

Continual Learning in Vision-Language Models via Aligned Model Merging

Ghada Sokar¹, Gintare Karolina Dziugaite¹, Anurag Arnab¹, Ahmet Iscen¹, Pablo Samuel Castro¹ and

Cordelia Schmid¹

¹Google DeepMind

Continual learning is conventionally tackled through sequential fine-tuning, a process that, while enabling adaptation, inherently favors plasticity over the stability needed to retain prior knowledge. While existing approaches attempt to mitigate catastrophic forgetting, a bias towards recent tasks persists as they build upon this sequential nature. In this work we present a new perspective based on model merging to maintain stability while still retaining plasticity. Rather than just sequentially updating the model weights, we propose merging newly trained task parameters with previously learned ones, promoting a better balance. To maximize the effectiveness of the merging process, we propose a simple mechanism that promotes learning aligned weights with previous ones, thereby avoiding interference when merging. We evaluate this approach on large Vision-Language Models (VLMs), and demonstrate its effectiveness in reducing forgetting, increasing robustness to various task orders and similarities, and improving generalization.

1. Introduction

Vision-language models (VLMs) have achieved remarkable success by leveraging pretraining on large-scale datasets ([Devlin, 2018](#); [Radford et al., 2021](#)). Fine-tuning these pretrained models on specific downstream tasks yields task-specific variants with improved performance. This approach, however, while effective for single-task optimization, falters in continual learning (CL) scenarios with a stream of diverse downstream tasks. Indeed, continual learning with large models poses two main challenges: (1) the model size, often reaching billions of parameters, can be prohibitively expensive for continual full fine-tuning; and (2) fine-tuning on a new task often leads to catastrophic forgetting ([Robins, 1995](#)), evidenced through performance degradation on previous tasks.

Parameter-efficient fine-tuning (PEFT) methods have emerged as an effective solution for addressing the first challenge – the prohibitive cost of continually fine-tuning large models – primarily in single-task settings ([Houlsby et al., 2019](#); [Hu et al., 2022](#); [Xu et al., 2023](#)). By updating only a small subset of pre-trained parameters or introducing a limited number of new trainable parameters, PEFT achieves accuracy comparable to full fine-tuning. This approach not only enables more computationally efficient training but also helps mitigate overfitting, which is often associated with fine-tuning large models on small datasets. However, as PEFT was originally developed for single-task fine-tuning, its direct application in continual learning scenarios often leads to challenges, particularly catastrophic forgetting. Nevertheless, recent research has begun exploring the potential of specific PEFT techniques, notably Low-Rank Adaptation (LoRA) ([Liu et al., 2024](#); [Wang et al., 2023](#); [Yu et al., 2024](#)), for continual learning. LoRA introduces trainable low-rank matrices to a model, and optimizes these matrices on downstream tasks. Despite being computationally efficient, previous approaches still face difficulties such as scalability due to continuous accumulation of parameters, the need for task identifiers at test time, and catastrophic forgetting.

We argue that catastrophic forgetting arises from the inherent bias of sequential fine-tuning towards the recently learned task, which favors plasticity over stability. We demonstrate that addressing the

continual learning problem through the lens of model merging can effectively address the stability-plasticity dilemma. Specifically, after learning a new task, the new knowledge is integrated with the prior one by merging the latest learned weights with the weights of a global, continuously evolving model. We learn each task using LoRA modules for its computational efficiency. We show that this simple change in the learning paradigm offers various advantages, addressing the shortcomings of conventional fine-tuning such as catastrophic forgetting, sensitivity to task order and similarity, and lower generalization (see Fig. 1 and Section 3 for full details). Furthermore, merging the weights in one global model addresses the parameter growth of previous techniques (Wang et al., 2023).

To further improve knowledge integration across tasks without interference, we propose PAM, a simple *during-training* Parameter Alignment method for effective Merging. Specifically, while training a new task, we encourage alignment between the LoRA weights being learned and the weights of a global, continuously evolving LoRA module representing previous tasks (see schematic illustration in Fig. 2). After training, the weights of the current task are merged into the global LoRA. We demonstrate that this simple change in the conventional learning paradigm yields high robustness against forgetting over standard sequential fine-tuning.

Our main contributions are:

- We demonstrate that leveraging model merging for continual learning, rather than the typical sequential fine-tuning, offers a solution to the latter’s inherent shortcomings, such as its bias towards recent tasks, sensitivity to task orders and similarities, and poorer generalization.
- We propose PAM (Parameter Alignment for Merging), a simple yet effective and computationally efficient approach that optimizes parameter alignment during task training to reduce inter-task interference and enable effective merging.
- We empirically demonstrate the effectiveness of PAM in improving performance on seen tasks and generalization to unseen ones across a diverse set of continual learning benchmarks. Furthermore, we show that it can be combined with existing continual learning techniques, bringing a further boost in performance.

2. Problem formulation

We are presented with a sequence of tasks $[T] = \{1, \dots, T\}$, where for each task t , we are given i.i.d. data D_t sampled from an unknown distribution \mathcal{D}_t . Our aim is to find a model $f(\cdot)$ which minimizes the expected loss \mathcal{L} averaged over all the tasks seen thus far:

$$\frac{1}{t} \sum_{t'=1}^t \mathbb{E}_{(x,y) \sim \mathcal{D}_{t'}} \mathcal{L}(f(x), y). \quad (1)$$

We consider models f with pretrained weights W_0 that are fixed, and a low rank adapter, LoRA, with weights W that are trained on the continual learning tasks. Specifically, let $W_0^l \in \mathbb{R}^{n \times m}$ be the pre-trained weights at layer l . Additional trainable parameters W^l are added to the layer, where W^l is composed of low rank matrices $A \in \mathbb{R}^{n \times r}$ and $B \in \mathbb{R}^{r \times m}$, and r is the LoRA rank. This rank is typically smaller than the original matrix dimensions (i.e., $r \ll \min(n, m)$). The forward pass of the data can then be written as

$$W_0^l x + W^l x = W_0^l x + BAx, \quad (2)$$

where x is the input of the layer.

