TACTFL: Temporal Contrastive Training for Multi-modal Federated Learning with Similarity-guided Model Aggregation

Guanxiong Sun[1]
guanxiong.sun@bristol.ac.uk
Majid Mirmehdi[2]
m.mirmehdi@bristol.ac.uk
Zahraa Abdallah[1]
zahraa.abdallah@bristol.ac.uk
Raul Santos-Rodriguez[1]
enrsr@bristol.ac.uk
Ian Craddock[1,3]
ian.craddock@bristol.ac.uk
Telmo de Menezes e Silva Filho[1]
telmo.silvafilho@bristol.ac.uk

[1] School of Engineering Mathematics and Technology
University of Bristol
Bristol, UK
[2] School of Computer Science
University of Bristol
Bristol, UK
[3] School of Civil, Aerospace and Design Engineering
University of Bristol
Bristol, UK

Abstract

Real-world federated learning faces two key challenges: limited access to labelled data and the presence of heterogeneous multi-modal inputs. This paper proposes TACTFL, a unified framework for semi-supervised multi-modal federated learning. TACTFL introduces a modality-agnostic temporal contrastive training scheme that conducts representation learning from unlabelled client data by leveraging temporal alignment across modalities. However, as clients perform self-supervised training on heterogeneous data, local models may diverge semantically. To mitigate this, TACTFL incorporates a similarity-guided model aggregation strategy that dynamically weights client models based on their representational consistency, promoting global alignment. Extensive experiments across diverse benchmarks and modalities, including video, audio, and wearable sensors, demonstrate that TACTFL achieves state-of-the-art performance. For instance, on the UCF101 dataset with only 10% labelled data, TACTFL attains 68.48% top-1 accuracy, significantly outperforming the FedOpt baseline of 35.35%. Code will be released upon publication.

1 Introduction

Federated learning (FL) [35] is a decentralised machine learning paradigm for training models on client devices collaboratively, without transferring clients' raw data to a central server. This privacy-preserving setting has become increasingly important in data-sensitive applications such as healthcare [38, 47, 51], internet of things [14, 37], and finance [4, 34]. Traditional FL methods typically rely on two restrictive assumptions: (i) well-annotated client data,

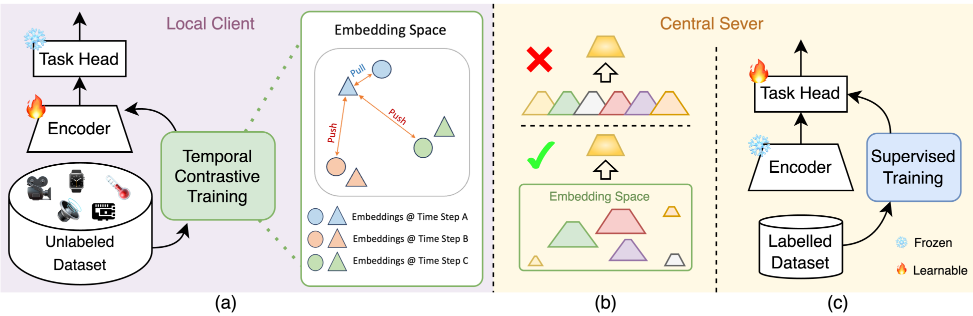

Figure 1: A conceptual overview of TACTFL. (a) On each client, encoders are updated using temporal contrastive training on unlabelled multi-modal data, while task heads remain frozen. The objective encourages embeddings from different modalities at the same time step to be pulled together in the feature space, and embeddings from different time steps to be pushed apart. (b) Unlike FedAvg [35], TACTFL adaptively weights each client model based on representational similarity. (c) The server trains the task head on a small labelled dataset with frozen encoders.

and (ii) single-modality input data. These assumptions often fail in practice due to annotation costs, privacy constraints, and diverse multi-modal data generation by modern devices (e.g., smartphones, wearables, smart cameras). Thus, recent studies have explored semi-supervised [12, 46], unsupervised [2, 19], as well as multi-modal FL (MMFL) [13, 55]. In this paper, we investigate a practically important yet challenging scenario: *semi-supervised MMFL*, where clients hold unlabelled multi-modal data, and a limited labelled dataset is available only on the server. Prior work has explored this scenario [26, 53, 54, 55], but significant challenges remain unresolved, specifically, learning effective multi-modal representations from unlabelled data and handling semantic misalignment in model aggregation.

To address these challenges, we propose TACTFL: TemporAl Contrastive Training for multi-modal Federated Learning, a unified framework that consists of two key components. The *first component* is a novel local temporal contrastive training strategy. Many semi-supervised MMFL methods [8, 25, 31, 54, 55] rely on reconstruction-based training, which is difficult to optimise and can lack semantic relevance [43, 48]. Therefore, recent methods explore contrastive learning [9, 10, 15] and achieve promising results. However, CreamFL [53] works on image-text pairs and requires expensive cross-client intermediate representations exchanges. Similarly, MFCPL [26] requires class labels to construct class-specific prototypes and introduces extra communication costs when exchanging class prototypes across clients. FUELS [32] is tailored for spatio-temporal forecasting tasks in the domain of traffic prediction, and thus, the input data should be well-structured, unimodal data. FTCCL [56] is designed for unimodal time series data, e.g., accelerometer signal, and it particularly focuses on the task of fault diagnosis. In contrast, TACTFL supports multi-modal data (e.g., video-audio streams or wearable sensor signals) and introduces a modal-agnostic supervision signal via temporal intersection-over-union (tIoU). Specifically, positive pairs are softly weighted based on their temporal overlap, providing richer, more flexible guidance for representation alignment across modalities. This departs from existing contrastive learning based MMFL approaches that rely on specific modalities [32, 56] and hard-labelled positives [26, 53], enabling more flexible, graded alignment across modalities. To our knowledge, this is the first use of tIoU as a soft contrastive signal for MMFL, making our framework modality-agnostic. The conceptual overview of local temporal contrastive training is shown in Figure 1(a). After local training, local models are sent to the central server without introducing any extra communication costs (intermediate representations or prototypes).

