

Unlocking Multi-Site Clinical Data: A Federated Approach to Privacy-First Child Autism Behavior Analysis

Guangyu Sun¹, Wenhan Wu³, Zhishuai Guo⁴, Ziteng Wang⁵, Pegah Khosravi^{1,2}, Chen Chen¹

¹ Institute of AI, University of Central Florida

² Department of Clinical Sciences, College of Medicine, University of Central Florida

³ University of North Carolina at Charlotte

⁴ Department of Computer Science, Northern Illinois University

⁵ Industrial and Systems Engineering, Northern Illinois University

guangyu@ucf.edu, wwu25@charlotte.edu, zguo@niu.edu, zwang3@niu.edu,
{pegah.khosravi, chen.chen}@ucf.edu

Abstract

Automated recognition of autistic behaviors in children is essential for early intervention and objective clinical assessment. However, the development of robust models is severely hindered by strict privacy regulations (e.g., HIPAA) and the sensitive nature of pediatric data, which prevents the centralized aggregation of clinical datasets. Furthermore, individual clinical sites often suffer from data scarcity, making it difficult to learn generalized behavior patterns or tailor models to site-specific patient distributions. To address these challenges, we observe that Federated Learning (FL) can decouple model training from raw data access, enabling multi-site collaboration while maintaining strict data residency. In this paper, we present the first study exploring Federated Learning for pose-based child autism behavior recognition. Our framework employs a two-layer privacy protection mechanism: utilizing human skeletal abstraction to remove identifiable visual information from the raw RGB videos and FL to ensure sensitive pose data remains within the clinic. This approach leverages distributed clinical data to learn generalized representations while providing the flexibility for site-specific personalization. Experimental results on the MMASD benchmark demonstrate that our framework achieves high recognition accuracy, outperforming traditional federated baselines and providing a robust, privacy-first solution for multi-site clinical analysis.

1. Introduction

Automated recognition of autism-related behaviors in children has emerged as a transformative tool for early screening, clinical diagnosis, and longitudinal monitoring [3, 5,

17, 19, 20]. Traditionally, the clinical assessment of Autism Spectrum Disorder (ASD) has relied on standardized observational instruments such as the Autism Diagnostic Observation Schedule (ADOS) and the Modified Checklist for Autism in Toddlers (M-CHAT) [3]. While these tools are clinically validated, they are inherently limited by human observer bias, significant inter-rater variability, and the substantial resource burden required for administration. Expert clinicians must often observe hours of behavioral interactions to identify subtle stereotypical motor movements or social-communicative deficits. By leveraging computer vision to quantify repetitive or atypical movements, automated systems can identify “digital biomarkers” that provide clinicians with objective, high-frequency measurements, effectively complementing traditional assessments with data-driven evidence.

Despite the promise of deep learning in this domain, its large-scale deployment is severely restricted by the sensitive nature of pediatric healthcare data. Under strict regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the US and the General Data Protection Regulation (GDPR) in the EU, raw RGB video recordings of children are considered primary biometric identifiers. These recordings contain sensitive Protected Health Information (PHI) and cannot be easily shared across institutions or aggregated in centralized cloud servers due to the risk of identity disclosure [22]. Furthermore, the Family Educational Rights and Privacy Act (FERPA) often applies when behavioral data is collected within educational settings, creating a complex multi-jurisdictional compliance burden for multi-site research. While skeleton-based representations have been proposed as a privacy-preserving alternative [9], motion-based biometric “fingerprints” still pose significant re-identification risks when large datasets

are centralized. Consequently, clinical data typically remains isolated in “data silos”, leading to small-scale studies that fail to capture the full spectrum of behavioral diversity across different demographics, age groups, and severity levels.

To overcome this obstacle, we propose a robust, privacy-preserving framework for pose-based autism behavior recognition using Federated Learning (FL) [15]. Our approach shifts the training paradigm from the traditional “data-to-model” architecture to a decentralized “model-to-data” paradigm. This enables collaborative training across distributed clinical sites without ever transferring raw patient information or sensitive pose sequences. To ensure high recognition accuracy in this decentralized setting, we adopt **FreqMixFormer** [26], a state-of-the-art skeleton-based human action recognition approach. By leveraging frequency-domain analysis, FreqMixFormer can capture discriminative temporal motion patterns while maintaining low computational overhead, making it ideal for deployment on clinical edge nodes with varying hardware capabilities.

The core of our framework is a **two-layer privacy protection strategy**. First, we utilize 3D skeletal representations of human bodies rather than raw RGB video, filtering out identifiable facial features, clothing, and environmental contexts. Second, the federated optimization protocol ensures that even these abstracted pose sequences remain localized within each clinical site’s secure infrastructure. Only anonymized model parameters are shared with a central aggregator, providing a double-layer defense against both visual and biometric identity disclosure while strictly aligning with the “data minimization” and “privacy-by-design” principles of modern healthcare regulations.