Although our models are aware when a new task arrives, we assume that (1) the data of previous tasks $\{D_1, \dots, D_{t-1}\}$ is not available when learning a new task t ; and (2) the task identity of an input is not available at inference time. These constraints are generally reflective of the types of challenges

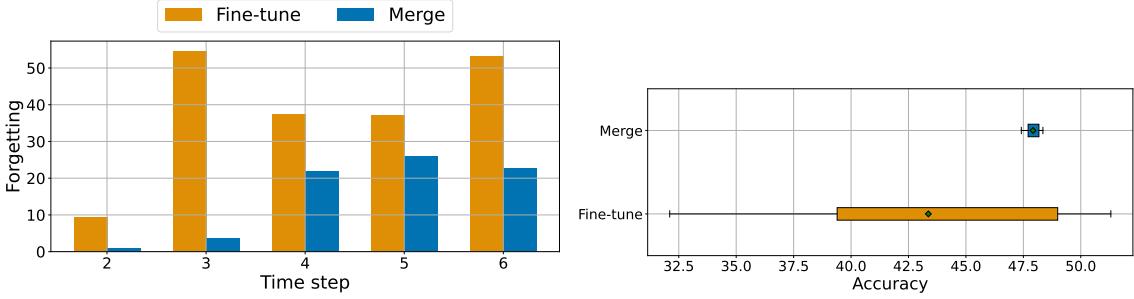


Figure 1 | (Left) Forgetting over time throughout training on visual QA tasks from the CoIN benchmark (Chen et al., 2024), measured as the negation of the backward transfer metric described in Section 5.1. At each time step, we report the average forgetting on seen tasks (lower is better). Merging demonstrates greater robustness to forgetting and task similarity compared to fine-tuning a single LoRA module, where performance fluctuates between steps. (Right) Average accuracy across all tasks for three different task orders. Fine-tuning shows higher sensitivity to task order.

real-world situations would induce and further complicate the direct optimization of Eq. (1), as we are unable to directly sample from, or iterate through, past tasks.

3. What can merging offer over fine-tuning?

The fundamental approach in continual learning involves continuous fine-tuning as new tasks arrive. On top of this core fine-tuning, CL methods integrate regularization or replay to improve stability (Buzzega et al., 2020; Kirkpatrick et al., 2017; Wang et al., 2024b). Sequential fine-tuning, however, suffers from catastrophic forgetting: it offers a relatively low expected loss on the most recent task t , but the expected loss of previous tasks is likely to increase. In this work, we provide another paradigm to address this challenge via merging, demonstrating its inherit ability to address the stability-plasticity dilemma. Just as fine-tuning, our approach can be combined with other continual learning approaches to further boost performance.

The modification to the conventional learning paradigm is rather simple: fine-tuning a temporary LoRA module on a new task, and then merging the learned parameters into a separate global LoRA module that evolves over time. Specifically, when a new task t arrives, the current global LoRA weights, denoted $W_{G,t-1}$, are cloned into a separate LoRA module (with parameters W') and fine-tuned on the new task data, D_t . After training on task t , the updated global LoRA weights, $W_{G,t}$, are obtained by merging $W_{G,t-1}$ and W' using element-wise averaging, a standard model merging technique (Choshen et al., 2022; Wortsman et al., 2022). This approach allows the model to integrate parameters learned from previous tasks while adapting to new tasks, effectively mitigating catastrophic forgetting in a parameter-efficient manner.

Fig. 1 demonstrates the impact of this approach on a recent continual instruction tuning benchmark with visual question answering (QA) tasks (Chen et al., 2024) (Details are provided in Section 5.1). We find that merging alone *effectively mitigates catastrophic forgetting* of previous tasks, even without incorporating sophisticated mechanisms. We also observe that in fine-tuning a single LoRA module, performance of some tasks can fluctuate depending on the most recently learned task, as demonstrated by the increases and decreases in forgetting over time. In contrast, merging is *less impacted by varying degrees of task similarity*. Notably, we find that merging exhibits *more robustness to task ordering*. This robustness to task ordering is a critical advantage over traditional fine-tuning, which, as shown in

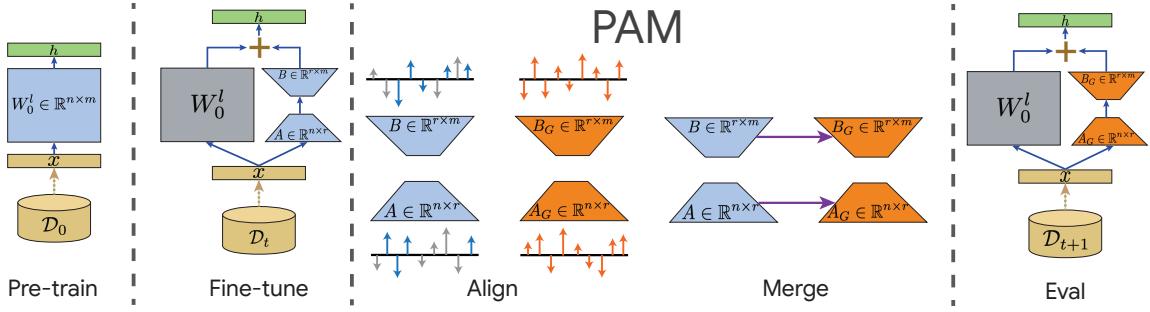


Figure 2 | Method overview. After pre-training W_0^l , PAM fine-tunes **temporary LoRA weights** $\{A, B\}$ on a new task, and periodically reinitializes a subset of the parameters which are misaligned with a **global LoRA** $\{A_G, B_G\}$. After training, the temporary LoRA $\{A, B\}$ is merged with the global LoRA $\{A_G, B_G\}$, which is ultimately used for evaluation.

Fig. 1 (right), is highly sensitive to the sequence of tasks.

4. Parameter Alignment and Merging (PAM)

Section 3 demonstrates the effectiveness of model merging in addressing the CL challenges with a simple post-training paradigm shift. However, while direct merging of weights learned from different tasks shows promise compared to standard fine-tuning, simply averaging these weights can be suboptimal as task-specific updates may interfere with each other. To address this, we propose to optimize for Parameter Alignment *during* training for effective Merging, leading to our method PAM.

PAM introduces a simple change to new task training to encourage alignment with prior tasks, as depicted in Fig. 2. During new task training, we *periodically* check the current task weights, identifying and re-initializing any that become misaligned with the important weights of the global LoRA weights, that capture knowledge of prior tasks. To measure alignment, we leverage previously published findings, demonstrating that the weight sign is a simple and effective tool for alignment in both weight and gradient spaces (Chen et al., 2020; Davari and Belilovsky, 2025; Yadav et al., 2024).

Specifically, during task training, we periodically compare the signs of W' with those in $W_{G,t-1}$, and re-initialize the weights in W' that have misaligned signs, provided their corresponding weights in $W_{G,t-1}$ are among the top $p\%$ in magnitude. The alignment percentage p acts as a control parameter, balancing the model’s stability and plasticity. A higher p

Algorithm 1 PAM

```

Input: pretrained weights  $W_0$ , Global LoRA  $W_{G,t-1}$ , alignment percent  $p$ , schedule  $s$ , number of tasks  $T$ 
for  $t = 2$  to  $T$  do
    Add a new LoRA with parameters  $W'$ 
     $W' = W_{G,t-1}$ 
     $A = \text{Sort } |W_{G,t-1}|$  in descending order
     $k = \lceil (p/100) \cdot \text{length}(A) \rceil$ 
     $\tau = A[k]$ 
    for  $step = 1$  to  $total\_steps$  do
        Update  $W'$ 
        if  $step \bmod s == 0$  then
            // Check misalignment
            for each weight  $w$  in  $W'$  do
                if  $\text{sign}(w') \neq \text{sign}(w_{G,t-1}) \wedge w_{G,t-1} > \tau$  then
                    Re-initialize  $w'$ 
                end if
            end for
        end if
    end for
    // Merge
     $W_{G,t} = \text{Avg}(W_{G,t-1}, W')$ 
end for

```

emphasizes preserving previously learned knowledge (stability) at the potential cost of reduced adaptability to the current task (plasticity) (see Section 5.3). After training on task t , the updated global LoRA weights, $W_{G,t}$, are obtained by merging $W_{G,t-1}$ and W' . We provide an outline in Algorithm 1.

