FEDSPEC: SHARED SPECTRAL STRUCTURES FOR HETEROGENEOUS FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Model heterogeneity poses a fundamental challenge to federated learning, as different model architectures produce incompatible parameter and representation spaces. Existing heterogeneous federated learning methods mainly rely on parameter sharing or feature alignment, often requiring compatible model components or explicit alignment mechanisms. To address this issue, we propose FedSpec, a heterogeneous Federated learning framework based on shared Spectral structures. FedSpec transforms heterogeneous local feature spaces into fixed-length normalized spectral structures through singular value decomposition, aggregates them into a global spectral structure, and uses the global spectrum as a structural prior for local spectral correction. The singular values characterize the relative importance distribution of principal representation directions, enabling a unified structural representation across heterogeneous models without requiring parameter compatibility or explicit feature alignment. We further analyze the convergence behavior and information reduction properties of FedSpec. Extensive experiments across diverse architectures and data modalities demonstrate that FedSpec consistently outperforms state-of-the-art methods, achieving up to 7.67% accuracy improvement with substantially lower communication cost.

1 INTRODUCTION

Federated learning (FL) McMahan et al. (2017) has emerged as a promising distributed learning paradigm that enables multiple clients to collaboratively train machine learning models without sharing raw data. By keeping local data on-device and only exchanging model-related information, FL effectively alleviates privacy concerns while leveraging distributed data resources Cheng et al. (2020); Randy et al. (2023); Qiang et al. (2020); Zeng et al. (2025); Zhou et al. (2025; 2024).

Traditional FL algorithms, such as FedAvg McMahan et al. (2017), generally assume that all participating clients employ an identical model architecture Zhu et al. (2021); Li et al. (2025); Moon et al. (2025); Qiao et al. (2024). However, this assumption is often violated in practice, where clients possess heterogeneous computational resources, memory capacities, communication budgets, and application requirements Horváth (2021); Diao (2021); Liang et al. (2025); Yuan et al. (2024); Hu et al. (2024); Li et al. (2024). As a result, different clients may adopt different model architectures, giving rise to heterogeneous federated learning (HFL) Ye et al. (2024). While HFL offers more deployment flexibility than homogeneous FL, it also introduces a key challenge: enabling effective collaboration across heterogeneous parameter and representation spaces while maintaining effective information sharing among clients.

Existing HFL methods mainly address model heterogeneity through parameter sharing and feature alignment. Parameter-sharing approaches Wu et al. (2022); Yi et al. (2024; 2025b) typically identify and exchange homogeneous components across heterogeneous models or introduce additional homogeneous modules for collaboration. Feature-alignment approaches Yi et al. (2025a); Tan et al. (2022) attempt to align intermediate representations or semantic prototypes into a common feature space. Although these methods have achieved encouraging performance, they often rely on shared parameter spaces or explicit feature alignment. Such requirements hinder the construction of a unified collaboration interface across heterogeneous models. This motivates us to seek a shared representation that captures homogeneous structural information from heterogeneous feature spaces while remaining independent of model architectures and feature dimensions.

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.

To address this challenge, we revisit heterogeneous federated collaboration from a structural perspective. Rather than directly sharing model parameters or aligning feature representations, we seek a compact abstraction of local feature spaces. The singular values of a feature space matrix characterize the relative importance of its principal representation directions, while the corresponding singular value vector reflects how representation information is distributed across them. This provides a compact spectral structure that is independent of model architectures and feature dimensions. By mapping heterogeneous feature spaces into fixed length spectral structures, clients can express local structural information in a unified form and facilitate effective collaboration across heterogeneous models. Based on this idea, we introduce **Shared Spectral Structures** as a new collaboration interface for heterogeneous federated learning.

To this end, we propose **FedSpec**, a novel heterogeneous **Fed**erated learning framework based on shared **Spec**tral structures. (1) *Spectral abstraction* transforms heterogeneous local feature spaces into fixed length normalized spectral structures. (2) *Spectral aggregation* integrates local spectral structures into a global spectral structure that captures common patterns across clients. (3) *Spectral correction learning* uses the global spectral structure as a structural prior to guide local representation optimization. Together, these components enable effective collaboration across heterogeneous models without requiring parameter compatibility and feature alignment.

The main contributions of this work are summarized as follows:

- We propose *shared spectral structures*, which characterize the importance and information distribution of principal representation directions, providing a unified structural information carrier for heterogeneous models.
- We propose *FedSpec*, which integrates spectral abstraction, spectral aggregation, and spectral correction to enable effective collaboration across heterogeneous models without parameter compatibility or explicit feature alignment.
- We theoretically analyze the convergence behavior and information reduction properties of FedSpec, and empirically demonstrate its effectiveness and communication efficiency across diverse architectures, datasets, and data modalities.

2 Related Work

Recent heterogeneous federated learning methods mainly enable collaboration across heterogeneous local models through *parameter sharing* or *feature alignment*.

Parameter sharing. A line of work addresses model heterogeneity by introducing shared homogeneous components for collaboration Liang et al. (2020); Chen et al. (2021); Collins et al. (2021); Oh et al. (2022); Pillutla et al. (2022); Jang et al. (2022); Liu et al. (2022). These methods typically decompose local models into heterogeneous and homogeneous parts and exchange only compatible parameters. More recent approaches further introduce auxiliary homogeneous models to bridge heterogeneous local models Shen et al. (2020); Kalra et al. (2023); Qin et al. (2023). Representative methods include FedKD Wu et al. (2022), which employs a shared distillation model for knowledge transfer, pFedES Yi et al. (2025b), which constructs a shared feature extractor for semantic interaction, and FedMRL Yi et al. (2024), which introduces a lightweight homogeneous model for representation learning. Although effective, these methods generally require additional homogeneous structures while increasing communication overhead.

Feature alignment. Another line of work enables heterogeneous collaboration through feature level interaction Jeong et al. (2018); Ahn et al. (2019; 2020); He et al. (2020). These methods exchange intermediate representations, semantic prototypes, logits, or label related statistics to establish comparable semantic representations across heterogeneous clients. For example, FedRAL Yi et al. (2025a) performs semantic collaboration through representation space alignment among heterogeneous models. Compared with parameter sharing methods, feature alignment provides greater flexibility across heterogeneous architectures, but usually requires explicit alignment.

Different from these approaches, FedSpec enables heterogeneous collaboration through shared spectral structures. It first abstracts heterogeneous local feature spaces into local spectral structures and then aggregates them into a global spectrum, which serves as a structural prior to guide local representation optimization without requiring parameter compatibility or explicit feature alignment.

3 PROBLEM DEFINITION

3.1 PROBLEM FORMULATION

Consider a heterogeneous federated learning system with N clients, denoted by $\mathcal{C} = \{C_1, C_2, \cdots, C_N\}$. Each client C_i owns a private local dataset $\mathcal{D}_i = \{(\boldsymbol{x}_{i,j}, y_{i,j})\}_{j=1}^{n_i} \sim p_i$, where n_i denotes the number of local samples and p_i denotes the local data distribution. In general, $p_i \neq p_j$, indicating non-independent and identically distributed (non-IID) data across clients. At communication round t, the server samples a subset of clients $\mathcal{S}_t \subseteq \mathcal{C}$ according to a participation rate ρ, and only the selected clients participate in local training and server collaboration.

Let $F_i(\cdot; \theta_i)$ denote the local model of client C_i, where θ_i represents its model parameters. The local model is decomposed into a feature extractor $f_i(\cdot)$ and a prediction head $h_i(\cdot)$:

$$F_i(\cdot; \theta_i) = h_i(f_i(\cdot)).$$

Different clients may adopt different model architectures according to their computational resources and deployment requirements. Hence, $\theta_i \neq \theta_j$ generally holds for $i \neq j$. The goal of heterogeneous federated learning is therefore to collaboratively optimize $\{\theta_1, \theta_2, \cdots, \theta_N\}$ without sharing raw data, while enabling effective collaboration across heterogeneous models.

3.2 DESIGN OBJECTIVES

The key challenge is to establish a unified collaboration interface that is independent of local model architectures and feature dimensions.

For a local sample $\boldsymbol{x}_{i,j}$ of client C_i, the feature extractor produces

$$\boldsymbol{r}_{i,j} = f_i(\boldsymbol{x}_{i,j}), \qquad \boldsymbol{r}_{i,j} \in \mathbb{R}^{d_i},$$

where d_i denotes the local representation dimension. The corresponding representation set is

$$\mathcal{R}_i = \{\boldsymbol{r}_{i,1}, \boldsymbol{r}_{i,2}, \cdots, \boldsymbol{r}_{i,n_i}\}.$$

Since heterogeneous models may produce representations with different dimensions, i.e., $d_i \neq d_j$, directly sharing or aligning local representations is nontrivial. We therefore seek a mapping

$$\phi : \mathcal{R}_i \to \boldsymbol{s}_i,$$

where $\boldsymbol{s}_i \in \mathbb{R}^m$ and $m \leq d_i$ is a predefined constant shared by all clients. Here, $\boldsymbol{s}_i$ denotes a *shared structural abstraction* that provides a unified collaboration interface across heterogeneous clients.

An effective structural abstraction should satisfy the following properties:

- **Architecture Independence:** It should remain independent of local model architectures and representation dimensions.
- **Structural Expressiveness:** It should preserve informative structural characteristics of local representation spaces for knowledge sharing.
- **Collaborative Compatibility:** It should support direct aggregation across clients, allowing the server to construct a shared global structure.

