

Federated Continual Instruction Tuning

Supplementary Material

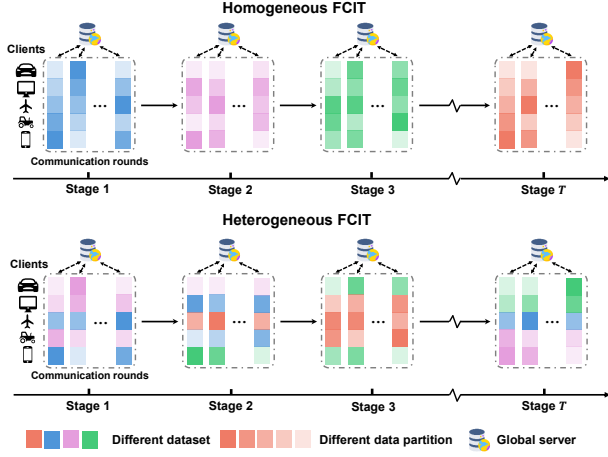


Figure A1. Illustration of Homogeneous FCIT and Heterogeneous FCIT settings.

A. More Details of FCIT Benchmark

A.1. Illustration display of FCIT

To better understand the proposed Homogeneous FCIT and Heterogeneous FCIT settings, we further provide an illustration detailing their entire process.

As shown in Figure A1, we use different colors to distinguish datasets and assume the continual learning stage lasts for T steps. In **Homogeneous FCIT** setting, the own data of the selected local clients in each stage belong to the same dataset, which exists in different proportions in different clients. Each stage comprises multiple communication rounds (set to 10 in our experiments), during which clients train the model on their own data (with 1 epoch per round), upload weights to the global server for aggregation, and receive the aggregated weights for the next round. Homogeneous FCIT setting extends continual instruction tuning to a federated learning setting with a non-IID data distribution, posing greater challenges for traditional methods.

Heterogeneous FCIT setting extends the former by allowing each client’s data to come from different datasets in each stage. This requires the global server to mitigate catastrophic forgetting across stages while resolving conflicts among datasets within the same stage. This setting is common in real-world scenarios, such as healthcare systems that need to simultaneously manage multiple disease outbreaks while continuously updating to track their progression and mitigate social risks, and our benchmark effectively addresses this real-world need, providing a comprehensive evaluation framework for such dynamic challenges.

A.2. Visualization of data heterogeneous

In federated learning tasks, data heterogeneity poses a core challenge in distributed training, primarily manifesting as varying proportions of private data across different clients. Therefore, we employ the Dirichlet distribution, a common approach in FL tasks, to model distributional differences among clients. In this paper, we use β to control the degree of distributional variation, as visualized in Figure A2.

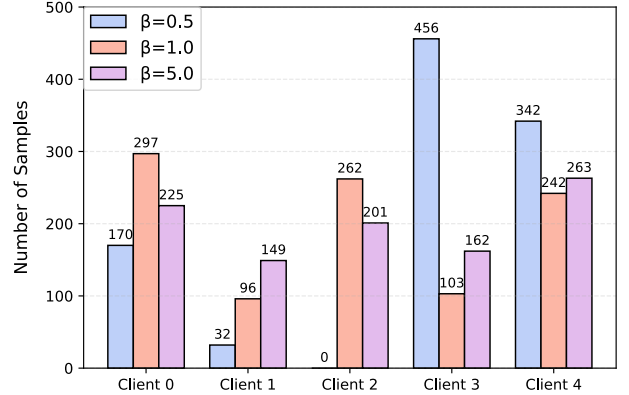


Figure A2. Visualization of Dirichlet distribution.

It can be seen that a smaller β leads to greater disparities in data distribution among clients, resulting in more extreme heterogeneity, whereas a larger value of β indicates a more uniform distribution.

A.3. Visualization of the dataset

Table A1 and Table A2 present the input images and instruction formats of the 12 selected datasets, which exhibit relatively low average zero-shot performance on the base model LLaVA-v1.5-7B, approximately **30%** lower than their fine-tuning performance (See Zero-shot and Centralized MTL in Table 4). This better ensures that these datasets remain unseen or unfamiliar to the base model during training, thereby reducing information leakage.

In our experiments, we design two dataset-level settings: Capability-related and Task-related. The capability-related setting categorizes the 12 datasets into four dimensions: general, math, chart, and other, where each capability consists of a mix of relevant datasets. The task-related setting selects 8 out of the 12 datasets for different stages of continual learning, evaluating how well different approaches mitigate forgetting over a long-stage learning setup. The specific data composition of these two settings can be found in Section 3.2.


