Federated Continual Instruction Tuning

Haiyang Guo[1,2], Fanhu Zeng[2,3], Fei Zhu[4], Wenzhuo Liu[2,3], Da-Han Wang[5],
Jian Xu[2,3], Xu-Yao Zhang[1,2,3]*, Cheng-Lin Liu[1,2,3]
[1]School of Advanced Interdisciplinary Sciences, UCAS [2]MAIS, CASIA
[3]School of Artificial Intelligence, UCAS [4]Centre for Artificial Intelligence and Robotics, HKISI-CAS
[5]FKLPRIU, School of Computer and Information Engineering, Xiamen University of Technology
{guohaiyang2023, zengfanhu2022, jian.xu}@ia.ac.cn, zhfei2018@gmail.com, {xyz, liucl}@nlpr.ia.ac.cn

Abstract

A vast amount of instruction tuning data is crucial for the impressive performance of Large Multimodal Models (LMMs), but the associated computational costs and data collection demands during supervised fine-tuning make it impractical for most researchers. Federated learning (FL) has the potential to leverage all distributed data and training resources to reduce the overhead of joint training. However, most existing methods assume a fixed number of tasks, while in real-world scenarios, clients continuously encounter new knowledge and often struggle to retain old tasks due to memory constraints. In this work, we introduce the Federated Continual Instruction Tuning (FCIT) benchmark to model this real-world challenge. Our benchmark includes two realistic scenarios, encompassing four different settings and twelve carefully curated instruction tuning datasets. To address the challenges posed by FCIT, we propose a dynamic knowledge organization to effectively integrate updates from different tasks during training and subspace selective activation to allocate task-specific output during inference. Extensive experimental results demonstrate that our proposed method significantly enhances model performance across varying levels of data heterogeneity and catastrophic forgetting. Code and dataset are released at `https://github.com/Ghy0501/FCIT`.

1. Introduction

Large Multimodal Models (LMMs) [2, 32, 34], which integrate Large Language Model [4, 20, 46, 49] with a visual encoder and multimodal projector to bridge visual and textual modalities, have exhibited impressive visual understanding and complex reasoning abilities. A crucial factor in this success is the supervised fine-tuning of LMMs using huge and diverse visual instruction-following data [32] to align with human preferences. However, collecting such vast amounts of training data and computational resources for joint fine-tuning is impractical for most researchers. Federated Learning (FL) [28, 31, 38, 43], as a decentralized paradigm, offers a viable alternative by leveraging distributed data and computational resources for local training while integrating local weights to produce a unified model. This paradigm accommodates constraints on storage and computation while ensuring privacy protection.

Generally, most existing FL frameworks [7, 55, 58] are modeled in static, closed-world scenarios [3, 59], where a fixed set of tasks is predefined and unchanged. However, real-world applications are dynamic [59, 63], requiring models to continuously acquire new knowledge while retaining previously learned tasks. Taking a realistic health emergency as an example, different hospitals act as local clients when a major disease outbreak occurs. Large hospitals can leverage FL to collaboratively train on their case data, building a comprehensive virus knowledge base, while smaller clinics update their own cases using this global knowledge. Over time, both large and small hospitals need to keep updating the global knowledge base through continual learning methods to cope with the event. Additionally, the emergence of new strains necessitates the simultaneous integration of knowledge and response strategies to minimize potential losses. In this scenario, traditional FL methods struggle with newly arrived knowledge, while Continual Learning (CL) [14, 15, 33, 56, 60] methods alone do not facilitate knowledge sharing between clients. Only an organic integration of both can address this real-world problem. Recently, numerous Federated Continual Learning methods have emerged to address this challenge in traditional image classification tasks [11, 12, 57]. However, their methodological and task-setting limitations make them insufficient for current LMM applications. Therefore, a more comprehensive benchmark is needed to better simulate the practical application of LMMs in real-world scenarios.

In this work, we first establish a *Federated Continual In-*

*Corresponding Author.

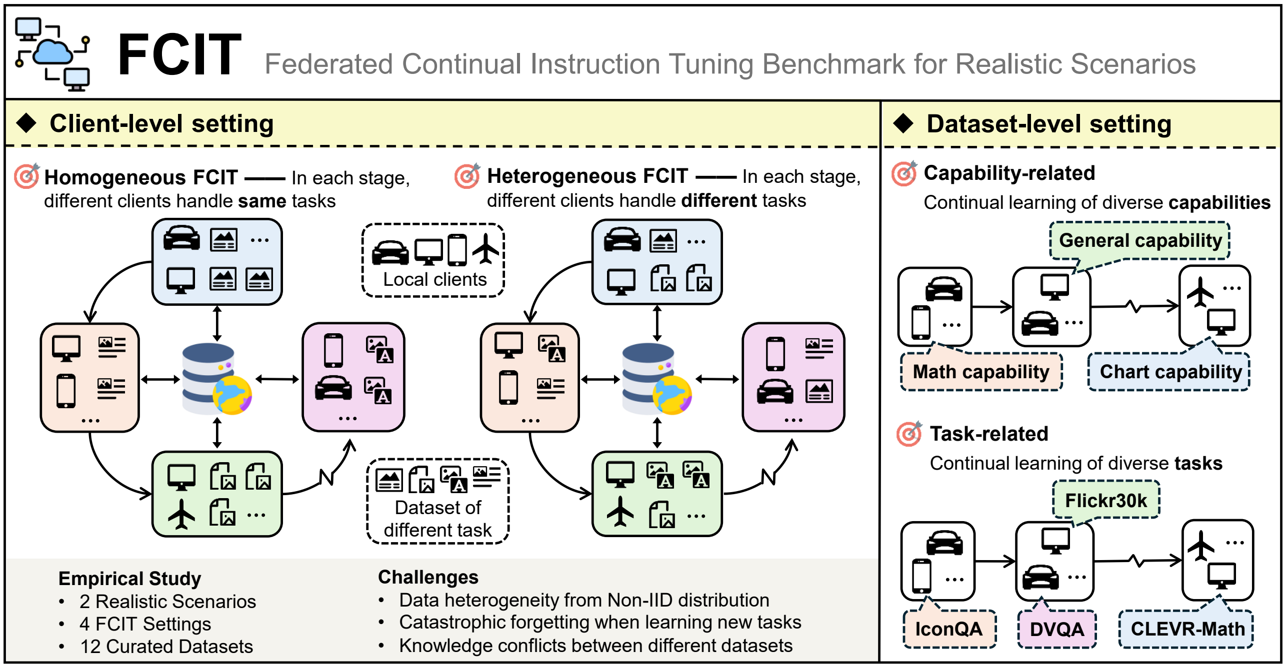

Figure 1. Overview of FCIT benchmark. FCIT encompasses 2 real-world scenarios, 4 FCIT settings, and 12 curated datasets, providing a comprehensive simulation of LMMs instruction-following training in real-world applications.

