

Analysis and Tuning of Knowledge Distillation for Efficient Collaborative Learning

Norah Alballa
KAUST

Ahmed M. Abdelmoniem
Queen Mary University of London

Marco Canini
KAUST

Abstract—Knowledge Distillation (KD) has become a crucial technique for efficient collaborative learning in distributed systems, especially under data and system heterogeneity. Despite considerable attention, few works systematically examine tuning strategies for KD hyperparameters or provide comprehensive empirical comparisons across different data distributions. In this paper, we conduct an in-depth study of multiple KD methods—Vanilla KD, Deep Mutual Learning (DML), Data Partitioning KD (DP-KD), and our Tuned KD approach—using diverse data partitions and transfer sets. We show that (1) hyperparameter tuning dramatically boosts KD performance, especially when teacher–student models have disparate accuracies, (2) different transfer set sizes and labeling conditions substantially affect outcomes. Our findings serve as practical guidelines for effectively applying KD to collaborative learning tasks.

Index Terms—Knowledge Distillation, Collaborative Learning, Data Heterogeneity, Hyperparameter Tuning

I. INTRODUCTION

Knowledge Distillation (KD), originally proposed for model compression [1]–[3], has become a key technique for transferring knowledge among pre-trained models [4]–[7]. In decentralized or federated learning, these pre-trained models often face performance drops due to data (non-IID) and system (bandwidth/compute) heterogeneity. KD mitigates such issues by exchanging model outputs instead of raw data, thereby reducing communication overhead and supporting diverse model architectures [4]–[13].

However, the literature lacks a unified understanding of KD’s efficacy in heterogeneous conditions, particularly for *pre-trained* models. Existing works on vanilla KD, deep mutual learning (DML) [4], [9], or data partitioning KD [14] do not (to our knowledge) systematically compare them across a wide range of data distributions. Moreover, tuning critical hyperparameters (temperature and weight) is often neglected, despite evidence of notable improvements [15].

Hence, we conduct a thorough investigation of KD under multiple data-partition strategies and transfer-set choices. Our contributions include:

- A broad comparison of KD methods (Vanilla, DML, DP-KD) across five data partitions, using both labeled and unlabeled transfer sets, plus hyperparameter tuning.
- Demonstration that systematically tuning temperature and weight boosts accuracy in non-IID settings and clarifies when such tuning matters most.
- Analysis of teacher-selection criteria in different distributions.

II. BACKGROUND AND RELATED WORK

This section introduces the concept of collaborative learning, elaborates on the foundations of KD, and discusses related work in these fields.

A. Collaborative Learning

Collaborative learning enables multiple models to improve collectively by selectively exchanging knowledge rather than raw data [16]. It encompasses centralized strategies (e.g., Federated Learning [17]) and decentralized peer-to-peer approaches [11], [14], [16], [18], [19]. While it addresses data privacy and bandwidth constraints, significant hurdles persist due to system heterogeneity (varying hardware and network conditions) and data heterogeneity (non-IID distributions). Such variations can lead to suboptimal training efficiency and poor generalization [20]–[23].

B. Knowledge Distillation

KD was originally introduced to train a student model to replicate a more complex teacher model’s output probabilities, often smoothed with temperature T [1]. Formally, for sample l and class i , the teacher’s output is:

$$p_i^t(z_{t,l}, T) = \frac{\exp(z_{t,i,l}/T)}{\sum_j \exp(z_{t,j,l}/T)}, \quad (1)$$

where $z_{t,i,l}$ is the teacher’s logit for class i . The student is trained via a weighted combination of cross-entropy and Kullback–Leibler divergence:

$$\mathcal{L} = (1 - \alpha) \mathcal{L}_{CE}(p^s, y) + \alpha \mathcal{L}_{KL}(p^s, p^t), \quad (2)$$

where α controls the teacher’s relative contribution. KD can also use labeled or unlabeled *transfer sets* (public data or the student’s own data) to facilitate knowledge exchange.

