DKDR: Dynamic Knowledge Distillation for Reliability in Federated Learning

Yueyang Yuan[1,2†], **Wenke Huang**[1,2†], **Guancheng Wan**[1†],
Kaiqi Guan[1], **He Li**[1], **Mang Ye**[1*]
[1] National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence,
Hubei Key Laboratory of Multimedia and Network Communication Engineering,
School of Computer Science, Wuhan University, Wuhan, China.
[2] Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
{yueyangyuan, wenkehuang, guanchengwan, yemang}@whu.edu.cn

Abstract

Federated Learning (FL) has demonstrated a promising future in privacy-friendly collaboration but it faces the data heterogeneity problem. Knowledge Distillation (KD) can serve as an effective method to address this issue. However, challenges arise from the unreliability of existing distillation methods in multi-domain scenarios. Prevalent distillation solutions primarily aim to fit the distributions of the global model directly by minimizing forward Kullback-Leibler divergence (KLD). This results in significant bias when the outputs of the global model are multi-peaked, which indicates **the unreliability of distillation pathway**. Meanwhile, cross-domain update conflicts can notably reduce the accuracy of the global model (teacher model) in certain domains, reflecting **the unreliability of the teacher model** in these domains. In this work, we propose **DKDR** (**D**ynamic **K**nowledge **D**istillation for **R**eliability in Federated Learning), which dynamically assigns weights to forward and reverse KLD based on knowledge discrepancies. This enables clients to fit the outputs from the teacher precisely. Moreover, we use knowledge decoupling to identify domain experts, thus clients can acquire reliable domain knowledge from experts. Empirical results from single-domain and multi-domain image classification tasks demonstrate the effectiveness of the proposed method and the efficiency of its key modules. The code is available at https://github.com/YueyangYuan/DKDR.

1 Introduction

Federated learning is a collaborative paradigm [21, 60, 27, 14–16, 62], enabling multiple clients to jointly train a shared global model [39, 28, 15] while ensuring privacy protection [52]. However, the distributed data is collected from different sources with diverse preferences and brings the non-independent and identically distributed (non-IID) characteristics. Knowledge distillation [13, 4] addresses this challenge effectively by aligning the outputs of local models with the global model. It brings the optimization objectives of each client closer together thus resolving the problem.

However, existing distillation methods [25, 11, 36, 5] typically use forward KLD to fit the distributions of the global model. We argue that this approach is unreliable in multi-domain scenarios. Given the global model distribution $Z(y|x)$ and the local model distribution $Z_w(y|x)$ parameterized by w, standard knowledge distillation objectives aim to minimize the forward KLD between them, denoted as $KL[Z|Z_w]$. This approach compels Z_w to encompass all modes of Z. However, we

[†] Equal Contribution.
[*] Corresponding Author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

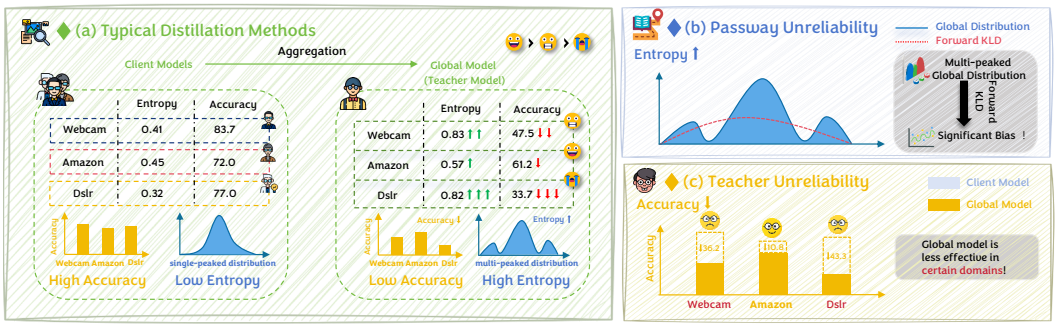


Figure 1: **Problem illustration. (a) Typical Distillation Methods** exist two unreliability problems as follows. **(b) Pathway Unreliability**: in multi-domain scenarios, the aggregated global model shows a significantly higher entropy compared to local models in the distributions over private datasets, exhibiting a multi-peak structure. In such case, minimizing forward KLD leads to significant bias; **(c) Teacher Unreliability**: after aggregation, the global model experiences catastrophic accuracy drops in certain domains. Thus in these areas, the global model is unable to serve as an effective teacher to provide high-quality guidance for clients.

notice that the global model will produce multi-peaked outputs in domains with limited data, as shown in Fig.1. In such situations, minimizing the forward KLD causes Z_w to assign unreasonably high probabilities to regions of low density in Z [38]. **Therefore, distillation on these domains will introduce significant bias, which is the unreliability of the distillation pathway**. This context naturally raises the following critical question: **I)** *how can we establish a reliable distillation pathway in federation?* Meanwhile, conflicts in update directions across different domains can significantly reduce the accuracy of the global model on some domains. **Thus, the global model is inherently less effective on these domains, which is the unreliability of the teacher model**. This situation prompts another intriguing question: **II)** *how can we get a reliable teacher for distillation?*

To address these challenges, we propose DKDR (**D**ynamic **K**nowledge **D**istillation for **R**eliability in Federated Learning). Concerning the issue of pathway unreliability mentioned in **I)**, we initially conduct a theoretical analysis of federated knowledge distillation (see Sec.3.2). The reverse KLD, denoted as $KL[Z_w||Z]$, is widely used in knowledge distillation. Reverse KLD promotes a mode-seeking behavior, leading Z_w to focus on a singular mode of Z [3, 53, 22], while forward KLD induces a mean-seeking behavior, encouraging Z_w to capture the overall distribution of Z. Therefore, when distilling the multi-peaked distributions, forward KLD will assign high probabilities to low-density regions. In contrast, the reverse KLD prioritizes confidence intervals. Both methods exhibit different types of bias. An intuitive idea is utilizing their characteristics to design a dynamic weighting method, aiming to minimize the bias introduced by distillation pathway as much as possible. We demonstrate from both experimental and theoretical perspectives that the forward KLD prioritizes fitting the dominant regions of the global distribution, while the reverse KLD prioritizes fitting the lower-probability segments. Based on their characteristics, we introduce **Dynamic Distillation**: dynamically allocating weights to forward and reverse KLD based on the knowledge discrepancies between the Dominant Knowledge Components (DKCs) and Ancillary Knowledge Components (AKCs) (see Sec.3.2). Thus clients can precisely fit the distributions of the teacher model.

In the second place, to get a reliable teacher mentioned in **II)**, we propose **Knowledge Decoupling** to get domain experts: we first modularize knowledge into shared and unique components and then use SVD to extract the main components of unique knowledge. Subsequently, clustering techniques are employed on the refined components to identify domain experts that specialize in distinct domains. In federation, clients learn from domain experts rather than the global model, thus gaining reliable domain knowledge. Experimental results reveal that our method consistently achieves better performance than others. The main contributions are summarized as:

❶ ***Re-examining KD in FL from a Reliability Perspective.*** Our findings indicate that existing federated distillation methods are unreliable in multi-domain scenarios, which results in significant distillation bias and less effective guidance for clients in domains with limited data.

❷ ***Novel Dynamic Multi-experts Distillation Framework for Reliability.*** Building on the phenomenon of unreliable KD in FL, we effectively mitigate distillation bias and comprehensively improve the performance across domains by addressing the pathway unreliability and teacher unreliability.

