UniFed-LoRA: Exploiting Semantic Task Correlation for Heterogeneous Multimodal Federated Fine-Tuning

Supplementary Material

6. Dataset and Splits Details

We evaluate on a 20,000-study subset of MIMIC-CXR-JPG comprising chest radiographs, radiology reports, and study-level labels. Each study may include one or more images together with report sections such as Findings and Impression, enabling both visual and textual tasks from the same clinical source. The dataset provides 14 labels—*Atelectasis*, *Cardiomegaly*, *Consolidation*, *Edema*, *Enlarged Cardiomediastinum*, *Fracture*, *Lung Lesion*, *Lung Opacity*, *No Finding*, *Pleural Effusion*, *Pleural Other*, *Pneumonia*, *Pneumothorax*, and *Support Devices*—with *positive*, *negative*, *uncertain*, and *unmentioned* states.

We instantiate three client tasks on this subset: (i) multi-label binary image classification with ViT from chest X-ray images, (ii) 4-way label-state classification with BERT from report text, and (iii) impression generation with GPT-2 from Findings text. The federation uses a disjoint split by unique study or subject identifier so that each sample belongs to exactly one client partition.

To support a federated setup, the selected subset is partitioned deterministically into disjoint client splits using study- and subject-level identifiers, ensuring that each sample is assigned to exactly one client. Each client retains only the samples associated with its assigned identifiers, and any images, reports, or labels belonging to other identifier groups are excluded from that client's local dataset.

Datasets for unrelated clients. For the unrelated-task stress tests, we replace one aligned client with AG News (text topic classification), CIFAR10 (image classification), or Tiny Shakespeare (language modeling) while keeping the remaining medical clients unchanged.

The AG News dataset is a classic English news topic classification benchmark used in NLP, consisting of 120,000 training samples and 7,600 test samples [39]. Each example is a short news item built from a title + description, and the task is to classify it into one of 4 categories: World, Sports, Business, or Sci/Tech.

Tiny Shakespeare is a small text corpus popular in character-level language modeling that contains roughly 40,000 lines of Shakespeare text compiled as a single plain-text file [18]. We use the dataset for plain next-token language modeling over raw text samples.

CIFAR10 is an image classification dataset that contains 60,000 images across 10 classes [19]. We use this dataset to simulate an unrelated client solving the task out of the primary medical domain.

7. Hypernetwork

We use a lightweight hypernetwork that predicts LoRA parameters from discrete metadata describing the target adaptation site, including task type, modality, layer depth, and layer type, with optional client-identity conditioning. Each attribute is represented by a learnable embedding followed by LayerNorm, and the resulting embeddings are concatenated into a single conditioning vector. This vector is processed by a shared MLP backbone consisting of an input mixing block, two residual MLP blocks, and a final LayerNorm-MLP projection, with SiLU activations and dropout throughout, before being mapped by a client-specific linear head to jointly generate LoRA adapters. This design enables a single forward pass to produce the full low-rank update for a given layer while supporting heterogeneous client backbones through separate output heads matched to each base-model dimension.

Initialization Given the sensitivity of hypernetworks to initialization, we follow [3] for model initialization of the output heads: linear layers are Xavier-uniform initialized with zero biases, embeddings are sampled from a Gaussian distribution with zero mean and 0.02 standard deviation. For the client-specific heads, we adopt a bias-based hypernetwork initialization in which head weights are zero-initialized, while the bias corresponding to the first half of the output $\boldsymbol{A}_c$ is initialized from a uniform distribution $U(-\frac{1}{d_c^{in} d_c^{out}}; \frac{1}{d_c^{in} d_c^{out}})$ and the second half $\boldsymbol{B}_c$ is initialized to zero, yielding stable, initial LoRA predictions.

In order to reduce the number of parameters on the server, we adopt the approach of [6] for the federated setting: the *M* (Table 1) variant predicts one LoRA factor at a time and conditions the hypernetwork on an additional A/B embedding, which provides a more parameter-efficient and modular decomposition of the generation process, while doubling the number of forward passes on the server.

