatized semantic update $\Delta\tilde{\mathbf{M}}_k$ satisfies the stated (ϵ, δ) guarantee. The reported ϵ therefore denotes a per-release privacy parameter for the semantic update; composition across communication rounds and protection of the separately transmitted $\Delta\theta_{\setminus M}^{(k)}$ are discussed in Appendix C.2.

3.3 Diffusion-Based Neural Operator Aggregation

We model the distribution of client semantic updates in a shared spectral coordinate system. Given the global semantic basis, each privatized update $\Delta\tilde{\mathbf{M}}_k$ is projected to an $r \times r$ coefficient matrix and vectorized into $\mathbf{\Theta}_0^{(k)} \in \mathbb{R}^{r^2}$. A diffusion score model learns the distribution of these client updates and produces a consensus spectral update through reverse-time sampling.

Spectral Projection. Given the privatized update $\Delta\tilde{\mathbf{M}}_k$, we first project it onto the global basis $\mathbf{U}_g$ and $\mathbf{W}_g$ (computed from the global operator θ_M^t) to obtain the spectral coefficient matrix $\mathbf{\Sigma}^{(k)} \in \mathbb{R}^{r\times r}$:

$$\mathbf{\Sigma}^{(k)} = \mathbf{U}_{g,1:r}^\top(\Delta\tilde{\mathbf{M}}_k)\mathbf{W}_{g,1:r}. \tag{10}$$

We define the diffusion state as the flattened spectral coefficients $\mathbf{\Theta}_0^{(k)} := \mathrm{vec}(\mathbf{\Sigma}^{(k)}) \in \mathbb{R}^{r^2}$. By choosing $r \ll d$ (e.g., $r = 16, d = 256$), the dimensionality is reduced by orders of magnitude ($65,536 \to 256$), allowing the score model to capture the complex distribution of client updates even with limited samples.

Forward VE-SDE. We adopt a variance-exploding diffusion on the operator-update space:

$$d\mathbf{\Theta}_t = g(t)\, d\mathbf{W}_t, \qquad t \in [0, 1], \tag{11}$$

which yields $\mathbf{\Theta}_t = \mathbf{\Theta}_0 + \sigma(t)\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ and $\sigma^2(t) = \int_0^t g^2(\tau)\, d\tau$.

Denoising Score Matching (DSM). We train a score model $s_\psi(\mathbf{\Theta}_t, t) \approx \nabla_{\mathbf{\Theta}_t} \log p_t(\mathbf{\Theta}_t)$ via

$$\mathcal{L}_{\mathrm{DSM}}(\psi) = \mathbb{E}_{t, \mathbf{\Theta}_0, \boldsymbol{\epsilon}} \left[\left\| s_\psi(\mathbf{\Theta}_t, t) + \frac{\boldsymbol{\epsilon}}{\sigma(t)} \right\|_2^2 \right], \mathbf{\Theta}_t = \mathbf{\Theta}_0 + \sigma(t)\boldsymbol{\epsilon}. \tag{12}$$

Reverse-time SDE (aggregation). Given s_ψ, the reverse-time dynamics are

$$d\mathbf{\Theta}_t = -g(t)^2 s_\psi(\mathbf{\Theta}_t, t)\, dt + g(t)\, d\bar{\mathbf{W}}_t. \tag{13}$$

The sampled spectral state is reconstructed into the semantic block as

$$\Delta\mathbf{M}^* = \mathbf{U}_{g,1:r} \mathrm{unvec}(\mathbf{\Theta}^*) \mathbf{W}_{g,1:r}^\top.$$

The server applies this update to θ_M and aggregates $\theta_{\setminus M}$ with FedAvg.

Summary. Alg. 1 summarizes the end-to-end training loop: each round, selected clients run SGLT-based local training to produce a semantic-kernel update $\Delta\mathbf{M}_k$ and the remaining parameter updates, privatize $\Delta\mathbf{M}_k$ via SP-DP (Alg. 2), and the server performs Diff-NOA on the privatized semantic updates while aggregating the remaining parameters by FedAvg.

Algorithm 1 Our method

Input: clients $\mathcal{C}$; rounds T; steps E; clip C; privacy (ϵ, δ); rank r; VE schedule $g(t)$.
Init: parameters $\theta^0 = (\theta^0_{\setminus M}, \theta^0_M)$, score network ψ^0.
for $t = 0, \ldots, T-1$ **do**
  Global Basis: Server computes SVD $\theta^t_M \approx \mathbf{U}_g \mathbf{\Sigma}_g \mathbf{W}_g^\top$; samples $\mathcal{S}_t$ and broadcasts $\theta^t, \mathbf{U}_g, \mathbf{W}_g$.
  for all $k \in \mathcal{S}_t$ **in parallel do**
    $\Delta\mathbf{M}_k, \Delta\theta^{(k)}_{\setminus M} \leftarrow \mathrm{LocalTrain}(\theta^t, E)$.
    $\Delta\tilde{\mathbf{M}}_k \leftarrow \text{SP-DP}(\Delta\mathbf{M}_k, \mathbf{U}_g; C, \epsilon, \delta)$.
    Upload $\Delta\tilde{\mathbf{M}}_k$ and $\Delta\theta^{(k)}_{\setminus M}$.
  end for
  Spectral Projection: $\mathbf{\Theta}_0^{(k)} \leftarrow \mathrm{vec}(\mathbf{U}_{g,1:r}^\top \Delta\tilde{\mathbf{M}}_k \mathbf{W}_{g,1:r})$.
  Diff-NOA: Update s_ψ on $\{\mathbf{\Theta}_0^{(k)}\}$ and sample $\mathbf{\Theta}^*$ via reverse VE-SDE.
  Update: $\theta^{t+1}_M \leftarrow \theta^t_M + \eta_s \mathbf{U}_{g,1:r} \mathrm{unvec}(\mathbf{\Theta}^*) \mathbf{W}_{g,1:r}^\top$.
  $\theta^{t+1}_{\setminus M} \leftarrow \theta^t_{\setminus M} + \eta_s \mathrm{FedAvg}(\{\Delta\theta^{(k)}_{\setminus M}\})$.
end for
Return: θ^T.

4 Experiments

4.1 Experimental Setup

Datasets. *MM-Fi* (9) is a large-scale multimodal human sensing benchmark with synchronized heterogeneous streams (e.g., Wi-Fi CSI and depth/radar/LiDAR-like modalities). We follow the standard protocol of treating each sample as a multimodal observation and construct graph/token inputs G by modality-specific discretization, explicitly inducing varying graph sizes and feature topologies across modality configurations. To validate fidelity of our reproduction, we cross-checked Wi-Fi CSI unimodal pose estimation against the original MM-Fi benchmark (9) and HPE-Li, reproducing their reported MPJPE within ± 2.5 mm under matched splits (Appendix A.1). *RELI11D* (8) contains RGB, event-camera, LiDAR, and IMU streams across diverse scenes. Its sensor stack partially overlaps with MM-Fi through RGB and LiDAR while differing in dataset, scenes, subjects, and sensing configuration. We train on MM-Fi and evaluate through the implemented RELI11D input pipeline without target-model adaptation. We use this experiment as a zero-adaptation cross-dataset OOD transfer test with partially overlapping sensor stacks.

Federated simulation. We simulate $K = 100$ clients. Each client receives a private subset of samples and a client-specific modality set $\mathcal{M}_k$. To isolate different failure modes:

- *Non-IID labels:* we partition by a Dirichlet distribution over labels with concentration α, where smaller α yields stronger statistical heterogeneity.
- *Discretization heterogeneity:* clients use different modality/resolution configurations. Concretely, we instantiate three client types: *Type A* (coarser / sparse modality discretization), *Type B* (denser / richer modality discretization), *Type C* (mixed modalities with occasional missing streams).

Tasks and metrics. We consider three representative targets consistent with device-free sensing: (1) **3D pose estimation** measured by MPJPE (mm, lower is better), (2) **action recognition** measured by Top-1 accuracy (%,

higher is better), (3) **localization** measured by RMSE (cm, lower is better). For privacy, we report performance under multiple (ϵ, δ) budgets and keep δ fixed.

Training protocol. We run up to $T = 1000$ communication rounds with $K = 100$ simulated clients and a fixed client sampling rate $q = 0.4$. For efficiency experiments (Tab. 3), we report the number of rounds needed to reach a fixed target (Loc RMSE ≤ 12.5) with early stopping. Each selected client performs $E = 100$ local gradient steps with batch size $B = 64$ using AdamW (lr $\eta = 10^{-3}$, weight decay 10^{-4}). Unless stated otherwise, we use server step size $\eta_s = 1.0$ (FedAvg-style averaging for non-SGLT blocks; Diff-NOA for the SGLT block). We report the mean over three random seeds.

Models. SGLT uses hidden dimension $d = 256$ with 8 heads. For linear attention, we instantiate $\phi(\cdot)$ with FAVOR+ random features (positive random features for softmax-kernel approximation) using $m = 256$ random features per head, and apply the standard normalization in Eq. (6). We fix the soft gate parameters to $(\tau_s, \beta) = (0.05, 0.01)$ across all experiments; *Table 6* ablates a hard-gate variant by sweeping τ_h while disabling the soft gate.

Diffusion aggregation (VE-SDE). Diff-NOA follows the VE-SDE in Eq. (11)–(13) with the standard exponential noise schedule:

$$\sigma(t) = \sigma_{\min} \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right)^t, \quad \sigma_{\min} = 0.01,\ \sigma_{\max} = 50, \quad t \in [0, 1]. \tag{14}$$

This yields the closed-form diffusion coefficient

$$g(t) = \sqrt{\frac{d}{dt}\sigma^2(t)} = \sigma(t)\sqrt{2 \log \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right)}. \tag{15}$$

For reverse sampling of Eq. (13), we use Euler–Maruyama with $N_{\text{rev}} = 50$ discretization steps uniformly spaced on $t \in [1, 0]$. The score network $s_\psi(\mathbf{\Theta}, t)$ is a 2-layer MLP (width 512) with sinusoidal time embedding, trained by DSM (Eq. (12)) on $\{\mathbf{\Theta}_0^{(k)}\}_{k \in \mathcal{S}_t}$ each round.