❸ ***Theoretical Guarantees and Experimental Validation.*** We provide theoretical guarantees for our framework, and further demonstrate the effectiveness of it through comprehensive experiments.

2 Related Work

2.1 Heterogeneous Federated Learning

A pioneering work proposed the currently most widely used algorithm, FedAvg [39]. However, it suffers from performance deterioration when applied to non-i.i.d data (data heterogeneity). Shortly thereafter, a substantial body of research [29, 46, 48, 28] emerged, focusing on non-i.i.d data. These methods primarily address label distribution skew, where non-i.i.d data [18] is created by partitioning existing data based on label space with limited domain shift. FedProx [29], FedCurv [47], pFedME [48], and FedDyn [2] calculate global parameter stiffness to control discrepancies. Besides, MOON[28], FedUFO [64], FedProto[49], and FedProc[42] maximize feature-level alignment of local model and global model. Moreover, SCAFFOLD [19] and FedDC [9] leverage global gradient calibration to control local drift. Nevertheless, when private data is sampled from different data domains, these works do not consider inter-domain performance, concentrating instead on learning an internal model. Recent studies have explored related issues in unsupervised domain adaptation for target domains [43, 30] and domain generalization on unseen domains [34]. However, collecting data in the target domain can be time-consuming and impractical, while considering performance on unknown domains represents an idealistic scenario. In more realistic settings, participants are likely to be more concerned with performance across other domains, as this could directly enhance economic benefits. Our method leverage the **Knowledge Decoupling** to capture domain-specific signals and identify domain experts. It focuses on improving performance in outer domains during distillation, learning a generalizable and stable global model during the federated learning process.

2.2 Federated Knowledge Distillation

Knowledge Distillation (KD) [13] is a technique that has been extensively studied and applied in various areas of machine learning. Currently, KD has found widespread applications in FL, which can be broadly categorized into three main areas: addressing data heterogeneity, enhancing generalization capabilities, and mitigating catastrophic forgetting [54, 66, 58, 26, 61, 57, 17]. In terms of addressing data heterogeneity, FedFTG [65] employs a data-free knowledge distillation method to fine-tune the global model, while FedDKD [31] introduces a decentralized knowledge distillation module to distill knowledge from local models. Moreover, FedUSL [6] employs a self-label reassigning method to rectify the global model predictions. Regarding the enhancement of generalization capabilities, FedX [11] utilizes a two-sided knowledge distillation approach with contrastive learning as a core component, enabling the federated system to operate without requiring clients to share any data features. Furthermore, FedMEKT [24] develops a distillation-based multimodal embedding knowledge transfer mechanism, which allows the server and clients to exchange joint multimodal embedding knowledge extracted from a multimodal proxy dataset. Finally, to address the issue of catastrophic forgetting, FedNTD [25] proposes a novel and effective algorithm, Federated Not-True Distillation, which preserves the global perspective on locally available data exclusively for the not-true classes. Additionally, CFeD [36] performs knowledge distillation on both the clients and the server to mitigate forgetting. And DFRD [35] maintains an exponential moving average copy of the generator on the server to overcome the catastrophic forgetting, using dynamic weighting and label sampling to accurately extract knowledge. It is worth noting that **all of these methods distill knowledge directly by utilizing forward KLD, resulting in significant bias in multi-domain scenarios.** In our work, we firstly introduce **Dynamic Distillation**, which dynamically weights forward and reverse KLD for different knowledge modules and establishes precise distillation.

3 Methodology

3.1 Preliminary

Generic Federated Learning. In general federated learning settings [39, 29, 28, 40, 41, 59, 15], there are K clients (indexed by k) each possessing its respective private data, denoted as $D_k = \{x_i, y_i\}_{i=1}^{N_k}$, where N_k represents the number of data points held by the k^{th} client. The global model parameters at the beginning of the t^{th} communication epoch are denoted as w^t. The server broadcasts these parameters to each client, assigning them as $w_k^t \leftarrow w^t$. Each client conducts local optimization and

uploads the updated parameters back to the server for weighted parameter aggregation:

$$w_k^t \leftarrow w_k^t - \eta \nabla \sum_{i \in B_k} l(w_k^t, \xi_i), \quad w^{t+1} = \sum_k \alpha_k w_k^t, \tag{1}$$

here, the B_k denotes the mini-batch sampled from the private data D_k, ξ represents the query instance, and η indicates the local learning rate. The optimization objective is to secure a well-performing global model through the federated learning process.

Domain shift. There exists domain shift among private data. Specifically, for the same label space, distinctive feature distributions exists among different participants, which can be defined as:

$$\mathbb{P}_i(x|y) \neq \mathbb{P}_j(x|y) \quad \mathbb{P}_i(y) = \mathbb{P}_j(y). \tag{2}$$

Federated Knowledge Distillation. In typical federated knowledge distillation settings [25, 14], clients use forward KLD to distill knowledge from the global model. Specifically, the global model w^{t-1} at the end of the $(t-1)^{th}$ round involves the knowledge learned from other participants. We calculate the distribution through the k^{th} client model and global model of the $(t-1)^{th}$ round on private data: $Z_{i,k}^t = f(w_k^t, x_i)$ and $Z_i^{t-1} = f(w^{t-1}, x_i)$ for private data x_i w.r.t its ground truth label y_i. The standard KD loss function of k^{th} client can be formulated as:

$$\mathcal{L}_{skd}(Z_{i,k}^t, Z_i^{t-1}) = \sigma(Z_i^{t-1}) \log(\frac{\sigma(Z_i^{t-1})}{\sigma(Z_{i,k}^t)}), \tag{3}$$

where σ denotes softmax function. The optimization objective is to mitigate the issues such as catastrophic forgetting and data heterogeneity problem.

KD Based on reverse KLD. KD based on minimizing reverse KLD has been widely applied in specific scenarios due to its mode-seeking characteristics [10, 20, 56]. Unlike standard KD, its distillation function can be expressed as:

$$\mathcal{L}_{rkd}(Z_{i,k}^t, Z_i^{t-1}) = \sigma(Z_{i,k}^t) \log(\frac{\sigma(Z_{i,k}^t)}{\sigma(Z_i^{t-1})}). \tag{4}$$

Research [10] suggests that due to its mode-seeking characteristics, this distillation method is more suitable for complex tasks compared to standard distillation methods.

3.2 Dynamic Knowledge Distillation (DKD)

Definition 3.1. (Knowledge Modules) *We take digits in Z in descending order and then cumulatively summed until the number of selected values surpasses μ, where μ is a hyperparameter and typically defined as 0.5. The selected values are defined as Dominant Knowledge Components (DKCs), while the remaining values are termed as Ancillary Knowledge Components (AKCs), formulated as:*

$$a(j) = \begin{cases} 0 & if z_{j,k}^t \in AKCs \\ 1 & if z_{j,k}^t \in DKCs \end{cases}, \tag{5a}$$

$$\min \sum_{j=0}^{n} a(j) \quad s.t. \quad \sum_{j=0}^{n} a(j) z_{j,k}^t \geq \mu. \tag{5b}$$