struction Tuning (*FCIT*) benchmark to fill this gap. Specifically, we simulate two client-level realistic scenarios: **(i) Homogeneous FCIT** refers to an FL system learning a series of instruction tuning data sequentially, with different clients learning the same task at each stage. **(ii) Heterogeneous FCIT**, on the other hand, requires the FL system to collaborate on different tasks simultaneously, with different clients potentially learning different tasks in the same stage. Building on this, we define two settings of datasets for each scenario: **Capability-related** and **Task-related**. The former evaluates the model's ability to integrate multiple dimensions of instruction following in a short period, while the latter assesses the model's performance during long phases of continual learning. For dataset selection, we curate twelve instruction tuning datasets unseen during LMM fine-tuning or with low zero-shot performance, preventing information leakage [13, 23]. In addition, we introduce varying degrees of data heterogeneity for each setting to challenge the model's performance in the non-IID situation [27]. An illustration of our FCIT benchmark is provided in Figure 1. To the best of our knowledge, this is the first work to introduce a comprehensive benchmark for federated learning of LMM in a continual learning setting.

To effectively address the challenges posed by FCIT, we propose a novel *Dynamic knowledge organIzation* and *Subspace seleCtive activatiOn* (*DISCO*) framework. Specifically, we identify FCIT challenges into two types: conflicts between different tasks within the same stage and conflicts between old and new tasks across different stages. The former can cause catastrophic forgetting by altering the parameter space of previous tasks when learning new ones without access to past data, while the latter requires the model to integrate and organize knowledge from different tasks to derive unified representations. Therefore, we first propose *Dynamic Knowledge Organization* (DKO), which leverages a dynamic cache at the global server to store task-specific parameters. Using a unique identity token matching mechanism, it systematically organizes knowledge for different tasks into corresponding subspaces within the cache, effectively mitigating two types of conflicts. To better utilize the organized task subspaces in the dynamic cache, we introduce *Subspace Selective Activation* (SSA), which selectively activates subspaces relevant to the test input while filtering out irrelevant outputs, leading to significant performance improvements. Consequently, these designs enable our framework to efficiently tackle the data heterogeneity and catastrophic forgetting in FCIT. In summary, our major contributions are:

- We present the first Federated Continual Instruction Tuning (FCIT) benchmark designed for real-world scenarios, providing a comprehensive evaluation of LMMs to continuously learn new knowledge using distributed data and training resources in real applications.
- We propose a novel DISCO framework that integrates dynamic knowledge organization and subspace selective activation to efficiently address data heterogeneity and catastrophic forgetting in FCIT settings.
- Extensive experiments demonstrate that our method sig-

nificantly enhances model performance under data heterogeneity while minimizing the catastrophic forgetting, and achieves state-of-the-art performance.

2. Related Work

Large Multimodal Models. With the grand unification of Large Language Models (LLMs) [4, 20, 49] for various NLP tasks, Large Multimodal Models (LMMs) [2, 9, 32, 34, 61] emerge by extending LLMs to combine with visual encoder and multimodal projectors, demonstrating exceptional visual understanding and complex reasoning abilities. To better align with human preferences, these LMMs typically undergo further fine-tuning on extensive instruction-following data, ensuring they meet the demands of real-world applications [32]. However, in real-world scenarios, improving the performance of LMMs on new downstream tasks becomes a significant challenge without access to sufficient training data and computational resources. In this paper, we introduce the first Federated Continual Instruction Tuning (FCIT) benchmark to bridge this gap from the perspective of distributed training and continual learning.

Federated Continual Learning. In classical vision tasks, Federated Continual Learning (FCL) [54] aims to adapt the global model to new data while maintaining the knowledge of the old task. From the perspective of LMMs, research in this setting faces two main challenges: (1) Most of the methods are designed for traditional vision tasks (*i.e.,* image classification). For instance, MFCL [1] employs a generative model to synthesize images of previously learned classes, thereby mitigating forgetting. PILoRA [12] introduces a prototype re-weight module to address the classifier bias caused by data heterogeneity and obtain unified knowledge through LoRA [18] fusion. Despite progress in classification tasks, complex designs remain challenging to adapt for LMMs research. (2) FCL follows a single data composition by splitting a dataset (*e.g.* ImageNet [10]) into different tasks based on classes, whereas LMMs face greater challenges in continual learning due to their diverse tasks with varying styles [5, 6, 56].

This work pioneers a Federated Continual Instruction Tuning setup for LMMs and establishes diverse scenarios to simulate real-world applications comprehensively. Notably, AFCL [45] is most relevant to our work, as it enables clients to continuously learn multiple tasks in different orders and asynchronous time slots. However, it is specifically designed for image classification tasks, whereas our study focuses on the more widely used LMM.

3. Problem Formulation

3.1. Preliminaries

Instruction tuning enhances LMM's ability to understand and execute human instructions by performing supervised fine-tuning of pre-trained models on extensive datasets comprising instructions and responses. Formally, the instruction data $\mathcal{D} = \{(\mathbf{x}_v^j, \mathbf{x}_{ins}^j, \mathbf{x}_{res}^j)_{j=1}^N\}$ consists of image input $\mathbf{x}_v$, instruction $\mathbf{x}_{ins}$ and response $\mathbf{x}_{res}$, where N represents the total number of samples. For clarity, given a simple image-instruction pair with a response of length L, the objective of an LMM is to predict the next token autoregressively based on all preceding tokens:

$$p(\mathbf{x}_{res}|\mathbf{x}_v, \mathbf{x}_{ins}) = \prod_{i=1}^{L} p_\theta(x_i|\mathbf{x}_v, \mathbf{x}_{ins}, \mathbf{x}_{res,<i}), \quad (1)$$

where θ denotes the trainable parameters during fine-tuning, $\mathbf{x}_{\text{res},<i}$ denotes all response tokens preceding the current prediction token x_i. Then, the loss function of fine-tuning LMMs can be expressed as:

$$\mathcal{L}_\theta = -\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{L_j} \log p_\theta(x_i^j|\mathbf{x}_v^j, \mathbf{x}_{ins}^j, \mathbf{x}_{res,<i}^j). \quad (2)$$

Federated learning framework typically comprises a global server and several local clients, all employing the same LMM with a shared homogeneous model architecture. In each communication round[1], local clients train their models on own data and upload the updated weights to the global server for aggregation, enabling collaborative optimization of the global model while preserving data privacy.

