# SMoLoRA: Exploring and Defying Dual Catastrophic Forgetting in Continual Visual Instruction Tuning

Supplementary Material

## A. Effect of Weight Ratio

To further investigate the impact of the weight ratio $[\alpha, \beta]$ within the adaptive fusion module, we randomly sampled the values of $\alpha$ and $\beta$ during inference across different layers of the fine-tuned model, as illustrated in Fig. 2. The analysis reveals that $\alpha$ and $\beta$ exhibit significant variations across layers. Furthermore, we computed the average values of $\alpha$ and $\beta$ for all layers. The resulting ratios indicate that the instruction following module plays a more pivotal role in the routing process compared to the visual understanding module.

## B. Results at Different Stages

As shown in Table 1-2, we present the experimental results of SeqLoRA and our SMoLoRA method at various stages from ScienceQA to VQAv2 on the CVIT benchmark. The results demonstrate that with each additional training stage, our approach consistently maintains stable and superior performance across different datasets.

## C. More Experiments on MLLMs

To further assess the versatility of our SMoLoRA method, we implemented it on another advanced MLLM, MiniGPT-4 [12], as shown in Table 3. While MiniGPT-4 exhibits a lower degree of forgetting compared to LLAVA, our method still yields significant overall performance enhancements compared to SeqLoRA. Notably, on the MIF metric, we observed improvements of 17.98% and 20.90% in single- and multi-type instruction settings, respectively. These results underscore the effectiveness of our method in mitigating forgetting in instruction following.

## D. Details of CVIT Benchmark

Table 4 provides an overview of the CVIT benchmark, detailing the instruction templates for each dataset and the number of samples.

**ScienceQA [6]:** ScienceQA is a comprehensive dataset consisting of science-related questions and answers designed to evaluate and enhance the reasoning and problem-solving capabilities of AI models in scientific domains.

**TextVQA [10]:** TextVQA is a visual question answering dataset that focuses on questions requiring models to read and understand text embedded within images to provide accurate answers.

**Flickr30k [8]:** Flickr30k is a large-scale image dataset containing over 30,000 photos sourced from Flickr, each annotated with multiple descriptive captions, widely used for training and evaluating image captioning and vision-language models.

**ImageNet [1]:** ImageNet is a large-scale visual dataset containing millions of annotated images across thousands of object categories, widely used for training and evaluating computer vision models.

**GQA [5]:** GQA is a visual question-response dataset comprising complex compositional questions about images, designed to evaluate and enhance AI models' reasoning and relational understanding capabilities in interpreting visual content.

**VQAv2 [2]:** VQAv2 is a widely-used visual question answering dataset that consists of images paired with diverse questions and multiple corresponding answers, designed to assess and enhance the ability of models to understand and reason about visual information.

**VizWiz [3]:** VizWiz originates from real-world visual question answering scenarios in which visually impaired individuals capture images and pose spoken questions about them. Each visual question is accompanied by 10 crowd-sourced answers, facilitating the development of assistive technologies for the visually impaired community.

**TextCaps [9]:** TextCaps is an image captioning dataset that requires models to generate descriptive captions by reading and interpreting text embedded within images, thereby enhancing the ability to incorporate textual information into visual descriptions.

**OCRVQA [7]:** OCRVQA is a visual question answering dataset designed to evaluate the ability of the models to read and comprehend text embedded within images to accurately answer related questions.

**Places365 [11]:** Places365 is a large-scale scene recognition dataset comprising over 1.8 million images across 365 diverse scene categories, widely used for training and evaluating computer vision models in understanding and classifying various environments.

## E. Additional Case Studies

As shown in Fig. 2, we present additional cases to validate the efficacy of our proposed SMoLoRA. Individual modules (VU and IF) are only capable of resolving their respective forgetting problems, whereas the combination of both can simultaneously mitigate the dual catastrophic forgetting.