Although many techniques have been recently proposed for improving model merging (Wang et al., 2024a; Yadav et al., 2024), they operate *post-training*, in contrast to our method that operates *during training*. For instance, a closely related method, TIES (Yadav et al., 2024), utilizes post-training alignment by *pruning* weights whose signs conflict with the sign of the highest total magnitude across all LoRAs. We argue that periodic alignment during training offers several advantages. It prevents LoRA weights from diverging excessively during fine-tuning, maintaining lower sign conflicts throughout training; moreover, periodic alignment allows weights to recover from initial misalignment and continue contributing to the learning process, avoiding the capacity loss associated with pruning in TIES, as discussed further in Section 5.3. This dynamic alignment strategy enables better knowledge integration and effective merging compared to a one-time post-training alignment.

5. Experiments

5.1. Experimental Details

Benchmarks We evaluate our method on CoIN (Chen et al., 2024), a recent CL benchmark for multimodal large language models. We considered the six image QA tasks: VQAV2 (Goyal et al., 2017), TextVQA (Singh et al., 2019), OCR-VQA (Mishra et al., 2019), ScienceQA (Lu et al., 2022), VizWizVQA (Gurari et al., 2018), and GQA (Hudson and Manning, 2019). For this benchmark, we evaluate three different task orderings: the original order and the alphabetically sorted order introduced in (Chen et al., 2024), where the former proved to be more challenging, and a randomly generated order. While we provide the main results for all three, our extended analysis is centered on the most challenging original task order. Furthermore, we extend CoIN by adding four image-based tasks—AI2D (Kembhavi et al., 2016), AOKVQA-MC Schwenk et al. (2022), OKVQA (Marino et al., 2019), and RSVQA-Ir (Lobry et al., 2020)—resulting in a longer sequence of ten tasks with higher diversity and task dissimilarity, including remote sensing VQA and multi-choice QA. Finally, we evaluate on a sequence of video QA that includes MSVD-QA (Xu et al., 2017), MSRVTT-QA (Xu et al., 2017), and ActivityNet-QA (Yu et al., 2019), further diversifying the evaluation beyond image-based tasks. See Appendix A for further details.

Model, code, and compute resources As our base model we use PaliGemma (Beyer et al., 2024), a representative and open-source VLM known for its effectiveness across a wide range of transfer tasks. Inputs are set to a resolution of 224×224 pixels, and we apply the best hyperparameters reported in (Beyer et al., 2024) for each task, except for the learning rate, which we adjust to 5×10^{-4} for improved performance with LoRA training. We use a LoRA rank of $r = 32$ and alignment percentage $p = 50\%$. Hyper-parameter details are presented in Appendix A. For all our experiments, we build on top of the big vision library (Beyer et al., 2022), a Jax and Flax-based framework. The library and the benchmark used have Apache2 license. All experiments were run on TPUs, with each individual task in the sequence utilizing 64 TPUs. The duration of training each task ranges from 1 hour to 10 hours depending on the size of the task, with a six-task sequence completing in roughly 20 hours.

Baselines We compare against several state-of-the-art methods. We start with standard approaches: (1) *Zero-shot*: tasks are evaluated using the pre-trained PaliGemma model without further training; (2) *Independent*: Separate LoRA modules are fine-tuned on each task independently; (3) *Multitask*: A single LoRA is trained concurrently on all tasks; (4) *Fine-tuning*: A single LoRA is trained sequentially

on all tasks. We also compare against continual learning methods including: (5) LWF ([Li and Hoiem, 2017](#)): a well-known regularization method that adds a distillation loss to restrict the change in previous knowledge; (6) I-LoRA ([Li et al., 2025](#)): the framework consists of two LoRA modules (faster LoRA learner and slower LoRA learner), and the knowledge is slowly consolidated during training in the slower learner. Note that this method initially incorporates experience replay of previous task data, but we exclude this part for fair and direct comparison against all rehearsal-free baselines; (7) O-LoRA ([Wang et al., 2023](#)): it adds new LoRA module for each new task, while optimizing its weights to be orthogonal to previously added LoRA modules to reduce interference. (8) MoELoRA ([Chen et al., 2024](#)): a mixture of experts is trained on the CL sequence with learned gate function to route the input to experts; (9) MaxMag ([Marczak et al., 2024](#)): a closely related work that merges knowledge from all previous learned models by picking the parameter with the highest magnitude; we applied this technique to LoRA weights. Finally, we compare against state-of-the-art model merging techniques proposed for multitask learning including (10) *Average*: LoRA weights are merged using a simple average; (11) *TIES* ([Yadav et al., 2024](#)): this post-training technique reduces interference by *pruning* the weights that are misaligned; (12) *TALL* ([Wang et al., 2024a](#)): this technique employs a post-training localization strategy to identify and merge important weights and can complement other merging techniques that employs parameter alignment. For the merging baselines, we maintain one core LoRA, which is copied as a new task arrives, fine-tuned on that task, and then merged back with the core. Note that these methods were initially designed with independent initialization for each task. However, we discovered that using the initialization of the core module as a starting point yields better results, as we demonstrate in Appendix B. Therefore, we will compare against the improved version of these baselines.

Evaluation metrics We do extensive evaluation by assessing four metrics to measure model capabilities: (1) *Accuracy* (ACC): the average accuracy of all tasks at the end of the sequence of T tasks; (2) *Backward transfer* (BWT) ([Lopez-Paz and Ranzato, 2017](#)): this metric measures the influence of learning task t on the performance of each previous task $i < t$. Thus, it assess the model robustness against forgetting; (3) *Forward transfer* (FWT) ([Lopez-Paz and Ranzato, 2017](#)): this metric measures the ability of a model to generalize to unseen tasks. It is estimated by the difference between the model performance at time step t on unseen tasks ($i > t$) and the zero-shot performance. (4) *Average task Accuracy* (A_t): this metric assess the plasticity of the model to learn a new task. It measures the performance of task t immediately after learning it. For merging based techniques, we also report A_m , which represents the accuracy after the merging step. The computation details of each metric can be found in the appendix.

5.2. Results

Motivated by recent observations of strong zero- and few-shot learning performance in large models ([Brown et al., 2020](#)), we start by evaluating the performance of the pretrained PaliGemma model on CoIN tasks. Despite its large pretraining data, we observe that its generalization to these downstream tasks is suboptimal, as evidenced by the zero-shot performance when compared to adapting an independent LoRA for each task or training single LoRA for all tasks jointly as shown in Table 1. Although sequential fine-tuning improves performance and has high plasticity (A_t), it suffers from catastrophic forgetting and poor generalization, as reflected by the lower BWT and FWT scores.