Our *second component* is then a similarity-guided model aggregation process to mitigate semantic misalignment between client models, an issue largely unaddressed in existing MMFL literature, which typically employs standard aggregation methods (e.g., FedAvg [35], FedOpt [42]). In TACTFL, the server adaptively weights each client's model contribution based on representational similarity. Specifically, we compute pairwise cosine similarities between client model representations obtained using the small labelled dataset on the server. Clients that produce representations more consistent with the majority (i.e., higher inter-client similarity) are considered more semantically aligned and are given greater weight during aggregation. The comparison between the standard aggregation technique and the similarity-guided model aggregation is shown as Figure 1(b). Finally, the central sever updates the task head by supervised training on the small labelled dataset, as shown in Figure 1(c).

The main contributions of TACTFL are as follows. First, we propose a novel temporal contrastive training scheme, which is modality-agnostic, explicitly leverages temporal alignment across modalities, and avoids additional intermediate representation exchanges. Second, we introduce similarity-guided model aggregation, a dynamic server-side strategy that adaptively weights client models based on representation similarity, effectively addressing semantic misalignment in heterogeneous data scenarios. Finally, we conduct extensive experiments across various multi-modal benchmarks spanning video, audio, and wearable sensor data, demonstrating that TACTFL consistently achieves state-of-the-art performance.

2 Related Work

Our work falls under the category of *semi-supervised, multi-modal federated learning*. In this section, we review relevant literature on FL with a particular focus on semi-supervised learning, MMFL, and contrastive learning approaches. Traditional FL studies mainly explore the challenge of data heterogeneity [11, 18, 41, 52], and aggregation strategies [28, 35, 42]. FedAvg [35] is the earliest FL algorithm. It aggregates local models by computing a weighted average, where each model's contribution is weighted by the size of its local dataset. FedOpt [42] extends FedAvg by incorporating adaptive optimisation techniques, i.e., momentum, during aggregation. In many real-world settings, local data usually lacks labels, so recent efforts have explored *semi-supervised* [12, 29, 33, 46] and *unsupervised* [2, 50] FL paradigms, where only a subset of data is labelled or none at all. Survey papers [21, 30, 44, 49] offer comprehensive overviews and trends in FL research.

MMFL extends FL to settings where data from multiple modalities – such as video, audio, and wearable sensors – are distributed across clients. FedMEKT [25], FedMAC [39], and earlier MMFL approaches [8, 54, 55] adopt deep autoencoder-based objectives to align or reconstruct cross-modal representations. However, reconstruction-based objectives are difficult to optimise [43, 48] and often misaligned with downstream tasks. Similarly, FedRecon [31] focuses on reconstructing missing modalities but assumes strong assumptions on modality mappings. MFCPL [26] addresses missing modalities and data heterogeneity via a cross-prototype learning framework, but operates under a supervised learning paradigm, requiring class labels to construct class-specific prototypes.

Contrastive learning has become popular for representation learning from unlabelled data [9, 15]. Recent MMFL methods adopt contrastive objectives to overcome the limitations of autoencoders. CreamFL [53] aligns image-text pairs through contrastive learning but requires sharing intermediate representations between clients and server, increasing communication overhead. MOON [27] and FedMAC [39] use model-contrastive objectives to align local and global representations, but they are designed for unimodal settings. More recent works, such as FUELS [32], target personalised federated spatio-temporal forecasting via dual semantic alignment but require both modality-specific encoders and global prototypes. Similarly, FTCCL [56] introduces temporal-context contrastive learning for fault diagnosis but focuses on supervised or pseudo-label training and does not consider modality alignment or missing modalities.

In contrast to prior works, by conducting temporal contrastive training, TACTFL combines the strengths of contrastive learning with a semi-supervised, modality-agnostic design. TACTFL avoids any intermediate representation or prototype exchange, which preserves privacy and minimises communication overhead. Moreover, TACTFL accommodates missing modalities during training and inference and can remain robust.

3 Method

TACTFL follows an iterative training workflow. First, the server initialises the global encoders and task head and distributes them to all clients. Each client then performs local encoder training using unlabelled multi-modal data with temporal contrastive training, keeping the task head frozen, as shown in Figure 1(a). After local training, clients upload their updated encoder weights to the server, which aggregates them using the similarity-guided model aggregation strategy, as shown in Figure 1(b). The server then trains the task head using the small labelled dataset while keeping the encoders fixed, as shown in Figure 1(c). Finally, the updated model (encoders and task head) is redistributed to all clients. This process is repeated multiple rounds. A detailed algorithm can be found in the Supplementary Material.

