When Heterogeneous Modalities Synchronize: Mitigating Desynchronization Attacks in Multimodal Online Federated Learning

Kainat Ibrar[1(✉)], Asim Ibrahim[2], and Francesco Palmieri[1]

[1] University of Salerno, Via Giovanni Paolo II, 132 - 84084 Fisciano, Salerno, Italy
{kibrar,fpalmieri}@unisa.it
[2] Department of Management Sciences, Bahria University Islamabad, Islamabad, Pakistan

Abstract. Multimodal online federated learning (MOFL) relies on accurate temporal alignment across heterogeneous data streams to enable effective cross-modal fusion. In this paper, we investigate temporal desynchronization as a practical adversarial attack surface in such systems. We show that even small temporal shifts between modalities can disrupt cross-modal alignment, leading to degraded model performance and increased attack effectiveness. Unlike traditional poisoning attacks, the proposed approach operates at the synchronization level, making it subtle and difficult to detect. We conduct experiments under both untargeted and label-conditioned (targeted) attack settings to comprehensively evaluate the impact of temporal misalignment. To mitigate this vulnerability, we propose a lightweight Cross-Client Consistency Scoring Mechanism (CCCSM) that reweights client updates based on their similarity during server-side aggregation. Experimental results on the UP-Fall dataset demonstrate that (i) the impact of the attack increases with the magnitude of temporal misalignment and (ii) the attack is more severe under targeted settings. For perturbation strength $k = 5$, the overall F1-score drops from 95.6% to 80.5%, while the attack success rate (ASR) increases up to 79.2% in targeted settings. The proposed CCCSM-based defense reduces ASR from 79.2% to 45.1% and improves overall F1 from 79.1% to 84.7% under targeted attacks with $k = 5$, while maintaining performance close to the clean baseline. These findings highlight the importance of synchronization in MOFL and the need for synchronization-aware defenses in real-world deployments.

Keywords: Online Federated Learning · Multimodal learning · Adversarial attacks · Cross-modal desynchronization attacks · Modality-Allignment attacks

1 Introduction

Federated learning (FL) has become a popular choice for modern intelligent systems, especially within safety-critical and distributed environments [15] where

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 363–378, 2027.
https://doi.org/10.1007/978-3-032-30521-3_23

data privacy and communication overhead constraints are of utmost importance for correct decision-making, such as in self-driving cars, industrial IoT, healthcare monitoring, and smart infrastructure. These systems use a variety of sensors, such as vision devices, inertial sensors, LiDAR, audio, and environmental sensors, to obtain additional information and improve decision-making [17] and are categorized as unimodal or multimodal FL systems, depending on the nature of the data possessed by the clients [7].

While traditional FL assumes static training rounds and synchronized updates, real-world deployments operate more and more frequently in Online Federated Learning (OFL) environments, in which clients continuously collect streaming data, train locally in sliding windows, and contribute model updates to a global server [20]. Moreover, in practical settings, the majority of the clients contain multimodal data, and this OFL's configuration introduces temporal drift, partial participation, and heterogeneous sampling rates across modalities, resulting in Multimodal Online Federated Learning (MOFL) systems heavily relying on accurate cross-modal synchronization to align heterogeneous sensor streams before fusion [20,21].

In the recent state-of-the-art (SOTA), synchronization mechanisms are assumed to be reliable and strong. The majority of current adversarial FL research focuses on backdoor attacks that target parameter updates, model replacement, and gradient poisoning [2,6], so that the temporal alignment layer remains an unexplored attack surface, especially in systems where synchronized heterogeneous modalities are crucial to fusion performance.

In this paper, we present a study that examines the exploitation of synchronization vulnerabilities in MOFL. We show that small and realistic temporal perturbations, such as timestamp shifts or window misalignments, can significantly degrade fusion quality and amplify adversarial effects. Unlike traditional poisoning attacks, these synchronization-based manipulations operate at the input alignment level, making them subtle, difficult to detect, and often indistinguishable from benign sensor delays or network jitter.

Specifically, we investigate temporal desynchronization attacks in MOFL, where adversarial clients introduce controlled temporal shifts between modalities to disrupt cross-modal alignment. By performing specific experiments on heterogeneous multimodal data, we demonstrate that even minor misalignments can lead to noticeable performance degradation and increased attack effectiveness under partial adversarial participation. To address this vulnerability, we propose a lightweight Cross-Client Consistency Scoring Mechanism (CCCSM), which reweights client updates based on their agreement during aggregation. Our results indicate that CCCSM effectively mitigates the influence of adversarial updates while maintaining performance on clean data.
Our work makes the following contributions:

1. We identify temporal synchronization in multimodal online federated learning as a potential adversarial attack surface.
2. We systematically study cross-modal desynchronization attacks and analyze their impact on multimodal fusion and model performance.

