Towards Performance Consistency in Multi-Level Model Collaboration

- Supplementary Material -

1. NeuLig under More Models

In Table 2 of the main manuscript, we evaluate the performance of NeuLig in scenarios involving the collaboration of up to five models. To further explore its scalability, we extend this investigation to scenarios with a greater number of models. Specifically, we incorporate two additional models fine-tuned on the STL10 and SVHN datasets, increasing the total to seven models. The experimental results are presented in Table s5.

Method	GTSRB			CIFAR100			RESISC45			CIFAR10			MNIST			STL10			SVHN			Avg		
Pre-trained	32.56			64.20			60.22			89.83			48.25			15.91			8.31			45.61		
Fine-tuned	98.95			84.22			94.13			97.13			99.56			96.09			96.80			95.27		
	Mer. ↑	Ens. ↑	Gap ↓	Mer. ↑	Ens. ↑	Gap ↓	Mer. ↑	Ens. ↑	Gap ↓	Mer. ↑	Ens. ↑	Gap ↓	Mer. ↑	Ens. ↑	Gap ↓	Mer. ↑	Ens. ↑	Gap ↓	Mer. ↑	Ens. ↑	Gap ↓	Mer. ↑	Ens. ↑	Gap ↓
Multi-Task Model Collaboration Methods																							
Simple-Averaging[36]	53.78	85.95	32.17	73.27	74.92	1.65	68.97	81.14	12.17	94.40	96.78	2.38	83.04	97.65	14.61	41.65	88.06	46.41	14.99	95.99	81.00	61.44	88.64	27.20
Task-Arithmetic[15]	57.97	88.30	30.33	62.29	75.92	13.63	53.60	73.90	20.30	91.84	96.56	4.72	89.99	99.22	9.23	67.64	90.58	22.94	29.84	94.07	64.23	64.74	88.36	23.62
Ties-Merging[38]	65.17	-	-	71.14	-	-	69.33	-	-	94.63	-	-	91.73	-	-	61.32	-	-	22.81	-	-	68.02	-	-
RegMean[17]	64.17	-	-	73.12	-	-	76.22	-	-	94.81	-	-	89.63	-	-	62.80	-	-	17.29	-	-	68.29	-	-
AdaMerging[39]	90.92	92.34	1.42	69.92	76.00	6.08	84.51	83.30	1.21	92.65	96.58	3.93	97.25	98.38	1.13	96.67	90.65	6.02	10.85	96.45	85.60	77.54	90.53	12.99
WeMoE[34]	91.36	92.80	1.44	72.45	74.30	1.85	86.50	86.98	0.48	94.24	96.58	2.34	97.80	98.12	0.32	93.48	97.52	4.04	26.48	96.53	70.05	80.33	91.83	11.50
Neural Ligand																								
Ours (Semi-Supervised)	99.05	99.10	**0.05**	85.39	85.62	0.23	93.87	94.05	**0.18**	**96.33**	**96.78**	0.45	**99.58**	99.57	**0.01**	**96.94**	96.08	0.86	96.82	96.79	**0.03**	95.43 (+15.10)	95.43 (+3.60)	**0.00** (-11.50)
Ours (Supervised)	**99.20**	**99.33**	0.13	**87.26**	**87.44**	**0.18**	**94.02**	**94.38**	0.36	96.10	96.32	**0.22**	99.44	**99.87**	0.43	96.35	**96.48**	**0.13**	**96.88**	**97.20**	0.32	**95.61** (+15.28)	**95.86** (+4.03)	0.09 (-11.41)

Table s5. Results of various methods across multiple datasets, including the merging performance, the ensembling performance, and the performance gap for CLIP-ViT-B/32.

As observed, even with an increased number of collaborating models, NeuLig consistently demonstrates exceptionally low performance gaps while significantly outperforming baseline methods. Remarkably, under the semi-supervised setting, the performance gap is entirely eliminated. These results further reinforce the validity of our findings and affirm the effectiveness of NeuLig as a robust validation framework.