Considering the excessive communication overhead of transferring the entire LMM between clients and the global server, we adopt LoRA [18] for efficient fine-tuning, balancing training overhead and implementation cost [55, 58]. Specifically, for a weight matrix $\mathbf{W}_0 \in \mathbb{R}^{d\times k}$, LoRA decomposes parameter updates $\Delta\mathbf{W}$ during fine-tuning into two low-rank subspaces:

$$\mathbf{W} = \mathbf{W}_0 + \Delta\mathbf{W} = \mathbf{W}_0 + \mathbf{BA}, \quad (3)$$

where $\mathbf{B} \in \mathbb{R}^{d\times r}$, $\mathbf{A} \in \mathbb{R}^{r\times k}$ and $r \ll \min\{d, k\}$.

Continual learning aims to minimize the loss on the current task while retaining knowledge from previous tasks. Formally, given a sequence of datasets $\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_T$, the optimization objective at task t is:

$$\begin{aligned} &\min_\theta \mathcal{L}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_t}[\mathcal{L}(f_{\theta_t}(x), y)] + \sum_{i=1}^{t-1}\epsilon_i, \\ &\text{s.t. } \mathbb{E}_{(x,y)\sim\mathcal{D}_i}[\mathcal{L}(f_{\theta_t}(x), y) - \mathcal{L}(f_{\theta_{t-1}}(x), y)] \leq \epsilon_i, \\ &\qquad \epsilon_i \geq 0; \forall i \in [1, \cdots, t-1], \end{aligned} \quad (4)$$

where ϵ_i is a slack variable that allows a small increase in the loss from the old datasets, providing tolerance for minor forgetting while focusing on learning the current task. In Eq. 4, x and y can be viewed as the multimodal inputs $(\mathbf{x}_v, \mathbf{x}_{ins})$ and the response $\mathbf{x}_{res}$, respectively.

[1]For clarification, we standardize the communication round in Section 4 to 1.

3.2. Federated Continual Instruction Tuning

As shown in Figure 1, we integrate federated and continual learning for LMMs within a unified framework and propose two client-level realistic scenarios: (1) **Homogeneous FCIT** (Hom-FCIT). In this scenario, clients sequentially learn a series of tasks, allowing the global server to continuously update its knowledge. Each client learns the same task at a given stage and can only access data from that stage. (2) **Heterogeneous FCIT** (Het-FCIT). In real-world applications, different clients may learn different tasks within the same stage, enabling the global server to respond more rapidly to diverse instructions. This requires the model not only to coordinate the knowledge of different tasks learned by clients in the current stage but also to mitigate forgetting during the learning process. For each scenario, we define two dataset-level settings: capability-related and task-related, detailed as follows.

- **Capability-related.** Following the dataset construction and division in LLaVA-OneVision [25], we classify the 12 datasets into 4 capabilities: *General*, *Math*, *Chart*, and *Other*. The *General* capability includes A-OKVQA [44], ImageNet-R [16], Grounding [37], and IconQA [35]; *Math* comprises CLEVR-Math [8], super-CLEVR [30], and TabMWP [36]; *Chart* involves ArxivQA [26], FigureQA [22], and DVQA [21]; and *Other* encompasses OCR-VQA [39] and Flickr30k [40]. We treat these four capabilities as different stages of continual learning, with each capability comprising a mixture of datasets.
- **Task-related.** To evaluate the performance of different methods in a long-phase continual learning situation, we selected 8 datasets: ImageNet-R, ArxivQA, IconQA, CLEVR-Math, OCR-VQA, Flickr30k, FigureQA, and super-CLEVR as distinct stages of continual learning.

In total, we propose 4 FCIT settings to evaluate different methods. Compared to existing works [5, 6, 55, 56, 58], we are the first to explore the organic integration of FL and CL in the context of LMM. We provide further illustrations on the settings and datasets in Appendix A.

4. The Proposed Framework: *DISCO*

Overview of the Method. As shown in Figure 2 and 3, our method primarily consists of: **(1) Dynamic Knowledge Organization (DKO)**, which dynamically integrates knowledge learned by different clients across stages during training, significantly reducing inter-task conflicts (Section 4.1); and **(2) Subspace Selective Activation (SSA)**, which selectively activates subspace outputs based on input features during inference, effectively filtering out irrelevant information (Section 4.2). We name this framework **DISCO**.

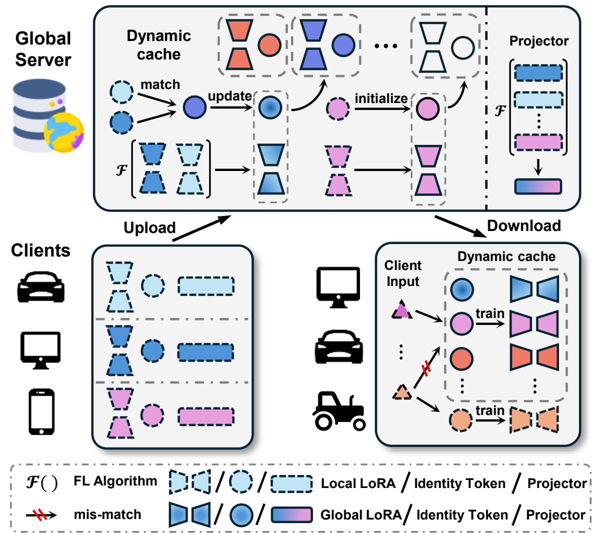

Figure 2. Illustration of the proposed DKO. Dynamic caches store the knowledge of each subspace while matching and updating through identity tokens.

4.1. Dynamic Knowledge Organization

The core challenge of FCIT is enabling the global server to effectively harmonize knowledge learned by clients, addressing both the forgetting of previous knowledge when acquiring new tasks and the conflicts that arise from integrating knowledge from different tasks within the same stage. To this end, we propose maintaining a dynamic cache at the global server to organize task-specific knowledge uploaded by clients at different stages, thereby preventing both forgetting and conflicts between knowledge. Specifically, each task is assigned a dedicated parameter space to store and update its corresponding knowledge:

$$\Delta \mathbf{W} = \mathbf{BA} \Leftrightarrow \underbrace{\{\mathbf{B}_1\mathbf{A}_1, \cdots, \mathbf{B}_T\mathbf{A}_T\}}_{\text{task-specific subspace}}, \tag{5}$$

where T is the number of tasks learned. At this point, it is crucial to aggregate the parameters uploaded by clients into their respective subspaces without privacy leakage.