Accordingly, the optimization objective is

$$\min_{\{\theta_i\}_{i=1}^N} \sum_{i=1}^{N} \frac{n_i}{\sum_{k=1}^{N} n_k} \mathcal{L}_i(\theta_i),$$

subject to

$$\boldsymbol{s}_i = \phi(\mathcal{R}_i), \qquad \boldsymbol{s}_i \in \mathbb{R}^m.$$

The central problem is to design $\phi(\cdot)$ such that heterogeneous local representation spaces can be transformed into a unified structural form that supports global aggregation and knowledge transfer.

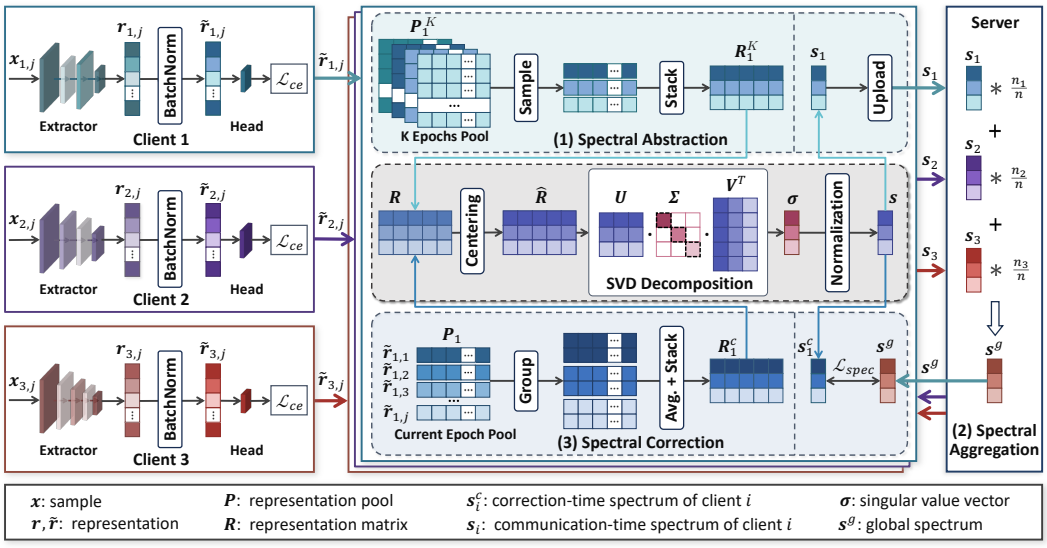


Figure 1: Overview of FedSpec. **(1) Spectral abstraction** transforms heterogeneous local representations into fixed length spectral structures. **(2) Spectral aggregation** integrates local spectra into a global spectral structure. **(3) Spectral correction learning** uses the global spectrum as a structural prior to guide local representation optimization. Together, these stages enable effective collaboration across heterogeneous models without parameter compatibility or explicit feature alignment.

4 The FedSpec Framework

This section presents the FedSpec framework, including an overview of the overall workflow, spectral abstraction for constructing local spectral structures, spectral aggregation for forming the global spectral structure, and spectral correction for guiding local representation optimization.

4.1 Overview

The overall framework is illustrated in Figure 1, and the complete training procedure is summarized in Algorithm 1 (Appendix). FedSpec operates through the following three key components:

(1) **Spectral Abstraction.** Each client transforms its local representation space into a fixed-length spectral structure that serves as a unified collaboration interface across clients.

(2) **Spectral Aggregation.** The server aggregates local spectral structures to construct a global spectral structure that captures common structural patterns shared across distributed clients.

(3) **Spectral Correction.** Each client utilizes the global spectral structure as a structural prior to guide local optimization and incorporate globally shared structural information.

Overall, FedSpec transforms heterogeneous local representation spaces into a unified spectral space, extracts shared structural knowledge from distributed clients, and injects the resulting global structural information into local optimization. Consequently, heterogeneous clients can effectively collaborate without requiring parameter compatibility or explicit feature-space alignment.

4.2 Spectral Abstraction

Spectral abstraction aims to transform heterogeneous local representation spaces into a unified spectral space. Although local representation spaces may differ in dimensionality, the importance distribution of their principal directions provides a dimension-independent characterization of the underlying representation structure. FedSpec therefore abstracts heterogeneous representation spaces into fixed-length spectral structures, enabling heterogeneous clients to exchange structural information

through a unified spectral space without parameter sharing or feature-space alignment. The spectral abstraction procedure is detailed as follows.

For a local sample $\boldsymbol{x}_{i,j}$ owned by client C_i, the local feature extractor produces a representation

$$\boldsymbol{r}_{i,j} = f_i(\boldsymbol{x}_{i,j}). \tag{1}$$

To reduce scale discrepancies across heterogeneous representations and ensure that the extracted spectral structure focuses on structural characteristics rather than feature magnitudes, Batch Normalization (BN) is first applied.

$$\tilde{\boldsymbol{r}}_{i,j} = \mathrm{BN}(\boldsymbol{r}_{i,j}). \tag{2}$$

FedSpec constructs a representation matrix $\boldsymbol{R}_i \in \mathbb{R}^{m \times d_i}$ from m representative normalized representations. The construction of $\boldsymbol{R}_i$ is flexible and can be adapted to different downstream objectives while preserving the same spectral abstraction process. The specific construction strategies adopted in this work are introduced in Sections 4.3 and 4.4.

To eliminate mean offsets, the representation matrix is centered as

$$\hat{\boldsymbol{R}}_i = \boldsymbol{R}_i - \mathbf{1}\boldsymbol{\mu}_i^\top, \tag{3}$$

where $\boldsymbol{\mu}_i$ denotes the column-wise mean vector.

FedSpec then performs Singular Value Decomposition (SVD)

$$\hat{\boldsymbol{R}}_i = \boldsymbol{U}_i \boldsymbol{\Sigma}_i \boldsymbol{V}_i^\top, \tag{4}$$

from which the singular value vector $\boldsymbol{\sigma}_i = [\sigma_{i,1}, \sigma_{i,2}, \ldots, \sigma_{i,m}]$ is extracted. The local spectral structure is then obtained through normalization

$$\boldsymbol{s}_i = \frac{\boldsymbol{\sigma}_i}{\sum_{k=1}^{m} \sigma_{i,k}}. \tag{5}$$

Definition 1 (Spectral Structure)**.** Given a local representation set $\mathcal{R}_i$, the corresponding spectral structure is defined as

$$\boldsymbol{s}_i = \phi(\mathcal{R}_i), \tag{6}$$

where $\boldsymbol{s}_i \in \mathbb{R}^m$.

The spectral structure possesses the following properties:

- **Architecture Independence.** Its dimension is fixed and independent of local model architectures and representation dimensions.
- **Structural Expressiveness.** It characterizes the relative importance of principal directions and preserves the dominant structural characteristics of local representation spaces.
- **Collaborative Compatibility.** It can be directly exchanged and aggregated across heterogeneous clients through a unified spectral space.

4.3 Spectral Aggregation

Spectral aggregation constructs a global spectral structure that captures common structural patterns across clients. To obtain a robust structural summary for communication, FedSpec adopts a *Stable Representation Sampling (SRS)* strategy to construct the communication-time representation matrix.

Specifically, after local training, client C_i collects all normalized representations generated during the last K local epochs and forms a representation candidate pool $\mathcal{P}_i^K$. Representations generated during the later stages of training are generally more stable and better reflect the converged local representation space. Therefore, utilizing representations from the last K local epochs reduces the influence of transient optimization fluctuations and provides a more reliable estimate of the underlying representation structure.

FedSpec then randomly samples m representations from $\mathcal{P}_i^K$ to construct the communication-time representation matrix

$$\boldsymbol{R}_i^K = \begin{bmatrix} \tilde{\boldsymbol{r}}_{i,1}^\top \\ \tilde{\boldsymbol{r}}_{i,2}^\top \\ \vdots \\ \tilde{\boldsymbol{r}}_{i,m}^\top \end{bmatrix} \in \mathbb{R}^{m \times d_i}. \tag{7}$$

Random sampling preserves diverse local structural characteristics while avoiding excessive storage and computational overhead. The spectral abstraction process described in Section 4.2 is then applied to $\boldsymbol{R}_i^K$ to obtain the uploaded local spectral structure $\boldsymbol{s}_i$.

After receiving local spectral structures from participating clients, the server performs sample-size-weighted aggregation like typical FedAvg:

$$\boldsymbol{s}^g = \sum_{i \in \mathcal{S}_t} \frac{n_i}{\sum_{k \in \mathcal{S}_t} n_k} \boldsymbol{s}_i, \tag{8}$$

where $\boldsymbol{s}^g$ denotes the global spectral structure.

The aggregated global spectral structure summarizes common structural patterns shared across distributed clients. After being broadcast in the subsequent communication round, it serves as a global structural prior for local optimization.

4.4 Spectral Correction

Spectral correction incorporates globally shared structural information into local optimization. While the global spectral structure captures common structural patterns shared across distributed clients, the local spectral structure reflects the current state of the local representation space. To progressively inject global structural knowledge into local representation learning, FedSpec performs spectral correction after standard local optimization within each local epoch. Specifically, each local epoch consists of the following two stages:

Stage 1: Local Optimization. Client C_i first updates its local model using the standard cross-entropy loss $\mathcal{L}_{ce}$. The updated local model is used to recompute representations in Stage 2.

