# HETEROGENEOUS SOCIALIZED LEARNING: ENABLING INDIVIDUAL GROWTH ACROSS DIVERSE MODEL ARCHITECTURES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Collaborative learning among heterogeneous agents offers a promising approach to overcoming the limited knowledge of individual agents. However, existing heterogeneous federated learning methods primarily focus on improving global or local task performance, rather than enabling each agent to acquire complementary knowledge beyond its original expertise. Although socialized learning emphasizes individual knowledge acquisition and preservation, its reliance on shared model architectures limits its applicability to heterogeneous agents. To address these challenges, we formulate Heterogeneous Socialized Learning, a new paradigm that enables agents with distinct model architectures to acquire complementary knowledge from their peers while preserving their original expertise. We further propose Heterogeneous Socialized Collaboration (HSC), an architecture-agnostic framework that facilitates cross-model knowledge exchange through compact knowledge packets constructed via dataset distillation. By integrating peer knowledge distillation with feature retention, HSC enables each agent to acquire new knowledge while retaining its existing expertise and model architecture. Extensive experiments on CIFAR-100 and Tiny ImageNet demonstrate that HSC consistently improves new knowledge acquisition and overall competence across diverse model architectures and agent populations while retaining substantial original expertise.

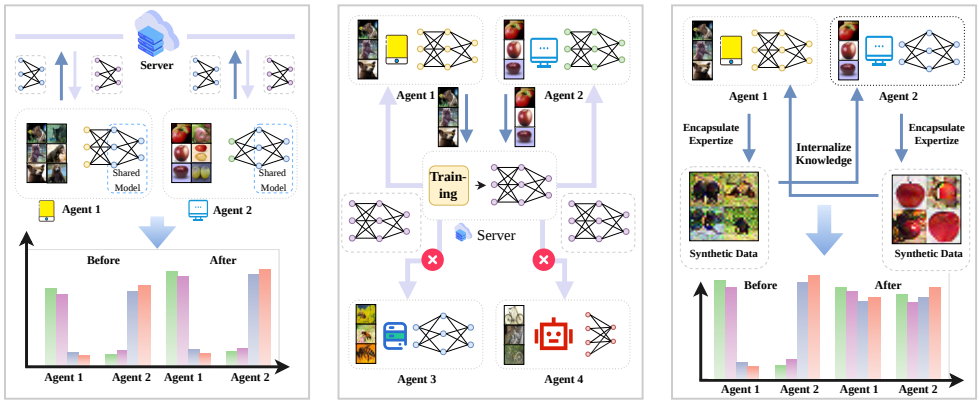
## 1 INTRODUCTION

With the proliferation of intelligent edge devices, collaborative learning among multiple heterogeneous agents has become an increasingly prevalent paradigm in real-world applications, ranging from autonomous driving and robotic systems to smart healthcare and the Internet of Things. In such environments, individual agents face diverse task requirements and local data distributions, leading to distinct knowledge and capabilities, while varying computational and storage constraints give rise to heterogeneous model architectures. Although each agent may perform well within its own domain of expertise, its limited knowledge and experience can restrict its ability to handle unfamiliar tasks or adapt to new environments. Collaborative learning offers a promising opportunity to overcome these limitations by enabling agents to acquire complementary knowledge and capabilities from one another. The central objective is therefore to enable each agent to acquire new knowledge from other agents and expand its capabilities while preserving its existing knowledge and competencies.

Heterogeneous federated learning provides an established framework for collaboration among agents with different model architectures. To bridge architectural differences, existing methods share compatible portions of local models, exchange knowledge through aligned representations or predictions, or introduce homogeneous auxiliary models to connect otherwise heterogeneous agents (Tan et al., 2022; Li et al., 2025b; Yi et al., 2025a; Huang et al., 2024; Yao et al., 2026b). However, these methods primarily focus on consolidating distributed knowledge into a global model or improving individual models on their local tasks, rather than enabling each agent to acquire new knowledge beyond its original expertise. Consequently, despite the knowledge exchange among agents, individual agents may still struggle to acquire knowledge from peers with substantially different task specializations, limiting their ability to perform tasks beyond their original domains.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.




(a) Heterogeneous Federated Learning (b) Canonical Socialized Learning (c) Heterogeneous Socialized Learning

Figure 1: Three collaboration paradigms. (a) Heterogeneous federated learning shares compatible knowledge to improve a global model or local task performance. (b) Canonical socialized learning grows individuals through a shared generalist, requiring a common model architecture. (c) HSL enables heterogeneous agents to acquire complementary expertise while retaining their own capabilities and architectures.

Socialized learning offers a promising paradigm for addressing this challenge by enabling individual agents to acquire complementary knowledge from their peers while preserving their original expertise (Thompson et al., 2022; Yao et al., 2024). In canonical socialized learning, individual experts contribute their knowledge to a shared generalist, which in turn transfers the integrated knowledge back to support individual growth. This collaboration relies on a common model structure, since knowledge must first be integrated within a shared representation space and then transferred back to structurally compatible individual models. As a result, agents are typically restricted to homogeneous or compatible architectures, limiting the applicability of canonical socialized learning to real-world multi-agent systems, where architectural heterogeneity is pervasive.

To address this challenge, we formulate Heterogeneous Socialized Learning (HSL), a new paradigm in which agents with diverse architectures and complementary expertise learn from one another while preserving their individual specialization. Starting from locally acquired expertise, each agent aims to extend its capabilities to categories mastered by its peers while retaining its original competence and model architecture. In this paper, we propose Heterogeneous Socialized Collaboration (HSC) to address these two challenges. For architecture-agnostic knowledge encapsulation, each agent distills its local expertise into a compact knowledge packet consisting of synthetic samples and expert soft labels. By expressing knowledge in the shared input and label spaces rather than model parameters or internal representations, these packets can be exchanged across heterogeneous architectures. For knowledge internalization with expertise retention, each receiver learns directly from peer knowledge packets on its own backbone: peer knowledge distillation transfers complementary knowledge from other agents, while feature retention anchors the receiver to its original representations and mitigates the degradation of existing expertise. Together, these mechanisms enable each heterogeneous agent to expand its competence while retaining its own model architecture and specialization. Extensive experiments on CIFAR-100 and Tiny ImageNet demonstrate that HSC substantially improves new knowledge acquisition and overall competence across heterogeneous architectures and agent populations while retaining considerable original expertise.