Definition 3.2. (Knowledge Discrepancy) *Knowledge discrepancy reflects the distance between two distributions. The knowledge discrepancy γ_s within AKCs and γ_l within DKCs are defined as:*

$$\gamma_s = \sum_{j=1}^{n} (1 - a(j)) |z_j^{t-1} - z_{j,k}^t|, \tag{6a}$$

$$\gamma_l = \sum_{j=1}^{n} a(j) |z_j^{t-1} - z_{j,k}^t|. \tag{6b}$$

Theoretical Analysis. Forward KLD's mean-seeking characteristics result in unreliable distillation when the global model has multi-peaked distributions. Conversely, reverse KLD will also lead to distinct bias due to its mode-seeking nature. Forward KLD and reverse KLD are adept at fitting different regions of the distribution. Therefore, how to balance forward and reverse KLD to minimize distilation bias becomes a key issue. This naturally leads us to reflect on the fundamental reasons behind the different behaviors of forward and reverse KLD. Let $Z_{i,k}^t = (z_{1,k}^t, z_{2,k}^t, ..., z_{n,k}^t)$ and $Z_i^{t-1} = (z_1^{t-1}, z_2^{t-1}, ..., z_n^{t-1})$, where n denotes the size of Z. The $\mathcal{L}_{skd}$ and $\mathcal{L}_{rkd}$ can be denoted as:

$$\mathcal{L}_{skd} = \sum_k z_j^{t-1} \log(\frac{z_j^{t-1}}{z_{j,k}^t}), \tag{7a}$$

$$\mathcal{L}_{rkd} = \sum_k z_{j,k}^t \log(\frac{z_{j,k}^t}{z_j^{t-1}}). \tag{7b}$$

The gradient for $z_{j,k}^t$ under forward and reverse KLD can be calculated by the chain rule as follows:

$$\frac{\partial \mathcal{L}_{skd}}{\partial z_{j,k}^t} = z_{j,k}^t - z_j^{t-1}, \tag{8a}$$

$$\frac{\partial \mathcal{L}_{rkd}}{\partial z_{j,k}^t} = z_{j,k}^t \log(\frac{z_{j,k}^t}{z_j^{t-1}}) - \mathcal{L}_{rkd}. \tag{8b}$$

Considering the converge condition of forward and reverse KLD:

$$\frac{\partial \mathcal{L}_{s(r)kd}}{\partial z_{j,k}^t} = 0, \forall j \in (1, 2, 3..., n), \tag{9}$$

we can infer that for both two methods, the sufficient and necessary condition for converge is:

$$z_{j,k}^t = z_j^{t-1}, \forall j \in (1, 2, 3..., n). \tag{10}$$

According to Eq.(10), both the forward and reverse KLD have the same optimal projective. Thus, the fundamental reason for their differing behaviors is the optimization process. Considering Eq.(7a), larger z_j^{t-1} means a larger weight in total loss and also more likely to generate a larger $\log(z_j^{t-1}/z_{j,k}^t)$. Hence, fitting the area with larger z_j^{t-1} is the priority of forward KLD. What's more, when $(z_j^{t-1}/z_{j,k}^t)$ goes to $+\infty$ the forward KLD goes to $+\infty$. Therefore, $z_{j,k}^t$ would try to cover as many peaks of z_j^{t-1} as possible, leading to the mean-seeking behavior of forward KLD. Similarly, considering Eq.(7b), $(z_{j,k}^t/z_j^{t-1})$ is easier to be $+\infty$ when z_j^{t-1} gets smaller, leading to a larger loss. Therefore, fitting the area with smaller z_j^{t-1} is the priority of reverse KLD. It avoids $(z_j^{t-1}/z_{j,k}^t)$ go to 0^+, which means $z_{j,k}^t$ shouldn't be too large when z_j^{t-1} is small, leading to mode-seeking behavior of reverse KLD.

Empirical Analysis. We perform distillation using forward and reverse KLD separately, calculating the average knowledge discrepancies between the client models and the global model across the two knowledge modules (DKCs and AKCs, defined in Eq.(5)) on Cifar-100 with 10 cilents for 100 communication epochs, as illustrated in Fig.2. The experimental results indicate that when using forward KLD, the differences in the DKCs are lower, while when using reverse KLD, the differences in the AKCs are lower.

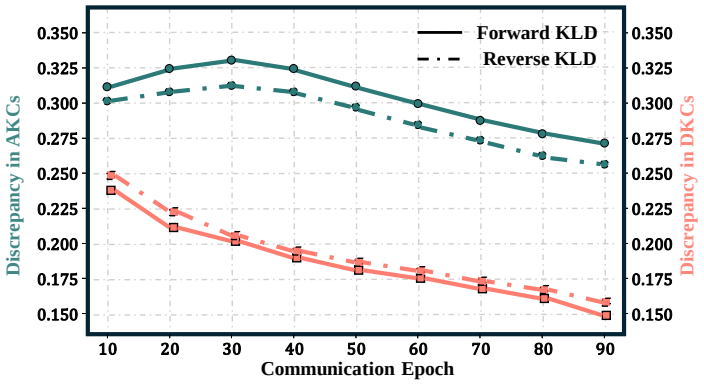


Figure 2: **Empirical Analysis.** The dashed line represents distillation based on reverse KLD, while the solid line denotes distillation based on forward KLD. For detailed information, please refer to the left main text.

Method. Based on these observations and analysis, it is intuitive to assign weight to the forward and the reverse KLD according to the knowledge discrepancies between the client model and the teacher model. In cases of disparity in DKCs, prioritize the forward KLD; Conversely, when there is a significant difference in AKCs, prioritize the reverse KLD. Combined with Eq.(3) and Eq.(4), the dynamic knowledge distillation loss function $\mathcal{L}_{dkd}$ is as follows:

$$\mathcal{L}_{dkd}(Z_{i,k}^t, Z_i^{t-1}) = \frac{\gamma_s}{\gamma_s + \gamma_l} \mathcal{L}_{rkd}(Z_{i,k}^t, Z_i^{t-1}) + \frac{\gamma_l}{\gamma_s + \gamma_l} \mathcal{L}_{skd}(Z_{i,k}^t, Z_i^{t-1}), \tag{11}$$

3.3 Knowledge Decoupling (KDP)

In prior knowledge distillation approaches, the aggregated global model has catastrophic accuracy drops, resulting in poor performance in some domains. Therefore, global model can not provide reliable guidance in these domains. We address this issue through **Knowledge Decoupling**: Decoupling

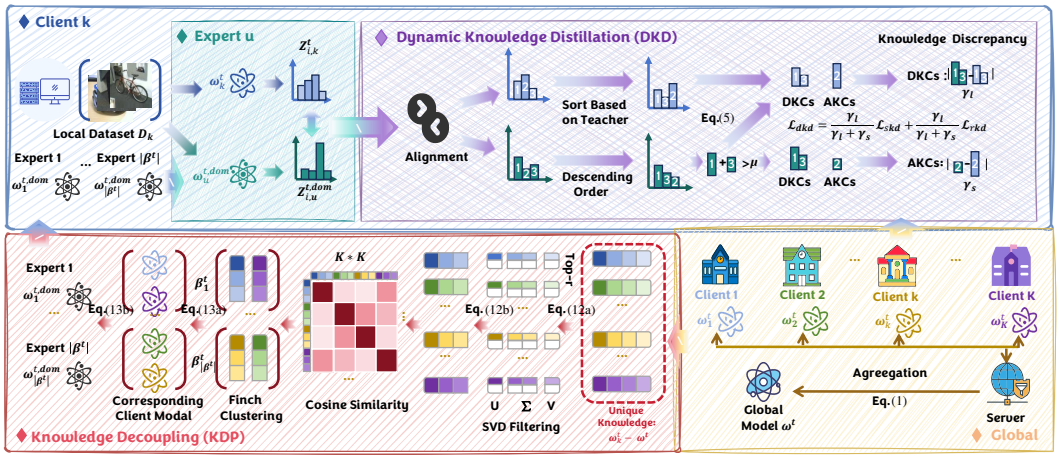


Figure 3: **Architecture illustration** of DKDR. DKDR consists of two core components: ❶ The top right box refers to Dynamic Knowledge Distillation (DKD), which adaptively weights forward and reverse KLD based on knowledge discrepancies in DKCs and AKCs (Sec.3.2). ❷ The bottom left box represents Knowledge Decoupling (KDP), where we separate shared and unique knowledge by SVD filtering and clustering to identify domain experts (Sec.3.3). Clients will distill knowledge dynamically and equally from each expert.

knowledge into shared knowledge and unique knowledge globally, and extracting domain signals from unique knowledge by SVD. Then we use Finch Clustering to get domain experts. Thus, Clients distill knowledge from these experts more efficiently.

