SPECTRAL DECOMPOSITION OF PROMPT KNOWLEDGE FOR FEDERATED CONTINUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated continual learning (FCL) requires clients to continuously learn new tasks while leveraging knowledge distributed across clients and historical tasks. Prompts provide a compact means of representing and aggregating such knowledge without repeatedly updating the entire model. However, existing methods typically select and aggregate historical prompts as a whole, overlooking that different structural directions may play distinct roles: some capture knowledge shared across tasks, while others may interfere with current knowledge integration. To address this issue, we propose SDPrompt, which organizes and exploits historical prompt knowledge from a spectral perspective. At the server, SDPrompt extracts a dominant spectral subspace from accumulated global prompts to preserve historical knowledge. At each client, it softly rectifies weak spectral components of historical prompts and extracts a task-conditioned subspace to facilitate relevant knowledge transfer. We further theoretically characterize the spectral concentration of shared knowledge and bound the effect of weak-spectrum rectification under explicit spectral separation conditions. Experiments on ImageNet-R, DomainNet, Cars196, and CIFAR-100 show that SDPrompt outperforms the compared methods in both Average Accuracy (Avg) and Average Incremental Accuracy (AIA), achieving 89.83% Avg on ImageNet-R with a 2.63% improvement. These results demonstrate the effectiveness of exploiting prompt knowledge at the spectral level for knowledge preservation and transfer in FCL.Code is available at `https://anonymous.4open.science/r/paper_-C76A/`.

1 INTRODUCTION

Federated Continual Learning (FCL) enables multiple clients to collaboratively learn from continuously evolving task streams without sharing local data(Li et al., 2024c; Piao et al., 2024). Compared with conventional federated learning, FCL must address not only data heterogeneity across clients but also knowledge forgetting and transfer under evolving tasks(Liang et al., 2024; Qi et al., 2023; Piao et al., 2024). Therefore, models need to adapt to new tasks while preserving and exploiting historical knowledge accumulated across clients and tasks. Recent advances in parameter-efficient learning provide a promising solution. In particular, Prompt Learning freezes the pretrained backbone and optimizes only a small set of learnable parameters, providing a compact carrier for continual knowledge accumulation and interaction in FCL(Wang et al., 2022b;a; Smith et al., 2023).

Following this paradigm, recent studies have explored how historical prompts can facilitate knowledge sharing across clients and tasks. Powder (Piao et al., 2024) leverages prompts for knowledge transfer across clients and temporal tasks, while C^2Prompt (Xu et al., 2025) employs local class distribution compensation and class-aware prompt aggregation to mitigate knowledge discrepancies and aggregation conflicts across clients. Despite different interaction mechanisms, these methods largely operate at the granularity of entire prompts, exploiting historical knowledge through prompt selection and aggregation(Piao et al., 2024; Xu et al., 2025).

However, *a historical prompt is not necessarily an indivisible unit of knowledge*. Different structural directions within a prompt may play distinct roles: some capture knowledge shared across tasks, some are particularly relevant to the current task, while others may interfere with knowledge integration. Treating a historical prompt as a whole inevitably transfers these components together, even when the prompt itself is relevant to the current task. This motivates a finer-grained question:

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.

beyond deciding which historical prompts to exploit, can we distinguish the knowledge structures within them to preserve shared knowledge, transfer relevant knowledge, and suppress potential interference?

We find that the *spectral structure* of historical prompts provides a natural perspective for addressing this question. Knowledge structures repeatedly supported across tasks tend to receive stronger spectral support and concentrate in dominant directions, whereas weakly supported structures exhibit relatively weaker spectral contributions (Skorobogat et al., 2025). Moreover, dominant structures extracted from accumulated historical prompts characterize globally shared knowledge, while those extracted from prompts relevant to the current input capture more task-relevant knowledge. These complementary views allow historical knowledge to be organized beyond the granularity of entire prompts.

Building on this observation, we propose **SDPrompt**, a spectral knowledge modeling approach for FCL. At the server, SDPrompt extracts a *Global Knowledge Subspace* from accumulated prompts to preserve dominant historical structures. At each client, it constructs a *Task-Conditioned Knowledge Subspace* from relevant historical prompts to facilitate knowledge transfer, while *Weak-Spectrum Rectification* softly suppresses selected weak spectral components before knowledge integration. The two subspaces jointly guide the learning of the current prompt, balancing historical knowledge preservation and relevant knowledge transfer. We further theoretically characterize the spectral concentration of shared knowledge under explicit separation conditions and bound the influence of weak-spectrum rectification on shared components.

Experiments on ImageNet-R, DomainNet, Cars196, and CIFAR-100 demonstrate that SDPrompt consistently improves both Average Accuracy (Avg) and Average Incremental Accuracy (AIA) over the compared methods. In particular, on ImageNet-R, SDPrompt achieves **89.83% Avg** and **88.54% AIA**, yielding improvements of **2.63%** and **1.51%**, respectively, over the best-performing compared methods. Further ablation studies demonstrate that modeling prompt knowledge at the spectral level enables more effective historical knowledge preservation and relevant knowledge transfer in FCL.

Our main contributions are summarized as follows:

- We revisit prompt knowledge in federated continual learning from a *spectral perspective*. Unlike existing methods that primarily select and aggregate entire prompts, we further explore the distinct roles of internal structural directions, enabling historical knowledge to be organized and exploited at a finer spectral granularity.
- We propose **SDPrompt**, which performs complementary spectral knowledge modeling for historical knowledge preservation and relevant knowledge transfer. SDPrompt captures globally shared structures through a global knowledge subspace, extracts task-relevant structures through a task-conditioned knowledge subspace, and mitigates potential interference via weak-spectrum rectification.
- Extensive experiments on ImageNet-R, DomainNet, Cars196, and CIFAR-100 demonstrate consistent improvements in both Average Accuracy (Avg) and Average Incremental Accuracy (AIA). Ablation studies and further analyses validate the effectiveness and complementarity of the proposed components.

2 Related Work

2.1 Federated Learning

Federated Learning (FL) enables multiple clients to collaboratively train a global model without directly sharing local data (Liao et al., 2024; Chen et al., 2024; Allouah et al., 2024; Qi et al., 2025). A central challenge in FL is data heterogeneity, where non-IID client distributions can lead to inconsistent local updates and degraded global performance (Zhang et al., 2024; McLaughlin & Su, 2024; M Ghari & Shen, 2024; Li et al., 2024a). Existing methods address this issue through local regularization (Li et al., 2020; Karimireddy et al., 2020; Mendieta et al., 2022; Acar et al., 2021; Gao et al., 2022), aggregation optimization (McMahan et al., 2017; Hsu et al., 2019; Wang et al., 2020; Li et al., 2021), or synthetic data generation (Zhu et al., 2021; Guo et al., 2023; Yoon et al., 2021b; Hu et al., 2022; Xiong et al., 2023). However, conventional FL typically assumes relatively stationary data and is not designed for continuously evolving client tasks.

2.2 Federated Continual Learning

In Federated Continual Learning (FCL), clients continuously learn private task streams, while the server integrates knowledge across clients and learning stages (Li et al., 2024c). Early methods mainly rely on replay or generative models to mitigate forgetting, often introducing additional computational overhead and privacy concerns (Liang et al., 2024; Babakniya et al., 2023; Li et al., 2024b; Qi et al., 2023; Zhang et al., 2023; Odena et al., 2017). Recent studies increasingly focus on parameter-efficient FCL. Powder (Piao et al., 2024) leverages prompts for knowledge transfer across clients and tasks, while PILoRA (Guo et al., 2024) and LoRM (Salami et al., 2025) employ low-rank adaptation for continual knowledge sharing. C^{2}Prompt (Xu et al., 2025) further introduces class-aware prompt aggregation to alleviate cross-client knowledge discrepancies and aggregation conflicts. Recent methods also explore adapters, multi-expert interaction, and self-distillation to facilitate knowledge interaction and transfer across clients and tasks (Yi et al., 2025).wledge interaction through class-aware prompt aggregation.

