# Fed-MultiFND: Federated Multimodal Learning for Imbalanced Short Video Fake News Detection

Jixuan Guo, Yihan Gao, Dabao Zhang, Yikun Liu, Boyue Wang*
School of Beijing University of Technology, Beijing, China
Email: guo040723@emails.bjut.edu.cn, aheadgyh@emails.bjut.edu.cn, wby@bjut.edu.cn

***Abstract*—Short video platforms have become a dominant medium for news dissemination, significantly amplifying the spread and impact of misinformation. Compared to traditional text-based news, short video fake news detection is more challenging due to heterogeneous multimodal data and stringent requirements on both accuracy and efficiency. Moreover, real-world deployment is constrained by data privacy regulations that prohibit centralized cross-platform data collection, while platform-specific data often exhibits uneven distributions of fake and real news, further hindering model learning. Motivated by these challenges, we propose a unified framework,Fed-MultiFND, for multimodal short video fake news detection that jointly addresses modeling complexity and cross-platform learning constraints. The framework integrates a lightweight yet effective multimodal base model with a federated learning architecture, enabling collaborative training without raw data sharing. Specifically, hierarchical attention is employed to progressively fuse textual, auditory, and visual information, while federated learning facilitates knowledge aggregation across platforms under skewed data distributions. Through this integrated design, the proposed approach achieves a practical improvement between detection effectiveness, robustness to distributional imbalance, and real-world deployment feasibility.**

***Index Terms*—Short Video Fake News Detection, Multi-Modal Learning, Attention Mechanism, Federated Learning, data imbalance**

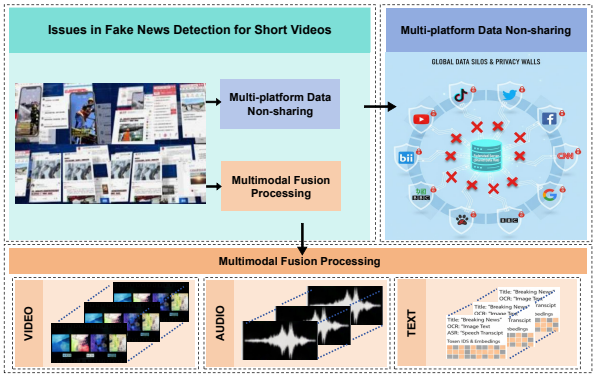


Fig. 1. This figure delineates the core bottlenecks impeding fake news detection in short-form videos, which are grouped into two primary challenges: multi-platform data non-sharing and multimodal fusion processing. The top-level overview synthesizes these two critical limitations. The right-hand panel illustrates the multi-platform data non-sharing issue: due to privacy constraints and platform-specific policies, the global ecosystem is fragmented into isolated data silos, with blocked data flows between major social media platforms. The lower panel details the multimodal fusion processing challenge, which addresses the need for coordinated semantic modeling across heterogeneous video, audio, and text modalities to capture the full contextual semantics of short-form content.

## I. Introduction

Short video platforms have rapidly emerged as a primary channel for information consumption, enabling news content to be disseminated at unprecedented speed and scale [5]. While this trend improves accessibility and user engagement, it also exacerbates the spread of fake news. Due to the vivid visual presentation and strong emotional appeal of short videos, misleading information can exert a more profound influence on public opinion than traditional text-based news. Overall, the current problems and challenges in short video fake news detection are concentrated in two aspects,multi-platform data non-sharing and multimodal fusion processing, which correspond to Fig. 1.

Detecting fake news in short videos poses significant challenges. First, short videos inherently contain heterogeneous multimodal data, including visual, auditory, and textual streams with distinct temporal structures and semantic properties. Effectively modeling cross-modal interactions while suppressing modal noise remains a non-trivial problem. Early studies have shown that naive feature concatenation or shallow fusion strategies are insufficient for capturing complex inter-modal dependencies, often leading to suboptimal detection performance [6], [19], [29]. Recent deep learning-based approaches attempt to enhance multimodal reasoning via attention mechanisms and adversarial learning [10], [27], yet they often incur high computational costs and face deployment challenges in real-world systems.

Beyond the challenges in multimodal fusion modeling, real-world deployment also faces critical constraints arising from the lack of cross-platform data interoperability. Short video platforms typically operate under strict data privacy regulations, which prohibit the sharing of short video data across platforms. As a result, data collected by different platforms remain isolated, limiting the applicability of centralized training strategies and hindering the development of detection models that can generalize well across heterogeneous data distributions.Moreover, the absence of cross-platform data sharing leads to inconsistent or even conflicting descriptions of the same event across platforms, further exacerbating inter-platform data imbalance. This isolation forces fake news detection to rely heavily on manual verification. Due to the inability to synchronize information in real time, such verification is often completed only after misinformation has already propagated for a prolonged period. This delay substantially

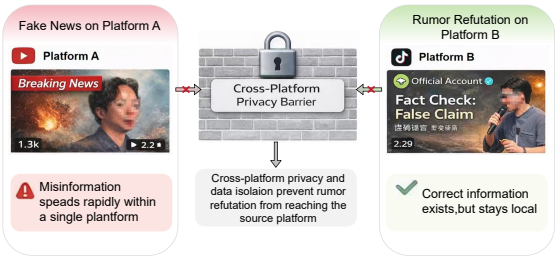


Fig. 2. Schematic diagram of fake news spread exacerbated by cross-platform data isolation. Specifically, the lack of data interoperability leads to divergent claims on the same event across platforms, and manual verification after prolonged dissemination further amplifies the spread of misleading content.

amplifies the spread of misleading content and magnifies its potential societal impact, as illustrated in Fig. 2.

