Model Merging in the Essential Subspace

Supplementary Material

A. Proofs

This section provides the proofs for the expected output error after truncation for both the standard SVD and the proposed Essential Subspace Decomposition (ESD), as well as the equivalence of whitening and Orthogonal Procrustes.

A.1. Proof for SVD Truncation Error

Theorem 1. *Given a task matrix $\Delta_W \in \mathbb{R}^{d_{out} \times d_{in}}$ with its singular value decomposition $\Delta_W = U\Sigma V^\top = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$. Let $\widehat{\Delta_W} = \sum_{i=1}^{k} \sigma_i u_i v_i^\top$ be its top-k rank approximation. For an input x drawn from a distribution $\mathcal{D}$, the expected squared L_2 error on the output activation is:*

$$\mathbb{E}_{x\sim\mathcal{D}}\left[\|\Delta_W x - \widehat{\Delta_W} x\|_2^2\right] = \sum_{i=k+1}^{r} \sigma_i^2 \cdot \mathbb{E}_{x\sim\mathcal{D}}\left[(v_i^\top x)^2\right].$$

Proof. The error matrix resulting from the truncation is the sum of the discarded components:

$$\Delta_W - \widehat{\Delta_W} = \sum_{i=k+1}^{r} \sigma_i u_i v_i^\top. \tag{18}$$

The error on the output activation for a given input x is:

$$(\Delta_W - \widehat{\Delta_W})x = \left(\sum_{i=k+1}^{r} \sigma_i u_i v_i^\top\right) x = \sum_{i=k+1}^{r} \sigma_i u_i (v_i^\top x). \tag{19}$$

Since $v_i^\top x$ is a scalar, we can rewrite this as a linear combination of the orthonormal vectors u_i. The squared L_2 norm of this error vector is:

$$\|(\Delta_W - \widehat{\Delta_W})x\|_2^2 = \left\| \sum_{i=k+1}^{r} (\sigma_i v_i^\top x) u_i \right\|_2^2. \tag{20}$$

Because the left singular vectors $\{u_i\}$ form an orthonormal set, the squared norm of their weighted sum is the sum of the squares of the weights:

$$\|(\Delta_W - \widehat{\Delta_W})x\|_2^2 = \sum_{i=k+1}^{r} (\sigma_i v_i^\top x)^2 = \sum_{i=k+1}^{r} \sigma_i^2 (v_i^\top x)^2. \tag{21}$$

By taking the expectation over the input distribution $\mathcal{D}$ and applying the linearity of expectation, we arrive at the final expression:

$$\mathbb{E}_{x\sim\mathcal{D}}\left[\|\Delta_W x - \widehat{\Delta_W} x\|_2^2\right] = \sum_{i=k+1}^{r} \sigma_i^2 \cdot \mathbb{E}_{x\sim\mathcal{D}}\left[(v_i^\top x)^2\right]. \tag{22}$$

This completes the proof. □

A.2. Proof for ESD Truncation Error

Theorem 2. *Given a task matrix $\Delta_W \in \mathbb{R}^{d_{out} \times d_{in}}$, let $\hat{P} \in \mathbb{R}^{d_{out} \times k}$ be the matrix whose columns are the top-k principal components (eigenvectors) derived from the activation shift matrix $\Delta_O = X_{proxy}\Delta_W^\top$. Let $\widehat{\Delta_W} = \hat{P}\hat{A} = \hat{P}(\hat{P}^\top \Delta_W)$ be the ESD reconstruction. For an input $x \sim \mathcal{D}$, the expected squared L_2 error on the output is proportional to the sum of the discarded eigenvalues:*

$$\mathbb{E}_{x\sim\mathcal{D}}\left[\|\Delta_W x - \widehat{\Delta_W} x\|_2^2\right] = \sum_{i=k+1}^{d_{out}} \lambda_i.$$