We summarize our contributions as follows.

- To our knowledge, we are the first to formulate the Heterogeneous Socialized Learning (HSL) paradigm, defining the task of enabling agents with distinct model architectures to acquire complementary knowledge from their peers while preserving their original expertise and model architecture.
- We propose Heterogeneous Socialized Collaboration (HSC), realizing HSL through architecture-agnostic knowledge encapsulation and expertise-preserving knowledge internalization. HSC packages synthetic samples and soft labels into transferable knowledge

packets, while peer distillation and feature retention enable receivers to internalize complementary knowledge on their own backbones while retaining original expertise.

- Extensive experiments on CIFAR-100 and Tiny ImageNet demonstrate that HSC enables effective new knowledge acquisition while retaining substantial original expertise across heterogeneous architectures. In particular, HSC improves Global accuracy over the strongest baseline by **5.58–10.45** and **10.17–16.29** percentage points, respectively, supporting its effectiveness in realizing the individual-growth objective of HSL.

# 2 Related Work

## 2.1 Heterogeneous federated learning

Model-heterogeneous federated learning addresses architectural incompatibility by establishing transferable interfaces among heterogeneous models (Ye et al., 2024; Zhang et al., 2025; Guo et al., 2025). Existing approaches either share structurally compatible portions of local models, exchange knowledge through predictions, logits, prototypes, or aligned representations (Li & Wang, 2019; Lin et al., 2020; Zhu et al., 2021; Weng et al., 2025; Tu et al., 2026; Chen et al., 2025; Yang et al., 2025; Yao et al., 2026a; Niu et al., 2026; Tan et al., 2022; Li et al., 2025b; Yi et al., 2024b;a; 2025b), or introduce shared modules, adapters, and auxiliary networks to bridge heterogeneous architectures (Diao et al., 2021; Yi et al., 2025a; 2026b;a; Huang et al., 2024; Yao et al., 2026b). These mechanisms mitigate parameter incompatibility and enable knowledge transfer without requiring all participants to use identical models.

Despite these advances, existing methods are typically optimized and evaluated under conventional federated-learning objectives. Server-oriented approaches aim to construct a stronger shared predictor, while personalized approaches primarily improve each client's performance on its local data distribution. Accordingly, cross-model knowledge serves mainly to enhance global aggregation or local-task performance, whereas explicitly enabling every participant to acquire complementary knowledge beyond its original expertise is rarely treated as the primary learning objective. Our setting preserves the need for cross-architecture knowledge transfer but shifts the unit of success to individual growth: each heterogeneous agent should internalize peer knowledge as its own competence while retaining its original expertise and model architecture.

## 2.2 Socialized learning

Socialized learning starts from a different premise: collaboration should make every participant better by allowing agents to learn from one another while retaining their own strengths (Thompson et al., 2022; Yao et al., 2024). In canonical protocols, agents first develop complementary expertise and contribute it to a shared generalist. The generalist accumulates broad competence and subsequently transfers the integrated knowledge back to support the growth of individual agents. This organization makes peer knowledge acquisition and the preservation of individual expertise central objectives of collaboration. However, such individual growth is mediated by the shared generalist. Knowledge from different experts must first be integrated within a common model or representation space and then transferred back to structurally compatible individual models. Canonical socialized learning therefore relies on homogeneous or compatible model architectures, limiting its applicability to real-world multi-agent systems where architectural heterogeneity is pervasive. Graph-based and cross-task extensions broaden the application domains of socialized learning (Li et al., 2025a; Yao et al., 2025), but do not directly address collaboration among independently structured models. HSL extends the individual-growth objective to the model-heterogeneous setting.

# 3 Method

## 3.1 Formulation and Overview

**Problem definition.** Consider a heterogeneous multi-agent system consisting of $N$ agents, denoted by $\mathcal{A} = \{a_i\}_{i=1}^{N}$. All agents operate over a shared input space $\mathcal{X}$ and a global label space $\mathcal{Y} = \{1, \ldots, C\}$, while each agent $a_i$ has access only to a private dataset $\mathcal{D}_i \subseteq \mathcal{X} \times \mathcal{C}_i$, where

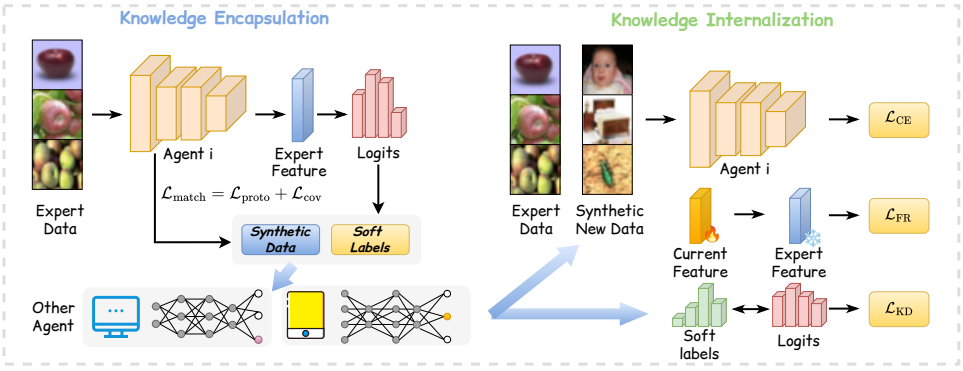



Figure 2: Overview of HSC. Each agent encapsulates its local expertise into a knowledge packet, and each receiver internalizes peer knowledge on its own model while retaining its original expertise.

$\mathcal{C}_i \subseteq \mathcal{Y}$ denotes the subset of classes on which it develops local expertise. We consider complementary, class-disjoint expertise across agents $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset \quad (i \neq j)$, $\bigcup_{i=1}^{N} \mathcal{C}_i = \mathcal{Y}$. Accordingly, $\mathcal{C}_i$ represents the knowledge initially available to agent $a_i$, whereas $\mathcal{Y} \setminus \mathcal{C}_i$ corresponds to complementary knowledge held by its peers.

