
Rethinking the Temperature for Federated Heterogeneous Distillation

Fan Qi¹ Daxu Shi¹ Chuokun Xu² Shuai Li¹ Changsheng Xu³

Abstract

Federated Distillation (FedKD) relies on lightweight knowledge carriers like logits for efficient client-server communication. Although logit-based methods have demonstrated promise in addressing statistical and architectural heterogeneity in federated learning (FL), current approaches remain constrained by suboptimal temperature calibration during knowledge fusion. To address these limitations, we propose **ReT-FHD**, a framework featuring: 1) Multi-level Elastic Temperature, which dynamically adjusts distillation intensities across model layers, achieving optimized knowledge transfer between heterogeneous local models; 2) Category-Aware Global Temperature Scaling that implements class-specific temperature calibration based on confidence distributions in global logits, enabling personalized distillation policies; 3) Z-Score Guard, a blockchain-verified validation mechanism mitigating 44% of label-flipping and model poisoning attacks. Evaluations across diverse benchmarks with varying model/data heterogeneity demonstrate that the ReT-FHD achieves significant accuracy improvements over baseline methods while substantially reducing communication costs compared to existing approaches. Our work establishes that properly calibrated logits can serve as self-sufficient carriers for building scalable and secure heterogeneous FL systems.

1 Introduction

In cross-device Federated Learning (FL), critical challenges arise from model heterogeneity (e.g., diverse architectures

across devices), statistical heterogeneity (non-IID data distributions across clients), prohibitive communication costs, and privacy risks due to exposure of sensitive visual data. Integrating knowledge distillation (KD) into FL frameworks addresses these issues by enabling architecture-agnostic knowledge transfer from heterogeneous local models to a lightweight global model, minimizing direct data transmission, and preserving data privacy through indirect gradient-free knowledge aggregation.

Early federated knowledge distillation frameworks, for example, the FedKD (Jeong et al., 2018) transmit mean logit vectors from clients to a server. These vectors are averaged into global consensus outputs and redistributed to clients, where they serve as teacher signals for distillation regularization during local training. Addressing the model heterogeneity, FedMD (Li & Wang, 2019) proposes the clients communicate based on their output class scores on samples from the public dataset. Later, additional knowledge carriers are transferred for resolving *model heterogeneity*, such as features (He et al., 2020; Wang et al., 2024) and proxy models (Lee et al., 2024). However, transferring additional information, whether as proxy models (Wang et al., 2025) or synthetic data (Huang et al., 2024; Liang et al., 2024), introduces both security risks and significant computational and communication overhead.

In this paper, we investigate whether the minimal carrier of knowledge—logits—can effectively address model heterogeneity in FL without public datasets and additional information. Recently, multi-level distillation, i.e., applying knowledge distillation at intermediate layers, has been demonstrated to enhance both the comprehensiveness and coherence of knowledge transfer (Khan et al., 2024; Hao et al., 2024; Huang et al., 2025b). However, the temperature, which is a critical hyperparameter in distillation, is typically simplified to static or monotonically decreasing (Hao et al., 2024; Huang et al., 2025b). This temperature strategy fails to adequately address the intricacies of data and model heterogeneity in FL, significantly limiting the efficiency and flexibility of distillation across heterogeneous models.

Therefore, we rethink the temperature for distillation and propose **ReT-FHD** framework in federated heterogeneous learning from theoretically and empirically: 1) *Model Heterogeneity*: We incorporate the elastic temperature that as-

¹School of Computer Science and Engineering, Tianjin University of Technology, Tianjin, China ²School of Computer Science and Engineering, Southeast University, Nanjing, China ³Institute of Automation, Chinese Academy of Sciences, Beijing, China. Correspondence to: Chuokun Xu <chuokunxu@seu.edu.cn>.

signs flexible temperatures to multi-level distillation. Theoretically, temperature scheduling balances staged logit entropy, enhancing distillation precision and inter-stage knowledge transfer. 2) *Data Heterogeneity*: We present global temperature scaling, which assigns category-specific temperatures, aligning with gradient disparity dynamics to mitigate class-wise update imbalances. Analysis shows this stabilizes convergence by harmonizing gradient variations. 3) *Security*: Integrated with blockchain-empowered FL (BC-FL), our framework employs Z-score verification to validate logit distributions against dynamic boundaries, enabling automated reward/punishment protocols that deter malicious behaviors. 4) Experiments across four benchmarks demonstrate ReT-FHD’s superiority over centralized and decentralized (DeL) baselines.

2 Preliminaries

2.1 Federated Distillation

Federated Distillation (FD) research addresses three core challenges: 1) *Data Heterogeneity*: Orthogonal efforts focus on refining knowledge carriers for non-IID data. Itahara et al. (2021) mitigate statistical heterogeneity via entropy reduction aggregation, while Zhang et al. (2022) synthesize hard samples to amplify KD effectiveness in data-free FL. 2) *Model Heterogeneity*: Foundational works include FedMD (Li & Wang, 2019), which aligns logits from public datasets for consensus, and FedDF (Lin et al., 2020), which ensembles soft labels from unlabeled data. FedGKT (He et al., 2020) further bridges model capacity gaps by distilling compact client features into a large server model. 3) *Security*: Chang et al. (2019) minimizes leakage via robust mean estimation on public predictions, Gong et al. (2022) employs quantized noise to perturb shared logits.

To defend model poisoning attacks, Khan et al. (2024) propose HYDRA-FL by offloading some of the KD loss to a shallow layer via an auxiliary classifier for local-to-local distillation. In contrast, our framework transfers logits instead of model parameters, introducing additional challenges due to compounded heterogeneity in both data distributions and model architectures.

2.2 Model-Heterogeneous Federated Learning

For generalized scenarios of heterogeneous models, two mainstream methods are *Prototype-based methods* and *Proxy model methods*. The *prototype-based* framework primarily facilitates knowledge transfer through client collaboration in constructing a unified class prototype. Tan et al. (2022) propose FedProto, a prototype-based method that addresses heterogeneity challenges in FL. Clients communicate local abstract class prototypes, which the server aggregates into a global prototype, eliminating model param-

eter communication. Zhang et al. (2024a) propose FedTGP, a framework using Adaptive Boundary Contrastive Learning (ACL) to learn trainable global prototypes (TGP) on the server, improving prototype separability while maintaining semantic meaning. On the other hand, the *proxy model* method focuses on training a compact model at each participant as a "proxy" for the primary model. Shen et al. (2020) propose FML, a federated learning framework enabling joint training of a universal model and personalized models, allowing custom designs for diverse applications beyond standard FL limitations. Wu et al. (2022) propose FedKD, a federated learning framework enabling mutual learning of client student-teacher models and student model sharing for knowledge transfer, alongside a dynamic gradient approximation method to reduce communication costs. Moreover, Wang et al. (2025) propose FedType, a bidirectional knowledge distillation strategy to optimize private and proxy models jointly, along with an uncertainty-based behavioral imitation method to improve proxy model guidance.

While the aforementioned methods address the challenges of Heterogeneous Federated Learning, the resulting increase in communication and computational demands poses significant limitations to the practical deployment of federated systems.

2.3 Dynamic Temperature Knowledge Distillation

Dynamic temperature strategies in knowledge distillation (KD) can be broadly divided into two categories: *Global scheduling methods* and *Local adaptive methods*. Global scheduling methods adjust a single temperature to control overall distillation difficulty. Liu et al. (2022) propose to learn the temperature by performing meta-learning on the extra validation set. Li et al. (2022b) dynamically adjust a shared temperature via curriculum learning to progressively increase distillation difficulty from easy to hard. Local adaptive methods assign temperatures at finer granularity. Guo et al. (2023) adjust sample-wise temperatures based on teacher confidence, while Wei & Bai (2024) minimize sharpness differences between teacher and student logits to derive separate temperatures for each.

Although these methods emphasize the effect of temperature on the distillation process, they are not suitable for knowledge distillation of heterogeneous models in FL by multi-client non-iid data.

3 Proposed Method

3.1 Background

Multi-level Knowledge Distillation: Compared to the original Federated Distillation approach (Li & Wang, 2019), Hybrid Knowledge Distillation (Khan et al., 2024; Hao et al.,

2024) has been shown to significantly enhance robustness and accuracy in Federated Learning. This improvement is achieved by integrating an auxiliary knowledge distillation constraint within a shallow layer of the local model. The hybrid loss function for FL clients is formulated as following:

$$\mathcal{L} = \mathcal{L}_{\text{CE}}(y_c, y) + \sum_{l=1}^L \mathcal{L}_{\text{KL}}\left(\sigma\left(\mathbf{z}_t^{(l)}, \tau\right), \sigma\left(\mathbf{z}_s^{(l)}, \tau\right)\right) \quad (1)$$

Here, σ denotes the softmax function, $\mathbf{z}_s^{(l)}$ denotes the logits of the student model in layer l , $\mathbf{z}_t^{(l)}$ denotes the logits of the teacher model in layer l . τ is called temperature and serve as the offsets of the softmax function, α and τ effectively adjusts the sharpness or softness of the probability distribution derived from the teacher's logits. The loss function contains two main elements:

- $\mathcal{L}_{\text{CE}}$ (Cross-Entropy Loss) is used to optimize the classification ability of the client model for good task performance.
- $\sum_{l=1}^L \mathcal{L}_{\text{KL}}^{(l)}$ (Multi-level Distillation Loss) is used to calculate the KL divergence between the teacher logits and the student logits at each level l to ensure that knowledge at different levels is adequately transferred.

Z-Score Distillation: Sun et al. (2024) propose a logit Z-Score standardization as a pre-process to enable student to focus on essential logit relations from teacher rather than requiring a magnitude match. To bridge the capacity gap of heterogeneous clients, we rewrite softmax based on the mean $\bar{z}$ and variance $\text{Var}(\cdot)$ as follows:

$$\sigma(\mathbf{z}; \bar{\mathbf{z}}, \text{Var}(\mathbf{z}))^{(c)} = \frac{\exp\left(\mathcal{Z}(\mathbf{z}; \tau)^{(c)}\right)}{\sum_{n=1}^{\mathcal{N}} \exp\left(\mathcal{Z}(\mathbf{z}; \tau)^{(n)}\right)}, \quad (2)$$

$$s.t. \mathcal{Z}(\mathbf{z}; \tau)^{(c)} = \frac{z^{(c)} - \bar{z}}{\text{Var}(\mathbf{z}) \tau}$$

where $z^{(c)}$ represents the c -th term of $\mathbf{z}$, which is the c -th class specific logits value, and $\mathcal{Z}$ is the Z-score function.

