FEDERATED SHARED–PRIVATE BASIS REALLOCATION FOR MULTIMODAL REPRESENTATION COLLAPSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal federated learning can overuse a dominant modality, while heterogeneous client updates suppress effective directions in weaker pathways. Existing approaches regulate modality selection, optimization, or distillation, but leave weak encoder capacity unprotected during aggregation. We introduce Federated Shared and Private Basis Reallocation (FedSBR), an encoder-level framework that separates shared semantic coordinates from private residual coordinates. FedSBR preserves weak-pathway capacity through collapse-aware regularization, transfers supervision only when loss, rank, and teacher reliability indicate dominance, and coordinates class semantics through reliability-weighted federated prototypes. Identity-preserving initialization reconstructs features without changing the initial predictor. Experiments on CREMA-D, ModelNet40, CrisisMMD, and MIMIC-IV under IID and Dirichlet non-IID partitions show improvements over FedAvg and competitive fusion performance. Pathway accuracy, participation rank, modality contribution, and class-conditional analyses collectively support the proposed mechanism under federated heterogeneity.

1 INTRODUCTION

Federated learning (FL) enables multiple clients to collaboratively train a model without sharing their private data (McMahan et al., 2017). This paradigm is attractive for multimodal applications, where audio, visual, textual, and sensor signals are distributed across privacy-sensitive devices or institutions. Multimodal models can exploit complementary information, but heterogeneous local optimization and global aggregation also create failure modes that are less visible in standard unimodal FL. Client drift and optimizer heterogeneity can already make FedAvg unstable (Li et al., 2020; Karimireddy et al., 2020; Reddi et al., 2021); in multimodal models, these effects interact with competition between pathways.

A central difficulty is modality collapse: a fusion model may rely mainly on one modality and neglect another, even when the neglected signal contains complementary information or is needed under missing-modality inference. Prior studies describe modality competition and optimization imbalance in multimodal learning (Wang et al., 2020b; Huang et al., 2022; Peng et al., 2022; Fan et al., 2023). A recent mechanistic account attributes collapse to cross-modal polysemantic interference, where noisy and predictive features become entangled through shared fusion neurons (Chaudhuri et al., 2025). Together with the low-rank bias of learned representations (Huh et al., 2023), this entanglement can turn a weak modality into a rank bottleneck. Collapse is therefore a representation-level capacity allocation problem rather than only an output loss imbalance.

We investigate when federation amplifies collapse beyond centralized multimodal training. In local training, a dominant modality is easier to optimize and can occupy useful fusion directions, leaving the weak modality with a narrow and entangled gradient path. Across clients, dominant-modality updates tend to be more consistent because they follow a stronger predictive signal, whereas weak-modality updates vary with local label skew, modality quality, and fusion state. Server averaging reinforces aligned directions and attenuates disagreeing ones, so weak directions can be cancelled even when they are useful on individual clients. The strength of this mechanism depends on client drift and modality heterogeneity, which we evaluate with pathway and representation diagnostics.

Multimodal FL studies address heterogeneous multimodal data through split multitask models, contrastive representation ensembles, and benchmark protocols (Chen & Zhang, 2022; Yu et al., 2023;

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.

Feng et al., 2023). Related work handles modal imbalance through modality selection and distillation (Fan et al., 2024; Han et al., 2026), and handles missing inputs through reconfigurable federated representations or, in centralized learning, shared-specific modeling and reconstruction (Nguyen et al., 2025; Wang et al., 2023b; Ma et al., 2021). Federated prototype learning further coordinates class semantics without sharing examples (Tan et al., 2022). Missingness removes an input; modality collapse instead suppresses an available pathway during joint optimization. Consequently, reconstruction and prototype-based semantic compensation leave weak-encoder capacity unresolved when all modalities are present. The open problem is how to preserve weak-modality capacity across local optimization and global aggregation. Explicit Basis Reallocation (EBR) (Chaudhuri et al., 2025) is the closest centralized precursor. FedSBR couples local capacity preservation with cross-client semantic coordination to address this federated failure mode.

Motivated by these observations, we propose *Federated Shared–Private Basis Reallocation* (FedSBR). Its design follows a simple federated division of labor: directions expected to be semantically common are shared and anchored, whereas directions tied to modality-specific variation remain private within the shared–private representation. FedSBR therefore separates each modality into shared semantic coordinates and private residual coordinates. The shared coordinates provide a client-aligned semantic interface, while the private coordinates preserve modality-specific directions that low-rank fusion may erase; identity-preserving initialization reconstructs the original feature and leaves the initial predictor unchanged. FedSBR addresses collapse at two coupled levels: Capacity-Preserving Reallocation (CPR) reserves directions for weak modalities, Dominance-Gated Transfer (DGT) transfers teacher supervision only when loss, rank, and reliability indicate a credible dominance state, and Federated Semantic Anchoring (FSA) aggregates reliability-weighted class prototypes to stabilize shared geometry and class-wise coverage across clients. All transfer is client-local, and semantic coordination uses class summaries without a public proxy dataset. CPR follows variance-covariance regularization for preventing degenerate representations (Bardes et al., 2022), while FSA draws on federated prototype learning (Tan et al., 2022). Together, these objectives connect local capacity preservation to cross-client semantic coordination within the shared–private representation.

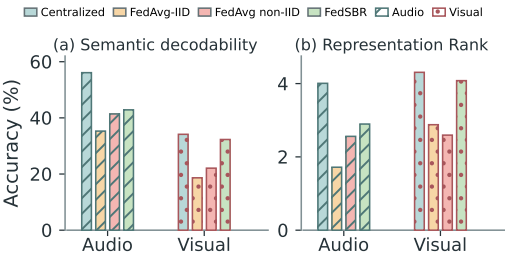

Figure 1: CREMA-D pathway and representation diagnostics. The comparison of fused prediction with unimodal decodability and representation rank motivates examining the weak pathway directly.

Our contributions are threefold:

- We characterize federated amplification of multimodal collapse through client-update inconsistency and its effect on weak-pathway representation capacity.
- We propose FedSBR, coupling shared–private basis reallocation with capacity preservation, gated transfer, and federated semantic anchoring.
- We evaluate weak-pathway preservation using fusion and pathway accuracy, participation rank, modality contribution, class-wise recall, and module ablations across four datasets and three independent seeds.

2 Why Federation Amplifies Multimodal Collapse

Consider modality encodings $\mathbf{z}_m = f_m(\mathbf{x}_m)$ and a fusion map φ that produces $\mathbf{h} = \varphi(\mathbf{z}_1, \ldots, \mathbf{z}_M)$ for prediction $g(\mathbf{h})$. We use modality collapse to describe a fused predictor that underuses an available pathway despite its potential predictive value, relying mainly on a subset $\mathcal{M}_c \subset \{1, \ldots, M\}$. Functional underuse concerns the fusion classifier, whereas information loss concerns the encoder and spectral concentration concerns feature geometry. We examine these complementary aspects through pathway prediction, modality contribution, and participation rank. Empirical studies report the same failure across fusion architectures and tasks (Javaloy et al., 2022; Wang et al., 2020b; Huang et al., 2022), and connect it to gradient imbalance and unequal learning rates across modalities.

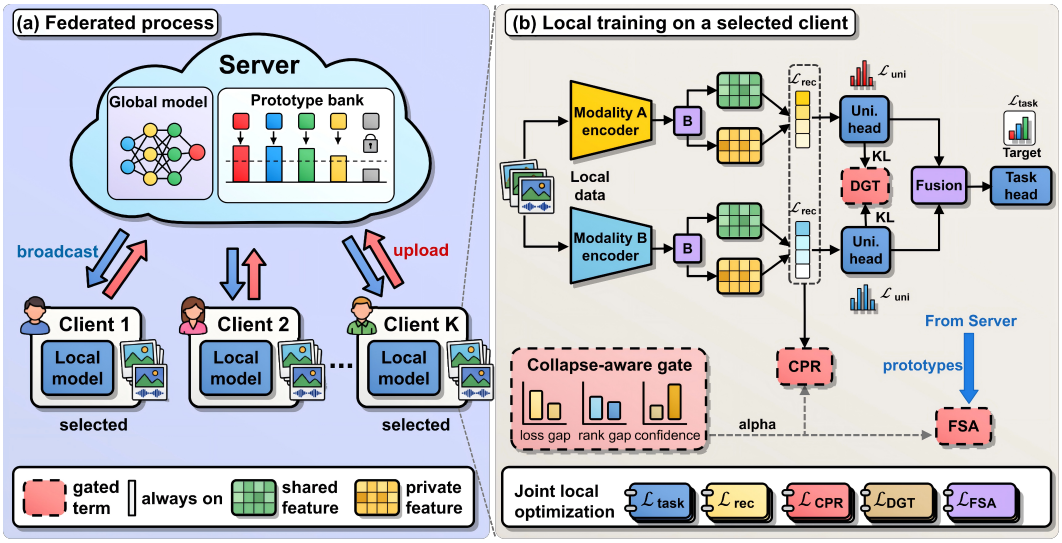

Figure 2: Overview of the FedSBR framework.

The mechanism identified by recent representation analyses is more specific than an output-level loss imbalance. Write a modality feature as $\mathbf{z}_m = \mathbf{z}_m^y + \mathbf{z}_m^\epsilon$, where the superscripts denote predictive and nuisance components. In a shared fusion head, a neuron with weight $\mathbf{w}_p$ can encode $\mathbf{z}_m^\epsilon$ together with $\mathbf{z}_{m'}^y$ from another modality; the resulting cross-modal polysemantic unit makes the marginal effect of $\mathbf{z}_m^y$ depend on the noise component. For fixed fusion width, the $M(M-1)$ ordered cross-modal pairs create more opportunities for interference as the number of modalities grows. Under interference, the weak pathway can lose effective gradient even when $\mathbf{z}_w^y$ remains present in the encoder. Chaudhuri et al. (Chaudhuri et al., 2025) formalize the interference and relate it to faulty capacity allocation; the low-rank simplicity bias of deep networks further favors a small set of directions (Huh et al., 2023). Modality-collapse analyses based on conflicting gradients and modality competition provide complementary empirical and optimization evidence (Javaloy et al., 2022; Huang et al., 2022).

Federated optimization can introduce an additional update obstruction after local fusion dynamics. Client k observes a distribution $P_k(X, Y)$ and obtains a local gradient $g_{b,k}$ for parameter block b. Dominant-modality gradients often share a direction across clients, whereas weak-modality gradients vary with local labels, modality quality, and the active polysemantic units. Weighted averaging therefore retains the coherent component and attenuates components with inconsistent signs. Writing the resulting update error as local drift, participation noise, and sampling error gives

$$\theta_{t+1} = \theta_t - \eta\big(\nabla F(\theta_t) + u_t\big), \qquad u_t = e_t^{\mathcal{S}} + a_t, \tag{1}$$

where $e_t^{\mathcal{S}}$ and a_t are defined for the reference FedAvg recursion in Appendix D. Extensions to other local update rules require accounting for their additional update error. Let P_b project onto parameter block b. For $S_{b,t} = \|g_b(\theta_t)\|^2$ and conditional error budget $\overline{\mathcal{A}}_{b,t}$, the expected block-update alignment obeys

$$\mathbb{E}_t\langle g_b, g_b + P_b u_t\rangle \geq S_{b,t} - \sqrt{S_{b,t}\overline{\mathcal{A}}_{b,t}}. \tag{2}$$

Eq. (2) makes the condition explicit: $\overline{\mathcal{A}}_{b,t} \gtrsim S_{b,t}$ can invalidate a guarantee of positive descent alignment for a weak block, while a coherent dominant block can continue to descend. Client label skew and partial participation also leave the weak pathway with uneven class coverage, the same client-drift setting addressed by FedProx, SCAFFOLD, and adaptive federated optimizers (Li et al., 2020; Karimireddy et al., 2020; Reddi et al., 2021). The ratio $\overline{\mathcal{A}}_{b,t}/S_{b,t}$ thus identifies weak blocks whose predictive gradient is most exposed to federated update error. The spectral analysis below connects this obstruction to rank contraction under explicit conditions on the covariance dynamics.

Cross-client cancellation operates on parameter updates, including with concatenation fusion. Under tail-prioritized contraction, Theorem 2 in Appendix D.5 shows that ordered extra federated contraction reduces participation rank relative to a centralized flow with the same initial spectrum. The

analysis separates retained representation capacity from its predictive use. Pathway accuracy, rank, and class-wise recall therefore complement fused performance when assessing the federated intervention.

3 FEDSBR METHOD

FedSBR couples a shared–private basis with CPR for local capacity, DGT for gated predictive supervision, and FSA for shared class semantics. It complements drift-correcting optimizers (Li et al., 2020; Karimireddy et al., 2020; Reddi et al., 2021) by regularizing the representations they update. Local capacity and aggregated-model geometry are assessed at their respective stages.

3.1 MULTIMODAL FEDERATED SETUP

Let $\mathcal{M} = \{1, \dots, M\}$ denote the modalities. Client k has n_k labeled examples in $\mathcal{D}_k$, with available modalities $\mathcal{M}_k \subseteq \mathcal{M}$. An encoder f_m produces each modality's feature, and a fusion operator φ and classifier g form the prediction:

$$\widehat{\mathbf{y}}_i^k = g\big(\varphi\big(\{f_m(\mathbf{x}_{m,i}^k) : m \in \mathcal{M}_k\}\big)\big). \tag{3}$$

All clients collaboratively train a single global model. The main experiments use complete modalities, $\mathcal{M}_k = \mathcal{M}$; the notation also accommodates clients with different modality sets.

In round t, the server broadcasts the global parameters to participating clients $\mathcal{S}_t$. Let $\mathcal{S}_t^m = \{k \in \mathcal{S}_t : m \in \mathcal{M}_k\}$ and let $n_{m,k}$ be the number of local examples contributing to modality m. The baseline aggregation rule for a modality-specific parameter block is

$$\boldsymbol{\theta}_m^{t+1} = \sum_{k \in \mathcal{S}_t^m} \pi_{m,k}^t \boldsymbol{\theta}_{m,k}^{t+1}, \qquad \pi_{m,k}^t = \frac{n_{m,k}}{\sum_{j \in \mathcal{S}_t^m} n_{m,j}}. \tag{4}$$

FedSBR retains these weights and applies the same collapse-aware intervention under the modality-specific aggregation. FSA supplies an explicit cross-client semantic signal. A block with no participating client retains its previous global value.

3.2 SHARED–PRIVATE BASIS REALLOCATION

For $m \in \mathcal{M}_k$, let $\mathbf{f}_{m,i}^k = f_m(\mathbf{x}_{m,i}^k) \in \mathbb{R}^d$. FedSBR learns a projection B_m and a reconstruction map R_m:

$$\begin{bmatrix} \mathbf{s}_{m,i}^k \\ \mathbf{p}_{m,i}^k \end{bmatrix} = B_m \mathbf{f}_{m,i}^k, \qquad \widetilde{\mathbf{f}}_{m,i}^k = R_m \begin{bmatrix} \mathbf{s}_{m,i}^k \\ \mathbf{p}_{m,i}^k \end{bmatrix}. \tag{5}$$

Here $\mathbf{s}_m \in \mathbb{R}^{d_s}$ carries shared semantics and $\mathbf{p}_m \in \mathbb{R}^{d_p}$ provides modality-specific coordinates, with $R_m = [R_{s,m}, R_{p,m}]$. Both parts are global parameters and are therefore subject to the same federated aggregation as the encoder block. We initialize B_m and R_m as identity-preserving permutations, with $d_s + d_p = d$ and $R_m B_m = I$. The fused predictor uses $\widetilde{\mathbf{f}}_{m,i}^k$ in place of $\mathbf{f}_{m,i}^k$ in Eq. (3), preserving its initial function. The shared block is coordinated across clients by FSA, while CPR preserves variation in the private block with variance–covariance regularization (Bardes et al., 2022).

3.3 COLLAPSE-AWARE LOCAL REALLOCATION

FedSBR identifies suppressed pathways using predictive risk and effective representation capacity, so that the same local signal used to trigger CPR and DGT can be interpreted consistently across clients. Let $\ell_{m,k}$ denote modality m's unimodal cross-entropy on client k. For a centered batch feature matrix $\mathbf{H}_{m,k}$ with $b_{m,k}$ rows, define $\mathbf{C}_{m,k} = b_{m,k}^{-1} \mathbf{H}_{m,k}^\top \mathbf{H}_{m,k}$. Its participation rank is

$$r_{m,k} = \begin{cases} \operatorname{tr}(\mathbf{C}_{m,k})^2 / \operatorname{tr}(\mathbf{C}_{m,k}^2), & \operatorname{tr}(\mathbf{C}_{m,k}) > 0, \\ 0, & \operatorname{tr}(\mathbf{C}_{m,k}) = 0. \end{cases} \tag{6}$$

This statistic measures how broadly feature energy is distributed across directions. It is invariant to uniform feature rescaling; we use the same covariance convention throughout.

For each client, the candidate teacher is the modality with the lowest loss, $a_k = \arg\min_{m\in\mathcal{M}_k}(\ell_{m,k}, m)$, using the modality index to break ties. The weak set is $\mathcal{W}_k = \{w \in \mathcal{M}_k \setminus \{a_k\} : \ell_{w,k} - \ell_{a_k,k} > \log\rho\}$, where $\rho > 1$. For each $w \in \mathcal{W}_k$, the loss gap and rank deficit are defined together with the gate:

$$d_{w,k} = [\ell_{w,k} - \ell_{a_k,k} - \log\rho]_+, \quad \delta^r_{w,k} = \frac{[r_{a_k,k} - r_{w,k}]_+}{r_{a_k,k} + \epsilon_{\text{gate}}}, \\ \alpha_{w,k} = \mathbf{1}^{\text{gate}}_{w,k} g^d_{w,k} g^r_{w,k} g^q_{a_k,k}, \qquad \mathbf{1}^{\text{gate}}_{w,k} = \mathbf{1}\{d_{w,k} > 0, \delta^r_{w,k} \geq \delta_{\min}, q_{a_k,k} \geq q_{\min}\}. \tag{7}$$

Teacher confidence is $q_{a_k,k} = [(\mathbb{E}_i \max_c \sigma(\mathbf{z}_{a_k,i})_c - 1/C)/(1-1/C)]^1_0$, where C is the number of classes and σ is softmax. The clipped factors are $g^d_{w,k} = [1 - e^{-d_{w,k}/\tau_g}]^1_0$, $g^r_{w,k} = [\delta^r_{w,k}/\delta_{\min}]^1_0$, and $g^q_{a_k,k} = [(q_{a_k,k} - q_{\min})/(1-q_{\min})]^1_0$. We require $\delta_{\min} > 0$, $0 \leq q_{\min} < 1$, and $\tau_g > 0$. On an active gate, $g^r_{w,k} = 1$; rank therefore acts as a hard eligibility threshold, while loss and confidence modulate the weight continuously. All gate statistics are detached. The capacity penalty, distillation, and weak-pathway semantic penalty are gated; reconstruction and direct unimodal/shared supervision remain active otherwise. Confidence supplies a continuous reliability score; prototype collection uses separate sample-quality weights. Gate statistics are recomputed on each local training minibatch.