2. NeuLig under Other Model Types

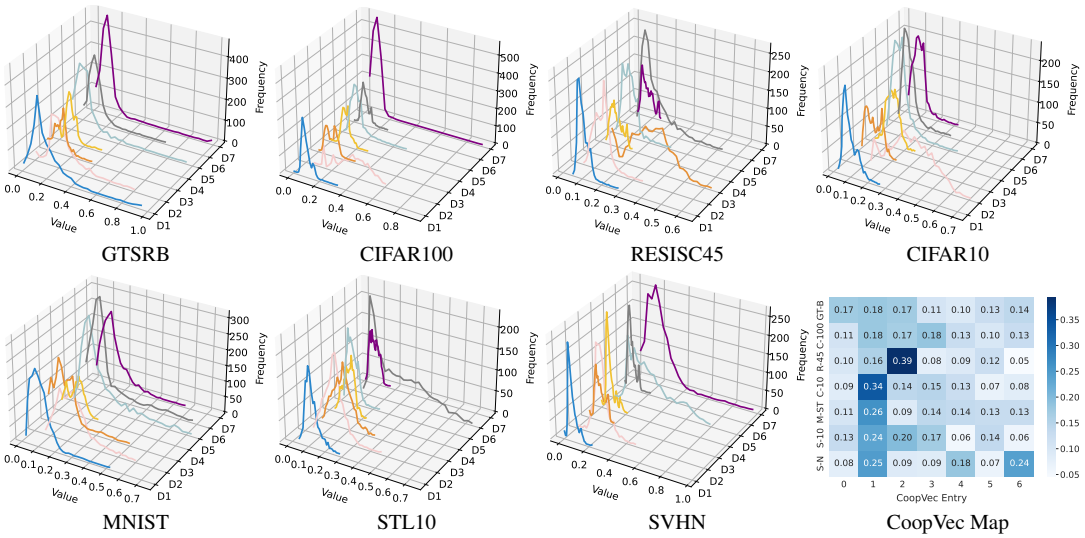
Method	GTSRB			CIFAR100			RESISC45			CIFAR10			MNIST			Avg		
Pre-trained	50.55			75.82			71.33			95.57			76.36			73.93		
Fine-tuned	99.11			91.64			96.05			98.80			99.70			97.06		
	Mer. ↑	Ens. ↑	Gap ↓	Mer. ↑	Ens. ↑	Gap ↓	Mer. ↑	Ens. ↑	Gap ↓	Mer. ↑	Ens. ↑	Gap ↓	Mer. ↑	Ens. ↑	Gap ↓	Mer. ↑	Ens. ↑	Gap ↓
Multi-Task Model Collaboration Methods																		
Simple-Averaging[36]	67.48	94.23	26.75	86.26	88.89	3.37	80.76	90.42	9.66	94.26	96.48	2.22	93.26	98.84	5.58	84.40	93.77	9.37
Task-Arithmetic[15]	68.23	94.15	25.92	85.46	89.13	3.67	80.48	90.91	10.43	93.92	97.56	3.64	93.78	98.92	5.14	84.37	94.13	9.76
Ties-Merging[38]	71.68	-	-	85.64	-	-	86.74	-	-	95.39	-	-	91.93	-	-	86.28	-	-
RegMean[17]	84.57	-	-	87.72	-	-	90.40	-	-	98.59	-	-	99.02	-	-	92.06	-	-
AdaMerging[39]	97.78	98.65	0.87	83.02	84.43	1.41	92.66	97.89	5.23	97.12	98.83	1.71	94.29	97.23	2.94	93.17	95.61	2.44
WeMoE[34]	97.90	98.56	0.66	85.86	87.22	1.36	92.69	95.43	2.74	96.97	98.71	1.74	97.44	98.80	1.36	94.17	95.74	1.57
Neural Ligand																		
Ours (Semi-Supervised)	**99.90**	**99.92**	**0.02**	**91.42**	**91.36**	**0.06**	**96.54**	96.60	**0.06**	98.97	99.12	0.15	**99.88**	**99.88**	**0.00**	**97.34** (+3.17)	97.38 (+1.64)	**0.04** (-1.53)
Ours (Supervised)	99.86	99.90	0.04	91.02	91.34	0.32	96.42	**96.65**	0.23	**99.68**	**99.62**	**0.06**	99.73	99.65	0.08	**97.34** (+3.17)	**97.43** (+1.69)	0.09 (-1.48)

Table s6. Results of various methods across multiple datasets, including the merging performance, the ensembling performance, and the performance gap for CLIP-ViT-L/14.