3 Preliminaries

Problem Setting and Baseline. Federated Continual Learning (FCL) involves K clients coordinated by a central server. Client k learns tasks $\{\mathcal{T}_k^t\}_{t=1}^T$ with datasets $\mathcal{D}_k^t$ and model parameters θ_k^t. We adopt C^{2}Prompt (Xu et al., 2025) as our baseline. Each task uses prompts $P_t \in \mathbb{R}^{N\times L_p\times D}$, where N, L_p, and D denote the prompt count, length, and feature dimension. The server maintains a global prompt pool $P \in \mathbb{R}^{MN\times L_p\times D}$ covering M tasks. For input x, Class-aware Prompt Aggregation (CPA) combines current-task prompts with prompts from the top-k relevant historical tasks:

$$p_x = \sum_{i=1}^{M_g} \alpha_i(x)[P^g]_i, \qquad \alpha_i(x) = \gamma(q(x) \odot [A^g]_i, [K^g]_i)\,, \tag{1}$$

where P^g contains M_g candidate prompts, $q(x)$ is the query, K^g and A^g are prompt keys and attention vectors, γ measures similarity, and $\odot$ denotes element-wise multiplication. The resulting $p_x \in \mathbb{R}^{L_p\times D}$ provides input-specific prompt knowledge.

Local Optimization. We retain Local Class Distribution Compensation (LCDC) and Non-Target Distillation (NTD). LCDC constructs class-dependent compensation prompts from local and global class statistics to mitigate client distribution shifts. NTD preserves non-target class knowledge by aligning the current model with the model saved at the beginning of each communication round:

$$\mathcal{L}_{\mathrm{NTD}} = -\sum_{\substack{j=1\\ j\neq y}}^{C_{\mathrm{cls}}} [\hat{y}^{\mathrm{tr}}]_j \log \frac{[\hat{y}^{\mathrm{cu}}]_j}{[\hat{y}^{\mathrm{tr}}]_j}, \tag{2}$$

where $\hat{y}^{\mathrm{cu}}$ and $\hat{y}^{\mathrm{tr}}$ denote the current and saved model predictions, y is the ground-truth label, and C_{cls} is the number of classes.

4 Method

We present SDPrompt in Fig. 1. The server extracts the dominant global knowledge subspace V_{global} from a matrix Q_G of accumulated global prompts. For each input, the client forms Q_T from relevant historical prompts, using its weak spectral directions V_R to softly rectify these prompts before fusion and its dominant directions to construct V_{task}. The two subspaces jointly constrain current-task prompt learning to preserve shared knowledge and transfer input-relevant historical knowledge.

4.1 Spectral Knowledge Modeling

Given a prompt knowledge matrix $Q \in \mathbb{R}^{L\times D}$, we apply Singular Value Decomposition (SVD):

$$Q = U\,\mathrm{diag}(\sigma_1, \ldots, \sigma_d)[v_1, \ldots, v_d]^\top, \tag{3}$$

where $d = \min(L, D)$, $\sigma_1 \geq \cdots \geq \sigma_d \geq 0$, and v_i is the right singular vector corresponding to σ_i.

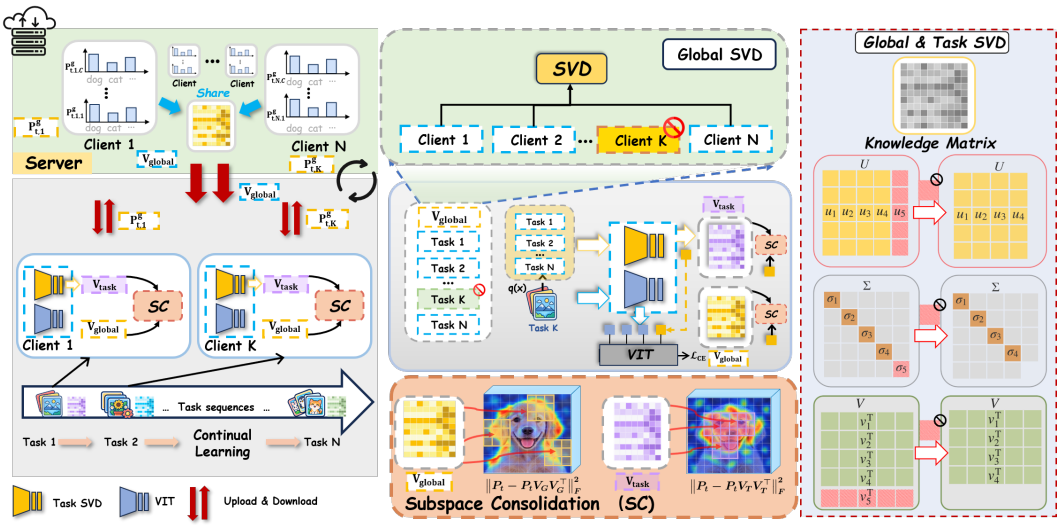

Figure 1: Overview of SDPrompt. The server extracts a global knowledge subspace from accumulated prompts, while each client constructs a task-conditioned subspace using relevant historical tasks. Spectral knowledge calibration suppresses weak structural components in historical prompts, and the two subspaces jointly guide current-task prompt learning.

We define the normalized and cumulative spectral energies as $e_i = \sigma_i^2 / \sum_{m=1}^{d} \sigma_m^2$ and $E_i = \sum_{\ell=1}^{i} e_\ell$, respectively. High-energy directions capture dominant structures, while low-energy directions represent weaker components. These quantities guide spectral rectification and knowledge subspace construction.

4.2 Global Knowledge Subspace

As new tasks arrive, the server retains their global prompts in its prompt pool. To characterize the dominant structure of the accumulated historical knowledge, we assemble these prompts task by task into a global knowledge matrix Q_G and perform spectral decomposition. We select the smallest number of leading spectral directions whose cumulative energy reaches τ_G, yielding

$$V_{\text{global}} = [v_{G,1}, \ldots, v_{G,r_G}], \qquad r_G = \min\{r \mid E_{G,r} \geq \tau_G\}. \tag{4}$$

Here, $E_{G,r}$ denotes the cumulative spectral energy of the first r directions of Q_G. The resulting V_{global} captures the dominant structure of historical prompts and provides a stable reference for client learning. The server distributes this subspace to participating clients to constrain current-task prompt updates during local optimization.

4.3 Spectral Knowledge Rectification

The global knowledge subspace provides an overall reference for historical knowledge. To account for input relevance, we select the top-k historical tasks $\mathcal{T}_x = \{t_1, \ldots, t_k\}$ using prompt relevance scores and assemble their relevance-weighted prompts into Q_T. To suppress components with limited cross-task support, we select directions with cumulative spectral energy in $(\tau_s, \tau_e]$ and individual energy below δ, forming

$$V_R = [v_{T,\ell}]_{\tau_s < E_{T,\ell} \leq \tau_e,\ e_{T,\ell} < \delta}. \tag{5}$$

Before knowledge fusion, we rectify the selected historical prompts P_h:

$$\widetilde{P}_h = P_h - \eta P_h V_R V_R^\top, \tag{6}$$

where η controls attenuation within the weak spectral subspace. The current-task prompt remains unchanged, while the rectified historical prompts are used for subsequent knowledge fusion.

4.4 DUAL-SUBSPACE KNOWLEDGE CONSOLIDATION

The global knowledge subspace V_{global} provides a structural reference drawn from all historical tasks. In parallel, the client constructs a task-conditioned knowledge subspace V_{task} from historical knowledge relevant to the current input, providing a more targeted structural constraint. Together, these subspaces characterize historical knowledge at global and task-conditioned levels and guide the learning of the current-task prompt.

Task-Conditioned Knowledge Subspace. Using the task-conditioned knowledge matrix Q_T defined above, we retain the smallest number of leading spectral directions whose cumulative energy reaches a threshold τ_T. These directions form the task-conditioned knowledge subspace

$$V_{\text{task}} = [v_{T,1}, \ldots, v_{T,r_T}], \qquad r_T = \min\{r \mid E_{T,r} \geq \tau_T\}. \tag{7}$$

Here, $E_{T,r}$ denotes the cumulative spectral energy of the first r directions of Q_T. For any direction $v_{T,\ell}$ selected for the weak spectral subspace V_R, its index satisfies $\ell > r_T$. Thus, V_{task} retains the dominant structure of the task-conditioned knowledge matrix, providing a reference from historical tasks that are relevant to the current input.