3.2 Rethink the Temperature in Heterogenous Federated Distillation

Despite the effectiveness against poisoning attacks (Kundu et al., 2021; Khan et al., 2024), the temperatures τ in Eq.1 are fixed, which neglects the inherent data and model heterogeneity present in the distillation process. Typically, higher τ makes the distribution sharper (more confident), while lower τ makes it softer (less confident), thus influencing the gradient and the learning process of the student model. However, in heterogeneous federated distillation, both data

and model heterogeneity necessitate finer temperature regulation to enhance the utilization effectiveness of global logits.

Multi-level Elastic Temperature : To address model heterogeneity, we dynamically adjust the temperature during knowledge distillation based on the hierarchical characteristics of the models, effectively bridging representation discrepancies across architectures at different levels. Our multi-level distillation framework employs stage-adaptive temperature scaling: each level's output features feed into the subsequent level, enabling dynamic temperature calibration based on the statistical divergence between successive stages. Specifically, if the Z-score difference between the logits of the next stage and the logits of the teacher's model is large, the temperature of the current stage should be increased to reduce the amount of knowledge in the logits of the current stage, and thus promote learning more information from the logits of the teacher's model. Conversely, when the difference in logits is small, lower the temperature so that the model focuses more on learning in other phases.

We rewrite τ_l in level l as $\tau_l = G_l(\Delta Z)$, $\Delta Z_l = \|Z_T - Z_{l+1}\|$. Z_T is the $Z(\cdot)$ distribution of the global logits. ΔZ_l represents the difference in Z-Scores between the global logits and the local logits, which is crucial for adjusting the temperature of the distillation process. The modulation of the effect of ΔZ is achieved through the function G , defined as a logarithmic function, as follows:

$$G_l(\cdot) = \xi + \frac{\gamma \log(1 + \Delta Z_l)}{\log(1 + \Delta Z_{\max})}, \quad (3)$$

where γ is a scaling factor that determines the sensitivity of the function, ξ is the lower bound of the temperature adjustment represented by the initial value of the temperature, and $\Delta Z_{\max}$ is used as a normalisation factor to determine the sensitivity of the function.

Category-Aware Global Temperature Scaling: To address data heterogeneity, we calculate distinct temperatures for each category in the global logits, enabling flexible adaptation to diverse client data characteristics, enhancing robustness and inter-category fairness while mitigating challenges from heterogeneous data distributions. Here, we give the formula for adjusting the temperature of the c -th category of global logits:

$$\tilde{\tau}_c = \tilde{\tau}_c - \beta \cdot \frac{\mathcal{L}_{\text{CE}}(f(\tilde{x}_c)) - \frac{1}{C} \sum_{i=1}^C \mathcal{L}_{\text{CE}}(f(\tilde{x}_i))}{\max\left(\left|\mathcal{L}_{\text{CE}}(f(\tilde{x}_c)) - \frac{1}{C} \sum_{i=1}^C \mathcal{L}_{\text{CE}}(f(\tilde{x}_i))\right|\right)} \quad (4)$$

where β controls the weight of the temperature update, $\max(|\cdot|)$ represents the absolute maximum value taken to control the stability of the update. Since cross entropy directly reflects the robustness of the category, it is chosen to regulate the temperature update.

Algorithm 1 Algorithm of our ReT-FHD

Input: Dataset $\{D_1, D_2, \dots, D_K\}$, $T \gg T$: Maximum communication rounds
 1: **Initialization:** $\mathcal{W} \leftarrow$ Worker, $\mathcal{V} \leftarrow$ Validator, $\mathcal{M} \leftarrow$ Miner, Global logits z^g ,
 Blacklist $\mathcal{B}$, Client neighbors set $\mathcal{N}$, Honest server $\mathcal{S}$, Client reward set $\mathcal{R}$
 2: **for** $t = 1 \rightarrow T$ **do**
 3: each $k \in \mathcal{K} \leftarrow \text{Randomly}\{\mathcal{W}, \mathcal{V}, \mathcal{M}\}$, each $k \in \mathcal{K} \leftarrow \{\mathcal{W}\}$
 4: **for** each $w \in [\mathcal{W}]$ **do**
 5: w downloads z_{t-1}^g from blockchain $\rightarrow$ **do** local update with Eq.6
 6: $\rightarrow$ sends z_t^w to $v \in \mathcal{V}$
 7: w downloads z_{t-1}^g from $\mathcal{S}$ and **do** local update with Eq.5 $\rightarrow$ sends z_t^w to $\mathcal{S}$
 8: w gathers z_{t-1}^n from $n \in \mathcal{N}_w$
 9: **end for**
 10: **for** each $v \in \mathcal{V}$ or $w \in \mathcal{W}$ or $\mathcal{S}$ **do**
 11: v evaluates $w \in \mathcal{W}$ with z_{t-1}^g by Eq.7
 12: $\rightarrow$ Auto set r for each $w \in \mathcal{W}$ by Eq.8
 13: v votes $w \in \mathcal{W} \rightarrow$ sends to $m \in \mathcal{M}$
 14: $\mathcal{S}$ evaluates $w \in \mathcal{W}$ by Eq.7
 15: $\rightarrow$ set r_k for each w by Eq.8 $\rightarrow$ gather z_{t+1} by $r_k \in \mathcal{R}$
 16: w evaluates z_{t-1}^n by Eq.7 $\rightarrow$ set r_w for each $n \in \mathcal{N}_w$ by Eq.8
 17: $\rightarrow$ gather z_t^w by $r_w \in \mathcal{R}$
 18: $\rightarrow$ **do** local update with Eq.6 $\rightarrow$ sends z_t^w to $n \in \mathcal{N}_w$
 19: **end for**
 20: $m \in \mathcal{M}$ aggregates z_t according to the weight of $\mathcal{R}$, makes $\mathcal{B}$ by $\sum_{i=1}^v \mathcal{V}^o$
 21: $\rightarrow$ mines a block $\rightarrow$ uploads $\mathcal{B}$, $\mathcal{R}$, z_t^g to blockchain
 22: **end for**

At the beginning of each training round, the client receives the global logits and then first adjusts the global logits according to the temperature $\tilde{\tau}$ calculated at the end of the previous round.

Local update: Ultimately, the FL optimization function can be expressed as:

$$\begin{aligned}
 \min_w \sum_{i=1}^K \frac{|D_i|}{|D|} \mathbb{E} \left[\mathcal{L}_{\text{CE}} + \sum_{l=1}^L \sum_{c=1}^C \mathcal{L}_{\text{KL}}^{(l)}(\sigma^t(\cdot), \sigma^s(\cdot)) \right] \\
 \text{s.t. } \sigma^t(\cdot) \sim \sigma(\mathbf{z}_t; \tilde{z}_t, \text{Var}(\mathbf{z}_t), G(\Delta Z), \tilde{\tau}_c), \\
 \sigma^s(\cdot) \sim \sigma(\mathbf{z}_s; \tilde{z}_s, \text{Var}(\mathbf{z}_s), G(\Delta Z), \cdot)
 \end{aligned} \quad (5)$$

where $\tilde{\tau}_c$ is the isolated temperature for category c in the global logits.

The DeL optimization function can be expressed as:

$$\begin{aligned}
 \theta_i^{t+1} = \theta_i^{t+\frac{1}{2}} - \sum_{j \in \mathcal{N}(i)} \sum_{l=1}^L \mathcal{L}_{\text{KL}}^{(l)}(\sigma^t(\cdot), \sigma^s(\cdot)) \\
 \text{s.t. } \sigma^t(\cdot) \sim \sigma(\mathbf{z}_t; \tilde{z}_t, \text{Var}(\mathbf{z}_t), G(\Delta Z), \tilde{\tau}), \\
 \sigma^s(\cdot) \sim \sigma(\mathbf{z}_s; \tilde{z}_s, \text{Var}(\mathbf{z}_s), G(\Delta Z), \cdot)
 \end{aligned} \quad (6)$$

3.3 Different FL Modes of Our ReT-FHD

The ReT-FHD framework supports distributed training through multilevel elastic distillation and statistical logit analysis, enabling intrinsic security and scalable FL deployment.

FL and DeL mode: In the Algorithm 1, the situation will not be so complicated for federated or decentralized learning. Each node belongs to the $w \in \mathcal{W}$. For FL mode, each client computes a multilayer distillation loss function with the server-side global logits, highlighted in yellow. For DeL mode, multilayer distillation training is performed after verifying the legitimacy of neighboring node logits, highlighted in pink.

BC-FL mode: Following the blockchain FL(Chen et al., 2021; Ying et al., 2024), we apply the POS voting-based consensus mechanism, highlighted in purple of Algorithm 1. In each round t , each $k \in \mathcal{K}$ is randomly assigned a role, including $w \in \mathcal{W}, v \in \mathcal{V}, m \in \mathcal{M}$, where $|\mathcal{W}| + |\mathcal{V}| + |\mathcal{M}| = |\mathcal{K}|$. w is responsible for performing computations related to f_i^w . Each w download global logits z_i from the latest block and train with private local data D_i . At this point the w node performs the E round local update and broadcasts the logits z_i^w , corresponding to the number of dimensions of the dataset categories, to the $v \in \mathcal{V}$ in the network. See the Appendix for more details.

Z-Score Guard: *Standardized Finding of Malicious Logits:* $\mathcal{Z}(\cdot, \cdot)$ plays a crucial role in evaluating logits for heterogeneous client models. This property likewise suggests that $\mathcal{Z}(\cdot, \cdot)$ can also play an important role in blockchain authentication sessions. In the verification process, validators $\mathcal{V}$ (verifiers) train locally and upload logits, while examinees $\mathcal{E}$ (validators) evaluate $\mathcal{V}$'s logits. Under data heterogeneity, label sets across nodes may partially overlap or diverge. We define the labeling relationship among nodes as: $|\mathcal{V} \cap \mathcal{H}| = \eta$, $|\mathcal{E} \cap \mathcal{H}| = \phi$, and $|\mathcal{V} \cap \mathcal{E}| = \psi$. Here, $\mathcal{H}$ represents the complete set of labels for a distributed classification task.