Merging aligned models via PAM demonstrates significant performance improvements over standard fine-tuning, exhibiting enhanced robustness against forgetting even without explicit forgetting mitigation techniques. Furthermore, PAM enhances the model’s generalization capabilities and increases forward transfer across tasks. Interestingly, PAM surpasses methods that expand parameters

Table 1 | Performance of different methods on the CoIN benchmark across four metrics that assess various model capabilities. Results are presented as the mean and standard deviation over three task orders, with higher values indicating better performance for all metrics.

METHOD	ACC	BWT	FWT	A_t
ZERO-SHOT	24.74	-	-	-
INDEPENDENT	76.46	-	-	-
MULTITASK	73.93	-	-	-
FINE-TUNE	43.36±8.18	-39.51±9.87	7.71±2.51	76.29±0.18
LWF	47.15±3.52	-33.66±3.84	9.67±1.22	75.20±0.32
I-LoRA	42.11±6.55	-40.95±8.13	7.89±2.60	76.24±0.35
O-LoRA	46.53±6.88	-32.08±7.87	9.45±3.02	73.27±1.01
MoELoRA	46.59±9.98	-36.40±11.97	7.79±2.24	76.93±0.27
MAGMAX	45.74±0.88	-22.68±6.51	4.75±3.36	76.29±0.18
PAM (OURS)	49.89±1.66	-19.45±0.95	11.11±0.09	76.31±0.03

Table 2 | Comparison of various merging methods on the CoIN benchmark. Results are presented as the mean and standard deviation over three task orders.

METHOD	ALIGNMENT/LOCALIZATION	ACC	BWT	FWT	A_t	A_m
AVERAGE	×	47.41±0.38	-22.40±0.80	10.50±2.24	76.31±0.15	65.95±1.31
PAM (OURS)	DURING TRAINING	49.89±1.66	-19.45±0.95	11.11±0.09	76.31±0.03	66.11±0.87
TIES	POST TRAINING	47.78±0.38	-22.87±0.68	9.92±2.05	76.35±0.14	66.85±0.26
TALL	POST TRAINING	49.22±2.53	-12.85±1.53	11.90±0.57	76.42±0.12	59.93±1.51
TALL+PAM	DURING & POST TRAINING	51.97±5.59	-9.27±5.53	11.96±0.97	76.28±0.25	59.70±1.23

over time (O-LoRA), or utilize a mixture of experts (MoELoRA), despite relying solely on a single evolving LoRA for all tasks. Moreover, we find that the averaging of aligned weights in PAM is more effective than selecting the parameter with the highest magnitude from each learned LoRA as in MagMax.

Chen et al. (2024) observed that task order influences model performance. This is likely attributable to the impact of the sequence on the learning dynamics and trajectory, which are shaped by task complexity, data size, and dissimilarity. We find that fine-tuning a single or mixture of LoRAs exhibits greater sensitivity to task order, as demonstrated by the high standard deviation in performance across different sequences. This is also highlighted by the fluctuations in task performance at each step during fine-tuning, as illustrated in Fig. 1 and Appendix A. In contrast, merging-based methods, specifically MagMax and PAM, demonstrate greater robustness to task order, delivering more stable performance regardless of the task sequence.

5.3. Empirical Analysis

In this section we analyze the effect of *during-training* alignment, which is the key algorithmic advancement behind our proposed method, PAM. Specifically, we present ablations investigating how during-training alignment affects PAM, revealing the role of this step in mitigating catastrophic forgetting. We also compare PAM against post-training interventions for improved merging, and examine the impact of PAM’s design choices.

Improved knowledge retention and reduced task interference Table 2 presents a comparison against state-of-the-art merging methods. The high performance of different strategies further

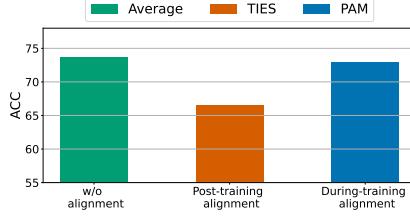


Figure 3 | Task accuracy before merging. Pruning misaligned parameters after training decreases performance, whereas optimizing for alignment during training maintains model plasticity.

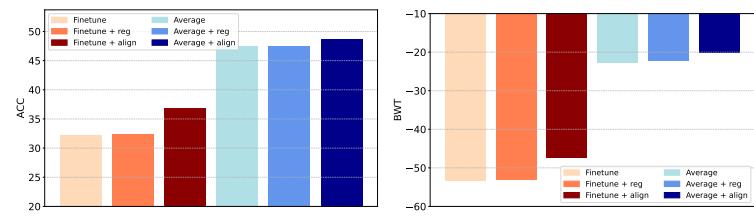


Figure 4 | Weight regularization provides only limited benefits in mitigating forgetting. In contrast, alignment improves performance of fine-tuning and average baselines.

confirms the gains by merging over the conventional fine-tuning. To assess the effect of during-training alignment in isolation, we compare against the average baseline, which is equivalent to PAM but with the alignment process ablated. As shown in the table PAM has superior performance over this baseline across all metrics. Comparing PAM and TIES, both utilizing parameter alignment, shows that aligning parameters during training improves knowledge retention and reduces task interference. Post-training alignment often favors one task when misalignment occurs, and pruning misaligned parameters can harm accuracy, as we will demonstrate next. In contrast, during-training alignment encourages the model to find parameters that aligned with previous tasks, leading to more effective merging of knowledge after training. Localizing and merging important weights between a global LoRA and a temporary one by TALL yields good average accuracy, but low A_m . This could be due to pruning selfish weights by TALL that are important to one task. Integrating PAM’s during-training alignment with TALL’s post-training localization effectively improves TALL’s performance and reduces forgetting, further demonstrating the effectiveness of during-training alignment.

Improved model plasticity and capacity We examine the impact of alignment on task accuracy and analyze the proportion of misaligned parameters. In this analysis, we focus on the final two tasks of one sequence, which demonstrates some level of dissimilarity. We compare the behavior of post-training alignment (TIES) with that of during-training alignment (PAM). Our findings reveal that, on average, 29.41% of parameters across all layers are misaligned between the global module and the most recent LoRA module trained using TIES. As illustrated in Fig. 3, just applying post-training alignment to the penultimate task prior to merging results in a 7.44% reduction in accuracy. In contrast, during-training alignment preserves model plasticity, as evidenced by PAM’s accuracy, which remains comparable to the case of without alignment.

Regularization does not have the same effect Regularization is a common CL strategy to preserve knowledge from past tasks by restricting the weights from moving too far from their original values

Alignment [%]	ACC	BWT	FWT	A_t	A_m
30%	47.74	-22.24	11.21	76.49	66.28
50%	48.16	-21.37	11.04	76.33	65.97
70%	48.32	-20.63	11.19	76.30	65.51

Table 3 | Effect of the percentage of alignment in PAM.