Subspace Consolidation. The two subspaces provide complementary structural constraints: V_{global} reflects knowledge accumulated across historical tasks, whereas V_{task} captures dominant knowledge from tasks relevant to the current input. We use both to constrain the current-task prompt P_t, with the aim of preserving historical knowledge while transferring relevant knowledge to the current task. The dual-subspace constraint is defined as

$$\mathcal{L}_{\text{sc}} = \left\|P_t - P_t V_{\text{global}} V_{\text{global}}^\top\right\|_F^2 + \left\|P_t - P_t V_{\text{task}} V_{\text{task}}^\top\right\|_F^2. \tag{8}$$

The first term penalizes components of P_t outside the global knowledge subspace, encouraging alignment with the dominant structure of accumulated historical knowledge. The second term encourages alignment with the dominant structure of input-relevant historical knowledge.

The overall training objective combines the classification loss, the non-target knowledge distillation loss, and the dual-subspace constraint:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \beta \mathcal{L}_{\text{NTD}} + \mathcal{L}_{\text{sc}}, \tag{9}$$

where $\mathcal{L}_{\text{CE}}$ is the classification cross-entropy loss, $\mathcal{L}_{\text{NTD}}$ is the non-target knowledge distillation loss, and β controls the weight of the distillation term.

4.5 THEORETICAL ANALYSIS

Following prior work on model merging (Skorobogat et al., 2025), we analyze spectral concentration and weak-spectrum rectification. Detailed assumptions and proofs are provided in Appendix A.

Definition 4.1 (Prompt Knowledge Decomposition)**.** For task i, let $P_i = P_i^{\text{common}} + P_i^{\text{specific}}$, where the two components represent shared and task-specific knowledge. Define

$$Q_G = \text{Concat}(P_1, \ldots, P_M), \qquad Q_T = \text{Concat}(\alpha_1 P_{t_1}, \ldots, \alpha_k P_{t_k}),$$

where α_i weights historical task t_i for the current input.

Proposition 4.2 (Global Shared Knowledge Concentration)**.** *Assume that common components share an r-dimensional right subspace, while specific subspaces are mutually orthogonal and orthogonal to it. With per-task common spectral energy bounded below by $m_G > 0$ and specific spectral energy bounded above by M_G,*

$$\frac{(\sigma_{G,1}^{\text{specific}})^2}{(\sigma_{G,r}^{\text{common}})^2} \leq \frac{M_G}{M m_G}. \tag{10}$$

For fixed m_G and M_G, this bound is $O(1/M)$.

Proposition 4.2 supports extracting V_{global} from dominant historical knowledge directions.

Definition 4.3 (Effective Task Support)**.** For relevance weights with at least one nonzero entry, define

$$k_{\text{eff}} = \frac{\sum_{i=1}^k \alpha_i^2}{\max_{1 \leq i \leq k} \alpha_i^2}, \qquad 1 \leq k_{\text{eff}} \leq k. \tag{11}$$

Proposition 4.4 (Task-Conditioned Knowledge Concentration)**.** *Under the same spectral separation assumptions, with corresponding per-task bounds $m_T > 0$ and M_T,*

$$\frac{(\sigma_{T,1}^{\text{specific}})^2}{(\sigma_{T,r}^{\text{common}})^2} \leq \frac{M_T}{m_T} \frac{1}{k_{\text{eff}}}. \tag{12}$$

Proposition 4.4 links the construction of V_{task} to effective task support rather than task count alone.

Remark 4.5 (Spectral Stability). For non-orthogonal representations, the dominant spectral structure remains stable under bounded Gram-matrix perturbations that preserve the common–specific gap; see Appendix A.

Definition 4.6 (Weak-Spectrum Rectification)**.** Let V_R have orthonormal columns orthogonal to V_{task}. For a historical prompt P_h, define

$$\widetilde{P}_h = P_h - \eta P_h V_R V_R^\top, \qquad 0 \leq \eta \leq 1,$$

and the shared knowledge residual

$$\rho_T = \left\| P_h^{\text{common}} (I - V_{\text{task}} V_{\text{task}}^\top) \right\|_F^2. \tag{13}$$

Proposition 4.7 (Bounded Weak-Spectrum Rectification)**.** *The rectified prompt satisfies*

$$\|\widetilde{P}_h V_R\|_F^2 = (1-\eta)^2 \|P_h V_R\|_F^2. \tag{14}$$

Applying the same rectification to its common component gives

$$\|\widetilde{P}_h^{\text{common}} - P_h^{\text{common}}\|_F^2 \leq \eta^2 \rho_T. \tag{15}$$

Proposition 4.7 justifies soft attenuation: weak directions may contain shared knowledge, but its disturbance is bounded by the residual ρ_T.

5 Experiments

5.1 Experimental Setups

We evaluate SDPrompt on four widely used FCL benchmarks: ImageNet-R (Hendrycks et al., 2021), DomainNet (Peng et al., 2019), Cars196 (Krause et al., 2013), and CIFAR-100 (Krizhevsky et al., 2009). We report Average Accuracy (Avg) and Average Incremental Accuracy (AIA) to evaluate final knowledge retention and performance throughout continual learning. In addition to the FCL benchmarks, we further evaluate SDPrompt under the Federated Class-Incremental Learning (FCIL) setting on ImageNet-R, where disjoint class-incremental tasks and different levels of client data heterogeneity are considered. Final Average Accuracy (FAA) is reported for this setting. Detailed benchmark configurations and metric definitions are provided in the appendix.

Compared Methods. For the FCL benchmarks, we compare SDPrompt with full fine-tuning methods, including FedWEIT (Yoon et al., 2021a), CFeD (Ma et al., 2022), GLFC (Dong et al., 2022), and FedSpace (Shenaj et al., 2023), as well as parameter-efficient methods including Fed-L2P (Wang et al., 2022b), Fed-Dual (Wang et al., 2022a), Fed-CODA (Smith et al., 2023), Fed-CPrompt (Bagwe et al., 2023), Powder (Piao et al., 2024), PILoRA (Guo et al., 2024), LoRM (Salami et al., 2025), C^{2}Prompt (Xu et al., 2025), and EWC-LoRA (Zheng et al., 2026). For the additional FCIL evaluation, we further include EWC (Kirkpatrick et al., 2017), LwF (Li & Hoiem, 2018), Fisher AVG (Matena & Raffel, 2022), RegMean (Jin et al., 2023), CCVR (Luo et al., 2021), L2P (Wang et al., 2022b), CODA-P (Smith et al., 2023), FedProto (Tan et al., 2022), and TARGET (Zhang et al., 2023). All experiments use official implementations with a ViT-B/16 backbone pretrained on ImageNet-21K.

Implementation Details. For the FCL benchmarks, following prior work (Piao et al., 2024), we set the discriminative prompt parameters N, L_p, and D to 10, 8, and 768, respectively. For Dual-Subspace Knowledge Consolidation, both τ_G and τ_T are set to 0.7. We use Adam with a learning rate of 0.005, five clients, and three communication rounds per task. For the additional FCIL experiment on ImageNet-R, we follow the experimental protocol adopted by LoRM (Salami et al., 2025).

The task construction, client data partitioning, communication schedule, local training configuration, data preprocessing, and evaluation protocol are detailed in Appendix C. All experiments are conducted on a single NVIDIA RTX 4090 GPU.

Table 1: Comparison with existing federated continual learning methods on ImageNet-R, DomainNet, Cars196, and CIFAR-100.

Method	Venue	ImageNet-R		DomainNet		Cars196		CIFAR-100	
		Avg↑	AIA↑	Avg↑	AIA↑	Avg↑	AIA↑	Avg↑	AIA↑
FedWEIT	*ICML2021*	71.10	74.30	67.84	69.63	65.69	65.08	95.17	95.61
CFeD	*IJCAI2022*	47.93	59.79	42.85	60.19	66.26	65.87	73.87	79.06
GLFC	*CVPR2022*	72.96	75.21	69.75	70.34	<u>71.39</u>	69.50	95.35	95.92
FedSpace	*CVPR2023*	72.27	73.36	68.98	70.71	66.45	<u>70.08</u>	94.17	94.87
Fed-L2P	*CVPR2022*	77.88	75.03	70.98	72.36	67.17	65.93	95.65	95.68
Fed-Dual	*ECCV2022*	76.85	74.91	71.90	72.15	67.28	65.54	95.35	95.08
Fed-CODA	*CVPR2023*	79.65	75.14	72.47	72.84	64.28	63.34	82.05	55.71
Fed-CPrompt	*ICML2023*	76.75	72.59	71.28	69.92	67.50	66.46	94.22	94.04
Powder	*ICML2024*	84.69	84.08	75.98	77.28	65.03	66.77	95.78	95.83
PILoRA	*ECCV2024*	45.43	48.72	31.22	40.76	45.63	44.60	76.21	82.31
LoRM	*ICLR2025*	58.00	67.78	23.18	28.49	56.20	51.90	77.42	80.11
C^2Prompt	*NeurIPS2025*	<u>87.20</u>	85.93	<u>78.88</u>	<u>77.55</u>	69.53	69.86	97.32	96.78
EWC-LoRA	*ICLR2026*	87.14	<u>87.03</u>	76.09	76.69	69.61	69.79	<u>97.39</u>	<u>97.84</u>
SDPrompt	*Ours*	**89.83**	**88.54**	**79.47**	**78.62**	**72.60**	**72.28**	**97.85**	**98.03**

5.2 Comparison Results

We follow the experimental settings of prior work (Piao et al., 2024). Table 1 summarizes the comparison results on ImageNet-R (Hendrycks et al., 2021), DomainNet (Peng et al., 2019), Cars196 (Krause et al., 2013), and CIFAR-100 (Krizhevsky et al., 2009). We report average accuracy (Avg) and average incremental accuracy (AIA). Avg measures overall performance on all learned tasks at the end of continual learning, while AIA measures average performance across learning stages. The best and second-best results are marked in bold and underlined, respectively.