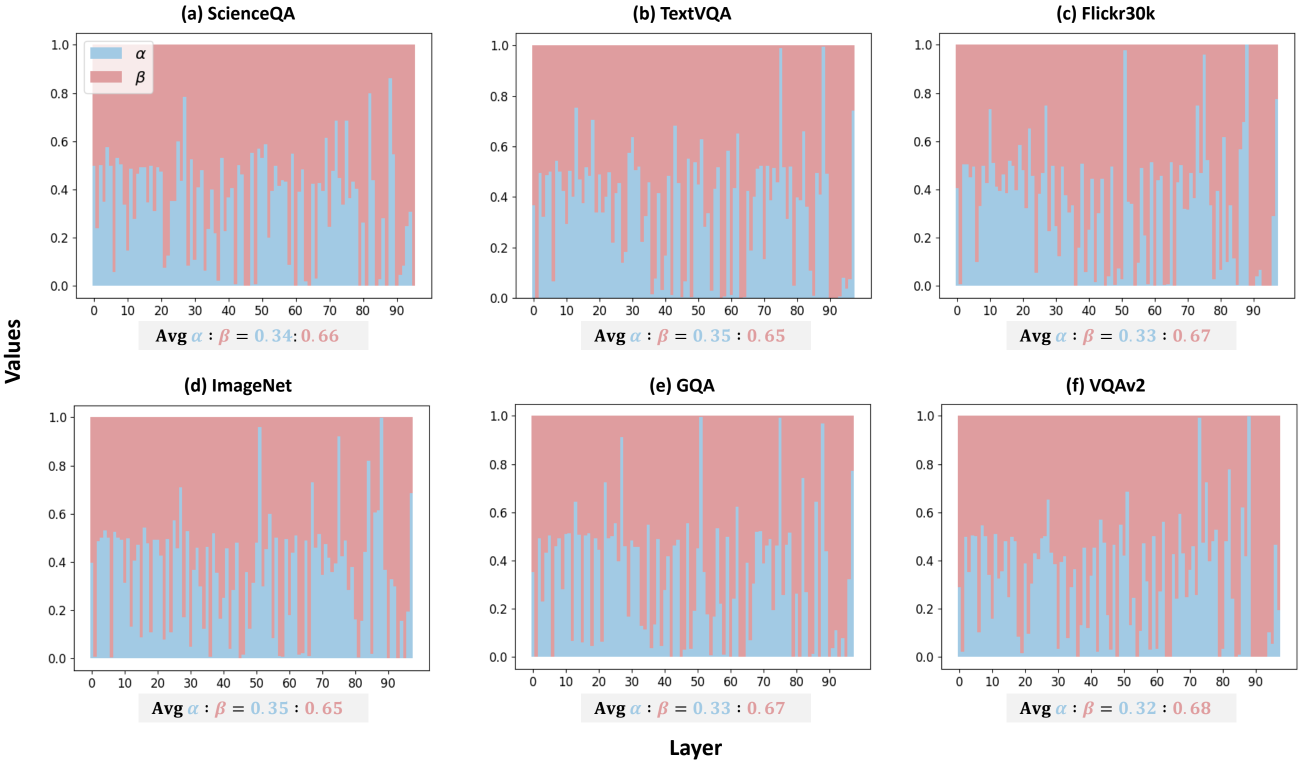


Figure 1. Distribution of weight ratio $[\alpha, \beta]$ across different layers. The visualization results show that the instruction following module ($\beta$) plays a more critical role in the routing process than the visual understanding module ($\alpha$).

Table 1. The evaluated results (%) of **SeqLoRA** for upstream continual learning **at different stages** of our CVIT benchmark. Each row represents a different training stage.

| | ScienceQA | TextVQA | Flickr30k | ImageNet | GQA | VQAv2 |
|---|---|---|---|---|---|---|
| **Single-type** | 83.75 | | | | | |
| | 66.21 | 49.95 | | | | |
| | 68.78 | 19.35 | 166.33 | | | |
| | 43.90 | 0.05 | 0.26 | 95.45 | | |
| | 53.29 | 34.74 | 25.76 | 11.84 | 57.69 | |
| | 55.31 | 50.22 | 33.89 | 22.73 | 50.52 | 64.37 |
| **Multi-type** | 83.85 | | | | | |
| | 69.18 | 50.24 | | | | |
| | 51.38 | 0.00 | 156.85 | | | |
| | 37.77 | 0.01 | 0.13 | 95.98 | | |
| | 53.01 | 33.78 | 3.77 | 10.18 | 58.01 | |
| | 59.21 | 50.80 | 20.99 | 20.30 | 49.98 | 64.41 |

Table 2. The evaluated results (%) of our **SmoLoRA** for upstream continual learning **at different stages** of our CVIT benchmark. Each row represents a different training stage.

| | ScienceQA | TextVQA | Flickr30k | ImageNet | GQA | VQAv2 |
|---|---|---|---|---|---|---|
| **Single-type** | 83.85 | | | | | |
| | 80.71 | 61.05 | | | | |
| | 81.99 | 61.20 | 150.72 | | | |
| | 73.80 | 51.90 | 140.71 | 96.28 | | |
| | 74.98 | 44.87 | 137.08 | 95.45 | 59.19 | |
| | 77.36 | 58.29 | 151.99 | 95.35 | 51.96 | 65.71 |
| **Multi-type** | 84.53 | | | | | |
| | 80.71 | 61.24 | | | | |
| | 81.58 | 60.24 | 162.78 | | | |
| | 74.44 | 44.28 | 136.92 | 96.14 | | |
| | 78.09 | 45.31 | 133.03 | 95.09 | 59.96 | |
| | 80.50 | 58.30 | 146.63 | 94.28 | 52.42 | 65.96 |

Table 3. The evaluated results (%) on upstream continual learning for our CVIT benchmark using **MiniGPT-4** after tuning on the final task.