To address these challenges, we propose **Fed-MultiFND**, a unified framework that jointly tackles multimodal fusion complexity and cross-platform data non-sharing. Our approach consists of two tightly coupled components: (1) a lightweight yet effective multimodal base model that performs hierarchical attention-based fusion of text, audio, and visual features; and (2) a privacy-preserving federated learning architecture that enables collaborative model training across platforms without exposing raw data [11]. Experiments on two real-world datasets demonstrate the superiority of the proposed Fed-MultiFND over eight baseline methods. Our main contributions are as follows:

- To the best of our knowledge, we are the first to systematically analyze short video fake news detection from both multimodal modeling and cross-platform deployment perspectives, identifying multimodal heterogeneity and data isolation as two fundamental challenges.
- We propose a hierarchical multimodal attention-based base model that progressively fuses text, audio, and visual information, achieving efficient and robust multimodal representation learning.
- We introduce a federated learning framework tailored for short video fake news detection, enabling privacy-preserving cross-platform collaboration under both balanced and imbalanced data distributions.

## II. Related Work

### A. Multimodal Fake News Detection for Short Videos

Early studies on video-based fake news detection primarily relied on handcrafted multimodal features and shallow fusion strategies. Hou et al. [29] and Medina Serrano et al. [19] explored misinformation detection in medical and COVID-19 YouTube videos using speech transcripts and textual features, while Shang et al. [6] extended this to TikTok short videos with speech-guided visual representations. Hybrid or shallow models also incorporated textual, visual, and social-context features [1], [12], but exhibited limited cross-modal reasoning [2], [24].

Recent work leverages deep learning to model complex cross-modal dependencies. Transformer-based frameworks [3], [14], [30] and process-aware approaches [4] improve semantic alignment and behavior modeling. Further advances include multimodal contrastive learning [13], [21], [31], multi-granularity semantic fusion [21], and large short-video datasets such as FakeSV [20] and MFND [23]. Despite improved accuracy, these models often rely on heavy pretrained networks, limiting scalability [7], [8].

### B. Federated and Privacy-Preserving Learning

Federated learning enables collaborative training across distributed clients without sharing raw data [11]. Methods like FedProx and FedBN address statistical heterogeneity [9], [15], and surveys summarize FL theory and challenges [25]. Privacy-preserving variants apply differential privacy, secure aggregation, and blockchain-based methods to multimodal scenarios [16], [17], [28].

### C. Limitations and Research Gap

FL-based fake news detection mainly focuses on textual data [18], and direct multimodal integration may exacerbate the issue of multimodal heterogeneity and reduce convergence stability [9], [15]. Our work addresses these gaps by combining hierarchical multimodal attention with federated optimization for privacy-preserving short video fake news detection.

## III. Methodology

To address the two core challenges of short video fake news detection, we propose the Fed-MultiFND architecture (Fig. 3) with two key innovations: 1) A federated learning (FL) mechanism for cross-platform collaborative training under privacy constraints, breaking data silos without raw data sharing; 2) A hierarchical attention-based multimodal fusion model with high training efficiency, which is fully compatible with FL scenarios.

### A. MultiFND: Hierarchical MultiModal Attention Fusion

Short video fake news contains heterogeneous text, audio and video modalities with inconsistent feature dimensions, sequence lengths and semantic structures. To exploit cross-modal complementarity and avoid semantic interference, MultiFND follows a unified pipeline: feature unification, intra-modal enhancement, hierarchical cross-modal interaction, and classification.

*1) Dynamic Routing for Feature Unification:* Dual encoders for each modality (BERT/X-CLIP for text, HuBERT/VGGish for audio, C3D/X-CLIP for video) output features with varying lengths and dimensions, hindering direct fusion. We introduce a dynamic routing mechanism to project all modalities into a unified representation of shape $[B, 32, 128]$, with attention weights adaptively adjusted by semantic relevance:

$$\beta_{i,j}^{(r+1)} = \beta_{i,j}^{(r)} + X_{\text{align},i}^{\top} W_{\text{cap}} X_{\text{align},j}, \quad W_{\text{cap}} \in \mathbb{R}^{128\times128} \tag{1}$$

$$\alpha_{i,j} = \text{softmax}(\beta_{i,j}^{(3)}), \quad X_{\text{fixed}} = \alpha X_{\text{align}} \tag{2}$$