A critical challenge in this decentralized setting is the inherent *heterogeneity* of multi-site clinical data. Variations in therapeutic protocols, room layouts, camera specifications, and patient demographics introduce significant domain shifts (Non-IID data) that standard federated algorithms often fail to resolve. To address this, we conduct a systematic investigation of various **Personalized Federated Learning (PFL)** strategies, including FedBN [12], FedPer [2], and Adaptive Personalized Federated Learning (APFL) [6]. Our framework evaluates how different personalization layers—ranging from local batch normalization to adaptive model mixing—can bridge the gap between global collaborative knowledge and site-specific behavioral nuances.

To validate our framework, we establish a simulated multi-site benchmark using the MMASD dataset [9]. Through extensive experiments across multiple clinical themes, including Robotic-assisted therapy, Rhythm-based activities, and Yoga-based poses, we demonstrate that while standard aggregation methods provide a collaborative base-

line, adaptive personalization is essential for handling severe clinical heterogeneity. Our results suggest that APFL, in particular, offers a superior balance of global representation and local adaptation, achieving performance comparable to centralized training while strictly maintaining data residency.

The primary contributions of this work are summarized as follows:

- We present the first comprehensive framework for pose-based child autism behavior recognition that integrates a **two-layer privacy strategy**, combining skeletal abstraction with decentralized federated optimization to comply with HIPAA/GDPR standards.
- We utilize the efficient **FreqMixFormer** architecture, incorporating frequency-aware attention to ensure robust behavioral analysis across distributed clinical nodes.
- We provide a **systematic investigation of personalized federated learning strategies** (FedBN, FedPer, APFL) to address cross-site clinical heterogeneity and domain shift in autism clinical data.
- We establish a **federated benchmark protocol** on the MMASD benchmark, demonstrating that **adaptive personalization** via APFL provides a robust pathway for multi-site clinical AI without compromising patient anonymity.

2. Related Work

2.1. Pose-based Autism Behavior Recognition

Automated assessment of Autism Spectrum Disorder (ASD) has transitioned from qualitative observations to quantitative metrics using advanced computer vision techniques. Early studies primarily focused on facial expression analysis and repetitive behavior detection (e.g., hand flapping, head banging) using raw RGB video [3, 17]. However, the use of raw video poses significant privacy risks for pediatric subjects, especially in rare disease contexts where background cues can lead to re-identification. To address this, skeleton-based action recognition has emerged as a robust, privacy-preserving alternative. Early graph-based methods such as ST-GCN [28] and AS-GCN [24] established the foundation for modeling spatiotemporal joint dynamics by representing the human body as a structured graph. More recently, Shift-GCN [4] and MS-G3D [13] have further improved recognition accuracy by introducing multi-scale aggregation and efficient shift operations.

These models have been adapted for ASD stimming behavior detection, capturing complex motion patterns without relying on identifiable appearance information [19, 20]. The introduction of the MMASD dataset [9] and its enhanced version MMASD+ [18] has provided the community with multimodal benchmarks (including 3D skeletons, optical flow, and body mesh) specifically tailored for in-

intervention analysis. Despite these advances, most existing models rely on centralized data collection, which is often prohibited in multi-site clinical research due to institutional data-sharing restrictions and regulatory hurdles.

2.2. Federated Learning in Clinical Settings

Federated Learning (FL) [15] offers a privacy-preserving paradigm that allows multiple institutions to collaboratively train a global model while keeping their raw data localized. In the healthcare domain, FL has been widely adopted to facilitate multi-institutional collaborations in medical imaging and genomics without compromising patient privacy [22, 27]. For ASD specifically, Shamseddine et al. [21] demonstrated the feasibility of FL for detecting autism using behavioral traits and facial features. Recent works have extended this decentralized approach to serious games [16] and smart assistant systems for behavioral monitoring [8].

A significant challenge in multi-site clinical FL is the risk of “gradient conflict”, where divergent data distributions across clinics (e.g., variations in age groups or therapeutic activities) pull the global model in opposing directions. Advanced aggregation algorithms, such as FedAP [14], have been proposed to mitigate these conflicts by incorporating adaptive batch normalization or gradient coordination. Our work builds on these foundations by proposing a decentralized framework that utilizes skeletal representations to provide a double layer of privacy protection tailored for autism clinical research.

2.3. Personalized Federated Learning

The primary bottleneck in clinical FL is the heterogeneity of data distributions across sites, often referred to as the Non-IID (not identically and independently distributed) problem. Standard FL algorithms like FedAvg [15] and the regularization-based FedProx [10] often struggle to generalize when site-specific domain shifts are severe. Personalized Federated Learning (PFL) [25] addresses this by allowing sites to maintain local model variants that adapt the global representation to their unique local distribution.

Seminal PFL approaches include Per-FedAvg [7], which utilizes meta-learning for rapid local adaptation, and Ditto [11], which regularizes local models toward the global representation while allowing for significant local specialization. Structural personalization techniques, such as FedBN [12] and FedPer [2], focus on keeping specific layers (e.g., batch normalization or classifier heads) local to capture site-specific statistics. In this work, we specifically leverage the Adaptive Personalized Federated Learning (APFL) algorithm [6], which adaptively learns the optimal mixing ratio between global and local knowledge. This ensures that our autism behavior recognition model can capture universal behavioral markers while remaining sensitive to the unique phenotypic expressions encountered at individual clinical

sites.