In addition to data heterogeneity, agents are equipped with heterogeneous model architectures. Specifically, agent $a_i$ employs a model $f_i = h_i \circ \phi_i$, where $\phi_i$ is its encoder and $h_i$ is a $C$-way classifier. Different agents may vary in network architecture and feature dimensionality, making their parameters and intermediate representations directly incompatible. Let $\mathcal{H}_i$ denote the model family determined by the architecture of $a_i$, and let $f_i^{\text{E}} = h_i^{\text{E}} \circ \phi_i^{\text{E}} \in \mathcal{H}_i$ denote the local expert trained on $\mathcal{D}_i$. Throughout collaboration, agent $a_i$ is updated within $\mathcal{H}_i$, allowing its competence to evolve without changing its underlying model architecture.

**Learning objective.** HSL aims to enable individual growth for every agent. Starting from its locally trained expert $f_i^{\text{E}}$, each agent $a_i$ collaborates with its peers to obtain a post-collaboration model $f_i^{+}$. The goal is for $f_i^{+}$ to acquire predictive competence on the complementary classes $\mathcal{Y} \setminus \mathcal{C}_i$ held by other agents, while retaining its original expertise on $\mathcal{C}_i$. Importantly, edge agents are typically subject to limited computational and storage budgets, and their model architectures are chosen to match these resource constraints. HSL therefore requires each agent to preserve its own model architecture throughout collaboration, i.e., $f_i^{\text{E}}, f_i^{+} \in \mathcal{H}_i$, rather than replacing it with a larger or structurally different model. The outcome is a population of individually grown models $\{f_i^{+}\}_{i=1}^{N}$, each extending its competence from its local expertise $\mathcal{C}_i$ to the global label space $\mathcal{Y}$.

**HSC overview.** Heterogeneous Socialized Collaboration (HSC) realizes the HSL objective through two complementary mechanisms. *Architecture-agnostic knowledge encapsulation* converts each agent's local expertise into a transferable form that can be consumed by peers with different model architectures. *Expertise-preserving knowledge internalization* then enables each receiver to absorb peer knowledge within its own model while anchoring the representation that supports its original expertise. As illustrated in Figure 2, HSC therefore supports cross-architecture individual growth without replacing the receiver's original model architecture.

### 3.2 Architecture-Agnostic Knowledge Encapsulation

Agents in HSL may belong to different model families $\mathcal{H}_i$, making their parameters and intermediate representations directly incompatible. Knowledge encapsulation therefore aims to express the local expertise of each agent through an exchange interface that does not depend on a shared parameter or feature space. To this end, HSC encapsulates the expertise of agent $a_i$ into a compact knowledge packet defined over the shared input space $\mathcal{X}$ and semantic label space $\mathcal{Y}$. Each packet consists of synthetic samples, their global class labels, and the corresponding expert logits. The synthetic samples provide compact training evidence for the expert classes, while the class-indexed logits convey the expert's fine-grained judgments on these samples. Since a receiver can process

the samples using its own encoder and interpret the supervision through the shared label semantics, the resulting exchange interface can be consumed across heterogeneous model architectures without sharing model parameters or internal features.

**Local expertise condensation.** Agent $a_i$ first trains its local expert $f_i^{\mathrm{E}}$ on $\mathcal{D}_i$. It then constructs $K$ synthetic samples for each class $c \in \mathcal{C}_i$ from its private data. Dataset condensation aims to replace a large training set with a compact synthetic set that preserves useful training signals (Wang et al., 2018; Zhao et al., 2021; Cazenavette et al., 2022). Following DSDM (Li et al., 2024), we match the feature statistics of real and synthetic samples using the local expert encoder $\phi_i^{\mathrm{E}}$. The expert model $f_i^{\mathrm{E}}$ subsequently provides annotations for the synthetic samples.

Let $(\boldsymbol{\mu}_{i,c}^r, \boldsymbol{\Sigma}_{i,c}^r)$ and $(\boldsymbol{\mu}_{i,c}^s, \boldsymbol{\Sigma}_{i,c}^s)$ denote the class prototype and covariance of real and synthetic features under $\phi_i^{\mathrm{E}}$, respectively. For clarity, we present the core distribution-matching terms:

$$\mathcal{L}_{\mathrm{match},i} = \sum_{c \in \mathcal{C}_i} \left[ \left\| \boldsymbol{\mu}_{i,c}^r - \boldsymbol{\mu}_{i,c}^s \right\|_2^2 + \frac{\lambda_{\mathrm{cov}}}{d_i^{\mathrm{E}}} \left\| \boldsymbol{\Sigma}_{i,c}^r - \boldsymbol{\Sigma}_{i,c}^s \right\|_F^2 \right]. \tag{1}$$

Here $d_i^{\mathrm{E}}$ denotes the feature dimension of the local expert encoder. Prototype matching aligns the class centers of real and synthetic data, while covariance matching preserves intra-class variation. The complete condensation procedure follows DSDM (Li et al., 2024).

The resulting labeled synthetic set is

$$\mathcal{S}_i = \{(s_{i,m}, y_{i,m})\}_{m=1}^{K|\mathcal{C}_i|}, \tag{2}$$

where $y_{i,m} \in \mathcal{C}_i$ is represented using the shared global label indexing.

**Expert annotation and packet construction.** Synthetic samples encode compact class-level training evidence, but hard labels alone cannot capture the fine-grained judgments of the local expert. HSC therefore augments each synthetic sample with the response of $f_i^{\mathrm{E}}$. Since $f_i^{\mathrm{E}}$ is trained only on $\mathcal{C}_i$, predictions outside this subset do not represent reliable local expertise. We thus retain only its logits over the expert classes:

$$\boldsymbol{z}_{i,m} = f_i^{\mathrm{E}}(s_{i,m})\big|_{\mathcal{C}_i} \in \mathbb{R}^{|\mathcal{C}_i|}. \tag{3}$$

The knowledge packet of agent $a_i$ is then defined as

$$\mathcal{P}_i = \{(s_{i,m}, y_{i,m}, \boldsymbol{z}_{i,m})\}_{m=1}^{K|\mathcal{C}_i|}. \tag{4}$$

Here, the hard label $y_{i,m}$ specifies the class identity in the shared semantic space, while $\boldsymbol{z}_{i,m}$ records the expert's relative judgment among classes in $\mathcal{C}_i$. Each logit vector remains paired with the synthetic sample that elicited it and is kept fixed after packet construction.