3.1 Problem Definition

We consider a federated learning scenario with C clients, each holding locally collected multi-modal data from heterogeneous sources such as smart cameras, wearable sensors, and mobile devices. Each client is assumed to possess at least two synchronised modalities, denoted as x^A and x^B. These may include RGB video frames (modality A), audio signals (modality B), or time-series data such as accelerometer and gyroscope measurements. In practice, some clients may experience missing modalities.

Let $\mathcal{D}_c = \{(x_i^A, x_i^B)\}_{i=1}^{N_c}$ represent the local unlabelled dataset on client $c \in \{1, \ldots, C\}$. All client datasets $\mathcal{D}_1, \ldots, \mathcal{D}_C$ are fully unlabelled, reflecting real-world challenges in data annotation at the edge. In contrast, a small labelled proxy dataset is available on the central server: $\mathcal{D}_S = \{(x_j^A, x_j^B, y_j)\}_{j=1}^{N_s}$, where y_j denotes task-specific labels. This server-side dataset shares the same modality structure as the client data and serves two key purposes: (i) training the downstream task head, and (ii) supporting similarity-based model aggregation to address inter-client semantic misalignment. The objective is to collaboratively learn a global model that can perform a downstream task (e.g., action recognition or emotion classification) by exploiting unlabelled multi-modal data from clients and the small labelled dataset on the

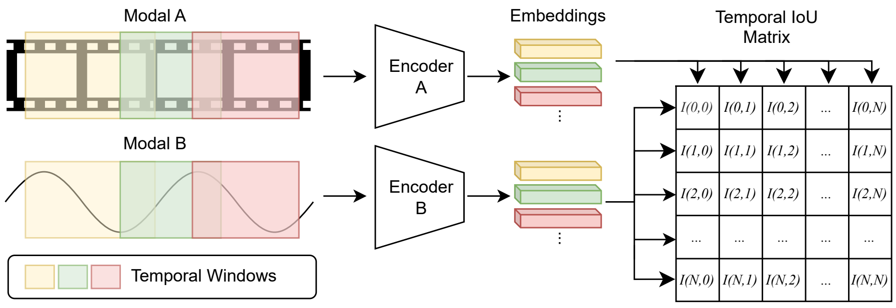

Figure 2: Details of the temporal contrastive training process.

server, without transferring raw data. This setup ensures strong privacy preservation while enabling effective and communication-efficient training.

Generalised Model Architecture. TACTFL adopts a modular and lightweight architecture composed of three parts: modality-specific encoders, a fusion module, and a downstream task head. For each input modality, an encoder E^m (e.g., E^A for RGB, E^B for audio), as shown in Figure 2, is used to extract high-level representations from raw data. The encoders can be implemented using modality and application-appropriate architectures, such as CNNs for visual data or LSTMs for wearable signals, and are trained locally on unlabelled data using our temporal contrastive training strategy, while the task head remains fixed during this stage.

The outputs of the encoders are aggregated by a lightweight fusion module that combines per-modality embeddings into a unified representation. This fusion can be implemented via simple concatenation or averaging, though more advanced approaches (e.g., attention-based fusion) are also compatible. The fused embedding is then passed to a prediction head H, which is responsible for producing task-related outputs. Unlike the encoders, the task head is trained exclusively on the central server using a small labelled proxy dataset. During this supervised training phase, encoder weights are frozen to preserve the representations learned during local contrastive training.

3.2 Temporal Contrastive Training

To enable local learning on unlabelled multi-modal data, we introduce a self-supervised contrastive objective that leverages temporal alignment as implicit supervision. The idea is that signals captured from different modalities within overlapping time windows are likely to describe the same underlying activity and should be embedded closely in the embedding space. Figure 1(a) shows some examples in the embedding space.

The temporal contrastive training process is shown as Figure 2. For each multi-modal sample, we segment the input sequence into two temporal chunks. The window size is defined as a proportion of the full input (e.g., 80% or 90%), and segments may overlap to ensure smooth transitions. Each segment preserves temporal coherence and can be treated as an individual training sample. For a client with two modalities (e.g., RGB video and audio), we denote the resulting sets of temporal segments as $X^A = \{x_k^A\}_{k=0}^{2N-1}$ and $X^B = \{x_k^B\}_{k=0}^{2N-1}$, where $2N$ is the total number of segments across the batch. For each pair (x_i^A, x_j^B), we compute their tIoU, a widely used metric in action localisation. The tIoU serves as a soft pseudo-label for the likelihood that the two segments represent the same content. We construct a matrix of tIoU scores $I(i,j)$ between all segment pairs from the two modalities, and normalise each row to produce a probability distribution. $T(i,\cdot)$ represents a soft target distribution for x_i^A across all x_j^B, and $T(i,j)$ is computed as $I(i,j)/\sum_{k=0}^{2N} I(i,k)$. The contrastive loss encourages the embeddings of segments with high temporal overlap to be closer in the feature space, pushing apart unrelated ones. The loss for a pair (i,j) is defined as

$$l(i,j) = -\log \frac{e^{\mathcal{S}(i,j)/\tau}}{\sum_{k=0}^{2N} e^{\mathcal{S}(i,k)/\tau}} T(i,j), \tag{1}$$

where τ is a temperature parameter and $\mathcal{S}(i,j)$ denotes the cosine similarity between encoder outputs as

$$\mathcal{S}(i,j) = \frac{E^A(x_i^A)^\top E^B(x_j^B)}{\|E^A(x_i^A)\| \cdot \|E^B(x_j^B)\|}. \tag{2}$$

The total contrastive loss is computed symmetrically over both modalities, aggregating all pairs (i,j) and (j,i) within each mini-batch. This formulation enables robust cross-modal representation learning without requiring labels, and naturally supports clients with missing modalities: when only a single modality is available, the same modality can be duplicated to form pseudo-pairs for training. Additionally, the temporal segmentation process implicitly augments data and improves generalisation across clients with limited local data.