3. Methodology

In this section, we detail our proposed framework for privacy-preserving autism behavior recognition. Our approach integrates a two-layer privacy protection mechanism that combines skeletal abstraction with decentralized optimization. To ensure both high recognition fidelity and computational efficiency, we adopt the **FreqMixFormer** [26] architecture as our core behavioral analysis backbone. Furthermore, we investigate several Personalized Federated Learning (PFL) strategies to handle the inherent heterogeneity in multi-site clinical data.

3.1. Problem Formulation

We consider a decentralized clinical network comprising N participating sites (clients). Each site $i \in \{1, \dots, N\}$ maintains a private dataset $\mathcal{D}_i = \{(x_{i,j}, y_{i,j})\}_{j=1}^{n_i}$, where $x_{i,j}$ denotes an abstracted behavioral sequence and $y_{i,j} \in \{1, \dots, C\}$ is the corresponding clinical label. Due to the sensitive nature of pediatric data, raw sequences $x_{i,j}$ are restricted to local storage within the site’s secure infrastructure. Our objective is to learn a set of personalized models $\{v_i\}_{i=1}^N$ that minimize the collective empirical risk:

$$\min_{\{v_1, \dots, v_N\}} \sum_{i=1}^N \frac{n_i}{n} F_i(v_i), \quad F_i(v_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} \ell(v_i; x_{i,j}, y_{i,j}), \quad (1)$$

where $n = \sum n_i$ and v_i represents the model deployed at site i . Depending on the personalization strategy, v_i may be a site-specific variant of a shared global model w .

3.2. Skeletal Abstraction for Identity Protection

The primary challenge in analyzing pediatric autism behaviors lies in the trade-off between clinical utility and patient privacy. Raw RGB video recordings are highly identifiable, containing facial features, home environments, and other sensitive visual cues that constitute Protected Health Information (PHI). To mitigate this risk, our framework adopts 3D skeletal representations $S \in \mathbb{R}^{T \times V \times 3}$ as the foundational data modality, where T denotes temporal length and V denotes the number of joint keypoints.

By abstracting behavioral sequences into 3D joint trajectories, we decouple critical motion patterns from identifiable appearance information. This transformation serves as the **first layer of privacy**, ensuring that identifiable patient characteristics are filtered out before any model training or collaborative optimization occurs. Furthermore, skeletal data is naturally invariant to environmental factors such as illumination, background clutter, and camera-specific color distributions, which often vary across clinical rooms. In this work, we leverage 3D joint coordinates from the

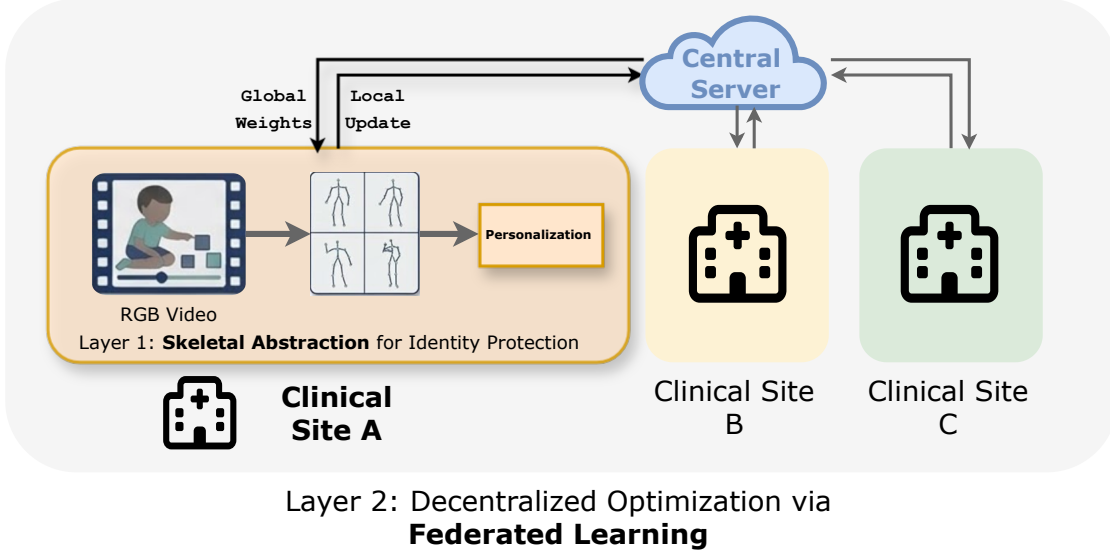


Figure 1. **Overview of the proposed Two-Layer Privacy framework.** The first layer achieves privacy via skeletal abstraction, filtering out raw biometric identifiers from videos. This is done at each clinical site for their local video data (a demonstration is shown for ‘Clinical Site A’ in this figure). The second layer employs decentralized optimization (FL) using the efficient FreqMixFormer backbone to maintain data residency. Adaptive personalization layers (FedBN, FedPer, or APFL) are evaluated to handle clinical heterogeneity by adaptively mixing global knowledge with local specialization.

MMASD benchmark [9], which features 71 keypoints (including SMPL, extra, and H36M joints). A visualization is presented in Figure 2. Note that our framework remains compatible with other advanced world-grounded pose estimation models like GVHMR [23] for enhanced temporal fidelity.

