

APPENDICES

The appendices include more ablation results (§A), additional phase-wise accuracy plots (§B), additional plots for α_η and α_ϕ values (§C), and the scaling weights (§D). Our open-source code will be public later.

A MORE ABLATION RESULTS

In Table S1, we supplement the ablation results for more settings. By differentiating the numbers of learnable parameters, we have three different types of blocks: By differentiating the numbers of learnable parameters, we have 3 block types: (1) “all” means learning all the convolutional weights and biases; (2) “scaling” means learning neuron-level scaling weights (?) on the top of a frozen base model θ_{base} ; and (3) “frozen” means using θ_{base} (frozen) as the feature extractor of the stable block. Please note that the classification layers are always learnable. “2×” (“4×”) means using an expanded network with two (four) branches with the same type residual blocks. We can observe that using two types of blocks (Rows 7-9) achieves better performance compared to using double-sized or quadruple-sized model using the same blocks.

Row	Ablation Setting	CIFAR-100		
		N=5	10	25
1	1× “all”	63.17	60.14	57.54
2	2× “all”	64.49	61.89	58.87
3	4× “all”	65.70	62.31	59.40
4	1× “scaling”	62.48	61.53	60.17
5	2× “scaling”	65.13	64.08	62.50
6	4× “scaling”	66.00	64.67	63.67
7	“all” + “scaling”	66.21	65.17	63.45
8	“all” + “frozen”	65.62	64.05	63.67
9	“scaling” + “frozen”	64.71	63.65	62.89

Table S1: More ablation results (%) on CIFAR-100.

B ADDITIONAL PHASE-WISE ACCURACY PLOTS

In Figures S1 and S2, we supplement phase-wise accuracy on ImageNet-Subset and ImageNet, respectively. “Upper Bound” shows the results of joint training with all previous data accessible in each phase. We can observe that our method achieves the highest average accuracy in all settings.

C ADDITIONAL PLOTS FOR α_η AND α_ϕ VALUES

In Figures S3-S7, we supplement the plots α_η and α_ϕ values on CIFAR-100 and ImageNet-Subset. All curves are smoothed with a rate of 0.8 for a better visualization.

D THE SCALING WEIGHTS

For stable blocks, we deploy the scaling weights ϕ , which specifically transfer the base model θ_{base} . The aim is to preserve the structural knowledge of θ_{base} and slowly adapt ϕ to the new class data. Specifically, we assume the q -th layer of θ_{base} contains R neurons, so we have R neuron weights as $\{W_{q,r}\}_{r=1}^R$. For conciseness, we denote them as W_q . For W_q , we learn R scaling weights denoted as ϕ_q . Let X_{q-1} and X_q be the input and output (feature maps) of the q -th layer. We apply ϕ_q to W_q as,

$$X_q = (W_q \odot \phi_q)X_{q-1}, \quad (1)$$

where \odot donates the element-wise multiplication. Assuming there are Q layers in total, the scaling weights are denoted as $\phi = \{\phi_q\}_{q=1}^Q$.

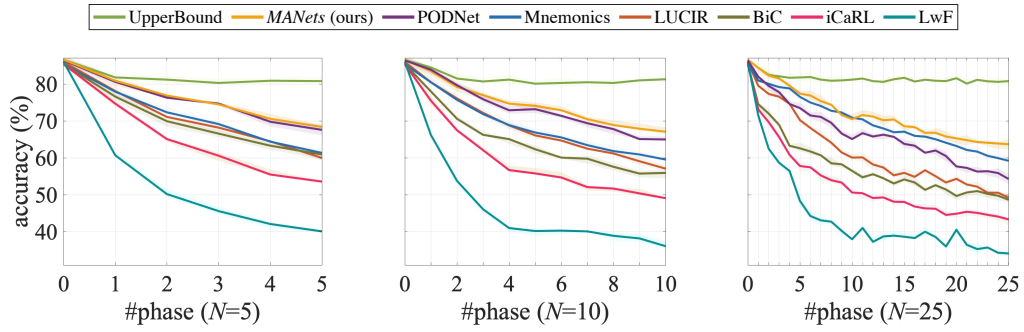


Figure S1: Phase-wise accuracy on ImageNet-Subset.