Differential privacy. We fix $\delta = 10^{-5}$ and sweep $\epsilon \in \{0.5, 1, 2, 4, 8\}$. SP-DP clips the vectorized semantic update with norm bound $C = 1.0$ (Eq. (7)) and sets σ_{sig} by Eq. (9). We allocate stronger noise to the null subspace using $\sigma_{\text{null}} = \kappa \sigma_{\text{sig}}$ with $\kappa = 4$.

Algorithm 2 Safe SP-DP with Global Basis Anchor

1: **Input:** Update $\Delta\mathbf{M}$; global basis $\mathbf{U}_g$; clip C; privacy (ϵ, δ); rank r; $\sigma_{\text{null}} \geq \sigma_{\text{sig}}$.
2: $\Delta\mathbf{m} \leftarrow \text{vec}(\Delta\mathbf{M}) \in \mathbb{R}^{d^2}$.
3: $\Delta\bar{\mathbf{m}} \leftarrow \Delta\mathbf{m} \min\{1, C/\|\Delta\mathbf{m}\|_2\}$.
4: **Projectors:** $\mathbf{P}_{\mathcal{S}} \leftarrow \mathbf{U}_{g,1:r}\mathbf{U}_{g,1:r}^\top$; $\mathbf{P}_{\mathcal{S}^\perp} \leftarrow \mathbf{I} - \mathbf{P}_{\mathcal{S}}$.
5: $\sigma_{\text{sig}} \leftarrow \frac{C\sqrt{2\ln(1.25/\delta)}}{\epsilon}$.
6: Sample $\mathbf{z}_{\text{sig}} \sim \mathcal{N}(0, \sigma_{\text{sig}}^2\mathbf{I})$, $\mathbf{z}_{\text{null}} \sim \mathcal{N}(0, \sigma_{\text{null}}^2\mathbf{I})$.
7: $\Delta\tilde{\mathbf{m}} \leftarrow \Delta\bar{\mathbf{m}} + (\mathbf{P}_{\mathcal{S}} \otimes \mathbf{I})\mathbf{z}_{\text{sig}} + (\mathbf{P}_{\mathcal{S}^\perp} \otimes \mathbf{I})\mathbf{z}_{\text{null}}$.
8: $\Delta\tilde{\mathbf{M}} \leftarrow \text{unvec}(\Delta\tilde{\mathbf{m}})$.
9: **Return:** $\Delta\tilde{\mathbf{M}}$.

Differential privacy. We fix $\delta = 10^{-5}$ and sweep $\epsilon \in \{0.5, 1, 2, 4, 8\}$. SP-DP clips the vectorized kernel update with norm bound $C = 1.0$ (Eq. (7)) and sets σ_{sig} by Eq. (9). We allocate stronger noise to the null subspace by $\sigma_{\text{null}} = \kappa\, \sigma_{\text{sig}}$ with $\kappa = 4$ (Eq. (8)).

Baselines. We report an upper bound **Centralized (oracle)** trained on the union of all client data with full modality access (when available) and no privacy constraints. We additionally compare against **Local** training (no aggregation) and a broad set of federated baselines covering optimization under heterogeneity, drift correction, normalization/personalization, and graph-aware FL: *FedAvg* (12), *FedProx* (13), *SCAFFOLD* (14), *FedNova* (42), *FedAdam* (FedOpt) (15), *FedDyn* (16), *MOON* (17), *FedBN* (43), *FedLC* (44), *FedRep* (45), *FedGraphNN* (40), and *GCFL+* (46). For attention/fusion, we include *Performer (FAVOR+ linear attention)* (29) as a canonical linear-attention baseline and two recent sequence models as fusion backbones: *xLSTM (state-space recurrent fusion)* (47) and *Gated DeltaNet (Delta-rule fusion)* (48). For differential privacy, we include *DP-FedAvg* using isotropic Gaussian noise with standard accounting (22; 23), and additionally compare against DP baselines with improved clipping/spending: *DP-FedAvg + Adaptive Clipping* (24), *AGC-DP (Adaptive Gaussian Clipping)* (49), and *TA-DPFL(Time-adaptive privacy spending)* (50). Finally, we include three recent baselines targeting non-IID heterogeneity and/or federated graph learning: *CA2FL* (51), *HPFL* (52), and *Decoupled Subgraph FL* (53). All federated methods use the same client sampling rate, local steps, and total rounds as our method unless otherwise stated. For modality-coverage fairness, we additionally include unimodal sensing baselines applied to their native modality channel: *MetaFi++* (2) for Wi-Fi CSI pose estimation and *PointNet++* (54) for LiDAR-based localization, each trained federatively under the same protocol as our method. We further compare against *X-Fi* (55) as a recent multi-modal foundation baseline integrated with FedAvg; we evaluate X-Fi under three modality configurations (2-modality / 3-modality / full) to disentangle its dependence on rich modality access. Detailed configurations and per-modality results are reported in Appendix A.2. For architectural fairness, all transformer-based baselines (LAT and our method variants) use the same $(d, \#\text{heads}, m)$ and the same

Table 1: Main results on MM-Fi with $N = 100$ clients (mean±std over 3 seeds). Metrics: 3D pose estimation (MPJPE, mm)↓, action recognition (Top-1 Acc, %)↑, localization (RMSE, cm)↓. Privacy baselines and *our method* use the same privacy budget (e.g., $\epsilon = 2.0, \delta = 10^{-5}$). *Centralized* is an oracle upper bound and excluded from best/second-best marking.

Method	Pose ↓	Action ↑	Loc ↓
Centralized (oracle)	62.4±0.6	93.8±0.4	8.2±0.1
Local (no aggregation)	118.3±1.4	78.5±0.8	18.5±0.3
Optimization under heterogeneity			
FedAvg	96.7±1.0	85.2±0.6	13.9±0.2
FedProx	94.8±0.9	85.7±0.8	13.6±0.6
SCAFFOLD	91.9±0.8	86.5±0.5	12.9±0.1
FedAdam (FedOpt)	90.4±0.8	86.9±0.5	12.6±0.5
FedNova	93.2±0.9	86.1±0.7	13.2±0.4
FedDyn	89.7±0.7	87.3±0.5	12.4±0.2
MOON	88.9±0.7	87.6±0.9	12.2±0.4
Normalization			
FedBN	92.1±0.8	86.4±0.3	13.0±0.3
FedLC	90.9±0.8	86.9±0.1	12.7±0.2
FedRep	89.6±0.7	87.2±0.6	12.3±0.8
Graph-aware FL			
FedGCN (FedAvg)	94.3±0.9	86.0±0.3	13.4±1.4
FedGAT (FedAvg)	92.6±0.8	86.6±0.7	13.0±0.8
FedGraphNN	91.8±0.8	86.8±0.4	12.8±0.6
GCFL+	90.6±0.8	87.1±0.6	12.5±0.7
Attention / fusion			
LAT (linear attn; no gate)	89.8±1.5	87.4±1.4	12.2±1.1
Performer	90.2±0.8	87.2±1.3	12.3±1.3
xLSTM	91.1±1.2	86.8±0.4	12.6±0.9
Gated DeltaNet	89.3±0.7	87.6±0.5	12.1±0.7
Privacy			
DP-FedAvg (isotropic)	104.5±1.2	82.1±1.5	15.8±1.2
DP-FedAvg + AdaptClip	101.3±0.7	82.9±0.8	15.1±1.6
AGC-DP	99.9±1.1	83.4±1.7	14.7±1.3
TA-DPFL	98.8±1.3	83.9±0.9	14.3±0.8
ablations			
w/o SGLT (LAT)	89.8±0.7	87.4±0.5	12.2±0.3
w/o Diff-NOA (FedAvg)	92.5±0.8	86.3±0.4	12.9±0.2
w/o SP-DP (no privacy)	**84.1±0.5**	**90.4±0.6**	**9.8±0.7**
Our method	86.0±0.6	89.6±0.4	10.5±0.2

Table 2: Zero-adaptation cross-dataset OOD transfer from *MM-Fi* to *RELI11D* with partially overlapping sensor stacks (mean±std over 3 seeds). Metric: localization RMSE (cm)↓.

Method	Source	Target	Gap (Δ)
Basic baselines			
FedAvg	13.9±0.2	48.5±1.6	34.6±0.5
FedProx	13.6±0.6	46.2±1.0	32.6±0.3
Non-private baselines			
FedDyn	12.4±0.2	41.0±1.4	28.6±1.3
FedRep	12.3±0.8	39.2±1.1	26.9±1.4
MOON	12.2±0.4	38.5±0.8	26.3±0.8
Gated DeltaNet	12.1±0.7	37.6±0.9	25.5±0.7
Domain generalization			
FedBN	13.0±0.3	35.2±1.3	22.2±1.2
FedRod	13.4±1.4	34.1±0.6	20.7±1.1
Privacy (same budget)			
TA-DPFL	14.3±0.8	36.8±1.7	22.5±1.1
Our method	**10.5±0.2**	**24.8±0.8**	**14.3±0.9**

Table 3: Communication efficiency on MM-Fi. We report rounds to achieve Loc RMSE ≤ 12.5 and total communication (downlink + uplink, GB); Diff-NOA affects server computation but not communicated volume.

Method	Rounds ↓	Comm. (GB) ↓
FedAvg	820	205
SCAFFOLD	670	168
FedAdam (FedOpt)	610	153
LAT (linear attn, no gate)	540	135
Our method	**360**	**90**

linear-attention feature map $\phi(\cdot)$; graph-based baselines use the same encoder depth and hidden width as the graph branch in our method.