Specifically, to get domain experts, we first examine the knowledge distillation process at a finer-grained knowledge perspective. We identify two types of critical knowledge: (1) **Shared knowledge**, which benefits **multiple domains**, and (2) **Unique knowledge**, which is useful only for a **specific domain**. In federated knowledge distillation, the mixing of shared knowledge and unique knowledge obscures domain-specific signals coming from unique knowledge. A natural idea is to separate shared knowledge and unique knowledge to clarify domain-specific signals. Therefore, during the t^{th} communication epoch, we consider the global model from the $(t-1)^{th}$ communication epoch as a natural placeholder to encapsulate the shared knowledge (denoted as w^{t-1}). Then we calculate the difference vector for each client w_k^t: $v_k^t = w_k^t - w^{t-1}$. This subtraction vector preserves domain-specific signals while diminishing the interference of shared knowledge. For practical use, we apply SVD [1] to filter redundant noise and clarify domain-specific signals within unique knowledge:

$$(v_1^t, v_2^t, \dots, v_K^t) = \quad \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{12a}$$

$$(v_1^{t,S}, v_2^{t,S}, \dots, v_K^{t,S}) = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^T. \tag{12b}$$

We apply truncated SVD to $(v_1^t, v_2^t, \dots, v_K^t)$, retaining the top r singular values to extract domain-specific signals. Then we use Finch Clustering on $(v_1^{t,S}, v_2^{t,S}, ..., v_K^{t,S})$ based on cosine similarity to capture domain-specific signals in unique knowledge. Then, we replace each v_k^S in each cluster with the corresponding client k to obtain client clusters β^t, where each $\beta_u^t \in \beta^t$ corresponds to a domain and contains clients belonging to this domain. Next we aggregate the clients within each β_u^t to identify domain experts $w_u^{t,dom}$. The process can be formulated as:

$$\begin{aligned} \boldsymbol{v} =& [\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \boldsymbol{v}_4, \boldsymbol{v}_5, \boldsymbol{v}_6] \\ \Downarrow\ & \mathit{Cluster} \\ =& [\underbrace{\boldsymbol{v}_1, \boldsymbol{v}_2}_{\text{Domain 1}}, \underbrace{\boldsymbol{v}_3, \boldsymbol{v}_4}_{\text{Domain 2}}, \underbrace{\boldsymbol{v}_5, \boldsymbol{v}_6}_{\text{Domain 3}}] \\ \Downarrow\ & \mathit{Map\ to\ client\ models} \\ \boldsymbol{\omega} =& [\underbrace{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2}_{\beta_1}, \underbrace{\boldsymbol{\omega}_3, \boldsymbol{\omega}_4}_{\beta_2}, \underbrace{\boldsymbol{\omega}_5, \boldsymbol{\omega}_6}_{\beta_3}], \end{aligned} \tag{13a}$$

$$w_u^{t,dom} = \frac{\sum\limits_{w_k^{t-1} \in \beta_u} \alpha_k w_k^{t-1}}{\sum\limits_{w_k^{t-1} \in \beta_u} \alpha_k}. \tag{13b}$$

Thus clients can get rich domain knowledge from these domain experts. We calculate the logits output through k^{th} client model and each domain expert $w_u^{t,dom}$ on private data x_i w.r.t its ground truth label y_i: $Z_{i,k}^t = f(w_k^t, x_i)$ and $Z_{i,u}^{t,dom} = f(w_u^{t,dom}, x_i)$. By inserting $Z_{i,k}^t$ and $Z_{i,u}^{t,dom}$ into the Eq.(11), We get the final defined knowledge distillation loss function $\mathcal{L}_{fkd}$:

$$\mathcal{L}_{fkd} = \sum_u \frac{1}{|\beta^t|}[\mathcal{L}_{dkd}(Z_{i,k}^t, Z_{i,u}^{t,dom})] \tag{14}$$

Eq.(14) assigns each expert the same weight thus mitigating the issue of domain skew. This dynamic knowledge distillation method reliably and efficiently distills knowledge for each domain. Combined with the cross-entropy loss $\mathcal{L}_{CE}$, the local loss of client k is now defined as:

$$\mathcal{L}_k = \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_k}(\mathcal{L}_{CE} + c\mathcal{L}_{fkd}), \tag{15}$$

where c represents the knowledge distillation intensity of the method.

3.4 Discussion and Limitation

Clustering Technical. A variety of clustering techniques have been proposed to discover natural grouping [55, 7, 8, 50, 33, 45]. The well-known methods, K-Means [37] and DBSCAN iteratively assign points to a fixed group number. However, they are sensitive to hyper-parameter selection under different scenarios. Thus we shift the gaze towards FINCH [45], which is parameter-free and thus suitable for heterogeneous federated learning. Specifically, we leverage the cosine similarity metric to evaluate the distance between any two client weights and view the weight with minimum distance as its "neighbor", sorted into the same set. After clustering, we aggregate all clients in the same cluster as they have related domain knowledge in order to get domain experts. We compare FINCH [45] with the well-known clustering methods, K-Means [37] and DBSCAN [8]. The results are shown in Tab.1.

Table 1: **Ablation** on popular clustering methods. Please refer to Sec.3.4 for details.

Methods	Office31			
	A	W	D	AVG
K-means	66.89	44.31	28.24	46.48
DBSCAN	65.28	36.82	28.76	43.62
FINCH (ours)	**67.06**	**54.82**	**29.84**	**50.57**

Conceptual Difference. Unlike conventional federated KD methods that depend solely on forward KLD and a single teacher model, DKDR introduces two distinctive innovations: dynamic KLD weighting and the use of multiple domain experts. The dynamic weighting reduces distillation bias by adapting to knowledge discrepancies, which is a departure from the static approaches of methods like FedNTD [25] or FedX [11]. It ensures precise alignment with the target distribution across diverse domains. Furthermore, by identifying domain experts, DKDR provides tailored guidance to clients, overcoming the limitation that the single teacher model may underperform in specific domains. This multi-expert paradigm enhances reliability and performance in multi-domain FL scenarios.

Limitation. While DKDR effectively enhances the reliability of distillation pathways and teacher models in multi-domain federated learning, it is not without drawbacks. The dynamic weighting mechanism involves SVD and clustering techniques, adding computational overhead that may significantly prolong training, particularly on resource-limited clients. Additionally, DKDR assumes that domains are sufficiently distinct for clustering to accurately identify experts; if domains overlap significantly, this assumption falters, potentially degrading overall performance.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate DKDR on two single-domain scenarios and two multi-domain scenarios.