3.3. Action Recognition Backbone: FreqMixFormer

To perform robust behavior analysis on the extracted skeletons, we employ **FreqMixFormer** [26] as our model backbone. Unlike traditional GCN-based methods that primarily operate in the spatial domain, FreqMixFormer utilizes a frequency-aware mixed transformer architecture that is specifically optimized for skeletal action recognition.

The model incorporates a frequency-aware attention module that leverages the Discrete Cosine Transform (DCT) [1] to capture discriminative motion patterns in the frequency domain. By analyzing the frequency components of joint trajectories, the framework effectively distinguishes between similar repetitive behaviors, such as subtle rhythmic hand movements versus stereotypical hand flapping. FreqMixFormer employs a mixed transformer architecture that balances global temporal dependencies with local spatial correlations while maintaining high computational and memory efficiency. This lightweight design is particularly advantageous for Federated Learning settings; by minimizing the parameter footprint, the model reduces the hardware overhead for on-device training at decentral-

ized clinical nodes and significantly lowers the communication costs associated with transmitting model weights or gradients across multiple FL rounds.

3.4. Decentralized Optimization via Federated Learning

Even with skeletal abstraction, the accumulation of large-scale clinical data at a single institution remains a significant regulatory and ethical bottleneck. We therefore employ Federated Learning (FL) as the **second layer of privacy**, shifting the training paradigm from centralized aggregation to decentralized optimization. This approach ensures that even anonymized pose sequences never leave the secure infrastructure of the participating clinical site, maintaining strict data residency and subject privacy.

The optimization proceeds in a server-client architecture. In each communication round t , a central server broadcasts the current global FreqMixFormer parameters w^t . Each clinical site i performs K local updates using Stochastic Gradient Descent (SGD) on its private data $\mathcal{D}_i$. To further mitigate system heterogeneity and stabilize training, we evaluate both standard FedAvg [15] and FedProx [10] aggregation schemes. Only the resulting model updates Δw_i^t are shared with the server, which aggregates them using a weighted averaging scheme:

$$w^{t+1} = \sum_{i=1}^N \frac{n_i}{n} (w^t + \Delta w_i^t). \quad (2)$$

This collaborative optimization enables individual sites to

benefit from a global representation of autistic behaviors without compromising institutional ownership of sensitive data.

3.5. Addressing Heterogeneity via Personalization

Clinical data is characterized by significant non-IID properties arising from diverse camera configurations, therapeutic protocols, and patient demographics. To resolve these domain shifts, we evaluate three prominent personalization strategies:

FedBN: Local Batch Normalization. FedBN [12] addresses feature-level heterogeneity by maintaining Batch Normalization (BN) layers locally at each site. While convolutional and linear weights are aggregated globally to learn shared behavioral representations, the local BN layers capture site-specific statistics (mean and variance), effectively performing a form of internal domain adaptation to the local clinical environment.

FedPer: Personalization Layers. FedPer [2] decomposes the model architecture into a shared backbone (base layers) and a local head (personalization layers). During the federated process, only the backbone parameters are aggregated, while each site optimizes its own classifier head tailored to its specific label distribution and behavioral nuances.

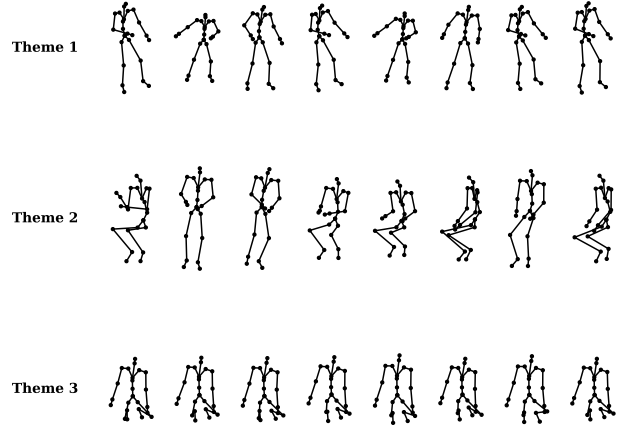
APFL: Adaptive Mixing. APFL [6] provides a more flexible approach by formulating the personalized model v_i as a weighted mixture of the global model w and a local-only model u_i : $v_i = \alpha_i u_i + (1 - \alpha_i)w$. The mixing parameter $\alpha_i \in [0, 1]$ is updated adaptively via gradient descent to minimize local validation loss:

$$\alpha_i^{t+1} = \alpha_i^t - \eta_\alpha \langle \nabla f_i(v_i), u_i - w \rangle. \quad (3)$$

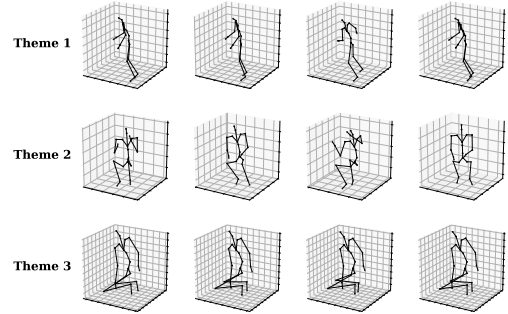
This mechanism allows sites to automatically determine the optimal balance between global knowledge and local specialization. As demonstrated in our experiments, this adaptive strategy consistently provides the best performance in heterogeneous clinical environments.