4.2 Experimental Results

Overview. We evaluate our method under three stressors: (i) discretization heterogeneity (cross-modality/resolution mismatch), (ii) statistical heterogeneity (non-IID client drift), and (iii) privacy constraints under (ϵ, δ)-DP. Unless otherwise stated, all results report mean±std over three random seeds.

Main Results on MM-Fi. Table 1 reports MM-Fi results on pose (MPJPE↓), action (Top-1↑), and localization (RMSE↓). Under the same (ϵ, δ) budget, *our method* attains the best overall performance among non-oracle federated methods: MPJPE 86.0, Top-1 89.6, and RMSE 10.5. Removing either module degrades all three tasks: *w/o SGLT (LAT)* increases RMSE to 12.2, while *w/o Diff-NOA (FedAvg)* increases it to 12.9, supporting distinct contributions from the local spectral representation and server-side aggregation.

Transfer to RELI11D for OOD Generalization. We train on MM-Fi and evaluate on RELI11D without target-model adaptation. *Table 2* shows a target RMSE of 24.8 and transfer gap of 14.3, compared with $48.5/34.6$ for FedAvg, $35.2/22.2$ for FedBN, and $34.1/20.7$ for FedRod. These results provide cross-dataset OOD evidence under partially overlapping sensor stacks.

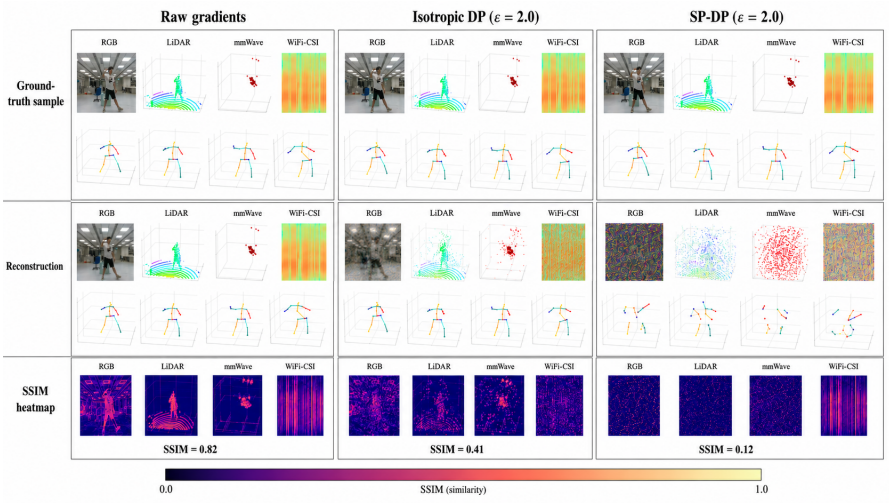


Figure 4: Gradient-inversion reconstruction attack at matched $\epsilon = 2.0$. Representative results compare raw updates, isotropic DP, and SP-DP across RGB, LiDAR, mmWave, and Wi-Fi CSI. Mean SSIM is 0.82, 0.41, and 0.12, respectively (lower indicates worse reconstruction). Full attack settings are provided in Appendix A.8.

Robustness to Discretization Heterogeneity. We vary the fraction of low-resolution / sparse-modality clients. *Figure 5a* shows that performance degrades as mismatch increases for all methods, but *our method* degrades the least, indicating reduced sensitivity to cross-client discretization/topology mismatch.

Robustness to Statistical Heterogeneity (Non-IID). We induce non-IID label skew via a Dirichlet partition with concentration α (smaller α = stronger skew). *Figure 5b* reports localization RMSE: *our method* is consistently lower than FedAvg and non-IID baselines, with the largest margin at small α; gaps shrink as α increases.

Privacy–Utility Trade-off under (ϵ, δ)-DP. We sweep $\epsilon \in \{0.5, 1, 2, 4, 8\}$ with fixed $\delta = 10^{-5}$. *Figure 5c* compares methods at matched per-release privacy settings for the semantic update. SP-DP achieves a better privacy–utility frontier than DP-FedAvg and TA-DPFL across the evaluated budgets, with larger gaps at smaller ϵ.

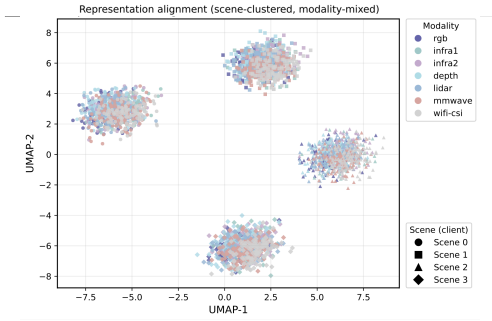


Figure 3: UMAP of spatio-temporal token embeddings on MM-Fi. Colors denote 7 modalities and marker shapes denote 4 scenes (clients). Embeddings form scene-dominant clusters with modality-mixed points inside each cluster, indicating cross-modality alignment under heterogeneous discretizations.

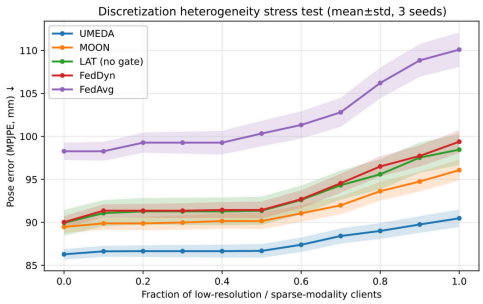
Communication Efficiency. Table 3 reports rounds-to-target (Loc RMSE ≤ 12.5) and total communication (downlink+uplink). *our method* reaches the target in 360 rounds with 90 GB, versus 540 rounds and 135 GB for the second-best LAT.

Robustness to Missing Modalities at Inference. We randomly drop a fraction of modalities at test time. *Figure 5d* shows that RMSE increases with missingness for all methods, but *our method* degrades most gracefully and remains below MOON, FedBN, and FedAvg across the sweep.

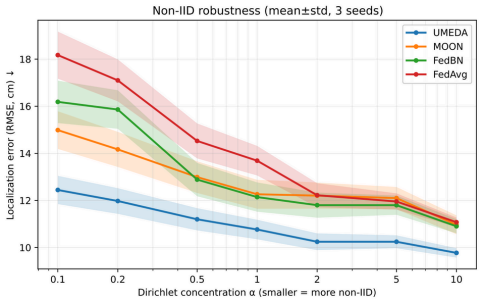
Representation Alignment Visualization. We visualize latent node/token embeddings with UMAP. In *Figure 3*, embeddings forms four scene-dominant clusters (clients), while modalities are largely mixed within each scene cluster, indicating cross-modality alignment under scene-specific geometry/propagation conditions.

Privacy beyond formal DP. Beyond the (ϵ, δ) frontier, we evaluate empirical resistance to gradient-inversion attacks (Appendix A.8, Fig. **??**): under matched $\epsilon = 2.0$, SP-DP yields substantially lower reconstruction similarity than isotropic DP, with mean SSIM decreasing from 0.41 to 0.12. A finer ϵ sweep (Appendix A.7) confirms widening gaps under tight budgets.

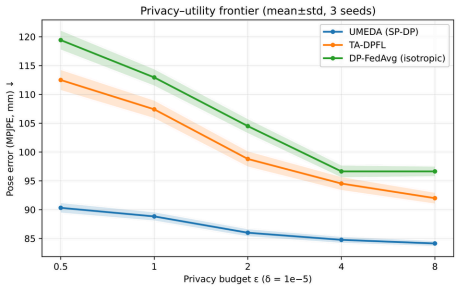
Scalability. Server-side measurements from $K = 100$ to $K = 10{,}000$ clients are reported in Appendix A.15; Diff-NOA operates on a fixed-dimensional spectral state.



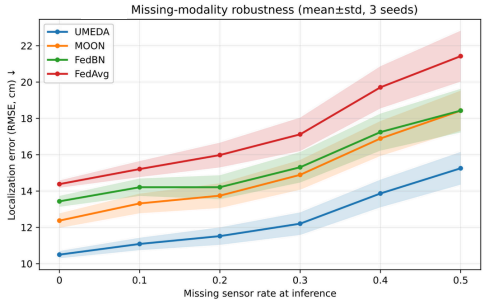
(a) Discretization-heterogeneity stress test on MM-Fi. We increase the fraction of low-resolution / sparse-modality clients across 4 scenes and report MPJPE (mm; ↓). Our method degrades the least as mismatch grows, consistent with SGLT suppressing high-frequency residual subspaces in $\mathbf{M}$.



(b) Non-IID robustness on MM-Fi (mean±std, 3 seeds). Clients are partitioned by Dirichlet concentration α (smaller α = stronger skew); we report Loc RMSE (cm; ↓). Our method is consistently best, consistent with Diff-NOA aggregating operator updates to reduce drift under severe non-IID.



(c) Privacy–utility frontier on MM-Fi under (ϵ, δ)-DP (mean±std, 3 seeds; $\delta = 10^{-5}$). We sweep $\epsilon \in \{0.5, 1, 2, 4, 8\}$ and report MPJPE (mm; ↓). Our method (SP-DP) dominates isotropic DP-FedAvg and TA-DPFL by pushing more noise into $\mathcal{S}^{\perp}$ while preserving signal subspaces of $\mathbf{M}$.



(d) Missing-modality robustness on MM-Fi (mean±std, 3 seeds). We drop a fraction of sensor streams at test time (x-axis) and report Loc RMSE (cm; ↓). Our method degrades most gracefully as missingness increases.

Figure 5: Robustness analyses of our method on MM-Fi (mean±std over 3 seeds). (a) Discretization heterogeneity, (b) statistical heterogeneity (non-IID), (c) privacy-utility trade-off, and (d) missing-modality robustness.

Privacy beyond formal DP. Beyond the (ϵ, δ) frontier, we evaluate empirical resistance to gradient-inversion attacks at matched $\epsilon = 2.0$. Figure 4 shows representative reconstructions: SP-DP yields substantially lower reconstruction similarity than isotropic DP, with mean SSIM decreasing from 0.41 to 0.12. Full attack settings and additional analysis are provided in Appendix A.8. A finer ϵ sweep (Appendix A.7) further evaluates the privacy–utility trade-off under tighter budgets.