For receiver $a_j$, we distinguish its own packet $\mathcal{P}_j$ from the peer knowledge available for acquisition:

$$\mathcal{P}_{-j} = \bigcup_{i \neq j} \mathcal{P}_i. \tag{5}$$

**Cross-architecture exchange.** The key property of $\mathcal{P}_i$ is that the exchanged knowledge is represented through the shared input and semantic label spaces rather than the sender's model space $\mathcal{H}_i$. Synthetic samples can be encoded by receivers using their own backbones, while their hard labels and expert logits provide class-indexed supervision without requiring matched feature dimensions or parameter structures. Knowledge encapsulation therefore transforms architecture-specific local expertise into a transferable interface that heterogeneous receivers can directly use.

### 3.3 Expertise-Preserving Knowledge Internalization

Knowledge internalization aims to transform the peer packets $\mathcal{P}_{-j}$ into predictive competence within the receiver's own model family $\mathcal{H}_j$, while mitigating the degradation of its original expertise. HSC realizes this process through three complementary signals. Full-class hard-label learning establishes discrimination over both local and peer classes, peer knowledge distillation transfers the fine-grained judgments of source experts, and feature retention anchors the representation supporting the receiver's original expertise.

**Receiver-space initialization.** Agent $a_j$ initializes its trainable model from the local expert $f_j^{\mathrm{E}}$. Its encoder $\phi_j$ remains trainable throughout internalization, while a frozen copy of the pre-collaboration expert encoder, denoted by $\bar{\phi}_j$, is retained only as a reference for feature retention. The classifier rows corresponding to the original classes $\mathcal{C}_j$ inherit their expert values, whereas rows for peer classes are initialized from peer packet samples in the receiver's own feature space.

Specifically, for each peer class $c \in \mathcal{Y} \setminus \mathcal{C}_j$, let $\mathcal{S}_c$ denote the synthetic samples of class $c$ contained in $\mathcal{P}_{-j}$. Their receiver-side prototype is

$$\boldsymbol{r}_{j,c} = \frac{1}{|\mathcal{S}_c|} \sum_{s \in \mathcal{S}_c} \phi_j(s), \qquad c \in \mathcal{Y} \setminus \mathcal{C}_j, \tag{6}$$

which is used to initialize the corresponding classifier row. Although the packet is constructed by a heterogeneous source model, this initialization maps the peer class directly into the representation space of receiver $a_j$. Subsequent optimization jointly updates the trainable encoder $\phi_j$ and the full $C$-way classifier, so all newly acquired competence remains within $\mathcal{H}_j$.

**Dual-stream learning.** Knowledge internalization requires both local evidence for maintaining the receiver's original classes and peer evidence for acquiring complementary classes. We therefore construct two training streams. Let $Q_j^{\mathrm{real}}$ denote the class-balanced distribution over the private data $\mathcal{D}_j$, and let $Q_j^{\mathrm{self}}$ denote the class-balanced distribution over the receiver's own packet $\mathcal{P}_j$. The local stream is defined as

$$Q_j^{\mathrm{loc}} = \rho Q_j^{\mathrm{real}} + (1-\rho) Q_j^{\mathrm{self}}, \tag{7}$$

while $Q_j^{\mathrm{peer}}$ is the class-balanced distribution over the peer packets $\mathcal{P}_{-j}$.

Let $\alpha_j = |\mathcal{C}_j|/C$. Both streams are trained with the $C$-way hard-label cross-entropy:

$$\mathcal{L}_{\mathrm{CE},j} = \alpha_j \mathbb{E}_{(x,y)\sim Q_j^{\mathrm{loc}}} \left[\ell_{\mathrm{CE}}(f_j(x), y)\right] + (1-\alpha_j) \mathbb{E}_{(i,s,y,\boldsymbol{z})\sim Q_j^{\mathrm{peer}}} \left[\ell_{\mathrm{CE}}(f_j(s), y)\right]. \tag{8}$$

Because both streams are class-balanced, $\alpha_j/|\mathcal{C}_j| = (1-\alpha_j)/(C-|\mathcal{C}_j|) = 1/C$, giving each class the same expected contribution to the hard-label objective. The local stream reinforces discrimination on the receiver's original classes, whereas the peer stream introduces supervision for classes absent from its private data.

**Peer knowledge distillation.** Hard labels specify class identity in the shared semantic space, but they do not capture the fine-grained judgments encoded by the source experts. HSC therefore further distills the expert logits stored in each peer packet. For a tuple $(i, s, y, \boldsymbol{z})$ sampled from $Q_j^{\mathrm{peer}}$, define

$$\boldsymbol{p}_i = \sigma(\boldsymbol{z}/\tau), \qquad \boldsymbol{p}_{j\leftarrow i} = \sigma\Big(f_j(s)|_{\mathcal{C}_i}/\tau\Big), \tag{9}$$

where $\sigma$ denotes the softmax function and $\tau$ is the temperature.

Since source agent $a_i$ provides reliable expertise only on $\mathcal{C}_i$, the receiver matches its predictions to the source expert only over these classes:

$$\mathcal{L}_{\mathrm{KD},j} = \tau^2 \mathbb{E}_{(i,s,y,\boldsymbol{z})\sim Q_j^{\mathrm{peer}}} \left[D_{\mathrm{KL}}(\boldsymbol{p}_i \,\|\, \boldsymbol{p}_{j\leftarrow i})\right]. \tag{10}$$

The source identity, synthetic sample, and expert logits remain aligned throughout distillation, allowing each peer to teach only the expertise it originally possesses. In this way, peer knowledge is internalized directly by the receiver's own model rather than through a shared teacher architecture.

**Feature retention.** Acquiring peer knowledge updates the same encoder that supports the receiver's original classes, which may cause representation drift on its local expertise. To mitigate this effect, HSC retains a frozen copy $\bar{\phi}_j$ of the pre-collaboration expert encoder and uses it as a representation anchor. Given a private sample $(x, y) \sim \mathcal{D}_j$ and the same sampled augmentation $a \sim \mathcal{T}$ for both branches, we define

$$\mathcal{L}_{\mathrm{FR},j} = \mathbb{E}_{\substack{(x,y)\sim\mathcal{D}_j \\ a\sim\mathcal{T}}} \left[1 - \frac{\langle \phi_j(a(x)), \bar{\phi}_j(a(x))\rangle}{\|\phi_j(a(x))\|_2 \left\|\bar{\phi}_j(a(x))\right\|_2}\right]. \tag{11}$$

Only $\bar{\phi}_j$ is frozen; the current encoder $\phi_j$ remains fully trainable. This constraint anchors the representation of local samples to the pre-collaboration expert while leaving the receiver free to adapt to complementary peer classes.