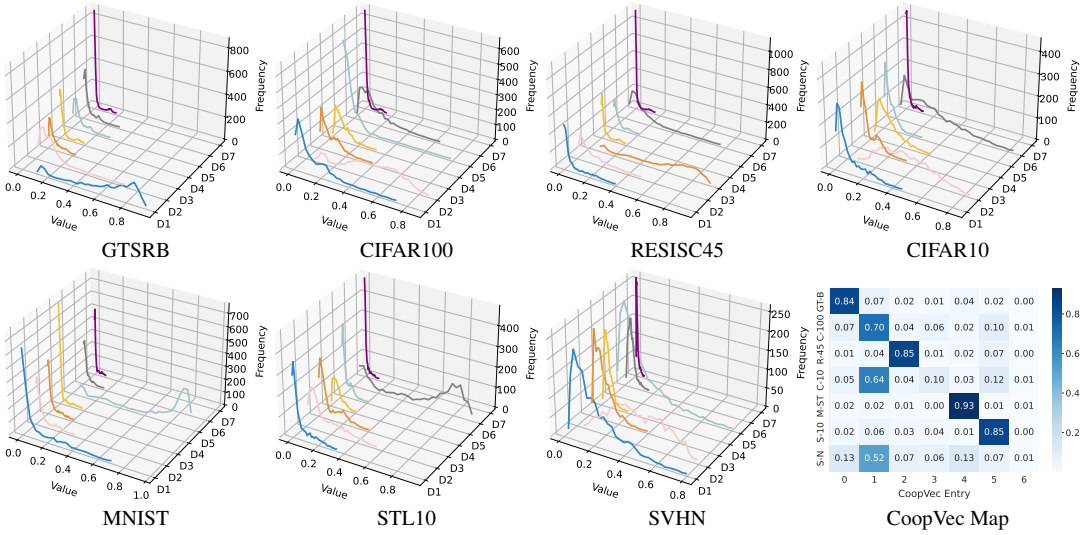
In the main manuscript, we employ two model architectures: CLIP-RN50 and CLIP-ViT-B/32. To further investigate the effectiveness of NeuLig with larger model architectures, we conduct additional experiments using CLIP-ViT-L/14 as the backbone. The results of these experiments are summarized in Table s6. It is evident that when using larger model architectures, NeuLig remains a highly effective validation framework. All baseline methods continue to exhibit relatively large performance gaps to varying degrees, whereas NeuLig consistently demonstrates minimal performance differences.

3. CoopVec Map under More Models

In Figure s8, we depict the distribution of CoopVecs at the initial training stage for a seven-model collaboration, and the final CoopVec Map derived from this distribution, while Figure s9 and Table s7 capture the variation in the diagonal elements of



(a) CLIP-RN50.



(b) CLIP-ViT-B/32.

Figure s8. CoopVec Distribution of seven different datasets and the corresponding CoopVec Map after training for one epoch. D1 to D7 represents the name of different datasets (D1-GTSRB, D2-CIFAR100, D3-RESISC45, D4-CIFAR10, D5-MNIST, D6-STL10, D7-SVHN).

Training Epoch	**Type**	**1**	**2**	**3**	**4**	**5**
CLIP-ViT-B/32 (Ori.)	Mer.	94.32	93.38	94.84	94.81	94.67
	Ens.	94.94	94.90	95.13	95.05	95.05
CLIP-ViT-B/32 (Trans.)	Mer.	93.65	93.81	94.08	94.27	94.33
	Ens.	94.34	94.26	94.21	94.43	94.42
Training Epoch	**Type**	**6**	**7**	**8**	**9**	**10**
CLIP-ViT-B/32 (Ori.)	Mer.	94.88	94.82	94.81	95.44	94.98
	Ens.	95.00	95.05	95.00	95.46	94.98
CLIP-ViT-B/32 (Trans.)	Mer.	94.16	94.36	94.02	94.11	94.01
	Ens.	94.25	94.35	94.30	94.20	94.26

Table s8. The transferability of `NeuLig`. Ori. refers to the original performance, while Trans. indicates the performance after directly applying Portland to the other group of models.

the CoopVec Map throughout training, alongside the performance achieved using CoopVec for model collaboration. These experimental results clearly demonstrate that the conclusions presented in Section 4.4 of the main manuscript remain broadly valid and applicable in scenarios involving collaboration among a larger number of models.