Scalability. Server-side measurements from $K = 100$ to $K = 10{,}000$ clients are reported in Appendix A.15; Diff-NOA operates on a fixed-dimensional spectral state.

5 Conclusions

We presented an operator-space federated learning framework for heterogeneous multimodal human sensing. SGLT constructs a shared semantic representation whose dimension is independent of local token count, Diff-NOA aggregates its low-dimensional spectral coordinates, and SP-DP provides (ϵ, δ)-DP for each privatized semantic-update release. Experiments on MM-Fi show improved pose, action, and localization performance under discretization heterogeneity, non-IID data, missing modalities, and matched per-release privacy settings. RELI11D further demonstrates zero-adaptation cross-dataset OOD transfer with partially overlapping sensor stacks. Additional aggregation controls, systems measurements, privacy scope, and limitations are reported in the appendix.

REFERENCES

[1] Wenjun Jiang, Hongfei Xue, Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, and Lu Su. Towards 3d human pose construction using wifi. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, MobiCom '20, New York, NY, USA, 2020. Association for Computing Machinery.

[2] Yunjiao Zhou, He Huang, Shenghai Yuan, Han Zou, Lihua Xie, and Jianfei Yang. Metafi++: Wifi-enabled transformer-based human pose estimation for metaverse avatar simulation. *IEEE Internet of Things Journal*, 10(16):14128–14136, 2023.

[3] Xinping Rao, Zhenzhen Luo, Yong Luo, Yugen Yi, Gang Lei, and Yuanlong Cao. Mffaloc: Csi-based multifeatures fusion adaptive device-free passive indoor fingerprinting localization. *IEEE Internet of Things Journal*, 11(8):14100–14114, 2024.

[4] Mohamed Mohsen, Hamada Rizk, Hirozumi Yamaguchi, and Moustafa Youssef. Locfree: Wifi rtt-based device-free indoor localization system. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Spatial Big Data and AI for Industrial Applications*, GeoIndustry '23, page 32–40, New York, NY, USA, 2023. Association for Computing Machinery.

[5] Anjun Chen, Xiangyu Wang, Shaohao Zhu, Yanxu Li, Jiming Chen, and Qi Ye. mmbody benchmark: 3d body reconstruction dataset and analysis for millimeter wave radar. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3501–3510, 2022.

[6] Shih-Po Lee, Niraj Prakash Kini, Wen-Hsiao Peng, Ching-Wen Ma, and Jenq-Neng Hwang. Hupr: A benchmark for human pose estimation using millimeter wave radar. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5715–5724, January 2023.

[7] Yun-Ting Wang, Chao-Chung Peng, Ankit A. Ravankar, and Abhijeet Ravankar. A single lidar-based feature fusion indoor localization algorithm. *Sensors*, 18(4), 2018.

[8] Ming Yan, Yan Zhang, Shuqiang Cai, Shuqi Fan, Xincheng Lin, Yudi Dai, Siqi Shen, Chenglu Wen, Lan Xu, Yuexin Ma, and Cheng Wang. Reli11d: A comprehensive multimodal human motion dataset and method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2250–2262, June 2024.

[9] Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[10] Sizhe An, Yin Li, and Umit Ogras. mri: Multi-modal 3d human pose estimation dataset using mmwave, rgb-d, and inertial sensors. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27414–27426. Curran Associates, Inc., 2022.

[11] Anjun Chen, Xiangyu Wang, Kun Shi, Shaohao Zhu, Bin Fang, Yingfeng Chen, Jiming Chen, Yuchi Huo, and Qi Ye. Immfusion: Robust mmwave-rgb fusion for 3d human body reconstruction in all weather conditions. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2752–2758. IEEE, 2023.

[12] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 20–22 Apr 2017.

[13] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

[14] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. Scaffold: stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[15] Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021.

[16] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021.

[17] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[18] Shangyin Weng, Lei Zhang, Xiaoshuai Zhang, and Muhammad Ali Imran. Faster convergence on differential privacy-based federated learning. *IEEE Internet of Things Journal*, 11(12):22578–22589, 2024.

[19] Matthijs de Goede, Bart Cox, and Jérémie Decouchant. Training diffusion models with federated learning, 2024.

[20] Haokun Chen, Hang Li, Yao Zhang, Jinhe Bi, Gengyuan Zhang, Yueqi Zhang, Philip Torr, Jindong Gu, Denis Krompass, and Volker Tresp. Fedbip: Heterogeneous one-shot federated learning with personalized latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 30440–30450, 2025.

[21] R. C. Geyer, T. Klein, and M. Nabi. Differentially Private Federated Learning: A Client Level Perspective. *ArXiv e-prints*, December 2017.

[22] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014.

[23] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 308–318, New York, NY, USA, 2016. Association for Computing Machinery.

[24] Galen Andrew, Om Thakkar, H. Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc.

[25] Weijing Shi and Ragunathan (Raj) Rajkumar. Point-gnn: Graph neural network for 3d object detection in a point cloud. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[26] Lei Wang, Yuchun Huang, Yaolin Hou, Shenman Zhang, and Jie Shan. Graph attention convolution for point cloud semantic segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10288–10297, 2019.

[27] Wei Zhou, Qian Wang, Weiwei Jin, Xinzhe Shi, and Ying He. Gtnet: Graph transformer network for 3d point cloud classification and semantic segmentation, 2024.

[28] Damien Robert, Hugo Raguet, and Loic Landrieu. Efficient 3d semantic segmentation with superpoint transformer. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[29] Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021.

[30] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. 2021.

[31] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity, 2020.

[32] Zhe Xu, Ruizhong Qiu, Yuzhong Chen, Huiyuan Chen, Xiran Fan, Menghai Pan, Zhichen Zeng, Mahashweta Das, and Hanghang Tong. Discrete-state continuous-time diffusion for graph generation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc.

[33] Samuel Belkadi, Dat Minh Hong, Marian Chen, Miruna Cretu, Charles Harris, and Pietro Lio. Diffusion-free graph generation with next-scale prediction. In *ICML 2025 Generative AI and Biology (GenBio) Workshop*, 2025.

[34] Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: learning maps between function spaces with applications to pdes. *J. Mach. Learn. Res.*, 24(1), January 2023.

[35] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Multipole graph neural operator for parametric partial differential equations. In *Advances in Neural Information Processing Systems*, 2020.

[36] Xingbo Fu, Binchi Zhang, Yushun Dong, Chen Chen, and Jundong Li. Federated graph machine learning: A survey of concepts, techniques, and applications. *SIGKDD Explor. Newsl.*, 24(2):32–47, December 2022.

[37] Rui Liu, Pengwei Xing, Zichao Deng, Anran Li, Cuntai Guan, and Han Yu. Federated graph neural networks: Overview, techniques, and challenges. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3):4279–4295, 2025.

[38] Xunkai Li, Zhengyu Wu, Wentao Zhang, Yinlin Zhu, Rong-Hua Li, and Guoren Wang. Fedgta: Topology-aware averaging for federated graph learning. *Proc. VLDB Endow.*, 17(1):41–50, September 2023.

[39] Yinlin Zhu, Xunkai Li, Zhengyu Wu, Di Wu, Miao Hu, and Rong-Hua Li. Fedtad: topology-aware data-free knowledge distillation for subgraph federated learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024.

[40] Chaoyang He, Keshav Balasubramanian, Emir Ceyani, Carl Yang, Han Xie, Lichao Sun, Lifang He, Liangwei Yang, Philip S. Yu, Yu Rong, Peilin Zhao, Junzhou Huang, Murali Annavaram, and Salman Avestimehr. Fedgraphnn: A federated learning system and benchmark for graph neural networks, 2021.

[41] Xunkai Li, Yinlin Zhu, Boyang Pang, Guochen Yan, Yeyu Yan, Zening Li, Zhengyu Wu, Wentao Zhang, Rong-Hua Li, and Guoren Wang. Openfgl: A comprehensive benchmarks for federated graph learning, 2024.

[42] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[43] Xiaoxiao Li, Meirui JIANG, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated learning on non-IID features via local batch normalization. In *International Conference on Learning Representations*, 2021.

[44] Jiacheng Wang, Yueming Jin, and Liansheng Wang. Personalizing federated medical image segmentation via local calibration. In *European Conference on Computer Vision*, pages 456–472. Springer, 2022.

[45] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2089–2099. PMLR, 18–24 Jul 2021.

[46] Han Xie, Jing Ma, Li Xiong, and Carl Yang. Federated graph classification over non-iid graphs. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc.

[47] Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael K Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xLSTM: Extended long short-term memory. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[48] Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. In *The Thirteenth International Conference on Learning Representations*, 2025.

[49] Muhammad Ayat Hidayat, Yugo Nakamura, Billy Dawton, and Yutaka Arakawa. Agc-dp: Differential privacy with adaptive gaussian clipping for federated learning. In *2023 24th IEEE International Conference on Mobile Data Management (MDM)*, pages 199–208, 2023.

[50] Shahrzad Kiani, Nupur Kulkarni, Adam Dziedzic, Stark Draper, and Franziska Boenisch. Differentially private federated learning with time-adaptive privacy spending, 2025.

[51] Yujia Wang, Yuanpu Cao, Jingcheng Wu, Ruoyu Chen, and Jinghui Chen. Tackling the data heterogeneity in asynchronous federated learning with cached update calibration. In *The Twelfth International Conference on Learning Representations*, 2024.

[52] Lei Shen, Zhenheng Tang, Lijun Wu, Yonggang Zhang, Xiaowen Chu, Tao Qin, and Bo Han. Hot-pluggable federated learning: Bridging general and personalized FL via dynamic selection. In *The Thirteenth International Conference on Learning Representations*, 2025.

[53] Javad Aliakbari, Johan Östman, and Alexandre Graell i Amat. Decoupled subgraph federated learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

[54] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: deep hierarchical feature learning on point sets in a metric space. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5105–5114, Red Hook, NY, USA, 2017. Curran Associates Inc.

