Adaptive Curriculum Learning for Bidirectional Federated Knowledge Distillation in Heterogeneous Client Environments

Sarim Malik
26100129 Lahore University of Management Sciences

Muhammad Nafees
26100029 Lahore University of Management Sciences

Abstract

Federated Learning (FL) has emerged as a promising paradigm for collaborative model training while preserving data privacy. However, performance significantly degrades in non-independent and identically distributed (non-i.i.d.) settings, presenting critical challenges for both server and client model generalizability. This paper introduces HiPer-FedKD, a novel heterogeneity-aware personalized federated knowledge distillation framework that implements adaptive curriculum learning to dynamically adjust training difficulty based on client data heterogeneity. Our approach combines bidirectional knowledge transfer between clients and server with personalized temperature scaling mechanisms that adapt to each client's unique data distribution. We address key limitations in existing methods through logit standardization to align knowledge representation across heterogeneous clients and exponential temperature scaling that responds optimally to distribution shifts. Extensive experiments demonstrate that HiPer-FedKD significantly outperforms baseline federated learning approaches, particularly in extreme non-i.i.d. scenarios, improving both server generalization and client personalization simultaneously.

Code Repository: https://github.com/s-malix21/HiPer-FedKD

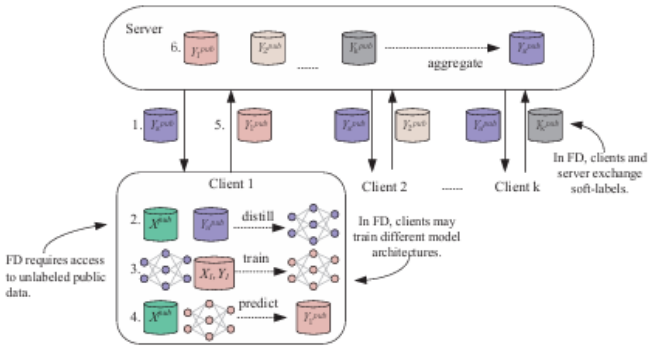

Figure 1: A general FD framework

1 Introduction

Federated Learning (FL) enables collaborative model training across decentralized edge devices while preserving data privacy, as clients only share model updates rather than raw data [9]. However, real-world federated learning scenarios commonly involve heterogeneous client data distributions (non-i.i.d. data), which significantly degrades model performance and creates challenges for both server generalization and client personalization [15]. This statistical heterogeneity causes clients to converge toward different local optima, resulting in weight divergence during aggregation and ultimately degrading the global model's performance. Knowledge Distillation (KD), first proposed by Hinton et al. [4], offers a promising approach by transferring knowledge through soft probability distributions, and recent works have integrated KD with FL to create Federated Knowledge Distillation (FedKD) techniques [7, 17] that reduce communication costs and mitigate some effects of data heterogeneity.

Despite these advances, critical limitations persist in current approaches. Methods such as FedBiKD [14] and FedBKD [12] attempt bidirectional knowledge distillation but lack mechanisms to effectively handle extreme data heterogeneity. Traditional KD methods apply uniform temperature scaling across all clients, disregarding varying degrees of client heterogeneity—a "one-size-fits-all" approach unsuitable for non-i.i.d. settings. During knowledge distillation in heterogeneous environments, clients with different data distributions produce logits with varying scales, creating inconsistencies during knowledge transfer and aggregation [16]. Furthermore, existing methods like He et al.'s group knowledge transfer [3] struggle to simultaneously improve both server generalization and client personalization, while approaches like FedBalancer [15] lack truly adaptive mechanisms that dynamically adjust the learning process based on each client's unique characteristics.

To address these challenges, we propose HiPer-FedKD (Heterogeneity-aware Personalized Federated Knowledge Distillation), which implements bidirectional heterogeneity-aware knowledge transfer between clients and server while accounting for client-specific data distributions. Building upon Lee et al.'s temperature scaling work [6], we introduce personalized logit chilling through an adaptive curriculum learning mechanism that dynamically computes client-specific temperature parameters based on the heterogeneity between a client's data distribution and the global distribution. We incorporate logit standardization techniques [16] to normalize logit distributions across heterogeneous clients, ensuring consistent scale during knowledge transfer. Additionally, we propose a novel exponential temperature scaling approach that responds more aggressively to higher heterogeneity, providing optimal adjustments for clients with extremely skewed data distributions. Our comprehensive empirical evaluation tests HiPer-FedKD against baseline methods across multiple Dirichlet parameter settings, representing varying degrees of non-i.i.d. conditions. The results will be expanded upon in the sections ahead.