The importance of ψ , η and ϕ arises particularly if $\mathcal{E}$ acts as a lazy node by re-uploading $\mathcal{Z}(z_g, \cdot)$, which impedes the verification process by $\mathcal{V}$. To facilitate this verification, we suggest plotting both $\mathcal{Z}(z_g, \tau)^{(\psi)}$ and $\mathcal{Z}(z_v, \tau)^{(\psi)}$ within a coordinate system:

$$\begin{aligned}
 \mathcal{D}_{\mathcal{E}} = \frac{\sum_{i=1}^H \left[\exp \left((\Delta Z_{\mathcal{E}, \mathcal{V}}^i)^{\psi} \right) + \exp \left((\Delta Z_{\mathcal{E}, g}^i)^{\phi} \right) \right]}{\sum_{i=1}^H \exp \left((\Delta Z_{\mathcal{V}, g}^i)^{\eta} \right)} \\
 > 2 \sum_{i=1}^H \exp \left(\mathcal{Z}(z_{\mathcal{E}}^i, \tau)^{(\psi)} - \frac{\mathcal{Z}(z_g^i, \tau)^{(\psi)} + \mathcal{Z}(z_v^i, \tau)^{(\psi)}}{2} \right) \\
 \text{s.t. } \Delta Z_{a,b}^i = \mathcal{Z}(z_a^i, \tau) - \mathcal{Z}(z_b^i, \tau)
 \end{aligned} \quad (7)$$

Here τ is a fixed temperature. This approach allows us to quantify the criteria necessary for a verified individual to be considered valid or to be subject to voting, based on their positional relationship within the coordinate system. When $Z(z_{\mathcal{E}}, \cdot)$ equilibrium between $Z(z_{\mathcal{V}}, \cdot)$ and $Z(z_{\mathcal{G}}, \cdot)$ is reached, v will evaluate highest scores and vote bad.

Logits-based Secure Validation: Logits, as a crucial medium for information transmission, can also indicate the confidence level of a sample, in particular, the confidence level can demonstrate the knowledge of the model and also reveal information about the security of the model, which can be solved by looking at the logits relationship in the distributed system. We dynamically assign rewards based on this confidence. Where R_b is the base reward, $\|D_i\|$ represents the number of samples per client, λ is the preset confidence range, we provide a base algorithm for measuring reward values as follows:

$$\mathcal{R} = \begin{cases} R_b \cdot \|D_i\| \cdot \max\{P(y = i | x)\}, & \text{if } \max\{P(\cdot)\} > \lambda \\ R_b \cdot \|D_i\| \cdot (1 - \max\{P(y = i | x)\}), & \text{otherwise} \end{cases} \quad (8)$$

4 Theoretical Analysis

In this section, we systematically demonstrate the effectiveness of the proposed Multi-level Elastic Temperature and Global Temperature Scaling mechanism from the perspectives of information entropy and gradient updating.

Referring to Zhao et al. (2024), we introduce the knowledge scale to quantify the level of knowledge of the predictive distribution, which can measure the scale of knowledge learnt from the probability distribution by the model with randomly initialized weights:

$$K^T = KL(I(x), T(x)) = \log C - H(P^T) \quad (9)$$

where $KL(\cdot)$ stands for Kullback–Leibler divergence, which calculates the cost of information required to travel from one distribution to another, C is the total number of categories and is a constant for the same dataset, and $H(\cdot)$ is the information entropy.

From the definition of knowledge scale, it can be found that the information entropy can be directly used to measure the knowledge level of teachers. Therefore, we obtain the following theorem by studying the partial derivative of information entropy with respect to temperature.

Theorem 4.1. *Temperature τ is positively correlated with the information entropy of the predictive distribution of logits.*

The proof of Theorem 4.1 can be found in Appendix A.1. Therefore, combining Theorem 4.1 and the definition of knowledge scale, the larger the temperature, the smaller the

knowledge scale, and the more local logits can learn from the global logits that can be learned.

In order to alleviate the problem of uneven distribution of category information due to data heterogeneity, we further introduce Global Logits Scaling, but before that, we need to discuss the effect of the previous Multi-level Elastic Temperature on categories.

Corollary 4.2. *For a model f with random distribution I in the initial state, for a deep neural network (DNN) model f with random distribution I in the initial state, the model has no preference for any sample. Thus for the model's predictive distribution $f(x; \theta_I) = \{p_1^I, p_2^I, \dots, p_c^I\}$, the following relationship exists:*

$$\mathbb{E}(p^I(x)) = \mathbb{E}(p_1^I(x)) = \mathbb{E}(p_2^I(x)) = \dots = \mathbb{E}(p_c^I(x)) \quad (10)$$

We assume that the model is optimised under the guidance of a soft label distribution with smoothing degree λ_1 , which is satisfied:

$$\begin{aligned} p_1^{\lambda_1}(x_1) &= p_2^{\lambda_1}(x_2) = \dots = p_c^{\lambda_1}(x_c) > \frac{1}{C} > \\ p_1^{\lambda_1}(x_{c \neq 1}) &= p_2^{\lambda_1}(x_{c \neq 2}) = \dots = p_c^{\lambda_1}(x_{c \neq c}) \end{aligned} \quad (11)$$

The following relationship exists for the partial derivatives of the knowledge scale for different classes of logits:

$$\frac{\partial KL(f(x; \theta_I), P_{\lambda_1})}{\partial z_1(x)} = \dots = \frac{\partial KL(f(x; \theta_I), P_{\lambda_1})}{\partial z_c(x)} \quad (12)$$

Corollary 4.2 shows that the overall change in knowledge scale has a balanced impact on different categories. The proof of Corollary 4.2 can be found in Appendix A.2.

Now we can discuss the theory of Global Logits Scaling. Here, we give Theorem 4.3 about the relationship between the category temperature of global logits and the category gradient update of local models.

Theorem 4.3. *Following the setup in Corollary 4.2, the temperature $\tilde{\tau}_c$ of category c is positively correlated with the gradient update $\nabla_{f \sim \lambda}^c$ of the local model category c .*

The proof of Theorem 4.3 can be found in Appendix A.3. Theorem 4.3 states that different temperatures affect the gradient update. Increasing the temperature of category c will cause the gradient update $\nabla_{f \sim \lambda}^c$ of that category to increase, and vice versa will cause the gradient update $\nabla_{f \sim \lambda}^c$ of the category to decrease, and by regulating the temperature of different categories, it will cause the gradient update of different categories to be more balanced.

Table 1. Testing accuracy (%) in FL model heterogeneous (a.) and homogeneous (b.) settings. *Dir-set*: Dirichlet distribution; *Pat-set*: Pathological distribution.

Methods		Tiny ImageNet		Flower102		Cifar-100		Cifar-10		
		Dir-set	Pat-set	Dir-set	Pat-set	Dir-set	Pat-set	Dir-set	Pat-set	
(a.)	Proto-based	FedProto	14.68 ±2.61	24.91±1.38	21.68±1.65	55.12±1.73	21.71±3.03	32.00±2.48	60.48±1.32	75.89±2.29
		FedTGP	14.94±0.51	40.56±1.42	27.15±1.46	68.76±1.42	26.24±0.72	62.08±1.37	68.08±1.06	89.00±1.26
		RethinkFL	15.21±1.65	38.47±0.80	30.36±1.19	65.24±1.42	28.37±0.39	63.11±2.44	68.21±2.16	88.96±3.24
	Proxy-models	FML	14.59±0.54	35.74±1.13	21.63±1.07	52.27±1.31	21.20±0.48	49.55±1.12	61.57±1.07	82.41±0.56
		FedKD	14.57±0.65	32.07±1.01	20.76±1.27	52.71±1.43	19.48±0.58	51.58±1.13	63.32±1.06	83.73±0.51
		FedType	16.05±0.64	38.84±0.92	23.53±0.86	60.89±1.24	24.21±0.31	52.51±1.34	66.43±1.23	84.28±0.52
		HYDRA-FL	14.56±1.62	30.87±1.09	22.27±2.69	55.18±2.24	20.82±1.94	48.50±0.92	61.74±2.71	79.37±1.75
		FedDistill	15.20±0.62	37.67±1.28	20.47±1.31	62.27±1.58	21.04±0.57	53.72±1.17	63.51±1.03	83.08±0.64
	Logits-based	FCCL	15.49±0.27	38.01±0.46	25.20±0.43	65.61±1.17	22.72±0.56	58.62±0.70	64.29±1.37	86.58±0.83
		ReT-FHD(Ours, FL)	17.24±0.53	42.35±0.98	33.16±1.46	72.91±0.97	30.77±0.47	68.56±0.66	72.76±0.77	91.23±0.65
(b.)	Proto-based	FedProto	31.48±1.64	37.12±0.78	53.14±1.02	66.35±1.24	48.01±0.49	63.49±0.58	86.32±1.26	85.04±0.46
		FedTGP	34.08±1.69	39.92±0.77	58.27±0.90	72.66±1.12	49.32±0.83	64.46±0.66	88.38±1.21	89.10±0.51
		RethinkFL	35.04±2.72	37.50±1.11	60.39±0.94	67.95±1.52	48.34±0.26	63.90±1.47	89.33±1.06	88.35±1.94
	Proxy-models	FML	34.13±1.49	37.50±0.79	53.58±1.04	66.21±1.04	45.61±0.56	60.68±0.49	87.56±1.14	87.26±0.50
		FedKD	33.99±1.56	37.97±0.77	53.82±1.05	66.06±1.05	48.62±0.42	63.98±0.46	88.38±1.10	89.05±0.53
		FedType	34.29±0.69	40.74±1.06	60.12±1.15	75.27±2.03	47.38±0.48	64.73±0.93	89.77±1.82	88.91±0.76
		HYDRA-FL	34.45±0.88	44.62±0.73	58.27±1.59	70.63±1.11	51.18±0.90	59.71±1.52	87.04±0.85	85.92±1.06
		FedDistill	33.71±1.58	37.73±0.80	55.67±0.95	65.37±1.05	49.39±0.76	63.51±0.48	88.56±1.13	88.19±0.50
	Logits-based	FCCL	33.95±0.94	38.36±1.83	56.77±0.49	69.75±1.15	51.52±0.78	64.42±0.57	88.29±1.45	87.98±0.16
		ReT-FHD(Ours, FL)	35.43±1.52	50.99±0.48	61.56±0.98	89.66±1.01	51.97±0.72	68.49±0.46	90.54±1.03	89.85±0.57

Table 2. Compared to decentralized methods in four datasets.