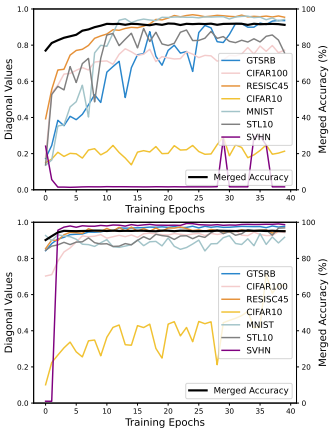


Figure s9. The variation of the diagonal values of CoopVec Map throughout the training process using CLIP-RN50 (top) and CLIP-ViT-B/32 (bottom).

Table s7. Performance when using the CoopVec Map.

Dataset	Performance			
	Mer.◆	Ens.●	Mer.◆	Ens.●
GTSRB	CLIP-RN50		CLIP-ViT-B/32	
	96.74	94.88	98.86	98.84
CIFAR100	CLIP-RN50		CLIP-ViT-B/32	
	76.84	77.33	85.97	85.59
RESISC45	CLIP-RN50		CLIP-ViT-B/32	
	92.06	92.54	93.63	93.84
CIFAR10	CLIP-RN50		CLIP-ViT-B/32	
	91.98	92.51	95.84	96.37
MNIST	CLIP-RN50		CLIP-ViT-B/32	
	99.50	99.64	99.57	99.57
STL10	CLIP-RN50		CLIP-ViT-B/32	
	89.29	92.95	96.21	96.06
SVHN	CLIP-RN50		CLIP-ViT-B/32	
	95.78	95.19	96.69	96.67
Avg.Acc	91.74 (+34.83)	92.15 (+3.57)	95.25 (+14.92)	95.28 (+3.45)

4. Transferability of Portland

In this experiment, we split each dataset into two equal-sized subsets and train two separate models on each subset, referred to as model-A and model-B for simplicity. The objective is to explore whether the Portland trained collaboratively using all model-As, can be directly transferred to scenarios where all model-Bs collaborate, thereby assessing Portland's transferability. The results are summarized in Table s8. Notably, despite the fact that Portland was not explicitly trained for the model-Bs, the performance before and after the transfer remains largely consistent, highlighting Portland's robust transferability.

5. Resilience of NeuLig With Varying Dataset Scales

In the main manuscript, we explore the performance variation of `NeuLig` under different visible dataset scales when five models collaborate. We observe that performance consistency is well-maintained even with very small dataset scales. Here, consistent with Table s5, we extend this investigation further by introducing two additional models, bringing the total model number to seven, which means we examine the impact of dataset scale on performance in the seven-model collaboration scenario. The results are presented in Table s9.

Data Scale	Type	0.01	0.05	0.1	0.15	0.2
CLIP-RN50	◆Mer.	73.63	84.63	89.63	90.46	91.05
	●Ens.	85.02	86.73	89.95	91.22	91.58
CLIP-ViT-B/32	◆Mer.	88.74	93.68	94.91	94.91	95.23
	●Ens.	93.17	94.71	95.10	95.19	95.24
Data Scale	**Type**	**0.3**	**0.4**	**0.6**	**0.8**	**1.0**
CLIP-RN50	◆Mer.	91.95	92.04	92.08	92.83	92.66
	●Ens.	91.88	92.57	92.43	92.96	92.85
CLIP-ViT-B/32	◆Mer.	95.30	95.00	94.92	94.91	95.43
	●Ens.	95.18	95.23	95.26	95.26	95.43

Table s9. The performance variation of `NeuLig` under the semi-supervised learning setup when datasets of different scales are used.

Aligned with the main manuscript, we evaluate 10 different dataset scales to cover a wide range of conditions. The experimental results reveal that `NeuLig` demonstrates strong resilience to data scale, even in scenarios involving a larger number of collaborating models. Remarkably, even under extreme conditions (e.g., at a scale of 0.1), it continues to achieve performance consistency and maintain superior performance.