Stage 2: Spectral Correction. Then, FedSpec constructs a local spectral structure from representations generated during the current epoch and utilizes it to perform structural correction.

To accurately estimate the current local representation structure, FedSpec adopts a *Group Representation Averaging (GRA)* strategy to construct the correction-time representation matrix.

Specifically, all normalized representations generated during the current local epoch are collected into a representation pool $\mathcal{P}_i$, which is randomly partitioned into m groups $\{\mathcal{G}_1, \ldots, \mathcal{G}_m\}$. For each group $\mathcal{G}_k$, the average representation is computed as

$$\bar{\boldsymbol{r}}_{i,k} = \frac{1}{|\mathcal{G}_k|} \sum_{\tilde{\boldsymbol{r}} \in \mathcal{G}_k} \tilde{\boldsymbol{r}}. \tag{9}$$

The group-level representations are used to construct the correction-time representation matrix

$$\boldsymbol{R}_i^c = \begin{bmatrix} \bar{\boldsymbol{r}}_{i,1}^\top \\ \bar{\boldsymbol{r}}_{i,2}^\top \\ \vdots \\ \bar{\boldsymbol{r}}_{i,m}^\top \end{bmatrix} \in \mathbb{R}^{m \times d_i}. \tag{10}$$

Compared with direct sampling, group averaging utilizes information from all representations generated during the current epoch and suppresses sample-level noise, thereby providing a more accurate estimate of the current local representation structure.

The spectral abstraction process described in Section 4.2 is then applied to $\boldsymbol{R}_i^c$ to obtain the correction-time spectral structure $\boldsymbol{s}_i^c$. FedSpec subsequently compares $\boldsymbol{s}_i^c$ with the global spectral structure $\boldsymbol{s}^g$ through the spectral correction loss

$$\mathcal{L}_{spec} = \lambda \|\boldsymbol{s}_i^c - \boldsymbol{s}^g\|_2^2. \tag{11}$$

The spectral correction loss is used to perform a second model update.

By minimizing $\mathcal{L}_{spec}$, the current local representation space progressively incorporates globally shared structural information. Importantly, spectral correction operates on spectral structures rather than local representations and therefore does not constrain the dimensionality or geometry of local representation spaces. Consequently, heterogeneous clients can benefit from shared structural knowledge while preserving their own parameter spaces and representation characteristics.

5 EXPERIMENTS

We implement FedSpec[1] and all baseline methods in PyTorch and conduct all experiments on NVIDIA RTX 3090 GPUs.

5.1 EXPERIMENTAL SETUP

Datasets and Data Partition. We conduct experiments on two widely used image classification benchmarks, CIFAR-100 with 100 classes[2] Krizhevsky et al. (2009) and Tiny-ImageNet with 200 classes[3] Chrabaszcz et al. (2017). To simulate pathological non-IID data heterogeneity, we adopt the commonly used label-skew partition strategy. For CIFAR-100, each client is assigned samples from 10 randomly selected classes, while for Tiny-ImageNet, each client is assigned samples from 20 randomly selected classes. Hence, clients possess highly heterogeneous local data distributions.

Models. Following recent HFL studies Yi et al. (2025b; 2024; 2025a), we evaluate FedSpec under heterogeneous model settings. Clients employ heterogeneous convolutional neural networks with different depths and parameter scales, resulting in distinct local model capacities.

Baselines. We compare FedSpec with representative state-of-the-art heterogeneous federated learning methods from different collaboration paradigms.

- *Standalone* independently trains each local model without inter-client collaboration while following the same client participation and local training schedule as the federated methods.

- *FedKD* Wu et al. (2022) performs heterogeneous collaboration through a homogeneous proxy model and knowledge distillation.

- *pFedES* Yi et al. (2025b) introduces a shared proxy homogeneous feature extractor to facilitate cross-client knowledge transfer.

- *FedMRL* Yi et al. (2024) performs collaboration through a shared proxy homogeneous model for representation learning.

- *FedRAL* Yi et al. (2025a) enables heterogeneous collaboration by representation space sharing.

Evaluation Metrics. We evaluate both model performance and communication efficiency. All experiments are repeated over three independent runs.

Test Accuracy. For model performance, we report the average test accuracy across all clients, computed as $\mathrm{Acc} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{Acc}_i$, Acc_i denotes the test accuracy of client i.

Communication Cost. For communication efficiency, we report the total communication cost required to reach a predefined target accuracy. Specifically, the total communication cost is computed as the product of the communication rounds required to reach the target accuracy and the per-round communication cost transmitted from participating clients to the server. This metric jointly reflects communication overhead and convergence efficiency.

Hyperparameter Settings. We consider a heterogeneous federated learning system with $N = 100$ clients. The client participation rate is set to $\rho \in \{20\%, 10\%, 5\%\}$ unless otherwise specified. The total communication rounds are fixed to $T = 500$, and each participating client performs $E = 10$ local epochs per communication round. All local models are optimized using SGD with a learning rate of 0.01, and the local batch size is set to 512. For fair comparison, all baseline methods use the officially recommended hyperparameter settings reported in their original papers. For FedSpec, the spectral correction weight is set to $\lambda = \{0.05, 0.3\}$, the Stable Representation Sampling window is set to $K = \{5, 10\}$, and the spectral dimension is set to $m = \{10, 100\}$ for the two datasets, respectively, unless otherwise specified.

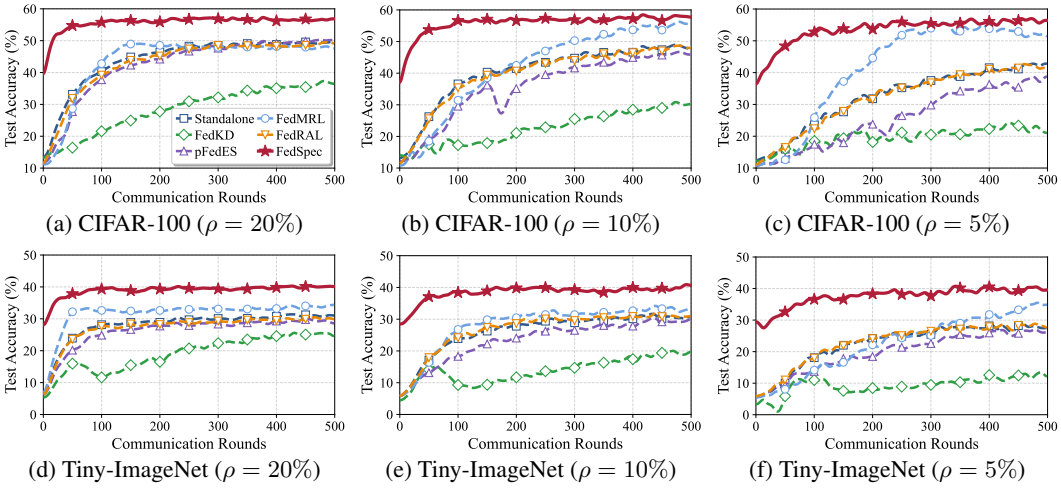
[1] https://anonymous.4open.science/r/FedSpec

[2] https://www.cs.toronto.edu/%7Ekriz/cifar.html

[3] http://cs231n.stanford.edu/tiny-imagenet-200.zip

Table 1: Average test accuracy (%) under pathological label-skew partitions with different client participation rates. The best and second-best results are highlighted in bold and underlined.

	$N = 100, \rho = 20\%$		$N = 100, \rho = 10\%$		$N = 100, \rho = 5\%$	
Method	CIFAR-100	Tiny-ImageNet	CIFAR-100	Tiny-ImageNet	CIFAR-100	Tiny-ImageNet
Standalone	<u>55.37±0.87</u>	34.93±0.36	53.39±0.71	35.36±0.59	50.68±0.24	35.47±0.82
FedKD Wu et al. (2022)	40.75±0.48	29.73±0.97	35.39±0.65	24.54±0.78	35.64±0.33	19.77±0.93
pFedES Yi et al. (2025b)	54.32±0.84	31.94±0.43	50.02±0.96	36.83±0.28	46.23±0.74	30.27±0.57
FedMRL Yi et al. (2024)	52.32±0.26	<u>36.49±0.81</u>	<u>60.26±0.41</u>	<u>37.42±0.68</u>	<u>62.44±0.22</u>	<u>42.23±0.89</u>
FedRAL Yi et al. (2025a)	53.69±0.76	33.67±0.31	53.32±0.92	35.31±0.52	53.09±0.99	34.97±0.38
FedSpec (Ours)	**61.94±0.17**	**43.10±0.44**	**63.70±0.29**	**45.09±0.21**	**64.56±0.35**	**47.30±0.14**



(a) CIFAR-100 ($\rho = 20\%$) (b) CIFAR-100 ($\rho = 10\%$) (c) CIFAR-100 ($\rho = 5\%$)

(d) Tiny-ImageNet ($\rho = 20\%$) (e) Tiny-ImageNet ($\rho = 10\%$) (f) Tiny-ImageNet ($\rho = 5\%$)

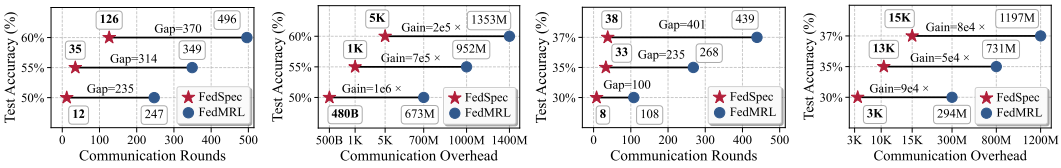
Figure 2: Convergence curves under pathological label-skew heterogeneous settings with different client participation rates. FedSpec consistently achieves faster convergence and higher accuracy.