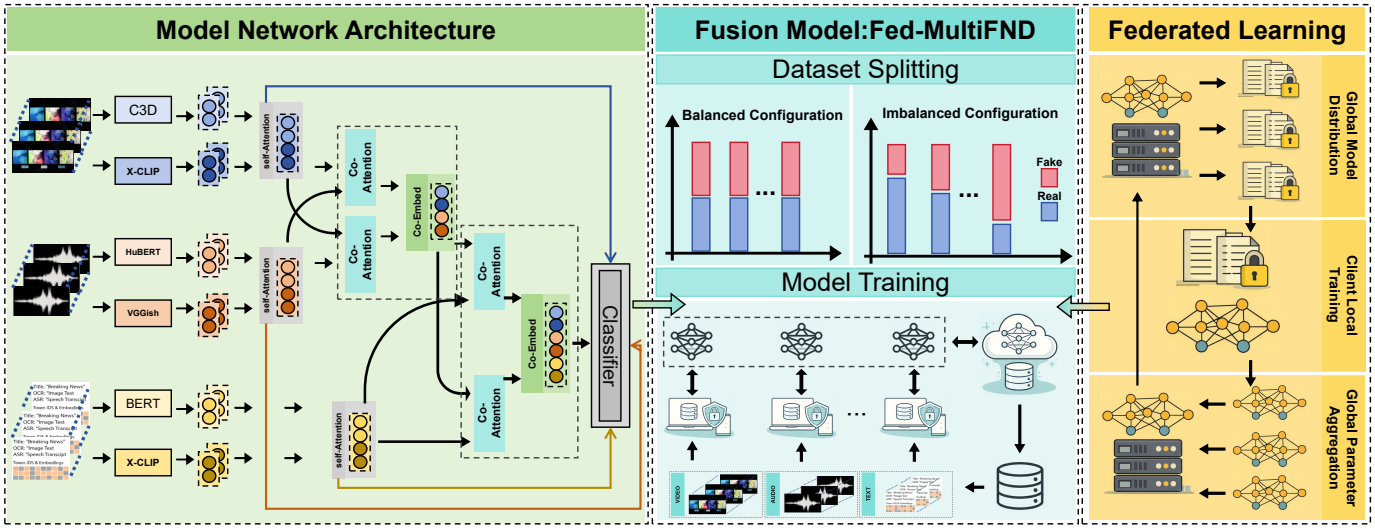


Fig. 3. Regarding the two core issues in short video fake news detection: privacy constraints that prevent data sharing across multiple platforms, and technical difficulties in effectively fusing multimodal (video/audio/text) features, this study innovatively adopts two key methods: first, a multimodal network architecture based on self-attention + multi-stage collaborative attention (to solve the multimodal fusion problem); second, a federated learning framework (to solve the data non-sharing problem, which includes three core processes: global model distribution, client-side local training, and parameter aggregation). By integrating the above two methods, a novel framework, termed Fed-MultiFND is finally constructed, which includes two key links: dataset allocation adapted to multi-platform scenarios (matching the constraint of data non-sharing), and federated training deployment combined with a multimodal architecture (realizing fake news detection under cross-platform data privacy protection).

where $\beta_{i,j}$ and $\alpha_{i,j}$ are routing coefficients and normalized weights, $X_{\text{fixed}}$ is the unified feature. This design ensures dimensional consistency while preserving fine-grained semantic cues.

*2) Dual-Encoder Intra-Modal Fusion:* Each modality contains complementary high-level semantic and low-level structural information. We concatenate unified features of dual encoders for each modality to form a comprehensive intra-modal representation:

$$X_{\text{text}} = [X^{\text{fixed}}_{\text{BERT}}; X^{\text{fixed}}_{\text{XCLIP-text}}] \in [B, 32, 256] \tag{3}$$

$$X_{\text{audio}} = [X^{\text{fixed}}_{\text{HuBERT}}; X^{\text{fixed}}_{\text{VGGish}}] \in [B, 32, 256] \tag{4}$$

$$X_{\text{video}} = [X^{\text{fixed}}_{\text{C3D}}; X^{\text{fixed}}_{\text{XCLIP-video}}] \in [B, 32, 256] \tag{5}$$

A TransformerBlock with 8 attention heads is then used to model temporal and semantic dependencies within each modality, generating enhanced intra-modal features $X^e_{\text{text}}, X^e_{\text{audio}}, X^e_{\text{video}}$ to strengthen single-modal discriminative power before cross-modal interaction.

*3) Hierarchical Cross-Modal Co-Attention Fusion:* Video and audio are inherently strongly coupled, while text provides explicit semantic guidance; direct three-modal fusion may cause semantic interference. We thus propose a hierarchical co-attention strategy for progressive semantic interaction:

1. Video-Audio Interaction: Bidirectional co-attention is applied to enhanced video and audio features to capture their inherent correlation:

$$X^{\text{va}}_{\text{video}}, X^{\text{va}}_{\text{audio}} = \text{CoAttn}(X^e_{\text{video}}, X^e_{\text{audio}}) \tag{6}$$

$$X_{\text{VA}} = W_{\text{va}}[X^{\text{va}}_{\text{video}}; X^{\text{va}}_{\text{audio}}] + b_{\text{va}}, \quad W_{\text{va}} \in \mathbb{R}^{4\times 256} \tag{7}$$

where $\text{CoAttn}(\cdot)$ is the bidirectional co-attention mechanism, and the linear layer $W_{\text{va}}$ projects concatenated features to a consistent dimension.

2. Audio-Visual-Text Interaction: Fused audio-visual features are interacted with text features to incorporate explicit semantic constraints:

$$X^t_{\text{VA}}, X^t_{\text{text}} = \text{CoAttn}(X_{\text{VA}}, X^e_{\text{text}}) \tag{8}$$

$$X_{\text{VAT}} = W_{\text{vat}}[X^t_{\text{VA}}; X^t_{\text{text}}] + b_{\text{vat}}, \quad W_{\text{vat}} \in \mathbb{R}^{4\times 256} \tag{9}$$

This two-stage fusion avoids direct mixing of weakly coupled modalities, ensuring clear semantic alignment and effective complementary information integration.

