ReSoFed: Reliability-Guided Model Souping for Robust Federated Learning in Heterogeneous Classroom Environments

Anonymous CVPR submission

Paper ID *****

Abstract

*Computer vision systems are increasingly used in educational environments to analyze classroom activities and support data-driven learning analytics. However, classroom visual data often contain sensitive student information and cannot be centralized across institutions due to privacy and governance constraints. Federated learning (FL) enables collaborative model training without sharing raw data, but its performance can degrade when visual data originate from heterogeneous capture environments. Under such conditions, unreliable client updates can introduce negative transfer during model aggregation. In this work, we propose **ReSoFed**[1], a privacy-preserving, reliability-guided federated aggregation framework designed to improve robustness under heterogeneous classroom environments. The framework estimates the cross-domain generalization capability of client models using a heterogeneous server-held validation set and incorporates this signal into a two-stage aggregation process. Specifically, a greedy model-soup procedure identifies a subset of reliable client models whose weight-space combination improves validation performance, followed by reliability-aware weighted aggregation. Experiments on the SCB dataset across multiple CNN and transformer backbones demonstrate that ReSoFed consistently outperforms standard federated learning baselines under heterogeneous visual conditions while preserving data privacy.*

1. Introduction

Computer vision technologies are increasingly integrated into educational environments to support classroom analytics, student engagement monitoring, and automated analysis of learning behaviors. By analyzing visual cues such as classroom activities, participation patterns, and student interactions, these systems can provide valuable insights for educators [24]. Recent research has also explored multimodal approaches that combine visual, behavioral, and contextual signals to detect collaborative learning states and support AI-assisted classroom analytics [2, 4]. However, deploying such vision-based educational systems across institutions presents a fundamental challenge: educational data is highly sensitive and subject to strict privacy regulations and institutional policies. Video streams, classroom recordings, and student behavior data cannot typically be centralized or shared across institutions due to ethical and privacy constraints.

Federated learning (FL) offers a promising solution by enabling collaborative model training across distributed institutions without exchanging raw data [11]. In this paradigm, each institution trains a model locally on its private data and shares only model parameters with a central server for aggregation [14], allowing knowledge sharing while preserving privacy. However, applying FL to real-world classroom vision systems introduces additional challenges. Classroom environments differ significantly across institutions in terms of camera hardware, lighting conditions, viewpoints, and recording setups, leading to substantial distribution shifts in visual data [7]. These variations produce highly non-IID client distributions [1], which can degrade the performance and stability of federated learning models and limit their ability to generalize across heterogeneous institutions.

A critical but often overlooked factor contributing to this issue is that not all client updates are equally reliable. Clients operating under degraded visual conditions may produce models that overfit to environment-specific artifacts. Conventional federated aggregation strategies, such as Federated Averaging (FedAvg) [18], implicitly assume that all client updates contribute equally useful information and therefore aggregate models using uniform or data-size–weighted averaging. Under heterogeneous visual conditions, this assumption can lead to negative transfer, where poorly generalizing client updates degrade the performance of the global model [22]. This challenge becomes particularly pronounced in educational settings, where classroom capture conditions can vary substantially across institutions.

[1] Codes and data are available at TBA

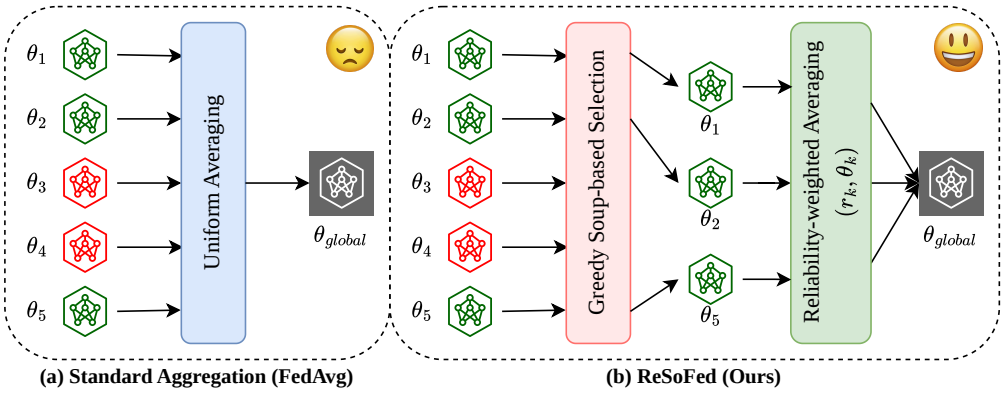


Figure 1. Comparison between (a) standard federated aggregation (FedAvg) and (b) the proposed **ReSoFed** framework. FedAvg uniformly aggregates all client updates $(\theta_1, \cdots, \theta_5)$, which may introduce negative transfer under heterogeneous visual conditions. In contrast, ReSoFed performs greedy model-soup–based client selection to filter unreliable clients and then applies reliability-aware weighted aggregation to produce the global model θ_{global}. Strongly generalizing clients are shown in green, while weak clients are marked in red.