Inspired by the widespread use of prototypes in FL fields [19, 47, 48], we propose to distinguish knowledge across tasks by using the feature mean of each client's training data as an **Identity Token**. Specifically, considering the uniqueness of textual inputs in the visual instruction tuning task, we introduce a text encoder f_{ins} for each client to extract the feature of the input instruction $\mathbf{x}^t_{ins,j}$ during training, and take the mean value μ^t_k as the local identity token of the k-th client on the task t at the end of training:

$$\mu^t_k = \frac{1}{n^t_k}\sum_{j=1}^{n^t_k} f_{ins}(\mathbf{x}^t_{ins,j}), \tag{6}$$

where n^t_k is the number of training samples of client k at

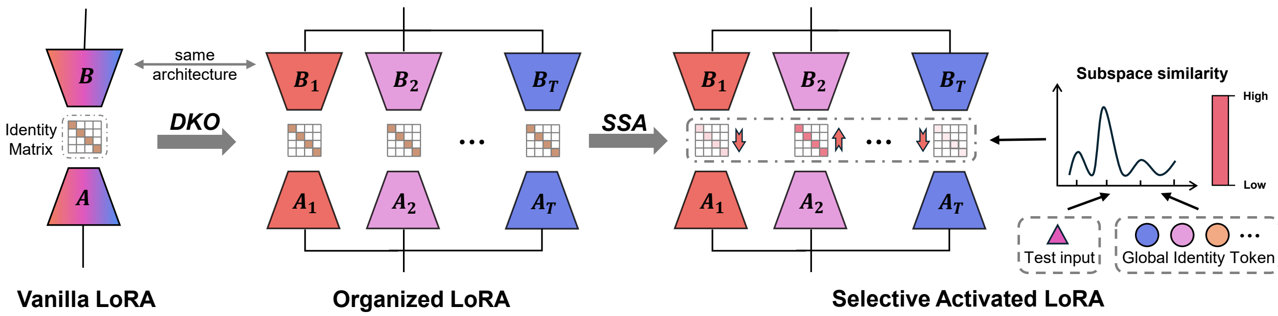

Figure 3. Illustration of the proposed SSA. Each subspace organized by DKO dynamically adjusts its output via the intrinsic activation matrix during inference, effectively filtering irrelevant information and enhancing model performance.

task t, and we use CLIP's text encoder [41] in our experiments. Notably, the text encoder f_{ins} remains frozen during training, which prevents additional training overhead.

Using the identity tokens, we first apply cosine similarity along with a threshold to match the uploaded local identity token μ_j^t with the global identity token $\tilde{\mathbf{z}}_i$:

$$\frac{\mu_j^t \cdot \tilde{\mathbf{z}}_i}{\|\mu_j^t\|\|\tilde{\mathbf{z}}_i\|} \geq \tau, \tag{7}$$

where τ denotes a pre-defined threshold. For paired local identity tokens, we update the corresponding global identity token. For mismatched local identity tokens, we pair them two by two using Eq. (7) and initialize new global identity tokens. This process can be formalized as:

$$\tilde{\mathbf{z}}_i^* = \begin{cases} \frac{n_{\tilde{\mathbf{z}}_i} \cdot \tilde{\mathbf{z}}_i + \sum_{j=1}^m n_j^t \cdot \mu_j^t}{n_{\tilde{\mathbf{z}}_i} + \sum_{j=1}^m n_j^t}, & \text{if } \tilde{\mathbf{z}}_i \text{ exists}, \\ \frac{\sum_{j=1}^m n_j^t \cdot \mu_j^t}{\sum_{j=1}^m n_j^t}, & \text{if } \tilde{\mathbf{z}}_i \text{ doesn't exist}, \end{cases} \tag{8}$$

where m, $n_{\tilde{\mathbf{z}}_i}$ and n_j^t denote the number of matched local identity tokens, the number of samples used to form the previous i-th global identity token, and the number of samples from the local client j at task t, respectively. Then, we leverage this matching process to guide the update or initialization of task-specific subspace $\{\theta_i = \mathbf{B}_i\mathbf{A}_i\}$ in Eq. (5):

$$\theta_i^* = \mathcal{F}(\theta_1^t, \cdots, \theta_m^t), \tag{9}$$

where $\mathcal{F}$ denotes the FL algorithm (*e.g.* FedAvg [38]) used to aggregate local weights [2]. As a result, we effectively prevent inter-task conflicts and integrate knowledge from different clients using the identity token matching mechanism.

After the global server completes the aggregation, it distributes the dynamic cache to each selected client. The client then matches the identity token (Using Eq. (7)) of each subspace with its own training data, deciding whether to update the corresponding subspace or reinitialize a new one. The entire DKO process is illustrated in Figure 2.

[2]In Section 5.4, we implement more FL algorithms to demonstrate the compatibility of our method.

4.2. Subspace Selective Activation

Section 4.1 effectively mitigates inter-task conflicts by disentangling and organizing complex knowledge into distinct subspaces. The key challenge then lies in how to leverage these task-specific subspaces during inference.

To address this, an intuitive method is to concatenate the subspaces of the dynamic cache in low-rank dimensions [50, 52], enabling the integration of knowledge across all task spaces. However, this may introduce redundant information unrelated to the current task during inference, potentially compromising the model's output. For instance, when the desired answer is a simple word (*e.g.*, what is the object in the picture), knowledge from other subspaces, such as those used for generating long-form descriptions, can introduce unnecessary information, leading to responses that do not align with the given instruction.

Drawing inspiration from LoRA's intrinsic space [53], we propose subspace selective activation (SSA) without additional training to filter out irrelevant subspace outputs, ensuring alignment between responses and instructions. In particular, A vanilla LoRA can be decomposed as the product of two low-rank subspaces (*i.e.,* $\mathbf{A}$, $\mathbf{B}$) and an intrinsic mixing matrix $\mathcal{W}$:

$$\Delta\mathbf{W} = \mathbf{B}\mathcal{W}\mathbf{A}, \tag{10}$$

where $\mathcal{W} \in \mathbb{R}^{r\times r}$ is typically an identity matrix and can thus be omitted. In this paper, we redefine $\mathcal{W}$ as the activation matrix that dynamically responds to the test input. Specifically, we treat $\mathcal{W}$ as the product of an identity matrix and an activation factor (*i.e.,* $\mathcal{W} = \alpha \cdot \mathbf{I}_{r\times r}$), where $\alpha = 1$ denotes full activation, and $\alpha = 0$ means that the output is fully masked. Therefore, for each subspace $\{\mathbf{B}_1\mathbf{A}_1, \cdots, \mathbf{B}_T\mathbf{A}_T\}$ in dynamic cache, we can control the activation or inhibition of the corresponding output by adjusting its activation factor $\{\alpha_1, \cdots, \alpha_T\}$.