3.3 Similarity-guided Model Aggregation

In contrast to canonical supervised FL, where local models are supervised by the same label distribution and tend to produce semantically aligned feature spaces, the use of self-supervised contrastive learning in TACTFL introduces a new challenge: semantic misalignment. Since each client independently optimises its encoders on local data using contrastive objectives, the resulting representations can diverge significantly across clients. For example, the same input sample may produce embeddings with a small cosine similarity when processed by encoders trained on different clients. This inconsistency makes standard aggregation techniques, e.g., FedAvg [35], unreliable.

The similarity-guided model aggregation strategy leverages the small labelled dataset on the central server D_S to assess the alignment of each client model. It assigns higher weights to those clients whose representations are more semantically consistent with the rest. Formally, given C client encoder models $\{m_0, m_1, \ldots, m_{C-1}\}$, the server first computes feature embeddings for the labelled dataset D_S by passing it through each client model. For each model m_i, the embeddings are averaged across all samples in D_S to produce a summary vector V_i. This results in C embedding vectors $\{V_0, V_1, \ldots, V_{C-1}\}$, one per client. The server then computes a pairwise similarity matrix $S \in \mathbb{R}^{C\times C}$, where each element $S(i,j)$ reflects the cosine similarity between the representations V_i and V_j, computed as $S(i,j) = V_i \cdot V_j / \|V_i\| \cdot \|V_j\|$. To determine how well-aligned each client is with the others, we compute a similarity-based weight vector $W \in \mathbb{R}^C$,

$$W(i) = \frac{\sum_{j=0}^{C-1} S(i,j)}{\sum_{i=0}^{k-1} \sum_{j=0}^{k-1} S(k,j)}, \tag{3}$$

ensuring that $\sum_{i=0}^{C-1} W(i) = 1$. These weights capture the relative alignment of each client's representations with the rest of the population. Finally, the global encoder model m' is

Dataset	α	Modalities	Metric	$r_l = 0.9$			$r_l = 0.8$		
				FedAvg [illegible]	FedOpt [illegible]	TACTFL	FedAvg [illegible]	FedOpt [illegible]	TACTFL
UCF101 [illegible]	0.1	Audio, Video	Acc	10.49	<u>35.35</u>	**68.48**	40.14	<u>44.43</u>	**71.67**
MiT10 [illegible]		Audio, Video	Acc	7.48	<u>19.86</u>	**48.87**	25.77	<u>36.90</u>	**51.77**
UCI-HAR [illegible]		Acce, Gyro	F1	33.78	<u>41.26</u>	**74.87**	35.11	<u>42.64</u>	**75.12**
UCF101 [illegible]	5.0	Audio, Video	Acc	6.41	<u>12.56</u>	**70.22**	40.71	<u>46.53</u>	**72.99**
MiT10 [illegible]		Audio, Video	Acc	9.21	<u>21.04</u>	**50.13**	27.60	<u>38.15</u>	**52.66**
UCI-HAR [illegible]		Acce, Gyro	F1	35.66	<u>45.21</u>	**79.49**	37.70	<u>44.50</u>	**76.41**

Table 1: Scenario 1: Accuracy/F1 (%) of TACTFL and two baseline methods, FedAvg/FedOpt, under two missing label rates r_l. α denotes the control factor of the Dirichlet distribution for data partitioning (best results in **bold**, second best <u>underlined</u>).

Datasets	Modality		FedIoT [illegible]	FedProx [illegible]	MOON [illegible]	CreamFL [illegible]	FedMEKT [illegible]	TACTFL
mHealth [illegible]	Acce-Gyro	Acce	63.83	67.86	64.47	<u>68.58</u>	68.28	**68.90**
		Gyro	63.62	64.36	64.12	64.34	<u>64.60</u>	**64.75**
	Acce-Mage	Acce	69.99	70.00	69.71	69.51	**71.16**	<u>71.09</u>
		Mage	68.49	69.21	68.79	69.18	<u>71.13</u>	**71.84**
	Gyro-Mage	Gyro	65.43	65.37	65.90	66.05	**67.10**	<u>67.21</u>
		Mage	68.28	68.75	68.13	68.00	<u>69.02</u>	**69.30**
UR-Fall [illegible]	Acce-RGB	Acce	61.70	65.66	65.89	66.82	**70.66**	<u>70.43</u>
		RGB	57.88	59.26	62.00	62.34	<u>66.70</u>	**66.92**
	Acce-Depth	Acce	67.24	68.79	68.08	71.61	**72.68**	<u>72.17</u>
		Depth	60.76	68.08	68.16	**75.33**	<u>75.22</u>	74.99
	RGB-Depth	RGB	69.88	75.60	73.38	**78.57**	77.87	<u>78.18</u>
		Depth	67.61	70.04	66.48	69.18	<u>70.57</u>	**70.86**
Opportunity [illegible]	Acce-Gyro	Acce	71.75	72.25	72.96	73.33	**73.50**	<u>73.37</u>
		Gyro	72.08	71.43	72.09	72.10	<u>72.15</u>	**73.12**

Table 2: Scenario 1: F1 scores on FedMEKT [25] benchmarks. Paired modalities are for training, and each individual modality is for testing (best in **bold**, second best <u>underlined</u>).

obtained as a weighted sum of the client models as $m' = \sum_{i=0}^{C-1} W(i) \cdot m_i$. By adaptively emphasising clients with better-aligned representations, the training convergence speed is accelerated. It also mitigates the impact of noisy, biased, poorly trained local models.