In this work, we argue that cross-domain reliability can serve as a key signal for federated aggregation in privacy-preserving educational vision systems. Client models that generalize well beyond their local classroom conditions are more reliable for constructing a robust global model. Motivated by this insight, we propose **ReSoFed** (**Re**liability-guided **So**up for **Fed**erated Learning), a federated aggregation framework that explicitly estimates and leverages client reliability during model aggregation. Instead of assuming uniform contributions from all clients, ReSoFed evaluates locally trained models on a heterogeneous server-held validation set to estimate their cross-domain generalization capability. Building on this signal, the framework performs a two-stage aggregation process. First, a greedy model-soup–based client selection procedure iteratively combines candidate models only when their inclusion improves validation performance, filtering out poorly generalizing updates. Second, reliability-aware weighted aggregation is applied over the selected clients, assigning greater influence to models with stronger cross-domain generalization. Figure 1 provides a high-level comparison between the standard FedAvg aggregation and the proposed ReSoFed framework. We evaluate ReSoFed on the SCB dataset [30] for classroom activity recognition, where client heterogeneity is simulated through diverse augmentations. Experimental results show that ReSoFed mitigates negative transfer while preserving privacy, requires no modification to local training objectives, and remains compatible with diverse architectures, including both convolutional and transformer-based models.

Our major contributions are as follows:

- We introduce **ReSoFed**, a privacy-preserving, reliability-guided federated aggregation framework that models client reliability via cross-domain validation performance under heterogeneous visual conditions.
- We propose a cross-domain reliability estimation strategy that evaluates client models on a heterogeneous server-held validation set to assess their generalization capability beyond local data distributions.
- A two-stage reliability-guided aggregation mechanism is developed that integrates greedy model-soup–based client selection with reliability-aware weighted averaging, enabling the server to filter weakly generalizing updates and prioritize robust client models during global aggregation.
- Extensive experiments demonstrate that ReSoFed consistently improves robustness and global model performance over standard federated learning baselines while remaining privacy-preserving, architecture-agnostic, and computationally lightweight.

2. Related Works

2.1. Federated Learning

Federated Learning (FL) is a distributed machine learning paradigm that enables multiple clients to collaboratively train a shared model without exchanging raw data, thereby preserving data privacy and reducing the need for centralized data collection [11, 18]. In a typical FL setup, each client performs local optimization on its private dataset and sends model weights to a central server, which aggregates them to form a global model. The most widely used baseline is Federated Averaging (FedAvg) [18], which aggregates client updates through data-size–weighted averaging. Despite its advantages, FL faces several challenges that are not present in traditional distributed learning settings. In particular, data heterogeneity across clients, often referred to as non-IID distribution, can significantly degrade model convergence and generalization performance [7, 14]. Such heterogeneity commonly arises in real-world deployments due to variations in device characteristics, environmental

conditions, or user behavior. To address the challenges posed by heterogeneous data, several extensions to FedAvg have been proposed. For example, FedProx introduces a proximal regularization term in the local optimization objective to stabilize training under heterogeneous conditions [15]. Other approaches focus on improving optimization and communication efficiency through modified aggregation strategies or adaptive optimization schemes [10, 23]. These methods highlight the importance of designing aggregation strategies that can effectively handle variability across distributed clients while maintaining the privacy-preserving benefits of federated learning.

2.2. Client Selection and Aggregation Strategies

Beyond the standard Federated Averaging (FedAvg) algorithm, a large body of work has focused on improving the aggregation and client participation strategies in federated learning. The aggregation rule plays a central role in federated optimization, as it integrates knowledge from distributed client models to produce a global model while addressing challenges such as statistical heterogeneity and communication constraints [19, 21]. One line of research focuses on improving aggregation through adaptive optimization strategies. Methods such as FedOpt introduce server-side adaptive optimization schemes to stabilize global model updates under heterogeneous client distributions [23]. Similarly, FedNova addresses objective inconsistency caused by varying local update steps across clients, improving convergence behavior in heterogeneous settings [28]. These approaches aim to improve the stability and efficiency of global aggregation without explicitly modeling client reliability. More recently, several works have investigated reliability-driven or contribution-aware aggregation mechanisms. These methods attempt to quantify the importance of each client update using metrics such as local loss, gradient similarity, or contribution scores, and adjust aggregation weights accordingly [9]. While these approaches demonstrate improvements under heterogeneous data conditions, many of them require additional information from clients or introduce significant computational overhead.

In contrast, our approach focuses on estimating cross-domain reliability using a server-held validation set and integrating this signal into the aggregation process through reliability-guided client selection and weighted model aggregation. This design enables improved robustness under heterogeneous visual conditions while maintaining compatibility with standard federated training pipelines.