Avg Comparison. SDPrompt outperforms existing state-of-the-art methods in Avg, with gains of 2.63 and 1.21 percentage points on ImageNet-R and Cars196, respectively, and improvements on the other datasets. On DomainNet, it achieves 79.47%, compared with 78.88% for C^2Prompt. On CIFAR-100, it achieves 97.85%, exceeding EWC-LoRA's 97.39% by 0.46 percentage points. These results indicate more effective long-term knowledge consolidation. Spectral knowledge calibration suppresses weak structural components in historical prompts, while the global knowledge subspace preserves dominant structures shared across tasks. The task-conditioned knowledge subspace retrieves and transfers relevant task knowledge based on the current input.

AIA Comparison. SDPrompt improves AIA over the strongest existing methods by 1.51 and 2.20 percentage points on ImageNet-R and Cars196, respectively, and also achieves the best results on DomainNet and CIFAR-100. On DomainNet, it achieves 78.62%, exceeding C^2Prompt's 77.55% by 1.07 percentage points. On CIFAR-100, it achieves 98.03%, compared with 97.84% for EWC-LoRA. These results suggest more stable performance throughout continual learning. By filtering and consolidating historical knowledge, SDPrompt preserves shared knowledge, suppresses weak structural components, and transfers relevant experience to new tasks, improving average performance across learning stages.

5.3 Robustness to Client Data Heterogeneity

We further evaluate SDPrompt under disjoint-class FCIL on ImageNet-R using three Dirichlet partitions, where smaller β indicates greater client data heterogeneity. As shown in Table 2, SDPrompt achieves the highest FAA at $\beta = 0.5$, 0.1, and 0.05, reaching 76.07%, 70.56%, and 71.02%, respectively, and outperforming C^2Prompt by 0.84, 0.82, and 2.28 percentage points. Although stronger

heterogeneity generally reduces performance, SDPrompt remains the best across all partitions, with the largest gain at $\beta = 0.05$. These results indicate improved robustness, consistent with weak-spectrum rectification suppressing weakly supported historical structures and the dual subspaces preserving shared knowledge while transferring input-relevant knowledge. Together with the preceding results, they demonstrate applicability to both asynchronous FCL and disjoint-class FCIL.

Table 2: FAA (%) comparison on ImageNet-R under varying levels of client data heterogeneity in the FCIL setting.

Method	Venue	$\beta = 0.5$	$\beta = 0.1$	$\beta = 0.05$
EWC	*NAS2017*	58.93	48.15	43.68
LwF	*PAMI2018*	54.03	41.02	46.07
Fisher AVG	*NeurIPS2022*	58.68	50.82	47.33
RegMean	*ICLR2023*	61.18	57.00	55.80
CCVR	*NeurIPS2021*	70.00	62.60	60.38
L2P	*CVPR2022*	42.08	23.85	16.98
CODA-P	*CVPR2023*	61.18	36.73	25.82
FedProto	*AAAI2022*	58.52	47.30	52.93
TARGET	*ICCV2023*	54.65	45.83	41.32
PILoRA	*ECCV2024*	53.67	51.62	49.37
Powder	*ICML2024*	74.62	67.14	62.26
LoRM	*ICLR2025*	72.48	63.83	66.45
C^2Prompt	*NeurIPS2025*	75.23	69.74	68.74
SDPrompt	*Ours*	**76.07**	**70.56**	**71.02**

5.4 Ablation Study and Supplementary Analysis

Table 3: Ablation comparison on ImageNet-R.

Method	D	G	T	ImageNet-R			
				Avg (	ΔAvg↑	AIA (	ΔAIA↑
Baseline	◯	◯	◯	87.20	–	85.93	–
+T	◯	◯	✓	87.96	+0.76	86.26	+0.33
+G	◯	✓	◯	88.05	+0.85	86.34	+0.41
+D	✓	◯	◯	88.07	+0.87	86.82	+0.89
DSKC	◯	✓	✓	88.41	+1.21	87.14	+1.21
Ours	✓	✓	✓	**89.83**	**+2.63**	**88.54**	**+2.61**

Effect of Model Components. We conduct component ablations on ImageNet-R using C^2Prompt as the baseline. Table 3 evaluates the global knowledge subspace (G), task-conditioned knowledge subspace (T), spectral knowledge calibration (D), G+T, and the full configuration. Adding T, G, and D individually improves Avg by 0.76, 0.85, and 0.87 percentage points and AIA by 0.33, 0.41, and 0.89 percentage points, respectively. Combining G and T improves both metrics by 1.21 percentage points. The full configuration achieves 89.83 Avg and 88.54 AIA, exceeding the baseline by 2.63 and 2.61 percentage points, respectively. Each component contributes positively, and combining them further improves overall performance.

Stage-wise Learning Dynamics. Figure 2 (left) compares the training losses of C^2Prompt and SDPrompt over three communication rounds. Unlike C^2Prompt, SDPrompt shows no pronounced loss increase at aggregation and maintains lower loss throughout most subsequent training. This trend is consistent with its design: spectral knowledge calibration suppresses weak structural components in historical prompts, while the global and task-conditioned knowledge subspaces preserve shared knowledge and incorporate input-relevant historical knowledge, respectively, jointly constraining prompt updates.

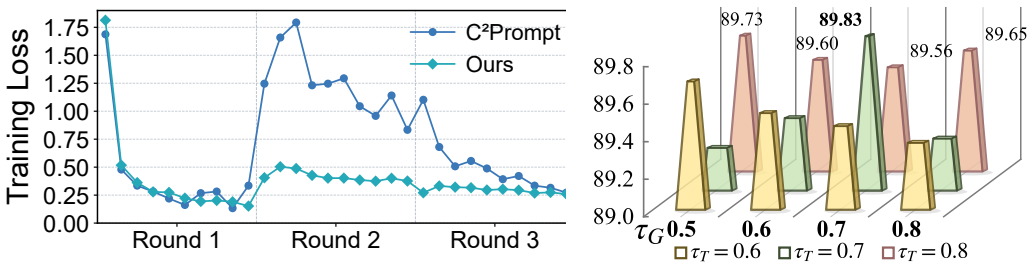

Figure 2: Left: Training loss of C^{2}Prompt and SDPrompt over three communication rounds on ImageNet-R. Right: Avg under different combinations of global and task-conditioned SVD energy thresholds on ImageNet-R.

Sensitivity to Subspace Thresholds. We analyze sensitivity to the SVD energy thresholds of both knowledge subspaces. These thresholds control the cumulative spectral energy retained in the global and task-conditioned subspaces. For a fixed knowledge matrix, increasing the threshold can include additional leading spectral directions. Keeping all other ImageNet-R settings fixed, we evaluate 12 combinations of $\tau_G \in \{0.50, 0.60, 0.70, 0.80\}$ and $\tau_T \in \{0.60, 0.70, 0.80\}$. As shown in Figure 2 (right), Avg ranges from 89.23 to 89.83, varying by only 0.60 percentage points, indicating robustness to both thresholds. Thus, none of the evaluated combinations causes a substantial accuracy drop relative to the best configuration. The highest Avg of 89.83 is achieved with $\tau_G = \tau_T = 0.70$, which we adopt for subsequent ImageNet-R experiments.