*4) Classification and Loss Function:* To fully utilize cross-modal interactive features and original intra-modal details, hierarchical fused features are concatenated with initial intra-modal features. After temporal correlation extraction via 1D convolution and dimensionality reduction via mean pooling, a multi-layer perceptron (MLP) outputs the final classification result:

$$\hat{y} = \text{MLP}(\text{MeanPool}(\text{Conv1d}(X_{\text{fused}}))) \in [B, 2]. \tag{10}$$

For binary fake/real news classification, binary cross-entropy loss is adopted for stable model convergence:

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B}\left[(1-y_i)\log \hat{y}_{i,0} + y_i \log \hat{y}_{i,1}\right], \tag{11}$$

where $y_i \in \{0, 1\}$ is the ground-truth label, $\hat{y}_{i,0}, \hat{y}_{i,1}$ are predicted probabilities of fake and real news.

### B. Federated Learning Framework

In practical deployment, short video platforms form data silos due to privacy regulations and commercial barriers, making centralized training infeasible. We design a privacy-preserving FL framework with a three-stage core workflow: global model distribution, client local training, and global parameter aggregation (Fig. 3), enabling cross-platform collaborative training without raw data sharing and adapting to non-IID cross-platform data characteristics.

Each FL communication round follows a closed loop: global model distribution $\rightarrow$ client local update $\rightarrow$ parameter aggregation for new global model. Details are as follows:

*1) Global Model Distribution:* At the start of communication round $t$, the server distributes the current global model parameter $\theta_t^c$ (superscript $c$ denotes central/global) to all $K$ clients. Client $k$ synchronizes the global parameter as its local model initial parameter:

$$\theta_t^k \leftarrow \theta_t^c, \tag{12}$$

where $\theta_t^k$ is the local model parameter of client $k$ at round $t$.

*2) Client Local Training:* Client $k$ optimizes model parameters using its private dataset $D_k$ (sample size $N_k$) and local loss $\mathcal{L}^*$. The core local update rule is:

$$\theta_{t+1}^k = \theta_t^k - \eta \nabla_\theta \mathcal{L}^* \left( f_{\theta_t^k}(\mathcal{X}_k), y_k \right), \tag{13}$$

where $\eta$ is the learning rate, $\nabla_\theta \mathcal{L}^*$ is the gradient of local loss with respect to model parameters, $\mathcal{X}_k$ and $y_k$ are private data and corresponding labels. Only model parameters are updated locally, ensuring strict privacy security.

*3) Global Parameter Aggregation:* After all clients complete local training, they upload updated parameters $\theta_{t+1}^k$ to the server. The server aggregates parameters via client sample size weighting to generate the global model parameter $\theta_{t+1}^c$ for round $t+1$:

$$\theta_{t+1}^c = \frac{\sum_{k=1}^K N_k \cdot \theta_{t+1}^k}{\sum_{j=1}^K N_j}. \tag{14}$$

The aggregated global model serves as the initial parameter for the next communication round, and the process repeats for $T$ rounds.

### C. Fusion Model: Fed-MultiFND

Fed-MultiFND is a unified framework that integrates the MultiFND base model as each client's local model with the privacy-preserving FL workflow. Each client deploys MultiFND's hierarchical multimodal attention architecture to process private short-video data, while the FL framework orchestrates cross-platform collaborative training without raw data sharing. This design simultaneously addresses the two core challenges of multimodal fusion and cross-platform data non-sharing.

*1) Dataset Splitting:* To evaluate the model's robustness under realistic cross-platform scenarios, we design two client-side data distribution configurations:

1) **Balanced Configuration**: Baseline ideal setting with a strict 1:1 fake-to-real news ratio for each client's local dataset, eliminating data imbalance-induced performance bias and providing a fair reference for heterogeneous scenario evaluation.
2) **Imbalanced Configuration**: Realistic setting with significant fake-to-real ratio variations across clients, simulating cross-platform data differences (e.g., fake news prevalence, user preferences) to validate the framework's adaptability and generalization ability under class imbalance.

All clients share the same validation and test sets to ensure fair and consistent performance comparison.

*2) Model Training:* The training process of Fed-MultiFND realizes deep integration of MultiFND's hierarchical multimodal attention architecture and the FL framework, with multimodal feature processing seamlessly embedded into the FL loop. Key design principles and formulations are as follows:

1) **Attention-Driven Local Multimodal Training** Each client's local training is built on the MultiFND attention architecture, with a task-specific local loss $\mathcal{L}^*$ and heterogeneous short-video inputs. The refined local update rule is:

$$\theta_{t+1}^k = \theta_t^k - \eta \nabla_\theta \mathcal{L}^* \left( \text{MultiFND}_{\theta_t^k} \left( x_k^{\text{text}}, x_k^{\text{audio}}, x_k^{\text{video}} \right), y_k \right), \tag{15}$$

where $\text{MultiFND}_{\theta_t^k}(\cdot)$ is the self-attention and hierarchical co-attention based MultiFND model parameterized by $\theta_t^k$.

2) **Attention-Weighted Aggregation for Heterogeneous Adaptation** To mitigate the negative impact of heterogeneous data on FL aggregation, we design an attention-weighted global aggregation rule that prioritizes parameters capturing universal fake news features:

$$\theta_{t+1}^c = \frac{\sum_{k=1}^K N_k \cdot \omega_k \cdot \theta_{t+1}^k}{\sum_{j=1}^K N_j \cdot \omega_j}, \tag{16}$$

where $\omega_k = \frac{1}{|\mathcal{S}_k|} \sum_{i \in \mathcal{S}_k} \alpha_i^k$, with $\mathcal{S}_k$ as the set of key multimodal attention weights for client $k$, and $\alpha_i^k$ as the attention weight of the $i$-th critical feature.