Method	ACC	BWT	FWT	A_t	A_m
reinit $W_{G,t-1}$	49.89 ± 1.66	-19.45 ± 0.95	11.11 ± 0.09	76.31 ± 0.03	66.11 ± 0.87
reinit zero	49.53 ± 1.61	-20.14 ± 1.19	10.81 ± 0.34	76.24 ± 0.15	66.32 ± 0.65

Table 4 | Effect of the reinitialization strategy in PAM.

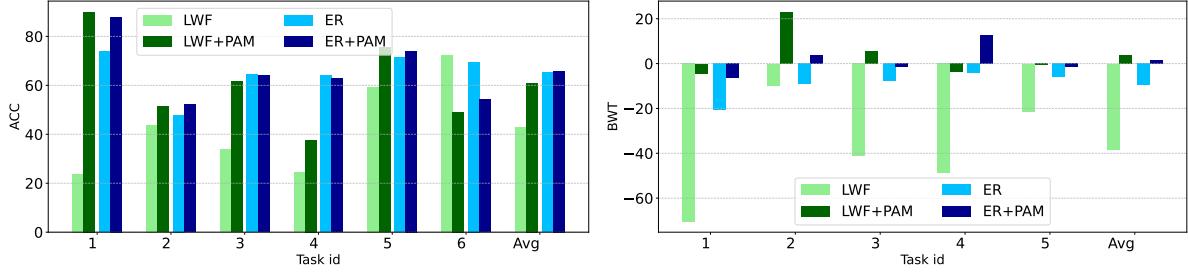


Figure 5 | Evaluating the performance of existing CL methods on the original order of the CoIN benchmark when combined with PAM. We report the ACC (left) and BWT (right) for each task after training on the whole sequence. PAM yields improved performance and less forgetting, particularly demonstrating its efficacy in challenging scenarios where past data is not available.

(Aljundi et al., 2018; Kirkpatrick et al., 2017). We investigate whether regularizing the weights would be sufficient to promote parameter alignment and mitigate forgetting.

Our results in Fig. 4 demonstrate that adding regularization yields a slight performance improvement, but it fails to effectively address catastrophic forgetting. Interestingly, we find that alignment can even *provide gains to the fine-tuning baseline and reduce forgetting*. This further demonstrates the effectiveness of alignment and its potential for improving performance of other techniques.

Control over the stability-plasticity trade-off During-training alignment, as opposed to post-training, allows controlling the percentage of aligned parameters. To investigate this effect, we analyze PAM with varying percentages of aligned parameters (30%, 50%, and 70%), and report the results in Table 3. We find that larger percentage of alignment favors previous tasks and reduces forgetting. Nevertheless, increasing the alignment percentage reduces the model’s plasticity as shown by the reduced A_t and A_m . This highlights the trade-off between stability and plasticity in CL and that PAM provides a valuable tool for navigating this trade-off.

Reinitialization strategy plays a role We examine two strategies for reinitializing misaligned weights: (1) zero initialization, and (2) using global weights $W_{G,t-1}$. Our findings indicate that the two strategies are effective, with re-initializing with global weights favors stability and reduces forgetting at a slight cost to current task performance after merging, as shown in Table 4.

5.4. PAM as a Complementary Continual Learning Technique

The simplicity of PAM facilitates its integration with existing CL techniques. In this section, we evaluate its performance when integrated with two representative methods from different strategies: LWF (Li and Hoiem, 2017), regularization-based, and Experience Replay (ER) (Chaudhry et al., 2019; Robins, 1995), a rehearsal based strategy.

Fig. 5 presents the accuracy and backward transfer on each task after learning the whole sequence of the CoIN benchmark. We find that PAM substantially improves the performance of earlier tasks, and reduces forgetting when integrated with LWF. Moreover, despite the effectiveness of ER in preserving performance on earlier tasks through replaying old samples, PAM still provides additional gains. Interestingly, the combination of PAM with the regularization-based method LWF achieves performance comparable to ER performance, notably without old data. These findings demonstrate the potential of our proposed shift in the learning paradigm for improved performance.

Table 5 | Performance on the extended CoIN and video QA benchmarks.

METHOD	EXTENDED COIN				VIDEO QA			
	ACC	BWT	FWT	A_t	ACC	BWT	FWT	A_t
ZERO-SHOT	23.08	-	-	-	22.06	-	-	-
INDEPENDENT	69.83	-	-	-	51.27	-	-	-
FINE-TUNE	49.73	-22.19	2.87	69.71	40.73	-16.66	10.61	51.84
PAM (OURS)	50.43	-14.10	9.38	69.64	46.06	-1.95	11.79	51.75
LWF	49.49	-20.81	3.56	68.23	41.55	-14.60	11.04	51.30
LWF + PAM	51.53	-5.08	7.60	67.90	46.12	-0.46	11.69	51.29

Evaluation on additional benchmarks Finally, we assess the generalization of our findings on two other benchmarks: extended version of CoIN with longer sequence of diverse tasks and video question answering tasks. Table 5 presents the results on the four studied metrics. Consistent with previous observations, PAM reduces forgetting and increase forward transfer. Moreover, it improves performance of LWF.

6. Related Work

Model merging Model merging has emerged as a promising technique for combining multiple models fine-tuned on different tasks. Several methods have shown to perform well, with a few dominating the landscape, such as averaging (Izmailov et al., 2018; Wortsman et al., 2022), model arithmetic (Ilharco et al., 2022), TIES (Yadav et al., 2024), and TALL-masks (Wang et al., 2024a). These methods have been primarily applied to merging entire models, and the success of these merging techniques has been shown to rely on linear mode connectivity (Frankle et al., 2020; Sharma et al., 2024). Recent concurrent work studied merging entire models over time (Dziadzio et al., 2024; Kleiman et al., 2025), with Dziadzio et al. (2024) revealing that current merging methods offer little benefit over simple averaging for temporal merging. More recently, merging fine-tuned LoRAs instead of full models has been explored. Initial studies suggest, however, that directly applying existing techniques to LoRA merging does not yield substantial gains. For instance, Prabhakar et al. (2024) investigated different merging strategies applied to LoRAs and found that simple concatenation often outperforms more sophisticated methods in a skill composition task. This highlights the need for further research to develop specialized methods tailored to the unique characteristics of LoRAs.