3.4 Capacity-Preserving Reallocation

CPR encourages weak-pathway capacity during local training by penalizing low private-coordinate variance and shared–private cross-covariance, while reconstructing the encoder feature. The capacity penalties are gated by collapse evidence; reconstruction and the supervised losses remain active when the gate is closed.

Local capacity preservation. Let $\mathcal{L}^k_{\text{cap}}$ denote the gated average of $\lambda_v \mathcal{L}_{\text{VC}}(\mathbf{p}_w) + \lambda_o \mathcal{L}_{\text{orth}}(\mathbf{s}_w, \mathbf{p}_w)$ over $w \in \mathcal{W}_k$, and let $\mathcal{L}^k_{\text{rec}}$ denote the reconstruction sum over available modalities. CPR is then

$$\mathcal{L}^k_{\text{CPR}} = \mathcal{L}^k_{\text{cap}} + \lambda_r \mathcal{L}^k_{\text{rec}}, \\ \mathcal{L}_{\text{VC}}(p) = d_p^{-1} \sum_j [\nu - \sqrt{\text{Var}(p_j) + \epsilon_0}]^2_+ + [d_p(d_p-1)]^{-1} \sum_{i\neq j} \text{Cov}(p_i, p_j)^2, \\ \mathcal{L}_{\text{orth}}(s,p) = (d_s d_p)^{-1} \| \text{Cov}(s,p) \|^2_F. \tag{8}$$

The reconstruction term is $\mathcal{L}^k_{\text{rec}} = \sum_{m\in\mathcal{M}_k} \|\widetilde{\mathbf{f}}_m - \text{sg}(\mathbf{f}_m)\|^2_2$. Following VICReg (Bardes et al., 2022), the variance term keeps private coordinates active, the covariance term discourages redundancy, and the cross-covariance term separates shared and private features. All covariances use centered minibatch features; ν sets the target private-coordinate standard deviation and ϵ_0 stabilizes the square root.

Federated realization. All modality-specific parameters follow Eq. (4) from a common identity-preserving initialization. CPR regularizes private coordinates locally, while FSA coordinates shared semantics. The capacity bounds in Appendix D.6 apply to the evaluated model through its covariance and reconstruction conditions, including after aggregation.

3.5 Dominance-Gated Transfer

Capacity becomes useful through predictive learning. DGT gives every modality direct label supervision and adds teacher guidance only when the gate is active. Specifically, $\mathcal{L}^k_{\text{uni}} = |\mathcal{M}_k|^{-1} \sum_{m\in\mathcal{M}_k} \text{CE}(\mathbf{z}_m, y)$ is the average unimodal cross-entropy, while $\mathcal{L}^k_{\text{dist}} = \max(1, |\mathcal{W}_k|)^{-1} \sum_{w\in\mathcal{W}_k} \alpha_{w,k} q_{a_k,k} \tau_d^2 \,\text{KL}(\sigma(\mathbf{z}_{a_k}/\tau_d) \,\|\, \sigma(\mathbf{z}_w/\tau_d))$ transfers the reliable teacher distribution to the weak pathways. Then

$$\mathcal{L}^k_{\text{DGT}} = \lambda_u \mathcal{L}^k_{\text{uni}} + \lambda_d \mathcal{L}^k_{\text{dist}}. \tag{9}$$

Here $\mathbf{z}_m$ denotes unimodal logits. The KL term follows standard distillation interfaces (Hinton et al., 2015; Li & Wang, 2019; Han et al., 2026), but remains client-local and conditional; direct supervision is always retained. With teacher targets frozen and KD aligned with the supervised gradient, the descent analysis quantifies the added predictive gain (Appendix D.7).

3.6 FEDERATED SEMANTIC ANCHORING

FSA connects the shared coordinates to a common class structure. The server maintains a modality-invariant prototype bank $\mathbf{P}^t = \{\mathbf{P}_c^t\}_{c=1}^C$. Clients use these prototypes as semantic references while contributing class summaries from their local data. This prototype interface follows the communication-efficient semantic coordination principle of FedProto (Tan et al., 2022); unlike modality-selection strategies such as BMSFed (Fan et al., 2024), it anchors the shared encoder space directly. The quality and support weights address multimodal agreement and class-skewed clients.

Collecting and updating semantic anchors. For sample i, the shared feature is $\bar{\mathbf{s}}_i = |\mathcal{M}_{k,i}|^{-1} \sum_{m \in \mathcal{M}_{k,i}} \mathrm{Norm}(\mathbf{s}_{m,i})$. Its quality is the confidence–agreement product $q_i = \max_c \sigma(\mathbf{z}_i)_c [|\mathcal{M}_{k,i}|^{-1} \sum_{m \in \mathcal{M}_{k,i}} \cos(\mathbf{s}_{m,i}, \bar{\mathbf{s}}_i)]_0^1$. For class c, client k sends the quality mass $Q_{c,k}^t = \sum_{i:y_i=c} q_i$ and the normalized weighted mean $\mathbf{c}_{c,k}^{t+1} = \mathrm{Norm}(\sum_{i:y_i=c} q_i \bar{\mathbf{s}}_i / Q_{c,k}^t)$ when $Q_{c,k}^t > 0$. Clients with positive class quality form $\mathcal{S}_t^c$. The server updates the raw prototype state using quality mass and class support. We set the round-specific update rate to $\eta_c^t = (1-\mu)\Xi_c^t$, where μ is prototype momentum and n_0 controls the class-support scale:

$$\begin{aligned}
\widehat{\mathbf{P}}_c^{t+1} &= \mathrm{Norm}\left(\sum_{k \in \mathcal{S}_t^c} \frac{Q_{c,k}^t}{\sum_{j \in \mathcal{S}_t^c} Q_{c,j}^t} \mathbf{c}_{c,k}^{t+1} \right), \\
\Xi_c^t &= \frac{\sum_k Q_{c,k}^t}{\sum_k n_{c,k}^t} \left(1 - e^{-\sum_k n_{c,k}^t / n_0} \right), \\
\widetilde{\mathbf{P}}_c^{t+1} &= (1-\eta_c^t)\widetilde{\mathbf{P}}_c^t + \eta_c^t \widehat{\mathbf{P}}_c^{t+1}.
\end{aligned} \tag{10}$$

The server retains the previous state if class support or total quality is zero, if $\Xi_c^t < \Xi_{\min}$, or if the proposed vector is zero before normalization. The raw state is retained for temporal updates, and the normalized copy $\mathbf{P}_c^{t+1} = \mathrm{Norm}(\widetilde{\mathbf{P}}_c^{t+1})$ is used in cosine losses. This moderates prototype changes supported by little data.

Learning discriminative shared features. The shared classifier is supervised by $\mathcal{L}_{\text{shared}}^k = |\mathcal{M}_k|^{-1} \sum_{m \in \mathcal{M}_k} \mathrm{CE}(h_s(\mathbf{s}_m), y)$. Let $\mathcal{L}_{\text{align}}^k$ be the average of $1 - \cos(\mathbf{s}_m, \mathbf{s}_{m'})$ over modality pairs, set to zero when fewer than two modalities are available. The semantic term is $\mathcal{L}_{\text{sem}}^k = \max(1, |\mathcal{W}_k|)^{-1} \sum_{w \in \mathcal{W}_k} \alpha_{w,k} [\lambda_a \mathcal{L}_{\text{align}}^k + \lambda_p (1 - \cos(\mathbf{s}_w, \mathbf{P}_y^t)) + \lambda_c \mathcal{R}_{\text{cov}}(\mathbf{s}_w, y; \mathbf{P}^t)]$. FSA combines them as

$$\mathcal{L}_{\text{FSA}}^k = \lambda_s \mathcal{L}_{\text{shared}}^k + \mathcal{L}_{\text{sem}}^k. \tag{11}$$

For supported classes, $\mathcal{R}_{\text{cov}}(\mathbf{s}, y; \mathbf{P}^t) = -a_y \log(e^{u_y} / \sum_{c \in \mathcal{C}_t} e^{u_c})$, with $u_c = \mathrm{Norm}(\mathbf{s})^\top \mathbf{P}_c^t / \tau_p$ and $a_y = (N_y^t + \epsilon_N)^{-1/2} / \sum_{c \in \mathcal{C}_t} (N_c^t + \epsilon_N)^{-1/2}$. Here $\mathcal{C}_t$ contains valid prototype classes, N_c^t is the server-maintained cumulative class support broadcast with the bank, and $\epsilon_N > 0$ stabilizes inverse support. FSA thus gives shared features a common class geometry while CPR preserves private variation.

3.7 JOINT OBJECTIVE AND FEDERATED OPTIMIZATION

The local objective at round t, initialized from the broadcast global model and prototype bank, is

$$\boxed{F_k(\boldsymbol{\theta}; \mathbf{P}^t) = \mathcal{L}_{\text{task}}^k + \mathcal{L}_{\text{CPR}}^k + \mathcal{L}_{\text{DGT}}^k + \mathcal{L}_{\text{FSA}}^k.} \tag{12}$$

The task loss supervises fused predictions. CPR regularizes private capacity locally, DGT combines direct supervision with gated transfer, and FSA organizes shared semantics. Aggregated-model diagnostics track the capacity retained across rounds. Clients return updated parameters and class-level prototype summaries, without public proxy data or sample-level logit exchange. The server averages parameters using Eq. (4) and updates prototypes using Eq. (10). The resulting model initializes the next round. Appendix C.5 gives the computation, memory, and communication costs of the added operations.

Table 1: Main results under IID and non-IID partitions (mean$_{SD}$, %). Green: best; yellow: second. Classification: fusion accuracy; MIMIC-IV: mean AUROC. †: personalized-client evaluation; ‡: 20% public proxy allocation (Appendix C.1).

	CREMA-D		ModelNet40		CrisisMMD		MIMIC-IV	
Method	IID	N-IID	IID	N-IID	IID	N-IID	IID	N-IID
FedAvg	48.83$_{(1.37)}$	45.39$_{(1.63)}$	78.93$_{(0.71)}$	78.53$_{(0.93)}$	75.46$_{(0.54)}$	74.25$_{(0.72)}$	78.56$_{(0.81)}$	75.13$_{(1.09)}$
FedProx	44.95$_{(0.93)}$	45.28$_{(1.18)}$	78.53$_{(0.58)}$	78.32$_{(0.74)}$	75.46$_{(0.47)}$	74.25$_{(0.61)}$	78.78$_{(0.73)}$	75.46$_{(0.95)}$
FedAdam	53.16$_{(1.48)}$	49.72$_{(1.87)}$	**86.18**$_{(0.79)}$	78.36$_{(1.08)}$	76.44$_{(0.66)}$	73.76$_{(0.83)}$	80.31$_{(0.94)}$	77.28$_{(1.27)}$
SCAFFOLD	37.40$_{(1.72)}$	37.74$_{(2.11)}$	57.54$_{(1.32)}$	58.47$_{(1.56)}$	73.31$_{(0.91)}$	73.94$_{(1.06)}$	76.45$_{(1.15)}$	72.84$_{(1.48)}$
FedProto	50.39$_{(1.35)}$	41.18$_{(1.74)}$	78.16$_{(0.67)}$	76.38$_{(0.89)}$	75.06$_{(0.53)}$	74.25$_{(0.68)}$	79.14$_{(0.78)}$	74.89$_{(1.04)}$
FedMSplit†	39.63$_{(1.41)}$	36.35$_{(1.96)}$	71.87$_{(0.94)}$	70.24$_{(1.21)}$	74.51$_{(0.78)}$	67.34$_{(0.99)}$	75.68$_{(1.03)}$	72.05$_{(1.41)}$
FedAvg-RMD	42.84$_{(0.96)}$	42.84$_{(1.32)}$	74.39$_{(0.61)}$	73.54$_{(0.86)}$	74.88$_{(0.45)}$	74.16$_{(0.59)}$	77.46$_{(0.68)}$	74.51$_{(0.88)}$
FedPMR	52.05$_{(1.23)}$	45.73$_{(1.54)}$	77.51$_{(0.83)}$	75.16$_{(1.07)}$	74.47$_{(0.63)}$	73.58$_{(0.77)}$	79.62$_{(0.86)}$	76.21$_{(1.16)}$
BMSFed	46.50$_{(1.58)}$	46.73$_{(1.79)}$	79.42$_{(0.85)}$	77.18$_{(1.09)}$	75.22$_{(0.65)}$	73.89$_{(0.83)}$	78.39$_{(0.91)}$	75.68$_{(1.23)}$
FedAMB†‡	31.70$_{(1.89)}$	24.18$_{(2.43)}$	62.13$_{(1.18)}$	49.66$_{(1.53)}$	73.34$_{(0.97)}$	70.83$_{(1.21)}$	74.12$_{(1.32)}$	69.31$_{(1.71)}$
FedSBR	**55.39**$_{(0.82)}$	**50.39**$_{(0.97)}$	86.16$_{(0.43)}$	**84.20**$_{(0.62)}$	**77.64**$_{(0.41)}$	**75.43**$_{(0.52)}$	**81.89**$_{(0.57)}$	**79.57**$_{(0.68)}$

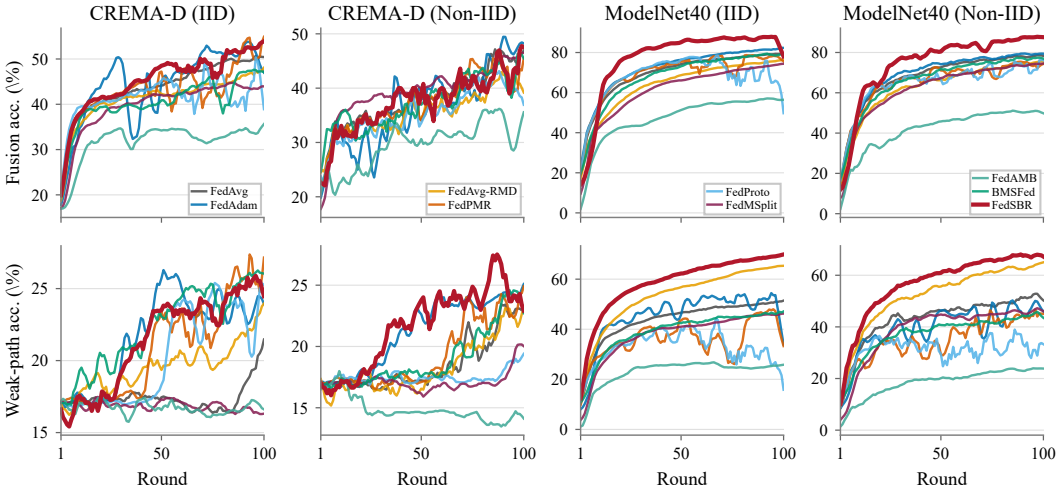

Figure 3: Fusion and weak-pathway validation accuracy over 100 communication rounds on CREMA-D and ModelNet40 under IID and non-IID partitions.

4 Experiments

4.1 Experimental Setup

Datasets and tasks. We cover audio-visual CREMA-D (Cao et al., 2014), multi-view ModelNet40 (Wu et al., 2015), social-media CrisisMMD (Alam et al., 2018), and clinical MIMIC-IV (Johnson et al., 2023). Dataset splits, inputs, and task definitions are given in Appendix C.1.

Federated protocol. Unless stated otherwise, each dataset uses 10 clients, 5 clients per round, 100 rounds, 2 local epochs, and batch size 128. IID clients have uniform class proportions; non-IID clients follow a Dirichlet partition with $\alpha = 0.5$. All methods share splits, schedules, preprocessing, and checkpoint rules. FedAMB alone uses a deterministic 20% public unlabeled proxy set. Details are listed in Appendix C.1. Appendix Fig. 5 visualizes the realized CREMA-D class allocation, exposing differences in local class support beyond the nominal partition parameter.

Models and optimization. Each modality uses an encoder followed by concatenation and a linear fusion classifier. FedSBR divides each 512-dimensional feature into 256 shared and 256 private coordinates. Architectural and optimization details are given in Appendix C.1.

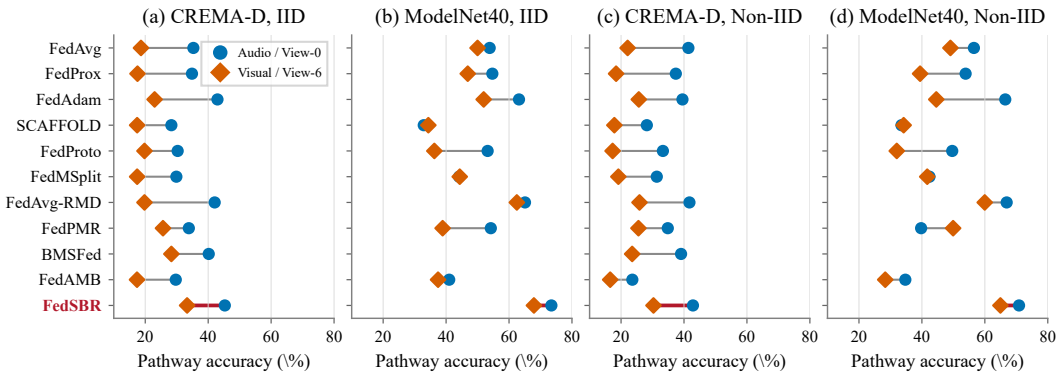

Figure 4: Per-modality pathway accuracy on CREMA-D and ModelNet40 under IID and non-IID partitions.

Baselines and evaluation. We compare FedAvg (McMahan et al., 2017), FedProx (Li et al., 2020), FedAdam (Reddi et al., 2021), and SCAFFOLD (Karimireddy et al., 2020) as generic federated optimizers. FedProto (Tan et al., 2022) and FedMSplit (Chen & Zhang, 2022) represent representation-heterogeneity methods. FedAvg-RMD (Alfasly et al., 2022), FedPMR (Fan et al., 2023), BMSFed (Fan et al., 2024), and FedAMB (Han et al., 2026) cover modality balancing, multimodal selection, and dominance-aware distillation. We set the FedProx coefficient to 0.01, the RMD probability to 0.5, and the FedPMR prototype coefficient to 1.0. FedAdam uses server learning rate 0.01, $(\beta_1, \beta_2) = (0.9, 0.99)$, and $\tau = 10^{-3}$. FedMSplit and FedAMB use mean personalized-client test performance. All other methods use the global model. For the three classification benchmarks, we report fusion and pathway top-1 accuracy, macro-F1, and per-class recall. MIMIC-IV uses mean AUROC, with task-level AUROC/AUPRC in the appendix. Representation contraction is measured by participation rank in Eq. (6). We also report the fraction of samples with negative weak-modality Shapley contribution, using negative cross-entropy as the utility. Unless otherwise stated, each reported mean is computed over three independent runs with seeds 42, 43, and 44. The uncertainty is the sample standard deviation across these runs. Appendix C.2 details evaluation units and matched-data controls.

4.2 Overall Comparison

Table 1 reports the primary test metric for each dataset and partition. FedSBR improves over FedAvg on all four datasets in both partitions, with gains of 6.56/5.00 points on CREMA-D, 7.23/5.67 on ModelNet40, 2.18/1.18 on CrisisMMD, and 3.33/4.44 on MIMIC-IV for IID/non-IID settings. FedAdam is slightly higher than FedSBR on ModelNet40 under IID (86.18 versus 86.16), leaving FedSBR within 0.02 points of the best entry. The convergence and pathway plots show that this fused performance is accompanied by a stronger weak-pathway trajectory in both benchmarks. Round-level computation is reported in Appendix Table 8.