Beyond compression, KD increasingly supports *pre-trained* teacher–student models that have encountered different data. This is especially relevant in decentralized and federated settings [5], [11], [14]. We focus on three key KD variants:

a) *Vanilla KD*: Uses Eq. (2) directly: the student distills knowledge from a single pre-trained teacher.

b) *Deep Mutual Learning (DML)*: Rather than a one-way teacher–student setup, two students s_1 and s_2 learn from each other:

$$\begin{aligned} \mathcal{L}_{s_1} &= \mathcal{L}_{CE}(p^{s_1}, y) + \mathcal{L}_{KL}(p^{s_1}, p^{s_2}), \\ \mathcal{L}_{s_2} &= \mathcal{L}_{CE}(p^{s_2}, y) + \mathcal{L}_{KL}(p^{s_2}, p^{s_1}), \end{aligned} \quad (3)$$

enabling a peer-to-peer learning process [4], [24], [25].

c) *Data Partitioning KD (DP-KD)*: DP-KD [14] partitions each data sample l based on which model is more confident—the teacher (t) or the student’s frozen state (st). Let m_l^t and m_l^{st} be binary masks; the distillation loss is

$$\mathcal{L}_{dist} = \frac{T^2}{n} \sum_{l=1}^n \left[m_l^t \mathcal{L}_{KL}(p^s, p^t) + m_l^{st} \mathcal{L}_{KL}(p^s, p^{st}) \right]. \quad (4)$$

In its unsupervised variant, the teacher or frozen student is selected purely by highest confidence, requiring no labels.

C. Related Work

KD’s ability to reduce communication overhead and enable heterogeneous architectures has inspired many collaborative learning frameworks [4], [7], [9], [10]. Approaches like FedMD [5], Cronus [11], and DS-FL [6] improve data-heterogeneous environments, while decentralized solutions (e.g., Def-KT [4] and ProxyFL [9]) remove central orchestration. However, most methods assess only one KD strategy or overlook hyperparameter tuning [15], and direct comparisons under varied data partitions remain limited. Our work fills this gap by providing the first large-scale head-to-head evaluation of four KD variants across five realistic non-IID partitions and three transfer-set sources, and by analyzing the impact of systematic (T, α) tuning.

III. METHODOLOGY

We seek to pinpoint *when and why* KD succeeds in heterogeneous learning environments. To that end we design a controlled one-teacher–one-student protocol that varies the following four independent factors while keeping every other training parameter fixed.

- 1) **KD algorithm.** *Vanilla KD* ($T=1, \alpha=0.5$), *DML* [4], *DP-KD* [14], and *Tuned KD*.
- 2) **Data partition.** Five canonical non-IID patterns: *Uniform*, *Quantity-Skew*, *Specialized Non-IID*, *Label-Skew (Chunks)*, and *Label-Skew (Dirichlet)* (details in §III-A).
- 3) **Transfer set.** (i) student data, (ii) public *unlabeled*, (iii) public *labeled*.
- 4) **Teacher strength.** Pre-KD gap $\Delta_{pre} = \text{acc}_t - \text{acc}_s$, later binned into *strong*, *peer*, and *weak*.

A. Evaluation Method

Experimental Setup. We consider $N = 10$ participants, each pre-training a ResNet-18 (up to 100 epochs) on its private dataset. One model (teacher) then distills knowledge to another (student) via various KD methods. The student’s dataset typically acts as the transfer set; we also experiment with public labeled or unlabeled data.

Data Partitioning Strategies. Each scenario yields 90 teacher–student pairs (excluding self-transfers):

- *Uniform*: Even splits among all classes.
- *Quantity Skew* [26]: Dataset sizes drawn from Dirichlet($\beta = 0.5$).
- *Specialized Non-IID*: $\sim 91\%$ from one dominant class, 9% from others.

- *Label Skew Random Chunks*: Random subsets of classes, chunked in blocks of 100 samples.
- *Label Skew Dirichlet*: Each class is distributed according to a Dirichlet vector (0.1/0.5).

Task, Models, and Datasets. We use CIFAR-10 (50k train, 10k test, plus 10% val) and also validate on CINIC-10 [27]. All participants train ResNet-18 with Adam (LR= 10^{-3} , weight decay= 4×10^{-4} , batch=32), stopping early if validation stalls for 10 epochs. Accuracy is measured on the test set.