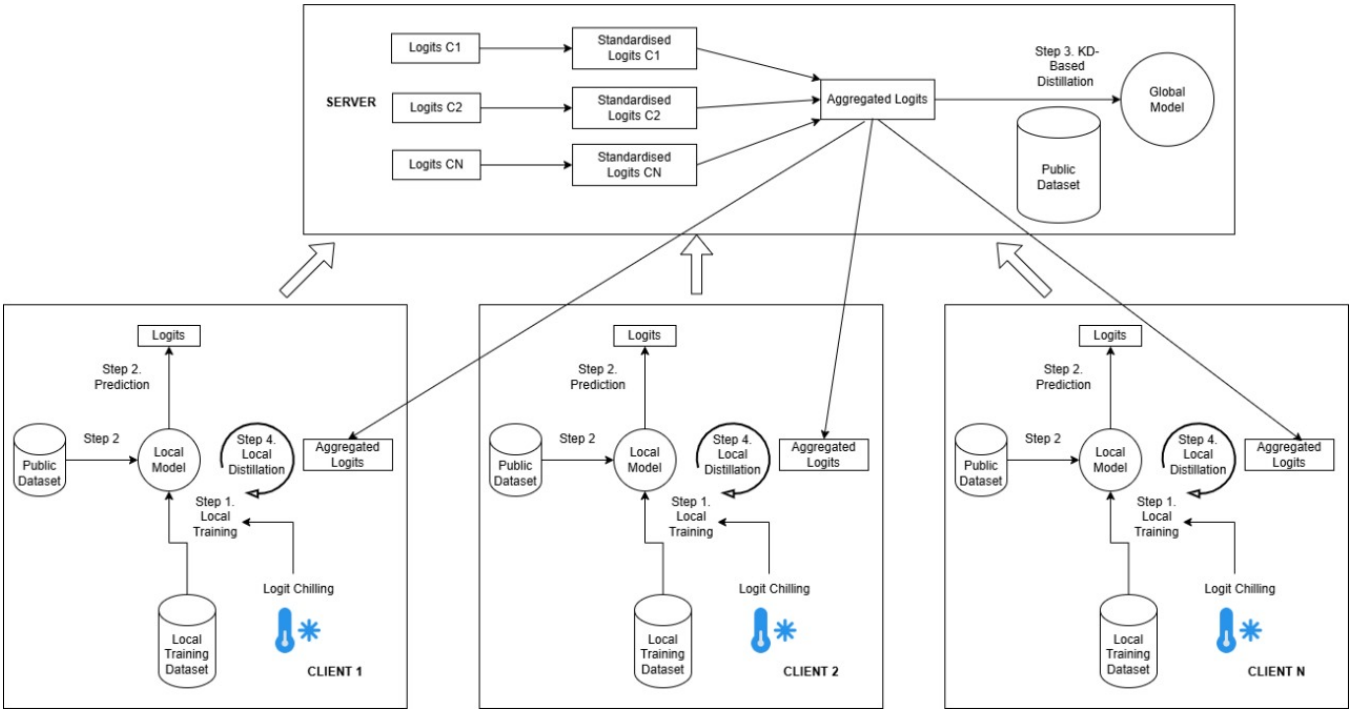

Figure 2: Proposed framework for HiPer-FedKD

2 Methodology

2.1 Experimental Setup

2.1.1 Problem Formulation We formulate the federated learning problem as a collaborative model training task among N heterogeneous clients, each with a local dataset $D_i = \{(x_j, y_j)\}_{j=1}^{n_i}$, where n_i is the number of samples at client i [12]. The objective is to learn a global model that generalizes well across all client distributions while simultaneously enabling each client to perform well on its local distribution.

In our bidirectional knowledge distillation approach, each client maintains its own local model f_i, while the server maintains a global model f_G. Knowledge is transferred between clients and the server through predicted logits on a shared unlabeled public dataset D_P. Specifically, the optimization involves three components: local training of client models on their private data, client-to-server knowledge transfer through aggregated logits, and server-to-client knowledge transfer through distillation from the global model [10].

The challenge lies in effectively handling heterogeneous data distributions across clients, where each client may have a significantly different class distribution compared to the global distribution, resulting in a non-i.i.d. setting that hinders traditional federated learning approaches.

2.1.2 Dataset Selection We use the CIFAR-10 dataset for our experiments due to its ideal balance of complexity and size for federated learning. CIFAR-10 consists of 60,000 color images across 10 classes, with 50,000 for training and 10,000 for testing. Its consistent image size (32×32), rich visual features, and class similarity structure make it well-suited for evaluating our bidirectional knowledge distillation framework. Moreover, its widespread use provides a reliable benchmark for comparison with other methods.

2.1.3 Partitioning Strategy and Client Distribution Configuration To simulate realistic federated learning heterogeneity, we employed a non-i.i.d. data partitioning strategy using the Dirichlet distribution with concentration parameter α. The 50,000 CIFAR-10 training images were split into 90% (45,000) for client training/validation and 10% (5,000) as a server-side public validation set for knowledge distillation. The client portion was then distributed among $N = 4$ clients using a Dirichlet distribution with $\alpha = 1.0$, inducing moderately heterogeneous local datasets.