3. We demonstrate that temporal misalignment can induce both global performance degradation and class-specific prediction bias.
4. We propose and evaluate a lightweight consistency-based defense that improves robustness against desynchronization attacks.

Our research suggests that in streaming systems, robustness can collapse when modalities go out-of-sync, even under minor perturbations, highlighting the urgent need to reconsider synchronization assumptions in these systems.

The rest of the article is organized as follows: Sect. 2 reviews studies relevant to the proposed approach. Section 3 presents the system model, threat model, attack design, and CCCSM defense mechanism. Section 4 describes the dataset, preprocessing, baseline model, attack settings, and defense implementation. Section 4.6 discusses the results and key findings. Finally, Sect. 5 summarizes the study and outlines future directions.

2 Related Work

Multimodal learning systems rely on the fusion of heterogeneous data sources, where temporal synchronization is essential for constructing meaningful cross-modal representations. Prior work has shown that even minor temporal inconsistencies can significantly degrade fusion quality. For instance, sequence contrastive learning approaches employ dynamic time warping to align video and sensor streams, demonstrating the importance of precise temporal coordination [11]. Similarly, latent space alignment methods in healthcare explicitly enforce synchronization across modalities such as imaging and biosignals to ensure consistent feature representations [18]. Despite these advances, synchronization remains an active challenge rather than a fully resolved problem.

Multimodal datasets require careful synchronization between different modalities such as IMU signals, video streams, and textual annotations, highlighting that maintaining temporal alignment in real-world scenarios is non-trivial [10]. Similarly, O. Anigala et al. explore that temporal misalignment is a hidden data pathology that negatively affects model robustness and downstream performance [1]. These challenges become more pronounced in streaming environments, where data arrives continuously and must be aligned in real time.

In OFL, clients continuously generate and process streaming data, introducing additional complexity due to temporal drift, asynchronous updates, and heterogeneous sampling rates. For example, online spatio-temporal FL for traffic flow forecasting relies on dynamic temporal correlations and period-aware aggregation, implicitly assuming reliable temporal ordering [8]. Similarly, MOFL frameworks addressing missing modalities construct prototypes over synchronized temporal windows, again assuming consistent alignment across modalities [20]. In practical deployments such as precision agriculture systems, synchronization under network delays and clock drift remains an open challenge, with no standardized mechanisms to ensure temporal consistency [16]. These works highlight that OFL systems depend heavily on temporal integrity but do not consider adversarial manipulation of timestamps.

Recent advances in transformer-based and attention-driven multimodal fusion further rely on accurate temporal relationships. Attention mechanisms learn dependencies across modalities based on temporal ordering, as demonstrated in multimodal transformer architectures for coordinated attack detection [13]. Similarly, domain-driven threat detection and cross-modal behavioral prediction frameworks integrate temporal information across multiple data sources but operate at coarse temporal granularity and assume trustworthy timestamps [4,14]. As a result, fine-grained temporal perturbations remain largely undetected in such systems.

Robustness in FL has primarily been addressed through parameter-space defenses, including Byzantine-resilient aggregation and anomaly detection of model updates. Methods such as robust aggregation and trajectory-based anomaly detection aim to identify malicious updates based on their deviation from benign behavior [2,3]. However, these approaches operate on model updates rather than input data and therefore cannot detect attacks that corrupt the temporal alignment before model training.

In addition, graph-based and spatiotemporal models depend critically on correct temporal relationships for constructing meaningful structures. Federated graph neural networks and dynamic spatiotemporal models rely on timestamp-based dependencies, which can be invalidated by temporal perturbations [9,19]. Similarly, imputation-based methods for handling missing modalities depend on accurate temporal context, and their performance degrades when alignment is compromised [22].

Overall, existing literature assumes that temporal synchronization is reliable and focuses primarily on noise reduction, missing data, or parameter-level robustness. The adversarial implications of temporal misalignment, particularly in MOFL, remain largely unexplored. This gap motivates our work, which systematically investigates temporal desynchronization as an attack vector and proposes a cross-client consistency-based defense mechanism.

3 Methodology

We model a realistic MOFL pipeline with synchronized heterogeneous modalities and investigate how temporal desynchronization can be exploited as an adversarial mechanism. Figure 1 illustrates the overall workflow of the proposed framework, highlighting both benign and adversarial client pipelines. Unlike traditional attacks that explicitly manipulate labels or gradients, the proposed approach targets the synchronization process, leading not only to performance degradation but also to systematic misalignment in learned cross-modal representations. To address this, we propose a lightweight synchronization-consistency defense that improves robustness in streaming federated environments.