2.3. Computer Vision for Education

Computer vision has increasingly been applied to educational environments to analyze student behaviors, engagement levels, and classroom interactions. Early research primarily focused on engagement recognition using facial expressions, head pose, and gaze patterns captured through webcams in online learning settings. Datasets such as DAiSEE have enabled the development of models that estimate student engagement levels from video streams by analyzing visual cues such as facial movements and attention patterns [5, 12]. More recent works leverage deep neural networks, including convolutional and transformer-based architectures, to improve engagement recognition using spatio-temporal visual signals [20, 25]. These studies demonstrate the potential of computer vision to support intelligent tutoring systems and provide educators with insights into student learning behavior. Another important research direction focuses on detecting and recognizing classroom activities from visual data. For example, the Student Classroom Behavior (SCB) dataset provides annotated images of common activities such as reading, writing, raising hands, and listening, enabling the development of automated behavior recognition systems [26]. Computer vision models based on object detection and action recognition frameworks have been proposed to analyze classroom scenes and identify student activities associated with engagement and participation [3]. These approaches aim to support classroom analytics and enhance teaching effectiveness by providing automated behavioral monitoring.

Despite significant progress in federated learning, aggregation strategies, and computer vision for education, several challenges remain when deploying vision models in real-world educational environments. Federated learning enables privacy-preserving collaborative training but suffers from degraded convergence and generalization under heterogeneous data distributions. Existing aggregation and client selection methods improve optimization stability but often do not explicitly capture cross-domain reliability of client models [13]. Meanwhile, most computer vision approaches for educational analytics, including engagement and behavior recognition, assume centralized training on pooled datasets [5, 26]. In practice, educational data are distributed across institutions and captured under highly heterogeneous visual conditions. To address these challenges, we propose a reliability-guided federated learning framework that combines cross-domain validation–based reliability estimation with model-soup–inspired client selection and reliability-aware aggregation, enabling robust and privacy-preserving learning across heterogeneous classroom environments.

3. Methodology

3.1. Problem Formulation

We consider a federated image classification setting with K distributed clients. Each client $k \in \{1, \dots, K\}$ possesses a private local dataset $\mathcal{D}_k = \{(x_i^k, y_i^k)\}_{i=1}^{n_k}$ of size n_k, where $x_i^k \in \mathcal{X}$ denotes an image and $y_i^k \in \{1, \dots, C\}$

denotes its corresponding label. In our setting, $C = 11$ for classroom action recognition. Due to heterogeneous visual capture conditions, the local datasets $\mathcal{D}_k$ are non-identically distributed (non-IID) across clients.

Let $f_\theta : \mathcal{X} \to \mathbb{R}^C$ denote a shared model parameterized by $\theta \in \mathbb{R}^d$. In standard federated learning, the objective is to minimize the global empirical loss:

$$\min_\theta \; F(\theta) = \sum_{k=1}^{K} \frac{n_k}{\sum_{j=1}^{K} n_j} F_k(\theta).$$

where $F_k(\theta)$ denotes the local empirical loss at client k.

Under heterogeneous visual conditions, direct aggregation of locally optimized parameters can cause negative transfer, as local models may overfit to client-specific distortions. To assess cross-domain generalization, we introduce a small server-held validation set $\mathcal{D}_{val}$ that reflects diverse capture conditions across clients. Importantly, $\mathcal{D}_{val}$ is used solely for aggregation-time evaluation, without accessing raw client data.

Our objective is to learn a global model θ_{global} that *(i)* achieves strong overall performance across heterogeneous clients, *(ii)* remains robust under domain shift, and *(iii)* preserves data privacy by sharing only model parameters.

3.2. Federated Learning

In standard federated learning, training proceeds in iterative communication rounds between a central server and K distributed clients. At round t, the server broadcasts the current global model parameters θ^t to participating clients. Each client performs local optimization on its private dataset $\mathcal{D}_k$, producing updated parameters θ_k^{t+1} after one or more local training epochs. The server then aggregates these locally optimized parameters to obtain the next global model.

The most common aggregation rule, known as Federated Averaging (FedAvg) [18], computes a data-weighted average of client updates:

$$\theta^{t+1} = \sum_{k=1}^{K} \frac{n_k}{\sum_{j=1}^{K} n_j} \theta_k^{t+1}.$$

This procedure approximates minimization of the global empirical objective while preserving data privacy, since only model parameters (or gradients) are communicated.

However, FedAvg implicitly assumes that client updates are equally reliable outside their local domains. Under heterogeneous visual conditions, local updates may be biased toward client-specific distortions, leading to suboptimal aggregation and negative transfer [16]. In such non-IID settings, uniform data-proportional weighting often fails to capture the true cross-domain generalization capability of individual clients. This limitation motivates the development of aggregation strategies that explicitly incorporate measures of client reliability during the update process.

3.3. ReSoFed

We introduce **ReSoFed**, a two-stage aggregation framework that explicitly incorporates cross-domain reliability into federated optimization. Rather than assuming uniform contribution from all clients, ReSoFed evaluates the generalization behavior of locally trained client models using a server-held validation set $\mathcal{D}_{val}$ and integrates this signal into the aggregation process. The framework consists of *(i)* greedy soup-based client selection and *(ii)* reliability-aware weighted aggregation, as illustrated in Figure 2.