4. Experiments

In this section, we conduct a systematic evaluation of our framework to address two primary research questions. First, we investigate whether standard Federated Learning (FL) can effectively capture autistic behavioral patterns across heterogeneous sites (RQ1). Second, we evaluate various personalization strategies to determine the optimal balance between global knowledge and site-specific adaptation (RQ2).



(a) Temporal trajectories across different clinical themes.



(b) 3D joint representation of an anonymized sample.

Figure 2. Visualization of 3D skeletal data from the MMASD benchmark. (a) shows representative sequences for Robotic-assisted therapy, Rhythm-based activities, and Yoga-based poses. (b) illustrates the 3D joint structure. This abstraction preserves kinetic motion for behavior analysis while raw biometric identifiers are removed.

4.1. Experimental Setup

Dataset: MMASD Benchmark. We evaluate our framework using the MultiModal ASD (MMASD) benchmark [9], a privacy-preserving dataset specifically designed for autism intervention analysis. The dataset consists of 1,315 segmented samples derived from over 100 hours of intervention recordings involving 32 children diagnosed with ASD. To facilitate robust behavior analysis, MMASD provides four primary modalities; our framework specifically utilizes the **3D skeleton** data, which features 71 body key-points extracted via the ROMP algorithm. Figure 2 illustrates representative 3D skeletal samples from the benchmark, demonstrating how behavioral motion is captured while filtering out identifiable visual appearance information.

Clinical Themes and Multi-site Partitioning. The behavioral samples in MMASD are categorized into three dis-

tinct clinical themes reflecting different therapeutic activities: (1) *Robotic-assisted therapy*, comprising ‘Arm Swing’, ‘Body Swing’, ‘Chest Expansion’, and ‘Squat’; (2) *Rhythm-based activities*, including ‘Drumming’, ‘Maracas Forward Shaking’, ‘Maracas Shaking’, and ‘Sing and Clap’; and (3) *Yoga-based poses*, such as ‘Frog Pose’, ‘Tree Pose’, and ‘Twist Pose’.

To simulate a distributed clinical learning environment, we adopt a **cross-silo Federated Learning** configuration. In cross-silo FL, a small number of trusted institutions (e.g., hospitals, clinics, or research centers) collaboratively train models while keeping patient data localized. Unlike cross-device FL, which typically involves thousands or millions of unpredictable edge devices, cross-silo settings usually consist of only a handful of participating sites with relatively larger datasets and stable computational resources. This setting is particularly relevant for clinical research, where collaboration typically occurs among a few verified medical institutions. Following this paradigm, we partition the dataset into *three simulated clients*, each corresponding to one clinical theme. This design reflects realistic healthcare collaborations in which specialized institutions often focus on distinct therapeutic protocols or patient cohorts. Consequently, each client exhibits a unique data distribution, introducing meaningful domain shifts across sites. By treating these theme-specific cohorts as individual silos, our setup captures the key challenge of cross-silo federated learning: leveraging complementary knowledge across a small number of heterogeneous clinical institutions while preserving strict data locality.

Model Backbone. All experiments utilize FreqMixFormer [26] as the shared action recognition backbone trained from scratch. Its frequency-aware attention and efficient parameterization allow for high-fidelity behavior analysis even on clinical edge devices with limited resources.

Implementation Details. We train our framework for a total of 30 communication rounds. In each round, every client performs 1 local epoch of training. We utilize Stochastic Gradient Descent (SGD) as the local optimizer with a learning rate of 0.1, a weight decay of 5×10^{-4} , and a momentum of 0.9. To ensure stable convergence in the early stages of decentralized optimization, we implement a warmup phase of 5 rounds for all federated and local baselines.

Baselines and Comparison Groups. We categorize the evaluated optimization strategies into three groups:

- **Baseline (Local):** A site-specific training scheme where models are trained independently using only local data. This represents the upper bound for purely local adaptation without collaborative knowledge.

- **Federated Learning (FL):** Standard decentralized optimization schemes, including FedAvg [15] and the regularization-based FedProx [10].
- **Personalized Federated Learning (PFL):** Methods designed to bridge the gap between global and local knowledge, including parameter-wise isolation (FedBN [12], FedPer [2]) and our primary adaptive strategy, APFL [6]. For APFL, the mixing parameter is initialized at $\alpha = 0.01$ to provide a strong global prior.

4.2. Performance Analysis

Effectiveness of Federated Learning (RQ1). Our comparative analysis reveals that traditional federated learning (FL) faces significant challenges in the clinical domain. As shown in Table 1 and Figure 3, standard FedAvg achieves an average accuracy of 70.30%, significantly lagging behind the 82.61% achieved by independent Local training. This gap is particularly severe in Theme 1, where FedAvg lags by nearly 17%. While FedProx improves robustness, it only manages to align with the local baseline rather than surpassing it. These findings confirm that simple model aggregation is often counterproductive for autism behavior recognition, as site-specific domain shifts (Non-IID data) mask universal behavioral markers. The results suggest that for autism research, a monolithic global model is insufficient to capture the full spectrum of behavioral diversity across heterogeneous clinical sites.