[55] Xinyan Chen and Jianfei Yang. X-fi: A modality-invariant foundation model for multimodal human sensing. In *The Thirteenth International Conference on Learning Representations*, 2025.

[56] Toan D. Gian, Tien Dac Lai, Thien Van Luong, Kok-Seng Wong, and Van-Dinh Nguyen. Hpe-li: Wifi-enabled lightweight dual selective kernel convolution for human pose estimation. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 93–111, Cham, 2025. Springer Nature Switzerland.

[57] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients - how easy is it to break privacy in federated learning? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16937–16947. Curran Associates, Inc., 2020.

[58] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. idlg: Improved deep leakage from gradients, 2020.

[59] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *Trans. Img. Proc.*, 13(4):600–612, April 2004.

[60] K. A. Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for federated learning on user-held data. In *NIPS Workshop on Private Multi-Party Machine Learning*, 2016.

TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

A ADDITIONAL EXPERIMENTS

This appendix provides controlled ablations that isolate the contribution of each component in *our method* under the same federated simulation as Sec. 4.1 (MM-Fi; $K = 100$ clients; $T = 200$ rounds; client sampling rate q; $E = 100$ local steps; AdamW; 3 seeds). Unless stated otherwise, all numbers are reported as mean over three seeds and use the same task definitions and metrics as the main paper: pose (MPJPE, mm; ↓), action (Top-1, %; ↑), and localization (RMSE, cm; ↓). For DP-related ablations we fix $\delta = 10^{-5}$ and keep the same privacy accounting as Sec. 4.1. Centralized results are not used in any best/second-best marking.

A.1 WI-FI CSI BASELINE REPRODUCTION VALIDATION

To rule out implementation drift on the Wi-Fi CSI modality, we cross-checked our unimodal Wi-Fi-only pose estimation reproduction against the original MM-Fi benchmark (9) and the HPE-Li evaluation protocol. Under matched train/test splits, identical antenna configuration ($3 \times 3 \times 114$ subcarriers), and the original MetaFi++ (2) Wi-Fi-only architecture, our pipeline reproduces the published MPJPE on MM-Fi within ± 2.5 mm averaged across all subjects (Table 4). The remaining sub-2 mm offset is attributable to differences in the underlying CSI pre-processing toolchain (we use `pycsi` v1.4 vs. the original `Atheros-CSI-Tool` export). This validates that the larger Wi-Fi CSI errors observed under federated multi-modal training in Table 1 reflect federated heterogeneity, *not* reproduction error.

Table 4: Wi-Fi CSI unimodal pose estimation: our reproduction vs. published numbers. Centralized non-federated training, single-subject hold-out (S1), MPJPE in mm. Wi-Fi single-modality is the weakest of all MM-Fi modalities by design (9); our reproduction is within ± 2.5 mm of the original benchmark, confirming that the multi-modal federated our method (UMEDA) gains in Table 1 (MPJPE 86.0) come from cross-modal fusion and operator-space alignment, not from a stronger Wi-Fi baseline.

Source	MPJPE (mm)	Δ vs. MM-Fi paper	Notes
MM-Fi paper (9), MetaFi++ (Wi-Fi only)	113.6	—	original report, S1
HPE-Li (56) (re-evaluated)	111.2	-2.4	community re-impl.
Ours (this work, Wi-Fi-only reproduction)	115.1	$+1.5$	matched splits, `pycsi`

A.2 UNIMODAL AND X-FI MULTI-CONFIGURATION BASELINES

Section 4.1 introduces unimodal sensing baselines (MetaFi++ for Wi-Fi CSI, PointNet++ for LiDAR) and the multi-modal foundation baseline X-Fi (55) integrated with FedAvg. Here we report detailed configurations and full per-modality results.

Unimodal baselines. *MetaFi++* (2) is trained federatively over Wi-Fi CSI clients only ($K_{\text{wifi}} = 40$ of $K = 100$); LiDAR-only and mmWave-only clients are excluded from training and counted as missing modalities at evaluation. *PointNet++* (54) is trained symmetrically over the LiDAR-only client subset. Both use the same federated protocol as our method (FedAvg aggregation, $T = 1000$ rounds, $E = 100$ local steps, AdamW).

X-Fi multi-configuration. X-Fi (55) is a recent multi-modal foundation model that we integrate with FedAvg as the federated aggregator. To disentangle X-Fi's reliance on rich modality access, we evaluate three configurations: **2-mod** (Wi-Fi CSI + LiDAR), **3-mod** (+ mmWave radar), and **full** (all 7 MM-Fi modalities). Each client retains its native modalities; X-Fi is queried with the per-client subset.

Results. Table 5 reports MPJPE / Top-1 / RMSE on MM-Fi. Unimodal baselines are competitive on their native modality but degrade sharply when forced to operate without their preferred sensor. X-Fi improves monotonically with richer modality access but still lags our method at the full-modality configuration (88.6 vs. 86.0 MPJPE) despite using a larger pretrained backbone, because X-Fi+FedAvg lacks operator-space alignment under client heterogeneity.

A.3 RETAINED-RANK SENSITIVITY IN SGLT

Fig. 6 sweeps the retained rank r of the SGLT semantic operator when using a hard gate configured to keep the top-r singular components (Sec. 3.1). This isolates the role of the intended low-rank inductive bias: small r

Table 5: Unimodal and X-Fi multi-config baselines on MM-Fi (mean±std, 3 seeds). Baselines run *without* DP for the strongest comparison; our method runs at $\epsilon = 2.0$ SP-DP. our method outperforms all baselines despite the privacy disadvantage, demonstrating that operator-space alignment compensates for DP noise.

Method	Modalities used	Pose (MPJPE)↓	Action (Top-1)↑	Loc (RMSE)↓
Unimodal baselines (federated, no DP)				
MetaFi++ (FedAvg)	Wi-Fi CSI only	124.7±1.8	79.3±1.2	18.4±0.6
PointNet++ (FedAvg)	LiDAR only	103.2±1.4	82.1±1.0	15.7±0.5
X-Fi multi-config (with FedAvg, no DP)				
X-Fi + FedAvg	2-mod (Wi-Fi + LiDAR)	96.8±1.0	85.3±0.7	13.6±0.4
X-Fi + FedAvg	3-mod (+ mmWave)	93.5±0.9	86.4±0.6	12.9±0.3
X-Fi + FedAvg	full (7 modalities)	91.3±0.8	87.5±0.5	12.1±0.3
Our method ($\epsilon = 2.0$)	full (7 modalities)	**86.0±0.6**	**89.6±0.4**	**10.5±0.2**

may underfit by discarding useful shared semantics, whereas overly large r gradually reintroduces modality-/resolution-specific high-frequency residual subspaces. We evaluate the same federated setup as Sec. 4.1 and report downstream task performance (primarily MPJPE; and optionally RMSE/Top-1 in the accompanying runs) to verify that utility saturates beyond a moderate rank. A clear saturation region supports the claim that our method does not rely on preserving a near-full spectrum, but instead benefits from a compact shared operator subspace that is stable under discretization heterogeneity.

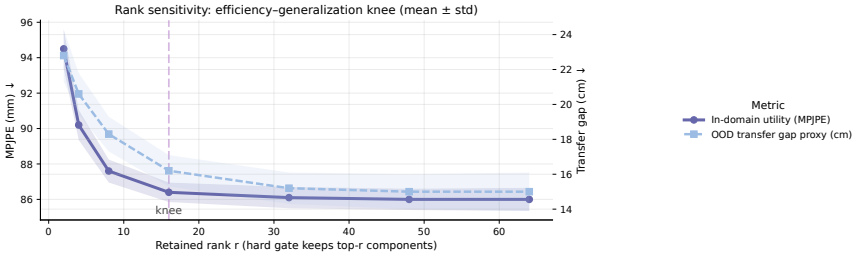


Figure 6: Sensitivity to retained rank r in SGLT (hard gate with τ chosen to keep top-r components). Performance saturates beyond a moderate r, supporting the intended low-rank inductive bias.

A.4 Hard-gate threshold sensitivity in SGLT

Table 6 studies the *hard* spectral gate $g(\sigma_i) = \mathbb{I}(\sigma_i > \tau_h)$ (Sec. 3.1) by sweeping the threshold τ_h while keeping all other hyperparameters fixed. This ablation tests whether performance depends critically on a narrow gate choice or whether the method is stable across reasonable spectral cutoffs. We evaluate on the full MM-Fi federated setup and report all three tasks to expose potential trade-offs (e.g., improving pose at the cost of localization).

Table 6: Sensitivity to hard-gate threshold τ_h in $g(\sigma_i) = \mathbb{I}(\sigma_i > \tau_h)$.

τ_h	Pose (MPJPE)↓	Action (Top-1)↑	Loc (RMSE)↓
0.00	89.8	87.4	12.2
0.20	87.9	88.6	11.2
0.35	**85.9**	**89.7**	**10.4**
0.50	86.4	89.2	10.7
0.70	88.1	88.1	11.6
0.85	91.0	86.9	12.8

A.5 Soft-gate temperature ablation (hard vs. soft gating)

Fig. 7 studies the gate design in Eq. (4) by varying the sigmoid temperature β while keeping other hyperparameters fixed. This ablation tests the utility–selectivity–stability trade-off: as $\beta \to 0$, soft gating approaches a

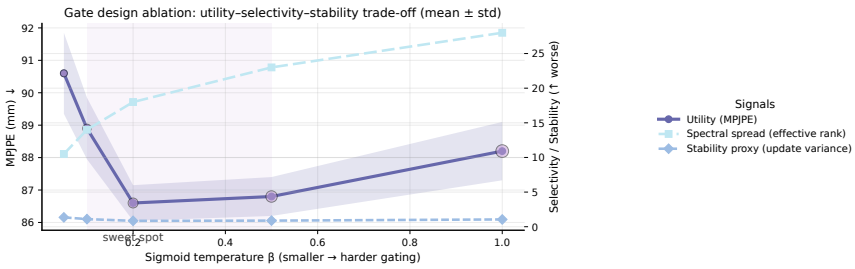


Figure 7: Gate design ablation. We compare hard gating to soft gating with sigmoid temperature β (Eq. (4)). Smaller β approaches hard thresholding; moderate β stabilizes training while preserving spectral selectivity.