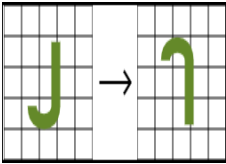

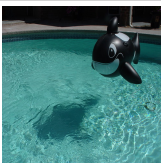
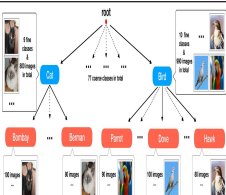
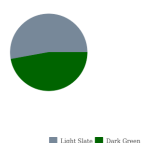
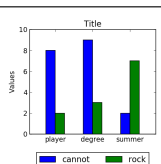
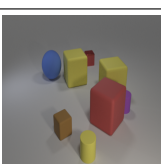
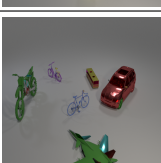
Dataset	Visual input	Question	Response
A-OKVQA		What is the man by the bags awaiting? 0. skateboarder 1. train 2. delivery 3. cab Answer with the option's letter from the given choices directly.	3
IconQA		What has been done to this letter? 0. turn 1. slide 2. flip Answer with the option's letter from the given choices directly.	2
Grounding		Please provide the bounding box coordinate of the region this sentence describes: kid on right teddy bear bib.	[0.65,0.44,0.88,0.98]
ImageNet-R		Question: What is the object in the image? Answer the question using a single word or phrase.	Killer whale
ArxivQA		How many coarse classes are represented in the figure? A) Less than 50 B) Exactly 77 C) More than 100 D) Exactly 99 Answer with the option's letter from the given choices directly.	B
FigureQA		Please answer the question and provide the correct option letter at the end. Question: Is Light Slate greater than Dark Green? Choices: (A) no (B) yes	B
DVQA		Please answer the question and provide the final answer at the end. Question: What is the value of the largest individual bar in the whole chart?	9
CLEVR-Math		Subtract all brown matte objects. Subtract all blue cylinders. How many objects are left? Answer the question using a single word or phrase.	7
super-CLEVR		Question: There is a matte thing that is in front of the small purple utility bike and behind the red metal thing; how big is it? Answer the question using a single word or phrase.	small

Table A1. Visualization of input images and instruction formats for each dataset in FCIT.


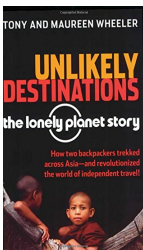
Dataset	Visual input	Question	Response											
TabMWP	Plants per garden	Question: The members of the local garden club tallied the number of plants in each person's garden. How many gardens have at least 47 plants? Answer the question using a single word or phrase.	13											
	<table><tr><th>Stem</th><th>Leaf</th></tr><tr><td>3</td><td>3 3 3 5 5</td></tr><tr><td>4</td><td>6</td></tr><tr><td>5</td><td>4 5 7 8</td></tr><tr><td>6</td><td>7 8</td></tr><tr><td>7</td><td>2 3 7 9</td></tr><tr><td>8</td><td>6 8 9</td></tr></table>			Stem	Leaf	3	3 3 3 5 5	4	6	5	4 5 7 8	6	7 8	7
Stem	Leaf													
3	3 3 3 5 5													
4	6													
5	4 5 7 8													
6	7 8													
7	2 3 7 9													
8	6 8 9													
Flickr30k		What is happening in the image? Generate a brief caption for the image.	Three older women are at a restaurant talking with other people											
OCR-VQA		Who wrote this book? Answer the question using a single word or phrase.	Tony Wheeler											

Table A2. Visualization of input images and instruction formats for each dataset in FCIT.

B. More Details of Experiments

B.1. Details of the comparison method

In this section, we present the underlying principles of the baseline methods used in our experiments.

LwF mitigates forgetting by applying knowledge distillation loss during new task learning. It preserves past knowledge by extracting soft labels from the frozen old model’s outputs and constraining the new model’s outputs to remain close, minimizing deviation from previous tasks.

EWC mitigates forgetting by restricting updates to important parameters of previous tasks. It computes parameter importance using the Fisher information matrix and penalizes significant changes, preserving past knowledge while learning new tasks.

L2P introduces a dynamic prompts pool, enabling the model to select and optimize relevant prompts based on similarity during training. Additionally, it applies a regularization loss to encourage task-specific prompt selection, mitigating catastrophic forgetting.

O-LoRA imposes an orthogonality constraint in parameter space, ensuring that the optimization of the current task occurs in a direction orthogonal to previous tasks, thereby minimizing task conflicts. During inference, it aggregates learned knowledge by concatenating the LoRA modules of all tasks along the specified dimension.