For each class c, a class-wise probability vector $p_c \sim \text{Dir}(\alpha)$ determined the fraction of its samples assigned to each client, resulting in imbalanced client datasets—some rich in certain classes, others lacking them entirely. Each client's data was further split into 80% for local training and 20% for local validation.

The parameter α governs heterogeneity: lower values create more skewed (non-i.i.d.) splits, while higher values yield balanced ones. Our choice of $\alpha = 1.0$ offers a realistic level of heterogeneity while remaining representative of practical use cases. We also varied α during ablation studies to assess robustness under different non-i.i.d. levels. This setup ensures each client has a distinct class distribution, the global distribution stays balanced, and heterogeneity remains controllable.

2.1.4 Network Architectures and Hyperparameters To isolate the effects of our proposed methods from architectural factors,

we used a consistent lightweight CNN across all clients and the server. The model, denoted `SimpleCNN`, is defined as:

$$\begin{aligned}&\text{Conv2d}(3, 32, 3, \text{padding} = 1) \rightarrow \text{ReLU} \rightarrow \text{MaxPool}(2, 2)\\&\text{Conv2d}(32, 64, 3, \text{padding} = 1) \rightarrow \text{ReLU} \rightarrow \text{MaxPool}(2, 2)\\&\text{Conv2d}(64, 128, 3, \text{padding} = 1) \rightarrow \text{ReLU} \rightarrow \text{MaxPool}(2, 2)\\&\text{Flatten} \rightarrow \text{Linear}(128 \times 4 \times 4, 512) \rightarrow \text{ReLU} \rightarrow \text{Dropout}(0.25)\\&\rightarrow \text{Linear}(512, 10)\end{aligned}$$

This architecture balances feature learning capacity with computational efficiency for federated learning on CIFAR-10. We trained all models using SGD with the following hyperparameters:

- **Batch size:** 64
- **Learning rate:** 0.01
- **Local training epochs per round:** 5 (per client)
- **Server distillation epochs per round:** 10
- **Client distillation epochs per round:** 5
- **Total communication rounds:** 10
- **Distillation temperature:** 3.0 (shared among both client and server models)

These hyperparameters were kept consistent across all experiments to ensure fair comparisons, except for the adaptively set client softmax temperatures based on our heterogeneity-aware personalization approach.

2.1.5 Baseline Model for Comparison Our baseline model implements a standard bidirectional federated knowledge distillation approach without our proposed improvements. The process follows these steps in each communication round: In the local training step, each client trains its local model on private data using standard cross-entropy loss with a uniform temperature of 1.0. For client-to-server knowledge transfer, clients generate logits on the server's public dataset, which the server aggregates through simple averaging. The global model is then updated via knowledge distillation from the aggregated logits. In server-to-client knowledge transfer, the server's aggregated logits are distilled back to the clients, and each client updates its local model through KL-divergence minimization. This baseline represents a typical approach to bidirectional knowledge distillation in federated learning, similar to methods proposed in existing literature, and serves as a reference point to evaluate the effectiveness of our personalized curriculum learning and logit standardization techniques.

2.1.6 Evaluation Metrics To comprehensively assess the performance of our approach, we employed multiple evaluation metrics targeting different aspects of the federated learning system:

- **Server Test Accuracy:** Accuracy of the global model on the held-out CIFAR-10 test set (10,000 images), measuring overall generalization performance.
- **Client Private Data Accuracy:** Accuracy of each client's model on its local validation set, measuring how well client models perform on their private data distributions.
- **Client Public Data Accuracy:** Accuracy of each client's model on the server's public dataset, measuring how well client models generalize beyond their private data.
- **Round-by-Round Convergence Analysis:** Tracking accuracy metrics across communication rounds to assess convergence speed and stability.
- **Pre/Post-Distillation Improvement:** Measuring the improvement in client accuracy before and after each distillation phase to quantify knowledge transfer effectiveness.

These metrics provide a multifaceted view of system performance, allowing us to assess overall system efficacy (server test accuracy), personalization benefits (client private data accuracy), generalization improvements (client public data accuracy), and knowledge transfer effectiveness (pre/post-distillation improvement). Together, they enable a thorough evaluation of how well our proposed approaches address the dual challenges of client personalization and server generalization in heterogeneous federated learning environments.

3 Key Innovations

3.1 Model Variant 1: Adaptive Curriculum Learning

For Model Variant 1, we introduce an adaptive curriculum that first quantifies each client's data heterogeneity and then tailors its local training dynamics accordingly. In Section 3.1.1 we describe our heterogeneity measurement mechanism, which computes a normalized divergence score between each client's class distribution and the global distribution. Building on those scores, Section 3.1.2 presents our personalized logit chilling technique, where we adjust the softmax temperature during local training to sharpen or soften predictions in line with each client's data skew.