KD Setup. We compare *Vanilla KD*, *DML*, *DP-KD* and *Tuned KD*, under three transfer-set options (public labeled, public unlabeled, or student data). By default, *Vanilla KD* sets $T = 1, \alpha = 0.5$. *Tuned KD* explores $\alpha \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$ and $T \in \{0.1, 0.5, 1, \dots, 7\}$. DP-KD runs unsupervised without labels; DML [4] trains two models jointly.

Metrics. We define *accuracy gain* = (post-KD) – (pre-KD). Negative values indicate a drop in performance.

IV. EVALUATION RESULTS

We compare KD methods under five data partitioning strategies and four transfer-set conditions. Fig 1 shows, for each teacher–student pair (x-axis: teacher minus student accuracy; y-axis: student’s accuracy gain after KD), how different KD approaches fare. *Positive knowledge transfers* (above the horizontal axis) indicate the student improved relative to its baseline, while *negative* ones suggest performance degradation. Transfer sets range from large (5,000 samples) to small (500 samples), either unlabeled or labeled, or the student’s own data.

A. Different Transfer Set Options

Finding (1): Larger transfer sets typically yield more positive knowledge transfers.

We observe that expanding the transfer set size improves KD results across all partitioning strategies (Fig. 1). Larger, more diverse sets better align the student’s parameters with the teacher’s [28], [29]. On the other hand, DML can exacerbate errors when both models lack labeled data, occasionally causing more *negative* transfers.

Finding (2): DML struggles without labeled data.

DML shows strong gains if the teacher is weaker and labeled or student data is available. However, in purely unlabeled scenarios, it can compound mutual errors, resulting in systematically lower accuracy gains compared to *Vanilla KD* or *DP-KD*.

B. DML and DP-KD vs. Vanilla KD

Finding (3): *Vanilla KD* often matches or surpasses *DP-KD*.

Vanilla KD’s simplicity proves robust. It consistently outperforms *DP-KD* in challenging setups like *Label Skew Dirichlet* and *Specialized Non-IID*, suggesting *DP-KD*’s vulnerability to overconfident teacher predictions in skewed distributions.