Table 4: Relationship between V_{global} and V_{task} on ImageNet-R.

Stage	θ	$\cos^2\theta$	$C_{T\to G}$	$C_{G\to T}$	r_G	r_T
1	26.60°	0.733	0.733	0.518	12.33	9.00
2	23.54°	0.779	0.779	0.496	12.67	8.33
3	26.95°	0.713	0.633	**0.641**	8.33	8.33
4	**16.87°**	**0.872**	**0.872**	0.453	10.33	5.67
5	37.01°	0.605	0.605	0.382	8.33	4.67
Mean	26.19°	0.740	0.724	0.498	10.40	7.20

Subspace Relationship Analysis. We examine the relationship between the global knowledge subspace V_{global} and the task-conditioned knowledge subspace V_{task} using stage-wise mean principal angles, symmetric overlap, directional containment, and effective ranks. As shown in Table 4, their relationship varies non-monotonically across stages. Stage 4 has the smallest mean principal angle (16.87°) and highest symmetric overlap (0.872), while $C_{G\to T}$ peaks at 0.641 in Stage 3. Symmetric overlap decreases to 0.605 in Stage 5. Across stages, the mean principal angle and symmetric overlap are 26.19° and 0.740, respectively. Mean directional containment is 0.724 for $C_{T\to G}$ and 0.498 for $C_{G\to T}$, while mean effective ranks are $r_G = 10.40$ and $r_T = 7.20$. These results suggest that the subspaces share knowledge structures while retaining distinct information, consistent with their complementary roles in cross-task knowledge retention and relevant task knowledge transfer.

6 Conclusion

This paper presents SDPrompt, a spectral knowledge modeling framework for federated continual learning. By modeling the internal spectral structure of historical prompts, SDPrompt preserves globally shared knowledge, facilitates task-relevant knowledge transfer, and mitigates interference from weak spectral components. Theoretical analyses characterize the underlying spectral properties, while extensive experiments demonstrate consistent improvements across multiple benchmarks. These results highlight the effectiveness of exploiting prompt knowledge at a finer spectral granularity for federated continual learning.

Reproducibility Statement

To facilitate reproducibility, we describe the SDPrompt framework, spectral knowledge calibration, construction of the global and task-conditioned knowledge subspaces, and the complete training objective in the main text. The assumptions, propositions, and complete proofs are provided in the appendix. Dataset splits, federated continual learning protocols, evaluation metrics, training configurations, and key hyperparameters are detailed in the main text and appendix. We also provide an anonymous code repository at `https://anonymous.4open.science/r/paper_-C76A/`, containing our implementation and experimental configurations.

AI Use Statement

Generative AI tools were used solely for language editing, including improving grammar, clarity, and readability. They were not used to develop research ideas, design the methodology or experiments, analyze data, interpret results, or formulate scientific claims. All AI-assisted edits were reviewed by the authors. The authors take full responsibility for the final content of this work.

Ethics Statement

All experiments use public visual classification benchmarks and involve neither human subjects nor the collection of personally identifiable information or sensitive personal data. We release no new sensitive or surveillance data. We follow the established research uses and evaluation protocols of the datasets. This work is intended for academic research in federated continual learning and visual recognition.

References

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=B7v4QMR6Z9w`.

Youssef Allouah, Abdellah El Mrini, Rachid Guerraoui, Nirupam Gupta, and Rafael Pinot. Fine-tuning personalization in federated learning to mitigate adversarial clients. *Advances in Neural Information Processing Systems*, 37:100816–100844, 2024.

Sara Babakniya, Zalan Fabian, Chaoyang He, Mahdi Soltanolkotabi, and Salman Avestimehr. A data-free approach to mitigate catastrophic forgetting in federated class incremental learning for vision tasks. *Advances in Neural Information Processing Systems*, 36:66408–66425, 2023.

Gaurav Bagwe, Xiaoyong Yuan, Miao Pan, and Lan Zhang. Fed-cprompt: Contrastive prompt for rehearsal-free federated continual learning. *arXiv preprint arXiv:2307.04869*, 2023.

Junbao Chen, Jingfeng Xue, Yong Wang, Zhenyan Liu, and Lu Huang. Classifier clustering and feature alignment for federated learning under distributed concept drift. *Advances in neural information processing systems*, 37:81360–81388, 2024.

Jiahua Dong, Lixu Wang, Zhen Fang, Gan Sun, Shichao Xu, Xiao Wang, and Qi Zhu. Federated class-incremental learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10154–10163. IEEE, 2022.

Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. Feddc: Federated learning with non-iid data via local drift decoupling and correction. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10102–10111. IEEE, 2022.

Haiyang Guo, Fei Zhu, Wenzhuo Liu, Xu-Yao Zhang, and Cheng-Lin Liu. Pilora: Prototype guided incremental lora for federated class-incremental learning. In *European Conference on Computer Vision*, pp. 141–159. Springer, 2024.

Tao Guo, Song Guo, and Junxiao Wang. Pfedprompt: Learning personalized prompt for vision-language models in federated learning. In *Proceedings of the ACM Web Conference 2023*, pp. 1364–1374, 2023.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329. IEEE, 2021.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Shengyuan Hu, Jack Goetz, Kshitiz Malik, Hongyuan Zhan, Zhe Liu, and Yue Liu. Fedsynth: Gradient compression via synthetic data in federated learning. *arXiv preprint arXiv:2204.01273*, 2022.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=FCnohuR6AnM.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pp. 5132–5143. PMLR, 2020.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Hongxia Li, Wei Huang, Jingya Wang, and Ye Shi. Global and local prompts cooperation via optimal transport for federated learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12151–12161. IEEE, 2024a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=6YEQUn0QICG.

Yichen Li, Qunwei Li, Haozhao Wang, Ruixuan Li, Wenliang Zhong, and Guannan Zhang. Towards efficient replay in federated incremental learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12820–12829. IEEE, 2024b.

Yichen Li, Wenchao Xu, Haozhao Wang, Yining Qi, Jingcai Guo, and Ruixuan Li. Personalized federated domain-incremental learning based on adaptive knowledge matching. In *European conference on computer vision*, pp. 127–144. Springer, 2024c.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12):2935–2947, 2018. doi: 10.1109/TPAMI.2017.2773081. URL https://doi.org/10.1109/TPAMI.2017.2773081.

Jinglin Liang, Jin Zhong, Hanlin Gu, Zhongqi Lu, Xingxing Tang, Gang Dai, Shuangping Huang, Lixin Fan, and Qiang Yang. Diffusion-driven data replay: A novel approach to combat forgetting in federated class continual learning. In *European Conference on Computer Vision*, pp. 303–319. Springer, 2024.

Xinting Liao, Weiming Liu, Pengyang Zhou, Fengyuan Yu, Jiahe Xu, Jun Wang, Wenjie Wang, Chaochao Chen, and Xiaolin Zheng. Foogd: Federated collaboration for both out-of-distribution generalization and detection. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 132908–132945. Curran Associates, Inc., 2024. doi: 10.52202/079017-4225. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/efd1e27afcb94addd03b9e14c8d9f78f-Paper-Conference.pdf.

Mi Luo, Fei Chen, Dapeng Hu, Yifan Zhang, Jian Liang, and Jiashi Feng. No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. *Advances in Neural Information Processing Systems*, 34:5972–5984, 2021.

Pouya M Ghari and Yanning Shen. Personalized federated learning with mixture of models for adaptive prediction and model fine-tuning. *Advances in Neural Information Processing Systems*, 37:92155–92183, 2024.

Yuhang Ma, Zhongle Xie, Jue Wang, Ke Chen, and Lidan Shou. Continual federated learning based on knowledge distillation. In Lud De Raedt (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 2182–2188. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/303. URL https://doi.org/10.24963/ijcai.2022/303. Main Track.

Michael S Matena and Colin Raffel. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716, 2022.

Connor J McLaughlin and Lili Su. Personalized federated learning via feature distribution adaptation. *Advances in Neural Information Processing Systems*, 37:77038–77059, 2024.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. Pmlr, 2017.

Matias Mendieta, Taojiannan Yang, Pu Wang, Minwoo Lee, Zhengming Ding, and Chen Chen. Local learning matters: Rethinking data heterogeneity in federated learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8387–8396. IEEE, 2022.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pp. 2642–2651. Pmlr, 2017.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1406–1415. IEEE, 2019.