4.3 Weak-Pathway Diagnostics

Fusion accuracy can conceal the degradation of an individual pathway. On CREMA-D, Table 2 shows that FedSBR improves weak-pathway macro-F1 and participation rank while reducing the fraction of negative Shapley contributions. Under IID, visual macro-F1 increases from 12.67 to 25.14, participation rank from 2.88 to 3.68, and the negative-contribution fraction decreases from 0.43 to 0.32. Under non-IID, the corresponding changes are 12.65 to 19.71, 2.59 to 4.08, and 0.45 to 0.38. The per-class recall plot in Appendix Fig. 7 shows that this macro-recall improvement comes from redistribution across classes, with gains concentrated on fear, happiness, and neutral. Appendix Table 11 pairs layer-wise rank with within- versus between-class similarity to assess capacity and class separation jointly. Appendix Fig. 6 further probes the trained fusion classifier by attenuating each modality's logit contribution without retraining, complementing the pathway and Shapley diagnostics.

Table 2: CREMA-D collapse diagnostics (mean$_{SD}$). F/V: fusion/visual; Neg: negative Shapley fraction; Rank: participation rank.

	IID				Non-IID ($\alpha = 0.5$)			
Method	F-F1	V-F1	V-Neg↓	V-Rank↑	F-F1	V-F1	V-Neg↓	V-Rank↑
FedAvg	48.00$_{(1.24)}$	12.67$_{(0.81)}$	0.43$_{(0.03)}$	2.88$_{(0.24)}$	43.17$_{(1.62)}$	12.65$_{(0.97)}$	0.45$_{(0.04)}$	2.59$_{(0.29)}$
FedAdam	50.44$_{(1.56)}$	14.91$_{(1.14)}$	0.53$_{(0.04)}$	2.54$_{(0.31)}$	44.43$_{(1.38)}$	16.93$_{(1.53)}$	0.45$_{(0.03)}$	2.20$_{(0.23)}$
FedProto	49.16$_{(1.41)}$	11.35$_{(0.69)}$	0.48$_{(0.04)}$	2.28$_{(0.19)}$	39.03$_{(1.87)}$	5.66$_{(0.58)}$	0.54$_{(0.05)}$	3.01$_{(0.41)}$
FedMSplit	37.09$_{(1.68)}$	8.38$_{(0.93)}$	0.52$_{(0.05)}$	3.51$_{(0.38)}$	31.60$_{(2.13)}$	11.82$_{(1.27)}$	0.48$_{(0.04)}$	3.18$_{(0.34)}$
FedAvg-RMD	41.57$_{(1.03)}$	17.18$_{(1.38)}$	0.44$_{(0.03)}$	2.13$_{(0.26)}$	40.50$_{(1.41)}$	18.71$_{(1.71)}$	0.50$_{(0.04)}$	1.95$_{(0.21)}$
FedPMR	51.92$_{(1.34)}$	21.73$_{(1.62)}$	0.43$_{(0.03)}$	1.98$_{(0.17)}$	41.16$_{(1.56)}$	16.11$_{(1.38)}$	0.43$_{(0.03)}$	2.76$_{(0.26)}$
BMSFed	45.11$_{(1.49)}$	**25.14**$_{(1.84)}$	0.37$_{(0.02)}$	2.02$_{(0.36)}$	43.65$_{(1.74)}$	16.11$_{(1.23)}$	0.45$_{(0.04)}$	2.43$_{(0.38)}$
FedAMB	25.84$_{(2.07)}$	10.01$_{(0.87)}$	0.47$_{(0.04)}$	3.28$_{(0.29)}$	15.36$_{(2.52)}$	6.07$_{(0.74)}$	0.55$_{(0.05)}$	2.01$_{(0.33)}$
FedSBR	**52.67**$_{(0.93)}$	**25.14**$_{(1.46)}$	**0.32**$_{(0.02)}$	**3.68**$_{(0.22)}$	**47.74**$_{(1.08)}$	**19.71**$_{(1.51)}$	**0.38**$_{(0.03)}$	**4.08**$_{(0.18)}$

MIMIC-IV provides a complementary modality-scaling test. We aggregate all $\binom{6}{q}$ subsets of the six EHR groups for each modality count q under the non-IID partition. FedSBR stays near 79% mean AUROC and retains 0.86 of the unimodal rank at five modalities. FedAvg and FedPMR fall to 0.47 and 0.63, respectively, as summarized in Appendix Table 9. Appendix C.11 additionally varies client count, label heterogeneity, and modality missingness on CREMA-D. These seed-42, low-budget sweeps characterize early-training sensitivity; the main comparisons use three seeds and 100 rounds.

4.4 Component Ablations

Table 3 isolates CPR, DGT, and FSA under non-IID partitions. Removing CPR reduces CREMA-D fusion and weak-pathway accuracy by 5.66 and 5.55 points, respectively, while removing DGT reduces ModelNet40 fusion and View-6 accuracy by 5.59 and 25.76 points. Removing FSA has a smaller effect on the weak pathway but consistently lowers fusion accuracy and rank on both datasets. The complete model achieves the highest fusion score, weak-pathway score, and rank in both cases. Their joint effect is particularly clear on ModelNet40: without DGT, weak-pathway accuracy is 40.19, compared with 45.46 for the structure-only control and 65.95 for the full model. Predictive transfer therefore makes preserved capacity and shared semantics useful together. Ranks here use the balanced validation subset; Appendix Table 16 pairs test-checkpoint rank with Shapley diagnostics.

Table 3: Component ablation under non-IID. Weak: Visual on CREMA-D; View-6 on ModelNet40.

	CPR	DGT	FSA	Fusion	Weak	Rank
CREMA-D	✓	✓	✓	**50.39**	**28.30**	**3.97**
	–	✓	✓	44.73	22.75	3.14
	✓	–	✓	47.95	27.75	3.51
	✓	✓	–	48.50	27.08	3.62
	–	–	–	43.17	20.53	2.67
ModelNet40	✓	✓	✓	**84.20**	**65.95**	**7.43**
	–	✓	✓	79.42	64.51	5.93
	✓	–	✓	78.61	40.19	6.04
	✓	✓	–	79.78	65.44	6.78
	–	–	–	78.24	45.46	5.09

5 Conclusion

We presented FedSBR, an encoder-level shared–private basis framework for multimodal federated learning. Local capacity regularization, gated predictive supervision, and reliability-weighted semantic anchoring jointly target weak pathways while retaining standard parameter averaging. Identity-preserving initialization preserves the initial predictor. Three-seed experiments on CREMA-D, ModelNet40, CrisisMMD, and MIMIC-IV show improvements over FedAvg in the primary test metric and competitive fusion performance. Pathway accuracy, participation rank, modality contribution, and class-conditional diagnostics reveal weak-pathway behavior beyond fused performance.

AI USE STATEMENT

We used generative AI tools to assist with writing and language refinement, information retrieval and related-work discovery, and consistency checks during revision. We carefully checked the AI-assisted content for factual errors, fabricated references, and unsupported claims, and revised it accordingly. We take full responsibility for the final manuscript.

ETHICS STATEMENT

The MIMIC-IV experiment uses de-identified clinical records accessed under the dataset's data-use requirements; no new patient contact or intervention is performed. The study does not make clinical deployment or treatment claims. Potential cohort, label, and demographic biases should be assessed before any clinical use, and the complete cohort construction and preprocessing record must accompany a reproducibility release.

REPRODUCIBILITY STATEMENT

The main tables report means and sample standard deviations over seeds 42, 43, and 44, with protocols and definitions given in Sec. 4 and the appendix.

REFERENCES

Firoj Alam, Ferda Ofli, and Muhammad Imran. CrisisMMD: Multimodal twitter datasets from natural disasters. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, pp. 465–473, 2018. doi: 10.1609/icwsm.v12i1.14983. URL https://doi.org/10.1609/icwsm.v12i1.14983.

Saghir Alfasly, Jian Lu, Chen Xu, and Yuru Zou. Learnable irrelevant modality dropout for multimodal action recognition on modality-specific annotated videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20208–20217, 2022. doi: 10.1109/CVPR52688.2022.01957. URL https://doi.org/10.1109/CVPR52688.2022.01957.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244. URL https://doi.org/10.1109/TAFFC.2014.2336244.

Abhra Chaudhuri, Anjan Dutta, Tu Bui, and Serban Georgescu. A closer look at multimodal representation collapse. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 7555–7577. PMLR, 2025. URL https://proceedings.mlr.press/v267/chaudhuri25a.html.

Jiayi Chen and Aidong Zhang. FedMSplit: Correlation-adaptive federated multi-task learning across multimodal split networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 87–96, 2022. doi: 10.1145/3534678.3539384. URL https://doi.org/10.1145/3534678.3539384.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. PMR: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20029–20038, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Fan_PMR_Prototypical_Modal_Rebalance_for_Multimodal_Learning_CVPR_2023_paper.html.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Fushuo Huo, Jinyu Chen, and Song Guo. Overcome modal bias in multi-modal federated learning via balanced modality selection. In *Computer Vision*

– *ECCV 2024*, pp. 178–195, 2024. doi: 10.1007/978-3-031-73004-7_11. URL https://doi.org/10.1007/978-3-031-73004-7_11.

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

Nuno C. Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *Computer Vision – ECCV 2018*, pp. 106–121, 2018. doi: 10.1007/978-3-030-01237-3_7. URL https://doi.org/10.1007/978-3-030-01237-3_7.

Seungjin Han, Juyeob Lee, Sangmin Lee, and Eunil Park. FedAMB: Adaptive modality balancing for dominance-robust multimodal federated distillation. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*, 2026. URL https://ijcai-preprints.s3.us-west-1.amazonaws.com/2026/4335.pdf. Accepted; proceedings pagination unavailable as of 2026-08-26.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. doi: 10.48550/arXiv.1503.02531. URL https://arxiv.org/abs/1503.02531.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? (provably). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9226–9259. PMLR, 2022. URL https://proceedings.mlr.press/v162/huang22e.html.

Minyoung Huh, Hossein Mobahi, Richard Zhang, Brian Cheung, Pulkit Agrawal, and Phillip Isola. The low-rank simplicity bias in deep networks. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=bCiNWDmlY2.

Fushuo Huo, Wenchao Xu, Jingcai Guo, Haozhao Wang, and Song Guo. C2KD: Bridging the modality gap for cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16006–16015, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huo_C2KD_Bridging_the_Modality_Gap_for_Cross-Modal_Knowledge_Distillation_CVPR_2024_paper.html.

Adrián Javaloy, Maryam Meghdadi, and Isabel Valera. Mitigating modality collapse in multimodal VAEs via impartial optimization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9938–9964. PMLR, 2022. URL https://proceedings.mlr.press/v162/javaloy22a.html.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo Anthony Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023. doi: 10.1038/s41597-022-01899-x. URL https://doi.org/10.1038/s41597-022-01899-x.

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*, pp. 5132–5143. PMLR, 2020. URL https://proceedings.mlr.press/v119/karimireddy20a.html.

Daliang Li and Junpu Wang. FedMD: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*, 2019. doi: 10.48550/arXiv.1910.03581. URL https://arxiv.org/abs/1910.03581. Preprint.

Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22214–22224, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Li_Boosting_Multi-Modal_Model_Performance_with_Adaptive_Gradient_Modulation_ICCV_2023_paper.html.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=KmykpuSrjcq.

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

Tao Lin, Lingjing Kong, Sebastian U. Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2351–2363, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/18df51b97ccd68128e994804f3eccc87-Abstract.html.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. SMIL: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2302–2310, 2021. doi: 10.1609/aaai.v35i3.16330. URL https://doi.org/10.1609/aaai.v35i3.16330.

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

Duong Nguyen, Nghia Hoang, Thanh Trung Huynh, Quoc Viet Hung Nguyen, and Phi Le Nguyen. Learning reconfigurable representations for multimodal federated learning with missing data. In *Advances in Neural Information Processing Systems*, volume 38, pp. 32613–32650, 2025. doi: 10.52202/085713-0968. URL https://doi.org/10.52202/085713-0968.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8237, 2022. doi: 10.1109/CVPR52688.2022.00806. URL https://doi.org/10.1109/CVPR52688.2022.00806.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and H. Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=LkFG3lB13U5.

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

Haozhao Wang, Yichen Li, Wenchao Xu, Ruixuan Li, Yufeng Zhan, and Zhigang Zeng. DaFKD: Domain-aware federated knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20412–20421, 2023a. URL https://openaccess.thecvf.com/content/CVPR2023/html/Wang_DaFKD_Domain-Aware_Federated_Knowledge_Distillation_CVPR_2023_paper.html.

Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. In *International Conference on Learning Representations*, 2020a. URL https://arxiv.org/abs/2002.06440.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15878–15887, 2023b. doi: 10.1109/CVPR52729.2023.01524. URL https://doi.org/10.1109/CVPR52729.2023.01524.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12692–12702, 2020b. doi: 10.1109/CVPR42600.2020.01271. URL https://doi.org/10.1109/CVPR42600.2020.01271.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1912–1920, 2015. doi: 10.1109/CVPR.2015.7298801. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Wu_3D_ShapeNets_A_2015_CVPR_paper.html.

Dezhong Yao, Wanning Pan, Yutong Dai, Yao Wan, Xiaofeng Ding, Hai Jin, Zheng Xu, and Lichao Sun. FedGKD: Toward heterogeneous federated learning via global knowledge distillation. *IEEE Transactions on Computers*, 73(1):3–17, 2024. doi: 10.1109/TC.2023.3315066. URL https://doi.org/10.1109/TC.2023.3315066.

Qiying Yu, Yang Liu, Yimu Wang, Ke Xu, and Jingjing Liu. Multimodal federated learning via contrastive representation ensemble. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Hnk1WRMAYqq.

Haoran Zhang, Dongjun Kim, Seohyeon Cha, and Haris Vikalo. Fedrot-lora: Mitigating rotational misalignment in federated lora. *arXiv preprint arXiv:2602.23638*, 2026. URL https://arxiv.org/abs/2602.23638.

Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27456–27466, 2024. doi: 10.1109/CVPR52733.2024.02592. URL https://doi.org/10.1109/CVPR52733.2024.02592.

Yuchen Zhao, Payam Barnaghi, and Hamed Haddadi. Multimodal federated learning on IoT data. In *IEEE/ACM Seventh International Conference on Internet-of-Things Design and Implementation*, pp. 43–54, 2022. doi: 10.1109/IoTDI54339.2022.00011. URL https://doi.org/10.1109/IoTDI54339.2022.00011.

Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12878–12889. PMLR, 2021. URL https://proceedings.mlr.press/v139/zhu21b.html.

APPENDIX

The appendix provides additional theoretical, empirical, implementation, and responsible-use evidence for FedSBR. To make the supplementary material easier to navigate, Table 4 indexes the second-level headings and summarizes the main contents of each appendix.

Table 4: Organization of the appendix and its main contents.

Appendix	Main Contents
Appendix A	• Multimodal collapse and balancing (A.1) • Federated multimodal heterogeneity (A.2) • Federated distillation and prototypes (A.3)
Appendix B	• Reproducible configuration (B.1) • Communication and evaluation (B.2) • Round-level algorithm (B.3)
Appendix C	• Experimental details (C.1) • Comparison protocol and evaluation units (C.2) • FedAvg+EBR comparison (C.3) • Convergence (C.4) • Computation cost (C.5) • MIMIC-IV modality scaling (C.6) • Client heterogeneity (C.7) • Modality attenuation (C.8) • Layer-wise geometry (C.9) • Class-conditional diagnostics (C.10) • Sensitivity sweeps (C.11)
Appendix D	• Notation and assumptions (D.1) • Centralized baseline (D.2) • Exact FedAvg decomposition (D.3) • Cross-client cancellation (D.4) • Update cancellation and rank collapse (D.5) • Capacity-preserving reallocation (D.6) • Dominance-gated transfer (D.7) • Federated semantic anchoring (D.8) • Combined FedSBR certificate (D.9) • Theory-to-experiment connections (D.10)
Appendix E	• Social impact considerations (E)
Appendix F	• Limitations and scope (F)

A RELATED WORK

A.1 MULTIMODAL COLLAPSE AND MODALITY BALANCING

Multimodal models can reach a strong fused score while learning a representation that is effectively controlled by one modality. Early work characterizes modality dominance as unequal optimization speed, gradient magnitude, and modality competition (Wang et al., 2020b; Huang et al., 2022); subsequent methods modify the optimization signal through on-the-fly gradient modulation, adaptive gradient modulation, or alternating unimodal adaptation (Peng et al., 2022; Li et al., 2023; Zhang et al., 2024). Prediction-level and representation-level remedies, including prototypical modal rebalancing, irrelevant-modality dropout, and impartial optimization, improve the weak stream in particular regimes (Fan et al., 2023; Alfasly et al., 2022; Javaloy et al., 2022). The optimization-based approaches establish that dominance is a training-dynamics problem, yet they mainly regulate losses, gradients, or modality participation. A complementary mechanistic account attributes collapse to cross-modal polysemantic interference, while the low-rank simplicity bias explains why a small number of directions can absorb most of the usable capacity (Chaudhuri et al., 2025; Huh et al., 2023); variance–covariance regularization offers a general tool for preventing such degeneracy (Bardes et al., 2022). FedSBR adopts a representation-level perspective: CPR protects a private subspace, DGT supplies predictive pressure only when a teacher is reliable, and the evaluation reports rank, attenuation, and class-wise recall in addition to fused accuracy.

A.2 FEDERATED MULTIMODAL LEARNING UNDER HETEROGENEITY

Federated multimodal learning adds client drift, partial participation, modality availability differences, and label skew to the optimization imbalance already present in centralized training. FedM-

Split distributes modality-specific computation across correlated split networks (Chen & Zhang, 2022), while CREAM-FL and FedMultimodal provide contrastive aggregation and benchmark protocols for clients whose modalities and data distributions differ (Yu et al., 2023; Feng et al., 2023). Other studies target sensing or deployment settings in which only a subset of modalities is observed, including multimodal IoT learning and reconfigurable federated representations (Zhao et al., 2022; Nguyen et al., 2025). Shared–specific feature modelling and SMIL similarly reconstruct common factors while coping with severe missingness (Wang et al., 2023b; Ma et al., 2021). Prior federated multimodal work clarifies how to exchange or reconstruct multimodal information without centralizing records, but missingness and collapse are different: a modality can be present on every client and still lose effective directions during local fusion and server averaging. FedSBR keeps the standard federated interface and addresses the remaining capacity problem by separating shared semantic coordinates, which can be aligned globally, from private residual coordinates, which are explicitly regularized locally. The distinction also explains why pathway-level diagnostics are needed to complement aggregate federated accuracy.