3) **Global-Local Attention Synergy** The training follows a "local attention refinement + global attention aggregation" paradigm: - Local stage: Each client uses self-attention to refine single-modal features and hierarchical co-attention to capture cross-modal dependencies adapted to its private data distribution. - Global stage: The server aggregates cross-client attention module parameters, preserving universal attention weights for core fake news features while integrating client-specific fine-grained patterns. This synergy ensures the global model retains both cross-platform generalization and precise multimodal feature discrimination.
4) **Attention Module Integrity Preservation in FL Loop** Unlike naive FL that may degrade complex local mod-

els, Fed-MultiFND fully preserves the hierarchical co-attention mechanism during local updates and global aggregation. The loss function $\mathcal{L}^*$ jointly optimizes the attention module and FL compatibility:

$$\mathcal{L}^* = \mathcal{L}_{\text{MultiFND}} + \lambda \mathcal{L}_{\text{FL-consistency}}, \tag{17}$$

where $\mathcal{L}_{\text{MultiFND}}$ is the binary cross-entropy loss for fake news classification, and $\mathcal{L}_{\text{FL-consistency}}$ penalizes excessive divergence between local and global attention parameters.

## IV. Experiments

In this section, we conduct extensive experiments on two real-world datasets to verify the effectiveness of Fed-MultiFND by comparing it with two representative baselines and the Fed-MultiFND variants.

TABLE I
STATISTICS OF TWO DATASETS FOR EVALUATION

| Dataset | Split | Fake | Real | Total |
|---|---|---|---|---|
| FakeSV | Training | 1,225 | 1,225 | 2,450 |
| | Validation | 270 | 500 | 770 |
| | Test | 273 | 238 | 511 |
| FakeTT | Training | 480 | 480 | 960 |
| | Validation | 118 | 82 | 200 |
| | Test | 99 | 136 | 235 |

TABLE II
STATISTICS OF BALANCED CONFIGURATION

| Dataset | Client | Fake | Real | Total |
|---|---|---|---|---|
| FakeSV | Client(1-5) | 245 | 245 | 490 |
| FakeTT | Client(1-5) | 96 | 96 | 192 |

TABLE III
STATISTICS OF IMBALANCED CONFIGURATION

| Dataset | Client | Fake | Real | Total |
|---|---|---|---|---|
| FakeSV | Client1 | 98 | 392 | 490 |
| | Client2 | 172 | 318 | 490 |
| | Client3 | 245 | 245 | 490 |
| | Client4 | 318 | 172 | 490 |
| | Client5 | 392 | 98 | 490 |
| FakeTT | Client1 | 38 | 153 | 192 |
| | Client2 | 67 | 125 | 192 |
| | Client3 | 96 | 96 | 192 |
| | Client4 | 125 | 67 | 192 |
| | Client5 | 153 | 38 | 192 |

### A. Datasets

Two publicly available multimodal short-video fake news datasets are adopted for experimental validation, with details as follows:

- **FakeSV**: Collected from 2017 to 2022, contains 3,624 text/audio/video multimodal short-video samples. Its training set has balanced fake/real labels, while the validation and test sets are intentionally imbalanced to simulate realistic cross-platform data distributions.
- **FakeTT**: Spanning 2019 to 2024, contains 1,991 multimodal short-video samples with pronounced cross-platform data heterogeneity. It follows the same training/validation/test split strategy as FakeSV to ensure consistent and fair evaluation.

To simulate realistic cross-platform scenarios and support federated learning experiments, we design three data partitioning schemes:

- **Table I**: Full split statistics of the complete FakeSV and FakeTT datasets, serving as the baseline;
- **Table II**: Ideal balanced setting, with training data evenly distributed across 5 clients, each holding a strict 1:1 fake-to-real news ratio;
- **Table III**: Imbalanced non-IID setting mimicking cross-platform data differences, with systematic perturbations applied to data partitioning, resulting in fake-to-real ratios of (0.2, 0.35, 0.5, 0.65, 0.8) across the 5 clients.

The validation and test sets remained unchanged across all schemes to ensure the fairness of model evaluation.

### B. Baseline

Two types of comparative experiments are designed to validate the effectiveness of the proposed method: first, ablation studies on the base model under centralized settings to examine the performance of each base module; second, performance comparisons between the federated learning-enhanced model and relevant baseline methods.

To ensure fair comparisons, we select 8 representative multimodal fake news detection methods covering traditional handcrafted feature-based, neural network-based, and large-scale pre-trained models, with detailed descriptions as follows.

- **HCFC-Hou** [29]: Linguistic, acoustic and engagement features with SVM for fake news classification.
- **HCFC-Media** [19]: TF-IDF features from video text + SVM for social media misinformation detection.
- **TikTie** [6]: Fuses speech-guided visual and MFCC-guided speech text via co-attention for short-video fake news detection.
- **CAFE** [10]: Ambiguity-aware multimodal fusion with auxiliary unimodal tasks to reduce modal inconsistency misclassifications.
- **FANVM** [27]: Fuses keyframe visual and text features via topic modeling and adversarial networks for topic-agnostic fake news verification.
- **FakingRecipe** [4]: Process-aware dual-branch framework (MSAM for semantic/sentiment clues, MEAM for spatiotemporal editing patterns) with late fusion for short-video fake news detection.
- **GPT-4**: Closed-source large multimodal model, enabling zero-shot fake news discrimination via prompt engineering on video text.
- **Qwen-VL**: Open-source Chinese visual-language model with strong cross-modal alignment and scalability.