Optimization via Personalized Federated Learning (RQ2). To overcome clinical heterogeneity, we evaluate various PFL strategies. Parameter-wise isolation methods, such as FedBN (69.67%) and FedPer (76.71%), show limited efficacy. As illustrated in the convergence curves, these methods often exhibit unstable training dynamics and fail to consistently outperform the local baseline. This suggests that isolating specific layers like batch normalization or classifier heads is insufficient for capturing the complex spatiotemporal nuances of 3D skeletal data across diverse therapeutic themes.

The APFL algorithm emerges as the superior solution, achieving the highest average accuracy of 87.80%, which represents a significant improvement over the 82.61% average accuracy of the local baseline. While isolated local training provides a strong benchmark for site-specific performance, APFL consistently surpasses these results by adaptively learning the optimal mixing ratio between global representations and local specialization. This performance gain demonstrates the critical benefit of leveraging collective data knowledge from other clinical sites; by adaptively incorporating global information, APFL enables each clinical silo to benefit from behavioral markers learned across the entire network, effectively overcoming the limitations of purely local data. Unlike standard FL methods which may

Table 1. **Comparative performance summary across clinical themes.** We categorize methods into Baseline (Local), Standard FL, and Personalized FL (PFL). Accuracy results demonstrate that adaptive personalization (APFL) consistently outperforms traditional federated baselines and isolated local training.

	Method	Theme 1	Theme 2	Theme 3	Avg (%)
Baseline	Local	87.10	65.33	95.41	82.61
FL	FedAvg	70.16	52.67	88.07	70.30
	FedProx	79.03	70.00	98.17	82.40
Personalized FL	FedBN	66.13	78.67	64.22	69.67
	FedPer	63.71	74.67	91.74	76.71
	APFL	92.74	78.00	92.66	87.80

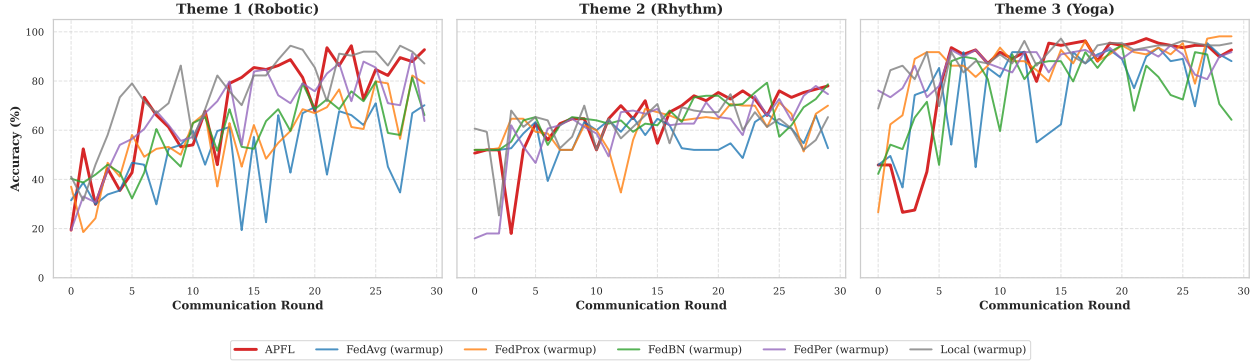


Figure 3. **Performance evolution across 30 communication rounds for different clinical themes.** The curves demonstrate the convergence characteristics of standard FL, parameter-wise PFL, and our adaptive personalization approach. APFL (red) consistently achieves superior stability and final recognition accuracy across all therapeutic domains.

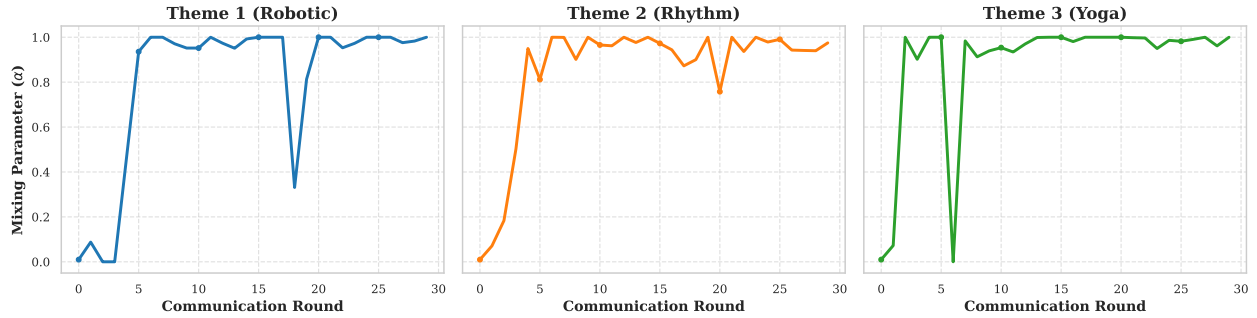


Figure 4. **Evolution of the adaptive mixing parameter α across different clinical themes.** The parameter is initialized with a low value and adaptively increases, indicating that the model initially prioritizes global knowledge synthesized from the collaborative network before gradually incorporating site-specific behavioral nuances from local data.

be degraded by clinical heterogeneity, our adaptive personalized approach recovers and enhances local performance, proving that collaboration across clinical silos is highly advantageous when appropriate personalization mechanisms are employed. Furthermore, APFL achieves a peak accuracy of 92.74% in Theme 1 and maintains stable convergence, as shown in Figure 3. This adaptive mechanism proves essential for multi-site clinical research, enabling the

framework to overcome the limitations of both monolithic global models and isolated local training.