3.1.1 Heterogeneity Measurement The cornerstone of our adaptive approach is the accurate quantification of data heterogeneity at each client [7]. We develop a comprehensive heterogeneity measurement mechanism that compares each client's class distribution against the global distribution.

For each client, we calculate the heterogeneity score by examining the class-wise distribution divergence. Specifically, for each class, we compute the ratio between the client's class probability and the global class probability. The absolute difference between this ratio and 1 indicates the degree of divergence for that class. By summing these differences across all classes and normalizing the result, we obtain a score that ranges from 0 to 1.

A score of 0 indicates perfect alignment with the global distribution—the client's data perfectly represents the overall system's data distribution. Conversely, a score approaching 1 signifies extreme heterogeneity, where the client's data distribution differs substantially from the global average. This scoring mechanism provides a fine-grained understanding of each client's position within the heterogeneity spectrum.

The global distribution is approximated by averaging the class distributions across all clients, with adjustment mechanisms to ensure that classes with very low representation still maintain a minimum probability. This approach provides a robust reference point against which individual client distributions can be compared.

3.1.2 Personalized Logit Chilling Based on heterogeneity scores, we implement a novel personalized logit chilling technique that

adapts softmax temperature during local training according to each client's specific needs [6]. This approach addresses the fundamental challenge of optimizing local models with skewed distributions in federated learning environments.

We derive a personalized temperature for each client using a linear mapping from heterogeneity scores, with higher heterogeneity corresponding to lower temperatures between 0.05 and 0.99. During local training, clients apply their personalized temperature to the softmax operation when computing cross-entropy loss.

Lower temperatures increase contrast between logit values, causing softmax to assign higher probabilities to the highest logit [4]. This sharpening effect amplifies confident predictions, propagates stronger gradients during backpropagation, and creates an implicit curriculum by emphasizing distinctive features in heterogeneous data [6]. Our approach diverges from traditional federated learning by tailoring temperatures to each client's data characteristics rather than using identical softmax temperatures across the federation [10].

3.2 Model Variant 2: Logit Standardization

Building on the Adaptive Curriculum Learning improvements introduced in Model Improvement 1, Model Variant 2 tackles a complementary challenge in our federated distillation pipeline: the scale inconsistency of logits across clients. In the following two subsections, we first diagnose how heterogeneous temperature settings and diverse data distributions lead to misaligned logit magnitudes (Scale Inconsistency Problem), then present a Z-score normalization scheme that standardizes all client outputs into a unified space before aggregation and distillation (Z-score Normalization Approach).

3.2.1 Scale Inconsistency Problem Knowledge distillation in federated settings encounters a fundamental challenge: scale inconsistency when integrating knowledge from multiple sources [16]. This occurs when different models produce logits with varying magnitudes and distributions. Our experiments reveal this issue arises when different temperature parameters are applied to the softmax function during training.

The problem intensifies in heterogeneous federated environments where diverse data distributions naturally produce logits with different characteristics. Lower-temperature models generate logits with wider spreads and more extreme values, while higher-temperature models produce more moderate, closely clustered logits.

During aggregation, these scale differences bias the collective knowledge representation toward clients with more extreme logit values, regardless of their knowledge quality. This scale inconsistency results in biased aggregation, unstable optimization during distillation, and poorer generalization in both client and server models, necessitating a normalization approach that harmonizes logit scales while preserving valuable relative relationships between class predictions [16].

3.2.2 Z-score Normalization Approach To address the scale inconsistency problem, we apply Z-score normalization to standardize logits from different clients into a common space with zero mean and unit variance before aggregation and during distillation [16].

The Z-score normalization of logits is formulated as:

$$\hat{z}_i = \frac{z_i - \mu_z}{\sigma_z + \epsilon}$$

where z_i represents the original logit value, μ_z is the mean across classes, σ_z is the standard deviation, and ϵ is a small constant for numerical stability.

For client logit aggregation, we first standardize each client's logits individually:

$$L'_k = \frac{L_k - \mu_{L_k}}{\sigma_{L_k} + \epsilon}$$

This per-client standardization ensures that aggregation treats each client equally, regardless of its original temperature setting or data distribution.

The standardized client logits are then averaged to produce the aggregated knowledge:

$$L_{\text{agg}} = \frac{1}{K} \sum_{k=1}^{K} L'_k$$

During knowledge distillation, we apply this same standardization to the server's output logits before computing the KL-divergence loss with the standardized aggregated logits. This ensures both teacher and student distributions exist in the same scale space, complementing classical temperature scaling techniques used in distillation [4]. This approach effectively reconciles scale divergence, ensuring balanced contributions from all clients while preserving relative relationships between class predictions. It also enables fair comparison between student and teacher distributions during distillation [16].