TABLE IV
PERFORMANCE COMPARISON UNDER CENTRALIZED LEARNING

| Dataset | Model | Accuracy | F1-score | Precision | Recall |
|---|---|---|---|---|---|
| FakeSV | HCFC-Hou | 74.91 | 73.61 | 75.58 | 73.30 |
| | HCFC-Media | 76.38 | 75.83 | 76.14 | 75.66 |
| | TikTie | 73.43 | 73.26 | 73.23 | 75.33 |
| | CAFE | 78.78 | 78.30 | 78.12 | 78.60 |
| | FANVM | 79.52 | 78.81 | 79.17 | 78.46 |
| | GPT-4 | 67.43 | 67.34 | 70.76 | 75.67 |
| | Qwen-VL | 74.17 | 76.35 | 78.47 | 74.34 |
| | FakingRecipe | 82.29 | 82.70 | **83.06** | 82.29 |
| | Ours | **83.06** | **83.06** | **83.06** | **83.06** |
| FakeTT | HCFC-Hou | 73.24 | 72.00 | 71.99 | 74.65 |
| | HCFC-Media | 66.24 | 62.23 | 65.58 | 66.84 |
| | TikTie | 62.52 | 65.08 | 65.84 | 67.87 |
| | CAFE | 62.87 | 62.46 | 65.32 | 66.89 |
| | FANVM | 71.57 | 70.21 | 70.22 | 72.63 |
| | GPT-4 | 61.45 | 60.66 | 63.28 | 65.31 |
| | Qwen-VL | 58.86 | 73.89 | 64.21 | 87.00 |
| | FakingRecipe | 77.45 | 76.77 | 76.91 | 76.67 |
| | Ours | **78.72** | **78.58** | **78.69** | **79.42** |

### C. Performance Comparison

Our experiments include three parts: performance evaluation of the centralized base model, and two comparative experiments of the federated learning (FL)-integrated model under balanced and imbalanced data settings. We simulate 5 independent clients to mimic cross-platform data isolation, and compare the proposed Fed-MultiFND and federated FakingRecipe with individual client training and centralized full-dataset training. Data statistics are detailed in Tables II and III.

*1)* ***Centralized Learning Evaluation****:* To establish a baseline for FL experiments, we train 8 representative baselines (handcrafted feature methods, multimodal models, vision-language models) and our MultiFND on FakeSV and FakeTT under identical settings (results in Table IV). Key findings: 1. Lightweight multimodal models and general VLMs have limited performance and robustness, with obvious degradation on FakeTT. 2. Specialized cross-modal models perform better, but still suffer from cross-dataset generalization issues. 3. Our MultiFND outperforms all baselines including FakingRecipe on both datasets, with stable superior performance via hierarchical attention fusion, laying a solid foundation for FL deployment.

*2)* ***Federated Learning Evaluation****:* We evaluate federated learning on 5 platform-simulating clients under balanced and imbalanced data configurations (data partitioning in Tables II and III), with validation/test sets consistent with the centralized setting.

Results in Tables V and VI show federated learning consistently boosts performance: compared with isolated client training, federated models achieve higher accuracy and lower performance variance, proving global aggregation mitigates single-platform learning limitations.

Federated models match centralized training performance under the balanced setting, and deliver even more prominent advantages in the imbalanced scenario (where local models degrade severely from skewed data) by maintaining stable

TABLE V
PERFORMANCE COMPARISON UNDER FEDERATED LEARNING BALANCED CONFIGURATION

| Dataset | Method | Client | Accuracy | F1-score | Precision | Recall |
|---|---|---|---|---|---|---|
| FakeSV | FakingRecipe | Base Model | 82.29 | 82.70 | 83.06 | 82.29 |
| | | Client1 | 79.83 | 79.53 | 81.72 | 79.84 |
| | | Client2 | 73.74 | 73.50 | 74.62 | 73.74 |
| | | Client3 | 74.16 | 74.02 | 74.69 | 74.16 |
| | | Client4 | 78.15 | 78.08 | 78.55 | 78.15 |
| | | Client5 | 75.63 | 75.61 | 75.72 | 75.63 |
| | | Fed | 78.67 | 78.67 | 79.16 | 79.09 |
| | Ours | Base Model | 83.06 | 83.06 | 83.06 | 83.06 |
| | | Client1 | 74.17 | 74.12 | 75.26 | 74.80 |
| | | Client2 | 72.80 | 72.76 | 72.79 | 72.90 |
| | | Client3 | 74.95 | 74.64 | 75.02 | 74.54 |
| | | Client4 | 70.25 | 70.21 | 71.18 | 70.84 |
| | | Client5 | 70.84 | 70.66 | 72.77 | 71.71 |
| | | Fed | **79.26** | **79.25** | **79.44** | **79.53** |
| FakeTT | FakingRecipe | Base Model | 77.45 | 76.77 | 76.91 | 76.67 |
| | | Client1 | 74.75 | 74.62 | 75.25 | 74.75 |
| | | Client2 | 72.22 | 72.06 | 72.75 | 72.23 |
| | | Client3 | 76.26 | 76.09 | 77.06 | 76.26 |
| | | Client4 | 68.69 | 68.53 | 69.06 | 68.69 |
| | | Client5 | 75.25 | 75.07 | 76.02 | 75.26 |
| | | Fed | 76.87 | 75.78 | **80.04** | 75.09 |
| | Ours | Base Model | 78.72 | 78.58 | 78.69 | 79.42 |
| | | Client1 | 67.23 | 67.23 | 68.71 | 68.81 |
| | | Client2 | 67.76 | 67.59 | 70.96 | 70.14 |
| | | Client3 | 70.21 | 70.20 | 71.40 | 71.65 |
| | | Client4 | 68.09 | 68.00 | 68.58 | 68.99 |
| | | Client5 | 70.64 | 70.15 | 70.06 | 70.37 |
| | | Fed | **77.45** | **77.21** | 77.14 | **77.77** |