5. Discussion

The experimental findings presented in Section 4 underscore several critical challenges and opportunities in the field of medical federated learning, particularly for pediatric autism research. In this section, we provide a deeper anal-

ysis of the trade-off between cross-site generalization and local specialization, the adaptive dynamics of the APFL algorithm, and the broader regulatory implications of our two-layer privacy framework.

Generalization vs. Personalization in ASD Recognition

A fundamental challenge in multi-site clinical AI is the tension between learning universal behavioral markers and adapting to local clinical nuances. The performance of our FreqMixFormer backbone under different optimization schemes suggests that autistic behaviors are highly context-dependent. The failure of traditional FedAvg to consistently outperform isolated local training indicates that monolithic “one-size-fits-all” models are ill-suited for the complex, non-IID distributions encountered in autism clinical data. This heterogeneity is likely driven by differences in clinical therapeutic protocols, individual patient support needs, and the high phenotypic variability inherent in Autism Spectrum Disorder.

Our observation that adaptive mixing (APFL) outperforms parameter-wise personalization methods (FedBN, FedPer) suggests that for high-dimensional temporal data like 3D skeletons, local adaptation must occur holistically. While methods like FedBN only adapt internal feature statistics, APFL allows the entire model representation to be adaptively weighted based on site-specific characteristics. This indicates that successful multi-site clinical AI requires frameworks that treat global collaborative knowledge as a robust prior that can be flexibly modulated by local behavioral evidence.

The Role of Adaptive Mixing in Clinical Heterogeneity

To further investigate the personalization mechanism, we analyze the evolution of the adaptive mixing parameter α throughout the federated training process. As illustrated in Figure 4, α starts with a relatively low value and gradually increases across all clinical themes. This trajectory provides two key insights. First, the low initial value indicates that the model relies heavily on the global representation in the early stages, effectively using the collaborative network to establish a stable behavioral foundation. Second, the adaptive increase toward 1 demonstrates that as the local model matures, the system automatically shifts its focus toward site-specific behavioral nuances. This confirms that the global model acts as a critical knowledge anchor that helps stabilize local learning, particularly for sites with limited data samples. The ability to automatically tune this balance makes APFL particularly robust for real-world clinical deployment, where the optimal degree of personalization may vary significantly across institutions.

Regulatory Implications and Visual Privacy Beyond performance metrics, the success of our Two-Layer Pri-

vacancy framework has significant implications for the future of clinical collaboration. By combining skeletal abstraction with decentralized optimization, we provide a mathematically and regulatorily sound pathway for building large-scale behavioral models.

The first layer of privacy—skeletal abstraction—addresses the primary biometric identifier (the human face) while the second layer—Federated Learning—addresses the administrative bottleneck of data residency. This dual-layer approach aligns strictly with HIPAA’s “privacy-by-design” and GDPR’s “data minimization” principles. By ensuring that raw identifiers are filtered at the source and that abstracted motion sequences never leave the secure clinical infrastructure, we address the “data silo” problem that has historically impeded progress in autism research. This framework establishes a standard for multi-site collaboration that respects patient autonomy while enabling the development of more powerful, objective diagnostic tools. Moving forward, the integration of such privacy-preserving technologies will be essential for democratizing access to high-quality clinical AI across diverse and under-resourced healthcare settings.

6. Conclusion

In this work, we introduced a robust, privacy-first framework for pose-based child autism behavior recognition. By integrating a **two-layer privacy protection strategy**—utilizing skeletal abstraction and cross-silo federated learning—we provide a secure pathway for collaboration among a trusted network of clinical institutions. Our systematic evaluation of the efficient **FreqMixFormer** architecture across various federated optimization schemes highlights the critical role of personalization in handling clinical data heterogeneity. We demonstrate that while traditional FL often struggles with non-IID behavioral data, adaptive personalization via **APFL** significantly outperforms isolated local training, proving the substantial benefit of leveraging shared knowledge across clinical sites while strictly maintaining data residency. Future research will explore the integration of *multimodal data streams* to further enhance diagnostic precision and behavioral insight. Specifically, incorporating *speech patterns*, *vocal prosody*, and the *conversational dynamics* between patients and clinicians or behavior analysts could provide deeper insights into social-communication nuances that are currently underserved by motion-only analysis. Furthermore, we aim to investigate the use of privacy-preserving large language models (LLMs) to analyze clinical notes and transcripts in a federated manner. Expanding the framework to handle such heterogeneous, high-dimensional data streams while maintaining rigorous privacy standards for sensitive audio-visual interactions remains a promising direction for the next generation of automated autism intervention and monitoring tools.