Continual learning The CL paradigm includes several key objectives, such as mitigating forgetting on past tasks, facilitating forward transfer to leverage prior knowledge for new tasks, and ensuring computational efficiency in learning, among others (Díaz-Rodríguez et al., 2018; Hadsell et al., 2020). Most prior work focuses on addressing forgetting, which can be categorized into regularization, architectural, and replay-based methods. Regularization methods minimize forgetting by constraining changes to important weights (Ahn et al., 2019; Aljundi et al., 2018; Kirkpatrick et al., 2017). Architectural methods modify the model’s structure, often using separate modules for different tasks, to reduce interference (Gurbuz and Dovrolis, 2022; Konishi et al., 2023; Sokar et al., 2021; Wortsman et al., 2020). Replay-based methods incorporate a subset of previous task data alongside new task data during training (Buzzega et al., 2020; Rebuffi et al., 2017; Shin et al., 2017). This paper introduces a perspective that is orthogonal to the three CL categories, focusing on model merging rather than standard fine-tuning which is the core component in these approaches. A few recent works explored adding new adapters over time (Liu et al., 2024; Wang et al., 2023; Yu et al., 2024). Biderman et al. (2024) showed that LoRA fine-tuning results in less forgetting of the base model’s

knowledge compared to full fine-tuning, having less impact on performance for tasks beyond the downstream task. [Marczak et al. \(2024\)](#) proposed integrating knowledge from each task model into one by selecting the parameters with the highest magnitudes across all models.

7. Discussion and Conclusion

In this work, we demonstrated key limitations of fine-tuning within the CL paradigm, such as a strong bias toward recent tasks, reduced generalization capability, and increased sensitivity to task order. To address these challenges, we propose a novel view that replaces standard fine-tuning with a merging-based strategy to better integrate new and previously learned knowledge. Our proposed method, PAM, incorporates two key components. First, separate fine-tuning on the latest task followed by merging with previously learned weights. Second, to address potential interference between tasks, we introduce periodic alignment of the current task parameters with those from prior tasks during training. To achieve computationally efficient training on a continuous data stream, we build our approach on LoRA.

We conducted evaluations across various vision-language benchmarks, considering differences in sequence length, complexity, and task diversity. The results indicate that PAM mitigates forgetting and enhances generalization. Additionally, it demonstrates greater robustness to variations in task order. Furthermore, our findings reveal that alignment during training surpasses state-of-the-art merging techniques that rely on post-training alignment, and even provides gains to sequential fine-tuning without merging. Lastly, we showed that PAM improves existing CL approaches.

Limitations and future work While the straightforward merging of data into a single LoRA offers an attractive solution for scaling in the CL paradigm, it may not adequately address scenarios where new data significantly differs from previously learned tasks. Additionally, selectively fine-tuning subsets of the weights could help mitigate forgetting and enhance the forward transfer of relevant knowledge.

Acknowledgements

The authors would like to thank Clare Lyle, Razvan Pascanu, Hugo Larochelle, Alexandre Ramé, Doina Precup, and the rest of the Google DeepMind team for valuable feedback on this work.

References

- H. Ahn, S. Cha, D. Lee, and T. Moon. Uncertainty-based continual learning with adaptive regularization. *Advances in neural information processing systems*, 32, 2019.
- R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pages 139–154, 2018.
- L. Beyer, X. Zhai, and A. Kolesnikov. Big vision. https://github.com/google-research/big_vision, 2022.
- L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *CoRR*, 2024.

- D. Biderman, J. Portes, J. J. G. Ortiz, M. Paul, P. Greengard, C. Jennings, D. King, S. Havens, V. Chiley, J. Frankle, C. Blakeney, and J. P. Cunningham. LoRA learns less and forgets less. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.
- T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- P. Buzzega, M. Boschini, A. Porrello, D. Abati, and S. Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020.
- A. Chaudhry, M. Rohrbach, M. Elhoseiny, T. Ajanthan, P. K. Dokania, P. H. Torr, and M. Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.
- C. Chen, J. Zhu, X. Luo, H. T. Shen, J. Song, and L. Gao. CoIN: A benchmark of continual instruction tuning for multimodel large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.
- Z. Chen, J. Ngiam, Y. Huang, T. Luong, H. Kretzschmar, Y. Chai, and D. Anguelov. Just pick a sign: Optimizing deep multitask models with gradient sign dropout. *Advances in Neural Information Processing Systems*, 33:2039–2050, 2020.
- L. Choshen, E. Venezian, N. Slonim, and Y. Katz. Fusing finetuned models for better pretraining. *arXiv preprint arXiv:2204.03044*, 2022.
- M. Davari and E. Belilovsky. Model breadcrumbs: Scaling multi-task model merging with sparse masks. In *European Conference on Computer Vision*, pages 270–287. Springer, 2025.
- J. Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- N. Díaz-Rodríguez, V. Lomonaco, D. Filliat, D. Maltoni, et al. Don’t forget, there is more than forgetting: new metrics for continual learning. In *Continual Learning Workshop at NeurIPS 2018*, pages 1–7, 2018.
- S. Dziadzio, V. Uinandarao, K. Roth, A. Prabhu, Z. Akata, S. Albanie, and M. Bethge. How to merge your multimodal models over time? *arXiv preprint arXiv:2412.06712*, 2024.
- J. Frankle, G. K. Dziugaite, D. Roy, and M. Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pages 3259–3269. PMLR, 2020.
- Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.
- D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.
- M. B. Gurbuz and C. Dovrolis. Nispa: Neuro-inspired stability-plasticity adaptation for continual learning in sparse networks. In *International Conference on Machine Learning*, pages 8157–8174. PMLR, 2022.

- R. Hadsell, D. Rao, A. A. Rusu, and R. Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in cognitive sciences*, 24(12):1028–1040, 2020.
- N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.
- E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.
- D. A. Hudson and C. D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.
- G. Ilharco, M. T. Ribeiro, M. Wortsman, S. Gururangan, L. Schmidt, H. Hajishirzi, and A. Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.
- P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.
- A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, and A. Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016.
- J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.
- A. Kleiman, G. K. Dziugaite, J. Frankle, S. Kakade, and M. Paul. Soup to go: mitigating forgetting during continual learning with model averaging. *arXiv preprint arXiv:2501.05559*, 2025.
- T. Konishi, M. Kurokawa, C. Ono, Z. Ke, G. Kim, and B. Liu. Parameter-level soft-masking for continual learning. In *International Conference on Machine Learning*, pages 17492–17505. PMLR, 2023.
- X. Li, W. Ren, W. Qin, L. Wang, T. Zhao, and R. Hong. Analyzing and reducing catastrophic forgetting in parameter efficient tuning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.
- Z. Li and D. Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.
- J. Liu, J. Wu, J. Liu, and Y. Duan. Learning attentional mixture of loras for language model continual learning. *arXiv preprint arXiv:2409.19611*, 2024.
- S. Lobry, D. Marcos, J. Murray, and D. Tuia. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566, 2020.
- D. Lopez-Paz and M. Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.
- P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