A.3 Federated distillation, prototypes, and parameter coordination

A broad federated literature reduces statistical or architectural mismatch through proximal objectives, control variates, adaptive server updates, distillation, or compact semantic summaries. FedAvg is the weighted-averaging baseline; FedProx, SCAFFOLD, and FedAdam address client drift and update noise (McMahan et al., 2017; Li et al., 2020; Karimireddy et al., 2020; Reddi et al., 2021). Distillation-based methods exchange predictive information across heterogeneous models or clients through logits, ensembles, global teachers, or domain-aware teachers (Li & Wang, 2019; Lin et al., 2020; Zhu et al., 2021; Yao et al., 2024; Wang et al., 2023a); modality-specific distillation and cross-modal gap bridging apply the same principle within multimodal networks (Garcia et al., 2018; Huo et al., 2024). Prototype learning instead communicates class-level feature summaries, reducing the need for shared examples (Tan et al., 2022; Li et al., 2021), while matched averaging and rotational alignment mitigate incompatible parameter coordinates (Wang et al., 2020a; Zhang et al., 2026). Recent multimodal federated methods combine selection or adaptive balancing with distillation (Fan et al., 2024; Han et al., 2026), but a stronger modality is not necessarily a safe teacher when its confidence is miscalibrated or its class support is narrow. FedSBR therefore combines a reliability-gated teacher with a prototype bank and a shared–private basis: transfer is activated only under a measured loss gap, rank deficit, and confidence threshold, while the private block is regularized locally rather than aligned to class prototypes. Both blocks are still subject to parameter averaging, which need not preserve private capacity. The resulting design places FedSBR alongside optimizer-, distillation-, and prototype-based methods as a representation-capacity intervention rather than a replacement for them.

B Implementation Details and Algorithm Flow

B.1 Reproducible configuration

All datasets use ten clients, five clients sampled without replacement in each communication round, 100 rounds, two local epochs, and batch size 128. The IID split preserves the global class proportions; the non-IID split samples client class proportions from a Dirichlet distribution with $\alpha = 0.5$. Unless a baseline requires a special optimizer, clients use SGD with learning rate 10^{-3}, momentum 0.9, and weight decay 10^{-4}. We use the three fixed seeds 42, 43, and 44 and report the sample standard deviation. Checkpoints are selected by validation fusion accuracy, or by mean validation AUROC for MIMIC-IV, and the selected checkpoint is evaluated once on the held-out test set.

The encoder dimension is split as $d = d_s + d_p$ with an identity-preserving initialization $R_m B_m = I$. The gate uses the thresholds and temperatures defined in Eq. (7); all gate statistics are detached before back-propagation. CPR uses the variance floor, covariance penalty, reconstruction term, and shared–private orthogonality term from Eq. (8); DGT combines the unimodal cross-entropy and temperature-scaled distillation losses in Eq. (9); FSA uses the shared classifier, cosine alignment, prototype attraction, and class-coverage term in Eq. (11). A client sends updated parameters, quality-weighted class prototypes, their quality masses, and class sample counts, as defined in Eq. (10). No raw example, pixel, waveform, note, or sample-level logit is sent.

Reference implementation defaults. Table 5 lists the common defaults exposed by the CREMA-D, ModelNet40, and CrisisMMD training entry points. The symbols map to the corresponding loss components in Sec. 3; these are implementation defaults, while the experimental protocol is specified above.

Table 5: Reference implementation defaults for loss coefficients and temperatures.

Role	Parameter	Default
Private variance–covariance	λ_v	0.05
Shared–private orthogonality	λ_o	0.01
Feature reconstruction	λ_r	0.05
Unimodal supervision	λ_u	0.30
Distillation	λ_d	0.30
Shared supervision	λ_s	0.20
Cross-modal alignment	λ_a	0.10
Prototype attraction	λ_p	0.10
Class coverage	λ_c	0.10
Dominance threshold	ρ	1.20
Loss-gate temperature	τ_g	0.25
Distillation temperature	τ_d	2.00
Prototype temperature	τ_p	0.20
Private standard-deviation target	ν	0.50
Prototype momentum	μ	0.90
Class-support scale	n_0	32
Variance stabilizer	ϵ_0	10^{-4}

B.2 Communication and evaluation

Evaluation units and resource allocation. Global-model entries evaluate the server model; FedMSplit and FedAMB use personalized-client evaluation. Appendix C.2 specifies these units and the private/public data allocation, including the matched comparison in Table 15.

Prototype state and pathway metrics. For modality block m, the server uses the sample-count weights in Eq. (4); a block with no participating client keeps its previous value. For each supported class, client k returns $(\mathbf{c}_{c,k}^{t+1}, Q_{c,k}^{t}, n_{c,k}^{t})$: the normalized quality-weighted class prototype, quality mass, and sample count from Eq. (10). Class counts are also reported when quality mass is zero; no normalized class vector is formed in that case. These quantities supply both the aggregation weights and the support-dependent update rate. The server retains the raw prototype state for temporal updates and broadcasts its normalized copy, validity mask, and cumulative class supports with the next model. Prototype values and supports start at zero with all classes invalid. Prototype attraction is evaluated only for valid target classes; coverage additionally requires at least two valid classes. Accepted observations activate their class entries. These masks keep uninitialized vectors out of cosine losses. The normalization analysis assumes a positive norm floor on the evaluated raw states. Pathway metrics are computed by replacing the other modality with zero input, while rank uses the scale-invariant statistic in Eq. (6).

B.3 Round-level algorithm

Algorithm 1 makes the order of gating, local optimization, aggregation, and prototype update explicit. The same procedure is used for IID and non-IID partitions; only the client sampler and local data distributions change.

Algorithm 1 FedSBR training

Require: clients $\{\mathcal{D}_k\}_{k=1}^{K}$, initial model θ^0, rounds T, local steps τ

Initialize prototype values and supports to zero, with all classes marked invalid.
for $t = 0, \ldots, T-1$ **do**
  Server samples $\mathcal{S}_t$ and broadcasts the model, prototype bank, validity mask, and class supports N_c^t.
  for each $k \in \mathcal{S}_t$ in parallel **do**
    Initialize local parameters from $(\theta^t, \mathbf{P}^t)$.
    for $s = 1, \ldots, \tau$ **do**
      Draw a local minibatch and compute unimodal losses, ranks, and teacher confidence.
      Select a_k and $\mathcal{W}_k$; detach the gate weights $\alpha_{w,k}$.
      Evaluate $\mathcal{L}_{\text{task}}^k + \mathcal{L}_{\text{CPR}}^k + \mathcal{L}_{\text{DGT}}^k + \mathcal{L}_{\text{FSA}}^k$ and take one optimizer step.
    end for
    Collect class summaries in an evaluation-mode pass over the local training data.
    Return updated parameters, class counts $n_{c,k}^t$, quality masses $Q_{c,k}^t$, and prototypes $\mathbf{c}_{c,k}^{t+1}$ where $Q_{c,k}^t > 0$.
  end for
  Aggregate modality blocks and fusion parameters with Eq. (4).
  Update $\mathbf{P}^{t+1}$ from returned class summaries with Eq. (10).
end for
return $\theta^T, \mathbf{P}^T$

C Additional Experiments

The experiments address three linked questions: how weak-pathway degradation changes under federation, what coordinated reallocation adds to a centralized remedy transferred into FedAvg, and how its components affect capacity and predictive use. Figure 1 supplies the initial centralized–federated diagnostic; Tables 6, 15, and 16 examine the transferred baseline, resource allocation, and component effects, respectively. Pathway, attenuation, layer-wise, and class-level results connect these comparisons to the behavior of the weak modality.

C.1 Experimental Details

CREMA-D (Cao et al., 2014) contains six emotion classes and uses an actor-disjoint split of 5,803/738/901 training, validation, and test clips. The audio input is a 257×299 short-time Fourier transform and the visual input is the center video frame. ModelNet40 (Wu et al., 2015) contains 40 object classes; views 0 and 6 are treated as modalities under the official 8,858/985/2,468 split. CrisisMMD (Alam et al., 2018) uses paired social-media images and text with the official version-2 Task-1 informative/non-informative labels and splits. MIMIC-IV (Johnson et al., 2023) provides six EHR groups (clinical notes, laboratory values, demographics, diagnoses, procedures, and medications); the tasks are 90-day mortality and 15-day readmission, evaluated by their mean AUROC on patient-disjoint splits.

Each dataset has 10 clients, with 5 sampled without replacement per round, 100 communication rounds, 2 local epochs, and batch size 128. IID partitions use uniform class proportions; non-IID partitions draw client proportions from a Dirichlet distribution with $\alpha = 0.5$. Proxy-free methods use the complete training split. FedAMB reserves a deterministic class-stratified 20% public unlabeled proxy set and distributes the remaining 80% to private clients. Checkpoints are selected by validation fusion accuracy, or by mean validation AUROC for MIMIC-IV, and evaluated once on the test set.

C.2 Comparison Protocol and Evaluation Units

Common training and selection rules. Comparisons within a dataset use the splits, preprocessing, backbone inputs, client partitions, round schedule, and validation-based checkpoint rule in Appendices B.1 and C.1. The primary budget is 100 rounds with five of ten clients participating and two local epochs. Method-specific optimizers and auxiliary objectives follow their stated config-

urations. Table 8 reports the resulting compute cost, while Appendix C.5 accounts for additional parameters and prototype traffic.

Prediction unit and data access. Table 1 marks personalized-client evaluation with † and public-proxy access with ‡. Global scores evaluate one server model; personalized scores average client-model performance. In that table, FedAMB reserves 20% of the training data as an unlabeled proxy set and uses the remaining 80% privately. Table 15 holds the private split at 80% for every method and labels the evaluation unit as G or P. Proxy-based methods additionally access the reserved 20%; FedAvg and FedSBR use only the common private split. The G rows compare server predictors, and the P rows compare personalized predictors. Together, these comparisons separate private-data volume, prediction unit, and proxy access when interpreting performance.

Controls for the federated intervention. FedAvg+EBR uses the same data and training protocol as FedSBR and transfers the centralized basis remedy directly into parameter averaging. Table 6 therefore tests the added value of the coordinated design. The task-only "None" control in Table 16 retains the shared–private parameterization; the component removals then identify how capacity regularization, predictive transfer, and semantic anchoring contribute within that structure.

Checkpoint and metric alignment. Table 16 pairs fusion accuracy, weak-pathway accuracy, participation rank, and negative Shapley fraction on the held-out test set at the validation-selected checkpoint. Table 3 reports rank on the balanced validation subset used for the original ablation diagnostic. The ModelNet40 task-only row uses the same checkpoint in both tables. These evaluation populations are stated explicitly so that numerical rank comparisons use the same population and checkpoint. The low-budget sensitivity sweeps in Appendix C.11 retain their separate seed-42 protocol.

C.3 Targeted FedAvg+EBR Comparison

The targeted baseline directly transfers the centralized Explicit Basis Reallocation (EBR) encoder (Chaudhuri et al., 2025) into FedAvg. The comparison uses the same data partitions, local training budget, checkpoint rule, and three seeds as the main experiments, while giving EBR the same tuning budget as the other baselines. Table 6 reports fusion performance, weak-modality performance, and participation rank alongside FedAvg and FedSBR. Fusion, weak-pathway, and rank metrics respectively measure the overall task result, retention of the weak pathway, and the change in representation capacity. FedAvg+EBR tests whether the federated coordination in FedSBR adds value beyond directly transplanting EBR. The centralized– federated diagnostics and conditional analysis evaluate federation's role in aggravating collapse.

Across datasets, FedAvg+EBR improves the weak pathway over FedAvg, while FedSBR gives the largest joint improvement in fusion quality and participation rank. On the non-IID CREMA-D split, for example, FedSBR reaches 50.39 fusion accuracy, 28.30 visual accuracy, and rank 4.08, compared with 46.91, 23.42, and 2.87 for FedAvg+EBR. The advantage over the transferred EBR baseline supports coordinating local capacity preservation with shared semantics during federation. The parameterization control and component ablations examine this coordination in Table 16.

CREMA-D frames and ModelNet40 views are resized to 224×224 and 112×112, respectively, without ImageNet pretraining. CrisisMMD uses cached ImageNet-pretrained ResNet-18 image features (512 dimensions) and BERT-base-uncased text features (768 dimensions); all methods train the same 512-dimensional modality adapters and fusion classifier. Unless a baseline specifies otherwise, clients use SGD with learning rate 10^{-3}, momentum 0.9, and weight decay 10^{-4}. SCAFFOLD uses zero client momentum. Reported means use three independent runs with sample standard deviations; all methods within a dataset share the same runs and checkpoint rule.

Table 6: FedAvg+EBR test-set comparison (mean$_{\mathrm{SD}}$, three runs). Weak: Visual (CREMA-D), View-6 (ModelNet40), Image (CrisisMMD), and weakest-modality AUROC averaged over two tasks (MIMIC-IV).

Dataset	Partition	Method	Fusion ↑	Weak ↑	Rank ↑
CREMA-D	IID	FedAvg	$48.83_{(1.37)}$	$18.76_{(1.12)}$	$2.88_{(0.24)}$
		FedAvg+EBR	$51.14_{(1.18)}$	$21.83_{(1.47)}$	$3.14_{(0.28)}$
		FedSBR	$55.39_{(0.82)}$	$28.30_{(1.27)}$	$3.68_{(0.22)}$
	Non-IID	FedAvg	$45.39_{(1.63)}$	$22.09_{(1.38)}$	$2.59_{(0.29)}$
		FedAvg+EBR	$46.91_{(1.47)}$	$23.42_{(1.64)}$	$2.87_{(0.33)}$
		FedSBR	$50.39_{(0.97)}$	$28.30_{(1.41)}$	$4.08_{(0.18)}$
ModelNet40	IID	FedAvg	$78.93_{(0.71)}$	$48.93_{(1.54)}$	$5.62_{(0.38)}$
		FedAvg+EBR	$81.47_{(0.62)}$	$55.86_{(1.31)}$	$6.27_{(0.34)}$
		FedSBR	$86.16_{(0.43)}$	$68.73_{(1.08)}$	$8.14_{(0.26)}$
	Non-IID	FedAvg	$78.53_{(0.93)}$	$46.28_{(1.74)}$	$5.21_{(0.43)}$
		FedAvg+EBR	$80.23_{(0.78)}$	$51.14_{(1.52)}$	$5.63_{(0.41)}$
		FedSBR	$84.20_{(0.62)}$	$65.95_{(1.04)}$	$7.58_{(0.31)}$
CrisisMMD	IID	FedAvg	$75.46_{(0.54)}$	$51.87_{(0.83)}$	$4.23_{(0.31)}$
		FedAvg+EBR	$76.18_{(0.47)}$	$53.24_{(0.76)}$	$4.52_{(0.27)}$
		FedSBR	$77.64_{(0.41)}$	$55.71_{(0.67)}$	$5.08_{(0.23)}$
	Non-IID	FedAvg	$74.25_{(0.72)}$	$50.14_{(1.02)}$	$3.89_{(0.37)}$
		FedAvg+EBR	$74.73_{(0.64)}$	$51.06_{(0.93)}$	$4.11_{(0.34)}$
		FedSBR	$75.43_{(0.52)}$	$53.42_{(0.79)}$	$4.67_{(0.26)}$
MIMIC-IV	IID	FedAvg	$78.56_{(0.81)}$	$71.34_{(1.12)}$	$5.93_{(0.42)}$
		FedAvg+EBR	$79.82_{(0.71)}$	$73.47_{(0.98)}$	$6.41_{(0.37)}$
		FedSBR	$81.89_{(0.57)}$	$76.52_{(0.81)}$	$7.38_{(0.29)}$
	Non-IID	FedAvg	$75.13_{(1.09)}$	$67.87_{(1.38)}$	$5.24_{(0.49)}$
		FedAvg+EBR	$76.78_{(0.94)}$	$69.93_{(1.21)}$	$5.63_{(0.44)}$
		FedSBR	$79.57_{(0.68)}$	$74.21_{(0.91)}$	$6.92_{(0.32)}$

C.4 Convergence

Table 7: CREMA-D convergence summary corresponding to Figure 3. Entries are five-round moving-average validation accuracies; the final column gives the best visual accuracy and round.

		Fusion				Visual pathway				
Partition	Method	$\overline{F}$	R25	R50	R100	$\overline{V}$	R25	R50	R100	Best (round)
IID	FedAvg	43.80	41.11	44.12	50.43	17.39	17.29	17.45	21.46	21.46 (100)
IID	FedSBR	45.27	41.54	48.51	53.90	21.47	17.26	23.50	24.47	25.88 (96)
Non-IID	FedAvg	37.82	34.74	34.04	46.91	18.53	16.88	17.24	23.14	23.44 (89)
Non-IID	FedSBR	37.44	33.98	39.00	47.56	21.66	18.10	23.77	22.93	27.51 (87)

Across the complete trajectory, FedSBR raises mean visual accuracy by 4.08 points under IID and 3.13 points under non-IID. The non-IID last-round value is close to FedAvg, while the trajectory mean and best round expose the larger weak-pathway recovery.

C.5 Computation Cost

The additional cost is dataset-dependent. Relative to FedAvg, FedSBR adds 5.07, 1.89, and 3.09 seconds per round on CREMA-D, CrisisMMD, and MIMIC-IV, respectively. On ModelNet40, the observed means are 25.65 seconds for FedSBR and 42.78 seconds for FedAvg, whose across-run and partition SD is 31.00 seconds. These timings describe the measured execution conditions; the

Table 8: Per-round computation over 100 communication rounds. Time/round is reported as mean_{SD} across runs and partitions, where the subscript is the sample standard deviation; total is compute time in GPU hours, excluding serialization and final evaluation.

	CREMA-D		ModelNet40		CrisisMMD		MIMIC-IV	
Method	Time/round (s)	Total (h)	Time/round (s)	Total (h)	Time/round (s)	Total (h)	Time/round (s)	Total (h)
FedAvg	$16.51_{(4.37)}$	0.46	$42.78_{(31.00)}$	1.19	$1.59_{(1.12)}$	0.04	$5.82_{(1.47)}$	0.16
FedProx	$12.75_{(1.29)}$	0.35	$36.17_{(5.61)}$	1.00	$1.40_{(0.19)}$	0.04	$5.47_{(0.83)}$	0.15
FedAdam	$24.31_{(44.91)}$	0.68	$48.12_{(13.28)}$	1.34	$1.33_{(0.16)}$	0.04	$7.14_{(3.21)}$	0.20
SCAFFOLD	$123.56_{(37.13)}$	3.43	$85.24_{(57.17)}$	2.37	$2.35_{(0.92)}$	0.07	$18.73_{(4.86)}$	0.52
FedProto	$20.47_{(2.98)}$	0.57	$28.26_{(2.45)}$	0.79	$2.54_{(1.67)}$	0.07	$6.89_{(1.15)}$	0.19
FedMSplit	$16.81_{(2.17)}$	0.47	$21.68_{(2.41)}$	0.60	$1.76_{(1.29)}$	0.05	$6.12_{(0.91)}$	0.17
FedAvg-RMD	$18.86_{(1.73)}$	0.52	$18.38_{(1.31)}$	0.51	$1.29_{(0.15)}$	0.04	$5.21_{(0.72)}$	0.14
FedPMR	$42.55_{(5.70)}$	1.18	$23.80_{(1.58)}$	0.66	$1.99_{(0.20)}$	0.06	$8.64_{(1.43)}$	0.24
BMSFed	$20.12_{(8.00)}$	0.56	$23.47_{(3.52)}$	0.65	$2.08_{(0.73)}$	0.06	$6.58_{(1.84)}$	0.18
FedAMB	$34.89_{(1.61)}$	0.97	$54.87_{(2.75)}$	1.52	$3.67_{(1.57)}$	0.10	$10.24_{(0.96)}$	0.28
FedSBR	$21.58_{(6.08)}$	0.60	$25.65_{(4.12)}$	0.71	$3.48_{(1.96)}$	0.10	$8.91_{(2.13)}$	0.25

operation counts below characterize algorithmic overhead. All measurements use the same 100-round protocol as the accuracy comparison.