Greedy Soup-Based Client Selection: Let $\{\theta_k\}_{k=1}^K$ denote the locally optimized client models received at the server after training. We first estimate the reliability of each client model using the validation set $\mathcal{D}_{val}$. Specifically, we compute a reliability score:

$$r_k = \text{Eval}(\theta_k, \mathcal{D}_{val}),$$

where $\text{Eval}(\cdot)$ denotes a validation metric (e.g., accuracy in our experiments). The reliability score r_k reflects how well client k generalizes beyond its local data characteristics.

Motivated by the model soup paradigm [29], which shows that weight-space averaging of independently fine-tuned models can improve generalization without increasing inference cost, we adapt greedy model souping to the federated aggregation setting. Unlike conventional model soups trained on centralized data, our approach performs reliability-guided souping over distributed client models.

Clients are sorted in descending order of reliability. We initialize the soup set S with the most reliable client:

$$S = \{k^\star\}, \quad k^\star = \arg\max_k r_k.$$

We then iteratively consider adding the remaining clients to S. For a candidate client j, we compute the uniformly averaged model over $S \cup \{j\}$:

$$\tilde{\theta}_{S\cup\{j\}} = \frac{1}{|S|+1} \sum_{k \in S\cup\{j\}} \theta_k.$$

If the validation performance improves, i.e.,

$$\text{Eval}(\tilde{\theta}_{S\cup\{j\}}, \mathcal{D}_{val}) > \text{Eval}(\tilde{\theta}_S, \mathcal{D}_{val}),$$

then client j is added to S; otherwise, it is discarded. This greedy procedure continues until all clients have been evaluated.

This stage serves as a hard filtering mechanism, preventing poorly generalizing or incompatible client updates from influencing the global model. By evaluating aggregated models directly on $\mathcal{D}_{val}$, the selection process explicitly favors combinations that improve cross-domain robustness.

Reliability-Aware Weighted Aggregation: After identifying the selected client subset S, we compute the final

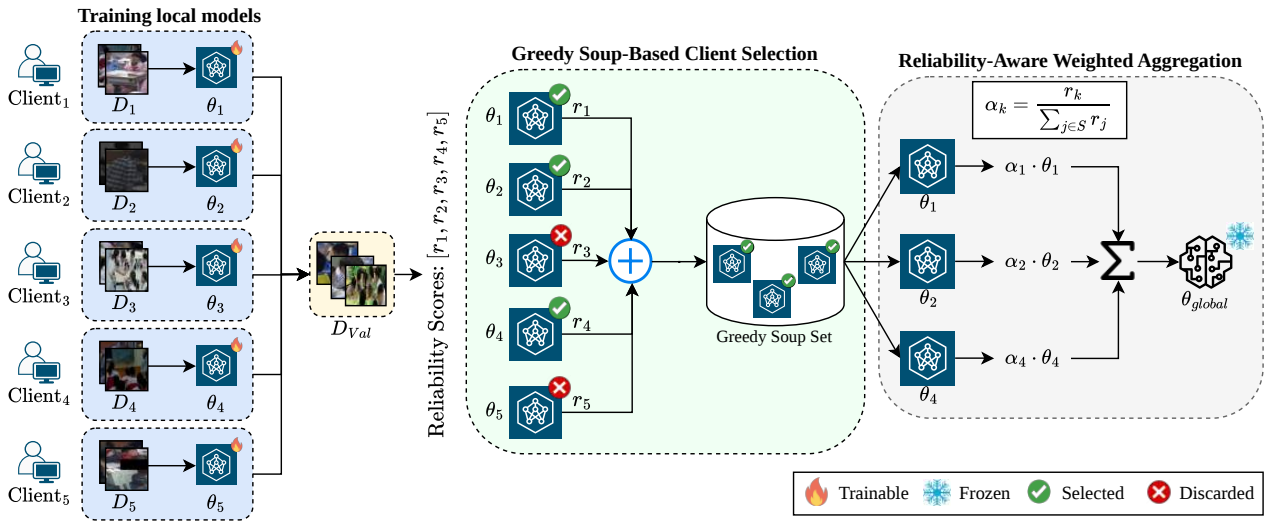


Figure 2. Overview of the proposed **ReSoFed** framework. Each client $k \in \{\text{Client}_1 \cdots \text{Client}_5\}$ trains a local model θ_k on its private dataset $\mathcal{D}_k$ under heterogeneous visual conditions (simulated via diverse augmentation regimes). Each locally trained model is evaluated on a diverse server-held validation set $\mathcal{D}_{val}$ to compute a reliability score r_k. **Greedy Soup-Based Client Selection:** Clients are sorted in descending order of reliability, and a soup set S is initialized with the most reliable client $k^\star = \arg\max_k r_k$. Additional clients are iteratively included only if the aggregated model improves validation performance. Poorly generalizing clients are discarded. **Reliability-Aware Weighted Aggregation:** The final global model is computed using reliability-proportional weights α_k resulting in $\theta_{\text{global}} = \sum_{k \in S} \alpha_k \theta_k$. Only model parameters are communicated, preserving data privacy while improving robustness under heterogeneous visual conditions.