After the model aggregation, the server conducts supervised central training in a standard way. The task head H is trained using the small labelled dataset D_S. During this stage, the encoders are frozen, and only H is optimised with supervised loss. The updated model containing the trained head is then sent to clients for the next round. More details of the central training can be found in the Supplementary Material.

4 Experiments

We evaluate TACTFL against state-of-the-art methods on diverse benchmark datasets covering various modalities and tasks. Following FedMEKT [25], we conduct experiments under two scenarios reflecting real-world challenges: *multi-modal* and *missing-modal* clients.

Detailed Settings. Multi-modal encoders extract embeddings from raw input. Following FedMM [13], MobileNetV2 [17], mel-frequency cepstral coefficients (MFCCs) [7], and MobileBERT [45] are used as encoders for videos/images, audios, and text, respectively. Following FedMEKT [25], 2 LSTM layers [16] are used as the encoder for wearable data. For our experiments, we follow FedMM [13] to conduct local training for 1 epoch in each of 200 rounds of global-local updates and fix the batch size for all datasets to 16. The number of clients for each dataset can be found in the Supplementary Material.

Scenario 1: Multi-modal Clients. Each client holds synchronised data from multiple

Datasets	r_m	FedOpt [32]	MOON [27]	FedProx [28]	FPL [20]	FedIoT [55]	FedMSplit [6]	MFCPL [26]	TACTFL
UCI-HAR [3]	0.5	74.79	74.87	70.14	75.16	71.57	74.76	<u>77.44</u>	**77.92**
	0.7	66.17	73.63	69.37	74.28	70.58	73.32	<u>75.61</u>	**75.83**
	0.8	56.45	72.78	68.77	73.46	68.62	72.77	**75.19**	<u>75.12</u>
Hateful Memes [23]	0.5	54.62	52.67	52.75	54.97	53.98	55.57	<u>56.30</u>	**56.89**
	0.7	51.61	51.40	50.49	53.79	52.72	53.61	**55.69**	<u>55.31</u>
	0.8	49.58	50.68	50.23	52.94	51.31	52.65	<u>55.04</u>	**59.47**
MELD [40]	0.5	51.58	50.66	51.92	52.11	51.81	52.97	<u>54.71</u>	**55.26**
	0.7	49.27	47.51	51.83	50.09	49.48	51.65	<u>53.62</u>	**54.10**
	0.8	47.54	46.94	50.48	49.49	49.61	50.57	<u>52.18</u>	**53.25**

Table 3: Scenario 2: Performance for various missing-modality rates (r_m) on UCI-HAR, Hateful Memes, MELD datasets (best in **bold**, second best <u>underlined</u>).

modalities, such as video-audio streams or wearable sensor signals. This setting reflects real-world conditions where multi-modal data is plentiful but labels are unavailable. We first follow FedMM [13] protocols to prepare datasets including UCF101 [22] and MiT-10 [36] for video-audio, and UCI-HAR [3] for accelerometer (Acce) and gyroscope (Gyro) data. To simulate federated non-IID distributions, data is partitioned using a Dirichlet distribution $\mathrm{Dir}(\alpha)$, where lower α values introduce higher heterogeneity. The label missing rate r_l controls the level of label availability. For example, when r_l=0.9, 90% data is treated unlabelled for local training, and 10% data is randomly sampled as the central labelled set D_S. Table 1 compares TACTFL against FedAvg and FedOpt. TACTFL consistently matches or exceeds baselines.

We further evaluate TACTFL under the FedMEKT [25] setup using three human activity datasets. The mHealth dataset [1] includes accelerometer, gyroscope, and magnetometer (Mage) signals. One participant is held out for testing, one for the labelled set D_S, and the rest for training. For UR-Fall [24], which includes RGB, depth, and accelerometer data, we randomly sample 10% for testing, 10% as D_S, and leave the remaining for training. The Opportunity dataset [5] includes Acce and Gyro signals from four participants performing kitchen tasks. Runs 4 and 5 from participants 2 and 3 are used for test, runs 4 and 5 from participants 1 and 4 are used as D_S. As shown in Table 2, TACTFL consistently achieves the best or second-best results across modality pairs, outperforming strong baselines like FedMEKT and CreamFL, especially in sensor-rich datasets like mHealth and UR-Fall where temporal and multi-modal cues are critical.

Scenario 2: Missing Modalities. In practice, some clients may lack access to certain modalities due to hardware limitations or sensor failures. This setting captures the challenge of learning from incomplete multi-modal inputs, with all client data remaining unlabelled. We evaluate this scenario on three datasets: (1) UCI-HAR [3], featuring Acce and Gyro data; (2) Hateful Memes [23], a binary classification dataset of 10k image-text pairs; and (3) MELD [40], a multimodal emotion recognition dataset with audio and text. We use four emotion classes (neutral, happy, sad, angry) with the most samples. Following FedMM [13] and MFCPL [26], we simulate missing modalities by randomly dropping entire modalities per client. For each client–modality pair, a Bernoulli trial determines whether that modality is retained or replaced with zeros. We experiment with missing modality rates $r_m \in \{0.5, 0.7, 0.8\}$. We report AUC/UAR/F1 metrics for the Hateful Memes/MELD/UCI-HAR datasets, respectively. As shown in Table 3, TACTFL achieves the best performance in the majority of cases, and ranks either first or second in every setting. Notably, it outperforms the best state-of-the-art method MFCPL [26], particularly on the MELD dataset. These results highlight TACTFL's robustness and strong generalisation ability in scenarios with missing modalities.