M-LoRA trains LoRA modules separately at each stage and mitigates forgetting by spatially merging them in parameter space during inference. Unlike O-LoRA, it does not incur additional memory overhead during training.

MoELoRA transforms the fine-tuning of individual LoRA into a Mixture-of-Experts framework, where a predefined set of LoRA modules serves as expert heads. During training, routers are optimized alongside expert selection, aiming to assign different tasks to distinct expert heads. This structured allocation helps mitigate forgetting by ensuring the router effectively distributes outputs across expert modules.

B.2. Details of evaluation

In our benchmark, tasks have different output formats, requiring tailored accuracy evaluation methods. For tasks with answering a single option or single word, we determine correctness using `pred.upper()` in `Response.upper()`. For captioning tasks, we adopt standard image captioning metrics, including `Bleu_1`, `Bleu_2`, `Bleu_3`, `Bleu_4`, `METEOR`, `ROUGE_L`, and `CIDEr`. The final results are computed as the average of these seven metrics.

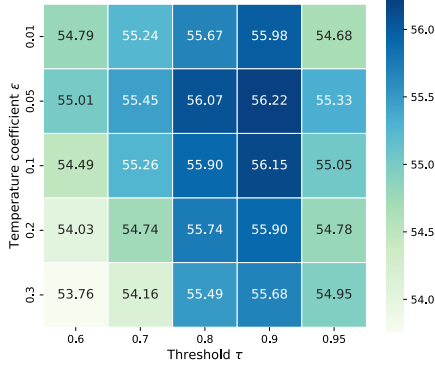


Figure B1. Ablation study of hyper-parameters in Hom-FCIT and task-related setting. The partition β is set to 1.0.

B.3. Ablation study of hyper-parameters

In this section, we conduct more ablation experiments on two key hyperparameters, the threshold τ and the temperature coefficient ϵ , with results shown in Figure B1. It can be seen that both excessively large and small temperature coefficients significantly affect the results. A larger coefficient leads to overly sharp activation assignments, increasing the likelihood of selecting mismatched subspaces, while a smaller coefficient incorporates excessive information from irrelevant subspaces, ultimately degrading model performance. Similarly, an excessively large threshold may filter out knowledge relevant to the same task, while a too-small threshold may misassign identity tokens to other subspaces, both leading to degraded performance.

Threshold τ	≤ 0.5	0.6	0.7	0.75	0.8	0.9	0.95
Num. of Subspace	1	3	6	7	8	8	10

Table B3. The effect of threshold selection on the number of subspaces in Hom-FCIT setting (task-related, $\beta = 1.0$).

We further investigate the effect of threshold selection on the number of subspaces formed. As shown in Table B3, a threshold that is too low fails to effectively separate task-specific knowledge, leading to subspace aggregation across tasks and potential knowledge conflicts. Conversely, an overly high threshold may split knowledge that belongs to the same subspace, resulting in redundant branches and degraded model performance.

B.4. Comparison on a single dataset

We further compare single-task performance between our proposed DISCO and Finetune. As shown in Figure B2 (Left), for the first learned task, our method significantly mitigates forgetting. Under Het-FCIT (Figure B2 Right), for task-specific OCR-VQA, DISCO not only enhances knowledge retention during learning but also maintains strong performance even when the task is absent. In contrast, Finetune suffers from severe inter-task conflicts,

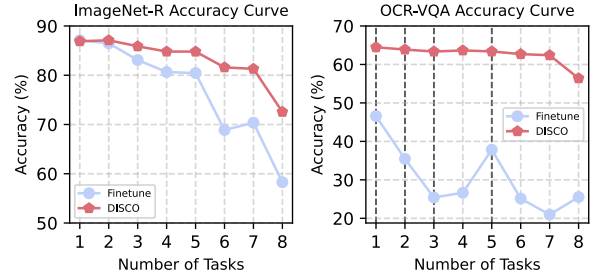


Figure B2. **Left.** Performance curve of first learned task ImageNet-R under Hom-FCIT and task-related settings; **Right.** Performance curve of OCR-VQA under Het-FCIT and task-related settings. The black dashed line indicates the stage where the model has learned OCR-VQA.

leading to continuous performance degradation even while learning the dataset (2-nd and 3-rd black dashed lines).

B.5. Evaluation of zero-shot capability and general benchmark performance

In addition to mitigating forgetting, we also evaluate the generalization ability of different methods on unseen tasks, as well as their performance on a general LMM benchmark. As shown in Table B4, our method improves zero-shot transfer performance on tasks without downstream supervision. Moreover, it minimizes negative transfer effects on the general LMM benchmark, demonstrating strong robustness and adaptability.