global model using reliability-aware weighted aggregation over the selected clients. Instead of uniform averaging, we assign weights α_k proportional to the reliability scores:

$$\alpha_k = \frac{r_k}{\sum_{j \in S} r_j}, \quad \text{for } k \in S.$$

The global model is then computed as:

$$\theta_{\text{global}} = \sum_{k \in S} \alpha_k \theta_k.$$

This soft weighting mechanism further emphasizes clients that exhibit stronger cross-domain generalization while still allowing multiple reliable clients to contribute. Unlike standard FedAvg, which weights clients solely based on data size, ReSoFed directly incorporates validation-driven reliability into the aggregation process.

Importantly, the proposed framework does not modify local training objectives or optimization at the client level. All reliability estimation and selection steps are performed at the server using model parameters only, thereby preserving privacy and introducing negligible computational overhead compared to standard federated aggregation. By combining greedy soup-based client selection with reliability-weighted aggregation, ReSoFed effectively filters poorly generalizing clients while prioritizing models that demonstrate stronger cross-domain reliability. This reliability-guided aggregation strategy mitigates negative transfer under heterogeneous visual conditions, leading to more stable global updates and improved overall performance.

4. Experiments

4.1. Setup

Dataset: We conduct experiments on the SCB (Student Classroom Behavior) dataset [30], which contains images collected from real classroom environments and annotated with bounding boxes representing student and teacher activities. Using the provided annotations, we extract object-level image crops to construct an image classification dataset. The processed dataset contains 14 behavior categories derived from classroom activities and scene elements. The dataset exhibits notable class imbalance across behaviors, where frequent activities contain significantly more samples compared to rarer behaviors, resulting in a long-tailed distribution. In total, the resulting dataset contains over 90,000 labeled instances. It also captures diverse classroom interactions, viewpoints, and environmental conditions, providing a realistic benchmark for evaluating robust classroom behavior recognition models. Representative samples from the dataset are illustrated in Figure 3.

Federated Client Construction: To construct the federated learning setting, the processed dataset is partitioned across five distributed clients, each representing an independent institution with its own local training data. Prior to partitioning, several low-frequency classes are removed due

Figure 3. Sample cropped instances extracted from the SCB dataset using bounding box annotations [30]. The dataset captures diverse classroom behaviors and visual conditions, including variations in viewpoint, occlusion, and lighting commonly observed in real educational environments.

to insufficient sample counts, ensuring that the remaining categories contain adequate instances for reliable training and evaluation. The filtered dataset is subsequently divided into training, validation, and tests subsets. The training portion is distributed across the five clients while maintaining balanced class representation, whereas the validation and test sets remain centralized at the server. This partitioning introduces client-specific data distributions, reflecting the non-IID characteristics commonly observed in federated learning scenarios. Under this configuration, each client performs local model optimization on its private dataset while sharing only model parameters with the central server, thereby preserving data privacy while enabling collaborative model training.

Heterogeneous Camera Simulation: To simulate heterogeneous visual capture conditions across educational institutions, each federated client is trained under a distinct augmentation regime designed to emulate common camera artifacts observed in real-world classroom recordings. These augmentations model variations frequently encountered in webcam-based data acquisition, including illumination changes, motion blur, compression artifacts, and partial occlusions arising from differences in camera hardware, lighting conditions, and recording environments. The specific augmentation configuration assigned to each client is summarized in Table 1. Such degradations are widely studied in robustness benchmarks and are known to affect computer vision models under distribution shift. In particular, corruption types such as brightness variation, blur, and compression artifacts are commonly used in robustness benchmarks such as ImageNet-C [8]. Similarly, occlusion-based perturbations have been extensively explored in visual recognition through techniques such as Cutout [6] and Random Erasing [31]. In our setup, each client is assigned a specific augmentation configuration that approximates a distinct camera condition while preserving the semantic content of the images. This strategy introduces controlled cross-client visual heterogeneity, enabling systematic evaluation of federated learning methods under realistic domain shifts while maintaining consistent evaluation using a shared validation and test set.

Client	Simulated Condition	Reference
C_1	Clean reference capture	Standard baseline setting
C_2	Brightness shift	ImageNet-C [8]
C_3	Motion / defocus blur	ImageNet-C [8]
C_4	Compression artifacts	ImageNet-C [8]
C_5	Partial occlusion	Cutout [6], Random Erasing [31]

Table 1. Client-specific visual conditions used to simulate heterogeneous camera environments across federated institutions. Each client is trained under a distinct augmentation regime reflecting commonly observed real-world capture artifacts.