Proof. The error on the output activation for an input x is:

$$\Delta_W x - \widehat{\Delta_W} x = \Delta_W x - \hat{P}\hat{P}^\top \Delta_W x = (I - \hat{P}\hat{P}^\top)\Delta_W x. \tag{23}$$

The matrix $(I - \hat{P}\hat{P}^\top)$ is the projection matrix onto the subspace spanned by the discarded eigenvectors $\{p_{k+1}, \ldots, p_{d_{\text{out}}}\}$. Let $y = \Delta_W x$ be the activation shift for input x. The error vector can be expressed as the projection of y onto this orthogonal subspace:

$$(I - \hat{P}\hat{P}^\top)y = \sum_{i=k+1}^{d_{\text{out}}} (p_i^\top y) p_i. \tag{24}$$

Since $\{p_i\}$ form an orthonormal basis, the squared L_2 norm is the sum of the squares of the projection coefficients:

$$\begin{aligned} \|\Delta_W x - \widehat{\Delta_W} x\|_2^2 &= \left\| \sum_{i=k+1}^{d_{\text{out}}} (p_i^\top y) p_i \right\|_2^2 \\ &= \sum_{i=k+1}^{d_{\text{out}}} (p_i^\top y)^2 \\ &= \sum_{i=k+1}^{d_{\text{out}}} (p_i^\top \Delta_W x)^2. \end{aligned} \tag{25}$$

Now, we take the expectation over the input distribution $\mathcal{D}$:

$$\begin{aligned} \mathbb{E}_{x\sim\mathcal{D}}\left[\|\Delta_W x - \widehat{\Delta_W} x\|_2^2\right] &= \mathbb{E}_{x\sim\mathcal{D}}\left[\sum_{i=k+1}^{d_{\text{out}}} (p_i^\top \Delta_W x)^2\right] \\ &= \sum_{i=k+1}^{d_{\text{out}}} \mathbb{E}_{x\sim\mathcal{D}}\left[(p_i^\top \Delta_W x)^2\right]. \end{aligned} \tag{26}$$

By the definition of Principal Component Analysis (PCA), the eigenvalue λ_i of the covariance matrix of activation

shifts corresponds to the variance of the activation shifts projected onto the i-th principal component p_i. Assuming the activation shifts are centered (or mean-subtracted during PCA), this variance is:

$$\begin{aligned}\lambda_i &= \mathrm{Var}(p_i^\top \Delta_W x) \\ &= \mathbb{E}_{x\sim\mathcal{D}}\left[(p_i^\top \Delta_W x)^2\right] - \left(\mathbb{E}_{x\sim\mathcal{D}}[p_i^\top \Delta_W x]\right)^2 \\ &= \mathbb{E}_{x\sim\mathcal{D}}\left[(p_i^\top \Delta_W x)^2\right].\end{aligned} \tag{27}$$

Substituting this result back into our error expression, we get:

$$\mathbb{E}_{x\sim\mathcal{D}}\left[\|\Delta_W x - \widehat{\Delta_W} x\|_2^2\right] = \sum_{i=k+1}^{d_{\text{out}}} \lambda_i. \tag{28}$$

This completes the proof, showing that the expected error is solely dependent on the magnitude of the discarded eigenvalues, which represent the functional variance captured by those directions. □

A.3. Equivalence of Whitening and Orthogonal Procrustes

Theorem 3. *The transformations $X \mapsto X(X^\top X)^{-1/2}$ (whitening) and $X \mapsto UV^\top$ (Orthogonal Procrustes), where $X = U\Sigma V^\top$ is the singular value decomposition (SVD) of X, are equivalent.*

Proof. Let $X = U\Sigma V^\top$ be the SVD of X, where Σ is a diagonal matrix of singular values, we have:

$$X^\top X = V\Sigma^2 V^\top. \tag{29}$$

It follows that:

$$(X^\top X)^{-1/2} = V\Sigma^{-1}V^\top. \tag{30}$$

Substituting this into the whitening transformation gives:

$$\begin{aligned}X(X^\top X)^{-1/2} &= (U\Sigma V^\top)(V\Sigma^{-1}V^\top) \\ &= U\Sigma\Sigma^{-1}V^\top \\ &= UV^\top.\end{aligned} \tag{31}$$