Method	Cif-10	Cif-100	Flow.	Tiny.
Local SGD	87.50 ± 1.37	55.47 ± 2.08	62.43 ± 1.13	35.26 ± 2.60
DPSGD	83.01 ± 1.31	40.56 ± 1.24	50.11 ± 1.58	23.57 ± 1.95
CGA	65.65 ± 2.37	30.81 ± 3.82	35.62 ± 2.86	18.85 ± 3.88
Coll-FL	87.14 ± 5.28	50.65 ± 3.52	64.21 ± 4.66	37.45 ± 3.59
L2C	90.14 ± 0.34	59.00 ± 0.42	65.92 ± 0.71	37.95 ± 0.58
DeSA	84.14 ± 0.51	60.12 ± 0.89	66.68 ± 1.64	38.05 ± 0.37
ReT-FHD(Ours, Del)	88.80 ± 5.07	64.11 ± 5.51	67.46 ± 5.16	38.21 ± 5.04

5 Experiment

5.1 Experimental Setup

Datasets and Heterogeneity Setting

Datasets: Cifar10 (Krizhevsky & Hinton, 2009), Cifar100 (Krizhevsky & Hinton, 2009), Tiny-ImageNet (Chrabaszcz et al., 2017), and Flower102 (Nilsback & Zisserman, 2008). All datasets are taken as 75% as training set and 25% as test set. *Data Heterogeneity*: For each dataset, we apply two main types of Non-IID setting (Zhang et al., 2024b): 1) Dirichlet distribution, we set $\alpha = \{0.1, 0.5\}$, it’s worth noting that α is the concentration parameter and smaller α corresponds to stronger data heterogeneity; 2) Pathological Setting, each client contains a fixed number of sample categories, but each client has a different sample category. *Model Heterogeneity*: We adopt the heterogeneous model setup from (Zhang et al., 2021), including AlexNet (Krizhevsky et al., 2017), ShuffleNetV2 (Ma et al., 2018), ResNet18 (He et al., 2016) and GoogleNet (Szegedy et al., 2015), which are randomly and evenly assigned to clients. For the DeL experimental setup, we use only 2-layer CNNs and apply two-stage elastic distillation for homogeneous models (Li et al., 2022a).

Baselines and Implementation

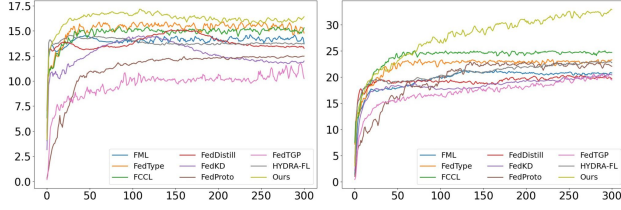
FL setting: We select eight baselines, excluding FedProto (Tan et al., 2022), FedTGP (Zhang et al., 2024a), RethinkFL (Huang et al., 2023), FML (Shen et al., 2020), FedKD (Wu et al., 2022), FedType (Wang et al., 2025), FedDistill (Jeong et al., 2023), HYDRA-FL (Khan et al., 2024) and FCCL (Huang et al., 2022). *DeL setting*: We add DPSGD (Lian et al., 2017), CGA (Esfandiari et al., 2021), Coll-FL (Zhu et al., 2023), L2C (Li et al., 2022a) and DeSA (Huang et al., 2025a) in a DL experiment. We evaluate the superiority of the algorithm by evaluating the average of all client accuracy and standard deviations throughout the training process. All experiments are implemented in Pytorch (i.e. Intel(R) Xeon(R) Platinum 8176 CPU @ 2.10GHz with NVIDIA GeForce RTX 3090 GPU).

Security Settings

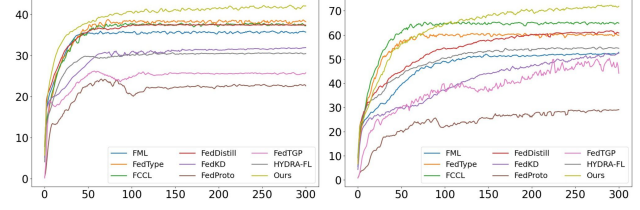
To test the reliability of validation in a blockchain-based FL setting, we simulate malicious nodes through three methods: label reversal attacks, model noising to emulate malicious models (Qin et al., 2024), and logits attacks (Khan et al., 2024), evaluating robustness under these adversarial conditions.

5.2 Comparison with State-of-the-Art Methods

Heterogeneous model. We conduct experiments across four datasets and two data partitioning schemes, with results detailed in Tab.1(a.). Our method consistently outperforms others by at least 2% across all experimental settings, with a 6.01% accuracy gain over FedTGP on Flower102 (Dirichlet 0.5). This highlights our temperature strategy’s effectiveness in optimizing client gradient descent and enhancing knowledge transfer across heterogeneous models. Moreover, as shown in Fig.1a, our method achieves optimal



(a) Accuracy curves for Dir (0.1) distribution tested on Tiny-ImageNet (left), Flower102 (right).



(b) Accuracy curves for PAT distribution tested on Tiny-ImageNet (right), Flower102 (left).

Figure 1. Accuracy curves for PAT distribution tested on Tiny-ImageNet, Flower102.

Table 3. The testing accuracy (%) of heterogeneous model combination on Cifar-10. $\mathcal{A}$. for AlexNet, $\mathcal{R}$. for ResNet18, $\mathcal{G}$. for googleNet, $\mathcal{S}$. for ShuffleNetv2

	$\mathcal{A}$.	$\mathcal{R}$.	$\mathcal{G}$.	$\mathcal{S}$.	$\mathcal{A} + \mathcal{R}$.	$\mathcal{G} + \mathcal{R}$.	$\mathcal{S} + \mathcal{R}$.	$\mathcal{A} + \mathcal{R} + \mathcal{G}$.	$\mathcal{A} + \mathcal{R} + \mathcal{S}$.	$\mathcal{U}(All)$.
Accuracy	89.28%	89.04%	89.38%	86.69%	88.94%	89.16%	87.84%	88.64%	87.49%	88.79%

Table 4. Ablation studies of our method on four datasets.

Method	Cifar10	Cifar100	Flower102	Tiny.
Single-level & Fixed τ	66.85	27.52	26.97	13.20
Multi-level & Fixed τ	69.93	29.12	30.26	14.56
Multi-level & Elastic τ	70.82	29.45	32.05	15.14
ReT-FHD(Ours, FL)	72.76	30.77	33.16	17.24

results with fewer communication rounds, reaching near-peak performance on Tiny-ImageNet in under 80 rounds. In Fig.1b, while our method exhibits slower initial growth on Flower102, it ultimately surpasses other baselines in later rounds.

Homogeneous model. Tab.1(b) shows our method outperforms leading FL baselines across four ResNet-based datasets, achieving a 19.03% accuracy gain over HYDRA-FL on Flower102 under PAT’s high-variability data. This highlights our framework’s strength in handling feature ambiguity via elastic temperature scaling.

DeL Experiments. Our method employs blockchain for decentralized federated training, eliminating the need for a central server. We compare it with state-of-the-art decentralized methods in Tab.2. On CIFAR-10, our method and DeSA fall short of L2C due to its efficient convergence on simpler datasets. However, on the more complex CIFAR-100, our method achieves the highest accuracy, demonstrating its strength in extracting meaningful patterns. Additionally, blockchain integration enhances privacy protection in distributed communication.

5.3 In-depth Analysis

Ablation studies. We design a series of ablation experiments to verify the effects of Multi-level distillation, Elastic Temperature τ and Global Logits Scaling on model per-

formance. The experiment removes these key components sequentially and observes their impact on the final performance. Specifically, we start from the full model (Ours: Multi-level + Elastic Temperature τ + Global Logits Scaling) and sequentially remove Global Logits Scaling, Elastic Temperature τ , and ultimately simplify it to a single-level fixed temperature (Single-level + Fixed τ) of the simplest version. As shown in Tab.4 and Fig.2(left), the full model performs the best out of all the setups, achieving a performance of 72.75%. At the same time, it shows that Multi-level Distillation can initially improve the quality of distillation, while Elastic Temperature τ and Gloabl Logits Scaling can provide more knowledge in distributed learning compared with Fixed τ .

Evaluation of Distillation Layers Number L . To investigate the impact of the number of distillation layers on model performance, we conduct experiment using the ResNet-18 architecture, varying the number of layers L from 1 to 5. Fig.2(right) shows that increasing distillation layers from 1 to 4 enhances accuracy: CIFAR-10 boosts from 67.21% to 72.75% (+5.54), CIFAR-100 from 28.01% to 30.77% (+2.76), and Tiny-ImageNet from 14.21% to 17.24% (+3.03). These findings reflect diminishing returns with excessive layers, highlighting the need to match distillation layers to model complexity for optimal performance.

Effect of different heterogeneous models on accuracy.

Tab. 3 shows the performance with heterogeneous model combinations on CIFAR-10. Multi-branch distillation achieves optimal accuracy (89.16%) using complementary architectures like GoogleNet ($\mathcal{G}$) + ResNet18 ($\mathcal{R}$). In contrast, combining all models ($\mathcal{U}$) degrades performance (88.79%), underperforming standalone $\mathcal{G}$. Results emphasize strategic model complementarity over quantity for effective FL.