3.1 System Model

We consider an OFL setting with N distributed clients, where each client processes locally available multimodal data streams. In practice, we simulate the

OFL setting using a pre-collected multimodal dataset by treating each subject as a client and sequentially feeding temporally ordered data to mimic streaming behavior. Each client $i \in \{1, \ldots, N\}$ possesses a set of modalities:

$$\mathcal{M} = \{1, 2, \ldots, M\} \tag{1}$$

where each modality $m \in \mathcal{M}$ corresponds to a distinct sensor stream (e.g., wearable inertial sensors and camera-based observations). Due to heterogeneous sensing mechanisms, modalities operate at different sampling rates and temporal resolutions.

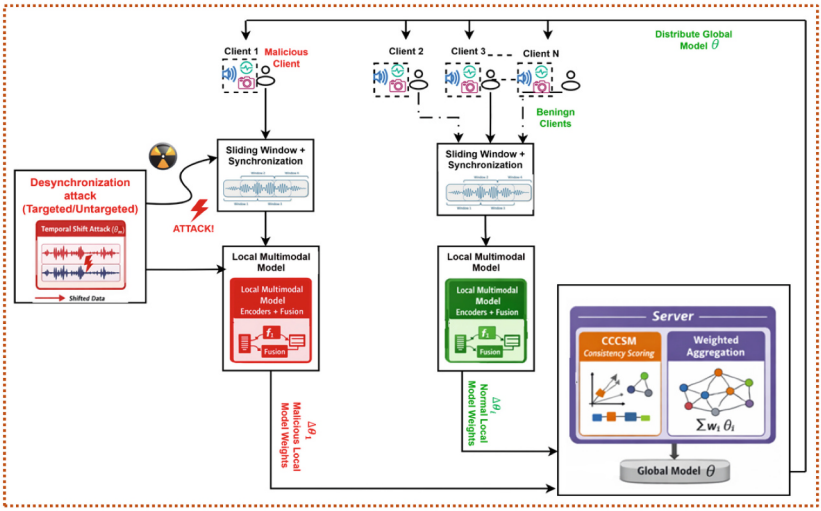

Fig. 1. The Overview of the Proposed Methodology

Temporal Windowing and Synchronization: To enable multimodal fusion, each client constructs aligned samples using a sliding window of size T. For modality m, the raw stream is denoted as

$$x_m(t) \tag{2}$$

A synchronization function $\mathcal{S}(\cdot)$ aligns modality-specific streams into a unified representation:

$$\tilde{x}_m^{(k)} = \mathcal{S}\big(x_m(t)\big), \quad t \in [t_k, t_k + T] \tag{3}$$

where k indexes the temporal window. This process involves resampling and timestamp alignment to ensure that multimodal observations correspond to the same underlying activity segment despite differing sampling rates.

Local Multimodal Model: Each client maintains modality-specific encoders:

$$z_m^{(k)} = f_m\big(\tilde{x}_m^{(k)}; \theta_m\big) \tag{4}$$

The modality embeddings are fused via a fusion function $\mathcal{F}(\cdot)$:

$$z^{(k)} = \mathcal{F}\big(z_1^{(k)}, z_2^{(k)}, \ldots, z_M^{(k)}\big) \tag{5}$$

The fused representation is passed to a task-specific head $h(\cdot)$ to produce predictions:

$$\hat{y}^{(k)} = h\big(z^{(k)}; \theta_h\big) \tag{6}$$

The local objective for client i is defined as:

$$\mathcal{L}_i = \mathbb{E}_{(x,y)}\big[\ell(\hat{y}, y)\big] \tag{7}$$

Online Federated Optimization: Clients update local models using temporally ordered data and periodically send model updates $\Delta\theta_i$ to the server. The server aggregates updates as:

$$\theta^{t+1} = \sum_{i \in \mathcal{C}_t} w_i \cdot \theta_i^t \tag{8}$$

where $\mathcal{C}_t$ denotes the set of participating clients at round t, and w_i represents aggregation weights.

3.2 Threat Model

We consider a subset of clients $\mathcal{A} \subset \mathcal{C}_t$ to be adversarial. The adversary has the following capabilities: (i) access to local multimodal data streams, (ii) ability to manipulate temporal alignment of modalities before fusion, and (iii) ability to participate intermittently in OFL. The adversary, however, does not control the server or the aggregation mechanism. The objective of the attacker is to degrade the global model performance, disrupt cross-modal fusion, induce systematic misclassification for specific activity patterns, and remain stealthy under standard aggregation rules.