5.2 Main Results

Accuracy Comparison. Table 1 reports the average test accuracy under pathological label-skew heterogeneous federated learning settings. FedSpec consistently achieves the best performance across both datasets and all client participation rates, outperforming the strongest baseline by up to 6.57% on CIFAR-100 and 7.67% on Tiny-ImageNet. The advantage remains consistent as the participation rate varies from 20% to 5%, demonstrating the effectiveness of spectral collaboration when only a subset of heterogeneous clients participates in each communication round. Moreover, FedSpec consistently outperforms FedKD, pFedES, FedMRL, and FedRAL, indicating that spectral structure sharing provides an effective alternative to distillation, auxiliary model sharing, and explicit representation space alignment for heterogeneous collaboration.

Convergence Analysis. Figure 2 visualizes the convergence behavior corresponding to Table 1. FedSpec consistently achieves faster convergence and higher final accuracy than the compared methods across both datasets and different participation rates. This indicates that the global spectral structure provides useful shared structural information for guiding local representation optimization, thereby improving the effectiveness of heterogeneous collaboration.

Communication Efficiency. Figure 3 compares the communication efficiency of FedSpec with the state-of-the-art FedMRL, which achieves the strongest overall performance among the compared baselines in Table 1. FedSpec consistently requires fewer communication rounds to reach the same target accuracy while transmitting only compact spectral structures instead of auxiliary homogeneous model parameters. Consequently, FedSpec reduces the cumulative communication cost by several orders of magnitude while maintaining superior model performance. These results demonstrate that spectral structure sharing enables effective heterogeneous collaboration with substantially lower communication overhead.



(a) Round Gap (b) Comm. Gain (c) Round Gap (d) Comm. Gain

Figure 3: Communication efficiency comparison between FedSpec and FedMRL under $\rho = 10\%$. (a)-(b) report results on CIFAR-100, while (c)-(d) report results on Tiny-ImageNet. "Round Gap" denotes the reduction in communication rounds required to reach the same accuracy threshold, and "Comm. Gain" denotes the multiplicative reduction in cumulative communication cost.

Table 2: Performance comparison (%) across diverse model architectures and data modalities. The best and second-best results are highlighted in bold and underlined, respectively.

Method	CIFAR-100	Fed-Heart-Disease	Stack Overflow
Modality	*CV*	*Tabular*	*NLP*
Standalone	45.58 ± 0.63	65.83 ± 0.74	16.35 ± 0.11
FedKD Wu et al. (2022)	32.43 ± 0.82	62.48 ± 1.07	17.53 ± 0.19
pFedES Yi et al. (2025b)	43.72 ± 0.47	64.91 ± 0.58	18.21 ± 0.14
FedMRL Yi et al. (2024)	$\underline{51.36 \pm 0.36}$	$\underline{67.54 \pm 0.81}$	$\underline{19.06 \pm 0.23}$
FedRAL Yi et al. (2025a)	46.95 ± 0.55	66.32 ± 0.49	18.47 ± 0.17
FedSpec (Ours)	$\mathbf{55.47 \pm 0.28}$	$\mathbf{70.12 \pm 0.66}$	$\mathbf{20.14 \pm 0.09}$

5.3 Generality across Architectures and Modalities

Settings. To further evaluate the generality of FedSpec beyond heterogeneous CNN based image classification, we consider three additional settings covering more diverse model architectures and data modalities. For CIFAR-100, we construct a highly heterogeneous model setting involving ResNet-{4,6,8,10,18,34,50,101,152}, CNN, and ViT models, introducing substantial differences in network architectures and model capacities. We further evaluate FedSpec on Fed-Heart-Disease, a cross silo healthcare tabular classification benchmark, and Stack Overflow, an NLP benchmark for next word prediction with heterogeneous LSTM models. We compare FedSpec with the same representative HFL baselines used in the main experiments with a $\rho = 10\%$ client participation rate.

Results. As shown in Table 2, FedSpec consistently achieves the best performance across all three settings. In particular, the heterogeneous CIFAR-100 setting shows that spectral collaboration remains effective across substantially different network architectures, while the results on Fed-Heart-Disease and Stack Overflow extend its effectiveness from vision to tabular and NLP modalities. These results provide further empirical evidence that shared spectral structures offer a unified collaboration interface that is not restricted to a specific model architecture or data modality. [4]

6 Conclusion

This paper proposed FedSpec, a heterogeneous federated learning framework based on shared spectral structures. FedSpec abstracts heterogeneous local feature spaces into compact spectral structures, aggregates them into a global spectral structure, and uses the global structure as a structural prior to guide local representation optimization. Theoretical analysis establishes the convergence behavior and information reduction properties of FedSpec. Extensive experiments demonstrate its effectiveness, communication efficiency, and robustness under different heterogeneous settings, while additional evaluations across diverse model architectures and data modalities further support its generality. These results demonstrate the potential of shared spectral structures as an effective collaboration interface for heterogeneous federated learning.

[4] The Appendix provides the complete algorithm description; computational and communication complexity analysis; detailed theoretical analysis of convergence and information reduction; and additional experiments on sensitivity and robustness, the effectiveness of spectral structures, and ablation studies.

AI USE STATEMENT

Generative AI tools were used solely for *language polishing* and improving the clarity of the manuscript. All AI-assisted content was carefully reviewed and revised by the authors, and all technical statements were verified against the methods, theoretical analysis, and experimental results. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work focuses on collaborative mechanisms for heterogeneous federated learning and does not involve the recruitment of human subjects or the collection of new sensitive data. All experiments are conducted on existing benchmark datasets. FedSpec communicates only spectral structures during collaboration and does not directly transmit raw data. The authors *adhere to the ICLR Code of Ethics* and take responsibility for the integrity of data usage, experimental design, result reporting, and research conclusions.

REPRODUCIBILITY STATEMENT

We provide the main information required to reproduce this work in the main paper and the appendix. The main paper describes the overall FedSpec framework, including spectral abstraction, spectral aggregation, and spectral correction, together with the datasets, data partitions, model settings, baselines, evaluation metrics, and major hyperparameters. The appendix further provides the complete training algorithm and additional results on sensitivity, robustness, effectiveness, and ablation studies. In addition, an anonymous link to the *source code* of FedSpec is provided in the experimental section to further facilitate the reproduction of the method and experimental results.

REFERENCES

Jin-Hyun Ahn et al. Wireless federated distillation for distributed edge learning with heterogeneous data. In *Proc. PIMRC*, pp. 1–6, Istanbul, Turkey, 2019. IEEE.

Jin-Hyun Ahn et al. Cooperative learning VIA federated distillation OVER fading channels. In *Proc. ICASSP*, pp. 8856–8860, Barcelona, Spain, 2020. IEEE.

Jiangui Chen et al. Fedmatch: Federated learning over heterogeneous question answering data. In *Proc. CIKM*, pp. 181–190, virtual, 2021. ACM.

Yong Cheng, Yang Liu, Tianjian Chen, and Qiang Yang. Federated learning for privacy-preserving ai. *Communications of the ACM*, 63(12):33–36, 2020.

Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the CIFAR datasets. *CoRR*, abs/1707.08819, 2017.

Liam Collins et al. Exploiting shared representations for personalized federated learning. In *Proc. ICML*, volume 139, pp. 2089–2099, virtual, 2021. PMLR.

Enmao Diao. Heterofl: Computation and communication efficient federated learning for heterogeneous clients. In *Proc. ICLR*, pp. 1, Virtual Event, Austria, 2021. OpenReview.net.

Chaoyang He et al. Group knowledge transfer: Federated learning of large cnns at the edge. In *Proc. NeurIPS*, pp. 1, virtual, 2020. .

S. Horváth. FjORD: Fair and accurate federated learning under heterogeneous targets with ordered dropout. In *Proc. NIPS*, pp. 12876–12889, Virtual, 2021. OpenReview.net.

Ming Hu et al. Fedcross: Towards accurate federated learning via multi-model cross-aggregation. In *Proc. ICDE*, pp. 2137–2150, Utrecht, The Netherlands, 2024.

Jaehee Jang et al. Fedclassavg: Local representation learning for personalized federated learning on heterogeneous neural networks. In *Proc. ICPP*, pp. 76:1–76:10, virtual, 2022. ACM.

Eunjeong Jeong et al. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. In *Proc. NeurIPS Workshop on Machine Learning on the Phone and other Consumer Devices*, pp. 1, virtual, 2018. .

Shivam Kalra et al. Decentralized federated learning through proxy model sharing. *Nature communications*, 14(1):2899, 2023.

Alex Krizhevsky et al. *Learning multiple layers of features from tiny images*. Toronto, ON, Canada, , 2009.

Xunkai Li et al. Adafgl: A new paradigm for federated node classification with topology heterogeneity. In *Proc. ICDE*, pp. 2517–2530, Utrecht, The Netherlands, 2024.

Yuting Li et al. pfssl-d: Generalization meets personalization in dual-phase federated semi-supervised learning. In *Proc. ICDE*, pp. 2188–2200, Hong Kong, 2025.