Ablation Study. We perform comprehensive ablation studies on the UCF101 dataset to

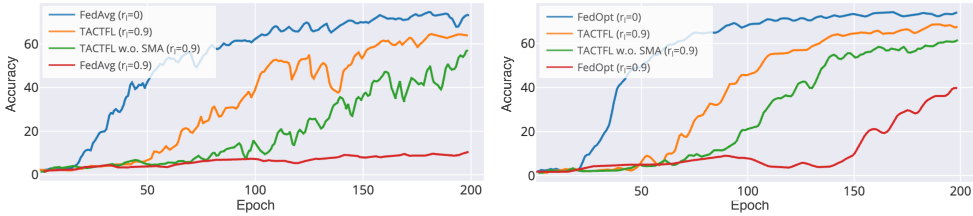

Figure 3: Test accuracy curve under four settings: 1. SSFL ($r_l = 0.9$); 2. TACTFL without SMA ($r_l = 0.9$); 3. Full TACTFL ($r_l = 0.9$); 4. Fully-supervised FL ($r_l = 0$).

assess the contribution of each component, including: (1) temporal contrastive training (TCT), (2) similarity-guided model aggregation (SMA), (3) robustness to varying label availability, (4) resilience to missing modalities, and (5) the impact of temporal window size.

Effects of TCT and SMA. We evaluate four configurations under both FedAvg and FedOpt settings. First, *SSFL* denotes semi-supervised FL with a missing label rate of $r_l = 0.9$, i.e., only 10% of the data is labelled and used as D_S on the server, while the remaining 90% of client data is discarded. Second, *TACTFL w/o SMA* includes TCT on unlabelled data but uses standard aggregation (FedAvg or FedOpt). Lastly, the full *TACTFL* method integrates both TCT and SMA. The test accuracy curves of all configurations are shown in Figure 3. More specifically, compared to SSFL baselines of FedAvg/FedOpt ($r_l = 0.9$), introducing TCT alone (TACTFL w/o SMA) improves accuracy from 10.49% to 54.58% and from 40.43% to 62.91%, respectively, demonstrating the effectiveness of TCT from unlabelled data. Moreover, incorporating SMA (TACTFL) yields additional performance gains, improving accuracy to 63.12% (FedAvg) and 69.60% (FedOpt). This confirms that SMA plays a critical role in mitigating semantic misalignment during aggregation, resulting in a more robust and semantically consistent global model. Furthermore, the test accuracy curve shown in Figure 3 demonstrates that SMA can speed up TACTFL's convergence speed, bringing it close to fully-supervised ($r_l = 0$) performance.

Robustness to label and modality sparsity. To assess TACTFL's robustness under practical constraints, we evaluate its performance under varying degrees of label scarcity and missing modality scenarios. In Table 4, we gradually increase the missing label and missing modality rates, r_l and r_m, from 0.1 to 0.9. TACTFL consistently outperforms FedAvg across all values of r_l, with the performance gap widening as fewer labels are available. Notably, at $r_l = 0.9$, TACTFL achieves 63.12% F1 score compared to just 10.49% for FedAvg. As for missing modalities, even with $r_m = 0.9$, TACTFL retains strong performance (72.22%) and consistently outperforms FedOpt. These results highlight TACTFL's robustness to both limited supervision and incomplete sensor inputs, making it suitable for deployment in real-world, resource-constrained FL scenarios.

Setting	**Method**	**Rate**				
		0.1	**0.3**	**0.5**	**0.7**	**0.9**
r_l	FedAvg [35]	71.02	70.81	67.28	60.25	10.49
	TACTFL	71.31	71.15	69.73	67.08	63.12
r_m	FedOpt [42]	74.48	73.20	71.46	68.56	56.32
	TACTFL	75.01	74.76	73.52	72.84	72.22

Table 4: Accuracy under varying levels of missing label and missing modality rates, r_l and r_m.

Effect of temporal window size. As shown in Table 5, moderate window sizes (e.g., 70–80%) yield the highest accuracy, while both shorter and longer windows slightly reduce performance. This suggests that appropriately-sized temporal context is crucial for effective temporal contrastive learning across modalities.

Temporal Window Size	90%	80%	70%	60%	50%
Accuracy	69.26	71.34	72.22	71.28	70.86

Table 5: Effect of temporal window size on performance (accuracy %).

5 Conclusion

In this work, we presented TACTFL, a unified framework for semi-supervised multi-modal federated learning. By introducing temporal contrastive training and similarity-guided model aggregation, TACTFL effectively leverages unlabelled multi-modal data while addressing semantic misalignment across clients. Extensive experiments demonstrate that our method achieves consistent and significant improvements over existing baselines across diverse benchmarks, highlighting its potential for real-world privacy-preserving applications. Despite its promising performance, TACTFL has several limitations. First, the current framework assumes access to synchronised multi-modal data, which may not always be feasible in real-world settings. Second, the temporal contrastive training currently relies on fixed window parameters, which might not optimally capture semantic dynamics for all modalities. Future work will explore dynamic temporal alignment techniques, support for asynchronous modalities, and personalised model adaptation for highly heterogeneous client populations.