The whitening operation is equivalent to solving the Orthogonal Procrustes problem, completing the proof. □

B. Method Details

B.1. Methodology Pseudocode

The pseudocode of our proposed model merging approach is presented in Algorithm 1. Steps highlighted in *orange* correspond to the essential subspace merging, while *blue* annotations indicate the polarized scaling. Our method decomposes each task matrix within its essential subspace and performs truncation, then concatenates the components from all task matrices to reconstruct a new merged matrix. Through three-level scaling, we ultimately obtain a fused model that effectively preserves essential task knowledge while maintaining minimal inter-task interference.

Algorithm 1 Essential Subspace Merging

Require: Task matrices $\{\Delta_{W_t}^{(\ell)}\}_{t=1}^{T}$ for all layers $\ell \in \mathcal{L}$, pre-trained weights θ_0, validation set $\mathcal{D}_{\text{val}}$
Ensure: Merged model parameters θ_M

▷ *Decomposition and Truncation*
for each task $t = 1$ to T and each layer ℓ **do**
  Obtain essential basis $P_t^{(\ell)}$
  Compute coordinate matrix: $A_t^{(\ell)} \leftarrow (P_t^{(\ell)})^\top \Delta_{W_t}^{(\ell)}$
  Truncate to top-k components: $\hat{P}_t^{(\ell)} \leftarrow P_t^{(\ell)}[:, 1:k]$, $\hat{A}_t^{(\ell)} \leftarrow A_t^{(\ell)}[1:k, :]$, where $k = \lfloor d_{\text{out}}/T \rfloor$
end for
▷ *Concatenation*
for each layer ℓ **do**
  $P_{\text{cat}}^{(\ell)} \leftarrow [\hat{P}_1^{(\ell)}, \hat{P}_2^{(\ell)}, \ldots, \hat{P}_T^{(\ell)}]$
  $A_{\text{cat}}^{(\ell)} \leftarrow [\hat{A}_1^{(\ell)}; \hat{A}_2^{(\ell)}; \ldots; \hat{A}_T^{(\ell)}]$
  ▷ *Inter-Task Scaling*
  for each task $t = 1$ to T **do**
    $\hat{A}_t^{(\ell)} \leftarrow s_t^{(\ell)} \cdot \hat{A}_t^{(\ell)}$
  end for
  ▷ *Inter-Dimension Scaling*
  for each column $j = 1$ to d_{in} **do**
    $a_j^{(\ell)} \leftarrow c_j^{(\ell)} \cdot a_j^{(\ell)}$
  end for
end for
▷ *Orthogonalization and Reconstruction*
for each layer ℓ **do**
  Compute SVD: $P_{\text{cat}}^{(\ell)} = U_P^{(\ell)} \Sigma_P^{(\ell)} (V_P^{(\ell)})^\top$
  Compute SVD: $A_{\text{cat}}^{(\ell)} = U_A^{(\ell)} \Sigma_A^{(\ell)} (V_A^{(\ell)})^\top$
  Orthogonalize: $\tilde{P}^{(\ell)} \leftarrow U_P^{(\ell)} (V_P^{(\ell)})^\top$, $\tilde{A}^{(\ell)} \leftarrow U_A^{(\ell)} (V_A^{(\ell)})^\top$
  Reconstruct: $\Delta_{\text{merged}}^{(\ell)} \leftarrow \tilde{P}^{(\ell)} \tilde{A}^{(\ell)}$
end for
▷ *Inter-Layer Scaling*
for each layer $\ell \in \mathcal{L}$ **do**
  $\Delta_{\text{merged}}^{(\ell)} \leftarrow \beta_\ell \cdot \Delta_{\text{merged}}^{(\ell)}$
end for
Find optimal α^* on $\mathcal{D}_{\text{val}}$ that maximizes performance
$\theta_M \leftarrow \theta_0 + \alpha^* \cdot \{\Delta_{\text{merged}}^{(\ell)}\}_{\ell=1}^{L}$
return θ_M