Juntao Liang et al. Fedecover: Fast and stable converging model-heterogeneous federated learning with efficient-coverage submodel extraction. In *Proc. ICDE*, pp. 2575–2587, Hong Kong, 2025.

Paul Pu Liang et al. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 1(1), 2020.

Chang Liu et al. Completely heterogeneous federated learning. *CoRR*, abs/2210.15865, 2022.

Brendan McMahan et al. Communication-efficient learning of deep networks from decentralized data. In *Proc. AISTATS*, volume 54, pp. 1273–1282, Fort Lauderdale, FL, USA, 2017. PMLR.

Jihoon Moon et al. Fedsdp: Federated self-derived prototypes for personalized federated learning. In *Proc. ICDE*, pp. 2796–2809, Hong Kong, 2025.

Jaehoon Oh et al. Fedbabu: Toward enhanced representation for federated image classification. In *Proc. ICLR*, pp. 1, virtual, 2022. OpenReview.net.

Krishna Pillutla et al. Federated learning with partial model personalization. In *Proc. ICML*, volume 162, pp. 17716–17758, virtual, 2022. PMLR.

Yang Qiang, Fan Lixin, and Yu Han. *Federated Learning: Privacy and Incentive*. Springer, Cham, 2020.

Pengpeng Qiao et al. Feed: Towards personalization-effective federated learning. In *Proc. ICDE*, pp. 1779–1791, Utrecht, The Netherlands, 2024.

Zhen Qin et al. Fedapen: Personalized cross-silo federated learning with adaptability to statistical heterogeneity. In *Proc. KDD*, pp. 1954–1964, Long Beach, CA, USA, 2023. ACM.

Goebel Randy, Yu Han, Faltings Boi, Fan Lixin, and Xiong Zehui. *Trustworthy Federated Learning*. Springer, Cham, 2023.

Tao Shen et al. Federated mutual learning. *CoRR*, abs/2006.16765, 2020.

Yue Tan et al. Fedproto: Federated prototype learning across heterogeneous clients. In *Proc. AAAI*, pp. 8432–8440, virtual, 2022. AAAI Press.

Chuhan Wu et al. Communication-efficient federated learning via knowledge distillation. *Nature Communications*, 13(1):2032, 2022.

Mang Ye et al. Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Comput. Surv.*, 56(3):79:1–79:44, 2024.

Liping Yi et al. Federated model heterogeneous matryoshka representation learning. In *Proc. NeurIPS*, pp. 1, Vancouver, Canada, 2024. .

Liping Yi et al. Federated representation angle learning. In *Proc. ICCV*, pp. 1314–1324, Honolulu, Hawai'i, USA, 2025a.

Liping Yi et al. pfedes: Generalized proxy feature extractor sharing for model heterogeneous personalized federated learning. In *Proc. AAAI*, pp. 22146–22154, Philadelphia, PA, USA, 2025b. AAAI Press.

Wei Yuan et al. Hetefedrec: Federated recommender systems with model heterogeneity. In *Proc. ICDE*, pp. 1324–1337, Utrecht, The Netherlands, 2024.

Zhihao Zeng et al. Heterogeneous-aware traffic prediction: A privacy-preserving federated learning framework. In *Proc. ICDE*, pp. 419–432, Hong Kong, 2025.

Ting Zhou et al. Robfl: Robust federated learning via feature center separation and malicious center detection. In *Proc. ICDE*, pp. 926–938, Utrecht, The Netherlands, 2024.

Xiaokai Zhou et al. Hounding data diversity: Towards participant selection in vertical federated learning. In *Proc. ICDE*, pp. 2810–2823, Hong Kong, 2025.

Hangyu Zhu et al. Federated learning on non-iid data: A survey. *Neurocomputing*, 465:371–390, 2021.

A Algorithm Details

Algorithm 1 Training Procedure of FedSpec.

Input: Client set $\mathcal{C}$, participation rate ρ, communication rounds T, local epochs E, stable epoch window K, spectral dimension m, correction coefficient λ.
Output: Trained heterogeneous models $\{F_i(\cdot;\theta_i)\}_{i=1}^{N}$.
Randomly initialize local models and a normalized global spectral structure $\boldsymbol{s}^g$.
for $t = 1$ to T **do**
  Sample participating clients $\mathcal{S}_t$ and broadcast $\boldsymbol{s}^g$.
  for each client $C_i \in \mathcal{S}_t$ in parallel **do**
    for $e = 1$ to E **do**
      Update local model using cross entropy loss $\mathcal{L}_{ce}$.
      ▷ **Spectral Correction**
      Construct $\boldsymbol{R}_i^c$ from current epoch representations using GRA.
      Obtain $\boldsymbol{s}_i^c$ according to Eqs. (3)–(5).
      Update local model using $\mathcal{L}_{spec}$ in Eq. (11).
    end for
    ▷ **Spectral Abstraction**
    Construct $\boldsymbol{R}_i^K$ from the last K local epochs using SRS.
    Obtain local spectral structure $\boldsymbol{s}_i$ according to Eqs. (3)–(5).
    Upload $\boldsymbol{s}_i$ to the server.
  end for
  ▷ **Spectral Aggregation**
  Compute global spectral structure $\boldsymbol{s}^g$ according to Eq. (8).
end for
return $\{F_i(\cdot;\theta_i)\}_{i=1}^{N}$.

B Complexity Analysis

We analyze the additional computational and communication costs introduced by FedSpec.

Spectral Abstraction Complexity. The main computational cost of spectral abstraction comes from SVD on the representation matrix $\boldsymbol{R}_i \in \mathbb{R}^{m\times d_i}$. For $m \le d_i$, its complexity is $O(m^2 d_i)$. Representation normalization, matrix construction, and centering require $O(md_i)$ operations. Therefore, each spectral abstraction has a computational complexity of $O(m^2 d_i)$.

Spectral Aggregation Complexity. Each participating client uploads a spectral structure $\boldsymbol{s}_i \in \mathbb{R}^m$ and receives the global spectrum $\boldsymbol{s}^g \in \mathbb{R}^m$, resulting in $O(m)$ communication per client in each round. On the server, aggregating spectra from $|\mathcal{S}_t|$ participating clients requires $O(|\mathcal{S}_t|m)$ computation. Since m is fixed and independent of model architectures, parameter sizes, and feature dimensions, the communication cost remains lightweight across heterogeneous models.

Spectral Correction Complexity. During each local epoch, GRA averages the local representations with a complexity of $O(n_i d_i)$, followed by spectral abstraction with a complexity of $O(m^2 d_i)$. Computing the spectral correction loss requires only $O(m)$. Therefore, the additional complexity of spectral correction in each local epoch is $O(n_i d_i + m^2 d_i)$.

Overall Complexity. Since spectral correction is performed for each of the E local epochs and one additional spectral structure is extracted for communication, the additional computational complexity per client in each communication round is

$$O\big(En_i d_i + (E+1)m^2 d_i\big)\,.$$

The communication complexity remains $O(m)$ per participating client. With a small predefined spectral dimension m, FedSpec enables collaboration across heterogeneous models with modest additional computation and communication costs.

C THEORETICAL ANALYSIS

This section analyzes the convergence behavior and information reduction properties of FedSpec.

C.1 CONVERGENCE ANALYSIS

FedSpec performs two consecutive updates within each local epoch: standard local optimization followed by spectral correction. We analyze its convergence under non-convex objectives while explicitly accounting for this two-stage update.

At communication round t, the server broadcasts the global spectral structure $\boldsymbol{s}^{g,t}$. Let $\mathcal{L}_i(\theta_i)$ denote the empirical objective of client C_i, and let $\mathcal{L}^t_{spec}(\theta_i)$ denote the spectral correction objective induced by $\boldsymbol{s}^{g,t}$. We define the round-wise composite objective as

$$\mathcal{J}^t_i(\theta_i) = \mathcal{L}_i(\theta_i) + \mathcal{L}^t_{spec}(\theta_i). \tag{12}$$

At local epoch e, FedSpec first performs standard local optimization

$$\tilde{\theta}^{t,e}_i = \theta^{t,e}_i - \eta \boldsymbol{g}^{t,e}_i, \tag{13}$$

where $\boldsymbol{g}^{t,e}_i$ is a stochastic gradient estimator of $\nabla\mathcal{L}_i(\theta^{t,e}_i)$. The subsequent spectral correction update is

$$\theta^{t,e+1}_i = \tilde{\theta}^{t,e}_i - \eta \boldsymbol{q}^{t,e}_i, \tag{14}$$

where $\boldsymbol{q}^{t,e}_i$ is a stochastic gradient estimator of $\nabla\mathcal{L}^t_{spec}(\tilde{\theta}^{t,e}_i)$. Equivalently,

$$\theta^{t,e+1}_i = \theta^{t,e}_i - \eta \boldsymbol{u}^{t,e}_i, \qquad \boldsymbol{u}^{t,e}_i = \boldsymbol{g}^{t,e}_i + \boldsymbol{q}^{t,e}_i. \tag{15}$$

For simplicity, the analysis below indexes the communication rounds in which client C_i participates. Under partial participation, the same analysis applies by reindexing the active rounds of each client. For notational simplicity, T denotes the number of active communication rounds of client C_i in the following analysis.