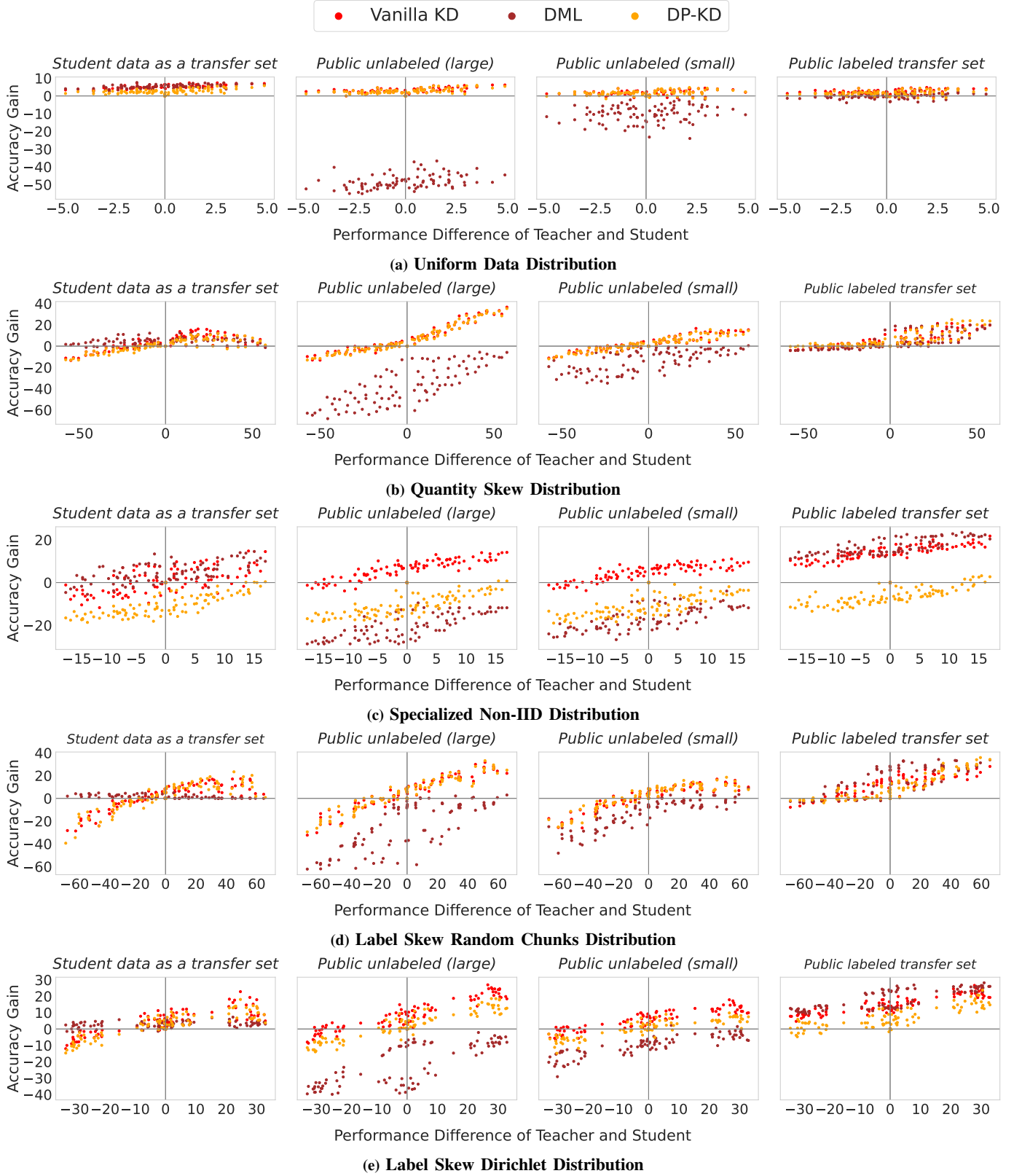


Fig. 1: Evaluation of KD techniques across various data partitioning strategies and transfer set options.