Implementation Details: All experiments are conducted using five simulated federated clients, each trained under distinct visual augmentation settings to emulate heterogeneous capture conditions. Specifically, client heterogeneity is introduced through different augmentation regimes. Each client trains a local model using its augmented data partition without sharing raw images. We evaluate the proposed framework across multiple backbone architectures, including three convolutional neural networks (CNNs) [17] (*ResNet-50*, *EfficientNet-V2-S*, and *ConvNeXt-B*) and three transformer-based models [27] (*Swin-V2-T*, *Swin-V2-B*, and *Vision Transformer B-16*). All models are fine-tuned on the SCB dataset for the classroom action recognition task. Images are resized and normalized using standard ImageNet statistics. Local models are trained using the Adam optimizer with a learning rate of 1×10^{-4}. A small server-held validation set $\mathcal{D}_{val}$ is used to estimate client reliability and guide the aggregation process in ReSoFed. This validation set is constructed to reflect diverse visual conditions and remains independent of all client training data. In our setup, reliability scores are computed using classification accuracy on $\mathcal{D}_{val}$. All experiments are implemented in *PyTorch* and executed on a single NVIDIA RTX 3090 GPU. We evaluate model performance using standard classification metrics, including Accuracy, Precision, Recall, and F1-score.

4.2. Results and Analysis

In this section, we present the evaluation results of the proposed framework, **ReSoFed**, on the SCB dataset for classroom action recognition. For comparison, we also report the performance of standard federated learning using Federated Averaging (FedAvg) under the same experimen-

Method	CNN Models					Transformer-based Models				
	Model	Acc (%)	Pre (%)	Rec (%)	F1 (%)	Model	Acc (%)	Pre (%)	Rec (%)	F1 (%)
Standard FL (FedAvg)	ResNet-50	73.82	77.34	73.67	72.83	ViT B-16	80.47	80.75	80.36	80.22
	EfficientNet-V2-S	78.93	80.31	78.79	78.45	Swin-V2-Tiny	81.23	82.56	81.10	80.80
	ConvNeXt-Base	81.82	81.97	80.78	81.37	Swin-V2-Base	81.96	83.26	81.73	81.41
ReSoFed	ResNet-50	78.50	78.38	78.44	78.06	ViT B-16	81.00	81.18	80.91	80.85
	EfficientNet-V2-S	81.61	81.49	81.57	81.22	Swin-V2-Tiny	81.68	83.01	81.58	81.18
	ConvNeXt-Base	**82.75**	**83.00**	**82.68**	**82.35**	Swin-V2-Base	**82.35**	**82.55**	**82.29**	**82.16**

Table 2. Performance comparison between standard federated learning (FedAvg) and the proposed ReSoFed framework on the SCB dataset. Results are reported across CNN and transformer architectures. **Bold** values indicate the best performance within each architecture group.

tal setting. Table 2 summarizes the classification performance across convolutional neural networks (CNNs) and transformer-based architectures to assess the effectiveness of ReSoFed across diverse model types.

The key observations from Table 2 are as follows: ***(1)*** ReSoFed consistently outperforms the standard federated aggregation (FedAvg) across all six models, including both CNN and transformer architectures, demonstrating the robustness and architecture-agnostic nature of the framework. ***(2)*** The performance improvements are particularly noticeable for CNN models. For instance, ResNet-50 shows a significant improvement in accuracy from 73.82% to 78.50%, while EfficientNet-V2-S improves from 78.93% to 81.61%, indicating that the proposed aggregation effectively mitigates negative transfer under heterogeneous visual conditions. ***(3)*** Although transformer architectures already achieve strong baseline performance under FedAvg, ReSoFed still provides consistent improvements, with Swin-V2-Base achieving the highest transformer performance at 82.35% accuracy. ***(4)*** Among all evaluated models, ConvNeXt-Base with ReSoFed achieves the best overall performance, reaching 82.75% accuracy and 82.35% F1-score, suggesting that modern CNN architectures benefit strongly from reliability-guided federated aggregation. ***(5)*** Overall, the consistent improvements across both architectures confirm that ReSoFed effectively handles client heterogeneity introduced by simulated camera conditions, validating the importance of reliability-guided client selection and weighted aggregation in federated learning settings.

4.3. Ablation Studies

To demonstrate the contribution of each component in the proposed ReSoFed framework, we conduct two ablation studies: (a) the effect of greedy soup-based client selection and (b) the effect of reliability-aware weighted aggregation. The ablations are performed on two representative backbone models, *ResNet-50* and *Swin-V2-Base*.

Effect of Greedy Soup-Based Client Selection: To evaluate the importance of the greedy client selection mechanism, we remove the greedy selection step and aggregate all client models using reliability-weighted averaging only. Figure 4a presents the accuracy and F1 scores achieved by the two models with and without the greedy selection stage. The results show that removing the greedy selection procedure leads to noticeable performance degradation for both models. This indicates that greedy soup-based client selection plays a crucial role in filtering weak generalizing client models before aggregation, thereby reducing negative transfer and improving the robustness of the global model.