3.3 Cross-Modal Desynchronization Attack

We introduce a synchronization-based attack that perturbs temporal alignment between modalities.

Attack Formulation: After synchronization, for a target modality m, the adversary applies a temporal shift:

$$x_m(t) \rightarrow x_m(t + \delta_m) \tag{9}$$

where δ_m is a small time offset relative to the sampling rate. The synchronized representation becomes:

$$\tilde{x}_m^{(k)} = \mathcal{S}\big(x_m(t + \delta_m)\big) \tag{10}$$

while other modalities remain unchanged.

Attack Impact: Fig. 2 illustrates the effect of temporal alignment on multimodal fusion, comparing synchronized and desynchronized modality streams. It shows that even small temporal shifts (δ_m) lead to misaligned signals, resulting in inconsistent fusion and degraded representation learning.

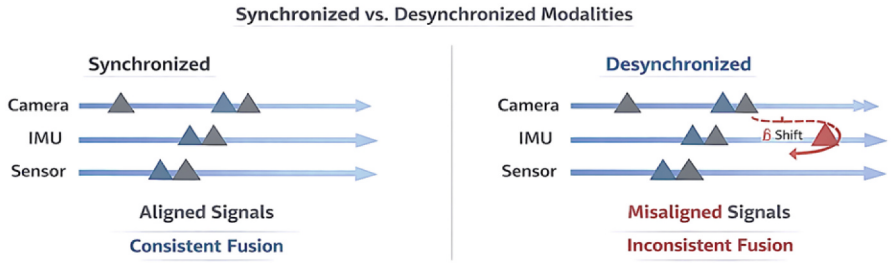

Fig. 2. The Comparison of Synchronized and Desynchronized Multimodal Signals

The intuition behind this attack lies in the assumption that multimodal fusion requires modality-specific encodings to represent the same underlying temporal event. Formally:

$$z_m^{(k)} \approx z_{m'}^{(k)}, \quad \forall m, m' \in \mathcal{M} \tag{11}$$

Desynchronization violates this assumption and illustrates that (a) embeddings correspond to different temporal contexts, (b) fusion introduces inconsistent cross-modal features and (c) the model learns spurious or incorrect relationships between modalities.

Although the perturbation is applied at the input level, its effect propagates through the learning process. In particular, it is executed in these two settings:

– **Global degradation:** In this setting, the temporal shift is applied to all samples at adversarial clients. Misaligned multimodal inputs introduce noise in representation learning, reducing overall model accuracy.

– **Implicit targeted effect:** In this setting, we apply temporal shifts selectively to samples belonging to a specific activity class; the model is exposed to consistently misaligned patterns for that class. Over multiple OFL rounds, this results in systematic confusion and biased predictions for those activities.

Moreover, the OFL setting itself is inherently susceptible to amplification of desynchronization effects due to continuous exposure to temporally ordered data, temporal drift in local data distributions, and asynchronous and partial client participation. Overall, our proposed attack exhibits the following characteristics:

1. Even small values of δ_m (e.g., on the order of milliseconds) are sufficient to significantly degrade model performance.
2. The perturbation closely resembles natural sensor delays, clock drift, or network-induced latency. As a result, it is difficult to distinguish adversarial manipulation from benign temporal inconsistencies.
3. It targets the synchronization layer rather than the model parameters or labels, thereby disrupting cross-modal representation learning.
4. It does not depend on a specific fusion strategy or model architecture. It can be applied to early, late, or attention-based fusion mechanisms, making it broadly applicable across different MOFL systems.

3.4 Cross-Client Consistency Scoring Mechanism (CCCSM)

The proposed Cross-Client Consistency Scoring Mechanism (CCCSM) is based on the observation that in a well-synchronized multimodal setting, benign client updates tend to exhibit consistent optimization directions. In contrast, desynchronization disrupts cross-modal alignment at the representation level, leading to inconsistent and noisy model updates.

Server-Side Consistency Scoring: CCCSM operates entirely at the server side and does not require access to raw data. For each participating client i, the server computes a consistency score based on the similarity [23] between its update $\Delta\theta_i$ and those of other clients in the same round:

$$c_i = \frac{1}{|\mathcal{C}_t|} \sum_{j \in \mathcal{C}_t} \cos(\Delta\theta_i, \Delta\theta_j) \tag{12}$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity.