6 Ethical Considerations and Broader Impact

The benchmark datasets used in our evaluation (e.g., UCF101, Hateful Memes, MELD) may contain inherent demographic and cultural biases, which can lead to models that perform unfairly across different user groups. Another potential risk of bias may be introduced by the Similarity-guided Model Aggregation (SMA). SMA could penalise client models trained on data from underrepresented populations, skewing the final global model. While TACTFL's federated design is inherently privacy-preserving, developers must be vigilant about these fairness issues. We strongly advocate for the careful auditing of proxy datasets to ensure they are balanced and representative. Future work should explore fairness-aware and personalised aggregation techniques to mitigate these potential biases before any real-world deployment.

7 Acknowledgements

This work was supported by the TORUS Project, EPSRC grant number EP/X036146/1. RSR is funded by the UKRI Turing AI Fellowship EP/V024817/1.

References

[1] Oresti Banos, Rafael Garcia, Juan A Holgado-Terriza, Miguel Damas, Hector Pomares, Ignacio Rojas, Alejandro Saez, and Claudia Villalonga. mhealthdroid: A novel frame-

work for agile development of mobile health applications. *Ambient Assisted Living and Daily Activities*, 8868(14):91–98, 2014.

[2] Guangyin Bao, Qi Zhang, Duoqian Miao, Zixuan Gong, Liang Hu, Ke Liu, Yang Liu, and Chongyang Shi. Multimodal federated learning with missing modality via prototype mask and contrast. *arXiv preprint arXiv:2312.13508*, 2023.

[3] Erhan Bulbul, Aydin Cetin, and Ibrahim Alper Dogru. Human activity recognition using smartphones. In *2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 1–6, 2018. doi: 10.1109/ISMSIT.2018.8567275.

[4] David Byrd and Antigoni Polychroniadou. Differentially private secure multi-party computation for federated learning in financial applications. In *Proceedings of the first ACM international conference on AI in finance*, pages 1–9, 2020.

[5] Ricardo Chavarriaga, Hesam Sagha, Alberto Calatroni, Sundara Tejaswi Digumarti, Gerhard Tröster, José del R Millán, and Daniel Roggen. The opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters*, 34(15):2033–2042, 2013.

[6] 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*, KDD '22, page 87–96, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3539384. URL https://doi.org/10.1145/3534678.3539384.

[7] Li-Wei Chen and Alexander Rudnicky. Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[8] Sijia Chen and Baochun Li. Towards optimal multi-modal federated learning on non-iid data with hierarchical gradient blending. In *IEEE Conference on Computer Communications (INFOCOM)*, 2022.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[10] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[11] Yutong Dai, Zeyuan Chen, Junnan Li, Shelby Heinecke, Lichao Sun, and Ran Xu. Tackling data heterogeneity in federated learning with class prototypes. In *AAAI*, 2023.

[12] Enmao Diao, Jie Ding, and Vahid Tarokh. Semifl: semi-supervised federated learning for unlabeled clients with alternate training. In *NeurIPS*, 2022.

[13] 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 *ACM KDD*, 2023.

[14] Meng Hao, Hongwei Li, Xizhao Luo, Guowen Xu, Haomiao Yang, and Sen Liu. Efficient and privacy-enhanced federated learning for industrial artificial intelligence. *IEEE Transactions on Industrial Informatics*, 16(10):6532–6542, 2019.

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[17] Andrew G Howard. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[18] Wenke Huang, Mang Ye, and Bo Du. Learn from others and be yourself in heterogeneous federated learning. In *CVPR*, 2022.

[19] Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. Rethinking federated learning with domain shift: A prototype view. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16312–16322. IEEE, 2023.

[20] Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. Rethinking federated learning with domain shift: A prototype view. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16312–16322, 2023. doi: 10.1109/CVPR52729.2023.01565.

[21] Yilun Jin, Yang Liu, Kai Chen, and Qiang Yang. Federated learning without full labels: A survey. *arXiv preprint arXiv:2303.14453*, 2023.

[22] Mubarak Shah Khurram Soomro, Amir Roshan Zamir. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[23] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33: 2611–2624, 2020.

[24] Bogdan Kwolek and Michal Kepski. Human fall detection on embedded platform using depth maps and wireless accelerometer. *Computer methods and programs in biomedicine*, 117(3):489–501, 2014.

[25] Huy Q Le, Minh NH Nguyen, Chu Myaet Thwal, Yu Qiao, Chaoning Zhang, and Choong Seon Hong. Fedmekt: Distillation-based embedding knowledge transfer for multimodal federated learning. *Neural Networks*, 183:107017, 2025.

[26] Huy Q Le, Chu Myaet Thwal, Yu Qiao, Ye Lin Tun, Minh NH Nguyen, Eui-Nam Huh, and Choong Seon Hong. Cross-modal prototype based multimodal federated learning under severely missing modality. *Information Fusion*, page 103219, 2025.

[27] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *CVPR*, 2021.

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

[29] Haowen Lin, Jian Lou, Li Xiong, and Cyrus Shahabi. Semifed: Semi-supervised federated learning with consistency and pseudo-labeling. *ArXiv*, 2021.