- D. Marczak, B. Twardowski, T. Trzciński, and S. Cygert. Magmax: Leveraging model merging for seamless continual learning. In *European Conference on Computer Vision*, pages 379–395. Springer, 2024.
- K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.
- S. I. Mirzadeh, A. Chaudhry, D. Yin, H. Hu, R. Pascanu, D. Gorur, and M. Farajtabar. Wide neural networks forget less catastrophically. In *International Conference on Machine Learning*, pages 15699–15717. PMLR, 2022.
- A. Mishra, S. Shekhar, A. K. Singh, and A. 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.
- A. Prabhakar, Y. Li, K. Narasimhan, S. Kakade, E. Malach, and S. Jelassi. Lora soups: Merging loras for practical skill composition tasks. *arXiv preprint arXiv:2410.13025*, 2024.
- A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.
- A. Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146, 1995.
- D. Schwenk, A. Khandelwal, C. Clark, K. Marino, and R. Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022.
- E. Sharma, D. M. Roy, and G. K. Dziugaite. The non-local model merging problem: Permutation symmetries and variance collapse. *arXiv preprint arXiv:2410.12766*, 2024.
- H. Shin, J. K. Lee, J. Kim, and J. Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017.
- A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.
- G. Sokar, D. C. Mocanu, and M. Pechenizkiy. Spacenet: Make free space for continual learning. *Neurocomputing*, 439:1–11, 2021.
- K. Wang, N. Dimitriadis, G. Ortiz-Jimenez, F. Fleuret, and P. Frossard. Localizing task information for improved model merging and compression. In *Forty-first International Conference on Machine Learning*, 2024a.
- L. Wang, X. Zhang, H. Su, and J. Zhu. A comprehensive survey of continual learning: theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024b.

- X. Wang, T. Chen, Q. Ge, H. Xia, R. Bao, R. Zheng, Q. Zhang, T. Gui, and X. Huang. Orthogonal subspace learning for language model continual learning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.
- M. Wortsman, V. Ramanujan, R. Liu, A. Kembhavi, M. Rastegari, J. Yosinski, and A. Farhadi. Supermasks in superposition. *Advances in Neural Information Processing Systems*, 33:15173–15184, 2020.
- M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR, 2022.
- D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017.
- L. Xu, H. Xie, S.-Z. J. Qin, X. Tao, and F. L. Wang. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv preprint arXiv:2312.12148*, 2023.
- P. Yadav, D. Tam, L. Choshen, C. A. Raffel, and M. Bansal. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36, 2024.
- J. Yu, Y. Zhuge, L. Zhang, P. Hu, D. Wang, H. Lu, and Y. He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23219–23230, 2024.
- Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, and D. Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019.

Table 6 | Hyperparameters used in training each task. We used the best hyperparameters reported by [Beyer et al. \(2024\)](#) on fine-tuning on each task independently.

TASK	EPOCHS	BATCH SIZE	WEIGHT DECAY	LABEL SMOOTHING
SCIENCEQA	20	128	0.0	0.0
TEXTVQA	5	256	0.0	0.0
GQA	1	256	0.0	0.0
VizWizVQA	10	256	0.0	0.0
VQAv2	10	256	1E-06	0.0
OCR-VQA	3	128	0.0	0.0
RSVQA-LR	3	256	0.0	0.2
AI2D	10	256	1E-06	0.0
OKVQA	10	128	0.0	0.0
AOKVQA-MC	15	128	0.0	0.0
ACTIVITYNET-QA	1	128	1E-06	0.0
MSRVTT-QA	1	128	0.0	0.0
MSVD-QA	1	128	3E-07	0.0

Table 7 | Performance of PAM across three different run on a sequence of three tasks.

	ACC	BWT
PAM	68.47±0.06	-2.08±0.03

A. Experimental Details

Tasks We evaluated our approach on six image-based question-answering tasks from the CoIN benchmark ([Chen et al., 2024](#)): VQAV2 ([Goyal et al., 2017](#)), TextVQA ([Singh et al., 2019](#)), OCR-VQA ([Mishra et al., 2019](#)), ScienceQA ([Lu et al., 2022](#)), VizWizVQA ([Gurari et al., 2018](#)), and GQA ([Hudson and Manning, 2019](#)). Following the sequences of tasks proposed by the authors, we also included an additional randomly shuffled order. The three task sequences are as follows:

- ScienceQA, TextVQA, GQA, VizWizVQA, VQAv2, and OCR-VQA
- GQA, OCR-VQA, ScienceQA, TextVQA, VizWizVQA, and VQAV2
- TextVQA, OCR-VQA, VizWizVQA, GQA, ScienceQA, and VQAV2

We introduced an extended version of the CoIN benchmark by incorporating four additional tasks (AI2D ([Kembhavi et al., 2016](#)), AOKVQA-MC [Schwenk et al. \(2022\)](#), OKVQA ([Marino et al., 2019](#)), and RSVQA-lr ([Lobry et al., 2020](#))), enhancing task diversity and creating a more challenging, longer sequence. The tasks were arranged in the following order: ScienceQA, TextVQA, AOKVQA-MC, OCR-VQA, RSVQA-lr, VizWizVQA, GQA, OKVQA, VQAv2, and AI2D. Fig. 6 illustrates examples from each task. We found that ScienceQA is one of the most challenging tasks, often causing interference with others, likely due to its higher degree of dissimilarity. This task involves reasoning and answering multiple-choice questions on topics spanning natural science, language science, and social science.

We also experiment with video question answering tasks, constructing a sequence of three tasks: MSVD-QA ([Xu et al., 2017](#)), MSRVTT-QA ([Xu et al., 2017](#)), and ActivityNet-QA ([Yu et al., 2019](#)).

TextVQA		Question: What is the name of the first app on the top right? Answer: flipshare
VQAv2		Question: What color is the bird's beak? Answer: black
ScienceQA		The diagrams below show two pure samples of gas in identical closed, rigid containers. Each colored ball represents one gas particle. Both samples have the same number of particles. Question: Compare the average kinetic energies of the particles in each sample. Which sample has the higher temperature? (A) sample A (B) neither; the samples have the same temperature, (C) sample B Answer: sample B
VizWizVQA		Question: What is this package? Answer: frozen food
OCR-VQA		Question: Who is the author of this book? Answer: James Dashner
GQA		Question: What animals are the trees behind of? Answer: horses
AI2D		Question: Which obtains energy from the sun? choose from: small mammals plants roadrunners snakes Answer: plants
RSVQA_LR		Question: Is the number of residential buildings equal to the number of square forests? Answer: no
OKVQA		Question: Where is this located? Answer: india
AOKVQA_MC		Question: What is the keyboard being plugged into? choose from: computer wall pen mouse Answer: computer

Figure 6 | Examples from each task in the extended CoIN benchmark.

Hyper-parameters We used the best hyperparameter found by [Beyer et al. \(2024\)](#) in fine-tuning each transfer task, except for the learning rate, where we used a value of 5e-4. We report the hyperparameters for all tasks in Table 6. We used a LoRA of a rank 32 for all experiments and analyzed other values in Fig. 8. LoRA modules are added to attention and feedforward layers. We perform the alignment procedure every 100 steps, and use alignment percentage of 50% in all experiments. We analyzed the effect of these hyper-parameters in Tables 3 and 8. Our main results characterize performance variability as the standard deviation across different task orders. We used one seed for our experiments due to resource constraints and the high computational cost of learning a sequence of tasks with VLMs. This is motivated by findings in ([Beyer et al., 2024](#)) showing low standard deviation across runs for most transfer tasks. To further validate this for our setup, we performed multiple runs with three seeds on a sequence of three tasks and provided the results in Table 7. Consistent with previous findings, we observe that the standard deviation across three run is very small.