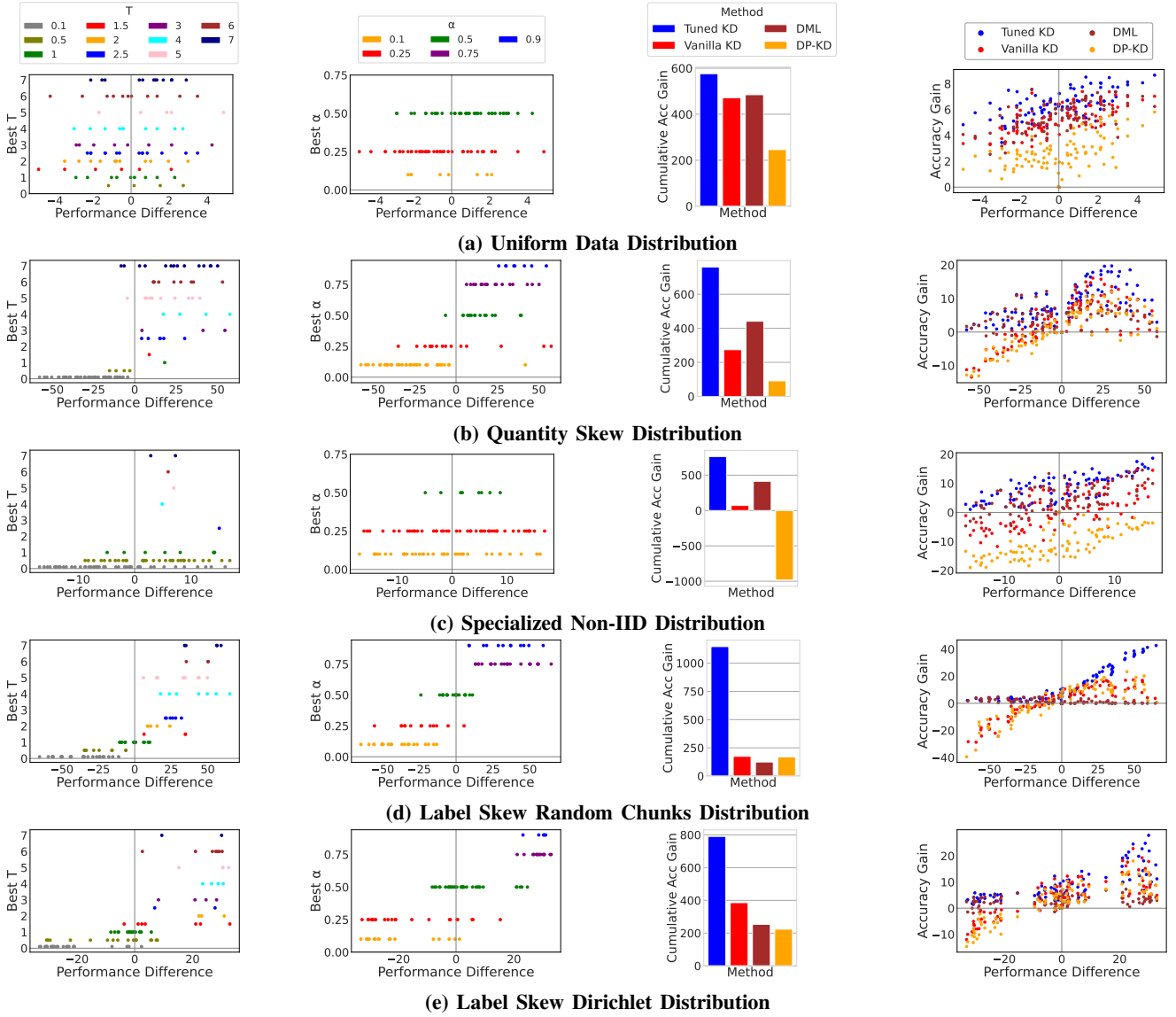


Fig. 2: Impact of varying T and α on KD performance across different data partitions (CIFAR-10). Columns show: (1) best T , (2) best α , (3) cumulative accuracy gains, and (4) per-pair gains.

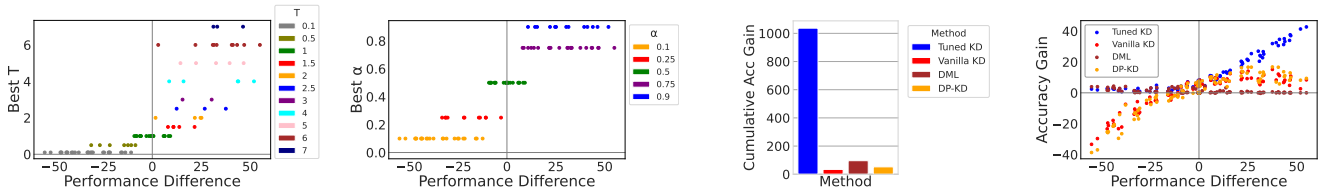


Fig. 3: Optimization of hyperparameters T and α in KD for CINIC-10 dataset using Label Skew Random Chunks Distribution.

Finding (4): DP-KD underperforms in Specialized Non-IID data.

Finding (5): Weak teacher + limited student data hamper Vanilla KD and DP-KD.

When one class dominates a participant's data, DP-KD often reinforces incorrect biases, especially if the teacher is overconfident in that majority class. This can harm the student's performance more than standard KD.

Both Vanilla KD and DP-KD rely on the teacher's correctness. If the teacher is substantially weaker than the student (or if the student's own data is small or unrepresentative), they risk negative transfers. DML, by contrast, can sometimes exploit mutual insights even when one model is weak, although it fails