References

- [1] Nasir Ahmed, T. Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1): 90–93, 1974. 4
- [2] Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019. 2, 3, 5, 6
- [3] Tal Barami, Liora Manelis-Baram, Hadas Kaiser, Michal Ilan, Aviv Slobodkin, Ofri Hadashi, Dor Hadad, Danel Waisengreen, Tanya Nitzan, Idan Menashe, et al. Automated analysis of stereotypical movements in videos of children with autism spectrum disorder. *JAMA network open*, 7(9): e2432851, 2024. 1, 2
- [4] Ke Cheng, Yifan Zhang, Xiangyu He, Weihang Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *CVPR*, pages 183–192, 2020. 2
- [5] Andong Deng, Taojiannan Yang, Chen Chen, Qian Chen, Leslie Neely, and Sakiko Oyama. Language-assisted deep learning for autistic behaviors recognition. *Smart Health*, 32:100444, 2024. 1
- [6] Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020. 2, 3, 5, 6
- [7] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In *NeurIPS*, 2020. 3
- [8] Vinit Hegiste, Vidit Goyal, Tatjana Legler, and Martin Ruskowski. Federated action recognition for smart worker assistance using fastpose, 2025. 3
- [9] Jicheng Li, Vuthea Chheang, Pinar Kullu, Eli Brignac, Zhang Guo, Anjana Bhat, Kenneth E Barner, and Roghayeh Leila Barmaki. Mmasd: A multimodal dataset for autism intervention analysis. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 397–405, 2023. 1, 2, 4, 5
- [10] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems (MLSys)*, pages 429–450, 2020. 3, 4, 6
- [11] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *ICML*, pages 6357–6368, 2021. 3
- [12] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. In *ICLR*, 2021. 2, 3, 5, 6
- [13] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *CVPR*, pages 143–152, 2020. 2
- [14] Wang Lu, Jindong Wang, Yiqiang Chen, Xin Qin, Renjun Xu, Dimitrios Dimitriadis, and Tao Qin. Personalized federated learning with adaptive batchnorm for healthcare. *IEEE Transactions on Big Data*, 10(6):915–925, 2024. 3
- [15] 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 the 20th International Conference on Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017. 2, 3, 4, 6
- [16] Nikolaos Pavlidis, Vasileios Perifanis, Eleni Briola, Christos-Chrysanthos Nikolaidis, Eleftheria Katsiri, Pavlos S Efraimidis, and Despina Elisabeth Filippidou. Federated anomaly detection for early-stage diagnosis of autism spectrum disorders using serious game data. *arXiv preprint arXiv:2410.20003*, 2024. 3
- [17] Sam Perochon, J Matias Di Martino, Kimberly LH Carpenter, Scott Compton, Naomi Davis, Brian Eichner, Steven Espinosa, Lauren Franz, Pradeep Raj Krishnappa Babu, Guillermo Sapiro, et al. Early detection of autism using digital behavioral phenotyping. *Nature Medicine*, 29(10):2489–2497, 2023. 1, 2
- [18] Pavan Uttej Ravva, Behdokht Kiafar, Pinar Kullu, Jicheng Li, Anjana Bhat, and Roghayeh Leila Barmaki. Mmasd+: A novel dataset for privacy-preserving behavior analysis of children with autism spectrum disorder. *arXiv preprint arXiv:2408.15077*, 2024. 2
- [19] Hany Said, Khaled Mahar, Shaymaa E Sorour, Ahmed Elsheshai, Ramy Shaaban, Mohamed Hesham, Mustafa Khadr, Youssef A Mehanna, Ammar Basha, and Fahima A Maghraby. Imitasd: Imitation assessment model for children with autism based on human pose estimation. *Mathematics*, 12(21):3438, 2024. 1, 2
- [20] Athmar NM Shamhan, Marwa Qaraqe, and Dena Al-Thani. Advancements in automated assessment and diagnosis of autism spectrum disorder through multi-modality sensing technologies: Survey of the last decade. *IEEE Transactions on Cognitive and Developmental Systems*, 2025. 1, 2
- [21] Hala Shamseddine, Safa Otoum, and Azzam Mourad. On the feasibility of federated learning for neurodevelopmental disorders: Asd detection use-case. In *GLOBECOM 2022-2022 IEEE Global Communications Conference*, pages 1121–1127. IEEE, 2022. 3
- [22] Micah J Sheller, Brandon Edwards, G Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R Colen, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10(1):12598, 2020. 1, 3
- [23] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia*, 2024. 4
- [24] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing*, 29:9532–9545, 2020. 2
- [25] Alysa Ziyang Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE transactions on neural networks and learning systems*, 34(12):9587–9603, 2022. 3

- [26] Wenhan Wu, Ce Zheng, Zihao Yang, Chen Chen, Srijan Das, and Aidong Lu. Frequency guidance matters: Skeletal action recognition by frequency-aware mixed transformer. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1632–1641, 2024. [2](#), [3](#), [4](#), [6](#)
- [27] Jie Xu, Benjamin S Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. Federated learning for healthcare informatics. *Journal of healthcare informatics research*, 5 (1):1–19, 2021. [3](#)
- [28] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. [2](#)