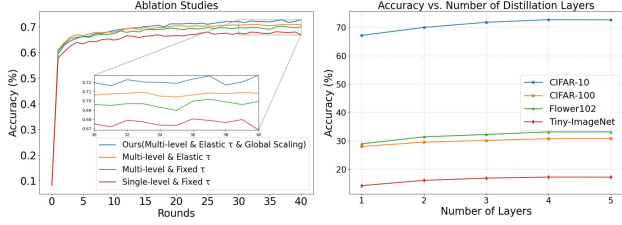


Figure 2. Ablation study (left) and Evaluation of Distillation Layers Number L (right) on Cifar-10.

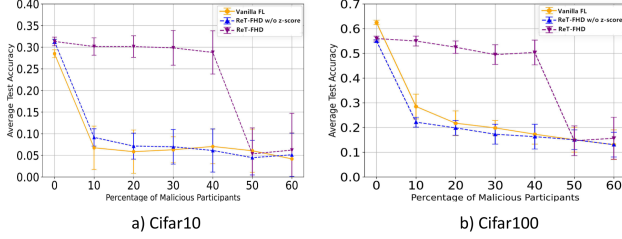


Figure 3. Security testing in a blockchain framework

Analysis of Complexity and Communication Volume. As shown in Tab.5, we analyze the communication upload and download data volumes, as well as the training complexity of all baseline methods. The experiments are conducted on the CIFAR-10 dataset, encompassing four different modalities in a heterogeneous setting (with the proxy model being ResNet-18). Overall, methods based on proxy models generate the highest communication volume due to the need to transmit model parameters. In contrast, our logits distillation method results in the least communication overhead. Regarding computational complexity, proxy model methods incur a significant computational burden during the local knowledge transfer between the teacher and student. In comparison, the prototype-based and logits distillation setups exhibit relatively lower computational demands during local training. In summary, our method outperforms other approaches in terms of both communication volume and computational complexity.

5.4 Security and Robustness Performance

To establish a reliable global logits benchmark, we configure 20 clients, each responsible for independent training on the CIFAR-10 dataset. We aggregate all logits of the same class using a weighted average to obtain the Z-scores for each category. These Z-scores are then visualized, with diagonal lines indicating predictions for the correct classes, demonstrating that the Z-scores for correct predictions are generally the highest.

Malicious Model. The experimental results in Tab.6 analyze the behavior of a malicious node in a simulated distributed network. By injecting noise into a ResNet model’s

Table 5. Comparison of the mean communication cost and amount of computation per round for each approach. “MB” are short for megabyte.

Methods	Comm.(MB)		Computation(Gflops)
	Up.	Down.	
FedProto	0.00256	0.00256	11340
FedTGP	0.00256	0.00256	11680
FML	10.45	10.45	23310
FedKD	10.45	10.45	4620
FedType	10.45	10.45	5160
FedDistill	0.003	0.003	11760
DeSA	0.6	0.6	24680
ReT-FHD(Ours)	0.0003	0.0004	3780

Table 6. Security Evaluation of Logits and Z-scores under Malicious Node and Label Reversal Attacks on CIFAR-10. This table presents the logits and their corresponding Z-scores under two adversarial attack scenarios: (1) malicious nodes injecting anomalous data for the "Airplane" and "Bird" classes, and (2) label reversal attacks flipping the labels between "Deer" and "Dog" classes.

		Malicious model			Label reversal attack		
		Air	Auto	Bird	Deer	Cat	Dog
$\mathcal{L}/C$	Airplane	681.74	248.60	1805.4	Deer	1.7862	0.9026
	Bird	0.0477	-1.0924	0.0016	Dog	1.7653	3.3860
Z_m	Airplane	0.7710	0.1043	2.5008	Deer	1.1481	0.5623
	Bird	0.5957	-0.1628	2.6090	Dog	1.1441	2.2295
	ΔZ_m^g	4.8235	3.6608	22.485	ΔZ_m^g	2.8462	10.203
	ΔZ_m^v	0.7203	-0.6765	-0.6595	Deer	0.7833	1.6304
Z_v	Bird	0.3182	-0.7181	-0.7223	Dog	0.7682	0.8572
	ΔZ_v^g	2.3946	4.0729	61.4704	ΔZ_v^g	2.8765	7.6342
	ΔZ_v^v	4.4746	3.0139	5.6661	ΔZ_v^v	2.0812	2.9867
							5.2146

Table 7. Testing accuracy(%) under attack and no attack.

Methods	no attack			attack		
	Flow.	Cif-100	Cif-10	Flow.	Cif-100	Cif-10
FedDistill	20.47	21.04	63.51	10.31	9.28	20.75
FCCL	25.20	22.72	64.29	12.88	9.06	31.18
DeSA	30.18	26.84	65.68	12.89	9.82	29.14
HYDRA-FL	22.27	20.82	61.74	9.65	11.04	36.94
ReT-FHD(Ours, FL)	33.16	30.77	72.76	20.11	19.28	38.75

weights, we examine its impact on logits and Z-scores for randomly selected clients. For the first three CIFAR-10 categories, the malicious node’s logits show irregular values, such as abnormally high scores for "Airplane." Similarly, its Z-scores Z_m exhibit unusual patterns. To quantify deviations, we use ΔZ_m^g and ΔZ_v^m (calculated via Eq.7), measuring differences between global, verification, and malicious Z-scores. While verification and global Z-scores remain normal, the malicious node’s scores are significantly elevated, with ΔZ_m^g and ΔZ_v^m confirming its aberrant behavior.

Label Reversal Attack. In Tab.6, we simulate label inversion by flipping "Deer" and "Dog" labels on a CIFAR-10 client. The malicious node shows skewed logits, with abnormally high confidence for "Dog" on deer samples, while the verification node’s Z-score deviation from the global average remains minimal. This divergence in logit distributions and Z-score alignment confirms our system’s efficacy in detecting label-flipping attacks.

Logits Attack. Clients under anomalous attacks produce logits conflicting with the optimization goal, disrupting federated learning distillation. Tab.7 analyzes test accuracy under client anomaly attacks on state-of-the-art logits distillation methods across three datasets. Our method outperforms others, with its multi-level distillation effectively mitigating logits attacks, as supported by prior research (Huang et al., 2025a). While HYDRA-FL also uses Multi-level distillation, it focuses on model homogeneity and underperforms in heterogeneous settings. This confirms that our dynamic temperature strategies enhance robustness and security in federated distillation.

Accuracy and Poisoning Tolerance. Model accuracy decrease as the percentage of malicious participants increased, with CIFAR-100 showing a sharper decline due to its complexity. In Fig. 3.a (CIFAR-10), accuracy dropped from 0.35 to below 0.1 with 60% malicious participants, while Fig. 3.b (CIFAR-100) shows a decline from 0.65 to under 0.2. These results demonstrate that server trust alone cannot mitigate adversarial impacts in federated learning, necessitating robust detection mechanisms and adaptive defenses to preserve performance under attack.

6 Conclusion

In this paper, we propose a novel dynamic temperature control strategy for federated distillation (ReT-FHD) and implement it in all FL, Del and blockchain-driven FL frameworks. Through comprehensive experiments, our approach demonstrates substantial improvements across three critical dimensions: model performance, communication efficiency, and security robustness. While our method addresses data and model heterogeneity, the current framework does not evaluate compatibility with attention-based architectures such as Vision Transformers (ViT) (Dosovitskiy et al., 2021). Future efforts will extend validation to Transformer frameworks given their prevalence in visual tasks.

Acknowledgements

This work was supported by NSFC (No.62206200, 62206137, 62036012, 62376196 and U23A20387).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

- Chang, H., Shejwalkar, V., Shokri, R., and Houmansadr, A. Cronus: Robust and heterogeneous collaborative learning with black-box knowledge transfer. *arXiv preprint arXiv:1912.11279*, 2019.
- Chen, H., Asif, S. A., Park, J., Shen, C.-C., and Bennis, M. Robust blockchained federated learning with model validation and proof-of-stake inspired consensus. *arXiv preprint arXiv:2101.03300*, 2021.
- Chrabaszcz, P., Loshchilov, I., and Hutter, F. A downsampled variant of imagenet as an alternative to the cifar datasets, 2017.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houshy, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL <https://arxiv.org/abs/2010.11929>.
- Esfandiari, Y., Tan, S. Y., Jiang, Z., Balu, A., Herron, E., Hegde, C., and Sarkar, S. Cross-gradient aggregation for decentralized learning from non-iid data. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3036–3046. PMLR, 2021. URL <http://proceedings.mlr.press/v139/esfandiari21a.html>.
- Gong, X., Sharma, A., Karanam, S., Wu, Z., Chen, T., Doermann, D., and Innanje, A. Preserving privacy in federated learning with ensemble cross-domain knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 11891–11899, 2022.
- Guo, Z., Zhang, C., Fan, Y., Tian, Y., Zhang, C., and Chawla, N. Boosting graph neural networks via adaptive knowledge distillation, 2023. URL <https://arxiv.org/abs/2210.05920>.
- Hao, Z., Guo, J., Han, K., Tang, Y., Hu, H., Wang, Y., and Xu, C. One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation. *Advances in Neural Information Processing Systems*, 36, 2024.
- He, C., Annavaram, M., and Avestimehr, S. Group knowledge transfer: Federated learning of large cnns at the edge. *Advances in Neural Information Processing Systems*, 33: 14068–14080, 2020.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30*,