Clients Weighting:

The computed consistency scores are used to reweight client contributions during aggregation by using the following Eq. (13)

$$w_i = \frac{c_i}{\sum_{j \in \mathcal{C}_t} c_j} \tag{13}$$

A lower consistency score indicates potential desynchronization-induced deviation. Therefore, clients with lower consistency scores are down-weighted, reducing their influence on the global model.

The key properties of our proposed defense are that it is model-agnostic, relies only on lightweight consistency scoring over client updates, and does not require access to labeled validation data or additional local training, thereby incurring low computational overhead and making it suitable for wearable and IoT-based MOFL deployments with limited memory, battery, and processing resources.

However, the framework has certain limitations as well. In particular, CCCSM assumes that benign clients constitute the dominant portion of participating updates in each round. Under coordinated malicious-majority settings, update similarity alone may be insufficient for reliable discrimination.

4 Experimental Setup

4.1 Dataset

We conduct our evaluation using the UP-Fall dataset [12], a multimodal human activity recognition dataset designed for fall detection and daily activity monitoring. The dataset consists of 11 activities, including six daily activities and five types of falls, performed by 17 subjects under controlled conditions.

The dataset provides synchronized multimodal recordings, including wearable sensor signals (accelerometer and gyroscope) and vision data captured from multiple cameras. These modalities exhibit different temporal resolutions and sensing characteristics, making the dataset suitable for studying cross-modal synchronization in MFL.

4.2 Preprocessing and Feature Construction

Following the feature construction pipeline provided with the UP-Fall dataset [12], the raw multimodal streams are segmented using sliding windows of duration $T \in \{1, 2, 3\}$ seconds with 50% overlap. Unless otherwise stated, we report results for $T = 2$ seconds as a balanced configuration between temporal resolution and contextual information.

For the sensor modality, handcrafted features are extracted from each windowed segment. Specifically, 18 features per signal are used, including temporal statistics (e.g., mean, standard deviation, quartiles) and frequency-domain features (e.g., spectral energy and entropy). Given the multiple sensor channels, this results in a high-dimensional feature representation per window.

For the vision modality, we adopt the Horn and Schunck optical-flow-based representation [5,12]. Motion between consecutive frames is computed, transformed into magnitude representations, resized to a fixed spatial resolution, and aggregated within each window to obtain a compact feature vector.

The resulting modality-specific features are normalized and used as inputs to the multimodal learning model.

4.3 Baseline Implementation Model

We employ a lightweight multimodal architecture consisting of modality-specific encoders and a fusion module. Each modality is processed using a dedicated multilayer perceptron (MLP) encoder to obtain latent embeddings. The embeddings are fused via concatenation and passed to a classification head for activity prediction. This design enables effective integration of heterogeneous modalities while maintaining computational efficiency, which is important in FL environments.

We simulate an MOFL environment by treating each subject as an individual client, resulting in $N = 17$ clients. Although the dataset is pre-collected, we emulate streaming behavior by feeding temporally ordered windowed samples sequentially to each client. Training is performed over 100 communication rounds. At each round, a subset of clients (10%) is randomly selected to participate in training. Each selected client updates its local model using newly available data segments and sends model updates to the server for aggregation.

To simulate realistic non-IID data distributions, we partition the data using a Dirichlet distribution over activity labels. We set the concentration parameter α, which governs the degree of data heterogeneity across clients, to 0.3, representing a moderately high level of non-IID distribution. Moreover, the local model is trained using the Adam optimizer with a learning rate of 0.001 and a batch size of 32. The hidden dimension of the MLP encoders is set to 128 with ReLU activation. The proposed CCCSM defense is applied at the server side during aggregation by reweighting client updates based on consistency scores.

4.4 Attack Configuration

The attack is realized by perturbing the temporal alignment of a selected modality at the raw signal level before the synchronization and windowing process. The detailed procedure of the proposed cross-modal desynchronization attack is presented in Algorithm 1.

For a given modality m, the adversarial transformation is defined as Eq. (9). Rather than defining δ_m as an absolute time value, we express it relative to the sampling interval of the modality:

$$\delta_m = k \cdot \Delta t_m \tag{14}$$

Here, Δt_m denotes the sampling interval and k is the number of shifted time steps. In our experiments, we consider small values of $k \in \{1, 2, 3, 5\}$. In our dataset, wearable sensor signals are sampled at approximately 50 Hz, corresponding to a sampling interval of $\Delta t_m \approx 20$ ms. Therefore, these values of k result in temporal offsets in the range of approximately 20–100 ms. For the vision modality, which is captured at a fixed frame rate, the perturbation corresponds to shifting the sequence by a small number of frames, leading to minor frame-level misalignment across modalities.

