A. Additional Settings

We experiment on two further CIFAR100 settings with distinct cardinality of base task classes:

- CIFAR100/20 Base, with 20 base task classes followed by 8 incremental tasks with 10 classes each,
- CIFAR100/50 Base, with 50 base task classes followed by 5 incremental tasks with 10 classes each.

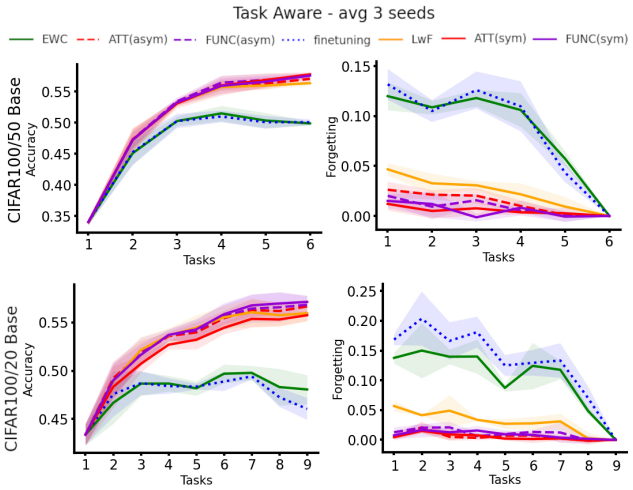

Figure 6. Mean and standard deviation of task-aware accuracy and forgetting scores for the additional CIFAR100/20 and CIFAR100/50 settings (over 3 random runs).

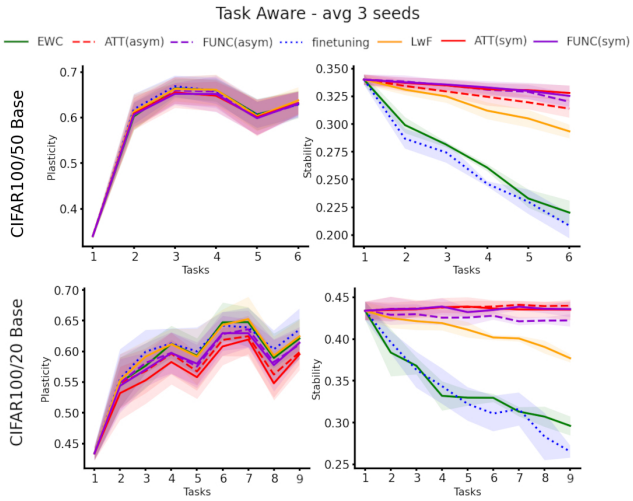

Figure 7. Mean and standard deviation of task-aware plasticity-stability scores for the additional CIFAR100/20 and CIFAR100/50 settings (over 3 random runs).

The task aware accuracy and forgetting scores on these are shown in Figure 6. We find the PAD-based losses to consistently outperform other regularization approaches with LwF being the closest tie. Along the direction of plasticity-stability trade off (see Figure 7), we observe that: (a) the attention-based $\mathcal{L}_{\text{(a)sym}}$ losses retain better rigidity than their functional counterparts, and (b) the asymmetric variants are more plastic than their symmetric counterparts across these settings. These trends further validate our hypotheses in sections 3.2 and 3.3, respectively.

B. Task Agnostic Results

Figure 8 depicts the task-agnostic accuracy and forgetting scores for the settings mentioned in the main paper as well as in Appendix A. Given the contradictory terms of resource-scarce *exemplar-free* CL and data-hungry ViTs, task-agnostic evaluations can be seen to be particularly challenging. The further avoidance of heavier data augmentations in our training settings give rise to two major repercussions across the task-agnostic accuracies: (a) the scores remain consistently low, and (b) the models show smaller yet consistent variations in performances across all settings.

That said, we find functional $\mathcal{L}_{\text{(a)sym}}$ losses to be performing the best on all but CIFAR100/50 setting. The larger proportion of base task classes in the latter setting can be seen to be greatly benefiting the learning of LwF (the least parameterized loss term). Further on the class proportions, we observe that an equal spread of classes across the tasks can be seen to have a smoothing effect on the variations of scores across different methods.

On the contrary, the CIFAR100/50 setting leads to low variability of task-agnostic forgetting scores across the methods. This can again be attributed to the fact that a larger first task better leverages the generalization capabilities of ViTs thus advancing them at avoiding forgetting over the subsequent incremental steps. This further adds to our reasoning regarding the natural resilience of ViTs to CL settings. When compared across methods, the attentional variants of $\mathcal{L}_{\text{(a)sym}}$ can be seen to display the least amount of forgetting followed by their functional counterparts.

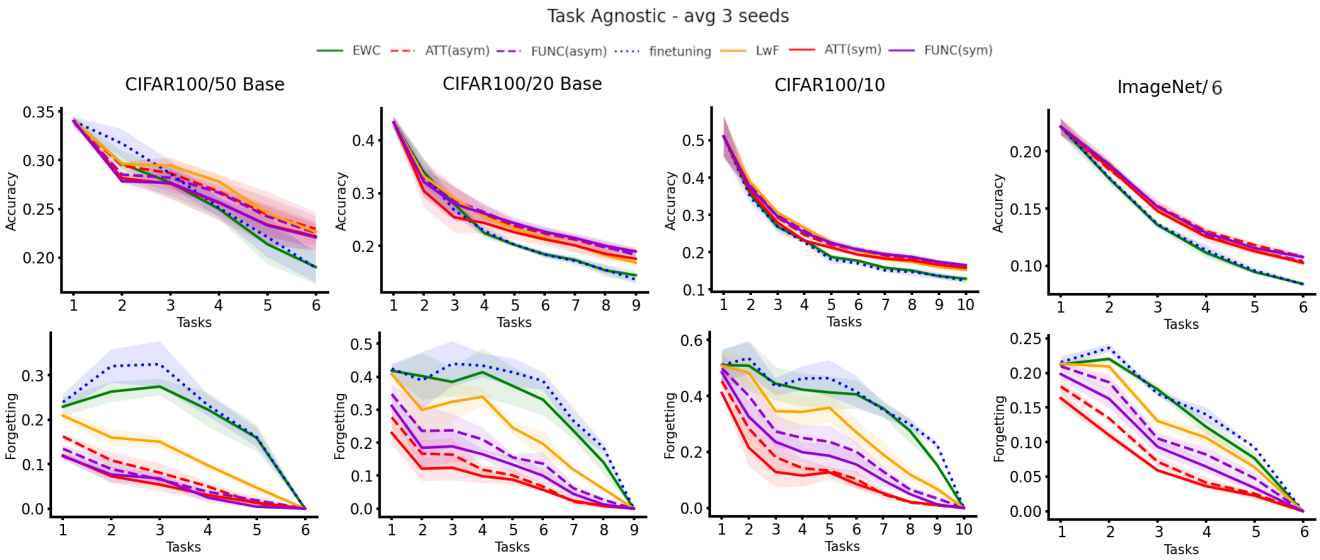

Figure 8. Mean and standard deviation of task-agnostic accuracy and forgetting scores for CIFAR100/10, CIFAR100/20, CIFAR100/50, and ImageNet/6 settings (over 3 random runs). The larger proportion of base task classes (for example, CIFAR100/50) gives rise to higher variations of accuracies and lower variation of forgetting scores across methods – with the latter indicating the inclination of ViTs towards better generalization and preservation of knowledge.