Baselines We used a LoRA rank of 32 for all baselines. For TIES. ([Yadav et al., 2024](#)), we tuned the percentage of weights retained before alignment (i.e., during the pruning of redundant weights, as shown in Figure 1 in ([Yadav et al., 2024](#))). We explored values from the grid [0.5, 0.7, 0.9] and found that 0.9 yielded the best results. TALL ([Wang et al., 2024a](#)) includes a hyperparameter that determines the amount of information the mask extracts from the multi-task vector. Lower values lead to more parameters being selected for a task. We searched over the grid [0.1, 0.2, 0.4] and found that 0.1 performed the best.

Table 8 | Effect of alignment schedule on PAM performance.

SCHEDULE	ACC	BWT	A_t
10	48.25	-21.21	65.93
100	48.16	-21.37	65.97
300	47.82	-21.78	65.97

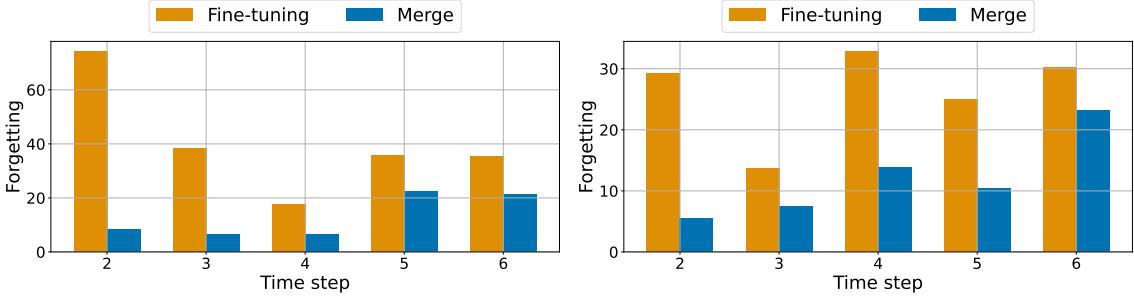


Figure 7 | Forgetting over time on the CoIN benchmark (Chen et al., 2024) for the second (left) and third (right) orders. At each time step, we report the average forgetting on seen tasks so far (lower is better).

Evaluation metrics We do extensive evaluation by assessing four metrics to measure model capabilities.

(1) *Average accuracy (ACC)*: Let $A_{t,i}$ be the accuracy of task i after learning task t . The average accuracy of all tasks at the end of the sequence of T tasks is defined as:

$$ACC = \frac{1}{T} \sum_{i=1}^T A_{T,i}. \quad (3)$$

(2) *Backward transfer (BWT)* (Lopez-Paz and Ranzato, 2017): This metric measures the influence of learning task t on the performance of each previous task $i < t$. Thus, it assesses the model robustness against forgetting. BWT is defined as:

$$BWT = \frac{1}{T-1} \sum_{i=1}^{T-1} A_{T,i} - A_{i,i}. \quad (4)$$

(3) *Forward transfer (FWT)* (Lopez-Paz and Ranzato, 2017): This metric measures the ability of a model to generalize to unseen tasks. It is estimated by the difference between the model performance at time step t on unseen tasks ($i > t$) and the zero-shot performance. Specifically, let \bar{a}_i be the zero-shot accuracy of task i , FWT is defined as:

$$FWT = \frac{1}{T-1} \sum_{i=2}^{T-1} A_{i-1,i} - \bar{a}_i. \quad (5)$$

(4) *Average task Accuracy (A_t)*: this metric assesses the plasticity of the model to learn a new task. It measures the performance of task t immediately after learning it, and is defined as follows:

$$A_t = \frac{1}{T} \sum_{i=1}^T A_{i,i}. \quad (6)$$

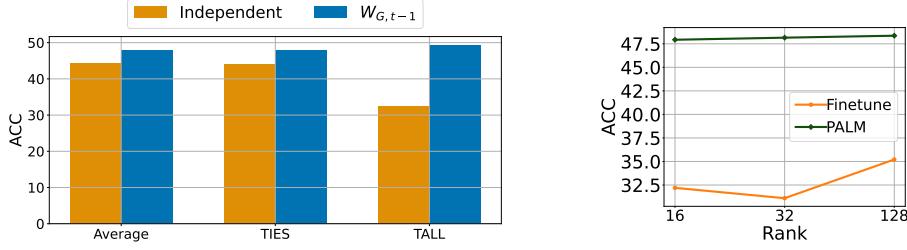


Figure 8 | Impact of weight initialization. Using global LoRA weights to initialize new task weights is more effective across different merging methods (left). Effect of LoRA rank on Performance. PAM can use low rank without sacrificing performance, enabling more computational efficient training over time (right).

B. Additional Experiments

Detailed results on CoIN benchmark Figure 1 presents the forgetting over time for the standard order of CoIN. We illustrate the results for the other two considered orders in Figure 7. Consistent with previous observations, fine-tuning is more prone to forgetting and the performance of a task varies over time; some tasks have a greater impact on previous task performance than others, likely due to varying degrees of task similarity. In contrast, merging demonstrates greater robustness to these task variations.

Impact of LoRA initialization Merging methods typically operate on independently trained models. In the context of merging LoRAs, we compare performance of initializing new LoRA module with the global LoRA module weights, versus starting from an independent initialization. To isolate the effect of initialization, we analyze the performance of the average merging baseline without the periodic alignment process. As illustrated in Fig. 8 left, initializing a new LoRA module from the global one leads to better performance compared to randomly initializing this module. We find that this observation is consistent across different merging algorithms including TIES and TALL. This finding highlights the importance of initialization strategies and provides insights for designing effective LoRA merging mechanisms.

Impact of LoRA Rank Previous studies have explored the impact of network width on continual learning performance, observing that wider networks generally exhibit reduced catastrophic forgetting (Mirzadeh et al., 2022). This phenomenon can be attributed to the increased capacity of wider networks, allowing them to accommodate new information without overwriting previously learned knowledge. We investigated whether this trend holds true for LoRA, where the rank of the adapter can be considered to be analogous to network width. While smaller ranks are more efficient, we hypothesized that they might exacerbate forgetting due to their limited capacity.

Our results in Fig. 8 right confirm that higher ranks (e.g., 128) lead to improved performance for fine-tuning. This suggests that increasing the rank provides additional capacity, similarly as in full fine-tuning, allowing the model to better retain previous knowledge. However, our proposed model, PAM, maintains strong performance even at low ranks, demonstrating its efficiency. This makes PAM a particularly attractive option for resource-constrained environments or applications where model size is a critical factor.