By standardizing logits via Z-score normalization, Model Improvement 2 ensures that every client—regardless of its local softmax temperature or data skew—contributes on equal footing to the global knowledge pool. This approach dovetails with our prior adaptive curriculum strategy: while Model 1 personalized the learning pace per client, Model 2 harmonizes the raw outputs they share. Together, these two innovations yield a more stable, fair, and robust federated learning process.

3.3 Model Variant 2.1: HiPer-FedKD

This model introduces **HiPer-FedKD**, a unified framework combining personalized curriculum learning with standardized knowledge transfer for federated learning in heterogeneous environments. Building on our earlier work, HiPer-FedKD retains personalized temperature scaling and logit standardization, while introducing exponential scaling for finer control over client-specific learning. This allows dynamic adaptation to client diversity with consistent bidirectional knowledge transfer.

Its strength lies in balancing personalization and generalization: exponential scaling tailors learning to client heterogeneity, while logit standardization supports effective server-side aggregation. Unlike prior methods that trade off one goal for the other, HiPer-FedKD shows these objectives can reinforce each other, especially under highly non-iid conditions. Furthermore, we have disclosed the exact algorithmic working of HiPer-FedKD in Algorithm 1.

Algorithm 1 HiPer-FedKD

Require: Number of rounds R; Clients K; Global model θ; Client models $\{\theta_k\}_{k=1}^{K}$; Public dataset $\mathcal{D}_{pub}$; Private datasets $\{\mathcal{D}_k\}_{k=1}^{K}$

Initialization Phase
Calculate global class distribution G from all clients
for each client $k = 1$ to K **do**
  Calculate heterogeneity score s_k relative to G
  Set temperature $T_k = 0.99 \cdot e^{-3s_k}$
end for
for each round $r = 1$ to R **do**
  Local Training Phase
  for each client $k = 1$ to K in parallel **do**
    Train local model θ_k on $\mathcal{D}_k$ using temperature-scaled CE:
      $\mathcal{L} = -\sum_i y_i \log(\text{softmax}(z_i/T_k))$
    Generate logits L_k on $\mathcal{D}_{pub}$
    Standardize: $L_k = \frac{L_k - \mu(L_k)}{\sigma(L_k)+\epsilon}$
  end for
  Server Aggregation and Update Phase
  Compute aggregated knowledge: $L_{agg} = \frac{1}{K}\sum_{k=1}^{K} L_k$
  Generate logits L_θ on $\mathcal{D}_{pub}$ from global model
  Standardize L_θ using Z-score normalization
  Update θ by minimizing:
    $\mathcal{L}_{KD} = KL(\text{softmax}(L_{agg}/T) \| \text{softmax}(L_\theta/T))$
  Client Update
  for each client $k = 1$ to K in parallel **do**
    Generate logits L_{θ_k} from client model on $\mathcal{D}_{pub}$
    Standardize L_{θ_k} using Z-score normalization
    Update θ_k by minimizing:
      $\mathcal{L}_{KD} = KL(\text{softmax}(L_{agg}/T) \| \text{softmax}(L_{\theta_k}/T))$
  end for
end for
return Updated global model θ and client models $\{\theta_k\}_{k=1}^{K}$

3.3.1 Exponential Temperature Scaling The effectiveness of temperature scaling in knowledge distillation depends heavily on how well the temperature parameter reflects task complexity. While our initial linear mapping from heterogeneity scores to temperature offered a solid starting point, further analysis revealed room for refinement.

Exponential temperature scaling introduces a non-linear relationship between client heterogeneity and temperature values, enabling stronger adjustments for highly heterogeneous clients and smoother scaling for moderate ones. It is defined as:

$$T(\text{softmax}) = 0.99 \cdot e^{-k \cdot \text{score}}$$

where k is a decay factor controlling curve steepness, and score denotes the client's heterogeneity score. Empirically, we found $k = 3$ to yield optimal performance.

We found this approach to offer three main advantages over linear scaling:

- It rapidly approaches low temperatures (near 0.05) for highly skewed clients, enhancing the learning signal.
- It ensures gradual transitions for clients with moderate heterogeneity, allowing finer discrimination.
- It retains temperatures close to 1.0 for clients with low heterogeneity, preserving standard training behavior.

In practice, exponential scaling improves personalization by delivering a more precisely calibrated training curriculum based on client-specific data characteristics. Experimental results show consistent accuracy gains over linear scaling, both for individual clients and in the global model's generalization, highlighting the value of refined local learning and knowledge aggregation.