Assumption 1 (Lower Boundedness)**.** *For any client C_i and communication round t, the objective $\mathcal{J}^t_i(\theta_i)$ is lower bounded by a finite constant $\mathcal{J}^\star$.*

Assumption 2 (Smoothness)**.** *For any client C_i and communication round t, the composite objective $\mathcal{J}^t_i(\theta_i)$ is L_J-smooth, and the spectral correction objective $\mathcal{L}^t_{spec}(\theta_i)$ is L_s-smooth.*

Assumption 3 (Stochastic Gradient Conditions)**.** *Let $\mathcal{F}_{t,e}$ denote the optimization history up to $\theta^{t,e}_i$. The stochastic gradients satisfy*

$$\mathbb{E}\left[\boldsymbol{g}^{t,e}_i \mid \mathcal{F}_{t,e}\right] = \nabla\mathcal{L}_i(\theta^{t,e}_i), \tag{16}$$

and

$$\mathbb{E}\left[\boldsymbol{q}^{t,e}_i \mid \tilde{\theta}^{t,e}_i\right] = \nabla\mathcal{L}^t_{spec}(\tilde{\theta}^{t,e}_i). \tag{17}$$

Moreover, there exist constants $G_c > 0$ and $U > 0$ such that

$$\mathbb{E}\left[\left\|\boldsymbol{g}^{t,e}_i\right\|_2^2 \mid \mathcal{F}_{t,e}\right] \le G_c^2, \tag{18}$$

and

$$\mathbb{E}\left[\left\|\boldsymbol{u}^{t,e}_i\right\|_2^2 \mid \mathcal{F}_{t,e}\right] \le U^2. \tag{19}$$

Assumption 4 (Bounded Inter-round Variation)**.** *The cumulative positive variation of the round-wise objective caused by changes in the global spectral structure is bounded. Specifically, there exists a constant $\Omega < \infty$, independent of T, such that*

$$\sum_{t=0}^{T-2} \mathbb{E}\left[\left[\mathcal{J}^{t+1}_i(\theta^{t+1,0}_i) - \mathcal{J}^t_i(\theta^{t,E}_i)\right]_+\right] \le \Omega, \tag{20}$$

where $[x]_+ = \max(x, 0)$.

Theorem 1 (Finite-time Convergence Bound)**.** *Under Assumptions 1–4, the two-stage local optimization of FedSpec satisfies*

$$\begin{aligned}&\frac{1}{TE}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\mathbb{E}\left[\left\|\nabla\mathcal{J}_i^t(\theta_i^{t,e})\right\|_2^2\right]\\&\leq\frac{2\left(\mathcal{J}_i^0(\theta_i^{0,0})-\mathcal{J}^\star+\Omega\right)}{\eta TE}+L_J\eta U^2+L_s^2\eta^2G_c^2.\end{aligned}\tag{21}$$

By choosing $\eta = O(1/\sqrt{TE})$, the average stationarity residual satisfies

$$\frac{1}{TE}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\mathbb{E}\left[\left\|\nabla\mathcal{J}_i^t(\theta_i^{t,e})\right\|_2^2\right]=O\left(\frac{1}{\sqrt{TE}}\right).\tag{22}$$

For a fixed number of local epochs E, the convergence rate with respect to the number of participating communication rounds is $O(1/\sqrt{T})$.

Proof. For simplicity, denote $\nabla\mathcal{J}_i^t(\theta_i^{t,e})$ by $\nabla\mathcal{J}_i^{t,e}$. Since the spectral correction gradient is evaluated at the intermediate parameter $\tilde{\theta}_i^{t,e}$, the two-stage direction $\boldsymbol{u}_i^{t,e}$ is not an unbiased estimator of $\nabla\mathcal{J}_i^{t,e}$. We first bound this discrepancy.

Conditioned on $\mathcal{F}_{t,e}$, Assumption 3 gives

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{u}_i^{t,e}\mid\mathcal{F}_{t,e}\right]=&\nabla\mathcal{L}_i(\theta_i^{t,e})\\&+\mathbb{E}\left[\nabla\mathcal{L}_{spec}^t(\tilde{\theta}_i^{t,e})\mid\mathcal{F}_{t,e}\right].\end{aligned}\tag{23}$$

Define the bias term

$$\begin{aligned}\boldsymbol{b}_i^{t,e}=&\mathbb{E}\left[\boldsymbol{u}_i^{t,e}\mid\mathcal{F}_{t,e}\right]-\nabla\mathcal{J}_i^{t,e}\\=&\mathbb{E}\left[\nabla\mathcal{L}_{spec}^t(\tilde{\theta}_i^{t,e})-\nabla\mathcal{L}_{spec}^t(\theta_i^{t,e})\mid\mathcal{F}_{t,e}\right].\end{aligned}\tag{24}$$

By the L_s-smoothness of $\mathcal{L}_{spec}^t$ and Eq. (13),

$$\begin{aligned}\left\|\boldsymbol{b}_i^{t,e}\right\|_2&\leq L_s\mathbb{E}\left[\left\|\tilde{\theta}_i^{t,e}-\theta_i^{t,e}\right\|_2\mid\mathcal{F}_{t,e}\right]\\&=L_s\eta\mathbb{E}\left[\left\|\boldsymbol{g}_i^{t,e}\right\|_2\mid\mathcal{F}_{t,e}\right]\\&\leq L_s\eta G_c,\end{aligned}\tag{25}$$

where the last inequality follows from Eq. (18) and Jensen's inequality.

By the L_J-smoothness of $\mathcal{J}_i^t$ and Eq. (15),

$$\begin{aligned}\mathcal{J}_i^t(\theta_i^{t,e+1})\leq&\mathcal{J}_i^t(\theta_i^{t,e})-\eta\left\langle\nabla\mathcal{J}_i^{t,e},\boldsymbol{u}_i^{t,e}\right\rangle\\&+\frac{L_J\eta^2}{2}\left\|\boldsymbol{u}_i^{t,e}\right\|_2^2.\end{aligned}\tag{26}$$

Taking the conditional expectation with respect to $\mathcal{F}_{t,e}$ and using Eq. (24), we obtain

$$\begin{aligned}&\mathbb{E}\left[\mathcal{J}_i^t(\theta_i^{t,e+1})\mid\mathcal{F}_{t,e}\right]\\\leq&\mathcal{J}_i^t(\theta_i^{t,e})-\eta\left\|\nabla\mathcal{J}_i^{t,e}\right\|_2^2-\eta\left\langle\nabla\mathcal{J}_i^{t,e},\boldsymbol{b}_i^{t,e}\right\rangle+\frac{L_J\eta^2}{2}U^2.\end{aligned}\tag{27}$$

Using Young's inequality

$$-\langle\boldsymbol{a},\boldsymbol{b}\rangle\leq\frac{1}{2}\|\boldsymbol{a}\|_2^2+\frac{1}{2}\|\boldsymbol{b}\|_2^2,\tag{28}$$

together with Eq. (25), gives

$$\begin{aligned}&\mathbb{E}\left[\mathcal{J}_i^t(\theta_i^{t,e+1}) \mid \mathcal{F}_{t,e}\right]\\ \leq&\mathcal{J}_i^t(\theta_i^{t,e}) - \frac{\eta}{2}\left\|\nabla\mathcal{J}_i^{t,e}\right\|_2^2 + \frac{L_s^2\eta^3}{2}G_c^2 + \frac{L_J\eta^2}{2}U^2.\end{aligned} \tag{29}$$

Taking the total expectation and summing Eq. (29) over $e = 0, \ldots, E-1$ and $t = 0, \ldots, T-1$, we obtain

$$\begin{aligned}&\frac{\eta}{2}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\mathbb{E}\left[\left\|\nabla\mathcal{J}_i^{t,e}\right\|_2^2\right]\\ \leq&\sum_{t=0}^{T-1}\left(\mathbb{E}\left[\mathcal{J}_i^t(\theta_i^{t,0})\right] - \mathbb{E}\left[\mathcal{J}_i^t(\theta_i^{t,E})\right]\right)\\ &+ \frac{TEL_s^2\eta^3}{2}G_c^2 + \frac{TEL_J\eta^2}{2}U^2.\end{aligned} \tag{30}$$

The first term on the right hand side satisfies

$$\begin{aligned}&\sum_{t=0}^{T-1}\left(\mathbb{E}[\mathcal{J}_i^t(\theta_i^{t,0})] - \mathbb{E}[\mathcal{J}_i^t(\theta_i^{t,E})]\right)\\ \leq&\mathcal{J}_i^0(\theta_i^{0,0}) - \mathcal{J}^\star + \Omega,\end{aligned} \tag{31}$$

where we use Assumptions 1 and 4 to bound the objective variation between consecutive communication rounds.

Substituting Eq. (31) into Eq. (30) yields

$$\begin{aligned}&\frac{\eta}{2}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\mathbb{E}\left[\left\|\nabla\mathcal{J}_i^{t,e}\right\|_2^2\right]\\ \leq&\mathcal{J}_i^0(\theta_i^{0,0}) - \mathcal{J}^\star + \Omega + \frac{TEL_s^2\eta^3}{2}G_c^2 + \frac{TEL_J\eta^2}{2}U^2.\end{aligned} \tag{32}$$

Dividing both sides by $\eta TE/2$ gives Eq. (21). With $\eta = O(1/\sqrt{TE})$, the three terms on the right hand side scale as $O(1/\sqrt{TE})$, $O(1/\sqrt{TE})$, and $O(1/(TE))$, respectively. Therefore, Eq. (22) follows.

□

Theorem 1 shows that the two-stage optimization of FedSpec achieves the standard $O(1/\sqrt{TE})$ first-order stationarity rate under the stated assumptions. The bias introduced by evaluating spectral correction at the intermediate model contributes only an $O(\eta^2)$ term and does not change the asymptotic convergence rate.