TABLE VI
PERFORMANCE COMPARISON UNDER FEDERATED LEARNING IMBALANCED CONFIGURATION

| Dataset | Method | Client | Accuracy | F1-score | Precision | Recall |
|---|---|---|---|---|---|---|
| FakeSV | FakingRecipe | Base Model | 82.29 | 82.70 | 83.06 | 82.29 |
| | | Client1 | 73.78 | 73.65 | 75.56 | 74.54 |
| | | Client2 | 77.89 | 77.89 | 78.37 | 78.31 |
| | | Client3 | 77.89 | 76.97 | 80.04 | 76.88 |
| | | Client4 | 76.91 | 76.31 | 77.77 | 76.18 |
| | | Client5 | 60.67 | 51.10 | 73.56 | 57.91 |
| | | Fed | 82.19 | 80.89 | 82.72 | 81.69 |
| | Ours | Base Model | 83.06 | 83.06 | 83.06 | 83.06 |
| | | Client1 | 80.04 | 80.04 | 80.45 | 80.43 |
| | | Client2 | 77.50 | 77.49 | 77.98 | 77.91 |
| | | Client3 | 69.86 | 69.51 | 72.66 | 70.93 |
| | | Client4 | 74.95 | 74.74 | 77.57 | 75.94 |
| | | Client5 | 77.10 | 76.93 | 77.04 | 76.87 |
| | | Fed | **82.78** | **82.77** | **82.85** | **82.99** |
| FakeTT | FakingRecipe | Base Model | 77.45 | 76.77 | 76.91 | 76.67 |
| | | Client1 | 73.62 | 68.46 | 82.09 | 68.96 |
| | | Client2 | 60.00 | 59.94 | 60.71 | 61.09 |
| | | Client3 | 74.04 | 73.27 | 73.48 | 73.18 |
| | | Client4 | 77.87 | 77.76 | 77.98 | 78.69 |
| | | Client5 | 53.19 | 50.04 | 64.84 | 58.60 |
| | | Fed | 75.32 | 74.54 | 74.72 | 74.42 |
| | Ours | Base Model | 78.72 | 78.58 | 70.94 | 83.84 |
| | | Client1 | 67.23 | 62.69 | 68.33 | 63.31 |
| | | Client2 | 65.11 | 62.93 | 64.03 | 62.85 |
| | | Client3 | 61.70 | 61.70 | 63.60 | 63.48 |
| | | Client4 | 58.72 | 57.62 | 66.27 | 62.83 |
| | | Client5 | 63.83 | 63.38 | 69.63 | 67.24 |
| | | Fed | **76.17** | **75.42** | **75.60** | **75.29** |

performance. These results verify the robustness and generalization of federated learning in real cross-platform scenarios.

*3) **Case Study**:* In centralized training, our MultiFND achieves state-of-the-art performance, outperforming FakingRecipe and other baselines with strong multimodal fusion and noise resistance. FakingRecipe serves as a strong competitive benchmark.

Further experiments validate federated learning's effectiveness for cross-platform fake news detection: both Fed-MultiFND and federated FakingRecipe outperform most individual client models, with performance close to centralized training, confirming federated learning can break cross-platform data silos without privacy violations.

Notably, Fed-MultiFND consistently outperforms federated FakingRecipe in both balanced and imbalanced settings, benefiting from the integration of hierarchical multimodal attention and attention-aware aggregation for stronger federated robustness. Overall, this case study validates that Fed-MultiFND offers a more reliable and effective solution for real-world cross-platform short video fake news detection.

## V. Conclusion

From a practical cross-platform standpoint, we innovatively integrate federated learning into multimodal models to address inter-platform data cannot be shared for training. We propose Fed-MultiFND, which tackles the dual challenges of multimodal heterogeneity and cross-platform data isolation. Furthermore, to simulate the potential data imbalance issues induced by cross-platform data, we construct two representative scenarios and compare them with the currently best performing model. Validated on FakeSV and FakeTT datasets, the framework outperforms individual client training under both balanced and imbalanced settings, delivering a privacy-compliant solution for cross-platform short video fake news detection.

## References


[1] Z. Jin, J. Cao, E. Guo, Y. Zhang, and J. Luo,"Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs,"in *Proc. International Conference on Multimedia Retrieval (ICMR)*, 2017, pp. 218–225.

[2] X. Zhou and R. Zafarani,"A survey of fake news: Fundamental theories, detection methods, and opportunities,"*ACM Computing Surveys*, vol. 53, no. 5, 2020.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin,"Attention Is All You Need,"in *Proc. NeurIPS*, 2017, pp. 5998–6008.

[4] Y. Bu, Q. Sheng, J. Gao , P. Qi, D. Wang, and J. Li,"FakingRecipe: Detecting Fake News on Short Video Platforms from the Perspective of Creative Process,"in *Proc. ACM Int. Conf. Multimedia*, 2024.