This perturbation is applied prior to the synchronization function $\mathcal{S}(\cdot)$, ensuring that the resulting multimodal windows contain temporally inconsistent information across modalities.

The attack is executed locally at adversarial clients during each communication round. For each selected client, the temporal perturbation is applied consistently across its local training data according to the attack configuration. After perturbation, the client performs standard local training and sends model updates to the server. Consequently, the adversarial effect is embedded implicitly in the learned representations and propagates through the federated aggregation process.

Untargeted and Label-Conditioned Settings To comprehensively evaluate the impact of desynchronization, we consider two attack settings:

Untargeted Desynchronization: In this setting, the temporal shift is applied to all samples at adversarial clients. The primary effect is a global degradation in model performance due to disrupted cross-modal alignment. The model receives inconsistent multimodal inputs, leading to noisy representations and reduced predictive accuracy.

Label-Conditioned Desynchronization: In this setting, the perturbation is applied selectively to samples belonging to a specific source class c_s. Only samples with label $y = c_s$ are temporally shifted, while all other samples remain unchanged. Over multiple communication rounds, this selective exposure introduces a systematic bias in the learned representations for class c_s. The model associates inconsistent multimodal patterns with this class, leading to increased confusion with other activities. This behavior resembles a targeted adversarial effect, even though no explicit trigger or label manipulation is introduced.

4.5 Defense Implementation

We apply the proposed CCCSM during server-side aggregation at each communication round. Upon receiving client updates $\{\Delta\theta_i\}_{i\in\mathcal{C}_t}$, the server first flattens each update into a vector form and computes pairwise cosine similarity between updates [23]. The consistency score for each client i calculated as Eq. (12), is defined as the average similarity with respect to all participating clients:

To ensure stable aggregation, the scores are normalized using a softmax function:

$$w_i = \frac{\exp(c_i)}{\sum_{j\in\mathcal{C}_t}\exp(c_j)} \tag{15}$$

These normalized weights are used to reweight client contributions during aggregation, thereby reducing the influence of inconsistent or adversarial updates.

Algorithm 1 Cross-Modal Desynchronization Attack in MOFL

Require: Participating clients $\mathcal{C}_t$, adversarial subset $\mathcal{A}_t$, target modality m, shift step k, sampling interval Δt_m, optional source class c_s
Require: Window size T, synchronization function $\mathcal{S}(\cdot)$, global model θ^t
Compute temporal shift $\delta_m \leftarrow k \cdot \Delta t_m$
for all client $i \in \mathcal{C}_t$ **do**
  Initialize local model with global parameters θ^t
  for all local samples (x, y) at client i **do**
    if $i \in \mathcal{A}_t$ **then**
      if c_s is not specified **or** $y = c_s$ **then**
        Apply temporal shift to modality m:

$$x_m(t) \leftarrow x_m(t + \delta_m)$$

      end if
    end if
    Construct synchronized windowed sample using $\mathcal{S}(\cdot)$ over interval $[t_k, t_k + T]$
  end for
  Perform local training on the resulting multimodal windows
  Compute local update $\Delta\theta_i^t$
  Send $\Delta\theta_i^t$ to server
end for
Server aggregates updates to obtain θ^{t+1}

In implementation, the mechanism is applied at every communication round without introducing additional communication overhead, as it operates solely on received model updates.

4.6 Results and Discussion

We evaluate the system under three experimental settings. First is the *no attack* scenario, where standard MOFL is performed without any adversarial intervention, serving as a baseline for comparison. Second is the *desynchronization attack* setting, in which a subset of adversarial clients introduces temporal shifts in one modality during local training. Finally, we consider the *attack with CCCSM* setting, and apply proposed defense mechanism at the server during aggregation to mitigate the impact of such adversarial updates.

Performance is evaluated using attack-oriented metrics widely adopted in the literature [6], including the attack success rate (ASR), F1-score on clean samples (Normal F1), F1-score on perturbed samples (Attack F1), and overall F1-score.

Table 1 presents a detailed evaluation of the proposed desynchronization attack under varying temporal shift strengths and attack configurations.

First, we observe that even small temporal shifts ($k = 1$) lead to noticeable degradation in performance, confirming the sensitivity of multimodal fusion to temporal alignment. As the shift magnitude increases, the attack success rate consistently rises, while both Attack F1 and Overall F1 decline. This trend indi-

Table 1. Attack-oriented evaluation under untargeted and label-conditioned desynchronization for different shift strengths k. Results are averaged across clients on the UP-Fall dataset.