- **Cifar-10** [23] contains 50k training images and 10k test images with 32×32 for 10 classes.
- **Cifar-100** [23] contains $50k$ and $10k$ images with 32×32 for 100 classes.
- **Office31** [44] consists of three domains: Amazon (A), Webcam (W) and DSLR (D). In total, the dataset contains 4,110 images with 256×256 across 31 categories, shared among the three domains.
- **Office Home** [51] consists of four domains: Art (A), Clipart (C), Product (P), and Real World (R). It contains 15,500 images with 256×256 across 65 categories, shared among the four domains.

Data Heterogeneity. As for the data heterogeneity simulation, we utilize the Dirichlet distribution, $Dir(\zeta)$ to simulate the label skew, as previous methods [29, 28, 63]. The smaller ζ is, the more imbalanced the local distribution is.

Counterparts. We compare our method with state-of-the-art (SOTA) federated knowledge distillation and federated learning approaches: FedAvg [39], FedProx [29], FedDyn [2], Scaffold [19], FedProto [49],MOON [28], FedNTD [25],FedDf [32].

Implement Details. We conduct communication epoch for $E = 200$ and local updating round $T = 5$, where all federated learning approaches have little or no accuracy gain with more communications. We use the SGD optimizer with the learning rate $lr = 1e-3$. The corresponding weight decay is $1e-5$ and momentum is 0.9. The training batch size is 64 for single-domain tasks and 16 for multi-domain tasks. The client number K is 20 for different datasets. We conduct experiments with ResNet-10 [12] on single-domain scenarios and ResNet-18 [12] on multi-domain scenarios. We fix the random seed to ensure reproduction and conduct experiments on the NVIDIA 3090Ti.

Table 2: **Comparison with the state-of-the-art method in the Office31 and Office Home** with domain skew. Best in bold and second with underline. Please refer to Sec.4.2 for further explanations.

Methods	Office31				Office Home				
	A	W	D	AVG	A	C	P	R	AVG
FedAvg	48.04	32.91	22.45	34.47	39.24	61.29	74.58	58.45	58.39
FedProx	45.55 ↓ 2.49	26.58 ↓ 6.33	19.39 ↓ 3.06	30.51 ↓ 3.96	39.67 ↑ 0.43	61.12 ↓ 0.17	74.52 ↓ 0.06	59.66 ↑ 1.21	58.74 ↑ 0.35
FedDyn	56.58 ↑ 8.54	20.25 ↓ 12.66	23.47 ↑ 1.02	33.43 ↓ 1.04	39.26 ↑ 0.02	60.21 ↓ 1.08	73.84 ↓ 0.74	59.08 ↑ 0.63	58.10 ↓ 0.29
Scaffold	46.80 ↓ 1.24	34.18 ↑ 1.27	23.47 ↑ 1.02	34.82 ↑ 0.35	39.88 ↑ 0.64	61.24 ↑ 0.05	74.75 ↑ 0.17	58.85 ↓ 0.60	58.68 ↑ 0.29
FedProto	51.60 ↑ 3.56	36.71 ↑ 3.80	**31.63** ↑ 9.18	39.98 ↑ 5.51	40.29 ↑ 1.05	63.65 ↑ 2.36	75.31 ↑ 0.73	60.34 ↑ 1.89	59.90 ↑ 1.51
MOON	49.47 ↑ 1.43	32.28 ↓ 0.63	26.53 ↑ 4.08	36.09 ↑ 1.62	38.78 ↓ 0.46	62.09 ↑ 0.80	74.27 ↓ 0.31	58.90 ↑ 0.45	58.51 ↑ 0.12
FedNTD	48.22 ↑ 0.18	35.44 ↑ 2.53	22.45 ↓ 0.00	35.37 ↑ 0.90	39.28 ↑ 0.04	63.07 ↑ 1.78	75.20 ↑ 0.62	59.20 ↑ 0.75	59.18 ↑ 0.79
FedDf	45.26 ↓ 2.78	32.78 ↓ 0.13	24.72 ↑ 2.27	34.25 ↓ 0.22	37.93 ↓ 1.31	60.53 ↓ 0.76	71.94 ↓ 2.64	57.83 ↓ 0.62	57.06 ↓ 1.33
DKDR	**67.06** ↑ 19.02	**54.82** ↑ 21.91	29.84 ↑ 7.39	**50.57** ↑ 16.10	**42.68** ↑ 3.44	**65.99** ↑ 4.70	**78.22** ↑ 3.64	**63.03** ↑ 4.58	**62.48** ↑ 4.09

4.2 Comparison to State-of-the-Arts

Performance Comparison The Tab.2 , Tab.3 and Tab.4 present the final accuracy metric by the end of the federated learning process with popular SOTA methods. It depicts that our method outperforms all other baselines in seven out of the eight settings, which confirms that the knowledge distillation of DKDR is reliable, efficient, and possesses superior domain generalization capabilities.

Table 3: **Comparison with the state-of-the-art method in the Cifar-10** with skew ratio $\zeta \in \{0.1, 0.3, 0.5\}$. Please refer to Sec.4.2 for details.

Methods	Cifar-10		
	$\zeta = 0.1$	$\zeta = 0.3$	$\zeta = 0.5$
FedAvg	76.91	79.86	80.34
FedProx	70.43 ↓ 6.48	74.14 ↓ 5.72	75.11 ↓ 5.23
FedDyn	78.77 ↑ 1.86	80.79 ↑ 0.93	81.26 ↑ 0.92
Scaffold	**79.62** ↑ 2.71	80.99 ↑ 1.13	81.40 ↑ 1.06
FedProto	78.45 ↑ 1.54	80.13 ↑ 0.27	81.49 ↑ 1.15
MOON	75.24 ↓ 1.67	79.83 ↓ 0.03	80.71 ↑ 0.37
FedNTD	77.18 ↑ 0.27	80.36 ↑ 0.50	80.94 ↑ 0.60
FedDf	78.53 ↑ 1.62	80.24 ↑ 0.38	81.37 ↑ 1.03
DKDR	79.46 ↑ 2.55	**81.93** ↑ 2.07	**82.63** ↑ 2.29

Table 4: **Comparison with the state-of-the-art method in the Cifar-100** with skew ratio $\zeta \in \{0.1, 0.3, 0.5\}$. Please refer to Sec.4.2 for details.

Methods	Cifar-100		
	$\zeta = 0.1$	$\zeta = 0.3$	$\zeta = 0.5$
FedAvg	43.76	46.66	48.57
FedProx	33.92 ↓ 9.84	37.81 ↓ 8.85	39.79 ↓ 8.78
FedDyn	46.21 ↑ 2.45	48.82 ↑ 2.16	50.23 ↑ 1.66
Scaffold	46.32 ↑ 2.56	50.33 ↑ 3.67	51.76 ↑ 3.19
FedProto	44.21 ↑ 0.45	49.88 ↑ 3.22	51.34 ↑ 2.77
MOON	42.96 ↓ 0.80	45.73 ↓ 0.93	48.58 ↑ 0.01
FedNTD	44.12 ↑ 0.36	47.10 ↑ 0.44	48.99 ↑ 0.42
FedDf	45.12 ↑ 1.36	46.62 ↓ 0.04	48.97 ↑ 0.40
DKDR	**46.80** ↑ 3.04	**51.84** ↑ 5.18	**53.18** ↑ 4.61

Convergence Analysis Fig.4 shows the curves of the average test accuracy during the training process across three random runs of three datasets (Cifar-100, Office31, Office Home) representing single-domain and multi-domain scenarios, including the results of various baselines. Traditional FL methods such as FedAvg performs poorly in heterogeneous scenarios while methods designed specifically for heterogeneous problem such as Scaffold and FedProto achieve much better performance.