B.2. Details on Polarized Scaling

In the empirical evidence presented in Section 3.3, individual or merged task matrices are loaded in different sequences. To isolate the phenomenon from the distinct characteristics of different layer types, our analysis focuses solely on the four principal layer types within the ViT: namely, the QKV projection and the Output projection in the attention module, along with the Up- and Down-projections in the MLP. Furthermore, when merging layers

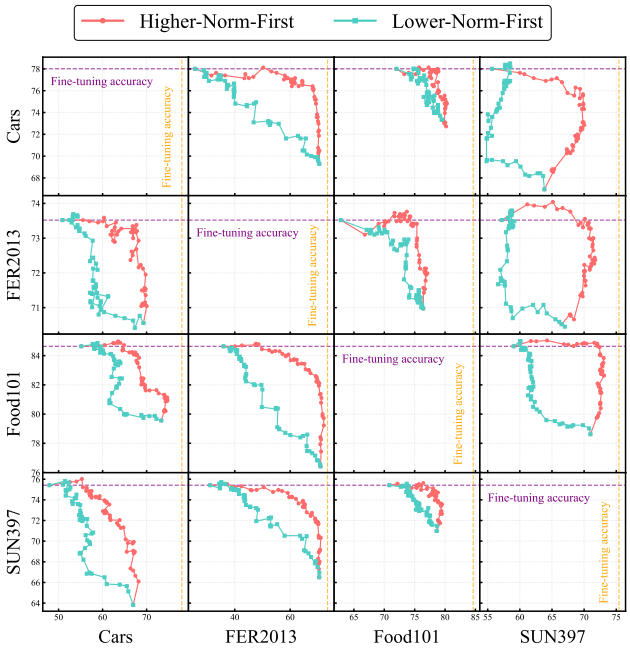


Figure 9. Illustration of task invasion: using the task matrices from one task to invade the fine-tuned model of another, performed layer-by-layer based on the norm order. Rows represent the invaded task, and columns represent the invading task.

sequentially, we group the layers by these four layer types. The merging process then proceeds by cyclically selecting task matrices from each group in a fixed order. This design ensures that the observed effects are attributable to the merging strategy rather than to variations across layer types.

Empirical Evidence: Pairwise Task Interaction. We further analyze the pairwise influence between task matrices. As shown in Figure 9, each column represents how the performance of two tasks changes when the task matrix of the column's task is added in different orders to the fine-tuned model of the row's task. The results show that adding task matrices in descending order of their norms yields a better Pareto optimal trade-off than adding them in ascending order. Introducing a large number of low-norm updates brings little improvement to the target task but significantly harms the performance of others. This highlights the importance of suppressing less critical or noisy updates while emphasizing the most essential ones in model merging.

B.3. Merging Non-Matrix Parameters

While most parameters in the ViT are 2D matrices merged using our proposed ESM within the Essential Subspace, the network also includes other parameter types. For non-matrix parameters such as bias vectors, layer normalization parameters, and the convolutional stem, we follow the standard practice in [11] and apply simple averaging.

Table 6. Performance on the 8-task benchmark. "ID": the average accuracy of the merged model across all tasks. "OOD": the result of merging fine-tuned models from 7 tasks and testing on the remaining unseen task. Each task is alternately used as the test task, and the reported OOD score is the average across all such settings.

Method	ViT-B/32		ViT-B/16		ViT-L/14	
	ID	OOD	ID	OOD	ID	OOD
Baseline	85.9	49.9	89.0	54.7	93.0	65.9
ESM (Ours)	**88.4**	**51.3**	**91.8**	**55.5**	**94.8**	**66.4**

Table 7. Scaling coefficient α selected using the validation set. T represents the total number of merged task-specific expert models.