To provide better flexibility in assigning activation factors, we use global identity tokens and test input features extracted by the text encoder for similarity matching:

$$s_i = \frac{\tilde{\mathbf{z}}_\mathbf{i} \cdot f_{ins}(\mathbf{x}_{ins}^{test})}{\|\tilde{\mathbf{z}}_\mathbf{i}\| \cdot \|f_{ins}(\mathbf{x}_{ins}^{test})\|}, \tag{11}$$

Dataset setting	Capability-related (4 task)						Task-related (8 task)					
Partition	$\beta = 0.5$		$\beta = 1.0$		$\beta = 5.0$		$\beta = 0.5$		$\beta = 1.0$		$\beta = 5.0$	
Methods	Last	Avg	Last	Avg	Last	Avg	Last	Avg	Last	Avg	Last	Avg
Zero-shot	30.57	-	30.57	-	30.57	-	29.08	-	29.08	-	29.08	-
Individual	61.75	-	59.62	-	60.27	-	64.21	-	63.85	-	64.07	-
Centralized MTL	63.83	-	63.83	-	63.83	-	66.60	-	66.60	-	66.60	-
Finetune	48.68	60.05	49.67	59.01	50.40	58.08	46.24	68.93	47.20	68.79	48.00	69.97
EWC	49.12	60.13	49.46	59.28	49.89	58.76	47.51	69.04	47.92	69.22	48.15	70.27
LwF	48.87	60.32	50.02	59.88	50.15	59.06	47.89	69.57	47.62	69.14	48.26	70.30
L2P	48.22	59.79	49.56	59.34	49.62	59.25	47.30	69.31	48.08	69.65	48.42	70.16
O-LoRA	51.65	60.19	49.13	57.93	50.29	58.18	52.87	71.54	49.87	70.26	47.76	70.84
M-LoRA	49.04	60.08	50.39	60.76	50.56	58.06	50.68	71.94	48.53	71.58	48.38	71.21
MoELoRA	49.69	61.00	50.90	60.18	50.43	59.11	49.23	70.96	49.02	70.65	48.82	71.08
DISCO	**53.73**	**62.00**	**55.47**	**62.07**	**55.06**	**60.53**	**57.69**	**74.03**	**56.22**	**73.03**	**55.58**	**72.64**

Table 1. Last and Avg performance of different methods on **Hom-FCIT** setting. The best performance is shown in bold.

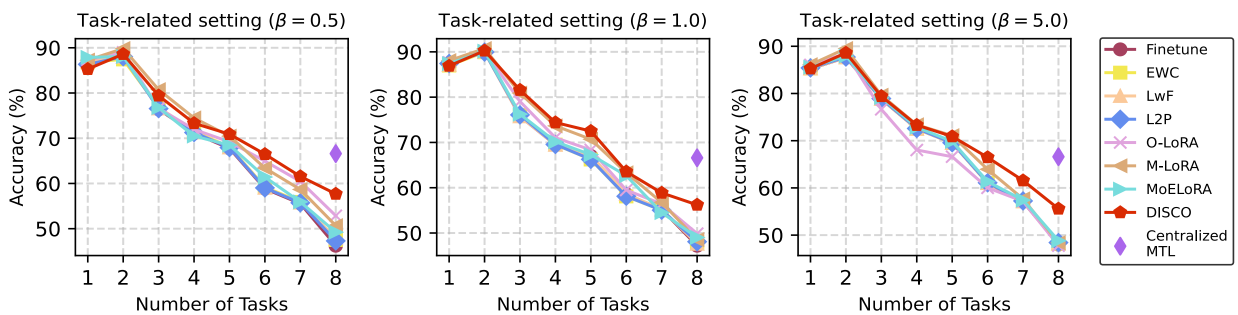

Figure 4. Performance curves of different methods in **Hom-FCIT** across seen tasks under varying data heterogeneity. We plot the average performance across all seen tasks at each stage.

where s_i denotes the similarity between the i-th global identity token and the feature of the test instruction $\mathbf{x}_{ins}^{test}$. The activation factor α_i is then computed by applying softmax normalization over all similarities:

$$\alpha_i = \frac{\exp(s_i/\varepsilon)}{\sum_{j=1}^{T} \exp(s_j/\varepsilon)}, \tag{12}$$

where ε is the temperature coefficient.

Overall, our proposed SSA can be formulated as follows:

$$\begin{aligned} \Delta \mathbf{W} &= \bar{\mathbf{B}} \bar{\mathcal{W}} \bar{\mathbf{A}} \\ &= \bar{\mathbf{B}} \begin{bmatrix} \alpha_1 \cdot \mathbf{I}_{r\times r} & \mathbf{0}_{r\times r} & \cdots & \mathbf{0}_{r\times r} \\ \mathbf{0}_{r\times r} & \alpha_2 \cdot \mathbf{I}_{r\times r} & \cdots & \mathbf{0}_{r\times r} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{r\times r} & \mathbf{0}_{r\times r} & \cdots & \alpha_T \cdot \mathbf{I}_{r\times r} \end{bmatrix} \bar{\mathbf{A}}, \end{aligned} \tag{13}$$

where $\bar{\mathbf{B}} \in \mathbb{R}^{d\times(r\cdot T)}$ represents the concatenation of all $\{\mathbf{B}_1, \cdots, \mathbf{B}_T\}$ in the dynamic cache along the low-rank dimension r, and $\bar{\mathbf{A}} \in \mathbb{R}^{(r\cdot T)\times k}$ follows the same structure. The activation factors $\{\alpha_1, \cdots, \alpha_T\}$, computed via Eq. (11) and Eq. (12), control each subspace's output by amplifying matched subspaces while suppressing irrelevant ones. The schematic of SSA is provided in Figure 3.

5. Experiments

5.1. Experimental Setup

Datasets. The dataset composition is detailed in Section 3.2. For the experimental setup, we implement two types of dataset-level settings: Capability-related and Task-related, based on the two client-level realistic scenarios, Hom-FCIT and Het-FCIT, respectively. At each stage, the own data of local clients is partitioned according to the Dirichlet distribution [27], with three partitions β for each setting to model different levels of data heterogeneity. More details are provided in Appendix A.

