

Fine-tuning Multimodal Transformers on Edge: A Parallel Split Learning Approach

Timo Fudala · Vasileios Tsouvalas · Nirvana Meratnia

Eindhoven University of Technology

FedGenAI @ IJCAI 2025

01 Motivation

The problem

Multimodal transformers exceed typical edge training budgets. Parameter-efficient fine-tuning (PEFT) in federated learning (FL) reduces trainable parameters but still executes the backbone on each client, so fewer updated parameters do not necessarily mean less client computation.

Contributions

- **Multimodal splitting:** extend parallel split learning (PSL) with one shared modality-agnostic encoder.
- **Shared optimization:** aggregate local losses for one server backward pass, keeping raw labels local.
- **Cross-task evidence:** evaluate performance and client cost across seven paired-modality datasets.

Paired inputs, distinct tasks

The study covers vision-text, vision-audio and audio-text pairs. Classification and retrieval impose different fusion and modality-alignment demands.

Where compute is measured

Client compute is measured in billions of floating-point operations (GFLOPs) per input. Offloading shifts computation to the server, so this metric alone does not describe total-system cost.

Scope

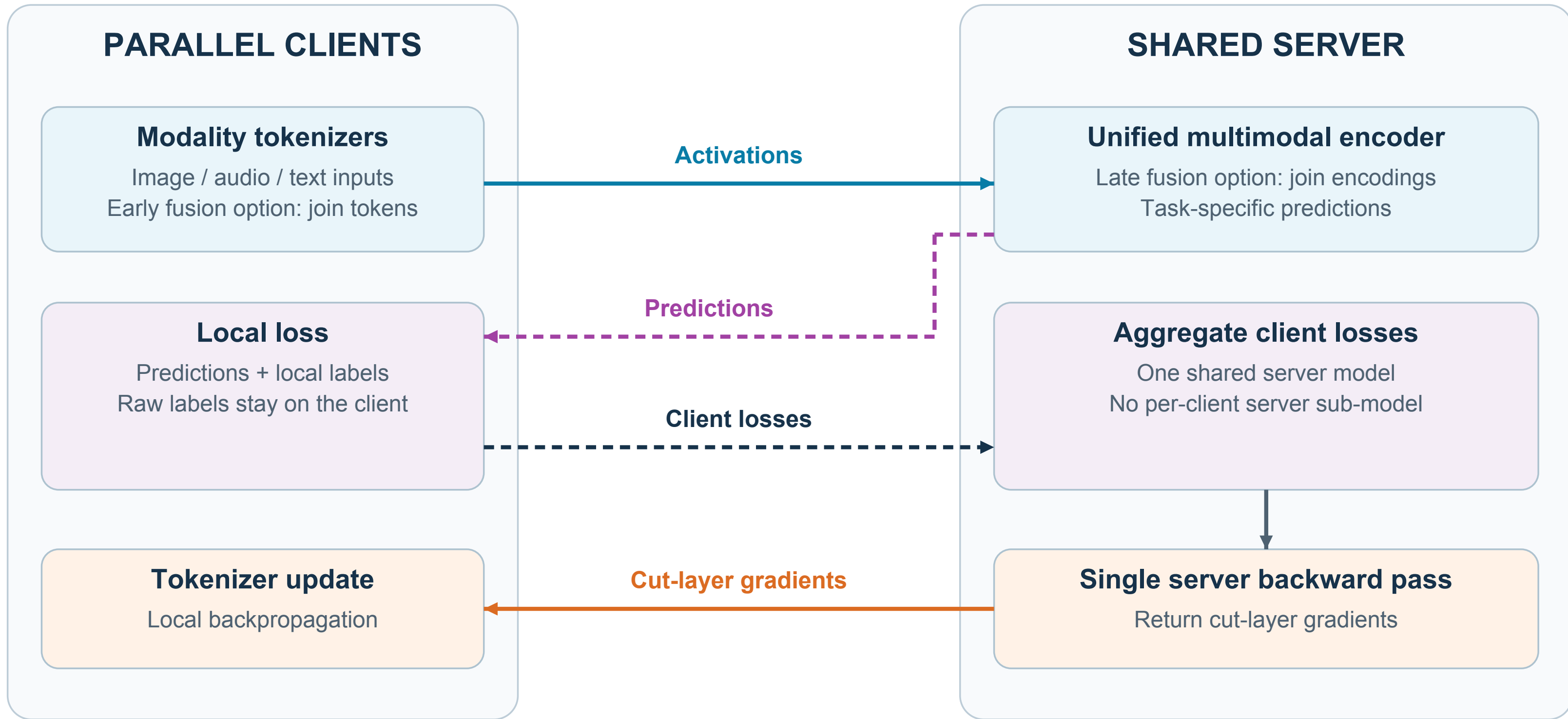
Evaluation averages client tokenizers into one model. Missing-modality robustness is untested; classification gains vary by task, and retrieval remains below centralized fine-tuning.

Notation

multimodal parallel split learning (MPSL) · Dirichlet distribution (Dir)
· vision transformer (ViT) · recall at rank 1 (R@1)

02 MPSL

One shared encoder, lightweight clients



Technical core

$$\mathcal{L}_S = \sum_{n \in \mathcal{N}} \frac{|\mathcal{B}_n|}{|\mathcal{B}|} \mathcal{L}_{C_n}$$

$\mathcal{L}_{C_n}$: local client loss; $\mathcal{B}_n$: client batch; $\mathcal{B}$: combined batch. Minimize this batch-weighted loss in one server backward pass.



Paper



Code

03 Evaluation

0.1 GFLOPs

average client compute

Per input, five classification tasks; FedAvg: 25.8.

58.2 ± 0.9%

MELD accuracy

100 clients; FedAvg: 48.9 ± 0.0%.

Client efficiency with task-dependent accuracy

Method	COCO-QA	UCF101	MELD	Client GFLOPs
Centralized	61.7 ± 0.1	91.3 ± 0.2	61.4 ± 0.4	25.8
FedAvg	53.5 ± 0.2	89.8 ± 0.9	48.9 ± 0.0	25.8
FedCLIP	13.1 ± 2.6	56.4 ± 1.2	50.7 ± 0.7	25.8
MPSL (Ours)	46.2 ± 0.5	89.3 ± 1.0	58.2 ± 0.9	0.1

Accuracy (%), reported ±, 3 seeds. Distributed methods: 100 clients, Dir(0.1); Meta-Transformer ViT-B/16, last 6 blocks trained. Client GFLOPs/input average five classification tasks.

Image-to-text retrieval remains challenging

Method	MS-COCO R@1	Flickr30K R@1	Client GFLOPs
Centralized	56.8 ± 2.4	66.8 ± 5.3	8.65
FedAvg	1.5 ± 0.1	11.2 ± 1.7	8.65
FedCLIP	0.1 ± 0.0	0.1 ± 0.0	8.65
MPSL (Ours)	21.1 ± 0.8	26.0 ± 0.5	0.1

Image-to-text R@1 (%), reported ±, 3 seeds. Distributed methods: 100 clients; full ViT-B/16 encoder fine-tuning. GFLOPs are per client input.

Takeaway

MPSL lowers average client compute to **0.1 GFLOPs per input**, with a MELD gain and COCO-QA deficit versus FedAvg. Retrieval improves over the tested federated baselines but remains below centralized training.