Computational and communication complexity. Let b denote batch size, M the number of available modalities, $d = d_s + d_p$ the encoder width, and C the number of classes. Dense basis projection and reconstruction add $O(Mbd^2)$ operations per batch and $2Md^2$ trainable weights. The shared classifier adds $Cd_s + C$ parameters. For W gated weak pathways, private covariance and shared–private cross-covariance require $O(Wb(d_p^2 + d_s d_p))$ operations and $O(W(d_p^2 + d_s d_p))$ matrix storage. Participation rank can be computed exactly from either $H^\top H$ or $HH^\top$, at $O(Mbd \min(b, d))$ time and $O(M \min(b, d)^2)$ matrix storage, using $\operatorname{tr}((H^\top H)^2) = \|HH^\top\|_F^2$. Shared classification, distillation, and prototype scoring add $O(MbCd_s + WbC)$ operations; shared-feature alignment costs $O(M^2 bd_s)$ for all modality pairs.

Prototype accumulation over n_k local examples costs $O(n_k M d_s)$ after feature extraction, with $O(Cd_s + C)$ persistent bank storage. Collecting summaries after local optimization also requires an evaluation-mode encoder pass over these examples. The class-summary upload contains at most $C(d_s + 2)$ scalars per participating client: a shared prototype, its quality mass, and its sample count. The broadcast bank, supports, and validity mask use $O(Cd_s + C)$ storage and communication. Parameter exchange additionally includes the $2Md^2 + Cd_s + C$ basis and shared-classifier parameters. With $M = 2$, $d = 512$, and $d_s = 256$, these extra parameters occupy about 4.01 MiB for six classes in 32-bit representation per model transmission; the class-summary upload occupies about 6.05 KiB. For 40 classes, the corresponding sizes are 4.04 MiB and 40.31 KiB. These counts separate prototype traffic from the additional trainable model state; optimizer-state memory scales with the latter.

C.6 MIMIC-IV Modality Scaling

MIMIC-IV provides a controlled test of how performance changes with the number of available modalities. We aggregate all $\binom{6}{q}$ subsets of the six EHR groups for each modality count q under the non-IID partition.

FedSBR stays near 79% mean AUROC as the modality count increases from two to five. Rank retention decreases from 0.93 to 0.86, while FedAvg and FedPMR fall to 0.47 and 0.63, respectively.

Table 9: MIMIC-IV performance as the number of modalities increases under the non-IID partition. Entries are formatted as mean_{SD}, where the subscript is the sample standard deviation over three runs. Mean AUROC averages mortality and readmission; rank retention is normalized by the unimodal reference.

	Mean AUROC ↑				Rank Retention ↑			
Method	2	3	4	5	2	3	4	5
FedAvg	$76.34_{(1.02)}$	$75.61_{(1.18)}$	$74.73_{(1.43)}$	$73.48_{(1.36)}$	$0.83_{(0.05)}$	$0.75_{(0.07)}$	$0.62_{(0.06)}$	$0.47_{(0.08)}$
FedAvg-RMD	$75.87_{(0.93)}$	$75.18_{(1.06)}$	$74.32_{(1.31)}$	$73.05_{(1.19)}$	$0.81_{(0.04)}$	$0.73_{(0.06)}$	$0.60_{(0.07)}$	$0.45_{(0.05)}$
FedPMR	$77.15_{(1.08)}$	$76.53_{(0.96)}$	$75.91_{(1.33)}$	$74.82_{(1.47)}$	$0.87_{(0.04)}$	$0.82_{(0.05)}$	$0.73_{(0.06)}$	$0.63_{(0.07)}$
BMSFed	$76.78_{(1.14)}$	$76.14_{(0.98)}$	$75.47_{(1.29)}$	$74.56_{(1.52)}$	$0.85_{(0.06)}$	$0.79_{(0.04)}$	$0.69_{(0.08)}$	$0.58_{(0.06)}$
FedSBR	$78.63_{(0.64)}$	$79.18_{(0.57)}$	$79.52_{(0.71)}$	$79.41_{(0.68)}$	$0.93_{(0.03)}$	$0.91_{(0.02)}$	$0.89_{(0.04)}$	$0.86_{(0.03)}$

C.7 Realized Client Heterogeneity

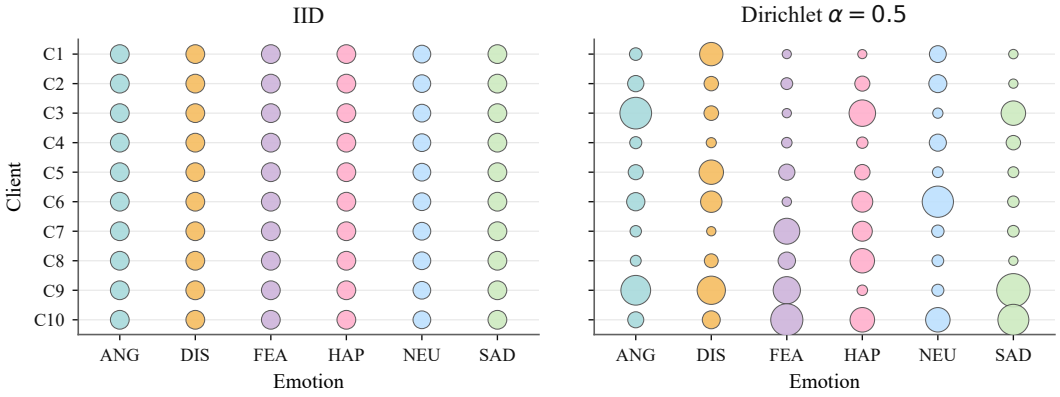

Figure 5: CREMA-D training-sample allocation across clients. Marker area is proportional to class count.

Marker areas show that the non-IID split combines clients with concentrated class support and clients with broader coverage. The client allocation creates the client-level update diversity used in the later heterogeneity and sensitivity analyses, while the shared prototype update provides a common semantic anchor.

C.8 Continuous Modality Attenuation

For a trained checkpoint, let the linear fusion logit be $z = z_a + z_v + b$. We evaluate $z(\rho_a, \rho_v) = \rho_a z_a + \rho_v z_v + b$ without retraining, with one coefficient varied over $\{0, .25, .50, .75, 1\}$ and the other fixed to one.

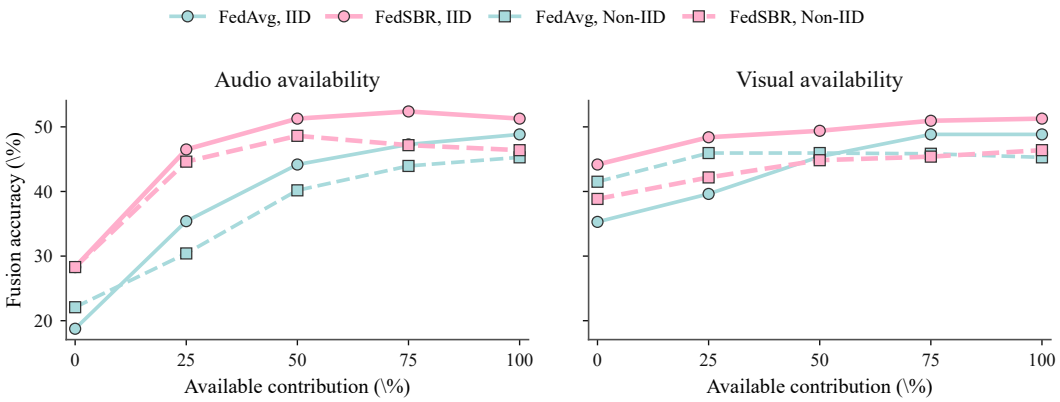

Figure 6: CREMA-D fusion accuracy when one modality contribution is continuously attenuated on the test set.

Table 10: CREMA-D endpoint metrics under modality attenuation. Visual-only sets $\rho_a = 0$. Audio-only sets $\rho_v = 0$. Full input sets $\rho_a = \rho_v = 1$. Ret. denotes accuracy retention, $100 \times \mathrm{Acc.}_{\mathrm{pathway}}/\mathrm{Acc.}_{\mathrm{full}}$.

		Full input		Visual only			Audio only		
Partition	Method	Acc.	F1	Acc.	F1	Ret. (%)	Acc.	F1	Ret. (%)
IID	FedAvg	48.83	48.00	18.76	12.67	38.41	35.29	29.38	72.27
IID	FedSBR	55.39	52.67	28.30	25.14	51.10	44.17	43.25	79.74
Non-IID	FedAvg	45.39	43.17	22.09	12.65	48.67	38.85	36.52	85.59
Non-IID	FedSBR	50.39	47.74	28.30	19.71	56.16	41.51	39.79	82.38

The attenuation curve and endpoint table show that FedSBR preserves useful predictions as either stream is weakened. Under the non-IID split, visual-only accuracy rises from 22.09 for FedAvg to 28.30 for FedSBR, and the retained visual contribution increases from 48.67% to 56.16%. Audio-only performance also remains competitive, indicating that the gain comes from a more balanced fusion geometry rather than a simple preference for the visual stream.

C.9 Layer-Wise Representation Geometry

Table 11: Layer-wise effective rank and class-similarity gap on a balanced CREMA-D validation subset. The gap is mean within-class cosine similarity minus mean between-class similarity.

		Effective rank				Class-similarity gap			
Setting	Modality	L1	L2	L3	L4	L1	L2	L3	L4
Centralized	Audio	2.33	2.88	5.12	7.19	.313	.321	.315	.446
	Visual	3.15	4.30	6.09	6.70	−.015	−.008	.026	.176
FedAvg IID	Audio	2.06	2.49	3.37	4.11	.293	.357	.338	.329
	Visual	2.94	3.84	5.34	8.78	−.018	−.003	.020	.048
FedAvg non-IID	Audio	2.08	2.70	5.26	8.61	.279	.362	.290	.267
	Visual	2.71	3.38	4.00	5.40	−.019	−.004	.033	.086
FedSBR IID	Audio	2.06	2.62	4.33	6.21	.284	.362	.319	.296
	Visual	2.90	4.55	5.75	7.36	−.021	−.001	.053	.113
FedSBR non-IID	Audio	2.06	2.78	6.44	10.96	.294	.322	.253	.202
	Visual	2.58	3.56	4.07	7.20	−.020	−.005	.045	.116

The layer-wise statistics separate capacity from class structure. FedSBR's visual L4 representation has a class-similarity gap of .113 under IID and .116 under non-IID, compared with .048 and .086 for FedAvg. The corresponding rank values remain substantial, showing that the recovered directions also carry class-relevant organization.

Effective rank alone does not imply semantic separation: the FedAvg-IID visual L4 rank is 8.78, but its class gap is only .048. FedSBR increases the visual L4 gap to .113 under IID and .116 under non-IID.

C.10 Class-Conditional Visual Diagnostics

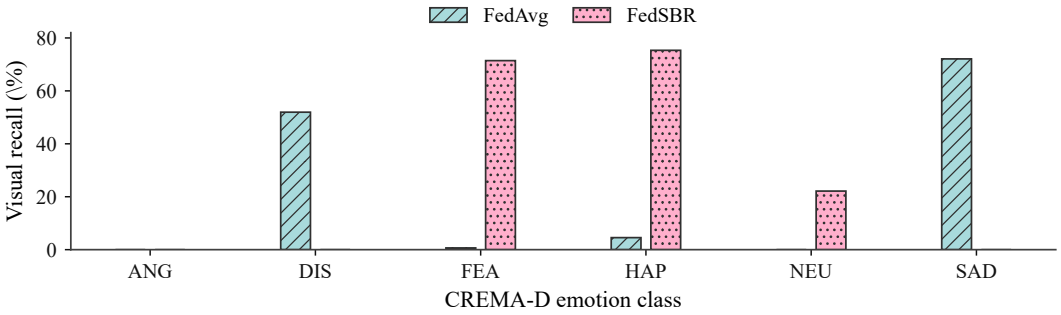

Figure 7: Per-class visual-pathway recall on the non-IID CREMA-D test set.

FedSBR raises visual macro-recall from 21.54 to 28.15 through a redistribution across classes. Recall increases most on fear, happiness, and neutral, while anger, disgust, and sadness have little or no visual-pathway recall.

C.11 Federated Sensitivity Sweeps

Client exposure. With K clients, m sampled uniformly per round, and T rounds, expected participation per client is Tm/K and the probability of at least one selection is $1 - (1 - m/K)^T$. Thus five rounds with five sampled clients cover an expected 22.6% of a 100-client population. The short-budget sweep characterizes early training under changing client exposure. A converged scaling comparison should hold expected participation fixed: $T = 10K$ with $m = 5$ matches the main protocol's 50 expected selections per client.

We vary one federation factor at a time on CREMA-D while keeping the actor split, optimizer, and evaluation protocol fixed. All three sweeps use a single seed (seed 42), one local epoch, and batch size 512. Most conditions use five communication rounds, with some client-count conditions limited to one round. The reported values are single-run test accuracies rather than three-seed means; no across-seed standard deviations are reported. These low-budget experiments provide exploratory sensitivity diagnostics alongside the 100-round main experiments. The missing-modality rate is the fraction of clients that are audio-only or visual-only, with the remaining clients observing both modalities.

Table 12: Sensitivity to client count on CREMA-D under non-IID ($\alpha = 0.5$). Entries are single-run test accuracies (%) using seed 42 and a low training budget; FedMSplit reports personalized-client performance.

	Fusion accuracy (%)							Visual accuracy (%)						
Method	5	10	15	20	30	50	100	5	10	15	20	30	50	100
FedAvg	34.18	30.53	27.87	24.53	21.42	17.87	18.09	20.42	19.09	18.42	17.87	16.42	15.09	15.31
FedProx	34.42	30.76	28.09	24.76	21.53	18.64	17.87	20.20	18.87	18.87	17.64	16.53	15.20	15.31
FedAdam	37.87	34.42	31.53	28.31	23.87	20.42	18.53	21.09	19.76	19.42	18.53	17.42	15.87	16.20
SCAFFOLD	28.42	25.09	25.42	21.87	18.64	17.42	16.76	18.09	17.20	16.87	16.42	15.42	14.87	14.64
FedProto	32.42	28.20	25.42	22.09	19.53	17.42	17.64	18.87	17.64	17.42	16.53	15.87	14.98	15.20
FedMSplit	30.87	27.42	25.53	22.87	20.09	17.98	16.64	19.42	18.20	17.42	17.64	15.87	15.31	15.53
FedAvg-RMD	33.64	29.42	26.53	23.42	20.09	17.87	16.87	21.31	19.87	19.20	18.42	17.09	15.76	16.09
FedPMR	36.09	32.09	29.20	25.64	22.09	18.76	17.98	22.20	20.53	19.76	19.09	17.98	16.42	16.87
BMSFed	35.31	31.42	28.42	25.31	21.20	18.64	17.42	21.53	20.09	19.42	18.42	18.76	16.09	15.87
FedSBR	32.64	28.87	26.64	24.53	21.87	19.42	18.64	22.97	21.76	20.98	19.87	18.87	17.42	17.64

Table 13: Sensitivity to label heterogeneity on CREMA-D with 10 clients. Smaller Dirichlet α indicates stronger class skew. Entries are single-run test accuracies (%) using seed 42 and a low training budget; FedMSplit reports personalized-client performance.

	Fusion accuracy (%)							Visual accuracy (%)						
Method	0.05	0.1	0.3	0.5	1.0	2.0	5.0	0.05	0.1	0.3	0.5	1.0	2.0	5.0
FedAvg	24.53	26.87	28.76	30.53	33.14	35.09	36.53	17.53	18.20	18.64	19.09	20.31	20.87	21.42
FedProx	24.76	27.09	28.98	30.76	33.42	35.31	36.87	17.31	17.98	18.42	18.87	20.09	20.64	21.20
FedAdam	28.09	30.42	32.42	34.42	37.31	38.87	39.42	18.20	18.87	19.31	19.76	20.98	21.42	22.09
SCAFFOLD	20.42	22.31	23.76	25.09	27.42	29.31	30.87	16.09	16.53	16.87	17.20	18.09	18.76	19.42
FedProto	22.53	24.87	26.53	28.20	30.64	32.42	33.76	15.76	16.20	16.76	17.64	18.98	19.64	20.20
FedMSplit	23.53	25.42	26.87	27.42	29.53	31.09	32.42	16.53	17.09	17.53	18.20	19.42	20.09	20.64
FedAvg-RMD	23.87	25.87	27.76	29.42	31.98	33.64	35.20	18.42	18.98	19.42	19.87	21.20	21.76	22.42
FedPMR	26.09	28.31	30.20	32.09	34.87	36.53	38.42	18.98	19.64	20.09	20.53	21.76	22.42	22.98
BMSFed	25.64	27.64	29.64	31.42	34.09	35.87	37.64	18.76	19.42	19.87	20.09	21.42	21.98	22.64
FedSBR	22.87	25.64	27.53	28.87	31.53	33.87	35.42	20.87	21.42	21.64	21.76	22.53	22.87	23.20

Table 14: Sensitivity to missing-modality rate on CREMA-D under non-IID ($\alpha = 0.5$) with 10 clients. The rate is the fraction of audio-only or visual-only clients. Entries are single-run test accuracies (%) using seed 42 and a low training budget; FedMSplit reports personalized-client performance.

	Fusion accuracy (%)						Visual accuracy (%)					
Method	0	10	25	50	75	90	0	10	25	50	75	90
FedAvg	30.53	29.42	27.87	24.87	21.98	18.76	19.09	19.20	17.42	15.53	13.87	12.20
FedProx	30.76	29.64	27.98	25.09	22.20	18.87	18.87	17.98	17.20	15.64	13.76	11.87
FedAdam	34.42	33.20	31.42	28.76	25.64	22.09	19.76	19.87	18.09	16.42	14.53	12.76
SCAFFOLD	25.09	24.20	24.31	21.87	19.64	17.42	17.20	17.31	16.09	14.87	13.20	11.42
FedProto	28.20	27.20	25.64	22.98	19.87	17.09	17.64	16.98	15.87	14.20	12.42	10.64
FedMSplit	27.42	26.64	24.87	22.64	19.76	16.98	18.20	17.42	16.53	14.98	12.98	11.20
FedAvg-RMD	29.42	28.42	26.87	24.42	21.42	18.20	19.87	20.09	18.31	16.53	14.76	12.87
FedPMR	32.09	30.98	29.20	26.53	23.42	20.09	20.53	20.64	18.87	17.20	15.31	13.42
BMSFed	31.42	30.31	28.64	26.09	22.98	19.42	20.09	20.31	18.42	16.87	14.87	13.09
FedSBR	28.87	28.64	28.20	27.53	26.42	24.76	21.76	21.53	20.87	19.87	18.87	17.42

The sweep tracks whether the weak visual pathway follows the fused score as federation conditions change. The visual-pathway accuracy columns make changes visible that fusion accuracy alone can hide.