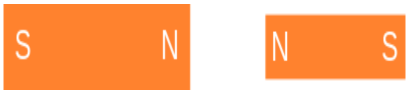
| **Method** | | **Accuracy on Each Task** | | | | | | **Overall Results** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ScienceQA | TextVQA | Flickr30k | ImageNet* | GQA | VQAv2 | AP ↑ | MAP ↑ | BWT ↑ | MIF ↑ |
| **Single-type** | Zero-shot | 41.10 | 0.00 | 0.03 | 26.32 | 0.00 | 0.00 | 11.24 | - | - | 0.96 |
| | SeqLoRA [4] | **54.56** | 36.18 | 134.83 | 38.93 | 40.85 | 34.16 | 56.59 | 56.78 | **9.22** | 58.42 |
| | **SMoLoRA(Ours)** | 54.16 | **38.61** | **135.60** | **48.46** | **44.11** | **48.35** | **61.55** | **62.05** | 3.39 | **76.40** |
| **Multi-type** | Zero-shot | 42.72 | 0.00 | 0.01 | 26.83 | 0.00 | 0.00 | 11.59 | - | - | 1.18 |
| | SeqLoRA [4] | 54.61 | 34.24 | **116.09** | 30.32 | 40.65 | 28.91 | 50.80 | 51.76 | -2.37 | 54.49 |
| | **SMoLoRA(Ours)** | **57.72** | **39.60** | 112.43 | **43.27** | **44.65** | **38.43** | **56.02** | **60.67** | **1.84** | **75.39** |

Table 4. Details of datasets in our CVIT Benchmark.

| **Dataset** | **Instruction template** | **Train Number** | **Test Number** |
|---|---|---|---|
| **ScienceQA** | <question><instruction>: Answer with the option's letter from the given choices directly.<br><question><instruction>: Select the correct answer by choosing the corresponding letter from the options provided.<br><question><instruction>: Select the correct letter from the given options to answer the question.<br><question><instruction>: Identify the correct answer by choosing the appropriate letter from the choices.<br><question><instruction>: Pick the correct answer by selecting the letter associated with the correct choice. | **12726** | **4241** |
| **TextVQA** | <question><instruction>: Answer using only one word or a short, descriptive phrase.<br><question><instruction>: Use a single word or a short phrase to respond to the question.<br><question><instruction>: Use one word or a concise phrase to respond to the question.<br><question><instruction>: Answer the question with just one word or a brief phrase.<br><question><instruction>:Answer the question with a single word or a brief, descriptive phrase. | **34602** | **5000** |
| **Flickr30k** | <instruction>: What is depicted in the displayed picture? Summarize it using a single, concise sentence.<br><instruction>: What is happening in the presented picture? Please describe it in one complete sentence.<br><instruction>: What does the image display clearly and succinctly? Provide a full sentence explaining it.<br><instruction>: How would you interpret the scene in the picture? Express your answer in one informative sentence.<br><instruction>: What is the captured scene about? Explain it clearly in one simple sentence. | **145000** | **1014** |
| **ImageNet** | <instruction>: What is the main object present in the image? Provide your answer using a word or brief phrase.<br><instruction>: Which specific object does the image depict? Give your answer in one word or a short phrase.<br><instruction>: What category does the object in the image belong to? Answer using a single word or phrase.<br><instruction>: What is the object in the image? Answer briefly with a word or a short phrase.<br><instruction>: What is the primary object visible in the image? Answer briefly with a word or a short phrase. | **117715** | **5050** |
| **GQA** | <question><instruction>: Answer using only one word or a short, descriptive phrase.<br><question><instruction>: Use a single word or a short phrase to respond to the question.<br><question><instruction>: Use one word or a concise phrase to respond to the question.<br><question><instruction>: Answer the question with just one word or a brief phrase.<br><question><instruction>:Answer the question with a single word or a brief, descriptive phrase. | **72140** | **12578** |
| **VQAv2** | <question><instruction>: Answer using only one word or a short, descriptive phrase.<br><question><instruction>: Use a single word or a short phrase to respond to the question.<br><question><instruction>: Use one word or a concise phrase to respond to the question.<br><question><instruction>: Answer the question with just one word or a brief phrase.<br><question><instruction>:Answer the question with a single word or a brief, descriptive phrase. | **82783** | **53588** |
| **VizWiz** | <question><instruction>: Answer using only one word or a short, descriptive phrase.<br><question><instruction>: Use a single word or a short phrase to respond to the question.<br><question><instruction>: Use one word or a concise phrase to respond to the question.<br><question><instruction>: Answer the question with just one word or a brief phrase.<br><question><instruction>:Answer the question with a single word or a brief, descriptive phrase. | **0** | **4319** |
| **TextCaps** | <instruction>: What is depicted in the displayed picture? Summarize it using a single, concise sentence.<br><instruction>: What is happening in the presented picture? Please describe it in one complete sentence.<br><instruction>: What does the image display clearly and succinctly? Provide a full sentence explaining it.<br><instruction>: How would you interpret the scene in the picture? Express your answer in one informative sentence.<br><instruction>: What is the captured scene about? Explain it clearly in one simple sentence. | **0** | **15830** |
| **OCRVQA** | <question><instruction>: Answer using only one word or a short, descriptive phrase.<br><question><instruction>: Use a single word or a short phrase to respond to the question.<br><question><instruction>: Use one word or a concise phrase to respond to the question.<br><question><instruction>: Answer the question with just one word or a brief phrase.<br><question><instruction>:Answer the question with a single word or a brief, descriptive phrase. | **0** | **99926** |
| **Places365** | <instruction>: What is the background of the image? Answer the question using a single word or phrase.<br><instruction>: What is the background depicted in the image? Provide your answer using a word or brief phrase.<br><instruction>: Which type of background does the image show? Give your answer in one word or a short phrase.<br><instruction>: What category best describes the background in the image? Answer using a brief phrase or single word.<br><instruction>: What is the primary background visible in the image? Answer briefly with a word or a short phrase. | **5 & 10** | **36500** |