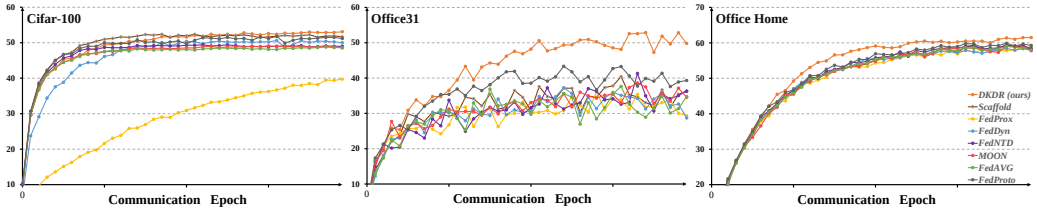
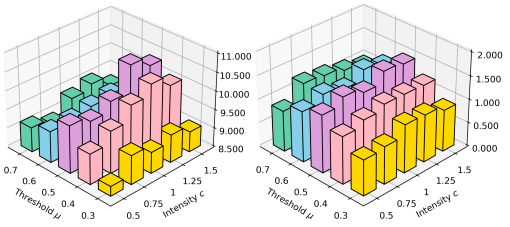


Figure 4: **Visualization** of training curves of the average test accuracy of DKDR and various baselines on three datasets (Cifar-100, Office31, Office Home). Please refer to Sec.4.2 for further explanations.

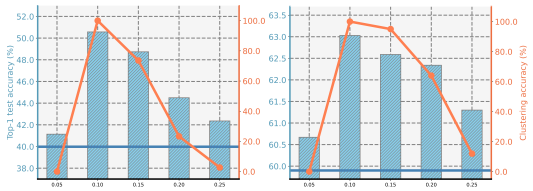
4.3 Sensitivity

Hyper Parameters c and μ. In the single-domain task Cifar-100 and the multi-domain task Office31, the optimal values of c and μ remain stable at 1.25 and 0.5 across different scenarios. This aligns with our theory (Sec.3.2): when $\mu < 0.5$, DKD gradually degenerates into forward KLD, while when $\mu > 0.5$, it gradually degenerates into reverse KLD. Both cases introduce different biases.



(a) Office31 for μ and c (b) Cifar-100 for μ and c

Figure 5: **Sensitivity analysis for μ and c** on Office31 and Cifar-100. The z-axis represents the performance improvement relative to the baselines. Please refer to Sec.4.3 for further detailed explanations.



(a) Office31 for r (b) Office Home for r

Figure 6: **Sensitivity analysis for r** across two multi-domain tasks by Top-1 test Accuracy and Clustering accuracy. The horizontal line represents the baseline. Please refer to Sec.4.3 for further explanations.

Hyper Parameter r. In assessing the filtering strength r of SVD, we utilize the average accuracy of clustering within each domain additionally. If a cluster includes clients from more than one domain, it is considered a clustering failure for all clients in that cluster. As shown in Fig.6a and Fig.6b, the optimal setting of r remains stable at 0.1 for both two datasets. It is worth noting that when too few singular values are retained, the domain signals tend to converge towards a common low-dimensional subspace, which can lead to the failure of clustering, just as when $r = 0.05$.

4.4 Effectiveness.

Effects of Key Components Mechanism of DKDR. To substantiate its robustness and stability, we meticulously evaluate the performance across both single-domain and multi-domain scenarios. As illustrated in Tab.5, compared to FedAVG, DKD achieves a consistent performance enhancement in both tasks, whereas KDP demonstrates more pronounced effectiveness in complex multi-domain tasks, as domain-specific experts can provide reliable guidance for clients in each domain.

Ablation Study of DKD. We compare DKD with KD using forward and reverse KLD separately to validate the effectiveness of DKD. As shown in Tab.6, DKD is more suitable for federated learning tasks in both single-domain and multi-domain tasks compared to using forward or reverse KLD alone.

Table 5: Ablation study of the key components in DKDR on four datasets (Office31, Office Home, Cifar-10, Cifar-100). Please see Sec.4.4 for details.

DKD	KDP	Office31	Office Home	Cifar-10	Cifar-100
✗	✗	34.47	58.39	79.86	46.66
✓	✗	36.83	59.52	80.72	48.79
✗	✓	48.59	61.03	80.63	50.33
✓	✓	**50.57**	**62.48**	**81.23**	**51.84**

Table 6: Ablation study of the DKD of four datasets (Office31, Office Home, Cifar-10, Cifar-100). Please refer to Sec.4.4 for further detailed explanations.

FKD	RKD	DKD	Office31	Office Home	Cifar-10	Cifar-100
✗	✗	✗	34.47	58.39	79.85	46.66
✓	✗	✗	34.88	58.70	80.59	47.52
✗	✓	✗	35.67	58.61	80.13	48.03
✗	✗	✓	**36.83**	**59.52**	**80.72**	**48.79**

5 Conclusion

In this paper, we address two significant problems in existing federated knowledge distillation methods: the unreliability of distilling pathway and teacher model. We empirically and theoretically analyze the fundamental differences between forward and reverse KLD, which leads us to propose a dynamic distillation approach that minimize distillation bias. To get reliable guidance, we employed knowledge decoupling to identify domain experts. Based on these insights, we propose the DKDR framework, which is strategically designed to achieve robust performance across diverse tasks. The effectiveness of DKDR has been validated with many sota methods over various classification tasks.

Acknowledgement

This work is supported by National Natural Science Foundation of China under Grant (62361166629, 623B2080, 62506269), the Major Project of Science and Technology Innovation of Hubei Province (2024BCA003, 2025BEA002), and the Innovative Research Group Project of Hubei Province under Grants 2024AFA017. The supercomputing system at the Supercomputing Center of Wuhan University supported the numerical calculations in this paper. This research was financially supported by the Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), under Grant No.GML-KF-24-10.

References

[1] Hervé Abdi. Singular value decomposition (svd) and generalized singular value decomposition. *Encyclopedia of measurement and statistics*, 907(912):44, 2007.

[2] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *ICLR*, 2021.

[3] Alan Chan, Hugo Silva, Sungsu Lim, Tadashi Kozuno, A Rupam Mahmood, and Martha White. Greedification operators for policy optimization: Investigating forward and reverse kl divergences. *Journal of Machine Learning Research*, 23(253):1–79, 2022.

[4] Huajun Chen. Large knowledge model: Perspectives and challenges. *arXiv preprint arXiv:2312.02706*, 2023.

[5] Yiqiang Chen, Wang Lu, Xin Qin, Jindong Wang, and Xing Xie. Metafed: Federated learning among federations with cyclic knowledge distillation for personalized healthcare. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[6] Zongyi Chen, Sanchuan Guo, Liyan Shen, Xi Zhang, and Zhuonan Chang. Improving knowledge distillation for federated learning on non-iid data. In *2023 IEEE International Conference on Big Data (BigData)*, pages 598–607. IEEE, 2023.

[7] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE TIT*, pages 21–27, 1967.

[8] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *ACM SIGKDD*, pages 226–231, 1996.

[9] Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. Feddc: Federated learning with non-iid data via local drift decoupling and correction. In *CVPR*, 2022.

[10] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[11] Sungwon Han, Sungwon Park, Fangzhao Wu, Sundong Kim, Chuhan Wu, Xing Xie, and Meeyoung Cha. Fedx: Unsupervised federated learning with cross knowledge distillation. In *ECCV*, 2022.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[14] Wenke Huang, Mang Ye, and Bo Du. Learn from others and be yourself in heterogeneous federated learning. In *CVPR*, 2022.