Baselines & Evaluation Metrics. We compare our method with continual learning methods such as LwF [29], EWC [24], L2P [51], O-LoRA [50], MoELoRA [6], and also with the LoRA merging method from the federated continual learning approach PILoRA [12], referred to as M-LoRA. We follow a rehearsal-free continual learning setting [62], where only the data for the current task is available. All comparison methods are carefully calibrated to ensure the fairness of evaluations.

For the evaluation metrics, we report the standard metrics to measure the model performance: ***Last*** refers to the

Dataset setting	Capability-related (4 task)						Task-related (8 task)					
Partition	$\beta = 0.5$		$\beta = 1.0$		$\beta = 5.0$		$\beta = 0.5$		$\beta = 1.0$		$\beta = 5.0$	
Methods	Last	Avg	Last	Avg	Last	Avg	Last	Avg	Last	Avg	Last	Avg
Zero-shot	30.57	-	30.57	-	30.57	-	29.08	-	29.08	-	29.08	-
Individual	61.75	-	59.62	-	60.27	-	64.21	-	63.85	-	64.07	-
Centralized MTL	63.83	-	63.83	-	63.83	-	66.60	-	66.60	-	66.60	-
Finetune	55.65	55.82	56.34	56.85	56.74	57.15	58.04	53.05	57.96	54.22	58.17	54.87
EWC	55.21	54.58	55.13	55.76	55.69	56.02	58.74	53.76	57.96	54.18	57.44	54.30
LwF	55.92	56.18	56.42	56.80	56.81	57.26	58.82	53.77	58.01	54.34	58.22	54.78
L2P	56.10	56.63	56.76	57.02	56.95	57.36	58.84	53.80	58.73	54.67	58.39	54.66
O-LoRA	58.24	58.58	58.32	58.65	58.60	58.94	59.61	54.55	59.74	55.20	59.51	54.97
M-LoRA	57.76	58.02	57.65	57.89	57.80	58.17	59.76	54.11	58.82	54.03	59.35	54.89
MoELoRA	57.68	57.95	57.77	58.00	58.15	58.44	59.02	54.25	59.14	54.69	58.86	54.50
DISCO	**59.40**	**60.01**	**59.94**	**59.91**	**59.71**	**60.16**	**62.08**	**59.64**	**63.25**	**61.99**	**62.78**	**60.87**

Table 2. Last and Avg performance of different methods on **Het-FCIT** setting. The best performance is shown in bold.

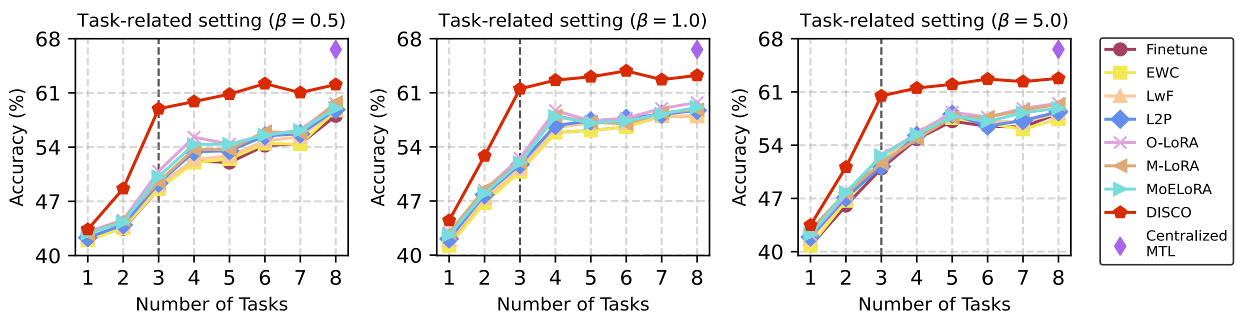

Figure 5. Performance curve of different methods in the **Het-FCIT** under varying degrees of data heterogeneity. We plot the average performance across all tasks at each stage. The black dashed line marks the stage where the model has encountered all tasks.

average result across all learned tasks after the completion of learning the final task. ***Avg*** is based on Last, which tracks the performance of the learned tasks at each stage of the continual learning process and reports the average results.

Implementation Details. We choose LLaVA-1.5-7b [32] as the base LMM due to its structural simplicity and adopt its LoRA fine-tuning strategy, training only the LoRAs and multimodal projector. The LoRA rank r at each stage is set to 8, with a learning rate of 2e-4, and it is embedded solely in the FFN layers of each block. The learning rate for the projector is set to 2e-5. For identity token extraction, we use the frozen CLIP text encoder [41]. The threshold τ and temperature coefficient ε are set to 0.9 and 0.05, respectively.

For the setting of FCIT, we set the epoch to 1 and communication rounds to 10 in each stage, and the Dirichlet distribution coefficients β are set to $\{0.5, 1.0, 5.0\}$. In each round, the global server randomly selects 5 clients from a pool of 50 to participate in the training, and we use FedAvg [38] as the base FL aggregation algorithm.

5.2. Main Results

Results are shown in Table 1, Table 2, Figure 4, and Figure 5. Under Hom-FCIT setting, our proposed DISCO achieves the best performance in both capability-related and task-related dataset settings, surpassing the best baseline by an average of **4.84%** and **1.43%** in the Last and Avg metrics, respectively. As illustrated in Figure 4, DISCO effectively mitigates the forgetting of previous tasks while learning new ones, outperforming other methods in the challenging long-stage continual learning setting.

In Het-FCIT setting, we plot the upward curve of the average performance across all tasks in Figure 5. As shown in the figure. Before completing all tasks (Left of the black dotted line), our method maintains the fastest rate of improvement, significantly outperforming other methods. In the subsequent phases (Right of the black dashed line), it continues to consolidate previously learned knowledge and further enhance performance, demonstrating strong adaptability in dynamic real-world scenarios. On the Last and Avg metrics, our method outperforms the best comparison method by an average of **2.17%** and **3.62%**, respectively.

Additionally, our method demonstrates superior robustness to varying degrees of data heterogeneity, adapting well to diverse distributions and maintaining strong performance even as heterogeneity increases, highlighting its reliability in dynamic real-world scenarios.