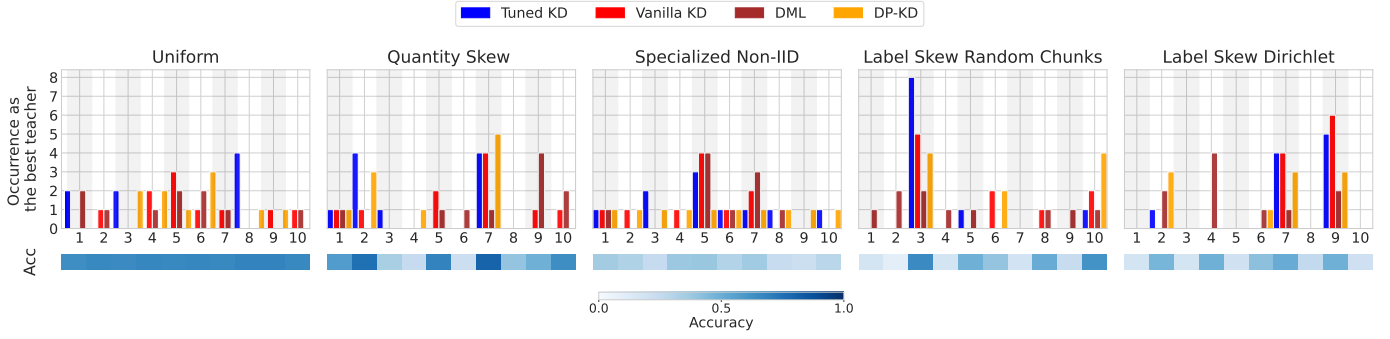


Fig. 4: Frequency of each participant being the “best teacher” (top row) vs. participant accuracy (bottom row). Results are shown for four KD approaches across five data distributions.

if neither model has labels for self-correction.

C. Optimizing Hyperparameters: Tuned KD

Our grid search over T and α (cf. Section III-A, $T \in 0.1 \dots 7, \alpha \in 0.1 \dots 0.9$) reveals that *Tuned KD* significantly improves performance relative to the fixed-parameter approaches. Unless otherwise noted, we use the student’s own data as the transfer set for these experiments.

Finding (6): Tuned KD outperforms Vanilla KD, DML, and DP-KD across all data partitions.

Fig. 2 (last column) shows that tuning T and α yields consistently higher *accuracy gains*, even in cases with weak teachers, where other methods often fail. This underscores the value of systematically optimizing KD hyperparameters.

Finding (7): Tuning matters most when teacher–student performance differs greatly.

When teacher and student accuracies are close, Tuned KD performs comparably to Vanilla KD. However, if the teacher is much stronger (or weaker), the correct T/α settings produce substantially higher accuracy gains.

Finding (8): Insights on best T/α combinations.

Across Quantity Skew, Label Skew, and Dirichlet splits, a lower temperature helps if the teacher is weaker, while a higher temperature (e.g., $T > 2$) and larger α benefit strong teachers. In Specialized Non-IID, both T and α should remain relatively small. Fig. 3 corroborates these trends on CINIC-10, suggesting their generality in diverse benchmarks.

D. Best Teachers Across KD Methods and Data Distributions

Finding (9): Participant accuracy alone does not ensure the best KD teaching.

Fig. 4 compares how often each participant becomes the top teacher (per accuracy gain) against its own accuracy. While highly accurate participants often serve as good teachers, certain partitioning scenarios reveal that other factors (e.g., data diversity) also matter. Notably, DML sometimes relies on teachers with slightly lower accuracy, hinting that richer

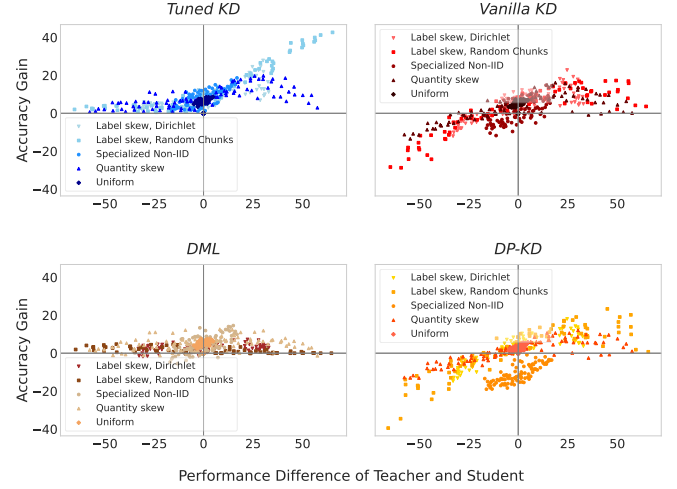


Fig. 5: Knowledge distillation techniques against varied data distribution characteristics.

class coverage or complementary knowledge may trump raw accuracy.

Overall, these observations suggest that teacher selection in KD-based collaborative learning can benefit from more nuanced metrics than simply picking the participant with the highest accuracy.