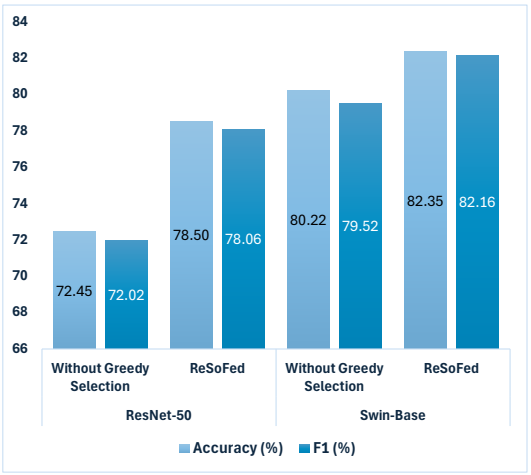
Effect of Reliability-Aware Weighted Aggregation: We next evaluate the impact of reliability-aware weighted aggregation. In this setting, we retain the greedy client selection stage but replace reliability-based weighting with uniform averaging among the selected clients. Figure 4b presents the performance before and after applying reliability weights during aggregation. The results demonstrate that reliability weighting provides additional performance gains beyond greedy selection alone. This confirms that reliability-aware weighting effectively prioritizes client models with stronger cross-domain generalization ability, further improving the performance of the aggregated model.

Overall, these ablations verify that both components of ReSoFed, greedy client selection and reliability-guided aggregation, contribute meaningfully to the final performance improvements observed in the proposed framework.

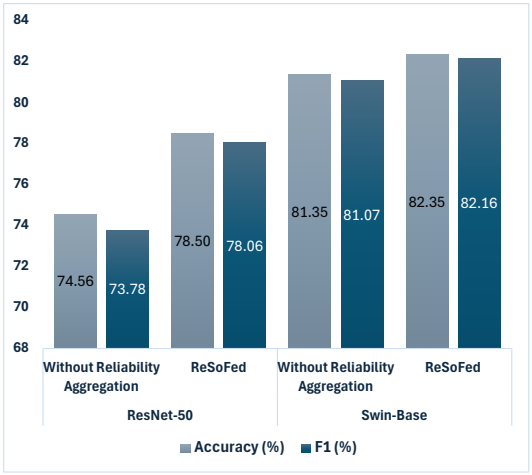
5. Discussion

The experimental results demonstrate that incorporating reliability-aware aggregation into federated learning can significantly improve model performance under heterogeneous visual conditions. By combining greedy model-soup–based client selection with reliability-weighted aggregation, ReSoFed mitigates the negative transfer effects that often arise when locally biased client models are aggregated indiscriminately. The consistent improvements observed across both convolutional and transformer-based backbones indicate that the proposed strategy is architecture-agnostic and can generalize across diverse model families.

Another important observation is that reliability estimation using a diverse server-held validation set provides an effective proxy for cross-domain generalization. Clients



(a) Effect of greedy soup-based model selection



(b) Effect of reliability-aware weighted aggregation

Figure 4. Ablation study of the ReSoFed framework. **(a)** Effect of greedy soup-based client selection, where removing greedy selection significantly degrades performance for both ResNet-50 and Swin-Base models. **(b)** Effect of reliability-aware weighted aggregation, showing that incorporating reliability scores during aggregation improves both accuracy and F1-score compared to uniform averaging.

whose locally trained models generalize well to this validation set are more likely to contribute positively to the global model. This allows ReSoFed to selectively emphasize reliable clients while discarding harmful updates originating from locally overfitted models. As a result, the global model benefits from better overall classification performance.

Furthermore, the proposed approach operates entirely at the aggregation stage without modifying the local training procedure. This design choice makes ReSoFed easy to integrate into existing federated learning pipelines. Since reliability estimation and client selection occur only at the server using model parameters, the approach preserves the core privacy advantages of federated learning while introducing minimal computational overhead.

Limitations: Despite its effectiveness, the proposed framework has several limitations. First, the reliability estimation depends on a server-held validation set representing diverse visual conditions; if this set is not representative of the client distributions, the reliability scores may not accurately capture cross-domain generalization. Second, our evaluation is conducted in a simulated setting where heterogeneity is introduced through controlled augmentations. Although these augmentations are based on established robustness benchmarks, real-world deployments may exhibit more complex distribution shifts. Third, the greedy client selection stage requires additional validation-time evaluation during aggregation, which may increase computational cost as the number of clients grows.

6. Conclusion

In this work, we introduced **ReSoFed**, a reliability-guided federated aggregation framework designed to improve robustness under heterogeneous visual conditions. The proposed approach combines greedy model-soup–based client selection with reliability-aware weighted aggregation, enabling the server to prioritize clients that demonstrate stronger cross-domain generalization. By leveraging a server-held validation set to estimate client reliability, ReSoFed mitigates the negative transfer that often arises when locally biased models are aggregated indiscriminately. Experimental results on the SCB dataset for classroom action recognition demonstrate that ReSoFed consistently outperforms standard federated learning, FedAvg, across multiple backbone architectures, including both convolutional neural networks and transformer-based models. These improvements highlight the effectiveness of reliability-guided aggregation for handling client heterogeneity while preserving the privacy of raw client data.