ViT-B/32			ViT-B/16			ViT-L/14		
T=8	T=14	T=20	T=8	T=14	T=20	T=8	T=14	T=20
0.88	0.76	0.69	0.82	0.76	0.67	0.91	0.72	0.64

C. Experiment Details

C.1. OOD Performance

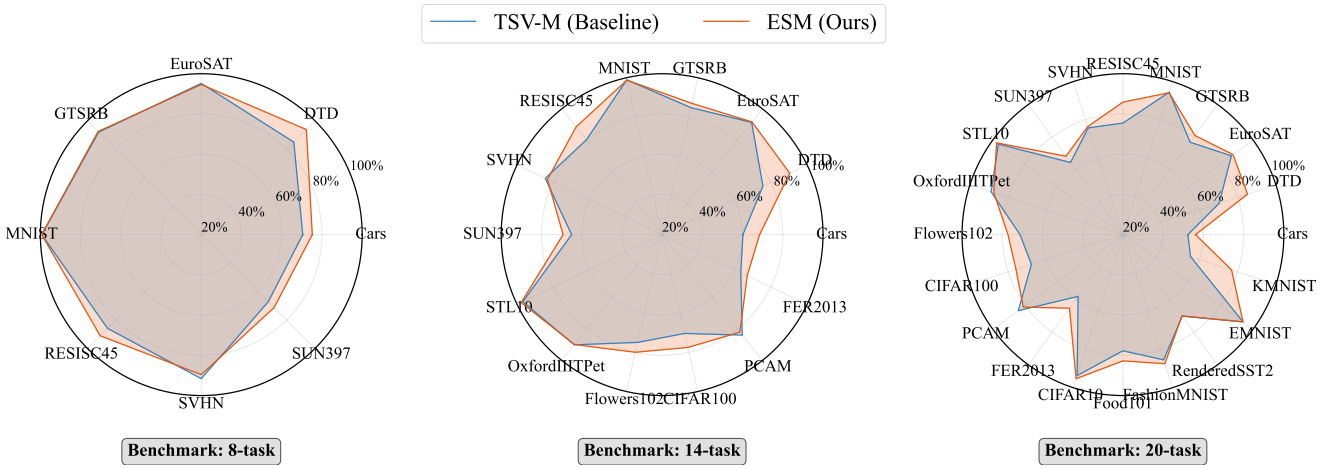
We evaluate the out-of-domain (OOD) generalization of our model merging method and that of the baseline [11], as presented in Table 6. Based on the 8-task benchmark, we treat each task in turn as the test task, merge the fine-tuned models from the remaining seven tasks, and evaluate the merged model on the held-out task. The final OOD score is the average performance across all test tasks. The results show that our method not only achieves substantial improvements in conventional in-domain evaluations but also demonstrates superior generalization in the out-of-domain setting.