Hongming Piao, Yichen Wu, Dapeng Wu, and Ying Wei. Federated continual learning via prompt-based dual knowledge transfer. In *Forty-first International Conference on Machine Learning*, 2024.

Daiqing Qi, Handong Zhao, and Sheng Li. Better generative replay for continual federated learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=cRxYWKiTan.

Zhuang Qi, Lei Meng, Zhaochuan Li, Han Hu, and Xiangxu Meng. Cross-silo feature space alignment for federated learning on clients with imbalanced data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19986–19994, 2025.

Riccardo Salami, Pietro Buzzega, Matteo Mosconi, Jacopo Bonato, Luigi Sabetta, and Simone Calderara. Closed-form merging of parameter-efficient modules for federated continual learning. In *International Conference on Learning Representations*, volume 2025, pp. 86612–86637, 2025.

Donald Shenaj, Marco Toldo, Alberto Rigon, and Pietro Zanuttigh. Asynchronous federated continual learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 5055–5063. IEEE, 2023.

Ronald Skorobogat, Karsten Roth, and Mariana-Iuliana Georgescu. Subspace-boosted model merging. *arXiv preprint arXiv:2506.16506*, 2025.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11909–11919. IEEE, 2023.

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.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European conference on computer vision*, pp. 631–648. Springer, 2022a.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 139–149. IEEE, 2022b.

Yuanhao Xiong, Ruochen Wang, Minhao Cheng, Felix Yu, and Cho-Jui Hsieh. Feddm: Iterative distribution matching for communication-efficient federated learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16323–16332. IEEE, 2023.

Kunlun Xu, Yibo Feng, Jiangmeng Li, Yongsheng Qi, and Jiahuan Zhou. Cˆ2prompt: Class-aware client knowledge interaction for federated continual learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 44109–44139. Curran Associates, Inc., 2025. doi: 10.52202/085713-1470. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3e8f5d2924f1bff724588eb0f8f5e62d-Paper-Conference.pdf.

Xiaoyang Yi, Yang Liu, Binhan Yang, and Jian Zhang. Federated continual learning via orchestrating multi-scale expertise. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 22785–22814. Curran Associates, Inc., 2025. doi: 10.52202/085713-0768. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/20de741d21f1a038093c6e3ee7c09481-Paper-Conference.pdf.

Jaehong Yoon, Wonyong Jeong, Giwoong Lee, Eunho Yang, and Sung Ju Hwang. Federated continual learning with weighted inter-client transfer. In *International conference on machine learning*, pp. 12073–12086. PMLR, 2021a.

Tehrim Yoon, Sumin Shin, Sung Ju Hwang, and Eunho Yang. Fedmix: Approximation of mixup under mean augmented federated learning. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=Ogga20D2HO-.

Hao Zhang, Chenglin Li, Nuowen Kan, Ziyang Zheng, Wenrui Dai, Junni Zou, and Hongkai Xiong. Improving generalization in federated learning with model-data mutual information regularization: A posterior inference approach. *Advances in Neural Information Processing Systems*, 37: 136646–136678, 2024.

Jie Zhang, Chen Chen, Weiming Zhuang, and Lingjuan Lyu. Target: Federated class-continual learning via exemplar-free distillation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4759–4770. IEEE, 2023.

Yaoyue Zheng, Yin Zhang, Joost van de Weijer, Gido van de Ven, Shaoyi Du, Xuetao Zhang, and Zhiqiang Tian. Revisiting weight regularization for low-rank continual learning. In *International Conference on Learning Representations*, volume 2026, pp. 42895–42919, 2026.

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

A SPECTRAL THEORY

This section provides complete proofs of the spectral properties presented in the main text. For an original prompt $P_i \in \mathbb{R}^{N\times L_p\times D}$, we flatten its first two dimensions and retain the notation $P_i \in \mathbb{R}^{L_i\times D}$, where $L_i = NL_p$. This operation only rearranges prompt tokens and does not change the right singular subspace along the feature dimension D. For multiple tasks, $\mathrm{Concat}(\cdot)$ denotes row-wise concatenation.

For any symmetric matrix A, we order its eigenvalues as $\lambda_1(A) \geq \lambda_2(A) \geq \cdots$.

For task i, we write

$$P_i = P_i^{\mathrm{common}} + P_i^{\mathrm{specific}},$$

and denote

$$C_i = P_i^{\mathrm{common}}.$$

To distinguish the ideal spectral structure from actual non-orthogonal representations, we use $S_i^{(0)}$ to denote the task-specific component in the ideal reference model. Deviations of the actual prompt Gram matrix from this ideal reference structure are represented as spectral perturbations.

A.1 GLOBAL SHARED KNOWLEDGE CONCENTRATION

Suppose the common components of different tasks share an r-dimensional right subspace $V_C \in \mathbb{R}^{D\times r}$, satisfying

$$V_C^\top V_C = I, \qquad C_i = C_i V_C V_C^\top.$$

For any unit vector $v \in \mathrm{span}(V_C)$, assume that each task's spectral contribution along shared directions satisfies

$$\|C_i v\|_2^2 \geq m_G > 0.$$

In the ideal reference model, suppose the specific components $S_i^{(0)}$ occupy mutually orthogonal right subspaces that are also orthogonal to V_C, and satisfy

$$\|S_i^{(0)}\|_2^2 \leq M_G.$$

Define

$$C_G = \mathrm{Concat}(C_1, \ldots, C_M), \qquad S_G^{(0)} = \mathrm{Concat}(S_1^{(0)}, \ldots, S_M^{(0)}).$$

Proposition 1 (Global Shared Knowledge Concentration). Under the ideal reference model above,

$$\frac{(\sigma_{G,1}^{\mathrm{specific}})^2}{(\sigma_{G,r}^{\mathrm{common}})^2} \leq \frac{M_G}{Mm_G}.$$

Proof. Since

$$C_G^\top C_G = \sum_{i=1}^{M} C_i^\top C_i,$$

for any unit vector $v \in \mathrm{span}(V_C)$, we have

$$v^\top C_G^\top C_G v = \sum_{i=1}^{M} \|C_i v\|_2^2 \geq Mm_G.$$

Therefore,

$$\lambda_{\min}\left(V_C^\top C_G^\top C_G V_C\right) \geq Mm_G,$$

which gives

$$(\sigma_{G,r}^{\mathrm{common}})^2 \geq Mm_G.$$

Meanwhile, the mutually orthogonal right subspaces of the $S_i^{(0)}$ induce independent spectral blocks in $(S_G^{(0)})^\top S_G^{(0)}$. Thus,

$$\|S_G^{(0)}\|_2^2 = \max_i \|S_i^{(0)}\|_2^2 \leq M_G.$$

Hence,

$$(\sigma_{G,1}^{\text{specific}})^2 \leq M_G.$$

Combining the two bounds yields

$$\frac{(\sigma_{G,1}^{\text{specific}})^2}{(\sigma_{G,r}^{\text{common}})^2} \leq \frac{M_G}{Mm_G}.$$

This completes the proof.

When $Mm_G > M_G$, a positive gap separates the common and specific spectra. In particular, when m_G and M_G do not grow with M, the ratio above decays as $O(1/M)$.

A.2 Approximate Global Spectral Separation

Define the ideal reference Gram matrix

$$A_G^{(0)} = C_G^\top C_G + (S_G^{(0)})^\top S_G^{(0)}.$$

The actual global knowledge matrix is

$$Q_G = \text{Concat}(P_1, \ldots, P_M),$$

with Gram matrix

$$A_G = Q_G^\top Q_G.$$

Define the total deviation from the ideal reference structure as

$$E_G = A_G - A_G^{(0)},$$

so that

$$A_G = A_G^{(0)} + E_G, \qquad \|E_G\|_2 \leq \epsilon_G.$$

Here, E_G captures non-orthogonality among the actual specific subspaces, common–specific coupling, and other deviations from the ideal spectral structure.

Define the ideal spectral gap

$$\Delta_G = Mm_G - M_G.$$

Proposition 2 (Approximate Global Spectral Separation). If

$$\Delta_G > 0, \qquad \epsilon_G < \frac{\Delta_G}{2},$$

then

$$\lambda_r(A_G) \geq Mm_G - \epsilon_G,$$

and

$$\lambda_{r+1}(A_G) \leq M_G + \epsilon_G.$$

Thus, a positive gap remains between the perturbed common-dominated spectrum and the remaining spectral directions.