The three sweeps show a consistent ordering of pathway retention. With 10 clients and $\alpha = 0.5$, FedSBR obtains 21.76 visual accuracy, and with 90% of clients missing one modality it retains 17.42 visual accuracy versus 12.20 for FedAvg. Under stronger label skew ($\alpha = 0.05$), the visual score remains 20.87 for FedSBR, above the 17.53 obtained by FedAvg. The observed trends connect the main result to changes in participation, label balance, and modality availability.

C.12 Controlled Comparisons and Task-Level Results

The following comparisons connect the federated design to three empirical checks: performance under a common private-data budget, component effects on jointly evaluated weak-pathway metrics, and discrimination on each clinical task. The prediction units and data-access rules follow Appendix C.2.

Table 15: Matched-protocol comparison on CREMA-D (%). All methods train on the same 80% private split. G/P: global/personalized predictor. †: personalized-client evaluation; ‡: 20% public proxy allocation.

		IID			Non-IID ($\alpha = 0.5$)		
Method	Unit	Fusion	Visual	V-F1	Fusion	Visual	V-F1
FedAvg	G	46.84	17.31	11.62	43.62	20.51	11.28
FedSBR	G	53.28	26.41	22.87	48.85	25.73	17.94
FedMSplit†	P	37.96	15.73	7.82	34.72	16.58	10.35
FedAMB†‡	P	31.53	14.24	9.76	24.06	11.73	5.84
CREAM-FL‡	G	49.37	19.86	14.41	45.91	21.36	13.72
CREAM-FL†‡	P	47.62	18.34	13.15	44.38	20.47	12.58

Under the common 80% private-data protocol, FedSBR improves global fusion accuracy over FedAvg by 6.44/5.23 points and visual accuracy by 9.10/5.22 points under IID/non-IID. Both use the same private examples and no public proxy. Among global predictors, FedSBR also exceeds CREAM-FL despite the latter's additional proxy access. The P rows report the corresponding personalized methods. Together with Table 6, the global-model comparisons support a gain from coordinated reallocation beyond private-data volume and the direct transfer of a centralized basis remedy.

Table 16: Checkpoint-matched component diagnostics under non-IID ($\alpha = 0.5$). All metrics are computed on the held-out test set at the validation-selected checkpoint. Weak: Visual (CREMA-D), View-6 (ModelNet40). Neg: negative weak-modality Shapley fraction. "None" retains the shared–private parameterization with task loss only.

Configuration	Fusion (%)	Weak (%)	Rank↑	Neg↓
CREMA-D				
Full FedSBR	50.39	28.30	4.08	0.38
w/o CPR	44.73	22.75	3.24	0.43
w/o DGT	47.95	27.75	3.58	0.41
w/o FSA	48.50	27.08	3.71	0.40
None	43.17	20.53	2.73	0.46
ModelNet40				
Full FedSBR	84.20	65.95	7.58	0.18
w/o CPR	79.42	64.51	6.07	0.23
w/o DGT	78.61	40.19	6.18	0.28
w/o FSA	79.78	65.44	6.91	0.21
None	78.24	45.46	5.21	0.30

The checkpoint-matched results connect capacity to predictive use. On CREMA-D, removing CPR lowers rank from 4.08 to 3.24 and increases the negative Shapley fraction from 0.38 to 0.43. On ModelNet40, removing DGT lowers View-6 accuracy by 25.76 points and raises the negative frac-

tion from 0.18 to 0.28. Removing FSA lowers fusion accuracy and rank on both datasets, linking the shared semantic signal to the quality of the federated predictor. The full model improves all four metrics over the task-only basis control. These complementary effects support the joint design: preserved private capacity needs predictive supervision, while shared semantics need coordination across clients.

Table 17: Task-level MIMIC-IV evaluation (%). All entries use the paper's patient-disjoint splits, 90-day mortality and 15-day readmission definitions, and main training protocol.

	90-day mortality		15-day readmission	
Method	AUROC	AUPRC	AUROC	AUPRC
		IID		
FedAvg	82.34	48.72	74.78	41.35
FedAdam	84.18	51.43	76.44	43.87
FedProto	83.42	49.58	74.86	42.14
FedPMR	82.78	50.16	76.46	43.91
BMSFed	81.53	47.85	75.25	41.62
FedSBR	85.47	54.63	78.31	47.85
		Non-IID ($\alpha = 0.5$)		
FedAvg	79.62	44.38	70.64	37.14
FedAdam	81.74	48.26	72.82	40.51
FedProto	79.45	43.72	70.33	36.58
FedPMR	80.53	46.81	71.89	39.04
BMSFed	80.14	45.37	71.22	37.86
FedSBR	83.72	52.18	75.42	44.93

FedSBR achieves the highest AUROC and AUPRC on both clinical tasks under both partitions. On non-IID data, its AUPRC exceeds FedAdam by 3.92 points for mortality and 4.42 points for readmission. Reporting each task separately shows that the improvement in mean AUROC extends to both outcomes and to precision–recall performance. These clinical results complement the pathway and component diagnostics on the classification benchmarks.

D Analysis of Federated Collapse and Mitigation

The analysis follows the federated mechanism in Sec. 2. First, local drift and partial participation introduce blockwise update errors; an exact cancellation identity quantifies how much client update energy survives aggregation. Second, a spectral comparison specifies when additional federated contraction lowers weak-pathway participation rank. Third, the CPR, DGT, and FSA bounds connect private capacity, predictive descent, and cross-client semantic coordination. The error decomposition and cancellation identity are algebraic; the spectral and mitigation results state their additional conditions explicitly. This develops the representation perspective of Chaudhuri et al. (2025) and Huh et al. (2023) around the local-update and aggregation cycle.

D.1 Notation and Assumptions

Let $\theta = (\theta_1, \ldots, \theta_M, \omega)$ collect the modality-specific parameter blocks and the remaining shared parameters. For FedSBR, each modality block includes its encoder and basis/reconstruction maps; ω includes the fusion and shared-classifier parameters. The symbols a_k and w in the main text are dynamic client-side roles. For the blockwise analysis below, a and w denote a fixed modality pair selected before the analysis window. Set $\mathcal{B} = \{1, \ldots, M, h\}$ and $\theta_h = \omega$. The global objective is

$$F(\theta) = \sum_{k=1}^{K} p_k F_k(\theta), \qquad p_k \geq 0, \qquad \sum_{k=1}^{K} p_k = 1. \tag{13}$$

The reference FedAvg analysis uses a fixed differentiable objective. When applying a local bound to FedSBR, we condition on a fixed prototype bank, teacher targets, modality roles, and detached loss

coefficients. The smoothness and moment assumptions apply to this conditioned objective; changes in the frozen quantities enter an additional residual. Multi-round telescoping uses a common fixed objective, while the FedSBR certificate concerns one frozen step. For the coordinate projector P_b, define

$$g_{b,k}(\theta) = P_b \nabla F_k(\theta), \qquad g_b(\theta) = \sum_{k=1}^{K} p_k g_{b,k}(\theta), \tag{14}$$

$$H_b^2(\theta) = \sum_{k=1}^{K} p_k \|g_{b,k}(\theta) - g_b(\theta)\|^2. \tag{15}$$

The weighted variance identity gives

$$\begin{aligned} \sum_{k=1}^{K} p_k \|g_{b,k}\|^2 &= \sum_{k} p_k \|g_b + (g_{b,k} - g_b)\|^2 \\ &= \|g_b\|^2 + H_b^2, \end{aligned} \tag{16}$$

because $\sum_k p_k (g_{b,k} - g_b) = 0$.

Assumption 1 (Block smoothness and lower bound)**.** *For all k, F_k is differentiable, $F(\theta) \geq F_* > -\infty$, and*

$$F(\theta') \leq F(\theta) + \langle \nabla F(\theta), \theta' - \theta \rangle + \frac{L}{2} \|\theta' - \theta\|^2, \tag{17}$$

$$\|P_b(\nabla F_k(\theta') - \nabla F_k(\theta))\|^2 \leq \sum_{c \in \mathcal{B}} L_{bc}^2 \|P_c(\theta' - \theta)\|^2. \tag{18}$$

Assumption 2 (Stochastic gradients)**.** *Let $B \geq 1$ denote the local minibatch size, distinct from the block set $\mathcal{B}$. Writing*

$$\widehat{g}_{b,k}(\vartheta; \xi) = g_{b,k}(\vartheta) + \zeta_{b,k}(\vartheta; \xi), \tag{19}$$

we assume, conditionally on the full local-step filtration $\mathcal{G}_{t,k,s}$,

$$\mathbb{E}[\zeta_{b,k}(\theta_{k,s}^t; \xi_{k,s}^t) \mid \mathcal{G}_{t,k,s}] = 0, \qquad \mathbb{E}[\|\zeta_{b,k}(\theta_{k,s}^t; \xi_{k,s}^t)\|^2 \mid \mathcal{G}_{t,k,s}] \leq \frac{\sigma_b^2}{B}. \tag{20}$$

We additionally assume the mixed-state moment bound used by the local-drift proof directly:

$$\sum_{k} p_k \mathbb{E}_t \|\widehat{g}_{b,k}(\theta_{k,s}^t)\|^2 \leq \Gamma_{b,t}, \qquad 0 \leq s < \tau. \tag{21}$$

Here $\mathbb{E}_t$ conditions on the state before client selection. The bounds $\Gamma_{b,t}$ and σ_b^2/B are assumed for the chosen stochastic gradient estimator, including batch-dependent regularizers, at the local states visited during the round. We assume fresh, conditionally independent minibatch draws across clients; across local steps, the noise forms a martingale-difference sequence under the full history.

D.2 Centralized Baseline

The centralized update is

$$\theta_{t+1}^{\mathrm{C}} = \theta_t - \eta \nabla F(\theta_t). \tag{22}$$

Lemma 1 (Centralized descent)**.** *If $0 < \eta \leq 1/L$, then*

$$F(\theta_{t+1}^{\mathrm{C}}) \leq F(\theta_t) - \frac{\eta}{2} \|\nabla F(\theta_t)\|^2. \tag{23}$$

Consequently,

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla F(\theta_t)\|^2 \leq \frac{2\Delta_0}{\eta T}, \qquad \Delta_0 = F(\theta_0) - F_*. \tag{24}$$

Proof. Substitution into Eq. (17) gives $F(\theta_{t+1}^{\mathrm{C}}) \leq F(\theta_t) - \eta(1 - L\eta/2)\|\nabla F(\theta_t)\|^2$. Since $1 - L\eta/2 \geq 1/2$, summing over t and using $F(\theta_T) \geq F_*$ proves both statements. □

D.3 Exact FedAvg Error Decomposition

Client k starts from θ_t and performs τ local steps:

$$\theta_{k,0}^t = \theta_t, \qquad \theta_{k,s+1}^t = \theta_{k,s}^t - \eta_\ell \widehat{\nabla} F_k(\theta_{k,s}^t; \xi_{k,s}^t), \quad 0 \le s < \tau. \tag{25}$$

The reference recursion uses plain local SGD and a common number of local steps τ (McMahan et al., 2017; Karimireddy et al., 2020). Its decomposition below is exact. Applying it to momentum updates or unequal client step counts requires a separately bounded residual for those differences. Let $\eta = \tau\eta_\ell$ and

$$d_t = \frac{1}{\tau}\sum_{s=0}^{\tau-1}\sum_{k=1}^{K} p_k\big(\nabla F_k(\theta_{k,s}^t) - \nabla F_k(\theta_t)\big), \tag{26}$$

$$\xi_t = \frac{1}{\tau}\sum_{s=0}^{\tau-1}\sum_{k=1}^{K} p_k \zeta_k(\theta_{k,s}^t; \xi_{k,s}^t), \qquad e_t = d_t + \xi_t. \tag{27}$$

Then, without approximation,

$$\begin{aligned}\sum_k p_k \theta_{k,\tau}^t &= \theta_t - \eta_\ell \sum_{s=0}^{\tau-1}\sum_k p_k \widehat{\nabla} F_k(\theta_{k,s}^t) \\ &= \theta_t - \eta\big(\nabla F(\theta_t) + d_t + \xi_t\big),\end{aligned} \tag{28}$$

so full-participation FedAvg is an inexact centralized step.

For a sampled client set $\mathcal{S}_t$, take $p_k = n_k/\sum_j n_j$ and the normalized sample-count weights used in the main protocol:

$$\widetilde{p}_{k,t} = \frac{n_k \mathbf{1}\{k \in \mathcal{S}_t\}}{\sum_{j\in\mathcal{S}_t} n_j}.$$

These common weights apply when all modalities are present. With missing modalities, the decomposition uses the corresponding block weights throughout. The sampled local drift and minibatch-noise terms are

$$d_t^{\mathcal{S}} = \frac{1}{\tau}\sum_{s=0}^{\tau-1}\sum_{k\in\mathcal{S}_t} \widetilde{p}_{k,t}\big(\nabla F_k(\theta_{k,s}^t) - \nabla F_k(\theta_t)\big),$$

$$\xi_t^{\mathcal{S}} = \frac{1}{\tau}\sum_{s=0}^{\tau-1}\sum_{k\in\mathcal{S}_t} \widetilde{p}_{k,t}\zeta_k(\theta_{k,s}^t; \xi_{k,s}^t), \qquad e_t^{\mathcal{S}} = d_t^{\mathcal{S}} + \xi_t^{\mathcal{S}}.$$

Define

$$a_t = \sum_{k\in\mathcal{S}_t} \widetilde{p}_{k,t}\nabla F_k(\theta_t) - \nabla F(\theta_t) = b_t + \widetilde{a}_t, \qquad b_t = \mathbb{E}_t[a_t], \quad \mathbb{E}_t[\widetilde{a}_t] = 0. \tag{29}$$

The ratio estimator has a bias term when client sizes are unequal; the zero-bias identity applies in the equal-size special case. The sampled update is

$$\theta_{t+1} = \theta_t - \eta\big(\nabla F(\theta_t) + u_t\big), \qquad u_t = e_t^{\mathcal{S}} + a_t. \tag{30}$$

For $K > 1$, equal client weights, and uniform sampling without replacement,

$$\mathbb{E}_t\|P_b\widetilde{a}_t\|^2 = \frac{K - K_t}{K_t(K-1)} H_b^2(\theta_t) \triangleq V_{b,t}^{\mathrm{part}}. \tag{31}$$

For $K = K_t = 1$, $V_{b,t}^{\mathrm{part}} = 0$. For the sample-count estimator, define instead $V_{b,t}^{\mathrm{part}} = \mathbb{E}_t\|P_b\widetilde{a}_t\|^2$ and retain the bias term $\|P_b b_t\|^2$ explicitly.

Lemma 2 (Blockwise local-drift bound)**.** *Let*

$$c_\tau = \frac{1}{\tau}\sum_{s=0}^{\tau-1} s^2 = \frac{(\tau-1)(2\tau-1)}{6}, \qquad q_{2,t}^{\mathcal{S}} = \mathbb{E}_t \sum_{k\in\mathcal{S}_t} \widetilde{p}_{k,t}^2. \tag{32}$$

Let $\Gamma^{\mathcal{S}}_{c,t}$ bound the selected-client mixed-state moment $\mathbb{E}_t \sum_{k\in\mathcal{S}_t} \widetilde{p}_{k,t}\|\widehat{g}_{c,k}(\theta^t_{k,s})\|^2$ uniformly over $0 \le s < \tau$. Then

$$\mathbb{E}_t\|e^{\mathcal{S}}_{b,t}\|^2 \le 2\eta_\ell^2 c_\tau \sum_{c\in\mathcal{B}} L^2_{bc}\Gamma^{\mathcal{S}}_{c,t} + \frac{2q^{\mathcal{S}}_{2,t}\sigma_b^2}{\tau B} \triangleq \mathcal{A}^{\mathcal{S}}_{b,t}. \tag{33}$$

Under partial participation,

$$\mathbb{E}_t\|P_b u_t\|^2 \le 2\mathcal{A}^{\mathcal{S}}_{b,t} + 2V^{\mathrm{part}}_{b,t} + 2\|P_b b_t\|^2 \triangleq \overline{\mathcal{A}}_{b,t}. \tag{34}$$

Proof. Set $\delta^t_{k,s} = \theta^t_{k,s} - \theta_t$. For every block c,

$$P_c\delta^t_{k,s} = -\eta_\ell \sum_{r=0}^{s-1} \widehat{g}_{c,k}(\theta^t_{k,r}), \tag{35}$$

$$\mathbb{E}_t\|P_c\delta^t_{k,s}\|^2 \le \eta_\ell^2 s \sum_{r=0}^{s-1} \mathbb{E}_t\|\widehat{g}_{c,k}(\theta^t_{k,r})\|^2. \tag{36}$$

Multiply the pointwise Cauchy–Schwarz bound by the selected-client weights, sum over clients, and then take $\mathbb{E}_t$. Applying the selected-client mixed-state moment bound gives

$$\mathbb{E}_t\left[\sum_{k\in\mathcal{S}_t} \widetilde{p}_{k,t}\|P_c\delta^t_{k,s}\|^2\right] \le \eta_\ell^2 s^2 \Gamma^{\mathcal{S}}_{c,t}. \tag{37}$$

For $d^{\mathcal{S}}_{b,t} = P_b d^{\mathcal{S}}_t$, Jensen's inequality and Eq. (18) give

$$\begin{aligned}
\mathbb{E}_t\|d^{\mathcal{S}}_{b,t}\|^2 &\le \frac{1}{\tau}\sum_{s=0}^{\tau-1} \mathbb{E}_t\left[\sum_{k\in\mathcal{S}_t} \widetilde{p}_{k,t}\|g_{b,k}(\theta^t_{k,s}) - g_{b,k}(\theta_t)\|^2\right] \\
&\le \frac{1}{\tau}\sum_{s=0}^{\tau-1}\sum_{c\in\mathcal{B}} L^2_{bc}\mathbb{E}_t\left[\sum_{k\in\mathcal{S}_t} \widetilde{p}_{k,t}\|P_c\delta^t_{k,s}\|^2\right] \\
&\le \eta_\ell^2 c_\tau \sum_{c\in\mathcal{B}} L^2_{bc}\Gamma^{\mathcal{S}}_{c,t}.
\end{aligned} \tag{38}$$

Conditional centering and the fresh-noise assumptions eliminate cross terms. The random participation weights remain inside the expectation:

$$\mathbb{E}_t\|\xi^{\mathcal{S}}_{b,t}\|^2 = \frac{1}{\tau^2}\sum_{s=0}^{\tau-1} \mathbb{E}_t\left[\sum_{k\in\mathcal{S}_t} \widetilde{p}^2_{k,t}\|\zeta_{b,k,s}\|^2\right] \le \frac{q^{\mathcal{S}}_{2,t}\sigma_b^2}{\tau B}. \tag{39}$$

Finally, $\|x+y\|^2 \le 2\|x\|^2 + 2\|y\|^2$ proves Eqs. (33) and (34). □

Theorem 1 (Federated descent bound)**.** *If $0 < \eta \le 1/(4L)$ and $\mathcal{E}_t = \mathbb{E}_t\|u_t\|^2$, then*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(\theta_t)\|^2 \le \frac{4\Delta_0}{\eta T} + \frac{4}{T}\sum_{t=0}^{T-1} \mathbb{E}\mathcal{E}_t. \tag{40}$$