Hom-FCIT $\beta = 1.0$	Zero-shot capability				General LMM benchmark			
	Task2	Task3	Task4	Avg	MME	POPE	MMBench	SEED
LLaVA-1.5	46.97	17.05	27.13	30.38	1476.9	86.4	66.1	60.1
O-LoRA	41.89	16.04	26.88	28.27	1354.6	80.7	59.4	59.0
M-LoRA	45.33	15.90	<u>31.26</u>	<u>30.83</u>	1376.8	79.8	60.1	59.5
MoELoRA	43.16	16.86	29.09	29.70	1358.3	82.8	60.5	59.2
DISCO	<u>46.14</u>	<u>17.00</u>	36.09	33.08	<u>1436.6</u>	<u>83.9</u>	<u>62.4</u>	<u>60.1</u>

Table B4. Zero-shot transfer and general LMM benchmark results in the **Hom-FCIT** setting (capability-related, $\beta = 1.0$).

B.6. Efficiency and storage analysis.

Methods	Efficiency		LoRA memory cost			
	Speed (\uparrow)	FLOPs (\downarrow)	Hom-FCIT	Δ	Het-FCIT	Δ
Finetune	3.46 it/s	8.55 T	22.2 M	0.16%	22.2 M	0.16%
O-LoRA	3.42 it/s	<u>9.35 T</u>	<u>88.7 M</u>	<u>0.63%</u>	<u>177.0 M</u>	<u>1.26%</u>
MoELoRA	3.36 it/s	9.77 T	93.4 M	0.68%	186.0 M	1.33%
DISCO	<u>3.45 it/s</u>	<u>9.35 T</u>	88.9 M	<u>0.63%</u>	178.0 M	<u>1.26%</u>

Table B5. All efficiency comparisons are conducted under identical conditions. Δ : relative percentage to the backbone model.

As shown in Table B5, our method demonstrates the most substantial performance gains compared to fine-tuning and other baselines, while preserving competitive inference efficiency and requiring only modest additional LoRA storage overhead.

B.7. Detailed Results of DISCO.

DISCO	General	Other	Chart	Math
Task1	71.26			
Task2	67.72	57.23		
Task3	64.47	52.75	50.10	
Task4	62.92	53.14	46.43	56.15

Table B6. Results matrix of DISCO in Hom-FCIT setting (capability-related, $\beta = 1.0$)

DISCO	General	Other	Chart	Math
Task1	69.87	55.51		
Task2	66.46	54.75	47.81	51.87
Task3	68.73	56.30	50.00	54.47
Task4	68.01	55.49	54.19	57.88

Table B7. Results matrix of DISCO in Het-FCIT setting (capability-related, $\beta = 1.0$)

DISCO	ImageNet-R	ArxivQA	IconQA	CLEVR-Math	OCRVQA	Flickr30k	FigureQA	super-CLEVR
Task1	86.90							
Task2	87.10	93.40						
Task3	85.88	93.35	65.47					
Task4	84.81	93.79	58.70	60.28				
Task5	84.80	93.96	59.14	60.84	63.76			
Task6	81.60	93.02	58.62	56.24	37.22	54.57		
Task7	81.30	92.45	58.27	56.03	25.82	54.46	43.70	
Task8	72.56	92.34	55.59	47.72	35.97	52.49	42.92	50.16

Table B8. Results matrix of DISCO in Hom-FCIT setting (task-related, $\beta = 1.0$)

DISCO	ImageNet-R	ArxivQA	IconQA	CLEVR-Math	OCRVQA	Flickr30k	FigureQA	super-CLEVR
Task1		87.99	58.37		64.76	53.39	41.52	
Task2		88.35	54.87	53.68	64.35	53.52	42.08	47.04
Task3	76.46	88.1	61.35	57.22	64.24	55.98	41.85	46.88
Task4	84.45	89.28	61.27	55.33	64.04	55.94	40.75	50.00
Task5	83.90	90.45	63.83	58.23	64.07	53.97	39.92	50.18
Task6	85.46	90.71	64.14	62.53	63.73	53.68	40.23	50.14
Task7	84.75	90.65	68.36	60.28	63.17	42.90	40.38	51.08
Task8	73.95	91.75	68.33	59.55	62.52	55.18	43.55	51.16

Table B9. Results matrix of DISCO in Het-FCIT setting (task-related, $\beta = 1.0$)