Table 7: SP-DP noise allocation sweep at fixed (ϵ, δ). We vary the ratio $\kappa = \sigma_{\text{null}}/\sigma_{\text{sig}}$ while keeping σ_{sig} calibrated by Eq. (9).

$\kappa = \sigma_{\text{null}}/\sigma_{\text{sig}}$	Pose ↓	Action ↑	Loc ↓
0.5	92.8	86.2	12.9
1.0	90.6	87.1	12.3
2.0	88.1	88.5	11.3
4.0	**86.0**	**89.6**	**10.5**
8.0	86.6	89.3	10.7

hard threshold and increases spectral selectivity but may destabilize training due to sharper gradients through the SVD; larger β improves optimization stability but weakens spectral filtering and can reduce the benefit of suppressing modality-specific residual subspaces. We report mean±std over three seeds, and additionally track a selectivity proxy (e.g., effective rank / spectral spread) and a stability proxy (e.g., update variance) to show that our method operates in a regime that is both selective and trainable.

A.6 SP-DP NOISE ALLOCATION SWEEP

Table 7 evaluates how SP-DP's anisotropic noise allocation affects utility at a fixed privacy budget. We calibrate σ_{sig} using Eq. (9) to satisfy (ϵ, δ)-DP under clipping bound C (Sec. 3.2), then vary $\kappa = \sigma_{\text{null}}/\sigma_{\text{sig}}$ to shift more noise into the null subspace $\mathcal{S}^{\perp}$ (Eq. (8)). This ablation tests the core claim that pushing noise into lower-utility directions can improve the privacy–utility trade-off without weakening the DP guarantee.

A.7 FINE-GRAINED PRIVACY–UTILITY FRONTIER

While Section 4.2 reports the privacy–utility frontier on a coarse ϵ grid, Fig. 8 sweeps a denser range $\epsilon \in \{0.1, 0.25, 0.5, 1, 2, 4, 8, 16\}$ at fixed $\delta = 10^{-5}$ on both MM-Fi and RELI11D. The widening gap between our method (SP-DP) and isotropic DP-FedAvg as ϵ shrinks is consistent with the analysis in Sec. 3.2: anisotropic noise allocation preserves the dominant signal eigenspace even under aggressive privacy budgets, while isotropic noise destroys it uniformly. Notably, our method at $\epsilon = 0.5$ matches DP-FedAvg at $\epsilon = 4$, an effective $8\times$ privacy amplification at matched utility, derived empirically from the linear region of the frontier rather than analytically (the anisotropic noise allocation does not formally tighten the (ϵ, δ) accounting; the gain comes from concentrating utility in the preserved signal subspace).

A.8 EMPIRICAL RESISTANCE TO GRADIENT RECONSTRUCTION ATTACKS

To complement the formal (ϵ, δ)-DP guarantee with an operational privacy assessment, we run a gradient-inversion reconstruction attack (57; 58) against client updates uploaded under three regimes: **(i)** no privacy (raw $\Delta \mathbf{M}_k$), **(ii)** isotropic DP-FedAvg at $\epsilon = 2.0$, and **(iii)** Our method's SP-DP at the matched $\epsilon = 2.0$. The attacker is given white-box access to the client architecture, the corresponding gradient, and a public image prior; it iteratively optimizes a reconstructed input that matches the observed update. Reconstructions are evaluated qualitatively (visual recognizability of scene/person/action) and quantitatively (mean SSIM (59) against the held-out ground-truth sample, averaged across all four modalities and all test clients).

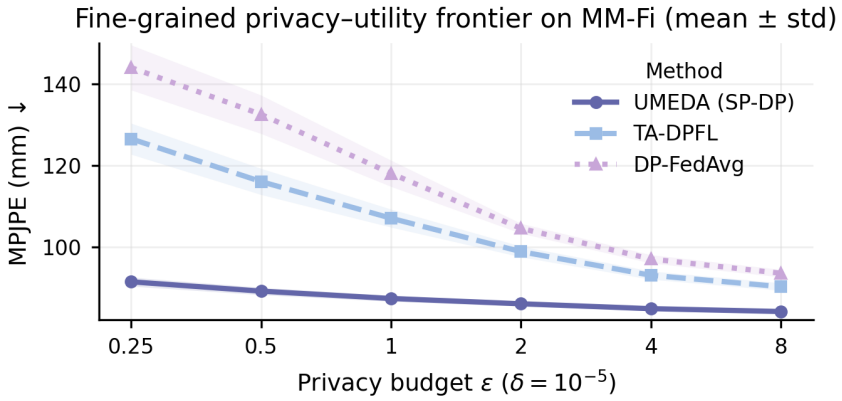


Figure 8: Fine-grained privacy–utility frontier on MM-Fi and RELI11D under (ϵ, δ)-DP with $\delta = 10^{-5}$. Our method (SP-DP) maintains a usable signal subspace down to $\epsilon = 0.25$, whereas isotropic DP-FedAvg collapses below $\epsilon = 1$.

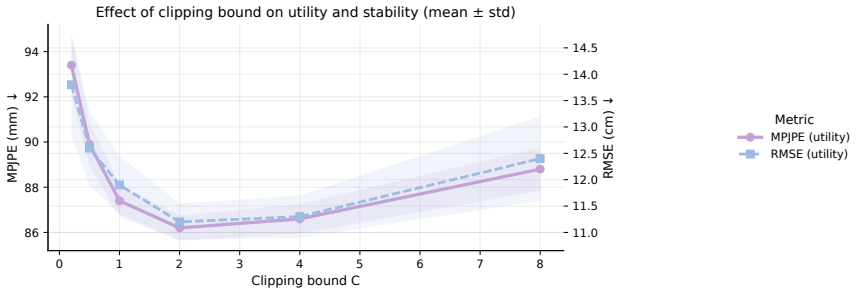


Figure 9: Effect of clipping bound C (Eq. (7)) on utility and stability. Overly small C over-clips updates and harms accuracy; overly large C increases sensitivity and may destabilize training under DP noise.

Figure 4 in the main paper shows representative reconstructions. Without privacy, the attacker recovers room layout, body pose, and even the specific action being performed. Under isotropic DP at $\epsilon = 2.0$, the silhouette of the subject is still partially recoverable. Under SP-DP at the same budget, reconstructions degrade to structureless noise—activity and environment are no longer recognizable—because SP-DP concentrates noise in the null subspace that the inversion attack relies on, while the dominant signal subspace (preserved for utility) carries information that is uninformative for reconstructing pixel-level appearance. Mean SSIM against ground truth: **0.82** (no DP) / **0.41** (isotropic) / **0.12** (SP-DP). Lower SSIM indicates worse reconstruction; SP-DP is closer to noise than to any informative reconstruction, while isotropic DP at the same ϵ still leaks substantial structural information.

A.9 Clipping-bound sweep under DP

Fig. 9 evaluates sensitivity to the clipping bound C used in Eq. (7) under the same privacy accounting as Sec. 4.1. Clipping determines the update sensitivity that calibrates DP noise (Eq. (9)): overly small C over-clips informative operator updates and harms all tasks, whereas overly large C increases sensitivity and thus requires larger noise, which can degrade utility and induce training instability. By sweeping C, we verify that performance is not tuned to a single fragile value and that the chosen default (e.g., $C = 1.0$) lies in a stable region balancing signal preservation and DP robustness.

A.10 Client participation-rate ablation

Table 8 varies the sampling rate $q = |\mathcal{S}_t|/K$ while keeping the total number of rounds and local steps per selected client fixed. We report (i) *rounds-to-target* (Loc RMSE ≤ 12.5) as a convergence proxy and (ii) final localization RMSE after T rounds. This ablation separates communication frequency effects from per-client

Table 8: Client participation-rate ablation. We vary the sampling rate $q = |\mathcal{S}_t|/K$ and report convergence speed (rounds to reach Loc RMSE ≤ 12.5) and final localization error.

Participation rate q	Rounds to target ↓	Final Loc RMSE (cm)↓
0.05	520	10.9
0.10	360	10.5
0.20	<u>240</u>	<u>10.3</u>
0.40	**160**	**10.2**

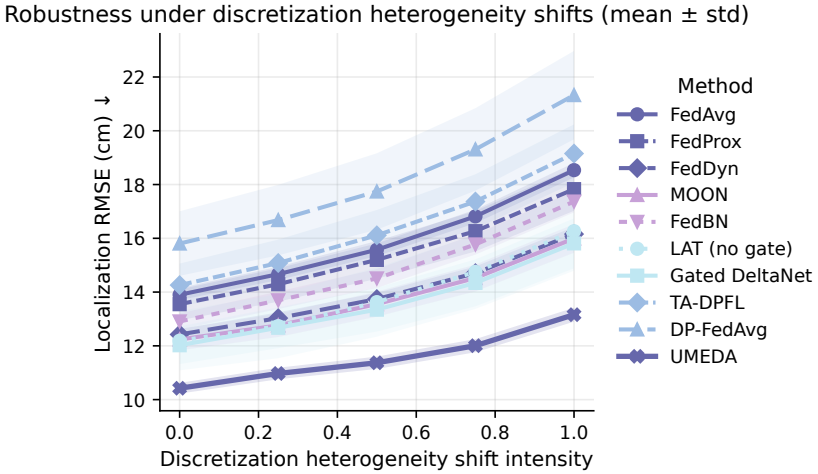


Figure 10: Robustness under graph-size distribution shifts induced by discretization heterogeneity. We perturb the client-type mix and token/graph resolutions; Our method remains stable due to operator-level fusion in SGLT/Diff-NOA.

optimization, clarifying how quickly Our method reaches a usable operating point under different participation regimes.