[30] Yi-Ming Lin, Yuan Gao, Mao-Guo Gong, Si-Jia Zhang, Yuan-Qiao Zhang, and Zhi-Yuan Li. Federated learning on multimodal data: A comprehensive survey. *Machine Intelligence Research*, 20(4):539–553, 2023.

[31] Junming Liu, Guosun Zeng, Ding Wang, Yanting Gao, and Yufei Jin. Fedrecon: Missing modality reconstruction in distributed heterogeneous environments. *arXiv preprint arXiv:2504.09941*, 2025.

[32] Qingxiang Liu, Sheng Sun, Yuxuan Liang, Min Liu, and Jingjing Xue. Personalized federated learning for spatio-temporal forecasting: A dual semantic alignment-based contrastive approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 12192–12200, 2025.

[33] Xinyang Liu, Pengchao Han, Xuan Li, and Bo Liu. Semidfl: A semi-supervised paradigm for decentralized federated learning, 2024.

[34] Guodong Long, Yue Tan, Jing Jiang, and Chengqi Zhang. Federated learning for open banking. In *Federated learning: privacy and incentive*, pages 240–254. Springer, 2020.

[35] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of International Conference on Artificial Intelligence and Statistics*. PMLR, 2017.

[36] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–8, 2019. ISSN 0162-8828. doi: 10.1109/TPAMI.2019.2901464.

[37] Dinh C Nguyen, Ming Ding, Pubudu N Pathirana, Aruna Seneviratne, Jun Li, Dusit Niyato, and H Vincent Poor. Federated learning for industrial internet of things in future industries. *IEEE Wireless Communications*, 28(6):192–199, 2021.

[38] Dinh C Nguyen, Quoc-Viet Pham, Pubudu N Pathirana, Ming Ding, Aruna Seneviratne, Zihuai Lin, Octavia Dobre, and Won-Joo Hwang. Federated learning for smart healthcare: A survey. *ACM Computing Surveys (Csur)*, 55(3):1–37, 2022.

[39] Manh Duong Nguyen, Trung Thanh Nguyen, Huy Hieu Pham, Trong Nghia Hoang, Phi Le Nguyen, and Thanh Trung Huynh. Fedmac: Tackling partial-modality missing in federated learning with cross-modal aggregation and contrastive regularization. In *2024 22nd International Symposium on Network Computing and Applications (NCA)*, pages 278–285. IEEE, 2024.

[40] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.

[41] Liangqiong Qu, Yuyin Zhou, Paul Pu Liang, Yingda Xia, Feifei Wang, Ehsan Adeli, Li Fei-Fei, and Daniel Rubin. Rethinking architecture design for tackling data heterogeneity in federated learning. In *cvpr*, 2022.

[42] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konecny, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

[43] Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pages 5628–5637. PMLR, 2019.

[44] Zixing Song, Xiangli Yang, Yifei Zhang, Xinyu Fu, Zenglin Xu, and Irwin King. A systematic survey on federated semi-supervised learning. In *IJCAI*, 2024.

[45] Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. Mobilebert: a compact task-agnostic bert for resource-limited devices. *arXiv preprint arXiv:2004.02984*, 2020.

[46] Arvin Tashakori, Wenwen Zhang, Z. Wang, and Peyman Servati. Semipfl: Personalized semi-supervised federated learning framework for edge intelligence. *IEEE Internet of Things Journal*, 2023.

[47] Zhen Ling Teo, Liyuan Jin, Nan Liu, Siqi Li, Di Miao, Xiaoman Zhang, Wei Yan Ng, Ting Fang Tan, Deborah Meixuan Lee, Kai Jie Chua, et al. Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture. *Cell Reports Medicine*, 5(2), 2024.

[48] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Leon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

[49] Jie Wen, Zhixia Zhang, Yang Lan, Zhihua Cui, Jianghui Cai, and Wensheng Zhang. A survey on federated learning: challenges and applications. *International Journal of Machine Learning and Cybernetics*, 14(2):513–535, 2023.

[50] Hao Yan and Yuhong Guo. Lightweight unsupervised federated learning with pretrained vision language model, 2024.

[51] Tian Yang, Xinhui Yu, Martin J McKeown, Z Jane Wang, et al. When federated learning meets medical image analysis: A systematic review with challenges and solutions. *APSIPA Transactions on Signal and Information Processing*, 13(1), 2024.

[52] Mang Ye, Xiuwen Fang, Bo Du, Pong C Yuen, and Dacheng Tao. Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Computing Surveys*, 56(3):1–44, 2023.

[53] Qiying Yu, Yang Liu, Yimu Wang, Ke Xu, and Jingjing Liu. Multimodal federated learning via contrastive representation ensemble. *ICLR*, 2023.

[54] Yuchen Zhao, Hanyang Liu, Honglin Li, Payam M. Barnaghi, and Hamed Haddadi. Semi-supervised federated learning for activity recognition. *ArXiv*, abs/2011.00851, 2020. URL `https://api.semanticscholar.org/CorpusID:226226498`.

[55] Yuchen Zhao, Payam Barnaghi, and Hamed Haddadi. Multimodal federated learning on iot data. In *IEEE/ACM International Conference on Internet-of-Things Design and Implementation (IoTDI)*, 2022.

[56] Haowen Zheng, Hui Liu, Zhenyu Liu, and Jianrong Tan. Federated temporal-context contrastive learning for fault diagnosis using multiple datasets with insufficient labels. *Advanced Engineering Informatics*, 60:102432, 2024.