Moreover,

$$\begin{aligned}
\mathcal{E}_t &= \sum_{b\in\mathcal{B}} \mathbb{E}_t\|P_b u_t\|^2 \\
&\le \sum_{b\in\mathcal{B}} \overline{\mathcal{A}}_{b,t} \triangleq \mathcal{A}^{\mathrm{fed}}_t.
\end{aligned} \tag{41}$$

Proof. For $g_t = \nabla F(\theta_t)$, smoothness of the reference objective gives

$$F(\theta_{t+1}) \le F(\theta_t) - \eta\langle g_t, g_t + u_t\rangle + \frac{L\eta^2}{2}\|g_t + u_t\|^2.$$

Using $|\langle g, u\rangle| \le (\|g\|^2 + \|u\|^2)/2$ and $0 < \eta \le 1/(4L)$ yields

$$F(\theta_{t+1}) \le F(\theta_t) - \frac{\eta}{4}\|g_t\|^2 + \eta\|u_t\|^2. \tag{42}$$

Taking expectations and telescoping with $F(\theta_T) \ge F_*$ gives the rate. The block budget follows from the orthogonal decomposition $\|u_t\|^2 = \sum_b \|P_b u_t\|^2$ and Eq. (34). □

D.4 Cross-Client Cancellation of Weak-Modality Updates

Let $v_{b,k,t}$ denote the expected local update of block b before server averaging and let $\bar{v}_{b,t} = \sum_k p_k v_{b,k,t}$. Define

$$E_{b,t}^{\mathrm{loc}} = \sum_k p_k \|v_{b,k,t}\|^2, \qquad V_{b,t}^{\mathrm{dir}} = \sum_k p_k \|v_{b,k,t} - \bar{v}_{b,t}\|^2. \tag{43}$$

Expansion around $\bar{v}_{b,t}$ gives the exact cancellation identity

$$E_{b,t}^{\mathrm{loc}} = \|\bar{v}_{b,t}\|^2 + V_{b,t}^{\mathrm{dir}}. \tag{44}$$

For $E_{b,t}^{\mathrm{loc}} > 0$, the fraction of local update energy surviving aggregation is

$$\rho_{b,t} = \frac{\|\bar{v}_{b,t}\|^2}{E_{b,t}^{\mathrm{loc}}} = 1 - \frac{V_{b,t}^{\mathrm{dir}}}{E_{b,t}^{\mathrm{loc}}} \in [0,1]. \tag{45}$$

If $E_{b,t}^{\mathrm{loc}} = 0$, all weighted client updates vanish; we set the coherence ratio to zero by convention, without interpreting it as cancellation. The fraction is relative to the energy of the included client updates. A comparison with centralized training requires matched states and budgets. In particular, one exact full-participation local step equals the centralized gradient step, even when the coherence ratio is small. For one exact local step, $v_{b,k,t} = -\eta_\ell g_{b,k}(\theta_t)$, and

$$\rho_{b,t} = \frac{\|g_b(\theta_t)\|^2}{\|g_b(\theta_t)\|^2 + H_b^2(\theta_t)}. \tag{46}$$

For any unit direction $q_{b,j}$,

$$h_{b,j,t}^2 = \sum_k p_k \langle g_{b,k} - g_b, q_{b,j}\rangle^2, \tag{47}$$

$$\rho_{b,j,t} = \frac{\langle g_b, q_{b,j}\rangle^2}{\langle g_b, q_{b,j}\rangle^2 + h_{b,j,t}^2}. \tag{48}$$

The directional ratio is also set to zero when its denominator vanishes. Large local gradient norms can coexist with a vanishing server update: if $\langle g_b, q_{b,j}\rangle = 0$ but $h_{b,j,t} > 0$, then $\rho_{b,j,t} = 0$.

Proposition 1 (Weak-block signal obstruction)**.** *For a block with $S_{b,t} > 0$, let*

$$S_{b,t} = \|g_b(\theta_t)\|^2, \qquad \delta_{b,t} = \sqrt{\frac{\overline{\mathcal{A}}_{b,t}}{S_{b,t}}}. \tag{49}$$

The block update obeys

$$\mathbb{E}_t\langle g_b, g_b + P_b u_t\rangle \geq S_{b,t} - \sqrt{S_{b,t}\overline{\mathcal{A}}_{b,t}} = S_{b,t}(1 - \delta_{b,t}). \tag{50}$$

When $S_{b,t} = 0$, the inner product is zero and no positive alignment is certified. For $S_{a,t}, S_{w,t} > 0$, if

$$\overline{\mathcal{A}}_{w,t} \geq S_{w,t}, \qquad \overline{\mathcal{A}}_{a,t} \leq c_1^2 S_{a,t}, \qquad 0 < c_1 < 1, \tag{51}$$

the dominant block has a positive certified descent alignment of at least $(1 - c_1)S_{a,t}$, whereas the weak-block lower bound is nonpositive. This is a loss of a sufficient certificate, not proof that the realized weak-block alignment is nonpositive.

Proof. Conditional Cauchy–Schwarz and Jensen give

$$\begin{aligned}
\mathbb{E}_t\langle g_b, P_b u_t\rangle &\geq -\|g_b\|\mathbb{E}_t\|P_b u_t\| \\
&\geq -\sqrt{S_{b,t}}\sqrt{\mathbb{E}_t\|P_b u_t\|^2} \\
&\geq -\sqrt{S_{b,t}\overline{\mathcal{A}}_{b,t}}.
\end{aligned} \tag{52}$$

Substitution yields the result. □

D.5 From Update Cancellation to Participation-Rank Collapse

Motivated by the low-rank simplicity bias of deep representations (Huh et al., 2023), let $Z_m \in \mathbb{R}^{n\times d}$ be the centered feature matrix of modality m on a fixed reference population and define $C_m = n^{-1}Z_m^\top Z_m$. For a nonzero covariance, use the same scale-invariant rank as in the main text. For the zero-covariance case set the rank to zero:

$$r(C_m) = \begin{cases} \operatorname{tr}(C_m)^2/\operatorname{tr}(C_m^2), & \operatorname{tr}(C_m) > 0, \\ 0, & \operatorname{tr}(C_m) = 0. \end{cases} \tag{53}$$

Suppose one infinitesimal optimization step changes Z to $Z^+ = Z - \gamma D$, where D is the centered feature-displacement field per unit step. The following formulas concern positive, differentiable eigenvalue branches; repeated eigenvalues require a compatible directional eigenbasis. Then

$$C^+ - C = -\frac{\gamma}{n}(Z^\top D + D^\top Z) + \frac{\gamma^2}{n}D^\top D, \tag{54}$$

$$\dot{C} = -\frac{1}{n}(Z^\top D + D^\top Z). \tag{55}$$

For $Cq_j = \lambda_j q_j$, $\|q_j\| = 1$, and $\lambda_j > 0$,

$$\dot{\lambda}_j = q_j^\top \dot{C} q_j = -\frac{2}{n}\langle Zq_j, Dq_j\rangle = -\kappa_j\lambda_j, \tag{56}$$

$$\kappa_j = \frac{2\langle Zq_j, Dq_j\rangle}{n\lambda_j}, \qquad \lambda_j(s) = \lambda_j(0)e^{-\int_0^s \kappa_j(u)\,\mathrm{d}u}. \tag{57}$$

Lemma 3 (Tail-selective spectral contraction)**.** *Let $\lambda_1 \geq \cdots \geq \lambda_d > 0$ and $\dot{\lambda}_j = -\kappa_j\lambda_j$ with $0 \leq \kappa_1 \leq \cdots \leq \kappa_d$. For*

$$\mathbb{E}_{\lambda^q}[\kappa] = \frac{\sum_j \lambda_j^q \kappa_j}{\sum_j \lambda_j^q}, \qquad q \in \{1, 2\}, \tag{58}$$

we have

$$\frac{\mathrm{d}}{\mathrm{d}s}\log r(C(s)) = 2\left(\mathbb{E}_{\lambda^2}[\kappa] - \mathbb{E}_\lambda[\kappa]\right) \leq 0. \tag{59}$$

The inequality is strict if some $i < j$ satisfy $\lambda_i > \lambda_j$ and $\kappa_i < \kappa_j$.

Proof. Let $S_q = \sum_j \lambda_j^q$. Direct differentiation gives

$$\frac{\mathrm{d}}{\mathrm{d}s}\log\frac{S_1^2}{S_2} = -2\frac{\sum_j \kappa_j\lambda_j}{S_1} + 2\frac{\sum_j \kappa_j\lambda_j^2}{S_2}. \tag{60}$$

Moreover,

$$\mathbb{E}_\lambda[\kappa] - \mathbb{E}_{\lambda^2}[\kappa] = \frac{1}{S_1 S_2}\sum_{i<j}\lambda_i\lambda_j(\lambda_i - \lambda_j)(\kappa_j - \kappa_i) \geq 0. \tag{61}$$

Substitution into Eq. (60) proves the claim. □

Let D_w^{C} and D_w^{F} denote the centralized and federated feature-displacement fields evaluated at a common state. Their instantaneous rate difference is

$$\Delta_{w,j} = \kappa_{w,j}^{\mathrm{F}} - \kappa_{w,j}^{\mathrm{C}} = \frac{2\langle Z_w q_{w,j}, (D_w^{\mathrm{F}} - D_w^{\mathrm{C}})q_{w,j}\rangle}{n\lambda_{w,j}}. \tag{62}$$

Since $\|Z_w q_{w,j}\|^2 = n\lambda_{w,j}$,

$$|\Delta_{w,j}| \leq \frac{2\|(D_w^{\mathrm{F}} - D_w^{\mathrm{C}})q_{w,j}\|}{\sqrt{n\lambda_{w,j}}}. \tag{63}$$

If the feature map has directional parameter sensitivity $J_{w,j}$, i.e.,

$$\mathbb{E}_t\|(D_w^{\mathrm{F}} - D_w^{\mathrm{C}})q_{w,j}\|^2 \leq J_{w,j}^2\overline{\mathcal{A}}_{w,t}, \tag{64}$$

then

$$\mathbb{E}_t|\Delta_{w,j}| \le \frac{2J_{w,j}}{\sqrt{n\lambda_{w,j}}}\sqrt{2\mathcal{A}^{\mathcal{S}}_{w,t} + 2V^{\mathrm{part}}_{w,t} + 2\|P_w b_t\|^2}. \tag{65}$$

For fixed directional sensitivity $J_{w,j}$, the rate-perturbation bound grows as $\lambda_{w,j}$ decreases. The following theorem states the additional sign and ordering conditions that turn this sensitivity into a rank comparison. Uniform covariance scaling leaves participation rank unchanged.

Theorem 2 (Conditional rank amplification)**.** *Assume a common initial weak spectrum, a common eigenbasis for the centralized and federated covariance flows, and that the centralized ordering is preserved so that $\lambda^{\mathrm{C}}_{w,1}(s) \ge \cdots \ge \lambda^{\mathrm{C}}_{w,d}(s) > 0$ for all s in the analysis window. Assume constant rates on this window and let*

$$\lambda^{\mathrm{C}}_{w,j}(s) = \lambda_{w,j}(0)e^{-\kappa^{\mathrm{C}}_{w,j}s}, \qquad \lambda^{\mathrm{F}}_{w,j}(s) = \lambda_{w,j}(0)e^{-(\kappa^{\mathrm{C}}_{w,j}+\Delta_{w,j})s}. \tag{66}$$

If

$$0 \le \Delta_{w,1} \le \cdots \le \Delta_{w,d}, \qquad \Delta_{w,1} < \Delta_{w,d}, \tag{67}$$

that is, if the additional federated contraction is tail-prioritized, then the federated rank is no larger than the centralized rank:

$$r(C^{\mathrm{F}}_w(s)) \le r(C^{\mathrm{C}}_w(s)), \qquad s \ge 0, \tag{68}$$

throughout the analysis window, with strict inequality for $s > 0$ under the stated positive-spectrum and nonconstant-extra-rate assumptions. These rate assumptions are additional conditions on the two flows, not consequences of the FedAvg error bound.

Proof. Set $x_j = \lambda^{\mathrm{C}}_{w,j}(s)$, $q_j = e^{-\Delta_{w,j}s}$, and

$$p^{\mathrm{C}}_j = \frac{x_j}{\sum_i x_i}, \qquad p^{\mathrm{F}}_j = \frac{x_j q_j}{\sum_i x_i q_i}. \tag{69}$$

The ordering assumption makes both x_j and q_j nonincreasing in j. For every $1 \le r < d$,

$$\sum_{j=1}^{r} p^{\mathrm{F}}_j - \sum_{j=1}^{r} p^{\mathrm{C}}_j = \frac{\sum_{i\le r}\sum_{j>r} x_i x_j (q_i - q_j)}{(\sum_i x_i q_i)(\sum_i x_i)} \ge 0. \tag{70}$$

Thus p^{F} majorizes p^{C}. Since

$$r(C) = \frac{1}{\sum_j p_j^2} \tag{71}$$

is Schur-concave, Eq. (68) follows. For $s > 0$, $q_1 > q_d$ and $x_1 x_d > 0$ make the prefix inequality at $r = 1$ strict. The two ordered probability vectors are therefore distinct. Strict convexity of $\sum_j p_j^2$ gives $\sum_j (p^{\mathrm{F}}_j)^2 > \sum_j (p^{\mathrm{C}}_j)^2$, proving the strict rank inequality. For a spectrum with zero eigenvalues, the same argument applies on its positive support: strictness requires two positive initial eigenvalues with different extra contraction rates. A constant rate on that support rescales the covariance and leaves participation rank unchanged. □

D.6 Capacity-Preserving Reallocation

The variance–covariance constraint follows the decorrelation principle used by Bardes et al. (2022). Write the reconstruction in Eq. (5) as

$$\widetilde{f}_m = R_{s,m}s_m + R_{p,m}p_m, \qquad B_m = \begin{bmatrix} B_{s,m} \\ B_{p,m} \end{bmatrix}, \qquad R_m = \begin{bmatrix} R_{s,m} & R_{p,m} \end{bmatrix}. \tag{72}$$

Proposition 2 (Function-preserving initialization)**.** *If $R^0_m B^0_m = I$, then for every feature f_m,*

$$\widetilde{f}^0_m = R^0_m B^0_m f_m = f_m, \qquad \frac{\partial \widetilde{f}^0_m}{\partial f_m} = I. \tag{73}$$

Therefore the initial fused logits, task loss, and encoder task gradient are identical before and after reparameterization.

Proof. Substituting $R_m^0 B_m^0 = I$ leaves the reconstructed features and hence the fused logits and task loss unchanged. The chain rule multiplies the task gradient by $(R_m^0 B_m^0)^\top = I$, preserving the encoder task gradient. □

Fix one model and one evaluation distribution, and let $C_{p_w} = \mathrm{Cov}(p_w)$ have dimension d_p. Suppose its measured variance–covariance quantities satisfy

$$\frac{1}{d_p}\sum_{j=1}^{d_p}\left[\nu - \sqrt{(C_{p_w})_{jj} + \epsilon_0}\right]_+^2 \le \varepsilon_v, \qquad \|\operatorname{off}(C_{p_w})\|_\infty \le \varepsilon_o. \tag{74}$$

Here $\|\operatorname{off}(C)\|_\infty$ is the maximum absolute row sum, not the maximum entry. On the same evaluation covariance, suppose the normalized off-diagonal penalty in Eq. (8) is at most $\varepsilon_{\mathrm{cov}}$ and the shared–private penalty is at most $\varepsilon_{\mathrm{orth}}$. For $d_p > 1$, row-wise Cauchy–Schwarz and the spectral-to-Frobenius norm bound give

$$\begin{aligned}\|\operatorname{off}(C_{p_w})\|_\infty &\le \sqrt{d_p - 1}\,\|\operatorname{off}(C_{p_w})\|_F \le (d_p - 1)\sqrt{d_p \varepsilon_{\mathrm{cov}}},\\ \|\operatorname{Cov}(s_w, p_w)\|_2 &\le \|\operatorname{Cov}(s_w, p_w)\|_F \le \sqrt{d_s d_p \varepsilon_{\mathrm{orth}}}.\end{aligned}$$

These expressions connect the loss normalizations to the spectral tolerances. Every diagonal element then obeys

$$(C_{p_w})_{jj} \ge \left(\nu - \sqrt{d_p \varepsilon_v}\right)_+^2 - \epsilon_0 \triangleq \underline{v}. \tag{75}$$

Assume $\underline{v} > \varepsilon_o$ and separately require $\|\operatorname{Cov}(s_w, p_w)\|_2 \le \varepsilon_{sp}$. Gershgorin's theorem then gives the eigenvalue bound below; the cross-covariance inequality is an assumption, not a consequence of that theorem:

$$\lambda_{\min}(C_{p_w}) \ge \underline{v} - \varepsilon_o \triangleq \underline{\lambda}_p > 0, \qquad \|\operatorname{Cov}(s_w, p_w)\|_2 \le \varepsilon_{sp}. \tag{76}$$

Under the additional spectral upper bound $\lambda_{\max}(C_{p_w}) \le \overline{\lambda}_p$, one obtains

$$r(C_{p_w}) = \frac{(\sum_j \lambda_j)^2}{\sum_j \lambda_j^2} \ge d_p \left(\frac{\underline{\lambda}_p}{\overline{\lambda}_p}\right)^2. \tag{77}$$

These are conditional bounds on a measured covariance, not automatic consequences of optimizing a finite-weight penalty. In particular, for a centered batch of size b, $\mathrm{rank}(C_p) \le b-1$, so a positive full-private-space eigenvalue floor cannot hold on that batch when $b \le d_p$. The positive-floor condition must instead be checked on a suitable evaluation population or a sufficiently large reference set. The variance penalty encourages a diagonal floor but does not enforce the spectral upper bound: $C_p = \mathrm{diag}(a^2, 1, \ldots, 1)$ can satisfy the variance floor and have zero off-diagonal penalty while $r(C_p) \to 1$ as $a \to \infty$. Positive variance therefore differs from high participation rank. Assuming also that R_p has full column rank,

$$C_{\tilde{f}_w} = R_s C_s R_s^\top + R_p C_p R_p^\top + R_s C_{sp} R_p^\top + R_p C_{sp}^\top R_s^\top, \tag{78}$$

$$U_p = R_p (R_p^\top R_p)^{-1/2}, \qquad U_p^\top U_p = I, \tag{79}$$

$$\lambda_{\min}\left(U_p^\top C_{\tilde{f}_w} U_p\right) \ge \sigma_{\min}(R_p)^2 \underline{\lambda}_p - 2\|R_s\|_2 \|R_p\|_2 \varepsilon_{sp}. \tag{80}$$

The bound applies to the nonzero private reconstruction range. When R_p has fewer than d columns, it gives no full-space eigenvalue bound. The evaluated model retains nonzero weak-private directions whenever the right-hand side of Eq. (80) is positive.