4 Results

4.1 Experimental Findings

Our experimental results demonstrate significant performance improvements across varying degrees of data heterogeneity (Table 1). In the mild heterogeneity setting ($\alpha = 1.0$), our final model (Improvement 2.1) achieves a server test accuracy of 56.99%, representing a substantial 10.24 percentage point increase over the baseline (46.75%). The exponential temperature scaling approach consistently outperforms the linear scaling method, with all clients showing enhanced private accuracy ranging from 59.11% to 68.75%. This improvement is particularly notable as it maintains strong public validation accuracy (58.96%–63.92%), indicating effective knowledge transfer across the federation despite data distribution differences.

The benefits of our approach become even more pronounced in moderate heterogeneity scenarios. At $\alpha = 0.5$, Improvement 2.1 achieves a server test accuracy of (51.00%), surpassing the baseline by 12.35 percentage points. In the highly non-IID setting ($\alpha = 0.1$), which represents an extremely challenging and somewhat pathological data distribution unlikely in many practical scenarios, our exponential temperature scaling approach still maintains a server accuracy of (32.58%) compared to the baseline's (29.31%). Although absolute performance in this extreme setting is lower across all models, the relative improvement of our approach over the baseline remains significant. The overall decline in server accuracies as α decreases reflects the fundamental challenge of learning from highly skewed distributions with limited communication rounds. With only 10 communication rounds, the models have insufficient opportunities to generalize across the fragmented knowledge landscape created by extreme data partitioning. Client private accuracies in this challenging scenario show dramatic improvements, with some clients (e.g., Client 0) achieving up to 20.31 percentage points gain over baseline performance, though this specialized knowledge struggles to transfer effectively to the global model.These results validate our hypothesis that adaptive temperature scaling effectively compensates for data heterogeneity by providing more aggressive adjustments for clients with skewed distributions.

4.2 Performance Analysis

Examining the results more closely for the $\alpha = 1.0$ setting provides key insights into our framework's effectiveness. In the baseline model, client private data accuracy varies significantly (51.69% to 60.29%), highlighting the challenges of non-i.i.d. data, even under moderate heterogeneity. Public validation accuracy remains more

Method	Model	Alpha 1.0			Alpha 0.5			Alpha 0.1		
		Private Acc	Public Val Acc	Server Test Acc	Private Acc	Public Val Acc	Server Test Acc	Private Acc	Public Val Acc	Server Test Acc
Baseline	Client 0	51.71%	52.58%	NA	31.32%	44.50%	NA	59.38%	27.58%	NA
	Client 1	51.69%	51.46%	NA	49.54%	41.78%	NA	28.79%	31.70%	NA
	Client 2	60.29%	50.74%	NA	50.37%	50.38%	NA	49.24%	34.88%	NA
	Client 3	52.63%	52.64%	NA	63.26%	40.32%	NA	26.86%	31.50%	NA
	Server	NA	NA	46.75%	NA	NA	38.65%	NA	NA	29.31%
Improvement 1	Client 0	58.11%	58.96%	NA	62.77%	47.52%	NA	73.44%	17.48%	NA
	Client 1	64.41%	60.68%	NA	73.15%	47.92%	NA	24.50%	34.12%	NA
	Client 2	68.51%	56.74%	NA	57.79%	58.36%	NA	75.46%	37.96%	NA
	Client 3	60.38%	61.08%	NA	67.37%	38.48%	NA	18.29%	21.96%	NA
	Server	NA	NA	43.88%	NA	NA	27.15%	NA	NA	20.91%
Improvement 2	Client 0	58.90%	57.16%	NA	56.80%	52.92%	NA	78.12%	22.24%	NA
	Client 1	64.46%	57.60%	NA	68.94%	52.06%	NA	37.12%	34.18%	NA
	Client 2	66.99%	56.66%	NA	59.52%	58.38%	NA	67.78%	38.62%	NA
	Client 3	62.41%	61.00%	NA	62.89%	46.36%	NA	43.70%	32.88%	NA
	Server	NA	NA	53.20%	NA	NA	47.85%	NA	NA	31.91%
Improvement 2.1	Client 0	59.11%	61.16%	NA	57.16%	55.54%	NA	79.69%	25.54%	NA
	Client 1	67.17%	62.50%	NA	67.69%	52.86%	NA	41.64%	37.26%	NA
	Client 2	68.75%	58.96%	NA	61.09%	59.16%	NA	66.17%	40.34%	NA
	Client 3	65.67%	63.92%	NA	68.00%	48.56%	NA	43.34%	33.88%	NA
	Server	NA	NA	56.99%	NA	NA	51.00%	NA	NA	32.58%

Table 1: Comparison of Client and Server Performances across Different Model Improvements and Data Heterogeneity Levels (CIFAR-10 Dataset)

uniform (around 51–52%) but is notably lower than what improved knowledge transfer could yield.