Future research directions include extending ReSoFed to multi-round federated optimization settings, exploring more advanced reliability estimation strategies beyond validation accuracy, and evaluating the framework on larger real-world federated settings with naturally occurring client heterogeneity. Additionally, exploring alternative reliability estimation strategies and more advanced client selection mechanisms could further enhance the robustness of federated aggregation under diverse real-world data conditions.

References

[1] Kasra Borazjani, Payam Abdisarabshali, Naji Khosravan, and Seyyedali Hosseinalipour. Redefining non-iid data in federated learning for computer vision tasks: Migrating from labels to embeddings for task-specific data distributions. *arXiv preprint arXiv:2503.14553*, 2025. 1

[2] Mariah Bradford, Ibrahim Khebour, Nathaniel Blanchard, and Nikhil Krishnaswamy. Automatic detection of collaborative states in small groups using multimodal features. In *International Conference on Artificial Intelligence in Education*, pages 767–773. Springer, 2023. 1

[3] H. Chen et al. Student behavior detection in classroom videos based on improved yolov8. *Sensors*, 2023. 3

[4] Clayton Cohn, Caitlin Snyder, Joyce Horn Fonteles, Ashwin TS, Justin Montenegro, and Gautam Biswas. A multimodal approach to support teacher, researcher and ai collaboration in stem+ c learning environments. *British Journal of Educational Technology*, 56(2):595–620, 2025. 1

[5] Wen Dai et al. Daisee: Towards user engagement recognition in the wild. In *ACM International Conference on Multimodal Interaction*, 2019. 3

[6] Terrance DeVries and Graham Taylor. Improved regularization of convolutional neural networks with cutout. In *arXiv preprint arXiv:1708.04552*, 2017. 6

[7] Dashan Gao, Xin Yao, and Qiang Yang. A survey on heterogeneous federated learning. *arXiv preprint arXiv:2210.04505*, 2022. 1, 2

[8] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019. 6

[9] Jiyue Huang, Chi Hong, Lydia Y. Chen, and Stefanie Roos. Is shapley value fair? improving client selection for mavericks in federated learning. In *International Conference on Machine Learning (ICML)*, 2021. 3

[10] Muhammad Rafsan Kabir, Rashidul Hassan Borshon, and Riasat Khan. Federated learning for human activity recognition: Balancing privacy, efficiency, and accuracy through innovative aggregation techniques. *Array*, page 100462, 2025. 3

[11] Peter Kairouz and H Brendan McMahan. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1-2):1–210, 2021. 1, 2

[12] S. N. Karimah et al. Automatic engagement estimation in smart education: A review. *IEEE Access*, 2022. 3

[13] J. Li et al. Client selection strategies in federated learning: A comprehensive survey. *Computer Networks*, 2024. 3

[14] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020. 1, 2

[15] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of MLSys*, 2020. 3

[16] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations (ICLR)*, 2020. 4

[17] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):6999–7019, 2021. 6

[18] 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 International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1273–1282. PMLR, 2017. 1, 2, 4

[19] Mohammad Moshawrab et al. Reviewing federated learning aggregation algorithms. *Electronics*, 12(10):2287, 2023. 3

[20] Sindhuja Penchala et al. Learning in focus: Detecting behavioral and collaborative engagement using vision transformers. In *International Conference on Computer Vision in Education*, 2025. 3

[21] Peng Qi et al. Model aggregation techniques in federated learning: A systematic review. *Future Generation Computer Systems*, 2024. 3

[22] Liangqiong Qu, Yuyin Zhou, Paul Pu Liang, Yingda Xia, Feifei Wang, Ehsan Adeli, Li Fei-Fei, and Daniel Rubin. Rethinking architecture design for tackling data heterogeneity in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10061–10071, 2022. 1

[23] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations (ICLR)*, 2021. 3

[24] Diganta Sengupta, Soumya Suvra Khan, Surajit Das, and Debashis De. Fedel: Federated education learning for generating correlations between course outcomes and program outcomes for internet of education things. *Internet of Things*, 25:101056, 2024. 1

[25] Farhad M. Shiri et al. Detection of student engagement in e-learning environments using hybrid deep learning models. *Journal of Artificial Intelligence*, 2024. 3

[26] B. Sun, Y. Wu, K. Zhao, et al. Student classroom behavior dataset: A dataset for recognizing classroom behaviors. *Neural Computing and Applications*, 2021. 3

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 6

[28] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

[29] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning (ICML)*, pages 23965–23998. PMLR, 2022. 4

[30] Fan Yang. Scb-dataset: A dataset for detecting student classroom behavior. *arXiv preprint arXiv:2304.02488*, 2023. 2, 5, 6

[31] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020. 6