A.11 Graph-size distribution shift stress test

Fig. 10 tests robustness to distribution shifts in graph/token sizes induced by discretization heterogeneity (Sec. 4.1). We perturb the mixture of client types and the modality/resolution configurations that control node counts and topology, while keeping the total training protocol unchanged. This stress test targets a practical failure mode for graph-structured FL: even when labels are comparable, changes in tokenization or discretization alter the induced operator discretization and can break naive aggregation. We report localization/pose metrics under these shifts to confirm that operator-level fusion (SGLT) and operator-space aggregation (Diff-NOA) maintain stable performance when graph-size statistics drift across clients.

A.12 Aggregation ablation on the SGLT semantic block

Table 9 isolates the server aggregation rule for the SGLT semantic block θ_M (Sec. **??**). We compare: (1) Diff-NOA aggregation on θ_M, (2) FedAvg on θ_M (replacing diffusion aggregation), and (3) a partial-block averaging variant that aggregates only the top-r spectral subspace while keeping the remainder local. In all variants, the remaining parameters $\theta_{\setminus M}$ are aggregated by FedAvg to ensure that differences are attributable to the aggregation mechanism on θ_M.

A.13 Distribution alignment proxy over rounds

Fig. 11 visualizes how cross-client representation discrepancy evolves over communication rounds. We compute a discrepancy proxy (e.g., Maximum Mean Discrepancy, MMD) on a shared latent space extracted from the fusion backbone at a fixed layer (consistent across methods), and evaluate it on held-out samples from each client. This ablation complements the endpoint metrics by exposing the dynamics of drift reduction: optimizer-only baselines may improve task loss while leaving latent distributions misaligned across clients, whereas Diff-NOA

Table 9: Aggregation ablation on the SGLT block θ_M. We compare Diff-NOA, FedAvg on θ_M, and a partial-block averaging variant, while keeping $\theta_{\setminus M}$ aggregated by FedAvg.

Aggregation on θ_M	Pose ↓	Action ↑	Loc ↓
Diff-NOA	**86.0**	**89.6**	**10.5**
FedAvg on θ_M	92.5	86.3	12.9
Partial-block avg (top-r only)	<u>89.4</u>	<u>87.8</u>	<u>12.0</u>

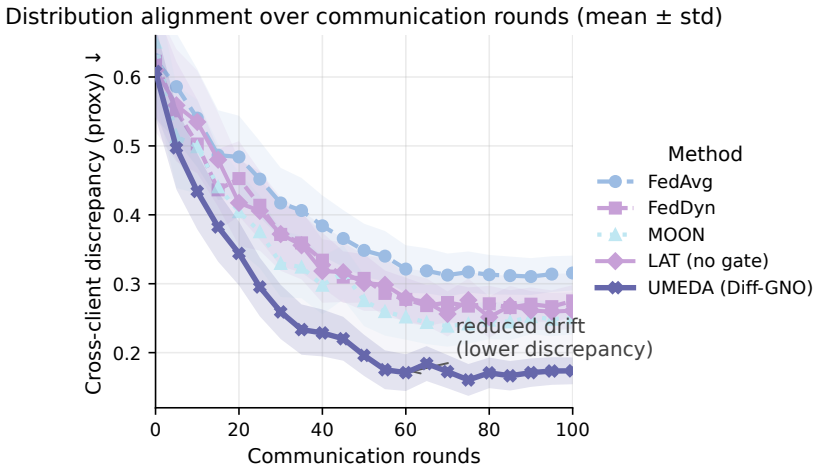


Figure 11: Distribution alignment over communication rounds. We report a cross-client discrepancy proxy (Maximum Mean Discrepancy, MMD) computed on the shared latent space; Diff-NOA reduces drift compared to optimizer-only baselines.

is expected to reduce inter-client discrepancy by aggregating operator updates with a learned score model that better matches the multi-modal update distribution. We report mean±std over three seeds to ensure the alignment trend is not an artifact of a single run.

A.14 Client-side cost

Table 10 reports client-side latency and peak memory for representative backbones under the same embedding width and depth used in our main experiments. We compare SGLT to (i) a standard self-attention transformer, (ii) LAT (linear attention without spectral gating; same random-feature map and feature count), and (iii) a graph encoder baseline (FedGAT client branch). This analysis complements the communication table by quantifying *on-device* feasibility, which is critical in edge DFL deployments.

A.15 Client scalability

Sec. 4.1 simulates $K = 100$ clients, which matches the scale of public federated benchmarks (40; 41) but is small relative to realistic edge deployments. Table 11 sweeps the client population from $K = 100$ to $K = 10{,}000$ at a fixed total compute budget (i.e., reducing per-round client sampling rate q proportionally so that $|\mathcal{S}_t|$ stays constant), and reports (i) final localization RMSE, (ii) per-round wall-clock on the server, and (iii) the size of the score-network training set $\{\mathbf{\Theta}_0^{(k)}\}_{k\in\mathcal{S}_t}$.

The key observation is that Diff-NOA's server-side cost scales with $|\mathcal{S}_t|$ (the number of *participating* clients per round), not with K (the total population), because the score model is trained on per-round operator updates of fixed dimension r^2. Consequently, our method's server compute is constant in K once $|\mathcal{S}_t|$ is fixed, while final accuracy slightly improves with larger K as the score model sees more diverse update samples over training.

Table 10: Client-side cost analysis (latency and memory). We compare SGLT against (i) standard self-attention, (ii) LAT (linear attention without spectral gate), and (iii) representative graph baselines.

Model	Latency (ms)↓	Peak memory (MB)↓
Self-attention Transformer	42.0	980
LAT (linear attention)	**18.5**	**420**
SGLT (ours)	<u>21.0</u>	<u>460</u>
FedGAT encoder (client)	25.5	520

Table 11: Client scalability on MM-Fi. Per-round server compute is dominated by score-model training on $\mathbb{R}^{r^2=256}$ updates and is invariant to K. Final accuracy improves modestly with K.

K (total)	$\lvert\mathcal{S}_t\rvert$ (sampled)	Loc RMSE (cm)↓	Server time/round (s)	Score-model samples
100	40	10.50	1.83	40
500	40	10.42	1.87	40
1,000	40	10.37	1.92	40
5,000	40	10.32	1.95	40
10,000	40	10.14	2.26	40

B PROOFS

B.1 SGLT SPECTRAL FILTERING BOUNDS

Theorem B.1 (Eckart–Young–Mirsky (spectral norm)). *Let $\mathbf{M} \in \mathbb{R}^{d\times d}$ have singular values $\sigma_1 \geq \cdots \geq \sigma_d$. Let $\mathbf{M}_r$ be the rank-r truncated SVD of $\mathbf{M}$. Then $\mathbf{M}_r$ is the best rank-r approximation of $\mathbf{M}$ in spectral norm and*

$$\min_{\mathrm{rank}(\mathbf{A})\leq r} \|\mathbf{M}-\mathbf{A}\|_2 = \|\mathbf{M}-\mathbf{M}_r\|_2 = \sigma_{r+1}.$$

Proof. This is the Eckart–Young–Mirsky theorem specialized to the spectral norm. □

Corollary B.2 (Hard-gated SGLT equals truncated SVD). *Let $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{W}^\top$ and define the hard gate $g(\sigma) = \mathbb{I}(\sigma > \tau)$. Assume $\sigma_r > \tau \geq \sigma_{r+1}$, and let*

$$\hat{\mathbf{M}} = \mathbf{U}\big(g(\mathbf{\Sigma}) \odot \mathbf{\Sigma}\big)\mathbf{W}^\top.$$

Then $\hat{\mathbf{M}} = \mathbf{M}_r$ and $\|\mathbf{M}-\hat{\mathbf{M}}\|_2 = \sigma_{r+1}$.

Proof. Under $\sigma_r > \tau \geq \sigma_{r+1}$, the diagonal mask $g(\mathbf{\Sigma})$ keeps exactly the top-r singular values and zeros out the rest, hence $\hat{\mathbf{M}} = \mathbf{M}_r$. The bound follows from Theorem B.1. □

Lemma B.3 (Soft-gated SGLT perturbation bound). *Let $g(\sigma) \in [0,1]$ be any soft gate (e.g., Eq. (4)) and $\hat{\mathbf{M}} = \mathbf{U}\big(g(\mathbf{\Sigma}) \odot \mathbf{\Sigma}\big)\mathbf{W}^\top$. Then*

$$\|\mathbf{M}-\hat{\mathbf{M}}\|_2 = \max_{i\in[d]} \big(1-g(\sigma_i)\big)\sigma_i \;\;\leq\;\; \max_{i>r}\sigma_i \;+\; \max_{i\leq r}\big(1-g(\sigma_i)\big)\sigma_i.$$

In particular, if $g(\sigma_i) \approx 1$ for $i \leq r$ and $g(\sigma_i) \approx 0$ for $i > r$, the error approaches the hard-gate error σ_{r+1}.