Improvement 1, which introduces personalized logit chilling, creates a more tailored learning experience, yielding substantial gains in client private accuracy. Client 2 improves from 60.29% to 68.51% (+8.22%), while Client 1 sees the largest jump—from 51.69% to 64.41% (+12.72%). Public validation accuracy also improves across all clients (+5.92% to +9.22%), indicating better generalization. However, server test accuracy drops to 43.88% (−2.87%), likely due to scale inconsistencies introduced by client-specific temperature parameters that hinder effective aggregation.

Improvement 2 mitigates this issue through logit standardization, normalizing client logits before aggregation. This recovers server performance (53.20%, +6.45% over baseline), while client private performance remains strong, though public validation accuracies slightly decline—possibly reflecting reduced richness in client-specific representations.

The final variant, HiPer-FedKD (Improvement 2.1), combines exponential temperature scaling with logit standardization, achieving the best balance. Client private accuracies average 65.18% (+11.10%), public validation 61.64% (+9.80%), and server test accuracy peaks at 56.99% (+10.24%). The reduced performance gap among clients (59.11% to 68.75%) indicates improved consistency and personalization. These findings validate our dual approach of heterogeneity-aware curriculum learning and standardized distillation, enabling client-level improvements to translate effectively into server gains. The consistent performance gains across different heterogeneity levels further demonstrate the robustness and adaptability of our proposed HiPer-FedKD framework in diverse federated learning scenarios.

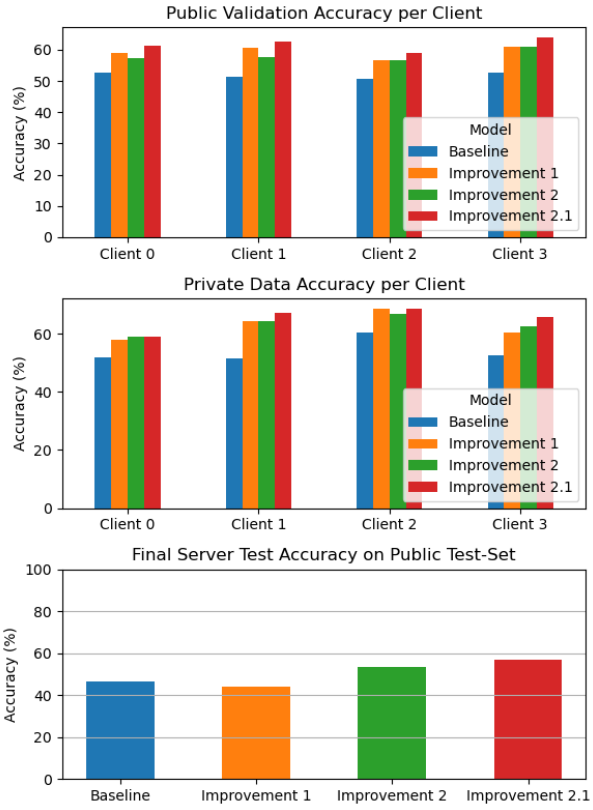

Figure 3: Visualization of the Results (at Dirichlet parameter ALPHA = 1.0)

5 Discussion and Conclusion

5.1 Challenges and Future Considerations

Assumptions of Public Data Availability: A significant limitation in our current HiPer-FedKD framework and many existing federated knowledge distillation approaches is the reliance on public datasets for effective knowledge transfer. This assumption often doesn't hold in practical scenarios where truly representative public data is scarce or unavailable, particularly in specialized domains like healthcare or finance where privacy concerns are paramount. The performance of knowledge distillation methods can degrade substantially when the available public data differs significantly from the private client data distributions [19]. Future research should explore synthetic data generation techniques that can create representative proxy datasets while preserving privacy guarantees [1]. Alternatively, developing distillation methods that can operate effectively with minimal or no public data through techniques such as model-generated pseudo-samples or self-distillation frameworks would substantially increase the practical applicability of HiPer-FedKD [11]. This challenge represents a critical gap between theoretical federated learning frameworks and their real-world deployments where public data assumptions rarely hold.

Scalability and System Heterogeneity in Federated Learning: While HiPer-FedKD effectively addresses statistical heterogeneity through personalized knowledge distillation and adaptive temperature scaling, system heterogeneity remains a significant hurdle for real-world deployment. Clients in federated environments often have varying computational capabilities, memory constraints, and communication bandwidths [18]. It is worth noting that our HiPer-FedKD approach offers greater flexibility than standard parameter-sharing methods since we utilize logit-based knowledge transfer, which means model heterogeneity and system heterogeneity challenges can be handled with only minor adaptations to the framework [2]. Future research should explore integrating model compression techniques with heterogeneity-aware knowledge distillation to create architecturally flexible frameworks. Specifically, developing adaptive model pruning mechanisms that consider both statistical and system heterogeneity could allow resource-constrained clients to participate effectively without compromising performance [8]. This would extend HiPer-FedKD's personalization capabilities beyond client heterogeneity to system-level constraints.