Proof. By Proposition 1 and the orthogonality between the ideal common and specific subspaces,

$$\lambda_r(A_G^{(0)}) \geq Mm_G,$$

and

$$\lambda_{r+1}(A_G^{(0)}) \leq M_G.$$

By Weyl's eigenvalue perturbation inequality, for any j,

$$\left|\lambda_j(A_G) - \lambda_j(A_G^{(0)})\right| \leq \|E_G\|_2 \leq \epsilon_G.$$

Therefore,

$$\lambda_r(A_G) \geq Mm_G - \epsilon_G,$$

and

$$\lambda_{r+1}(A_G) \le M_G + \epsilon_G.$$

Since

$$\epsilon_G < \frac{\Delta_G}{2},$$

we have

$$Mm_G - \epsilon_G > M_G + \epsilon_G.$$

The perturbed spectral gap therefore remains positive. This completes the proof.

These results show that ideal orthogonality is used only to characterize the reference spectral structure. The concentration of globally shared knowledge in the dominant spectrum persists as long as perturbations in the actual representations do not close the common–specific spectral gap.

A.3 Task-Conditioned Knowledge Concentration

For the current input x, let the retrieved historical tasks be

$$\mathcal{T}_x = \{t_1, \dots, t_k\},$$

with corresponding nonzero prompt relevance weights $\{\alpha_1, \dots, \alpha_k\}$.

Define the ideal common knowledge matrix

$$C_T = \mathrm{Concat}\left(\alpha_1 C_{t_1}, \dots, \alpha_k C_{t_k}\right),$$

and the ideal specific knowledge matrix

$$S_T^{(0)} = \mathrm{Concat}\left(\alpha_1 S_{t_1}^{(0)}, \dots, \alpha_k S_{t_k}^{(0)}\right).$$

Assume that the common components of the relevant tasks share the same r-dimensional right subspace V_C, satisfying

$$C_{t_i} = C_{t_i} V_C V_C^\top,$$

and that, for any unit vector $v \in \mathrm{span}(V_C)$,

$$\|C_{t_i} v\|_2^2 \ge m_T > 0.$$

Also assume

$$\|S_{t_i}^{(0)}\|_2^2 \le M_T,$$

with the ideal specific subspaces mutually orthogonal and orthogonal to V_C.

Proposition 3 (Task-Conditioned Knowledge Concentration). Define the effective task support as

$$k_{\text{eff}} = \frac{\sum_{i=1}^k \alpha_i^2}{\max_i \alpha_i^2}.$$

Then

$$1 \le k_{\text{eff}} \le k,$$

and

$$\frac{(\sigma_{T,1}^{\text{specific}})^2}{(\sigma_{T,r}^{\text{common}})^2} \le \frac{M_T}{m_T} \frac{1}{k_{\text{eff}}}.$$

Proof. Since

$$C_T^\top C_T = \sum_{i=1}^k \alpha_i^2 C_{t_i}^\top C_{t_i},$$

for any unit vector $v \in \mathrm{span}(V_C)$, we have

$$v^\top C_T^\top C_T v = \sum_{i=1}^k \alpha_i^2 \|C_{t_i} v\|_2^2 \ge m_T \sum_{i=1}^k \alpha_i^2.$$

Therefore,

$$(\sigma_{T,r}^{\text{common}})^2 \geq m_T \sum_{i=1}^{k} \alpha_i^2.$$

Meanwhile, the mutual orthogonality of the ideal specific subspaces gives

$$\|S_T^{(0)}\|_2^2 = \max_i \alpha_i^2 \|S_{t_i}^{(0)}\|_2^2 \leq M_T \max_i \alpha_i^2.$$

Thus,

$$(\sigma_{T,1}^{\text{specific}})^2 \leq M_T \max_i \alpha_i^2.$$

It follows that

$$\frac{(\sigma_{T,1}^{\text{specific}})^2}{(\sigma_{T,r}^{\text{common}})^2} \leq \frac{M_T}{m_T} \frac{\max_i \alpha_i^2}{\sum_{i=1}^{k} \alpha_i^2} = \frac{M_T}{m_T} \frac{1}{k_{\text{eff}}}.$$

Since at least one α_i is nonzero,

$$\max_i \alpha_i^2 \leq \sum_{i=1}^{k} \alpha_i^2 \leq k \max_i \alpha_i^2.$$

Therefore,

$$1 \leq k_{\text{eff}} \leq k.$$

This completes the proof.

When all relevant tasks have equal absolute weights, $k_{\text{eff}} = k$, recovering

$$\frac{(\sigma_{T,1}^{\text{specific}})^2}{(\sigma_{T,r}^{\text{common}})^2} \leq \frac{M_T}{k m_T},$$

which gives the ideal $O(1/k)$ spectral concentration rate.

A.4 Approximate Task-Conditioned Spectral Separation

Define the ideal task-conditioned Gram matrix

$$A_T^{(0)} = C_T^\top C_T + (S_T^{(0)})^\top S_T^{(0)}.$$

The actual task-conditioned knowledge matrix is

$$Q_T = \text{Concat}\left(\alpha_1 P_{t_1}, \ldots, \alpha_k P_{t_k}\right),$$

and

$$A_T = Q_T^\top Q_T = A_T^{(0)} + E_T,$$

where

$$\|E_T\|_2 \leq \epsilon_T.$$

Define

$$L_T = m_T \sum_{i=1}^{k} \alpha_i^2, \qquad U_T = M_T \max_i \alpha_i^2,$$

and the ideal common–specific spectral gap

$$\Delta_T = L_T - U_T.$$

Note that

$$\Delta_T > 0 \quad \Longleftrightarrow \quad k_{\text{eff}} > \frac{M_T}{m_T}.$$

Proposition 4 (Task-Conditioned Spectral Stability). If

$$\Delta_T > 0, \qquad \epsilon_T < \frac{\Delta_T}{2},$$

then

$$\lambda_r(A_T) \geq L_T - \epsilon_T,$$

and

$$\lambda_{r+1}(A_T) \leq U_T + \epsilon_T.$$

Furthermore, let $\widehat{V}_C$ denote the dominant subspace formed by the leading r eigenvectors of A_T. Then

$$\left\|\sin\Theta(\widehat{V}_C, V_C)\right\|_2 \leq \frac{2\epsilon_T}{\Delta_T}.$$

Proof. By Proposition 3, the ideal common spectral block satisfies

$$\lambda_r(A_T^{(0)}) \geq L_T,$$

while the ideal specific spectral block satisfies

$$\lambda_{r+1}(A_T^{(0)}) \leq U_T.$$

By Weyl's inequality,

$$\lambda_r(A_T) \geq L_T - \epsilon_T,$$

and

$$\lambda_{r+1}(A_T) \leq U_T + \epsilon_T.$$

Since

$$\epsilon_T < \frac{\Delta_T}{2},$$

we have

$$L_T - \epsilon_T > U_T + \epsilon_T.$$

Thus, the perturbed dominant spectrum remains separated from the remaining spectral directions.

The gap between the common subspace and the remaining spectral directions of $A_T^{(0)}$ is at least Δ_T. Applying the Davis–Kahan $\sin\Theta$ subspace perturbation theorem with $\|E_T\|_2 \leq \epsilon_T$ and $\epsilon_T < \Delta_T/2$ gives

$$\left\|\sin\Theta(\widehat{V}_C, V_C)\right\|_2 \leq \frac{2\epsilon_T}{\Delta_T}.$$

This completes the proof.

Therefore, the task-conditioned dominant subspace remains stable relative to the ideal common subspace as long as actual non-orthogonal perturbations do not close the common–specific spectral gap.