E. Influence of Data Distribution and Design Choices for KD

We re-examine the performance of various KD methods, including Tuned KD, under a different light and unveil new insights into the adaptability and robustness of these methods. Fig. 5 juxtaposes the efficacy of KD methods against varied data distributions. This comparative visualization allows us to discern patterns and outliers in performance, highlighting the consistent strength of Tuned KD across multiple scenarios. By contrasting these methods side-by-side, we better understand how each method responds to diverse data conditions.

Finally, based on the previous findings, we create a decision tree for a strategic guide towards selecting suitable KD methods based on specific use cases and data environments, offering a practical approach for applying KD methods in various contexts. In principle, we break down the decisions into questions regarding: 1) affordability of the tuning process; 2) the strength of the teacher; 3) ability to approximate the

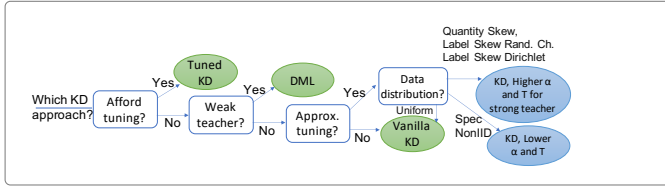


Fig. 6: Decision tree to determine the suitable KD approach.

tuning; and 4) the data distribution. These questions can help practitioners navigate KDs for their scenario to maximize the gains. The decision tree is shown in Fig. 6.

V. LIMITATION OF OUR WORK

Our study is deliberately scoped to *single teacher–student* distillation so that the impact of (T, α) , transfer-set source, and data partition can be isolated. Consequently, it does *not* address the practical dimension of multi-teacher or many-to-one settings. In addition, it does not address online or privacy-constrained tuning, where grid-search is too expensive. Exploring automatic hyper-parameter tuning, teacher selection, and communication-efficient protocols in cross-silo federated learning [30] is therefore left as promising future work.

VI. CONCLUSION AND FUTURE WORK

We presented a thorough study of KD for collaborative learning with pre-trained models, evaluating Vanilla KD, Tuned KD, DML, and DP-KD across multiple data partitions. Our experiments show that *Tuned KD* consistently achieves higher accuracy gains, especially when teacher–student performance differs widely. Moreover, systematic tuning of temperature and teacher-weight can boost accuracy by up to 20% points without extra communication. Overall, these findings provide practical insights for improving efficiency and scalability in distributed learning systems.

In our future work, we aim to study both multi-teacher and federated learning settings, where each client can act as both a teacher and a student. Such setups introduce challenges in coordinating knowledge transfer across heterogeneous and potentially unreliable participants. Also, decentralized or peer-to-peer KD schemes that respect FL constraints—e.g., by leveraging secure aggregation or differential privacy—could be explored. Moreover, there is a need for adaptive teacher selection based on model quality, data similarity, or network conditions to optimize distillation performance in dynamic FL.