C.2 Information Reduction Analysis

FedSpec does not provide a formal privacy guarantee. Instead, we analyze the information reduction induced by spectral abstraction during communication.

Recall that each client communicates only the local spectral structure

$$\boldsymbol{s}_i = \phi(\boldsymbol{R}_i^K), \tag{33}$$

where $\boldsymbol{R}_i^K$ is the communication-time representation matrix constructed by SRS, and $\phi(\cdot)$ denotes the spectral abstraction defined in Section 4.2. The corresponding information flow can be expressed as

$$\mathcal{D}_i \to \boldsymbol{R}_i^K \to \boldsymbol{s}_i. \tag{34}$$

Theorem 2 (Information Reduction Property)**.** *For any client C_i, the communicated spectral structure satisfies*

$$I(\mathcal{D}_i; \boldsymbol{s}_i) \leq I(\mathcal{D}_i; \boldsymbol{R}_i^K), \tag{35}$$

where $I(\cdot;\cdot)$ denotes mutual information.

Proof. Since $\boldsymbol{s}_i = \phi(\boldsymbol{R}_i^K)$ is obtained solely from $\boldsymbol{R}_i^K$, the variables form the Markov chain

$$\mathcal{D}_i \to \boldsymbol{R}_i^K \to \boldsymbol{s}_i. \tag{36}$$

According to the data processing inequality,

$$I(\mathcal{D}_i; \boldsymbol{s}_i) \leq I(\mathcal{D}_i; \boldsymbol{R}_i^K). \tag{37}$$

Therefore, spectral abstraction cannot increase the mutual information with the local dataset relative to communicating the underlying representation matrix. □

Theorem 2 shows that spectral abstraction does not increase the information exposed about local data compared with directly communicating local representation matrices.

Theorem 3 (Non-invertibility of Spectral Representation)**.** *The mapping from a centered representation matrix to its normalized spectral structure is non-injective. Specifically, there exist distinct centered representation matrices $\hat{\boldsymbol{R}} \neq \hat{\boldsymbol{R}}'$ that produce the same normalized spectral structure.*

Proof. Let

$$\hat{\boldsymbol{R}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \tag{38}$$

be the singular value decomposition of a centered representation matrix $\hat{\boldsymbol{R}} \in \mathbb{R}^{m \times d_i}$.

Consider an orthogonal matrix $\boldsymbol{Q} \in \mathbb{R}^{d_i \times d_i}$ satisfying

$$\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}, \tag{39}$$

and define

$$\hat{\boldsymbol{R}}' = \hat{\boldsymbol{R}}\boldsymbol{Q}. \tag{40}$$

Since $\hat{\boldsymbol{R}}$ is centered,

$$\mathbf{1}^\top \hat{\boldsymbol{R}}' = \mathbf{1}^\top \hat{\boldsymbol{R}}\boldsymbol{Q} = \mathbf{0}, \tag{41}$$

and therefore $\hat{\boldsymbol{R}}'$ is also centered.

Moreover,

$$\hat{\boldsymbol{R}}'\hat{\boldsymbol{R}}'^\top = \hat{\boldsymbol{R}}\boldsymbol{Q}\boldsymbol{Q}^\top\hat{\boldsymbol{R}}^\top = \hat{\boldsymbol{R}}\hat{\boldsymbol{R}}^\top. \tag{42}$$

Hence, $\hat{\boldsymbol{R}}$ and $\hat{\boldsymbol{R}}'$ have identical singular values and consequently produce the same normalized spectral structure.

For any orthogonal transformation $\boldsymbol{Q}$ such that $\hat{\boldsymbol{R}}\boldsymbol{Q} \neq \hat{\boldsymbol{R}}$, we therefore have

$$\hat{\boldsymbol{R}} \neq \hat{\boldsymbol{R}}', \qquad \phi(\hat{\boldsymbol{R}}) = \phi(\hat{\boldsymbol{R}}'). \tag{43}$$

Thus, the spectral representation is non-injective. □

Theorem 3 shows that the communicated spectrum does not uniquely determine the underlying representation matrix. Spectral abstraction retains only normalized singular values while discarding singular vectors and absolute spectral magnitude, thereby removing substantial representation-specific information.

Overall, Theorems 2 and 3 characterize the information reduction induced by spectral abstraction rather than providing a formal privacy guarantee. FedSpec communicates compact structural information while avoiding the direct transmission of local model parameters or feature representations.

D MORE EXPERIMENTS

All additional experiments in this section use a client participation rate of $\rho = 10\%$ by default.

D.1 SENSITIVITY AND ROBUSTNESS ANALYSIS

Impact of Data Heterogeneity. Figure 4 evaluates FedSpec under different pathological and practical non-IID settings. Under pathological partitions, the performance generally decreases as the number of classes per client increases, since local data distributions become less complementary and the benefit of cross-client collaboration is reduced. Under practical non-IID partitions, the performance varies moderately as the Dirichlet coefficient α changes. Nevertheless, FedSpec maintains competitive performance across a wide range of pathological and practical non-IID settings, demonstrating its robustness to different degrees and forms of data heterogeneity.

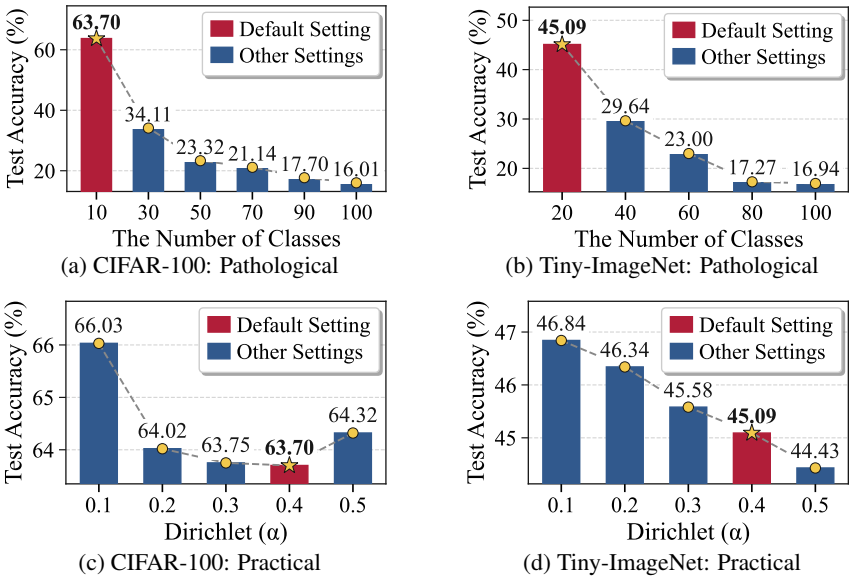


Figure 4: Impact of data heterogeneity on FedSpec. The first row evaluates pathological non-IID partitions by varying the number of classes per client, while the second row evaluates practical non-IID partitions generated by the Dirichlet distribution. The red bars denote the default settings.

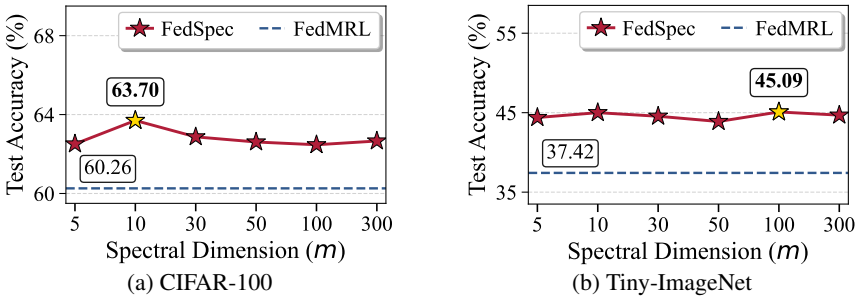
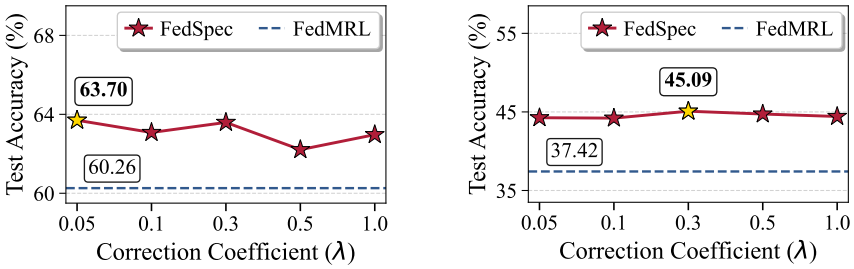


Figure 5: Impact of the spectral dimension m. The dashed lines denote the performance of FedMRL, while the yellow markers indicate the default settings.

Impact of Spectral Dimension. Figure 5 evaluates the impact of the spectral dimension m. FedSpec consistently outperforms FedMRL across all evaluated settings and maintains relatively stable performance over a wide range of spectral dimensions. This indicates that the effectiveness of spectral collaboration does not rely on a narrowly tuned value of m, and that the dominant structural characteristics of local representation spaces can be captured under different levels of spectral granularity. Notably, m is not determined by the number of classes: CIFAR-100 and Tiny-ImageNet contain 100 and 200 classes, while their default spectral dimensions are 10 and 100, respectively.