- 2016, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL <https://doi.org/10.1109/CVPR.2016.90>.
- Huang, C.-Y., Srinivas, K., Zhang, X., and Li, X. Overcoming data and model heterogeneities in decentralized federated learning via synthetic anchors. *arXiv preprint arXiv:2405.11525*, 2024.
- Huang, C.-Y., Srinivas, K., Zhang, X., and Li, X. Overcoming data and model heterogeneities in decentralized federated learning via synthetic anchors. In *Proceedings of the 41st International Conference on Machine Learning*, ICML’24. JMLR.org, 2025a.
- Huang, W., Ye, M., and Du, B. Learn from others and be yourself in heterogeneous federated learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10133–10143, 2022. doi: 10.1109/CVPR52688.2022.00990.
- Huang, W., Ye, M., Shi, Z., Li, H., and Du, B. Rethinking federated learning with domain shift: A prototype view. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16312–16322. IEEE, 2023.
- Huang, Y., Yan, Z., Shen, C., Fang, F., and Zhang, G. Harmonizing knowledge transfer in neural network with unified distillation. In *European Conference on Computer Vision*, pp. 58–74. Springer, 2025b.
- Itahara, S., Nishio, T., Koda, Y., Morikura, M., and Yamamoto, K. Distillation-based semi-supervised federated learning for communication-efficient collaborative training with non-iid private data. *IEEE Transactions on Mobile Computing*, 22(1):191–205, 2021.
- Jeong, E., Oh, S., Kim, H., Park, J., Bennis, M., and Kim, S.-L. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. *arXiv preprint arXiv:1811.11479*, 2018.
- Jeong, E., Oh, S., Kim, H., Park, J., Bennis, M., and Kim, S.-L. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data, 2023.
- Khan, M. A., Chandio, Y., and Anwar, F. M. Hydra-fl: Hybrid knowledge distillation for robust and accurate federated learning. *arXiv preprint arXiv:2409.19912*, 2024.
- Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 1(4), 2009.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, may 2017. ISSN 0001-0782. doi: 10.1145/3065386. URL <https://doi.org/10.1145/3065386>.
- Kundu, S., Sun, Q., FU, Y., Pedram, M., and Beerel, P. A. Analyzing the confidentiality of undistillable teachers in knowledge distillation. In *35th Neural Information Processing Systems*, 2021.
- Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.
- Lee, R., Fernandez-Marques, J., Hu, S. X., Li, D., Laskaridis, S., Dudziak, Ł., Hospedales, T., Huszár, F., and Lane, N. D. Recurrent early exits for federated learning with heterogeneous clients. *arXiv preprint arXiv:2405.14791*, 2024.
- Li, D. and Wang, J. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*, 2019.
- Li, S., Zhou, T., Tian, X., and Tao, D. Learning to collaborate in decentralized learning of personalized models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 9756–9765. IEEE, 2022a. doi: 10.1109/CVPR52688.2022.00954. URL <https://doi.org/10.1109/CVPR52688.2022.00954>.
- Li, Z., Li, X., Yang, L., Zhao, B., Song, R., Luo, L., Li, J., and Yang, J. Curriculum temperature for knowledge distillation, 2022b. URL <https://arxiv.org/abs/2211.16231>.
- Lian, X., Zhang, C., Zhang, H., Hsieh, C., Zhang, W., and Liu, J. Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5330–5340, 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/f75526659f31040afeb61cb7133e4e6d-Abstract.html>.
- Liang, P., Chen, J., Wu, Y., Pu, B., Huang, H., Chang, Q., and Ran, G. Data free knowledge distillation with feature synthesis and spatial consistency for image analysis. *Scientific Reports*, 14(1):27557, 2024.

- Lin, T., Kong, L., Stich, S. U., and Jaggi, M. Ensemble distillation for robust model fusion in federated learning. *Advances in neural information processing systems*, 33: 2351–2363, 2020.
- Liu, J., Liu, B., Li, H., and Liu, Y. Meta knowledge distillation, 2022. URL <https://arxiv.org/abs/2202.07940>.
- Ma, N., Zhang, X., Zheng, H., and Sun, J. Shufflenet V2: practical guidelines for efficient CNN architecture design. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y. (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, volume 11218 of *Lecture Notes in Computer Science*, pp. 122–138. Springer, 2018. doi: 10.1007/978-3-030-01264-9_8. URL https://doi.org/10.1007/978-3-030-01264-9_8.
- Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 722–729, 2008. doi: 10.1109/ICVGIP.2008.47.
- Qin, Z., Yan, X., Zhou, M., and Deng, S. Blockdfl: A blockchain-based fully decentralized peer-to-peer federated learning framework. In *Proceedings of the ACM on Web Conference 2024*, pp. 2914–2925, 2024.
- Shen, T., Zhang, J., Jia, X., Zhang, F., Huang, G., Zhou, P., Kuang, K., Wu, F., and Wu, C. Federated mutual learning, 2020.
- Sun, S., Ren, W., Li, J., Wang, R., and Cao, X. Logit standardization in knowledge distillation. *arXiv preprint arXiv:2403.01427*, 2024.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 1–9. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7298594. URL <https://doi.org/10.1109/CVPR.2015.7298594>.
- Tan, Y., Long, G., Liu, L., Zhou, T., Lu, Q., Jiang, J., and Zhang, C. Fedproto: Federated prototype learning across heterogeneous clients. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 8432–8440. AAAI Press, 2022. doi: 10.1609/AAAI.V36I8.20819. URL <https://doi.org/10.1609/aaai.v36i8.20819>.
- Wang, J., Zhao, C., Lyu, L., You, Q., Huai, M., and Ma, F. Bridging model heterogeneity in federated learning via uncertainty-based asymmetrical reciprocity learning. *arXiv preprint arXiv:2407.03247*, 2024.
- Wang, J., Zhao, C., Lyu, L., You, Q., Huai, M., and Ma, F. Bridging model heterogeneity in federated learning via uncertainty-based asymmetrical reciprocity learning. In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*. JMLR.org, 2025.
- Wei, Y. and Bai, Y. Dynamic temperature knowledge distillation, 2024. URL <https://arxiv.org/abs/2404.12711>.
- Wu, C., Wu, F., Lyu, L., Huang, Y., and Xie, X. Communication-efficient federated learning via knowledge distillation. *Nature Communications*, 13(1), April 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-29763-x. URL <http://dx.doi.org/10.1038/s41467-022-29763-x>.
- Ying, C., Xia, F., Wei, D. S., Yu, X., Xu, Y., Zhang, W., Jiang, X., Jin, H., Luo, Y., Zhang, T., et al. Bit-fl: Blockchain-enabled incentivized and secure federated learning framework. *IEEE Transactions on Mobile Computing*, 2024.
- Zhang, J., Guo, S., Ma, X., Wang, H., Xu, W., and Wu, F. Parameterized knowledge transfer for personalized federated learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 10092–10104, 2021. URL <https://proceedings.neurips.cc/paper/2021/hash/5383c7318a3158b9bc261d0b6996f7c2-Abstract.html>.
- Zhang, J., Liu, Y., Hua, Y., and Cao, J. Fedtgp: Trainable global prototypes with adaptive-margin-enhanced contrastive learning for data and model heterogeneity in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16768–16776, 2024a.
- Zhang, J., Liu, Y., Hua, Y., and Cao, J. An upload-efficient scheme for transferring knowledge from a server-side pre-trained generator to clients in heterogeneous federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.
- Zhang, L., Shen, L., Ding, L., Tao, D., and Duan, L.-Y. Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10174–10183, 2022.

Zhao, S., Wang, X., and Wei, X. Mitigating accuracy-robustness trade-off via balanced multi-teacher adversarial distillation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Zhu, T., He, F., Chen, K., Song, M., and Tao, D. Decentralized SGD and average-direction SAM are asymptotically equivalent. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 43005–43036. PMLR, 2023. URL <https://proceedings.mlr.press/v202/zhu23e.html>.

A Appendix

A.1 The proof of Theorem 4.1 in Sec.4

For a discrete random variable X with probability distribution $P(x)$, the information entropy is defined as:

$$H(X) = - \sum_{x \in X} P(x) \log P(x) \quad (15)$$

where $P(x)$ is obtained from the softmax function. The softmax function is defined as follows:

$$\text{softmax}(p_i) = \frac{e^{z_i/T}}{\sum_{j=1}^n e^{z_j/T}} \quad (16)$$

Firstly, we can calculate the derivative of the probability distribution p_i with respect to the temperature T :

$$\frac{\partial p_i}{\partial T} = \frac{\partial}{\partial T} \left(\frac{\exp(z_i/T)}{\sum_{j=1}^n \exp(z_j/T)} \right) \quad (17)$$

By calculation, it is possible to obtain:

$$\begin{aligned} \frac{\partial p_i}{\partial T} = & \left[\sum_{j=1}^n \exp\left(\frac{z_j}{T}\right) \exp\left(\frac{z_i}{T}\right) \left(-\frac{z_i}{T^2}\right) - \exp\left(\frac{z_i}{T}\right) \right. \\ & \left. \cdot \sum_{j=1}^n \exp\left(\frac{z_j}{T}\right) \left(-\frac{z_j}{T^2}\right) \right] / \left(\sum_{j=1}^n \exp\left(\frac{z_j}{T}\right) \right)^2 \end{aligned} \quad (18)$$

Simplifying and organizing gives:

$$\frac{\partial p_i}{\partial T} = p_i \left(\frac{\sum_{j=1}^n p_j z_j - z_i}{T} \right) \quad (19)$$

Secondly, information entropy is a fundamental concept used to measure the uncertainty of a random variable. For a discrete random variable X with probability distribution $P(x)$, the information entropy is defined as:

$$H(X) = - \sum_{x \in X} P(x) \log P(x) \quad (20)$$

The information entropy is derived with respect to temperature as follows:

$$\frac{\partial H}{\partial T} = - \sum_{i=1}^n \left(\frac{\partial p_i}{\partial T} \log p_i + p_i \frac{\partial}{\partial T} (\log p_i) \right) \quad (21)$$

Simplifying and organizing gives:

$$\begin{aligned} \frac{\partial H}{\partial T} &= - \sum_{i=1}^n \left(\frac{\partial p_i}{\partial T} \log p_i + \frac{\partial p_i}{\partial T} \right) \\ &= - \sum_{i=1}^n \frac{\partial p_i}{\partial T} (\log p_i + 1) \end{aligned} \quad (22)$$

By substituting the expression for $\frac{\partial p_i}{\partial T}$, we get:

$$\frac{\partial H}{\partial T} = - \sum_{i=1}^n p_i \left(\frac{\sum_{j=1}^n p_j z_j - z_i}{T} \right) (\log p_i + 1) \quad (23)$$

Based on the above derivation, the equation for the derivative of the entropy H with respect to the temperature T is known. Next, it is necessary to replace the conventional temperature T with our predefined elastic temperature τ in Eq. (6). To accurately determine the relationship between entropy and the elastic temperature τ , it is necessary to understand how τ varies with ΔZ , namely $\frac{\partial \tau}{\partial \Delta Z}$.