	Method	Hom-FCIT				Het-FCIT			
		Last	Δ	Avg	Δ	Last	Δ	Avg	Δ
(a)	Text	**56.22**	0.0	**73.03**	0.0	**63.25**	0.0	**61.99**	0.0
	Image	55.63	-0.59	72.13	-0.90	63.00	-0.25	61.80	-0.19
	Text & Image	55.96	-0.26	72.68	-0.35	63.02	-0.23	61.78	-0.21
(b)	Softmax	**56.22**	0.0	**73.03**	0.0	**63.25**	0.0	**61.99**	0.0
	Concatenate	51.74	-4.48	69.20	-3.83	60.36	-2.89	59.88	-2.11
	Cosine sim	52.83	-3.39	70.07	-2.96	60.92	-2.33	60.13	-1.86
	Argmax	55.74	-0.48	72.07	-0.96	62.88	-0.37	61.42	-0.57
(c)	FFN	**56.22**	0.0	**73.03**	0.0	**63.25**	0.0	**61.99**	0.0
	Attn	56.03	-0.19	72.88	-0.15	63.01	-0.24	61.83	-0.16
	FFN & Attn	56.30	+0.08	72.96	-0.07	63.18	-0.07	62.04	+0.05

Table 3. Ablation studies on (a) identity token extraction methods; (b) calculation of activation factors in SSA; (c) location of LoRA embedding. All experiments were conducted in the task-related setting with $\beta = 1.0$.

5.3. Ablation Study

The detailed ablation studies are provided in Table 3. In this paper, we use CLIP's text encoder to extract textual features as identity tokens. This section tests two alternatives: using only visual features from CLIP's visual encoder, and combining visual and textual features. From Table 3a, we observe that both alternatives degrade performance. This is potential because, in visual instruction tuning datasets, task similarities at the image level are higher than at the textual level (*e.g.* CLEVR-Math and super-CLEVR), making visual features less effective than textual ones.

Table 3b shows the ablation studies on how to calculate the activation factors in SSA. Direct concatenation, where the activation factors of all subspaces are set to 1.0, leads to significant performance degradation due to unfiltered outputs. Using cosine similarity (Eq.(11)) to filter the outputs also fails to solve the issue. While selecting the largest similarity (Argmax) can filter out irrelevant information, it risks fully activating a subspace that, though similar, is not directly related to the current task. This can negatively impact the output, as there are inherent similarities between the textual information of different tasks. In contrast, our method normalizes the cosine similarity to better focus on the relevant task output, achieving optimal performance.

We also test the location of LoRA embedding. As can be seen in Table 3c, embedding in the attention layer alone proved less effective than in the FFN layer. Moreover, embedding LoRA in every linear layer did not offer significant improvement and resulted in higher parameter transmission. Therefore, we chose to embed LoRA only in the FFN layer as a balanced solution. We provide more ablation and visualization results in Appendix B.

5.4. Further Analysis

Compatible with other FL algorithms. In this paper, we use the classical FedAvg as the FL algorithm for global server aggregation of local weights. Additionally, we implement other federated learning algorithms, including FedAvgM [17], FedAdam [42], FedAdagrad [42], and

Method	Hom-FCIT		Het-FCIT	
	Last	Avg	Last	Avg
FedAvg	**56.22**	73.03	**63.25**	**61.99**
FedAvgM	56.07	72.59	62.47	61.10
FedAdam	55.76	72.91	62.78	61.35
FedAdagrad	56.02	72.80	62.55	61.41
FedYogi	56.11	**73.16**	62.88	61.67

Table 4. Results of different FL algorithms. All experiments were conducted in the task-related setting with $\beta = 1.0$.

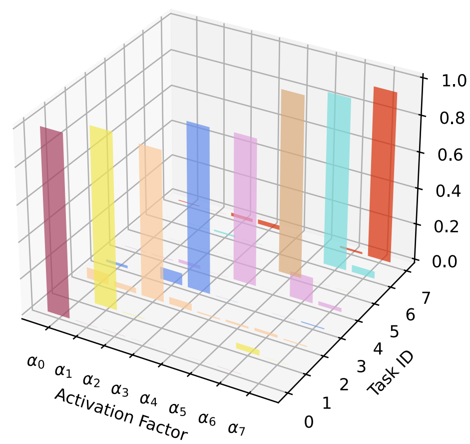

Figure 6. Visualization of activation factors during inference.

FedYogi [42], to extend our framework. As shown in Table 4, FedAvg achieves the best average performance.

Visualization of activation factors. In Figure 6, we plot the responses of activations to the test inputs of different tasks. As observed, only the activator corresponding to the current task is responsive, effectively activating the relevant subspace, while the others remain largely inhibited. This confirms the effectiveness of our proposed SSA.

6. Conclusion

We have explored for the first time the integration of federated learning and continual learning for the instruction tuning of LMMs, addressing the real-world challenge of dynamically acquiring new knowledge through distributed training resources and data. Our proposed FCIT benchmark encompasses 2 real-world scenarios, 4 distinct settings, and 12 curated datasets, providing a comprehensive evaluation of different methods. Additionally, we introduce the DISCO framework, which leverages dynamic knowledge organization (DKO) to decompose inter-task conflicts and subspace selective activation (SSA) to assign task-relevant outputs while suppressing irrelevant information. Extensive experiments demonstrate that our approach significantly improves the model's ability to learn new knowledge and handle data heterogeneity in real-world scenarios.

Acknowledgments

This work was supported by the National Science and Technology Major Project (2022ZD0116500), National Natural Science Foundation of China (62222609, 62320106010), CAS Project for Young Scientists in Basic Research (YSBR-083), Strategic Priority Research Program of Chinese Academy of Sciences under Grant (XDA0480200), Major Science and Technology Plan Project on the Future Industry Fields of Xiamen City (3502Z20241027), Unveiling and Leading Projects of Xiamen (3502Z20241011) and the InnoHK program.

References

[1] 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, 2024. 3

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1, 3

[3] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1893–1902, 2015. 1

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 3

[5] Meng Cao, Yuyang Liu, Yingfei Liu, Tiancai Wang, Jiahua Dong, Henghui Ding, Xiangyu Zhang, Ian Reid, and Xiaodan Liang. Continual llava: Continual instruction tuning in large vision-language models. *arXiv preprint arXiv:2411.02564*, 2024. 3, 4

[6] Cheng Chen, Junchen Zhu, Xu Luo, Hengtao Shen, Lianli Gao, and Jingkuan Song. Coin: A benchmark of continual instruction tuning for multimodel large language model. *arXiv preprint arXiv:2403.08350*, 2024. 3, 4, 6

[7] Haokun Chen, Yao Zhang, Denis Krompass, Jindong Gu, and Volker Tresp. Feddat: An approach for foundation model finetuning in multi-modal heterogeneous federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11285–11293, 2024. 1

[8] Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. In *16th International Workshop on Neural-Symbolic Learning and Reasoning, NeSy 2022, Windsor, UK, september 28-30, 2022*. CEUR-WS, 2022. 4