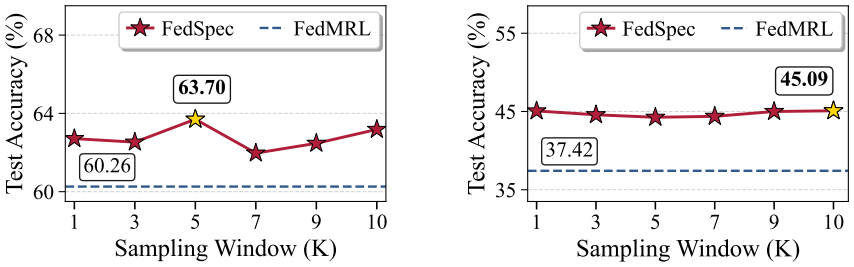
Impact of Spectral Correction Coefficient. Figure 6 evaluates the impact of the spectral correction coefficient λ. FedSpec consistently outperforms FedMRL across the evaluated range of λ, while its performance remains relatively stable under different correction strengths. These results demonstrate the effectiveness of spectral correction and show that FedSpec is not sensitive to the specific choice of λ.



(a) CIFAR-100 (b) Tiny-ImageNet

Figure 6: Impact of the spectral correction coefficient λ. The dashed lines denote FedMRL, while the yellow markers indicate the default settings.

Impact of Sampling Window. Figure 7 evaluates the impact of the sampling window K. FedSpec consistently outperforms FedMRL across different window sizes and maintains relatively stable performance on both datasets. This indicates that the structural information extracted by SRS remains reliable under different temporal sampling ranges, demonstrating the robustness of FedSpec to the specific choice of K.



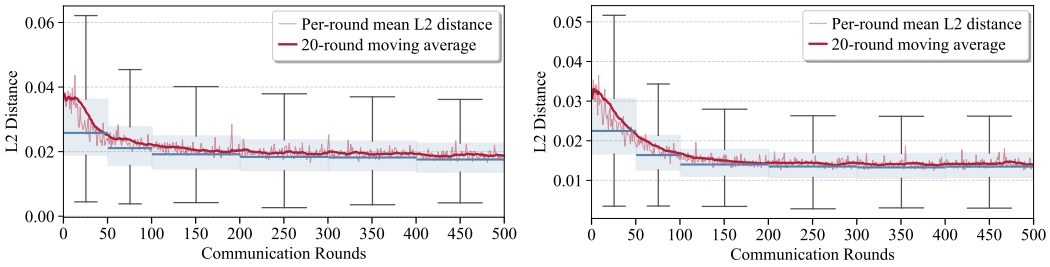
(a) CIFAR-100 (b) Tiny-ImageNet

Figure 7: Impact of the sampling window K. The dashed lines denote the performance of FedMRL, while the yellow markers indicate the default settings.

D.2 Understanding the Effectiveness of FedSpec

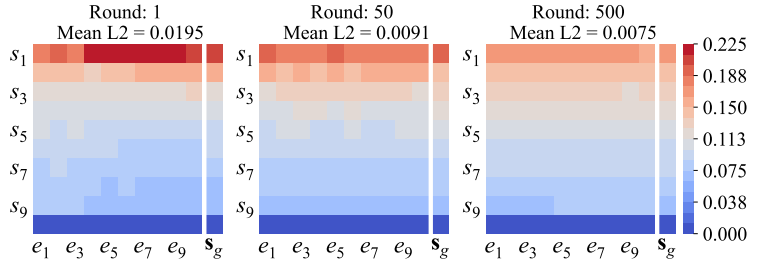
Evolution of Spectral Structures. Figure 8 illustrates the evolution of the discrepancy between local and global spectral structures throughout federated training. On both datasets, the average L_2 distance exhibits an overall decreasing trend as communication proceeds, indicating that local spectral structures progressively approach the aggregated global spectrum. Moreover, the spectral discrepancies become increasingly stable in later training stages, as reflected by the reduced fluctuations and more concentrated distance distributions. These observations show that the aggregated spectrum provides a meaningful shared structural reference that can be progressively incorporated into heterogeneous local representation learning through spectral correction.

Visualization of Spectral Structures. Figure 9 further visualizes the local and global spectral structures at the early (*Round 1*), middle (*Round 50*), and late (*Round 500*) stages of federated training. For CIFAR-100, each spectral vector consists of $m = 10$ normalized spectral components (s_1–s_{10}). For Tiny-ImageNet, $m = 100$ is adopted, while only the first ten components are visualized since the dominant spectral information is concentrated in the leading dimensions. Columns e_1–e_{10} denote the average local spectra across participating clients at different local epochs, while $\boldsymbol{s}_g$ denotes the aggregated global spectral structure.

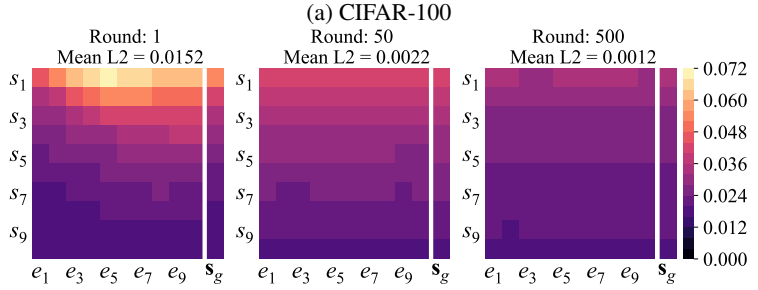


(a) CIFAR-100 (b) Tiny-ImageNet

Figure 8: Evolution of the spectral discrepancy between local and global spectral structures during training. The curves report the mean L2 distance between local spectra and the global spectrum averaged across all participating clients at each communication round. The box plots summarize the client-wise distance distributions at selected communication rounds.



(a) CIFAR-100



(b) Tiny-ImageNet

Figure 9: Visualization of local and global spectral structures at different training stages. Each heatmap shows normalized spectral vectors. For CIFAR-100, the spectral structure consists of $m = 10$ spectral elements (s_1–s_{10}). For Tiny-ImageNet, the spectral structure consists of $m = 100$ spectral elements, while only the first ten components are visualized since the dominant spectral information is concentrated in the leading dimensions. Columns e_1–e_{10} denote the average local spectra across all participating clients at different local epochs, while $\boldsymbol{s}_g$ denotes the aggregated global spectral structure. The color intensity represents normalized singular values. As training progresses from the early stage (*Round 1*) to the middle stage (*Round 50*) and the final stage (*Round 500*), local spectral structures become increasingly consistent with the global spectrum, as evidenced by the decreasing mean L2 distance.

Clear spectral decay patterns can be observed throughout training, with larger singular values concentrated in the leading components, reflecting the unequal importance distribution of principal representation directions. Moreover, within each communication round, later local epochs exhibit higher consistency with the global spectrum than earlier epochs.

Together with the continuous trend in Figure 8, these results support the role of the aggregated global spectrum as a stable structural prior. Importantly, FedSpec does not require the global spectrum to reconstruct a common representation space. Instead, it summarizes the shared relative importance pattern of principal representation directions and guides local representation optimization through spectral correction, thereby facilitating effective collaboration across heterogeneous models.

D.3 Ablation Study

Settings. To better understand the contribution of each design component in FedSpec, we construct four ablation variants by modifying one component at a time while keeping the others unchanged. Specifically, "QR" replaces the proposed SVD based spectral abstraction with QR decomposition to evaluate the effectiveness of singular value based spectral structures. "SRS+SRS" adopts Stable Representation Sampling (SRS) for both spectral aggregation and spectral correction, while "GRA+GRA" adopts Group Representation Averaging (GRA) at both stages, allowing us to examine whether different representation construction strategies are necessary for global aggregation and local correction. "One-Stage" replaces the proposed two-stage optimization with a one-stage formulation to evaluate the benefit of separating task optimization from spectral correction.

Table 3: Ablation study of FedSpec on CIFAR-100 and Tiny-ImageNet. Each variant modifies one component while keeping the others unchanged. ✓ and ✗ denote retained and modified components, respectively. The best results are highlighted in bold.

Variant	SVD	SRS	GRA	Two-stage	CIFAR.	Tiny.
QR	✗	✓	✓	✓	63.07	44.38
GRA+GRA	✓	✗	✓	✓	62.64	43.72
SRS+SRS	✓	✓	✗	✓	62.49	44.79
One-Stage	✓	✓	✓	✗	62.32	44.92
FedSpec	✓	✓	✓	✓	**63.70**	**45.09**

Note: "SVD" denotes Singular Value Decomposition, "SRS" denotes Stable Representation Sampling, "GRA" denotes Group Representation Averaging, "Two-stage" denotes the proposed two-stage optimization strategy.

Results. As shown in Table 3, all variants consistently underperform the complete FedSpec on both datasets, demonstrating the complementary contribution of the proposed components. In particular, replacing SVD with QR decomposition leads to consistent performance degradation while leaving the remaining framework unchanged. This result indicates that the relative importance distribution characterized by singular values provides useful structural information for heterogeneous collaboration, supporting the effectiveness of singular-value-based spectral abstraction.

Both "SRS+SRS" and "GRA+GRA" perform worse than FedSpec, confirming that SRS and GRA play different but complementary roles. SRS provides a stable estimate of local representation structure for global aggregation, whereas GRA characterizes the current local representation structure for spectral correction. The performance degradation of "One-Stage" further demonstrates the benefit of separating task optimization from structural correction, allowing the global spectral structure to serve as a structural prior during local learning.