By the definition of G , we can derive:

$$\begin{aligned} \frac{\partial \tau}{\partial \Delta Z} &= \frac{\partial}{\partial \Delta Z} \left(\xi + \frac{\gamma \log(1 + \Delta Z)}{\log(1 + \Delta Z_{\max})} \right) \\ &= \frac{\gamma}{\log(1 + \Delta Z_{\max})} \cdot \frac{1}{1 + \Delta Z} \end{aligned} \quad (24)$$

Ultimately, the derivative of the entropy H with respect to τ can be expressed as:

$$\begin{aligned} \frac{\partial H}{\partial \Delta Z} &= \frac{\partial H}{\partial \tau} \cdot \frac{\partial \tau}{\partial \Delta Z} \\ &= \left(- \sum_{i=1}^n p_i \left(\frac{\sum_{j=1}^n p_j z_j - z_i}{\tau} \right) (\log p_i + 1) \right) \\ &\quad \cdot \left(\frac{\gamma}{\log(1 + \Delta Z_{\max})} \cdot \frac{1}{1 + \Delta Z} \right) \end{aligned} \quad (25)$$

The condition $\frac{\partial H}{\partial \tau} > 0$ implies that an increase in the elastic temperature τ leads to a decrease in the system's entropy. This reduction in entropy is reflected as an increase in information, known as information gain. Consequently, appropriately increasing the elastic temperature can enhance information gain, allowing the model to learn more effectively from the training data.

Then the Theorem 4.1 is proved.

A.2 The proof of Corollary 4.2 in Sec. 3

For a model f with random distribution I in the initial state, for a deep neural network (DNN) model f with random distribution I in the initial state, the model has no preference for any sample. Thus for the model's predictive distribution $f(x; \theta_I) = \{p_1^I, p_2^I, \dots, p_c^I\}$, the following relationship exists:

$$\mathbb{E}(p^I(x)) = \mathbb{E}(p_1^I(x)) = \mathbb{E}(p_2^I(x)) = \dots = \mathbb{E}(p_c^I(x)) \quad (9)$$

Assumption A.1. We assume that the model is optimised under the guidance of a soft label distribution with smoothing degree λ_1 is satisfied:

$$\begin{aligned} p_1^{\lambda_1}(x_1) &= p_2^{\lambda_1}(x_2) = \dots = p_c^{\lambda_1}(x_c) > \frac{1}{C} > \\ p_1^{\lambda_1}(x_{c \neq 1}) &= p_2^{\lambda_1}(x_{c \neq 2}) = \dots = p_c^{\lambda_1}(x_{c \neq c}) \end{aligned} \quad (10)$$

For simplicity, we only consider a dataset that includes two categories, for which the categories are defined as follows.

Definition A.2. Let the cross-entropy loss of a sample be denoted as $\mathcal{L}_{CE}$, and the average cross-entropy loss over all samples be $\bar{\mathcal{L}}_{CE}$. The class ID is defined as:

$$\text{class ID} = \begin{cases} 1, & \text{if } \mathcal{L}_{CE} > \bar{\mathcal{L}}_{CE}, \\ 2, & \text{otherwise.} \end{cases}$$

Referring to the study of (Zhao et al., 2024), we introduce the knowledge scale to quantify the level of knowledge of the predictive distribution, which can measure the scale of knowledge learnt from the probability distribution by the model with randomly initialised weights:

$$K^T = KL(I(x), T(x)) = \log C - H(P^T) \quad (11)$$

where $KL(\cdot)$ stands for Kullback–Leibler divergence, which calculates the cost of information required to travel from one distribution to another, C is the total number of categories and is a constant for the same dataset, and $H(\cdot)$ is the information entropy.

First we calculate the derivative of the knowledge scale with respect to logits $z_c(x)$:

$$\begin{aligned} & \frac{\partial KL(f(x; \theta_I), P_{\lambda_1})}{\partial z_c(x)} \\ &= \sum_{i=1}^C \frac{\partial p_i}{\partial z_c(x)} \cdot \frac{\partial KL(f(x; \theta_I), P_{\lambda_1})}{\partial p_i} \\ &= \sum_{i=1}^C p_i^{\lambda_1}(x) p_c^I(x) - p_c^{\lambda_1}(x) \\ &= p_c^I(x) - p_c^{\lambda_1}(x) \end{aligned} \quad (12)$$

For simplicity, we will subsequently develop the proof using $C = 2$ as an example. So we can easily obtain:

$$\begin{aligned} & \mathbb{E} \left(\frac{\partial KL(f(x_1; \theta_I), P_{\lambda_1})}{\partial z_1(x_1)} \right) = \mathbb{E} \left(\frac{\partial KL(f(x_2; \theta_I), P_{\lambda_1})}{\partial z_2(x_2)} \right) \\ &= \mathbb{E} \left(p_c^I(x_c) - p_1^{\lambda_1}(x_1) \right) = \mathbb{E} \left(p_c^I(x_c) - p_2^{\lambda_1}(x_2) \right) \end{aligned} \quad (13)$$

$$\begin{aligned} & \mathbb{E} \left(\frac{\partial KL(f(x_1; \theta_I), P_{\lambda_1})}{\partial z_2(x_1)} \right) = \mathbb{E} \left(\frac{\partial KL(f(x_2; \theta_I), P_{\lambda_1})}{\partial z_1(x_2)} \right) \\ &= \mathbb{E} \left(p_c^I(x_c) - p_2^{\lambda_1}(x_1) \right) = \mathbb{E} \left(p_c^I(x_c) - p_1^{\lambda_1}(x_2) \right) \end{aligned} \quad (14)$$

Combined with the previous assumptions, we can obtain that the partial derivatives of knowledge scale changes are symmetric for all categories, which suggests that the overall change in knowledge scale has a balanced impact on different categories. Therefore, our elastic temperature focus on the adjustment of knowledge scale at different stages of the model and treat different categories equally, which helps us to bridge the structural differences between heterogeneous models of different clients.

Then the Corollary 4.2 is proved.

A.3 The proof of Theorem 4.3 in Sec.4

We explore the relationship between changes in the category gradient and temperature, using category 1 as an example. The gradient for category 1 is:

$$\nabla_{f \sim \lambda_1}^1 = \mathbb{E} \left[\left| \frac{\partial z_1}{\partial \theta} \cdot \frac{\partial \text{KL}(f_s(x_1; \theta_I), P_{\lambda_1})}{\partial z_1} \right| + \left| \frac{\partial z_2}{\partial \theta} \cdot \frac{\partial \text{KL}(f_s(x_1; \theta_I), P_{\lambda_1})}{\partial z_2} \right| \right] \quad (2)$$

We increase the temperature of category 1 to $\tilde{\tau}_1 + \Delta \tilde{\tau}_1$ to obtain the global logits distribution P_{λ_2} . We can then obtain:

$$\begin{aligned} & \nabla_{f \sim \lambda_2}^1 - \nabla_{f \sim \lambda_1}^1 = \\ & \mathbb{E} \left[\left| \frac{\partial z_1}{\partial \theta} \cdot \frac{\partial \text{KL}(f_s(x_1; \theta_I), P_{\lambda_2})}{\partial z_1} \right| + \left| \frac{\partial z_2}{\partial \theta} \cdot \frac{\partial \text{KL}(f_s(x_1; \theta_I), P_{\lambda_2})}{\partial z_2} \right| \right] \\ & - \mathbb{E} \left[\left| \frac{\partial z_1}{\partial \theta} \cdot \frac{\partial \text{KL}(f_s(x_1; \theta_I), P_{\lambda_1})}{\partial z_1} \right| + \left| \frac{\partial z_2}{\partial \theta} \cdot \frac{\partial \text{KL}(f_s(x_1; \theta_I), P_{\lambda_1})}{\partial z_2} \right| \right] \end{aligned} \quad (3)$$

where $\frac{\partial z}{\partial \theta}$ can be regarded as the impact of the class itself on model weight optimization, in a practical sense, this part reflects the inner relationship between DNN and sample, more specifically, it can reflect the difficulty of the sample itself for

the model and not relate to the optimization object. Therefore, the above equation can be simplified as:

$$\begin{aligned} & \nabla_{f \sim \lambda_2}^1 - \nabla_{f \sim \lambda_1}^1 = \\ & \mathbb{E} \left[\frac{\partial z_1}{\partial \theta} \cdot \left(p_1^{\lambda_1}(x_1) - p_1^{\lambda_2}(x_1) \right) + \frac{\partial z_2}{\partial \theta} \cdot \left(p_2^{\lambda_1}(x_1) - p_2^{\lambda_2}(x_1) \right) \right] \end{aligned} \quad (4)$$

Based on the effect of temperature on the probability distribution: $p_1^{\lambda_1}(x_1) - p_1^{\lambda_2}(x_1) > 0$, $p_2^{\lambda_1}(x_1) - p_2^{\lambda_2}(x_1) < 0$. At the same time, since category 1 is the main category, its gradient changes dominate. Therefore we can get:

$$\nabla_{f \sim \lambda_2}^1 - \nabla_{f \sim \lambda_1}^1 > 0 \quad (5)$$

The same can be obtained when we reduce $\Delta \tilde{\tau}_1$ for the category temperature to get the global logits distribution P_{λ_3} :

$$\nabla_{f \sim \lambda_3}^1 - \nabla_{f \sim \lambda_1}^1 < 0 \quad (6)$$

Then the Theorem 4.3 is proved.

A.4 Blockchain-based FL Details

In the blockchain network, the role of the v node is exclusively dedicated to the voting processes. This node evaluates transactions based on the established Equation x , which serves as a decision-making criterion. Depending on the results derived from applying this equation, the verification node can make its decisions, opting either for P (Positive) or N (Negative). This dichotomy allows the node to actively participate in the consensus mechanism by approving or rejecting transactions, thereby ensuring the reliability and security of the blockchain.