Evaluate the stated conditions on the model to which the conclusion is applied; parameter averaging can change them. For example, take $B_1 = R_1 = I$ and $B_2 = R_2 = \mathrm{diag}(I_{d_s}, -I_{d_p})$. With independent shared and private features and $C_p = I$, both clients reconstruct exactly and satisfy the local covariance conditions, but averaging B_k and R_k annihilates the private block. Shared prototypes cannot resolve that private sign ambiguity. An aggregation-level capacity guarantee therefore requires separate control of coordinate drift after initialization or direct verification of the covariance and reconstruction conditions on the aggregated model. Identity-preserving initialization alone does not supply such control, and the conditional bounds do not assert that the private range is predictive.

If the reconstruction term satisfies $n^{-1}\|\widetilde{Z}_w - Z_w\|_F^2 \le \varepsilon_r$, then

$$\begin{aligned}\|C_{\widetilde{f}_w} - C_{f_w}\|_2 &\le \frac{2}{n}\|Z_w\|_2\|\widetilde{Z}_w - Z_w\|_2 + \frac{1}{n}\|\widetilde{Z}_w - Z_w\|_2^2 \\ &\le 2\sqrt{\mathrm{tr}(C_{f_w})\varepsilon_r} + \varepsilon_r. \end{aligned} \tag{81}$$

For the centered error $E = \widetilde{Z}_w - Z_w$, the preceding bound follows from $C_{\widetilde{f}_w} - C_{f_w} = n^{-1}(Z_w^\top E + E^\top Z_w + E^\top E)$, $\|E\|_2/\sqrt{n} \le \sqrt{\varepsilon_r}$, and $\|Z_w\|_2/\sqrt{n} \le \sqrt{\mathrm{tr}(C_{f_w})}$. Centering is an orthogonal projection, so an uncentered reconstruction-error bound also bounds this centered error.

D.7 Dominance-Gated Transfer and Semantic Descent

DGT uses temperature-scaled knowledge distillation, restricted by the dominance and reliability gates (Hinton et al., 2015; Garcia et al., 2018). Freeze the prototype bank, teacher targets, gates, and all parameter blocks except θ_w. Let F^{sem} be the weighted sum of fusion and direct unimodal/shared supervised losses, excluding KD, CPR, and the remaining FSA penalties. With the frozen quantities fixed it is an ordinary local objective. Here L denotes a smoothness constant for F^{sem} in θ_w over the step under analysis; this smoothness is assumed separately from that of the complete objective. Set $g_{w,t}^{\mathrm{sem}} = \nabla_{\theta_w} F^{\mathrm{sem}}(\theta_t)$ and define

$$h_{w,t} = \sum_{k:w\in\mathcal{W}_k} p_k \frac{\lambda_d \alpha_{w,k} q_{a_k,k}}{\max(1, |\mathcal{W}_k|)} \nabla_{\theta_w} \mathcal{L}_{\mathrm{KD},w,k}(\theta_t), \qquad \beta_t = 1. \tag{82}$$

Here $\mathcal{L}_{\mathrm{KD},w,k}$ includes the temperature-squared factor. All client-specific weights and teacher roles are absorbed into $h_{w,t}$; $\beta_t = 1$ preserves the notation below. Assume independently that

$$\langle h_{w,t}, g_{w,t}^{\mathrm{sem}} \rangle \ge \kappa_t \|g_{w,t}^{\mathrm{sem}}\|^2, \qquad \|h_{w,t}\| \le B_t, \qquad \kappa_t \ge 0. \tag{83}$$

The loss, rank, and confidence gates select eligible batches; alignment in Eq. (83) is a separate condition on their gradients. With $\delta_{\min} > 0$, the rank factor equals one whenever the gate is active. The analysis fixes teacher distributions as well as gate statistics; changes in targets or shared parameters enter the residual budget below. The first-order gain from DGT is

$$\beta_t \langle h_{w,t}, g_{w,t}^{\mathrm{sem}} \rangle \ge \beta_t \kappa_t \|g_{w,t}^{\mathrm{sem}}\|^2. \tag{84}$$

For the ideal weak update $\theta_w^+ = \theta_w - \eta(g_{w,t}^{\mathrm{sem}} + \beta_t h_{w,t})$, smoothness gives the full quadratic expression

$$\begin{aligned} F^{\mathrm{sem}}(\theta_w^+) &\le F^{\mathrm{sem}}(\theta_w) - \eta\|g_{w,t}^{\mathrm{sem}}\|^2 - \eta\beta_t\langle g_{w,t}^{\mathrm{sem}}, h_{w,t}\rangle \\ &\quad + \frac{L\eta^2}{2}\|g_{w,t}^{\mathrm{sem}} + \beta_t h_{w,t}\|^2 \\ &\le F^{\mathrm{sem}}(\theta_w) - \eta(1+\beta_t\kappa_t)S_{w,t}^{\mathrm{sem}} + \frac{L\eta^2}{2}\left(\sqrt{S_{w,t}^{\mathrm{sem}}} + \beta_t B_t\right)^2, \end{aligned} \tag{85}$$

where $S_{w,t}^{\mathrm{sem}} = \|g_{w,t}^{\mathrm{sem}}\|^2$. Because $h_{w,t} = 0$ when all relevant gates close, no KD update is then added. The alignment condition can be evaluated through KD–supervised gradient inner products on the selected batches.

D.8 Federated Semantic Anchoring and Class Coverage

FSA extends class-prototype coordination in federated learning (Tan et al., 2022) with modality reliability and class-coverage weights. Assume $0 \le \mu < 1$ and a fixed target P_c^* over the analysis window. Let $\mathcal{G}_t$ contain the past state, selected clients, current quality masses, and the update acceptance decision. The random step size η_c^t and Q_c^t are $\mathcal{G}_t$-measurable. Define $Q_c^t = \sum_{k\in\mathcal{S}_t^c} Q_{c,k}^t$ and $w_{c,k}^t = Q_{c,k}^t/Q_c^t$ for $Q_c^t > 0$. Assume the normalized aggregate has the following conditional bias and variance bounds, with a uniform budget $\varsigma_c^2 \ge \sum_k w_{c,k}^t \varsigma_{c,k}^2$:

$$\begin{aligned} \widehat{P}_c^{t+1} = P_c^* + b_c^t + \varepsilon_c^t, \qquad \mathbb{E}[\varepsilon_c^t \mid \mathcal{G}_t] = 0, \qquad \|b_c^t\| \le b_c, \\ \mathbb{E}[\|\varepsilon_c^t\|^2 \mid \mathcal{G}_t] \le \sum_{k\in\mathcal{S}_t^c} (w_{c,k}^t)^2 \frac{\varsigma_{c,k}^2}{Q_{c,k}^t} \le \frac{\varsigma_c^2}{Q_c^t}\mathbf{1}\{Q_c^t > 0\}. \end{aligned} \tag{86}$$

Here $b_c^t = \mathbb{E}[\widehat{P}_c^{t+1} \mid \mathcal{G}_t] - P_c^*$ is $\mathcal{G}_t$-measurable. The conditional bias and variance bounds apply to the normalized, quality-weighted aggregate and are specified separately from independent raw-sample assumptions. This conditioning also accounts for selection effects from the update-acceptance rule. The indicator in Eq. (86) defines the ratio as zero when $Q_c^t = 0$. For rejected updates set $\eta_c^t = 0$ and interpret the variance contribution as zero. The raw-state update in Eq. (10) is

$$\widetilde{P}_c^{t+1} = (1 - \eta_c^t)\widetilde{P}_c^t + \eta_c^t \widehat{P}_c^{t+1}, \qquad \eta_c^t = \mathbf{1}\{\text{update accepted}\}(1-\mu)\Xi_c^t. \tag{87}$$

The bias term is retained because quality weighting and normalization generally produce a biased estimator. For any $\nu_b > 0$, subtracting P_c^* and applying Young's inequality yields

$$D_c^{t+1} \le \mathbb{E}\left[(1+\nu_b)(1-\eta_c^t)^2\|\widetilde{P}_c^t - P_c^*\|^2 + (1+\nu_b^{-1})(\eta_c^t)^2\left(b_c^2 + \frac{\varsigma_c^2}{Q_c^t}\mathbf{1}\{Q_c^t > 0\}\right)\right]. \tag{88}$$

Here $D_c^t = \mathbb{E}\|\widetilde{P}_c^t - P_c^*\|^2$; random coefficients remain inside the expectation, so no factorization with D_c^t is assumed. For rejected or zero-support rounds, $\eta_c^t = 0$ and the raw state is unchanged; the displayed bound is then conservative and does not imply contraction. For deterministic constants $0 < \eta_c \le 1$, $Q_c > 0$, and b_c, set $\kappa_b = (1+\nu_b)(1-\eta_c)^2 < 1$ and $J_c = (1+\nu_b^{-1})\eta_c^2(b_c^2 + \varsigma_c^2/Q_c)$. Iterating the scalar recursion yields

$$D_c^t \le \kappa_b^t D_c^0 + J_c\frac{1-\kappa_b^t}{1-\kappa_b}, \qquad \limsup_{t\to\infty} D_c^t \le \frac{J_c}{1-\kappa_b}. \tag{89}$$

The zero-bias case is recovered by setting $b_c = 0$. If stored states and their reference targets have norm at least $r_0 > 0$, then

$$\|\operatorname{Norm}(x) - \operatorname{Norm}(y)\| \le \frac{2}{r_0}\|x - y\|, \tag{90}$$

Thus the normalized copies are Lipschitz transforms of the stored state. The contraction bound concerns the stored pre-normalized state. Normalized copies enter through the preceding Lipschitz bound. The norm floor is an additional assumption, not a consequence of accepting only nonzero proposals: an EMA of two nonzero vectors can still approach or reach zero. Normalization and its Lipschitz bound apply only on a window where the raw states and targets satisfy the stated positive floor. A moving target would require an additional tracking-error term.

For a frozen, prototype-dependent local objective, assume each client gradient is L_P-Lipschitz in the stored raw bank, with normalization included under the positive norm floor. All other parameters and coefficients are held fixed in this comparison. Let $\widetilde{H}_{w,t}^2$ denote expected gradient heterogeneity and $H_{w,*}^2$ its value at the reference bank at the same parameter state. Bounding deviations from the reference gradients gives

$$\widetilde{H}_{w,t}^2 \le 2H_{w,*}^2 + 2L_P^2\sum_{c=1}^{C} D_c^t, \tag{91}$$

$$\mathcal{A}_{w,t}^{\mathcal{S},\mathrm{SBR}} = 2\eta_\ell^2 c_\tau \sum_{c\in\mathcal{B}} L_{wc}^2 \Gamma_{c,t}^{\mathcal{S},\mathrm{SBR}} + \frac{2q_{2,t}^{\mathcal{S}}\sigma_w^2}{\tau B}. \tag{92}$$

Here $\Gamma_{c,t}^{\mathcal{S},\mathrm{SBR}}$ bounds the selected-client mixed-state gradient second moment, and $q_{2,t}^{\mathcal{S}} = \mathbb{E}_t\sum_{k\in\mathcal{S}_t}\widetilde{p}_{k,t}^2$. The moment bound must be established for the frozen complete client objective; a heterogeneity bound alone is insufficient. When the recursion contracts, it controls a prototype-error term in the preceding upper bound. Actual gradient heterogeneity and drift require separate comparison measurements against training without FSA. The Lipschitz constant includes normalization under the stated norm floor; the gradient moment assumptions must hold for the same frozen objective and aggregation weights.

For normalized s and prototypes P_c, define

$$\gamma_y(s) = s^\top P_y - \max_{c\in\mathcal{C}_t\setminus\{y\}} s^\top P_c. \tag{93}$$

For $y \in \mathcal{C}_t$ and $|\mathcal{C}_t| \ge 2$, the coverage loss in Eq. (11) is exactly

$$\mathcal{R}_{\mathrm{cov}}(s, y; P) = a_y \log\left(1 + \sum_{c\in\mathcal{C}_t\setminus\{y\}} \exp\left[-\frac{s^\top P_y - s^\top P_c}{\tau_p}\right]\right). \tag{94}$$

Therefore, if $\mathcal{R}_{\mathrm{cov}}(s, y; P) \le \varepsilon_y$, then every $c \in \mathcal{C}_t \setminus \{y\}$ satisfies

$$s^\top P_y - s^\top P_c \ge -\tau_p \log\left(e^{\varepsilon_y/a_y} - 1\right), \tag{95}$$

which is strictly positive whenever $\varepsilon_y < a_y \log 2$. The guarantee is per sample and covers competitors present in the loss denominator; unseen-class coverage and population recall require separate evidence.

D.9 Combined FedSBR Certificate

Analyze one weak-block step with all other blocks, prototypes, teacher targets, and detached coefficients frozen. Decompose $r_{w,t} = e^{\mathrm{fed}}_{w,t} + a_{w,t}$, where $e^{\mathrm{fed}}_{w,t}$ is the deviation from the full-client gradient of that frozen objective and $a_{w,t}$ collects the remaining CPR and FSA gradients excluded from F^{sem}. Any change of the frozen quantities along actual local steps must also be included in $a_{w,t}$, as must deviations from the reference local-SGD recursion, including momentum, unequal step counts, and any different aggregation weights. Assume the conditional auxiliary-gradient budget $\mathbb{E}_t\|a_{w,t}\|^2 \le U_{w,t}$; no independence from $e^{\mathrm{fed}}_{w,t}$ is required, and put $\overline{\mathcal{A}}^{\mathrm{SBR}}_{w,t} = 2\mathcal{A}^{\mathcal{S},\mathrm{SBR}}_{w,t} + 2V^{\mathrm{part}}_{w,t} + 2\|P_w b_t\|^2$. Minkowski's inequality gives

$$\mathbb{E}_t\|r_{w,t}\|^2 \le \left(\sqrt{2\mathcal{A}^{\mathcal{S},\mathrm{SBR}}_{w,t} + 2V^{\mathrm{part}}_{w,t} + 2\|P_w b_t\|^2} + \sqrt{U_{w,t}}\right)^2 \triangleq \mathcal{R}_{w,t}. \tag{96}$$

Theorem 3 (Conditional FedSBR collapse certificate)**.** *Assume Eqs. (76), (83), and (88). Put $S = S^{\mathrm{sem}}_{w,t}$, $A = \mathcal{R}_{w,t}$, and*

$$\mathfrak{D}_t = (1 + \beta_t\kappa_t)S - \sqrt{SA} - \frac{L\eta}{2}\left(\sqrt{S} + \beta_t B_t + \sqrt{A}\right)^2. \tag{97}$$

Under the residual budget above and L-smoothness in the frozen weak block, if $\mathfrak{D}_t > 0$, then

$$\mathbb{E}_t F^{\mathrm{sem}}(\theta_{w,t+1}) \le F^{\mathrm{sem}}(\theta_{w,t}) - \eta\mathfrak{D}_t. \tag{98}$$

If additionally the restricted private-range bound in Eq. (80) is positive, the private component has nonzero variance in that range. Full-space collapse remains outside the stated bound. If, for a sample with $y \in \mathcal{C}_t$ and $|\mathcal{C}_t| \ge 2$, $\mathcal{R}_{\mathrm{cov}} \le \varepsilon_y < a_y \log 2$, then its class-$y$ prototype margin against supported competing classes is positive by Eq. (95). The capacity conditions apply to the evaluated model, including the server model if a server claim is intended. The descent statement conditions on fixed updates of the other blocks and the prototype bank.

Proof. The decomposed update is $\theta_{w,t+1} = \theta_{w,t} - \eta(g^{\mathrm{sem}}_{w,t} + \beta_t h_{w,t} + r_{w,t})$. Conditional Cauchy–Schwarz gives $\mathbb{E}_t\langle g^{\mathrm{sem}}_{w,t}, r_{w,t}\rangle \ge -\sqrt{SA}$; Minkowski in L_2 gives $(\mathbb{E}_t\|g^{\mathrm{sem}}_{w,t} + \beta_t h_{w,t} + r_{w,t}\|^2)^{1/2} \le \sqrt{S} + \beta_t B_t + \sqrt{A}$. Together with Eq. (83), conditional smoothness yields

$$\mathbb{E}_t F^{\mathrm{sem}}(\theta_{w,t+1}) \le F^{\mathrm{sem}}(\theta_{w,t}) - \eta(1 + \beta_t\kappa_t)S + \eta\sqrt{SA} + \frac{L\eta^2}{2}(\sqrt{S} + \beta_t B_t + \sqrt{A})^2. \tag{99}$$

This is Eq. (98). The private-range variance and supported-class margin follow from Eqs. (80) and (95), respectively. □

The three sufficient conditions act on distinct terms. In the deterministic constant-coefficient case, b_c^2 is the prototype-aggregate bias contribution:

$$\underbrace{\underline{\lambda}_p > 0}_{\mathrm{CPR:private-rangecapacity}}, \qquad \underbrace{\beta_t\kappa_t S}_{\mathrm{DGT:semantic\ gain}}, \qquad \underbrace{D_c^{t+1} \le \kappa_b D_c^t + (1 + \nu_b^{-1})\eta_c^2\left(\underbrace{b_c^2}_{\mathrm{bias}} + \frac{\varsigma_c^2}{Q_c}\right)}_{\mathrm{FSA:\ coordination}}. \tag{100}$$

D.10 Theory-to-Experiment Connections

The reference recursion separates local drift and sampling error from a centralized step. With full participation and one exact local step, the two updates coincide; additional federated effects arise through the local-update trajectory and participation pattern. Cross-client cancellation also applies with concatenation, whose modality Jacobian ranges are orthogonal. The rank theorem then requires ordered extra tail contraction; the update-error bound quantifies sensitivity without determining that ordering.

The module results apply to a fixed evaluation model or a frozen weak-block step. CPR's covariance and reconstruction conditions must hold on the evaluated server model for a server-capacity claim. DGT uses an independently checked gradient-alignment condition, and FSA uses the stated prototype bias, variance, and norm bounds. Changes in gates, targets, and update rules enter the complete residual budget. The component summary uses deterministic prototype coefficients; random quality weights retain Eq. (88).

Table 16 connects these roles to jointly measured capacity, predictive utility, and modality contribution. Its task-only basis control distinguishes the parameterization from the coordinated objectives; the FSA removal tests the value of the shared semantic signal across clients. Table 11 complements rank with class separation, while the attenuation and class-recall diagnostics assess how the predictor uses the available pathways. These outcomes test the method's effects; the sufficient conditions above concern the underlying update and covariance quantities.

E Social Impact

FedSBR keeps multimodal records on their originating devices and exchanges only model parameters and class-level feature summaries, which can reduce direct exposure of speech, images, text, and clinical measurements. Class-level summaries may still leak information through inversion or membership inference, while gate calibration and uneven class participation can produce subgroup disparities. Deployments should therefore use secure aggregation and access controls, report pathway and class-conditional metrics, audit subgroup performance, and retain human review for high-stakes decisions.

F Limitations

The main evaluation covers four datasets, three seeds, and a ten-client synchronous protocol. The analysis gives local sufficient conditions for a non-convex gated objective, with gate calibration, prototype support, and the evaluated population determining their applicability. CMU-MOSEI and UR-FUNNY provide natural next testbeds for language-centered modality imbalance. Further evaluation should examine larger and asynchronous deployments, broader modality scales and participation patterns, and privacy, poisoning robustness, and energy or communication costs.