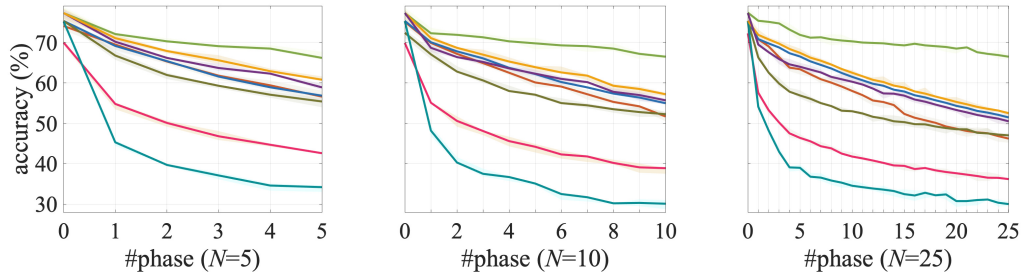
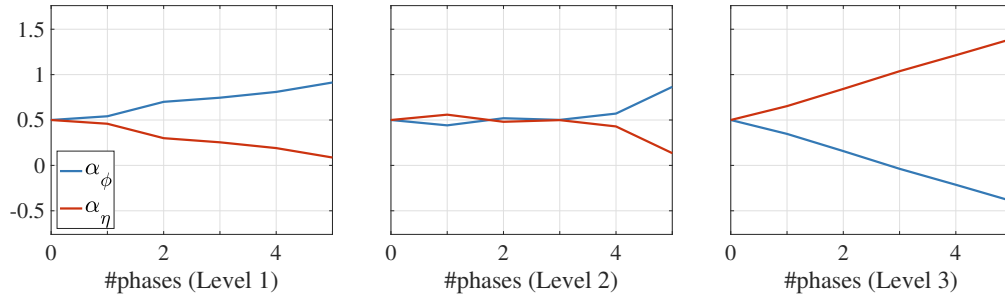
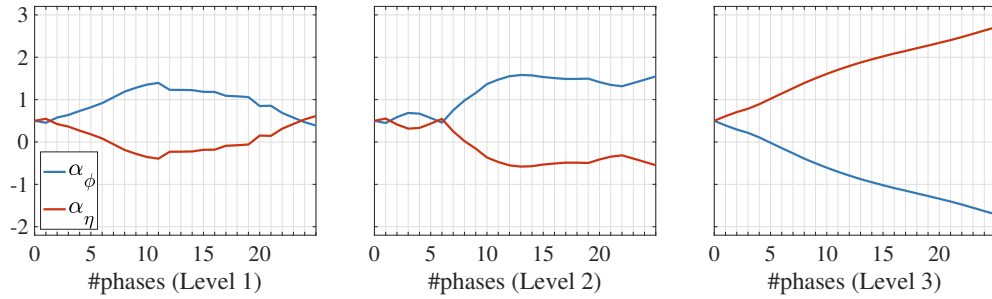


Figure S2: Phase-wise accuracy on ImageNet.

Figure S3: The changes of values for α_η and α_ϕ on CIFAR-100 ($N=5$).Figure S4: The changes of values for α_η and α_ϕ on CIFAR-100 ($N=25$).

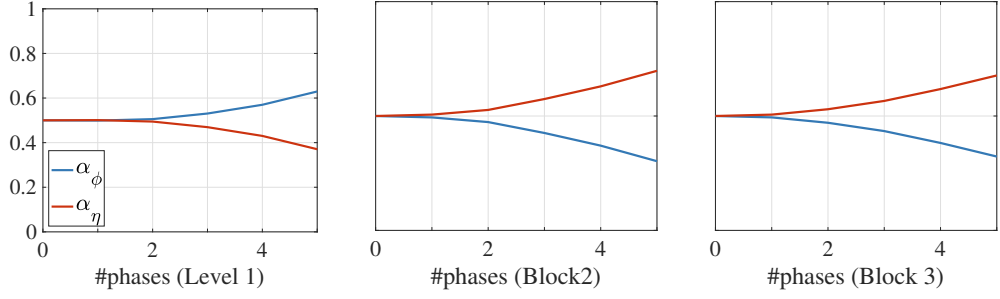


Figure S5: The changes of values for α_η and α_ϕ on ImageNet-Subset ($N=5$).

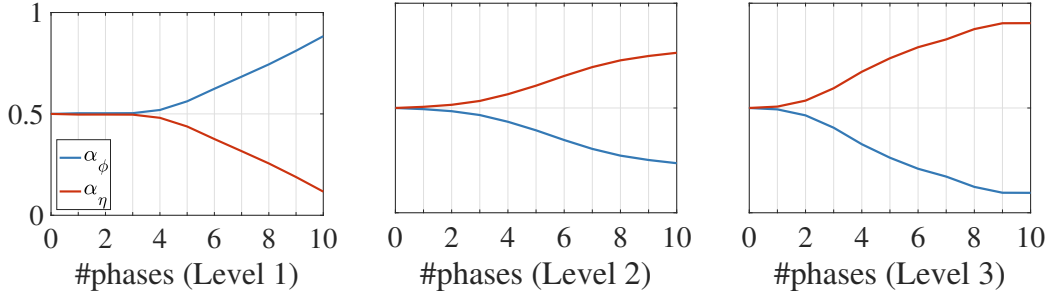


Figure S6: The changes of values for α_η and α_ϕ on ImageNet-Subset ($N=10$).

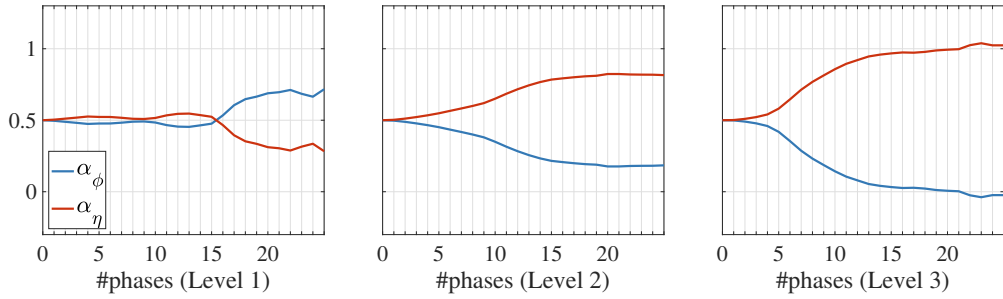


Figure S7: The changes of values for α_η and α_ϕ on ImageNet-Subset ($N=25$).