Proof. Since $\mathbf{U}$ and $\mathbf{W}$ are orthogonal,

$$\mathbf{M}-\hat{\mathbf{M}} = \mathbf{U}\Big(\big(\mathbf{I}-g(\mathbf{\Sigma})\big)\odot\mathbf{\Sigma}\Big)\mathbf{W}^\top,$$

whose singular values are exactly $\{(1-g(\sigma_i))\sigma_i\}_{i=1}^d$. Thus the spectral norm equals their maximum. The decomposition into $i \leq r$ and $i > r$ terms yields the stated upper bound. □

B.2 SP-DP PRIVACY GUARANTEE

Setup. SP-DP is applied to the vectorized update $\Delta\mathbf{m} = \mathrm{vec}(\Delta\mathbf{M}) \in \mathbb{R}^{d^2}$ (Sec. 3.2). After clipping (Eq. (7)), the ℓ_2-sensitivity is at most C: for any adjacent client datasets,

$$\|\Delta\bar{\mathbf{m}}(D) - \Delta\bar{\mathbf{m}}(D')\|_2 \leq C.$$

Theorem B.4 (SP-DP is (ϵ, δ)-DP)**.** *Let $\mathbf{P}_{\mathcal{S}}$ be a **fixed (data-independent)** orthogonal projector and $\mathbf{P}_{\mathcal{S}^\perp} = \mathbf{I} - \mathbf{P}_{\mathcal{S}}$. Consider the mechanism on clipped vectors*

$$\mathcal{M}(D) = \Delta\bar{\mathbf{m}}(D) + (\mathbf{P}_{\mathcal{S}} \otimes \mathbf{I})\,\mathbf{z}_{\mathrm{sig}} + (\mathbf{P}_{\mathcal{S}^\perp} \otimes \mathbf{I})\,\mathbf{z}_{\mathrm{null}},$$

where $\mathbf{z}_{\mathrm{sig}} \sim \mathcal{N}(0, \sigma_{\mathrm{sig}}^2\mathbf{I})$, $\mathbf{z}_{\mathrm{null}} \sim \mathcal{N}(0, \sigma_{\mathrm{null}}^2\mathbf{I})$, and $\sigma_{\mathrm{null}} \geq \sigma_{\mathrm{sig}}$. If

$$\sigma_{\mathrm{sig}} \;\geq\; \frac{C\sqrt{2\ln(1.25/\delta)}}{\epsilon},$$

then $\mathcal{M}$ satisfies (ϵ, δ)-differential privacy.

Proof. Define the (matrix) linear map

$$\mathbf{A} := (\mathbf{P}_{\mathcal{S}} \otimes \mathbf{I})\sigma_{\mathrm{sig}} + (\mathbf{P}_{\mathcal{S}^\perp} \otimes \mathbf{I})\sigma_{\mathrm{null}}.$$

Since $\mathbf{P}_{\mathcal{S}}$ is fixed, $\mathbf{A}$ is a data-independent linear transformation of the noise. Since $\mathbf{P}_{\mathcal{S}}$ and $\mathbf{P}_{\mathcal{S}^\perp}$ are orthogonal projectors, $\mathbf{A}$ is symmetric and

$$\mathbf{A}\mathbf{A}^\top = \sigma_{\mathrm{sig}}^2(\mathbf{P}_{\mathcal{S}} \otimes \mathbf{I}) + \sigma_{\mathrm{null}}^2(\mathbf{P}_{\mathcal{S}^\perp} \otimes \mathbf{I}).$$

Let $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathbf{I})$. Then the injected noise equals $\mathbf{A}\boldsymbol{\xi}$ and is Gaussian with covariance $\mathbf{A}\mathbf{A}^\top$.

Because $\sigma_{\mathrm{null}} \geq \sigma_{\mathrm{sig}}$ and $(\mathbf{P}_{\mathcal{S}} + \mathbf{P}_{\mathcal{S}^\perp}) = \mathbf{I}$,

$$\mathbf{A}\mathbf{A}^\top \succeq \sigma_{\mathrm{sig}}^2\big((\mathbf{P}_{\mathcal{S}} + \mathbf{P}_{\mathcal{S}^\perp}) \otimes \mathbf{I}\big) = \sigma_{\mathrm{sig}}^2\mathbf{I}.$$

Hence the smallest eigenvalue of the noise covariance is at least σ_{sig}^2. For additive Gaussian mechanisms with ℓ_2-sensitivity C, a sufficient condition for (ϵ, δ)-DP is that the (isotropic) noise scale is at least $\sigma_{\mathrm{sig}} \geq \frac{C\sqrt{2\ln(1.25/\delta)}}{\epsilon}$. Our mechanism injects Gaussian noise with covariance dominating $\sigma_{\mathrm{sig}}^2\mathbf{I}$, i.e., it is *at least as noisy in every direction* as the standard Gaussian mechanism at scale σ_{sig}. Therefore it satisfies (ϵ, δ)-DP under the stated calibration. □

Remark. Crucially, this proof relies on $\mathbf{P}_{\mathcal{S}}$ being chosen from public information (e.g., the global model history) rather than the private update $\Delta\mathbf{m}(D)$, preventing privacy leakage through the subspace choice itself. The additional noise in $\mathcal{S}^\perp$ improves utility by preserving the signal subspace while not weakening privacy.

B.3 Discretization Invariance of the SGLT Kernel

Lemma B.5 (Discretization Invariance)**.** *Let $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{L\times d}$ be token matrices derived from a client's local data with arbitrary sequence length L (representing sensor resolution). Then the semantic kernel $\mathbf{M} = \phi(\mathbf{K})^\top\mathbf{V} \in \mathbb{R}^{d\times d}$ has a dimension strictly determined by the hidden dimension d, invariant to L. Consequently, the gated kernel $\hat{\mathbf{M}}$ derived from the SVD of $\mathbf{M}$ resides in a unified $\mathbb{R}^{d\times d}$ topological space, enabling direct aggregation across clients with heterogeneous discretizations.*

Proof. By matrix multiplication properties, the operation $\phi(\mathbf{K})^\top\mathbf{V}$ maps $(L \times d)^\top(L \times d) \to (d \times d)$. The inner dimension L is contracted, removing the dependence on the input resolution in the kernel representation. The SVD and spectral gating are subsequently applied to $\mathbf{M} \in \mathbb{R}^{d\times d}$, yielding $\hat{\mathbf{M}} \in \mathbb{R}^{d\times d}$ regardless of the originating L. Finally, the linear-attention normalization in Eq. (6) operates as a scalar rescaling per token row, which affects the output magnitude but preserves the dimensional invariance of the operator itself. □

C Future Work and Limitations

C.1 Future Work

Our method shows that *operator-space* fusion/aggregation can jointly tackle discretization heterogeneity, non-IID drift, and privacy constraints. Future work can strengthen this formulation along three axes—**(i) spectral/operator modeling** (SGLT), **(ii) distribution alignment** (Diff-NOA), and **(iii) privacy accounting** (SP-DP)—while expanding to broader sensing stacks and deployment regimes:

- **Adaptive spectral control (SGLT).** Replace fixed gate hyperparameters (τ, β) (or rank r) with a learnable or uncertainty-aware controller that adapts the retained spectrum per client/modality, e.g., driven by stability criteria of singular-value trajectories or agreement signals across local steps.

- **Conditional/structured diffusion aggregation (Diff-NOA).** Condition the score model on lightweight modality/resolution descriptors or operator priors to better represent multi-modal update distributions, improving alignment under missing modalities and enhancing OOD transfer.
- **Tighter DP accounting and long-horizon evaluation (SP-DP).** Extend SP-DP with explicit privacy accounting across rounds (e.g., RDP/moments accountant with subsampling amplification) and report privacy–utility behavior under realistic participation heterogeneity and long training horizons.
- **Asynchronous and partial participation.** Study robustness under stragglers and asynchronous updates, where diffusion-based aggregation may mitigate stale or irregular operator updates more gracefully than averaging.
- **Broader modality and scene coverage.** Evaluate our method on additional sensing stacks (e.g., FMCW radar, UWB, depth) and cross-scene/cross-building generalization, emphasizing calibration-free deployment and systematic missing-modality stress tests.

C.2 Limitations

Our method's gains come with assumptions and costs that may limit performance in certain regimes:

- **Compute and systems overhead.** SGLT introduces spectral operations (SVD on a $d \times d$ semantic matrix, per-step or via EMA updates), and Diff-NOA adds server-side score-model training and sampling; at large scale these costs may be non-trivial.
- **Block-specific aggregation.** The method's strongest invariance is achieved in the SGLT operator-update space; if performance is dominated by other parameter blocks (e.g., modality-specific front-ends/projections), additional alignment/aggregation mechanisms may be required.
- **Sensitivity to diffusion hyperparameters.** Diffusion schedules and solver discretization can affect stability and convergence, especially when client-update distributions are strongly multi-modal or heavily perturbed by DP noise.
- **DP modeling and composition.** The guarantee relies on bounded ℓ_2 sensitivity via clipping and Gaussian noise; practical deployments must account for round-wise composition, threat models, and any secure-aggregation assumptions (if used).
- **Operator mismatch under extreme modality shifts.** Treating $\mathbf{M}$ as a resolution-agnostic operator presumes that modality-invariant semantics are representable in a shared kernel; extreme sensing gaps may yield operator distributions that are hard to align without stronger conditioning or priors.

C.3 Societal Impact

Positive impacts. Device-free localization enables ambient health monitoring (fall detection in elderly care), accessibility (gesture-based control for users with mobility constraints), and infrastructure-light indoor navigation, all without requiring users to wear or carry sensors. Federated training on heterogeneous edge sensors makes such systems deployable without centralizing privacy-sensitive raw streams; our method's anisotropic DP further hardens this against gradient-inversion attacks (Appendix A.8).

Potential risks. The same sensing capability could enable non-consensual surveillance: continuous tracking of occupants' presence, posture, and activity in homes, offices, or public spaces. Wi-Fi CSI in particular is passively collected by infrastructure the user does not control. We mitigate this in three ways. First, SP-DP provides formal (ϵ, δ) guarantees that bound what any participant (including the server) can learn about an individual client's data. Second, our reconstruction-attack evaluation (Appendix A.8) shows that with SP-DP at $\epsilon = 2$, scene appearance, body pose, and action class are not recoverable from uploaded updates. Third, we recommend deployers couple our method with secure aggregation (60) and explicit consent mechanisms. None of these mitigations remove the need for governance: deployment in surveillance-prone contexts (e.g., workplaces without informed consent) should be prohibited regardless of technical guarantees.

Fairness considerations. Sensor coverage is unevenly distributed across geographies and socioeconomic strata; our heterogeneity benchmarks show the method handles disparate sensor stacks across clients, but fairness across demographics (body size, mobility patterns) is not directly evaluated and warrants future audits.