C.2. Selection of the Global Scaling Coefficient α

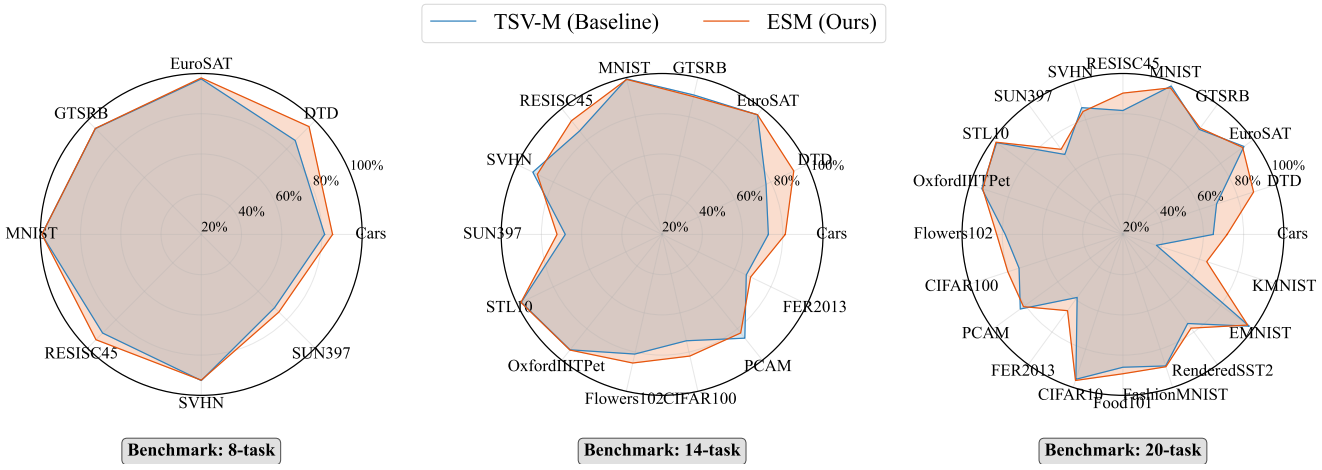
We report the global scaling coefficient α selected on the validation set, as shown in Table 7. Based on the empirical ranges used in previous model merging studies [11, 28], we set the search interval for α between 0.0 and 2.0 and perform binary search to determine the optimal value. The results show that the optimal α decreases as the number of tasks increases, likely because merging more tasks amplifies the norm of the combined updates. Consequently, a smaller scaling factor helps balance the output feature norm and achieves better validation performance.

C.3. Performance on Individual Tasks

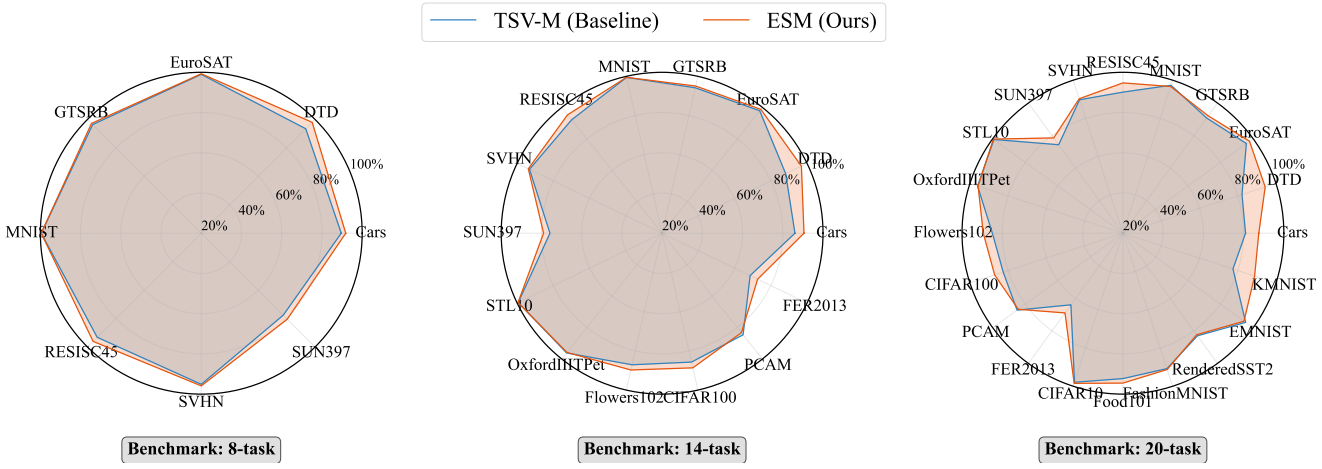
We present detailed model fusion results across multiple models and datasets, as shown in Figure 10. Our proposed ESM framework demonstrates consistent performance improvements across nearly all datasets and models.