[15] Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. Rethinking federated learning with domain shift: A prototype view. In *CVPR*, pages 16312–16322, 2023.

[16] Wenke Huang, Mang Ye, Zekun Shi, Guancheng Wan, He Li, Bo Du, and Qiang Yang. A federated learning for generalization, robustness, fairness: A survey and benchmark. *TPAMI*, 2024.

[17] Hai Jin, Dongshan Bai, Dezhong Yao, Yutong Dai, Lin Gu, Chen Yu, and Lichao Sun. Personalized edge intelligence via federated self-knowledge distillation. *IEEE Transactions on Parallel and Distributed Systems*, 34(2):567–580, 2022.

[18] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

[19] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for on-device federated learning. In *ICML*, 2020.

[20] Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*, 2024.

[21] Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

[22] Zhiqiang Kou, Si Qin, Hailin Wang, Mingkun Xie, Shuo Chen, Yuheng Jia, Tongliang Liu, Masashi Sugiyama, and Xin Geng. Label distribution learning with biased annotations by learning multi-label representation, 8 2025. Main Track.

[23] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[24] Huy Q Le, Minh NH Nguyen, Chu Myaet Thwal, Yu Qiao, Chaoning Zhang, and Choong Seon Hong. Fedmekt: Distillation-based embedding knowledge transfer for multimodal federated learning. *Neural Networks*, 183:107017, 2025.

[25] Gihun Lee, Minchan Jeong, Yongjin Shin, Sangmin Bae, and Se-Young Yun. Preservation of the global knowledge by not-true distillation in federated learning. In *NeurIPS*, 2022.

[26] Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation. In *NeurIPS Workshop*, 2019.

[27] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. In *ICDE*, pages 965–978, 2022.

[28] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *CVPR*, pages 10713–10722, 2021.

[29] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *MLSys*, 2020.

[30] Xiaoxiao Li, Yufeng Gu, Nicha Dvornek, Lawrence H Staib, Pamela Ventola, and James S Duncan. Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: Abide results. *MedIA*, page 101765, 2020.

[31] Xinjia Li, Boyu Chen, and Wenlian Lu. Feddkd: Federated learning with decentralized knowledge distillation. *Applied Intelligence*, 53(15):18547–18563, 2023.

[32] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. In *NeurIPS*, pages 2351–2363, 2020.

[33] Fuchang Liu, Yu Wang, Zheng Li, and Zhigeng Pan. Geikd: Self-knowledge distillation based on gated ensemble networks and influences-based label noise removal. *Computer Vision and Image Understanding*, 235:103771, 2023.

[34] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *CVPR*, pages 1013–1023, 2021.

[35] Kangyang Luo, Shuai Wang, Yexuan Fu, Xiang Li, Yunshi Lan, and Ming Gao. Dfrd: data-free robustness distillation for heterogeneous federated learning. *arXiv preprint arXiv:2309.13546*, 2023.

[36] Yuhang Ma, Zhongle Xie, Jue Wang, Ke Chen, and Lidan Shou. Continual federated learning based on knowledge distillation. In *IJCAI*, pages 2182–2188, 2022.

[37] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *BSMSP*, pages 281–297, 1967.

[38] Gales M Malinin A. Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. In *NeurIPS*, 2019.

[39] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, pages 1273–1282, 2017.

[40] Matias Mendieta, Taojiannan Yang, Pu Wang, Minwoo Lee, Zhengming Ding, and Chen Chen. Local learning matters: Rethinking data heterogeneity in federated learning. In *CVPR*, pages 8397–8406, 2022.

[41] Jiaxu Miao, Zongxin Yang, Leilei Fan, and Yi Yang. Fedseg: Class-heterogeneous federated learning for semantic segmentation. In *CVPR*, pages 8042–8052, 2023.

[42] Xutong Mu, Yulong Shen, Ke Cheng, Xueli Geng, Jiaxuan Fu, Tao Zhang, and Zhiwei Zhang. Fedproc: Prototypical contrastive federated learning on non-iid data. *arXiv preprint arXiv:2109.12273*, 2021.

[43] Xingchao Peng, Zijun Huang, Yizhe Zhu, and Kate Saenko. Federated adversarial domain adaptation. In *ICLR*, 2020.

[44] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226, 2010.

[45] M. Saquib Sarfraz, Vivek Sharma, and Rainer Stiefelhagen. Efficient parameter-free clustering using first neighbor relations. In *CVPR*, pages 8934–8943, 2019.

[46] Neta Shoham, Tomer Avidor, Aviv Keren, Nadav Israel, Daniel Benditkis, Liron Mor-Yosef, and Itai Zeitak. Overcoming forgetting in federated learning on non-iid data. In *NeurIPS Workshop*, 2019.

[47] Neta Shoham, Tomer Avidor, Aviv Keren, Nadav Israel, Daniel Benditkis, Liron Mor-Yosef, and Itai Zeitak. Overcoming forgetting in federated learning on non-iid data. In *NeurIPS Workshop*, 2019.

[48] Canh T. Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. In *NeurIPS*, pages 21394–21405, 2020.

[49] Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *AAAI*, 2022.

[50] Grant Van Horn, Rui Qian, Kimberly Wilber, Hartwig Adam, Oisin Mac Aodha, and Serge Belongie. Exploring fine-grained audiovisual categorization with the ssw60 dataset. In *ECCV*, 2022.

[51] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5018–5027, 2017.

[52] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, page 3152676, 2017.

[53] Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints. *arXiv preprint arXiv:2309.16240*, 2023.

[54] Haozhao Wang, Yichen Li, Wenchao Xu, Ruixuan Li, Yufeng Zhan, and Zhigang Zeng. Dafkd: Domain-aware federated knowledge distillation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 20412–20421, 2023.

[55] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *JASA*, pages 236–244, 1963.

[56] Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. f-divergence minimization for sequence-level knowledge distillation. *arXiv preprint arXiv:2307.15190*, 2023.

[57] Chen Wu, Sencun Zhu, and Prasenjit Mitra. Federated unlearning with knowledge distillation. *arXiv preprint arXiv:2201.09441*, 2022.

[58] Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Yongfeng Huang, and Xing Xie. Communication-efficient federated learning via knowledge distillation. *Nature communications*, 13(1):2032, 2022.

[59] Yuan-Yi Xu, Ci-Siang Lin, and Yu-Chiang Frank Wang. Bias-eliminating augmentation learning for debiased federated learning. In *CVPR*, pages 20442–20452, 2023.

[60] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM TIST*, pages 1–19, 2019.

[61] Dezhong Yao, Wanning Pan, Yutong Dai, Yao Wan, Xiaofeng Ding, Chen Yu, Hai Jin, Zheng Xu, and Lichao Sun. Fedgkd: Toward heterogeneous federated learning via global knowledge distillation. *IEEE Transactions on Computers*, 73(1):3–17, 2023.

[62] Mang Ye, Wei Shen, Bo Du, Eduard Snezhko, Vassili Kovalev, and Pong C Yuen. Vertical federated learning for effectiveness, security, applicability: A survey. *ACM Computing Surveys*, 57(9):1–32, 2025.

[63] Jie Zhang, Zhiqi Li, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Chao Wu. Federated learning with label distribution skew via logits calibration. In *ICML*, pages 26311–26329, 2022.

[64] Lin Zhang, Yong Luo, Yan Bai, Bo Du, and Ling-Yu Duan. Federated learning for non-iid data via unified feature learning and optimization objective alignment. In *ICCV*, pages 4420–4428, 2021.