The three learning signals are optimized jointly:

$$\mathcal{L}_j = \mathcal{L}_{\mathrm{CE},j} + \lambda_{\mathrm{KD}}\mathcal{L}_{\mathrm{KD},j} + \lambda_{\mathrm{FR}}\mathcal{L}_{\mathrm{FR},j}. \tag{12}$$

Through this process, agent $a_j$ evolves from $f_j^{\mathrm{E}} \in \mathcal{H}_j$ to $f_j^{+} \in \mathcal{H}_j$, extending its competence from $\mathcal{C}_j$ to $\mathcal{Y}$ while retaining its original model architecture and anchoring the representation that supports its local expertise.

# 4 Experiments

## 4.1 Experimental Setup

**Datasets and heterogeneous multi-agent settings.** We evaluate HSC on CIFAR-100 (Krizhevsky, 2009) and Tiny ImageNet (Le & Yang, 2015) in heterogeneous multi-agent systems with $N \in \{5, 10, 20\}$ agents. To instantiate model heterogeneity, agents are equipped with five backbone architectures: ConvNet-3, ConvNet-4, AlexNet (Krizhevsky et al., 2012), ResNet-10, and ResNet-18 (He et al., 2016). For $N = 5$, each architecture is assigned to one agent; for $N = 10$ and $N = 20$, each architecture appears twice and four times, respectively, maintaining a balanced architecture distribution as the number of agents increases.

Following the class-disjoint individual-growth setting of socialized learning (Yao et al., 2024), we evenly partition the label space across agents. Each agent $a_i$ is assigned a disjoint class subset $\mathcal{C}_i$ and has access only to training samples from these classes, so each agent starts with distinct local expertise while the agents collectively possess complementary knowledge. On CIFAR-100, each agent owns 20, 10, and 5 classes for $N = 5$, 10, and 20, respectively. On Tiny ImageNet, each agent owns 40, 20, and 10 classes. For knowledge encapsulation, each agent constructs $K = 10$ synthetic samples per class on CIFAR-100 and $K = 20$ on Tiny ImageNet.

**Compared methods and evaluation metrics.** We compare HSC with three categories of baselines. First, MASC (Yao et al., 2024) represents canonical socialized learning, where individual growth is achieved through a shared-architecture collaboration protocol. We report only its pre-collaboration Expert accuracy, because its shared-architecture collaboration cannot be carried out in our heterogeneous setting. Second, DeSA (Huang et al., 2024) and FedRE (Yao et al., 2026b) are representative model-heterogeneous federated learning methods that enable collaboration across different model architectures. Third, we include two data-based knowledge transfer baselines: Heuristic directly samples real images from each agent's local data, while FAST (Cui et al., 2026) constructs a frequency-domain coreset. Heuristic and FAST use the same transferred image budget and receiver training schedule as HSC for a controlled comparison.

We evaluate individual growth from three complementary perspectives. For each receiver, Expert measures accuracy on its original expert classes $\mathcal{C}_i$, reflecting the retention of local expertise; New measures accuracy on the complementary classes $\mathcal{Y} \setminus \mathcal{C}_i$, directly quantifying the acquisition of peer knowledge; and Global measures accuracy over the complete label space $\mathcal{Y}$, summarizing the receiver's overall competence after collaboration. For each seed, results are first averaged over all receivers; we then report the mean and standard deviation across three random seeds.

## 4.2 Main Results

Table 1: Results (%) on CIFAR-100, where the best and sub-optimal results are bold and underlined.

| | 5 agents | | | 10 agents | | | 20 agents | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | Expert | New | Global | Expert | New | Global | Expert | New | Global |
| MASC | 74.89 | – | – | 80.48 | – | – | 85.49 | – | – |
| DeSA | **68.51**$_{\pm 3.65}$ | 1.17$_{\pm 0.12}$ | 14.64$_{\pm 0.83}$ | **76.18**$_{\pm 0.78}$ | 2.31$_{\pm 0.13}$ | 9.70$_{\pm 0.07}$ | **84.64**$_{\pm 1.19}$ | 2.27$_{\pm 0.03}$ | 6.39$_{\pm 0.07}$ |
| FedRE | 58.41$_{\pm 0.12}$ | 0.00$_{\pm 0.00}$ | 11.69$_{\pm 0.02}$ | <u>63.06</u>$_{\pm 0.28}$ | 0.00$_{\pm 0.00}$ | 6.31$_{\pm 0.03}$ | <u>72.14</u>$_{\pm 0.94}$ | 0.00$_{\pm 0.00}$ | 3.61$_{\pm 0.05}$ |
| Heuristic | <u>64.40</u>$_{\pm 0.09}$ | <u>25.91</u>$_{\pm 0.75}$ | <u>33.60</u>$_{\pm 0.59}$ | 51.76$_{\pm 0.89}$ | <u>23.98</u>$_{\pm 0.51}$ | <u>26.76</u>$_{\pm 0.45}$ | 64.87$_{\pm 0.21}$ | <u>16.82</u>$_{\pm 0.66}$ | <u>19.22</u>$_{\pm 0.63}$ |
| FAST | 64.19$_{\pm 0.36}$ | 23.54$_{\pm 0.44}$ | 31.67$_{\pm 0.29}$ | 51.29$_{\pm 0.22}$ | 21.14$_{\pm 0.15}$ | 24.15$_{\pm 0.12}$ | 63.65$_{\pm 0.84}$ | 14.80$_{\pm 0.07}$ | 17.25$_{\pm 0.06}$ |
| **HSC** | 61.05$_{\pm 0.91}$ | **33.72**$_{\pm 0.53}$ | **39.18**$_{\pm 0.52}$ | 57.75$_{\pm 0.18}$ | **31.39**$_{\pm 0.20}$ | **34.02**$_{\pm 0.17}$ | 57.14$_{\pm 0.08}$ | **28.23**$_{\pm 0.22}$ | **29.67**$_{\pm 0.21}$ |

Table 2: Results (%) on Tiny ImageNet.