REFERENCES

- [1] G. Hinton, O. Vinyals, and J. Dean, “Distilling the Knowledge in a Neural Network,” 2015, arXiv 1503.02531.
- [2] C. Buciluă, R. Caruana, and A. Niculescu-Mizil, “Model Compression,” in *KDD*, 2006.
- [3] J. Schmidhuber, *Neural Sequence Chunkers*. Inst. für Informatik, 1991.
- [4] C. Li, G. Li, and P. K. Varshney, “Decentralized Federated Learning via Mutual Knowledge Transfer,” *IEEE Internet of Things Journal*, 2021.
- [5] D. Li and J. Wang, “FedMD: Heterogeneous Federated Learning via Model Distillation,” 2019, arXiv 1910.03581.
- [6] S. Itahara, T. Nishio, Y. Koda, M. Morikura, and K. Yamamoto, “Distillation-Based Semi-Supervised Federated Learning for Communication-Efficient Collaborative Training With Non-IID Private Data,” *IEEE Transactions on Mobile Computing*, 2021.
- [7] F. Sattler, A. Marban, R. Rischke, and W. Samek, “CFD: Communication-Efficient Federated Distillation via Soft-Label Quantization and Delta Coding,” *IEEE Transactions on Network Science and Engineering*, 2021.
- [8] F. Sattler, T. Korjakow, R. Rischke, and W. Samek, “FedAUX: Leveraging Unlabeled Auxiliary Data in Federated Learning,” *TNNLS*, 2021.
- [9] S. Kalra, J. Wen, J. C. Cresswell, M. Volkovs, and H. R. Tizhoosh, “Decentralized federated learning through proxy model sharing,” *Nature Communications*, 2023.
- [10] L. Hu, H. Yan, L. Li, Z. Pan, X. Liu, and Z. Zhang, “MHAT: An efficient model-heterogeneous aggregation training scheme for federated learning,” *Information Sciences*, 2021.
- [11] H. Chang, V. Shejwalkar, R. Shokri, and A. Houmansadr, “Cronus: Robust and Heterogeneous Collaborative Learning with Black-Box Knowledge Transfer,” 2019, arXiv 1912.11279.
- [12] A. M. Abdelmoniem, A. N. Sahu, M. Canini, and S. A. Fahmy, “REFL: Resource-Efficient Federated Learning,” in *Proceedings of ACM EuroSys*, 2023, p. 215–232.
- [13] A. Arouj and A. M. Abdelmoniem, “Towards energy-aware federated learning via collaborative computing approach,” *Computer Communications*, vol. 221, pp. 131–141, 2024.
- [14] K. Roth, L. Thede, A. S. Koepke, O. Vinyals, O. Hénaff, and Z. Akata, “Fantastic Gains and Where to Find Them: On the Existence and Prospect of General Knowledge Transfer between Any Pretrained Model,” 2023, arXiv 2310.17653.
- [15] N. Alballa and M. Canini, “A First Look at the Impact of Distillation Hyper-Parameters in Federated Knowledge Distillation,” in *EuroMLSys*, 2023.
- [16] H. Daga, P. K. Nicholson, A. Gavrilovska, and D. Lugones, “Cartel: A system for collaborative transfer learning at the edge,” in *Proceedings of the ACM Symposium on Cloud Computing*, 2019.
- [17] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *AISTATS*, 2017.
- [18] H. Daga, Y. Chen, A. Agrawal, and A. Gavrilovska, “Clue: Systems support for knowledge transfer in collaborative learning with neural nets,” *IEEE Transactions on Cloud Computing*, 2023.
- [19] Y. Lu, Y. Shu, X. Tan, Y. Liu, M. Zhou, Q. Chen, and D. Pei, “Collaborative learning between cloud and end devices: an empirical study on location prediction,” in *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, 2019.
- [20] M. Ye, X. Fang, B. Du, P. C. Yuen, and D. Tao, “Heterogeneous federated learning: State-of-the-art and research challenges,” *ACM Computing Surveys*, 2023.
- [21] A. M. Abdelmoniem, C.-Y. Ho, P. Papageorgiou, and M. Canini, “A comprehensive empirical study of heterogeneity in federated learning,” *IEEE Internet of Things Journal*, 2023.
- [22] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, “Advances and open problems in federated learning,” *Foundations and trends in machine learning*, 2021.
- [23] A. F. Khan, A. A. Khan, A. M. Abdelmoniem, S. Fountain, A. R. Butt, and A. Anwar, “Float: Federated learning optimizations with automated tuning,” in *ACM EuroSys*, 2024.
- [24] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, “Deep Mutual Learning,” in *CVPR*, 2018.
- [25] R. Anil, G. Pereyra, A. Passos, R. Ormandi, G. E. Dahl, and G. E. Hinton, “Large Scale Distributed Neural Network Training Through Online Distillation,” 2018, arXiv 1804.03235.
- [26] Q. Li, Y. Diao, Q. Chen, and B. He, “Federated Learning on Non-IID Data Silos: An Experimental Study,” in *ICDE*, 2022.
- [27] L. N. Darlow, E. J. Crowley, A. Antoniou, and A. J. Storkey, “CINIC-10 is not ImageNet or CIFAR-10,” 2018.
- [28] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.
- [29] M. Phuong and C. Lampert, “Towards Understanding Knowledge Distillation,” in *ICML*, 2019.
- [30] N. Alballa, W. Zhang, Z. Liu, A. M. Abdelmoniem, M. Elhoseiny, and M. Canini, “Query-based knowledge transfer for heterogeneous learning environments,” *arXiv preprint arXiv:2504.09205*, 2025.