[65] Lin Zhang, Li Shen, Liang Ding, Dacheng Tao, and Ling-Yu Duan. Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. In *CVPR*, pages 10174–10183, 2022.

[66] Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *International conference on machine learning*, pages 12878–12889. PMLR, 2021.

A Convergence Proof of DKDR

A.1 Assumptions

L-smoothness: For all k, $f_k(w)$ is L-smooth: $\|\nabla f_k(w_1) - \nabla f_k(w_2)\| \leq L\|w_1 - w_2\|$.

(Dynamic weights preserve smoothness; γ_s, γ_l changes Lipschitz: $|\gamma_s(w_1) - \gamma_s(w_2)| \leq \kappa\|w_1 - w_2\|$, yielding $L' \leq L + \kappa \max(G_{skd}, G_{rkd})$. Quasi-static: γ per round.)

Bounded gradients: $\|\nabla f_k(w)\|^2 \leq G^2$.

Bounded variance: $\mathbb{E}_k[\|\nabla f_k(w) - \nabla f(w)\|^2] \leq \sigma^2$.

Bounded distillation gradients: $\|\nabla L_k^{DKD}(w)\|^2 \leq D^2 \leq \max(G_{skd}^2, G_{rkd}^2)$ (logits clipped).

Non-convex: $f(w)$ lower-bounded by f^*.

Parameters: $\eta > 0$, rounds T, local steps E, $\eta \leq 1/(4L'E)$.

A.2 Theorem 1 (Convergence of DKD)

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(w^t)\|^2 \leq \frac{4(f(w^0) - f^*)}{\eta TE} + 6L'\eta(E-1)G^2 + \frac{\sigma^2}{K} + 8(L')^2 E\eta\left(G^2 + \frac{\Lambda(\gamma)}{K}\right) + O(\kappa^2\eta^2 EG^2),$$

where $\Lambda(\gamma) = \mathbb{E}_k[\|\nabla f_k(w) - \nabla f(w)\|^2]$.

Dynamic weights yield $\Lambda(\gamma) < \Lambda_0$ (static). For $\eta = O(\sqrt{K/(TEL')})$, RHS $O(1/\sqrt{KTE})$, converges to 0 as $T \to \infty$. Compared to FedAvg ($\Lambda = \sigma^2$) or static KLD, DKD tighter.

Key Lemma. Reduction in $\Lambda(\gamma)$ Forward KLD prioritizes γ_l (mean-seeking; $\partial L_{skd}/\partial z_{j,k} = z_{j,k} - z_j^{t-1}$); reverse prioritizes γ_s (mode-seeking; $\partial L_{rkd}/\partial z_{j,k} = z_{j,k}\log(z_{j,k}/z_j^{t-1}) - L_{rkd}$). Weights $\alpha = \gamma_l/(\gamma_s + \gamma_l)$ balance via gradient norm.

$$\nabla L_{dkd} = \frac{\gamma_s}{\gamma_s + \gamma_l}\nabla L_{rkd} + \frac{\gamma_l}{\gamma_s + \gamma_l}\nabla L_{skd}.$$

By Pinsker's (KL $\geq$ (1/2) TV^2):

$$\mathbb{E}[KL(Z_k\|Z_{global})] \approx \min_{path}\int\|\nabla L_{dkd}\|dt \leq \max(\mathbb{E}[KL_{skd}], \mathbb{E}[KL_{rkd}]) - \delta\frac{\gamma_s\gamma_l}{(\gamma_s + \gamma_l)^2},$$

$\delta > 0$ from complementarity. Specifically,

$$\|\nabla L_{dkd}\|^2 \leq \max(\|\nabla L_{skd}\|^2, \|\nabla L_{rkd}\|^2) - \beta\|\nabla L_{skd} - \nabla L_{rkd}\|^2\frac{\gamma_s\gamma_l}{(\gamma_s + \gamma_l)^2},$$

($\beta = \Theta(1)$). Thus,

$$\|\nabla f_k(w) - \nabla f(w)\|^2 \leq \Lambda_0 - \beta\frac{\gamma_s\gamma_l}{(\gamma_s + \gamma_l)^2},$$

$$\Lambda(\gamma) \leq \Lambda_0(1 - \epsilon)(\epsilon \approx 1/2 \text{ when balanced}) < \Lambda_0.$$

A.3 Proof Sketch

Step 1: Local:

$$w_k^{t,e+1} = w_k^{t,e} - \eta g_k^{t,e}, g_k^{t,e} = \nabla f_k(w_k^{t,e}).$$

Aggregate:

$$w^{t+1} = (1/K)\sum w_k^{t,E}.$$

Define

$$\bar{w}^{t,e} = (1/K)\sum w_k^{t,e}, \bar{g}^{t,e} = (1/K)\sum g_k^{t,e}.$$

Step 2: Descent:

$$\mathbb{E}[f(w^{t+1})] \leq f(w^t) + \langle \nabla f(w^t), w^{t+1} - w^t \rangle + (L'/2)\|w^{t+1} - w^t\|^2,$$

$$w^{t+1} - w^t = -\eta \sum_{e=0}^{E-1} \bar{g}^{t,e}/E,$$

yielding:

$$\leq f(w^t) - (\eta E/2)\|\nabla f(w^t)\|^2 + (\eta/(2E)) \sum_{e=0}^{E-1} \mathbb{E}\|\bar{g}^{t,e} - \nabla f(w^t)\|^2 + (L'\eta^2 E G^2/2)$$

.

Step 3: Bias:

$$\mathbb{E}\|\bar{g}^{t,e} - \nabla f(w^t)\|^2 \leq (\sigma^2 + \Lambda(\gamma))/K + 6(L')^2\eta^2 e(G^2 + \Lambda(\gamma)/K) + O(\kappa^2\eta^2 e G^2)$$

.

Step 4: Sum t=0 to T-1, divide by $\eta ET/2$; $\sum e \leq E^2/2$, constants to 8, yields bound.

B Notations Table

Symbol	Meaning	Symbol	Meaning
$Z_w(y \mid x)$	Local model distribution	$Z(y \mid x)$	Global model distribution
D_k	Private data of k-th client	N_k	Number of data points
w^t	Global model parameters at round t	w_k^t	Local model parameters of k-th client
$Z_{i,k}^t$	Client model output distribution	Z_i^{t-1}	Global model output distribution
σ	Softmax function	$\mathcal{L}_{\text{skd}}$	Standard KD loss
$\mathcal{L}_{rkd}$	Reverse KD loss	$a(j)$	Indicator function for DKCs/AKCs
μ	Threshold to define DKCs/AKCs	γ_s	Knowledge discrepancy in AKCs
γ_l	Knowledge discrepancy in DKCs	$\mathcal{L}_{dkd}$	Dynamic KD loss
v_k^t	Difference vector	$\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}^T$	SVD decomposition matrices
β^t	Client clusters	$w_v^{t,\text{dom}}$	Domain expert model
$\mathcal{L}_{fkd}$	Final KD loss	$\mathcal{L}_k$	Local loss of k-th client

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Please refer to Sec.1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to Sec.3.4.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Please refer to Sec.3.2.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please refer to Sec.4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is accessible in this paper.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Please refer to Sec.4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Please refer to Sec.4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please refer to Sec.4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Please see the supplementary file.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Justification: The research presented in this paper is foundational. It is not directly tied to any specific applications or deployments.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Please refer to Sec:4.1.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: None of the core methods in this paper rely on LLMs.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.