[5] J. Hendrickx, "From Newspapers to TikTok: Social Media Journalism as the Fourth Wave of News Production, Diffusion and Consumption," in *Blurring Boundaries of Journalism in Digital Media*, Cham: Springer, 2023, pp. 229–246.

[6] L. Shang, Z. Kou, Y. Zhang, and D. Wang, "A Multimodal Misinformation Detector for COVID-19 Short Videos on TikTok," in *Proc. IEEE Int. Conf. Big Data*, 2021, pp. 899–908.

[7] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, P. Askell, et al.,"Learning Transferable Visual Models from Natural Language Supervision,"in *Proc. ICML*, 2021.

[8] G. Bertasius, H. Wang, and L. Torresani,"Is Space-Time Attention All You Need for Video Understanding?",in *Proc. ICML*, 2021, pp. 813–824.

[9] T. Li, A. K. Sahu, M. Sanjabi, A. Talwalkar, and V. Smith,"Federated Optimization in Heterogeneous Networks,"in *Proc. MLSys*, 2020, pp. 429–438.

[10] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, and L. Shang, "Cross-modal Ambiguity Learning for Multimodal Fake News Detection," in *Proc. ACM Web Conf. (WWW)*, 2022, pp. 2897–2905.

[11] H.B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Aguera y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017, pp. 1273–1282.

[12] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao,"EANN: Event-Adversarial Neural Networks for Multi-Modal Fake News Detection,"in *Proc. AAAI Conf. Artificial Intelligence*, 2018.

[13] Y. Yang, X. Shi, H. Li, B. Fan, and Y. Xu, "Fake News Detection in Short Videos by Integrating Semantic Credibility and Multi-Granularity Contrastive Learning," *Applied Sciences*, 2025.

[14] J. Lu, D. Batra, D. Parikh, and S. Lee,"ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks,"in *Proc. NeurIPS*, 2019, pp. 13–23.

[15] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou,"FedBN: Federated Learning on Non-IID Features via Local Batch Normalization,"in *Proc. ICLR*, 2021.

[16] R. C. Geyer, T. Klein, and M. Nabi,"Differentially Private Federated Learning: A Client Level Perspective,"in *Proc. ICML*, 2017, pp. 2953–2962.

[17] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra,"Federated Learning with Non-IID Data,"in *Proc. ICLR Workshops*, 2018.

[18] K. Shu, A. Bhattacharjee, F. Altawi, T. Nazer, K. Ding, M. Karami, and H. Liu,"Combating Disinformation in a Social Media Age,"*ACM Comput. Surv.*, 2020.

[19] J. C. M. Serrano, O. Papakyriakopoulos, and S. Hegelich, "NLP-based Feature Extraction for the Detection of COVID-19 Misinformation Videos on YouTube," in *Proc. Workshop on NLP for COVID-19 at ACL*, 2020.

[20] P. Qi, Y. Bu, J. Cao, W. Ji, R. Shui, J.Xiao, D. Wang, T. Chua, "FakeSV: A Multimodal Benchmark with Rich Social Context for Fake News Detection on Short Video Platforms," *arXiv preprint*, 2022.

[21] S. Li, C. W. Q. Ruth, H. Xu, and F. Liu, "Multimodal Learning for Fake News Detection in Short Videos Using Linguistically Verified Data and Heterogeneous Modality Fusion," *arXiv preprint*, 2025.

[22] Y. Wang, L. Chen, Z. Qian, and P. Li, "Official-NV: An LLM-Generated News Video Dataset for Multimodal Fake News Detection," *Technical Report*, 2024.

[23] Y. Zhu, Y. Wang, and Z. Yu, "Multimodal Fake News Detection: MFND Dataset and Shallow-Deep Multitask Learning," *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2025.

[24] K. Shu, S. Wang, and H. Liu,"Beyond News Contents: The Role of Social Context for Fake News Detection,"in *Proc. ACM SIGKDD*, 2019, pp. 312–320.

[25] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, et al., "Advances and Open Problems in Federated Learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[26] R. Jadhav, V. Meshram, A. Bhosle, K. Patil, S. Dash, and S., "Explainable Multilingual and Multimodal Fake-News Detection: Toward Robust and Trustworthy AI for Combating Misinformation," *Frontiers in AI*, 2025.

[27] H. Choi and Y. Ko, "Using Topic Modeling and Adversarial Neural Networks for Fake News Video Detection," in *Proc. ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2021, pp. 2950–2954.

[28] A. Heidari, N. J. Navimipour, H. Dag, S. Talebi, and M. Unal, "A Novel Blockchain-Based Deepfake Detection Method Using Federated and Deep Learning Models," *Cognitive Computation*, 2024.

[29] R. Hou, V. Perez-Rosas, S. Loeb, and R. Mihalcea, "Towards Automatic Detection of Misinformation in Online Medical Videos," in *Proc. Int. Conf. Multimodal Interaction (ICMI)*, 2019, pp. 235–243.

[30] H. Tan and M. Bansal,"Lxmert: Learning Cross-Modality Encoder Representations from Transformers,"in *Proc. EMNLP*, 2019, pp. 5100–5111.

[31] J. Wang, J. Liu, N. Zhang, and Y. Wang, "Consistency-aware Fake Videos Detection on Short Video Platforms," *arXiv preprint*, 2025.