(a) ViT-B/32



(b) ViT-B/16

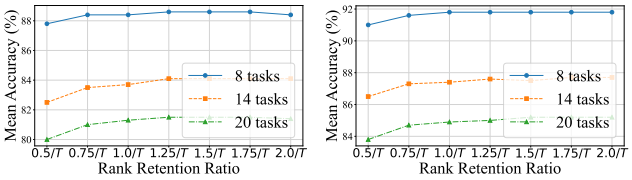


(c) ViT-L/14

Figure 10. Performance comparison between the baseline method TSV-M [11] and our proposed ESM framework.

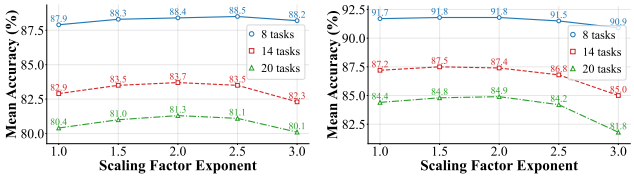
Table 8. Ablation study on the order of the three-level scaling in Polarized Scaling.

Scaling Order	ViT-B/32			ViT-B/16		
	8 tasks	14 tasks	20 tasks	8 tasks	14 tasks	20 tasks
(1) Inter-Dimension, (2) Inter-Task, (3) Inter-layer	$88.0_{(94.1)}$	$83.2_{(91.4)}$	$80.8_{(88.3)}$	$91.6_{(96.7)}$	$87.1_{(93.7)}$	$84.7_{(90.9)}$
(1) Inter-Task, (2) Inter-Dimension, (3) Inter-layer	$88.4_{(95.3)}$	$83.7_{(92.0)}$	$81.3_{(88.9)}$	$91.8_{(97.0)}$	$87.4_{(94.1)}$	$84.9_{(91.1)}$



(a) ViT-B/32 (b) ViT-B/16

Figure 11. Ablation study on the impact of rank retention ratio on merged model performance. T denotes the number of tasks.



(a) ViT-B/32 (b) ViT-B/16

Figure 12. Impact of the scaling factor exponent on merged model performance.

C.4. Order of Scaling Across Tasks, Dimensions, and Layers

As shown in Algorithm 1, the default order of Polarized Scaling in our experiments is: first across tasks, then across dimensions, and finally across layers. Since inter-layer scaling is designed to emphasize task consensus, it is applied by default to the final fused task matrices. The order of inter-task and inter-dimension scaling, however, can be interchanged. Therefore, we ablate the sequence of these two operations in Table 8. The results indicate that our default order, which applies inter-task scaling first followed by inter-dimension scaling, yields better performance.

C.5. Ablation Study on Rank k Selection

In our method and experiments, the default setting uses $k = \lfloor d_{\text{out}}/T \rfloor$ as the rank budget for low-rank decomposition of each task matrix, where T denotes the number of tasks and d_{out} represents the original output dimension. We conduct an ablation study on the selection of rank k, as shown in Figure 11. The results demonstrate that the merged model exhibits robustness to the choice of rank k, maintaining comparable performance across a wide range of values ($\lfloor 0.5 \cdot d_{\text{out}}/T \rfloor \sim \lfloor 2.0 \cdot d_{\text{out}}/T \rfloor$).

C.6. Ablation Study on the Exponent of the Scaling Factor

The default configuration of our method employs a power of 2 in the polarized scaling coefficient, i.e., $(\frac{\text{norm}}{\mathbb{E}[\text{norm}]})^2$. The rationale for this choice is to amplify significant parameters while suppressing redundant ones. To validate the sensitivity of our approach to this hyperparameter, we conducted an ablation study. As shown in Figure 12, the results indicate that model merging is robust across a range of exponents. The value of 2 was chosen as the default because it achieves optimal performance.

C.7. Calculation of Energy Retention

For the SVD-based method, energy retention is calculated as the ratio of the sum of squares of the retained singular values to the sum of squares of all singular values. For our ESD method, it is defined as the ratio of the sum of the retained eigenvalues to the sum of all eigenvalues. This is because the square of a singular value and an eigenvalue both correspond to the explained variance.