Privacy-Utility Trade-offs in Knowledge Distillation: Although knowledge distillation offers privacy benefits by sharing only model outputs rather than gradients, the soft probability distributions used in distillation may still leak sensitive information about client data characteristics. While our HiPer-FedKD framework focuses primarily on addressing heterogeneity challenges rather than enhancing privacy protections, the fundamental tension between knowledge transfer fidelity and privacy guarantees requires further investigation [13]. Additionally, exploring methods based on privacy metrics to measure and reduce information leakage during bidirectional knowledge transfer is an important direction for future research. [5]. These advances would help bridge the gap between performance optimization and privacy preservation in federated knowledge distillation systems operating in heterogeneous environments.

5.2 Final Remarks

This paper introduced HiPer-FedKD, a novel personalized federated KD framework that addresses the critical challenges of non-i.i.d. data in federated learning environments. By implementing personalized temperature scaling mechanisms that adapt to client-specific data distributions, our approach effectively bridges the gap between global model generalization and client personalization. Comprehensive empirical evaluation demonstrated that HiPer-FedKD significantly outperforms baseline federated learning approaches across various degrees of data heterogeneity, with particularly pronounced improvements in extreme non-i.i.d. scenarios. The bidirectional knowledge transfer approach enables simultaneous improvement of both server and client generalization, representing a significant advancement over previous methods that typically sacrifice one objective for the other. While challenges remain in addressing system heterogeneity, strengthening privacy guarantees, and overcoming public data availability assumptions, HiPer-FedKD provides a solid foundation for future research in heterogeneity-aware federated learning that can thrive in the complex environments characterizing real-world applications.

References

[1] M. R. Behera, S. Upadhyay, S. Shetty, S. Priyadarshini, P. Patel, and K. F. Lee. 2022. FedSyn: Synthetic Data Generation using Federated Learning.

[2] D. Gao, X. Yao, and Q. Yang. 2022. A Survey on Heterogeneous Federated Learning.

[3] C. He, M. Annavaram, and S. Avestimehr. 2020. Group Knowledge Transfer: Federated Learning of Large CNNs at the Edge.

[4] G. Hinton, O. Vinyals, and J. Dean. 2015. Distilling the Knowledge in a Neural Network.

[5] M. Jagielski, M. Nasr, C. Choquette-Choo, K. Lee, and N. Carlini. 2023. Students Parrot Their Teachers: Membership Inference on Model Distillation.

[6] K. Lee, S. Kim, and J. Ko. 2024. Improving Local Training in Federated Learning via Temperature Scaling.

[7] L. Li, J. Gou, B. Yu, L. Du, Z. Yi, and D. Tao. 2024. Federated Distillation: A Survey.

[8] X. Liu, T. Ratnarajah, M. Sellathurai, and Y. C. Eldar. 2024. Adaptive Model Pruning and Personalization for Federated Learning over Wireless Networks.

[9] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas. 2023. Communication-Efficient Learning of Deep Networks from Decentralized Data.

[10] X. Pang, J. Hu, P. Sun, J. Ren, and Z. Wang. 2024. When Federated Learning Meets Knowledge Distillation: An Experimental Study.

[11] B. Pfitzner and B. Arnrich. 2022. DPD-fVAE: Synthetic Data Generation Using Federated Variational Autoencoders With Differentially-Private Decoder.

[12] P. Qi, X. Zhou, Y. Ding, Z. Zhang, S. Zheng, and Z. Li. 2023. FedBKD: Heterogenous Federated Learning via Bidirectional Knowledge Distillation for Modulation Classification in IoT-Edge System.

[13] T. Qi, F. Wu, C. Wu, et al. 2023. Differentially private knowledge transfer for federated learning.

[14] E. Shang, H. Liu, Z. Yang, J. Du, and Y. Ge. 2023. FedBiKD: Federated Bidirectional Knowledge Distillation for Distracted Driving Detection.

[15] J. Shin, Y. Li, Y. Liu, and S.-J. Lee. 2022. FedBalancer: Data and Pace Control for Efficient Federated Learning on Heterogeneous Clients.

[16] S. Sun, W. Ren, J. Li, R. Wang, and X. Cao. 2024. Logit Standardization in Knowledge Distillation.

[17] C. Wu, F. Wu, L. Lyu, Y. Huang, and X. Xie. 2022. FedKD: Communication Efficient Federated Learning via Knowledge Distillation.

[18] H. Zhou, T. Lan, G. Venkataramani, and W. Ding. 2022. On the Convergence of Heterogeneous Federated Learning with Arbitrary Adaptive Online Model Pruning.

[19] Z. Zhu, J. Hong, and J. Zhou. 2021. Data-Free Knowledge Distillation for Heterogeneous Federated Learning.