Setting	k	ASR (%)	Normal F1 (%)	Attack F1 (%)	Overall F1 (%)
No Attack (Baseline)					
Baseline (MOFL)	–	0.0	95.6	–	95.6
Without Defense					
Untargeted	1	40.9	91.3	87.2	89.1
	3	56.8	89.4	80.9	85.0
	5	68.9	87.1	75.4	80.5
Targeted	1	52.6	90.8	85.1	87.6
	3	67.9	88.8	79.6	84.2
	5	79.2	86.5	73.0	79.1
With CCCSM Defense					
Untargeted + CCCSM	1	19.4	92.1	90.0	91.1
	3	28.1	90.9	86.4	89.0
	5	39.8	89.3	82.2	85.6
Targeted + CCCSM	1	25.2	91.6	88.7	90.2
	3	33.4	90.2	85.8	88.7
	5	45.1	88.8	81.0	84.7

cates that stronger temporal misalignment introduces progressively severe inconsistencies in the learned cross-modal representations.

We also note that the label-conditioned (targeted) setting consistently exhibits higher ASR compared to the untargeted case, along with slightly lower Attack F1 and Overall F1 values. This suggests that selectively applying temporal perturbations to specific activity classes induces stronger and more focused misclassification behavior, effectively acting as a targeted adversarial mechanism even in the absence of explicit trigger patterns.

Additionally, our results demonstrate that the proposed CCCSM defense significantly reduces ASR across all values of k and improves both Attack F1 and Overall F1. Notably, the improvement is more pronounced for smaller perturbations, while the recovery becomes less effective as the shift magnitude increases. This behavior indicates that consistency-based weighting effectively suppresses anomalous client updates and stabilizes the learning process, particularly under mild to moderate attack conditions.

However, it is important to mention that the performance does not fully recover to the no-attack baseline, especially for larger values of k. This suggests that temporal desynchronization introduces persistent distortions in cross-modal representations that cannot be entirely eliminated through aggregation alone.

To provide further insight into the effect of temporal desynchronization, Fig. 3 illustrates the variation of Overall F1 with respect to the shift strength k. From the plot, it is evident that model performance degrades consistently as k increases, highlighting the sensitivity of multimodal fusion to temporal misalignment. This degradation is more pronounced in the targeted setting, where we apply temporal perturbations on selected labeled classes.

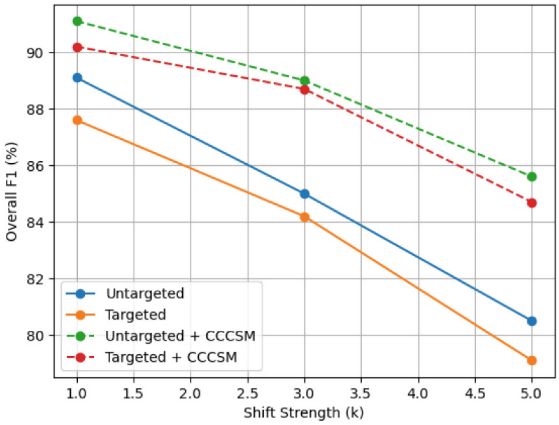

Fig. 3. Impact of temporal shift strength k on model performance.

Overall, the results demonstrate that temporal desynchronization constitutes a practical and effective attack vector in MOFL, while the proposed CCCSM defense provides a lightweight and effective mitigation strategy.

5 Conclusion and Future Work

In this paper, we investigated temporal desynchronization as a practical adversarial threat in MOFL and showed that even small temporal shifts can significantly degrade performance and induce systematic errors. We further demonstrated that the impact of the attack increases with the magnitude of misalignment and is more pronounced in targeted settings. To address this threat, we proposed CCCSM, a lightweight defense that effectively reduces the ASR and improves robustness without requiring additional data. Furthermore, since the proposed approach is modality-agnostic, it is not limited to the specific sensor–image modalities considered in this study and can be extended to other multimodal datasets and application domains.

However, our results also show that CCCSM does not fully recover performance under stronger perturbations, motivating the development of more adaptive synchronization-aware defenses. Moreover, the evaluation is conducted using a pre-recorded wearable-sensing dataset; streaming behavior is mimicked in a controlled and reproducible setting. Therefore, future work includes evaluating CCCSM on larger and more diverse datasets, comparing it against established FL defenses, and assessing its efficiency and robustness in real-world asynchronous federated environments.

Disclosure of Interests. There is no conflict of interest.