| Method | 5 agents | | | 10 agents | | | 20 agents | | |
|---|---|---|---|---|---|---|---|---|---|
| | Expert | New | Global | Expert | New | Global | Expert | New | Global |
| MASC | 65.13 | – | – | 72.88 | – | – | 78.99 | – | – |
| DeSA | 46.27$_{\pm 1.17}$ | 0.23$_{\pm 0.07}$ | 9.44$_{\pm 0.19}$ | **62.68**$_{\pm 1.16}$ | 0.40$_{\pm 0.03}$ | 6.63$_{\pm 0.09}$ | **69.65**$_{\pm 0.09}$ | 0.60$_{\pm 0.09}$ | 4.05$_{\pm 0.09}$ |
| FedRE | 6.80$_{\pm 0.68}$ | 0.00$_{\pm 0.00}$ | 1.36$_{\pm 0.14}$ | 15.49$_{\pm 0.68}$ | 0.00$_{\pm 0.00}$ | 1.55$_{\pm 0.07}$ | 27.53$_{\pm 0.38}$ | 0.00$_{\pm 0.00}$ | 1.38$_{\pm 0.02}$ |
| Heuristic | 53.92$_{\pm 2.40}$ | 7.37$_{\pm 0.45}$ | 16.68$_{\pm 0.80}$ | 57.74$_{\pm 1.35}$ | 7.01$_{\pm 0.10}$ | 12.08$_{\pm 0.08}$ | 58.87$_{\pm 0.61}$ | 4.35$_{\pm 0.10}$ | 7.08$_{\pm 0.06}$ |
| FAST | 51.38$_{\pm 1.64}$ | 6.55$_{\pm 0.62}$ | 15.52$_{\pm 0.24}$ | 55.84$_{\pm 0.93}$ | 6.84$_{\pm 0.14}$ | 11.74$_{\pm 0.19}$ | 58.04$_{\pm 1.11}$ | 4.19$_{\pm 0.26}$ | 6.88$_{\pm 0.19}$ |
| **HSC** | **59.92**$_{\pm 0.76}$ | **18.58**$_{\pm 1.17}$ | **26.85**$_{\pm 0.83}$ | 58.93$_{\pm 0.54}$ | **22.47**$_{\pm 1.33}$ | **26.12**$_{\pm 1.16}$ | 56.85$_{\pm 0.17}$ | **21.61**$_{\pm 0.62}$ | **23.37**$_{\pm 0.58}$ |

Tables 1 and 2 report the main results on CIFAR-100 and Tiny ImageNet, while Figure 3 further visualizes how New and Global accuracy change as the number of agents increases. From these results, we can draw the following observations.

**HSC enables effective acquisition of peer knowledge.** HSC achieves the highest New accuracy in all six dataset-agent settings, demonstrating a clear advantage in learning classes that are absent from each receiver's private data. On CIFAR-100, HSC obtains 33.72%, 31.39%, and 28.23% New accuracy with 5, 10, and 20 agents, respectively. The advantage is even more pronounced on Tiny ImageNet, where HSC reaches 18.58%, 22.47%, and 21.61%, while Heuristic achieves only 7.37%, 7.01%, and 4.35%. Since New is evaluated on complementary peer classes, these results show that receivers effectively acquire knowledge from their peers to recognize previously unseen classes.

**Peer knowledge acquisition leads to stronger overall performance.** HSC also achieves the highest Global accuracy in every setting. On CIFAR-100, HSC reaches 39.18%, 34.02%, and 29.67% Global accuracy for 5, 10, and 20 agents, compared with 33.60%, 26.76%, and 19.22% for Heuristic. Similar improvements are observed on Tiny ImageNet, where HSC improves Global accuracy from 16.68% to 26.85%, from 12.08% to 26.12%, and from 7.08% to 23.37% over Heuristic. Although HSC does not always achieve the highest Expert accuracy, it attains a better acquisition-retention balance by substantially improving peer-class performance while retaining strong original expertise.

**HSC remains effective as expertise becomes increasingly fragmented.** Increasing the number of agents divides the same global label space into progressively smaller local expertise sets, making individual growth more challenging because each receiver starts with fewer classes and must acquire a larger portion of the global class space from its peers. Figure 3 shows that HSC maintains a clear New and Global advantage. Its Global margin over Heuristic grows from 5.58 to 7.26 and 10.45 percentage points on CIFAR-100 and from 10.17 to 14.04 and 16.29 points on Tiny ImageNet as the number of agents increases. This suggests that HSC is particularly effective when local expertise is sparse and complementary knowledge is distributed across more heterogeneous agents.

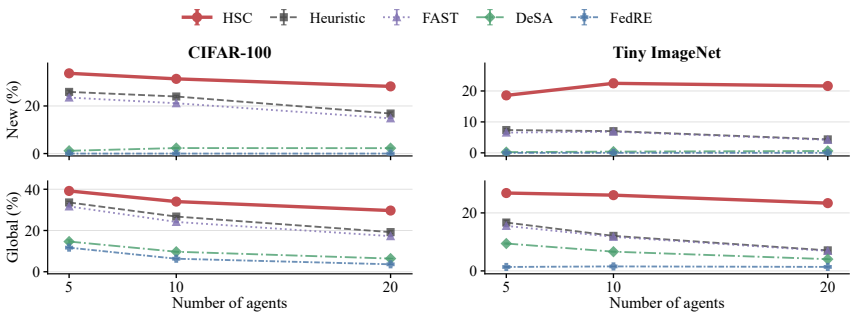



Figure 3: New and Global accuracies across different numbers of agents.

To further examine whether the overall improvements depend on particular model architectures, we break down the five-agent CIFAR-100 results by receiver backbone in Table 3. In this setting, the five

agents are equipped with ConvNet-3, ConvNet-4, AlexNet, ResNet-10, and ResNet-18, respectively, and the results for each backbone are aggregated over three random seeds. We can see that HSC achieves the highest New and Global accuracy on all five backbones, consistently outperforming Heuristic and FAST across all architectures. In terms of Global accuracy, the improvement over the sub-optimal method ranges from 0.42 percentage points on ResNet-10 to 8.52 points on ConvNet-4.