Subsequently, the node v will take on the responsibility of broadcasting all the votes it has received for the node w to all m nodes within the network. This dissemination ensures that every m node in our simulated network receives the set of values z belonging to the set $\mathcal{Z}$ for each w . This comprehensive distribution of vote information across the network not only enhances transparency but also enables each node to have a holistic view of the network's consensus on each w , facilitating more informed decision-making and maintaining the integrity of the overall system.

Each node m in the network meticulously processes all data extracted from the transactions sent by node v . Upon processing, each m node synthesizes this information into a series of candidate blocks. A typical candidate block, denoted as $block_j^m$, encapsulates not only the global logits z_i but also includes the result R and data pertaining to the blacklist $\mathcal{B}$. This blacklist is dynamically generated based on the collective voting outcomes, reflecting the network's consensus on disreputable nodes or transactions.

Both types of nodes, v from set $\mathcal{V}$ and m from set $\mathcal{M}$, receive rewards in alignment with their contributions to the network's security and integrity. The reward system is designed to incentivize activities that uphold the blockchain's operational standards and trustworthiness. Furthermore, the blockchain protocol is engineered to select the candidate block from the most affluent m nodes—the ones with the highest accrued rewards—as the preferred choice for linkage to the blockchain. This selection criterion ensures that the most reliable and trustworthy nodes, as determined by their wealth accumulation through valid and honest contributions, have their candidate blocks prioritized in the blockchain extension process.

To effectively simulate a real blockchain-based Federated Learning (FL) environment, our methodology includes the dissemination of pertinent information through the network. This is achieved by encapsulating the information within a transaction, which is then securely signed using the private key of the designated role. This approach ensures the integrity and authenticity of the data being broadcast across the network.

A.5 Additional Experimental Results

Effects of different E and K. In our experiments, we specifically assess the performance of our method across different setups, involving 50 and 100 clients. The results, detailed in Table 8, indicate that while there is a slight decline in accuracy as the number of clients increases, our method demonstrates relatively stable outcomes over multiple training cycles. For instance, with 50 clients, our method achieve an accuracy of 42.46% at $E = 20$, and with 100 clients, it maintains a respectable 37.57%, despite the increased complexity and potential for data discrepancies among a larger number of clients. In contrast to the general trend observed with other models, our proposed High-Dimensional Federated Learning (HDFL) approach managed to maintain performance across an extended number of client training cycles. This resilience

Table 8. The test accuracy (%) on Cifar100 in the practical setting using the heterogeneous model group with a different number of client training epochs (E) and different number of clients(K).

	$E = 5$	$E = 10$	$E = 20$	$K = 50$	$K = 100$
FML	39.63 ± 0.83	39.81 ± 0.82	39.45 ± 0.83	36.65 ± 1.62	33.78 ± 1.66
FedProto	37.61 ± 0.91	37.34 ± 1.01	35.71 ± 1.31	33.24 ± 2.14	31.79 ± 1.89
FedDistill	40.20 ± 0.98	39.82 ± 0.97	39.89 ± 1.01	37.77 ± 1.60	35.36 ± 1.66
FedKD	37.01 ± 1.11	36.96 ± 1.12	37.77 ± 1.07	32.64 ± 1.85	32.34 ± 1.64
FedTGP	42.87 ± 1.67	41.69 ± 1.84	40.94 ± 1.80	39.84 ± 2.03	36.75 ± 1.84
Ours	47.98 ± 0.62	48.38 ± 0.64	48.51 ± 0.63	42.46 ± 1.62	37.57 ± 1.86

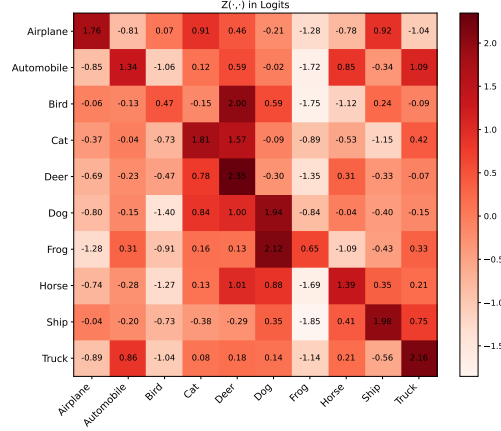


Figure 4. Global Logits Z-Score Visualization for Cifar-10 with 20 Clients

can be attributed to our model’s ability to effectively manage and synchronize heterogeneous data and learning strategies across diverse client environments. The capability to sustain high accuracy levels under varying client counts and training conditions highlights the robustness of our approach, making it particularly suitable for practical applications where client variability and training dynamics are challenging factors.

Global Logits Z-Score Visualization. Fig.4 visualizes the Z-Score of the global logits under the Cifar-10 dataset. It can be found that the data distribution of logits is more concentrated through the z-score processing.

Supplement of Tab.6. Tab.9 and Tab.10 extends the results of Tab.7 by providing a comprehensive display of the complete logits and Z-scores, offering deeper insights into the behavior of the models under the evaluated scenarios. These supplementary results aim to enhance the understanding and robustness of the conclusions drawn from the main experiments.

Complete Experimental Data on Malicious Model. Tab.9 shows that the complete logits and Z-scores obtained by injecting noise to the weights of the malicious node’s category airplane and category bird. The comprehensive analysis of logits and Z-scores across all categories in the supplementary data reveals distinct behavioral patterns between malicious and normal nodes. Specifically, the malicious node exhibits significant anomalies in the "Airplane" and "Bird" categories, with logits and Z-scores deviating markedly from the normal range (e.g., specific values in Z-scores for category airplane corresponding to category airplane and category bird: $[0.7710, 2.5008] \rightarrow [0.7203, -6595]$). In contrast, other categories such as "Car" and "Trucks" remain stable, with minimal deviations (e.g., $[-0.6385, -0.8651]$ vs. $[-0.7638, -0.7107]$, $\Delta = [0.1253, -0.1544]$). The differential measures ΔZ_m^g and ΔZ_v^m effectively quantify these deviations, with values for anomalous categories (e.g., Bird: $[22.485, 61.4704]$) significantly exceeding those of normal categories. These findings underscore the targeted nature of the attack and the robustness of the proposed detection metrics in identifying malicious activities.

Complete Experimental Data on Label Reversal Attack. In Tab. 10, we simulate a label inversion attack by flipping the labels of "Deer" and "Dog" on a selected CIFAR-10 client. The malicious node exhibits significant anomalies in its logits, particularly displaying abnormally high confidence for the "Dog" class when the true label is "Deer." For instance, the Z-Score for "Dog" on deer samples rise to 2.2411, compared to the normal range of $[-0.6547, 1.1481]$. In contrast, the

Table 9. Supplement to Table 6.(Malicious model)

		Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Trucks
$\mathcal{L}/C$	Airplane	681.74	248.60	1805.43	-233.92	-289.58	-133.81	117.65	105.17	-111.56	-381.13
	Bird	0.0477	-1.0924	0.00163	-0.03615	-0.6927	-0.0313	-0.78591	0.11322	0.17006	-0.0338
Z_m	Airplane	0.7710	0.1043	2.5008	-0.6385	-0.7242	-0.4844	-0.0973	-0.1165	-0.4501	-0.8651
	Bird	0.5957	-0.1628	2.6090	-0.7720	-0.2781	-0.6905	-0.2938	-0.1418	-0.1322	-0.7333
	ΔZ_m^g	4.8235	3.6608	22.485	7.0336	14.503	5.1918	8.1249	4.8195	5.7027	3.1847
Z_v	Airplane	0.7203	-0.6765	-0.6595	-0.7638	0.4510	1.0750	-0.6278	-0.8247	2.0166	-0.7107
	Bird	0.3182	-0.7181	-0.7223	-0.5224	0.4126	1.6685	-0.6531	-0.8371	1.7418	-0.6880
	ΔZ_v^m	2.3946	4.0729	61.4704	2.4421	5.5224	17.1949	3.2096	4.1907	20.675	2.2256
	ΔZ_v^g	4.4746	3.0139	5.6661	7.3089	6.3297	6.99128	5.1713	2.40193	8.0513	3.3053

Table 10. Supplement to Table 6.(Label reversal attack)

		Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Trucks
$\mathcal{L}/C$	Deer	-0.8718	-0.3399	-0.8518	1.7862	0.9026	3.4350	-0.9332	-0.9158	-0.7916	-0.8755
	Dog	-0.9006	-0.3496	-0.8857	1.7653	3.3860	0.9474	-0.9386	-0.9114	-0.7518	-0.7910
Z_m	Deer	-0.6140	-0.2614	-0.6008	1.1481	0.5623	2.2411	-0.6547	-0.6432	-0.5608	-0.6165
	Dog	-0.6413	-0.2723	-0.6313	1.1441	2.2295	0.5963	-0.6667	-0.6485	-0.5416	-0.5679
	ΔZ_m^g	2.2690	2.1762	3.3980	2.8462	10.203	18.7151	3.2746	4.6532	2.4289	3.3192
Z_v	Deer	0.3912	-0.7767	-0.6786	0.7833	1.6304	1.0069	-0.6776	-0.8615	0.3785	-0.6293
	Dog	0.3617	-0.7883	-0.6498	0.7682	0.8572	1.8016	-0.7156	-0.8255	0.3113	-0.5843
	ΔZ_v^m	5.6487	4.0295	2.4296	2.8765	7.6342	7.5652	2.4054	2.9043	5.02901	2.3278
	ΔZ_v^g	6.4078	4.3835	3.3734	2.0812	2.9867	5.2146	2.7634	6.7946	4.140	3.8648

verification node’s Z-scores remain closely aligned with the global average, with smaller deviations (e.g., ΔZ_v^g values within [2.7634, 6.4078]). This stark divergence in logit distributions and Z-score alignment between the malicious and verification nodes highlights the effectiveness of our system in detecting label-flipping attacks. Furthermore, the differential measures ΔZ_m^g and ΔZ_v^m quantify the malicious node’s deviations, with values for the flipped categories (e.g., [7.6342, 7.5652]) significantly exceeding those of normal categories. These results demonstrate that our approach not only identifies the presence of label-flipping attacks but also provides robust metrics for quantifying the extent of the adversarial manipulation.