a)

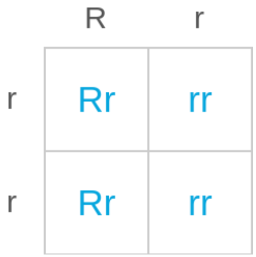
**User:** Two magnets are placed as shown. Will these magnets attract or repel each other?
A. repel B. attract
*Answer with the option's letter from the given choices directly.*



**w/ VU module:** repel
**w/ IF module:** B
**w/ VU+IF module(Ours):** A
**Ground Truth:** A

b)

**User:** In a group of cockatiels, some individuals have bright orange cheeks and others have pale orange cheeks. In this group, the gene for the cheek color trait has two alleles. The allele for bright orange cheeks (R) is dominant over the allele for pale orange cheeks (r). This Punnett square shows a cross between two cockatiels. What is the probability that a cockatiel produced by this cross will have bright orange cheeks?
A. 2/4 B. 1/4 C. 4/4 D. 0/4 E. 3/4
*Answer with the option's letter from the given choices directly.*

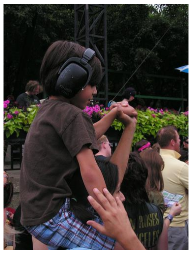


|   | R | r |
|---|---|---|
| r | Rr | rr |
| r | Rr | rr |

**w/ VU module:** 2/4
**w/ IF module:** E
**w/ VU+IF module(Ours):** A
**Ground Truth:** A

c)

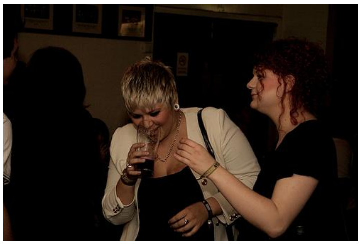
**User:** What is happening in the presented picture?
*Please describe it in one complete sentence.*



**w/ VU module:** Boy is wearing headphones
**w/ IF module:** A boy in a brown shirt is being held up by a man.
**w/ VU+IF module(Ours):** A boy wearing headphones is being held by a woman.
**Ground Truth:** A boy wearing headphones sits on a woman's shoulders.

d)

**User:** What is happening in the presented picture?
*Please describe it in one complete sentence.*



**w/ VU module:** Woman drinking
**w/ IF module:** A woman in a black shirt is talking to a woman in a black jacket.
**w/ VU+IF module(Ours):** A woman in a white jacket is drinking from a glass.
**Ground Truth:** A woman with blond-hair is drinking out of a glass.

Figure 2. Additional Case Studies on the Effectiveness of Separation in SMoLoRA.