Table 3: Receiver-backbone-specific results(%) on CIFAR-100 under the five-agent setting.

| | Heuristic | | | FAST | | | HSC | | |
|---|---|---|---|---|---|---|---|---|---|
| Backbone | Expert | New | Global | Expert | New | Global | Expert | New | Global |
| ConvNet-3 | **56.03**$_{\pm 1.06}$ | 24.53$_{\pm 1.13}$ | 30.83$_{\pm 0.70}$ | 54.35$_{\pm 2.11}$ | 21.69$_{\pm 0.01}$ | 28.22$_{\pm 0.42}$ | 51.60$_{\pm 1.75}$ | **35.49**$_{\pm 0.80}$ | **38.71**$_{\pm 0.82}$ |
| ConvNet-4 | **57.95**$_{\pm 1.66}$ | 27.12$_{\pm 0.56}$ | 33.29$_{\pm 0.52}$ | 57.72$_{\pm 1.20}$ | 24.13$_{\pm 0.60}$ | 30.85$_{\pm 0.26}$ | 57.42$_{\pm 2.91}$ | **37.90**$_{\pm 0.69}$ | **41.81**$_{\pm 0.05}$ |
| AlexNet | 66.52$_{\pm 2.68}$ | 22.40$_{\pm 0.60}$ | 31.23$_{\pm 0.16}$ | **66.57**$_{\pm 1.77}$ | 21.10$_{\pm 0.70}$ | 30.20$_{\pm 0.22}$ | 59.73$_{\pm 0.74}$ | **32.53**$_{\pm 0.32}$ | **37.97**$_{\pm 0.18}$ |
| ResNet-10 | **71.42**$_{\pm 1.64}$ | 29.25$_{\pm 1.11}$ | 37.69$_{\pm 1.20}$ | 70.92$_{\pm 1.95}$ | 26.74$_{\pm 1.12}$ | 35.57$_{\pm 1.27}$ | 69.82$_{\pm 4.26}$ | **30.18**$_{\pm 1.81}$ | **38.11**$_{\pm 1.95}$ |
| ResNet-18 | 70.08$_{\pm 0.68}$ | 26.22$_{\pm 0.91}$ | 35.00$_{\pm 0.73}$ | **71.38**$_{\pm 0.39}$ | 24.04$_{\pm 0.69}$ | 33.51$_{\pm 0.52}$ | 66.67$_{\pm 1.62}$ | **32.48**$_{\pm 1.85}$ | **39.32**$_{\pm 1.77}$ |

### 4.3 Ablation Study

We conduct ablation studies on CIFAR-100 under the five-agent setting to examine the roles of peer knowledge distillation (KD) and feature retention (FR). Starting from the full HSC model, we remove KD or FR individually while keeping all other training settings unchanged. Figure 4 reports the resulting changes in Expert, New, and Global accuracy relative to full HSC. Removing KD increases Expert accuracy by 1.95 percentage points, but decreases New and Global accuracy by 1.33 and 0.67 points, respectively. This indicates that KD primarily facilitates the acquisition of peer knowledge: without the soft supervision provided by peer experts, the receiver becomes more biased toward its original classes and learns complementary classes less effectively. In contrast, removing FR decreases Expert accuracy by 1.63 points while slightly increasing New accuracy by 0.37 points, with Global accuracy remaining nearly unchanged ($-0.03$ points). This shows that FR mainly serves to retain the receiver's original expertise by constraining excessive representation drift during peer-knowledge acquisition. Overall, KD and FR play complementary roles: KD strengthens the acquisition of new knowledge, whereas FR protects existing expertise, jointly supporting the acquisition-retention objective of HSC.

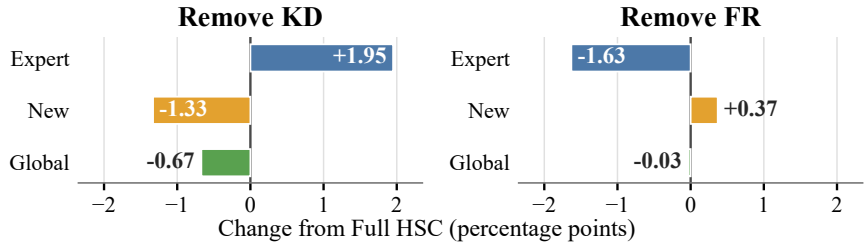



Figure 4: Changes from Full HSC after removing KD or FR (percentage points).

## 5 Conclusion

This paper is the first to formulate Heterogeneous Socialized Learning (HSL), bringing individual knowledge acquisition and expertise preservation to collaboration among architecturally diverse agents. Our Heterogeneous Socialized Collaboration (HSC) framework addresses two central challenges: making local expertise transferable across architectures through knowledge encapsulation, and integrating peer knowledge without losing existing expertise through knowledge internalization. Extensive experiments on CIFAR-100 and Tiny ImageNet demonstrate that HSC achieves stronger new-class acquisition and overall competence while retaining substantial original expertise. These findings establish an effective path for heterogeneous agents to grow through collaboration within their own model architectures.

## AI USE STATEMENT

AI tools assisted with drafting selected sections, refining language and presentation, and identifying potentially relevant literature. All AI-assisted content and suggested references were reviewed by the authors before inclusion. The authors determined the methodology, experimental design, and scientific claims and take full responsibility for the final manuscript and all reported results.

## REFERENCES

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A. Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4750–4759, 2022.

Haokun Chen, Hang Li, Yao Zhang, Jinhe Bi, Gengyuan Zhang, Yueqi Zhang, Philip Torr, Jindong Gu, Denis Krompass, and Volker Tresp. FedBiP: Heterogeneous one-shot federated learning with personalized latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30440–30450, 2025.

Jin Cui, Boran Zhao, Jiajun Xu, Jiaqi Guo, Shuo Guan, and Pengju Ren. FAST: Topology-aware frequency-domain distribution matching for coreset selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Enmao Diao, Jie Ding, and Vahid Tarokh. HeteroFL: Computation and communication efficient federated learning for heterogeneous clients. In *International Conference on Learning Representations*, 2021.

Yuanchun Guo, Bingyan Liu, Yulong Sha, and Zhensheng Xian. PracMHBench: Re-evaluating model-heterogeneous federated learning based on practical edge device constraints, 2025. URL https://arxiv.org/abs/2509.08750.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Chun-Yin Huang, Kartik Srinivas, Xin Zhang, and Xiaoxiao Li. Overcoming data and model heterogeneities in decentralized federated learning via synthetic anchors. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20111–20133. PMLR, 2024. URL https://proceedings.mlr.press/v235/huang24v.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge. CS231N Course Project, Stanford University, 2015.