7.1. Backbone Models

We evaluate four model families spanning text encoders, vision encoders, and generative language/vision-language models: BERT, ViT, GPT-2, and SmolVLM-256. Table 3 summarizes the models, parameter scales, input modality, and task.

Prediction Heads. For both BERT and ViT, we use the final CLS token representation produced by the pretrained encoder backbone. We use dropout followed by a linear

Table 3. Summary of models used in our experiments.

Model	Params	d^{in}/d^{out}	Input	Task
BERT	109,876,481	768	Text	Multi-label 4-way classification
ViT	86,192,897	768	Image	Multi-label binary classification
GPT-2	124,439,808	768	Text	Text generation / language modeling
SmolVLM-256	256,484,928	768	Image + Text	Unified multi-task baseline

projection to obtain one logit per label for the image encoder. For the text encoder, we use a linear classifier that maps the CLS token to 72 outputs, which are then reshaped into a tensor of size 18×4. The ViT is trained with binary cross-entropy with logits. To account for label imbalance, we use a positive-class weighting term. BERT, for each label, predicts one of four mutually exclusive classes. Training uses cross-entropy loss with class weights to mitigate class imbalance. When labels are missing or masked, the loss is computed only over valid entries.

8. Hyperparameters

Table 4. Training hyperparameters used for the experiments.

Hyperparameter	Value
Optimizer	AdamW
Learning rate η	2×10^{-5}
Weight decay	0.0
Batch size B	32
Number of local epochs E	1
Number of rounds K	300
Number of rounds K (baselines)[a]	100
Number of clients C	3
Number of training samples per client	5223
Number of validation samples per client	989
Number of testing samples per client	454
Learning rate scheduler	cosine
Warmup steps / ratio	0
Max gradient norm	1
$\text{Dropout}_{\text{BERT/ViT}}$	0.5
Maximum text length	512
Rank r	8
Scaling parameter α	32
$\text{Dropout}_{\text{LoRA}}$	0.05
Embedding dimension d^{emb}	32
Number of MLP layers	4
Hidden dimension	256
$\text{Dropout}_{\text{HN}}$	0.05

[a] All baselines select 3 clients per round, while our approach randomly selects one client.

9. Additional Discussion and Scope

Scope and applicability. UniFed-LoRA is intended for *domain-constrained* cross-silo federated settings in which clients may differ in modality, task, and backbone architecture, yet still operate over semantically related phenomena. Our results should therefore be interpreted as evidence that semantic task correlation is a useful transfer signal in aligned multi-heterogeneous federations, not as a claim for arbitrary heterogeneous FL without domain overlap. Consistent with this, Table 2 shows that replacing one aligned medical client with an unrelated task degrades the remaining in-domain clients.

Comparison scope of baselines. Our evaluation includes both directly comparable mixed-architecture baselines and contextual homogeneous reference runs. In particular, *Local LoRA*, *Centralized*, and *SmolVLM-256* are the most direct comparisons for our mixed-architecture setting, while *FedIT* and *FLoRA* are included as homogeneous reference points since they assume a shared model family and a shared downstream task space. We therefore interpret the latter as useful context for communication-efficient federated fine-tuning, but not as fully matched baselines for the multi-heterogeneous setting considered here.

Privacy profile. UniFed-LoRA avoids transmitting raw client data, logits, prototypes, or full model weights, and communicates only client descriptors together with adapter-space updates. This reduces the communicated object relative to prototype- or logit-sharing methods, but it does not constitute a formal privacy guarantee. In particular, we do not analyze reconstruction, inversion, or membership-inference attacks in the adapter space, and such questions remain open for future work.

Efficiency and architectural assumptions. Our communication analysis focuses on the size of the communicated LoRA parameters rather than on end-to-end wall-clock training time. Moreover, the current design assumes a client-specific output head for each supported backbone family, which requires the server to know the relevant tensor dimensions in advance. This makes the present framework most suitable for fixed cross-silo collaborations with a known set of participating architectures. Extending the approach to zero-shot onboarding of previously unseen backbone families remains future work.