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 3

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[11] Jiahua Dong, Lixu Wang, Zhen Fang, Gan Sun, Shichao Xu, Xiao Wang, and Qi Zhu. Federated class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10164–10173, 2022. 1

[12] 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*, pages 141–159. Springer, 2024. 1, 3, 6

[13] Haiyang Guo, Fei Zhu, Fanhu Zeng, Bing Liu, and Xu-Yao Zhang. Desire: Dynamic knowledge consolidation for rehearsal-free continual learning. *arXiv preprint arXiv:2411.19154*, 2024. 2

[14] Haiyang Guo, Fanhu Zeng, Ziwei Xiang, Fei Zhu, Da-Han Wang, Xu-Yao Zhang, and Cheng-Lin Liu. Hidellava: Hierarchical decoupling for continual instruction tuning of multimodal large language model. *arXiv preprint arXiv:2503.12941*, 2025. 1

[15] Haiyang Guo, Fanhu Zeng, Fei Zhu, Jiayi Wang, Xukai Wang, Jingang Zhou, Hongbo Zhao, Wenzhuo Liu, Shijie Ma, Xu-Yao Zhang, et al. A comprehensive survey on continual learning in generative models. *arXiv preprint arXiv:2506.13045*, 2025. 1

[16] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021. 4

[17] 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. 8

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[19] Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. Rethinking federated learning with domain shift: A prototype view. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16312–16322. IEEE, 2023. 4

[20] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 1, 3

[21] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2018. 4

[22] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Fig-

ureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017. 4

[23] Gyuhak Kim, Changnan Xiao, Tatsuya Konishi, and Bing Liu. Learnability and algorithm for continual learning. In *International Conference on Machine Learning*, pages 16877–16896. PMLR, 2023. 2

[24] 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. 6

[25] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 4

[26] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*, 2024. 4

[27] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. In *2022 IEEE 38th international conference on data engineering (ICDE)*, pages 965–978. IEEE, 2022. 2, 6

[28] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020. 1

[29] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 6

[30] Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14963–14973, 2023. 4

[31] Yi-Ming Lin, Yuan Gao, Mao-Guo Gong, Si-Jia Zhang, Yuan-Qiao Zhang, and Zhi-Yuan Li. Federated learning on multimodal data: A comprehensive survey. *Machine Intelligence Research*, 20(4):539–553, 2023. 1

[32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 3, 7

[33] Wenzhuo Liu, Fei Zhu, Haiyang Guo, Longhui Wei, and Cheng-Lin Liu. Llava-c: Continual improved visual instruction tuning. *arXiv preprint arXiv:2506.08666*, 2025. 1

[34] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 1, 3

[35] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2021. 4

[36] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2022. 4

[37] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 4

[38] 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*, pages 1273–1282. PMLR, 2017. 1, 5, 7

[39] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 4

[40] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 123 (1):74–93, 2017. 4

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 5, 7

[42] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020. 8

[43] Lorenzo Sani, Alex Iacob, Zeyu Cao, Bill Marino, Yan Gao, Tomas Paulik, Wanru Zhao, William F Shen, Preslav Aleksandrov, Xinchi Qiu, et al. The future of large language model pre-training is federated. *arXiv preprint arXiv:2405.10853*, 2024. 1

[44] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022. 4

[45] Donald Shenaj, Marco Toldo, Alberto Rigon, and Pietro Zanuttigh. Asynchronous federated continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5055–5063, 2023. 3

[46] Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Xiangyang Liu, Hang Yan, Yunfan Shao, Qiong Tang, Shiduo Zhang, et al. Moss: An open conversational large language model. *Machine Intelligence Research*, 21(5):888–905, 2024. 1

[47] 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, pages 8432–8440, 2022. 4

[48] Yue Tan, Guodong Long, Jie Ma, Lu Liu, Tianyi Zhou, and Jing Jiang. Federated learning from pre-trained models: A contrastive learning approach. *Advances in neural information processing systems*, 35:19332–19344, 2022. 4

[49] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1, 3

[50] Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning. *arXiv preprint arXiv:2310.14152*, 2023. 5, 6

[51] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 139–149, 2022. 6

[52] Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, Hongyi Wang, Lingjuan Lyu, and Ang Li. Flora: Federated fine-tuning large language models with heterogeneous low-rank adaptations. *arXiv preprint arXiv:2409.05976*, 2024. 5

[53] Taiqiang Wu, Jiahao Wang, Zhe Zhao, and Ngai Wong. Mixture-of-subspaces in low-rank adaptation. *arXiv preprint arXiv:2406.11909*, 2024. 5

[54] Xin Yang, Hao Yu, Xin Gao, Hao Wang, Junbo Zhang, and Tianrui Li. Federated continual learning via knowledge fusion: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2024. 3

[55] Rui Ye, Wenhao Wang, Jingyi Chai, Dihan Li, Zexi Li, Yinda Xu, Yaxin Du, Yanfeng Wang, and Siheng Chen. Openfedllm: Training large language models on decentralized private data via federated learning. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6137–6147, 2024. 1, 3, 4

[56] Fanhu Zeng, Fei Zhu, Haiyang Guo, Xu-Yao Zhang, and Cheng-Lin Liu. Modalprompt: Dual-modality guided prompt for continual learning of large multimodal models. *arXiv preprint arXiv:2410.05849*, 2024. 1, 3, 4

[57] Jie Zhang, Chen Chen, Weiming Zhuang, and Lingjuan Lyu. Target: Federated class-continual learning via exemplar-free distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4782–4793, 2023. 1

[58] Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Tong Yu, Guoyin Wang, and Yiran Chen. Towards building the federatedgpt: Federated instruction tuning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6915–6919. IEEE, 2024. 1, 3, 4

[59] Xu-Yao Zhang, Cheng-Lin Liu, and Ching Y Suen. Towards robust pattern recognition: A review. *Proceedings of the IEEE*, 108(6):894–922, 2020. 1

[60] Hongbo Zhao, Fei Zhu, Rundong Wang, Gaofeng Meng, and Zhaoxiang Zhang. Mllm-cl: Continual learning for multimodal large language models. *arXiv preprint arXiv:2506.05453*, 2025. 1

[61] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3

[62] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5871–5880, 2021. 6

[63] Fei Zhu, Shijie Ma, Zhen Cheng, Xu-Yao Zhang, Zhaoxiang Zhang, and Cheng-Lin Liu. Open-world machine learning: A review and new outlooks. *arXiv preprint arXiv:2403.01759*, 2024. 1