Daliang Li and Junpu Wang. FedMD: Heterogenous federated learning via model distillation, 2019. URL https://arxiv.org/abs/1910.03581.

Hongcheng Li, Yucan Zhou, Xiaoyan Gu, Bo Li, and Weiping Wang. Diversified semantic distribution matching for dataset distillation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7542–7550, 2024. doi: 10.1145/3664647.3680900.

Jialu Li, Yu Wang, Pengfei Zhu, Wanyu Lin, Xinjie Yao, and Qinghua Hu. Graphs help graphs: Multi-agent graph socialized learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/558d48c1f08675daa636e09bfe94a89e-Abstract-Conference.html.

Yichen Li, Xiuying Wang, Wenchao Xu, Haozhao Wang, Yining Qi, Jiahua Dong, and Ruixuan Li. Feature distillation is the better choice for model-heterogeneous federated learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/96dd63ab8fd79163a5399e0b385f208d-Abstract-Conference.html.

Tao Lin, Lingjing Kong, Sebastian U. Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2351–2363, 2020.

Ziru Niu, Hai Dong, and A. K. Qin. Bridging generalization gap of heterogeneous federated clients using generative models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=xiTrskqWph.

Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. FedProto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8432–8440, 2022. doi: 10.1609/aaai.v36i8.20819.

Bill Thompson, Bas van Opheusden, Theodore Sumers, and Thomas L. Griffiths. Complex cognitive algorithms preserved by selective social learning in experimental populations. *Science*, 376 (6588):95–98, 2022. doi: 10.1126/science.abn0915.

Jingke Tu, Lei Yang, Chao Ma, and Weigang Wu. Personalized data-free knowledge distillation for federated learning under heterogeneous models and data. *ACM Transactions on Knowledge Discovery from Data*, 2026. doi: 10.1145/3831243.

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. Dataset distillation, 2018. URL https://arxiv.org/abs/1811.10959.

Ziqiao Weng, Weidong Cai, and Bo Zhou. FedSKD: Aggregation-free model-heterogeneous federated learning via multi-dimensional similarity knowledge distillation for medical image classification, 2025. URL https://arxiv.org/abs/2503.18981.

Mingzhao Yang, Shangchao Su, Bin Li, and Xiangyang Xue. One-shot heterogeneous federated learning with local model-guided diffusion models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 71157–71176. PMLR, 2025. URL https://proceedings.mlr.press/v267/yang25ad.html.

Dezhong Yao, Tongtong Liu, Yuexin Shi, and Zhiqiang Xu. FedMHO: Heterogeneous one-shot federated learning towards resource-constrained clients. In *Proceedings of the ACM Web Conference 2026*, pp. 5686–5697, 2026a. doi: 10.1145/3774904.3792743.

Xinjie Yao, Yu Wang, Pengfei Zhu, Wanyu Lin, Jialu Li, Weihao Li, and Qinghua Hu. Socialized learning: Making each other better through multi-agent collaboration. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56927–56945. PMLR, 2024. URL https://proceedings.mlr.press/v235/yao24d.html.

Xinjie Yao, Yu Wang, Pengfei Zhu, Wanyu Lin, Ruipu Zhao, Zhoupeng Guo, Weihao Li, and Qinghua Hu. Socialized coevolution: Advancing a better world through cross-task collaboration. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 71780–71797. PMLR, 2025. URL https://proceedings.mlr.press/v267/yao25e.html.

Yuan Yao, Lixu Wang, Jiaqi Wu, Jin Song, Simin Chen, Zehua Wang, Zijian Tian, Wei Chen, Huixia Li, and Xiaoxiao Li. FedRE: A representation entanglement framework for model-heterogeneous federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b.

Mang Ye, Xiuwen Fang, Bo Du, Pong C. Yuen, and Dacheng Tao. Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Computing Surveys*, 56(3):79:1–79:44, 2024. doi: 10.1145/3625558.

Liping Yi, Han Yu, Chao Ren, Gang Wang, Xiaoguang Liu, and Xiaoxiao Li. Federated model heterogeneous matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7a9ee756fb71d401bcb5230db02c11db-Abstract-Conference.html.

Liping Yi, Han Yu, Zhuan Shi, Gang Wang, Xiaoguang Liu, Lizhen Cui, and Xiaoxiao Li. FedSSA: Semantic similarity-based aggregation for efficient model-heterogeneous personalized federated learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 5371–5379, 2024b. doi: 10.24963/ijcai.2024/594.

Liping Yi, Han Yu, Chao Ren, Gang Wang, Xiaoguang Liu, and Xiaoxiao Li. pFedES: Generalized proxy feature extractor sharing for model heterogeneous personalized federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22146–22154, 2025a. doi: 10.1609/aaai.v39i21.34368.

Liping Yi, Han Yu, Gang Wang, Xiaoguang Liu, and Xiaoxiao Li. Federated representation angle learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1314–1324, 2025b. URL https://openaccess.thecvf.com/content/ICCV2025/html/Yi_Federated_Representation_Angle_Learning_ICCV_2025_paper.html.

Liping Yi, Han Yu, Gang Wang, Xiaoguang Liu, and Qinghua Hu. pFedMoE: Data-level personalization with mixture of experts in model-heterogeneous personalized federated learning. *IEEE Transactions on Knowledge and Data Engineering*, 2026a. doi: 10.1109/TKDE.2026.3656194.

Liping Yi, Heng Zhang, Han Yu, Gang Wang, Xiaoguang Liu, and Qinghua Hu. pFedLoRA: Model-heterogeneous personalized federated learning with homogeneous low-rank adapter sharing on mobile edge devices. *IEEE Transactions on Mobile Computing*, 2026b. doi: 10.1109/TMC.2026.3674996.

Jianqing Zhang, Xinghao Wu, Yanbing Zhou, Xiaoting Sun, Qiqi Cai, Yang Liu, Yang Hua, Zhenzhe Zheng, Jian Cao, and Qiang Yang. HtFLlib: A comprehensive heterogeneous federated learning library and benchmark. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5900–5911, 2025. doi: 10.1145/3711896.3737379.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *International Conference on Learning Representations*, 2021.

Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12878–12889. PMLR, 2021. URL https://proceedings.mlr.press/v139/zhu21b.html.