A.5 Bounded Weak-Spectrum Rectification

Our method adaptively constructs the task-conditioned knowledge subspace V_{task} using cumulative spectral energy. Let

$$\Pi_T = V_{\text{task}} V_{\text{task}}^\top$$

denote its orthogonal projection matrix. For the common component P^{common} of any historical prompt, define the shared knowledge residual not retained by V_{task} as

$$\rho_T = \left\|P^{\text{common}}(I - \Pi_T)\right\|_F^2.$$

Let V_R denote a weak spectral subspace constructed from the spectral space not retained by V_{task}, such that

$$\text{span}(V_R) \subseteq \text{span}(V_{\text{task}})^\perp,$$

and

$$V_R^\top V_R = I.$$

Define

$$\Pi_R = V_R V_R^\top.$$

Then

$$\Pi_R \preceq I - \Pi_T.$$

For a historical prompt P, spectral rectification is defined as

$$\widetilde{P} = P - \eta P \Pi_R, \qquad 0 \le \eta \le 1.$$

Proposition 5 (Bounded Weak-Spectrum Rectification). The rectified weak spectral component satisfies

$$\left\| \widetilde{P} V_R \right\|_F^2 = (1-\eta)^2 \left\| P V_R \right\|_F^2 .$$

The energy of the common component within the weak spectral subspace satisfies

$$\| P^{\text{common}} V_R \|_F^2 \le \rho_T.$$

Consequently, the perturbation to the common component caused by spectral rectification satisfies

$$\left\| \widetilde{P}^{\text{common}} - P^{\text{common}} \right\|_F^2 \le \eta^2 \rho_T.$$

Proof. Since $V_R^\top V_R = I$,

$$\Pi_R V_R = V_R V_R^\top V_R = V_R.$$

Therefore,

$$\begin{aligned} \widetilde{P} V_R &= (P - \eta P \Pi_R) V_R \\ &= P V_R - \eta P V_R \\ &= (1-\eta) P V_R. \end{aligned}$$

Taking squared Frobenius norms gives

$$\left\| \widetilde{P} V_R \right\|_F^2 = (1-\eta)^2 \left\| P V_R \right\|_F^2 .$$

Moreover, since

$$\Pi_R \preceq I - \Pi_T,$$

we have

$$\begin{aligned} \| P^{\text{common}} V_R \|_F^2 &= \operatorname{tr}\left(P^{\text{common}} \Pi_R (P^{\text{common}})^\top \right) \\ &\le \operatorname{tr}\left(P^{\text{common}} (I - \Pi_T) (P^{\text{common}})^\top \right) \\ &= \| P^{\text{common}} (I - \Pi_T) \|_F^2 \\ &= \rho_T. \end{aligned}$$

Finally, the rectified common component satisfies

$$\widetilde{P}^{\text{common}} = P^{\text{common}} - \eta P^{\text{common}} \Pi_R.$$

Thus,

$$\widetilde{P}^{\text{common}} - P^{\text{common}} = -\eta P^{\text{common}} \Pi_R.$$

Using $\Pi_R = V_R V_R^\top$ and the orthonormality of V_R,

$$\| P^{\text{common}} \Pi_R \|_F^2 = \| P^{\text{common}} V_R \|_F^2 .$$

It follows that

$$\begin{aligned} \left\| \widetilde{P}^{\text{common}} - P^{\text{common}} \right\|_F^2 &= \eta^2 \| P^{\text{common}} V_R \|_F^2 \\ &\le \eta^2 \rho_T. \end{aligned}$$

This completes the proof.

This result shows that weak-spectrum rectification explicitly attenuates knowledge components in V_R, while its perturbation to shared knowledge is controlled by the residual ρ_T not retained by V_{task}. The conclusion therefore does not require V_{task} to exactly recover the unknown full common subspace.

B Details of the Datasets and Evaluation Metrics

B.1 Datasets

We evaluate our method on four image datasets: ImageNet-R, DomainNet, CIFAR-100, and Cars196. ImageNet-R contains 30,000 images from 200 classes, including challenging ImageNet examples and images collected in different visual styles. It contains 24,000 training images and 6,000 test images. We further reserve 20% of the training samples for validation and model parameter selection. DomainNet is a large-scale multi-domain dataset containing approximately 600,000 images from 345 classes across six visual domains. CIFAR-100 contains 100 classes, with 50,000 training images and 10,000 test images. Cars196 is a fine-grained car classification dataset containing 196 classes, with 8,144 training images and 8,041 test images.

B.2 Configuration of Federated Continual Learning Benchmarks

For ImageNet-R, DomainNet, CIFAR-100, and Cars196, we follow the benchmark configuration of C^2Prompt (Xu et al., 2025). This setting considers both cross-task transferability and asynchronous task arrivals across clients. Specifically, different continual learning tasks may share classes, while each task uses only a subset of the training data from its classes, creating task sequences with varying degrees of relatedness.

For ImageNet-R, each task randomly selects 20 classes and uses 20% of the training samples from each class. Tasks are assigned to different clients and arrive asynchronously. For DomainNet, each task randomly selects 35 classes and uses 2% of the training samples from each class, with all other settings matching ImageNet-R. The smaller sampling ratio accommodates differences between this dataset and the pretraining data distribution. CIFAR-100 and Cars196 use the same FCL benchmark settings as ImageNet-R. Following prior work, we control class overlap to construct tasks with varying degrees of relatedness rather than using the Dirichlet partitioning common in federated learning. In FCL, task-specific class sets continually change, making task relatedness difficult to control directly through a Dirichlet distribution.

We use five clients, each learning independent and evolving local tasks. Training is divided into stages of three communication rounds each. At the beginning of each stage, 40% of clients switch to new tasks to simulate asynchronous task arrivals. LCDC is trained with Adam when a new task arrives. Its parameters are then frozen, and it is used only for subsequent classification prompt training, not during testing. During relevant-task retrieval, the model selects the top-k tasks from those that have arrived, based on similarity weights for the current task. Let t denote the global stage index starting from 1, with three communication rounds per stage. We use $k = 4 + t$ for ImageNet-R, CIFAR-100, and Cars196, and $k = 3 + t$ for DomainNet, so k increases over time. For fair comparison, the classification prompt optimizer, learning rate, and number of local training epochs follow Powder (Piao et al., 2024).

B.3 Evaluation Metrics

We use three metrics to evaluate continual learning performance: Average Accuracy (Avg), Average Incremental Accuracy (AIA), and Final Average Accuracy (FAA).

Average Accuracy (Avg). Avg measures performance across all learned tasks at the end of training. Let $\mathcal{T}$ denote the set of tasks encountered throughout FCL, and let $a^t_{c,\max(\mathcal{R})}$ denote the accuracy on task $\mathcal{T}^t_c$ after the final communication round. Then

$$\text{Avg} = \frac{1}{|\mathcal{T}|} \sum_{\mathcal{T}^t_c \in \mathcal{T}} a^t_{c,\max(\mathcal{R})}. \tag{16}$$

Average Incremental Accuracy (AIA). AIA measures average performance on all encountered tasks throughout continual learning. Let $\mathcal{R}$ denote the set of communication rounds at which task switches occur, $\mathcal{T}_r$ the set of tasks encountered by round r, and $a^t_{c,r}$ the test accuracy on task $\mathcal{T}^t_c$ at that round. Then

$$\text{AIA} = \frac{1}{|\mathcal{N}|} \sum_{r \in \mathcal{R}} \sum_{\mathcal{T}^t_c \in \mathcal{T}_r} a^t_{c,r}. \tag{17}$$

Final Average Accuracy (FAA). FAA measures final knowledge retention and accumulation across all tasks after completing incremental learning. Let a_t denote the test accuracy on task t after the final incremental stage. Then

$$\mathrm{FAA} = \frac{1}{T}\sum_{t=1}^{T} a_t, \tag{18}$$

where T is the total number of incremental tasks.

C Results under Other Federated Incremental Learning Settings

Beyond the FCL setting in the main text, we evaluate our method under Federated Class-Incremental Learning (FCIL). Unlike the FCL setting, which allows class overlap across tasks, FCIL divides learning into incremental tasks with disjoint class sets, providing an additional setting for evaluating applicability. We follow C^{2}Prompt (Xu et al., 2025) and compare with existing continual learning, federated learning, and federated continual learning methods on ImageNet-R.

FCIL Setting. We divide learning into 10 incremental tasks with disjoint class sets. ImageNet-R contains 200 classes, yielding 20 classes per task. For each task, training data are partitioned among 10 clients using a Dirichlet distribution with concentration parameter $\beta \in \{0.5, 0.1, 0.05\}$ to simulate different levels of non-IID data distributions. Smaller β values produce less balanced client distributions and stronger data heterogeneity.

Training Details. We use an ImageNet-21K-pretrained ViT-B/16 backbone. Each incremental task involves five communication rounds, with five local training epochs per client per round. The complete 10-task sequence therefore comprises 50 communication rounds. Training images undergo random horizontal flipping and normalization. Test images are resized to 256×256 using bicubic interpolation, center-cropped to 224×224, and normalized. Except for client data partitioning, task splits and training configurations remain identical across β settings for fair comparison.