References

1. Anigala, O., Okolo, J.N., Son, J., Pack, C.: Mm-score: a two-factor framework for auditing multimodal dataset quality. In: Proceedings of the International Conference on Research in Adaptive and Convergent Systems, pp. 1–8 (2025)
2. Batur Şahin, C.: Securing UAV swarms with vision transformers: a byzantine-robust federated learning framework for cross-modal intrusion detection. Drones **10**(2), 125 (2026)
3. Ding, Z., et al.: Identifying alternately poisoning attacks in federated learning online using trajectory anomaly detection method. Sci. Rep. **14**(1), 20269 (2024)
4. Gu, Z., Guo, J., Xu, J., Sun, Y., Liang, W.: Domain knowledge-driven method for threat source detection and localization in the power internet of things. Electronics **14**(13), 2725 (2025)
5. Horn, B.K., Schunck, B.G.: Determining optical flow. Artif. Intell. **17**(1–3), 185–203 (1981)
6. Lai, Y., Wu, L., Lin, H., Liu, J.: Bad-MFL: a cross-modality bi-trigger backdoor attack against multi-modal federated learning. IEEE Internet Things J. (2025)
7. Li, L., Fan, Y., Tse, M., Lin, K.Y.: A review of applications in federated learning. Comput. Ind. Eng. **149**, 106854 (2020)
8. Liu, Q., Sun, S., Liu, M., Wang, Y., Gao, B.: Online spatio-temporal correlation-based federated learning for traffic flow forecasting. IEEE Trans. Intell. Transp. Syst. **25**(10), 13027–13039 (2024)
9. Liu, R., Xing, P., Deng, Z., Li, A., Guan, C., Yu, H.: Federated graph neural networks: Overview, techniques, and challenges. IEEE Trans. Neural Networks Learn. Syst. (2024)
10. Ma, J., et al.: T3Set: a multimodal dataset with targeted suggestions for LLM-based virtual coach in table tennis training. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, vol. 2. pp. 5686–5697 (2025)
11. Ma, Y., He, H., Wang, X., Wu, G.: Sequence contrastive learning for multi-modal alignment in heavy plate rolling. In: 2025 7th International Conference on Industrial Artificial Intelligence (IAI), pp. 1–6. IEEE (2025)
12. Martínez-Villaseñor, L., Ponce, H., Brieva, J., Moya-Albor, E., Núñez-Martínez, J., Peñafort-Asturiano, C.: Up-Fall detection dataset: a multimodal approach. Sensors **19**(9), 1988 (2019)
13. Nooraien, M.A., Kumar, P.D., Priya, M.B.L., Nadeem, P., et al.: A multimodal transformer architecture for detecting coordinated network attacks. In: 2025 5th International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN), pp. 992–997. IEEE (2025)
14. Pu, X., Ren, Y., Liang, J.: ACMBP: an adaptive graph-based meta-learning framework for cross-modal behavior prediction from IoT and social media streams. Informatica **50**(9) (2026)
15. Quan, M.K., et al.: Federated learning for cyber physical systems: a comprehensive survey. IEEE Commun. Surv. Tutorials (2025)
16. Rao, S., Neethirajan, S.: Computational architectures for precision dairy nutrition digital twins: a technical review and implementation framework. Sensors **25**(16), 4899 (2025)
17. Sahoo, S.: Sensor fusion and virtual sensor design for enhanced multi-sensor data accuracy in autonomous systems. Int. J. Smart Sustain. Intell. Comput. **1**(2), 21–39 (2024)

18. Shrivastava, A., Kholmatovna, U.N., Kannimuthu, S., MohamadAbbas, H., Thakur, D., Priyanka, T.P.: Cross-modal neural networks for real-time multi-sensory information fusion in smart healthcare. In: 2025 International Conference on Innovations and Emerging Technologies In AI & Communication Systems (IETACS), pp. 1567–1570. IEEE (2025)
19. Tan, C., Hu, J., Li, J., Miao, M., Hu, W., Wang, S.: DTS-MixNet: dynamic spatiotemporal graph mixed network for anomaly detection in multivariate time series. Big Data Cogn. Comput. **9**(10), 245 (2025)
20. Wang, H., Liu, X., Zhong, X., Chen, L., Liu, F., Zhang, W.: Multimodal online federated learning with modality missing in internet of things. IEEE Trans. Mobile Comput. (2025)
21. Wang, H., Zhong, X., Liu, K., Liu, F., Zhang, W.: Denoising and adaptive online vertical federated learning for sequential multi-sensor data in IIoT. IEEE Trans. Mobile Comput. (2025)
22. Wang, J., et al.: Patch: a plug-in framework of non-blocking inference for distributed multimodal system. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 7(3), pp. 1–24 (2023)
23. Xia, P., Zhang, L., Li, F.: Learning similarity with cosine similarity ensemble. Inf. Sci. **307**, 39–52 (2015)