NEURAL INFORMATION
PROCESSING SYSTEMS

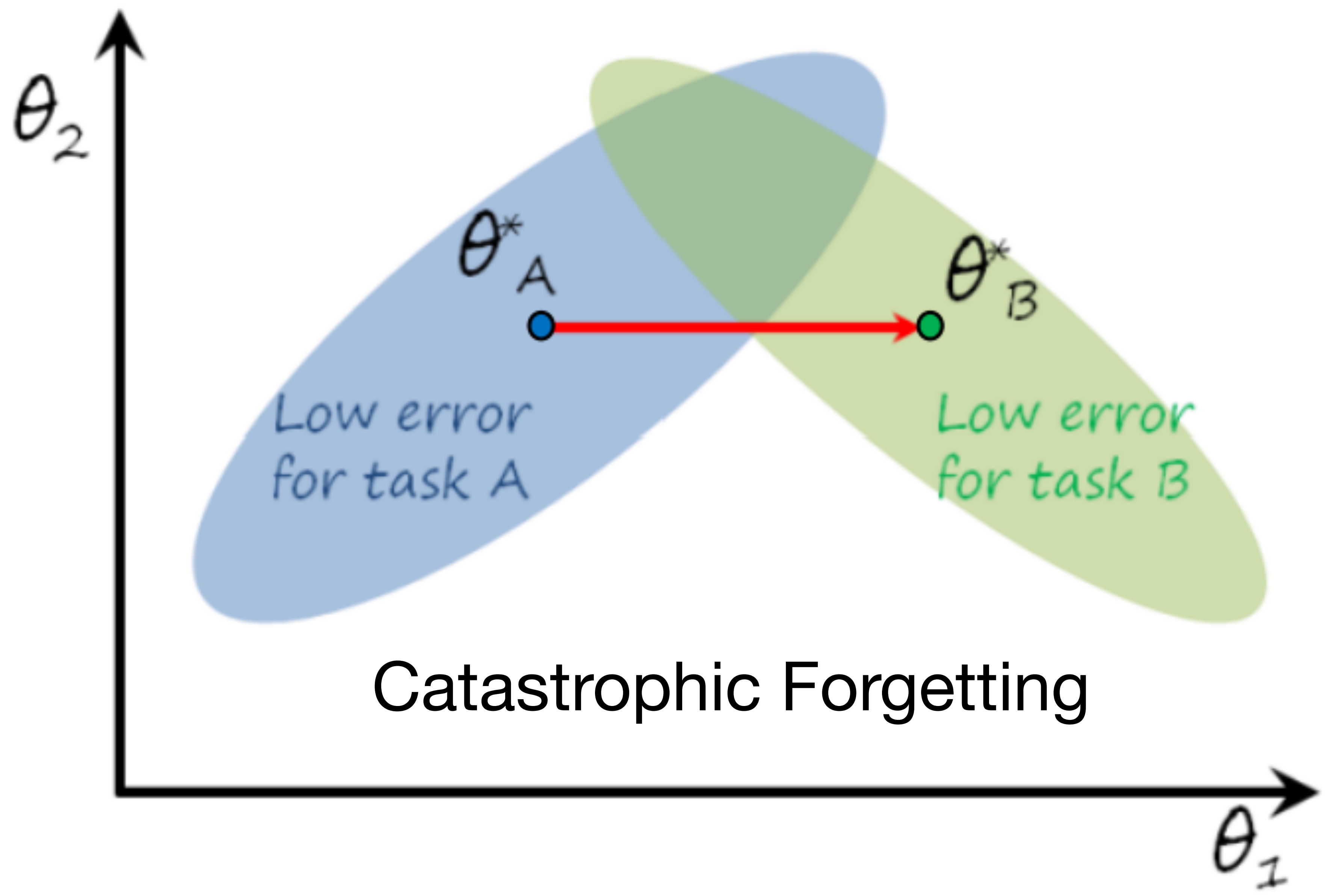
Hierarchical Decomposition of Prompt-Based Continual Learning: Rethinking Obscured Sub-optimality

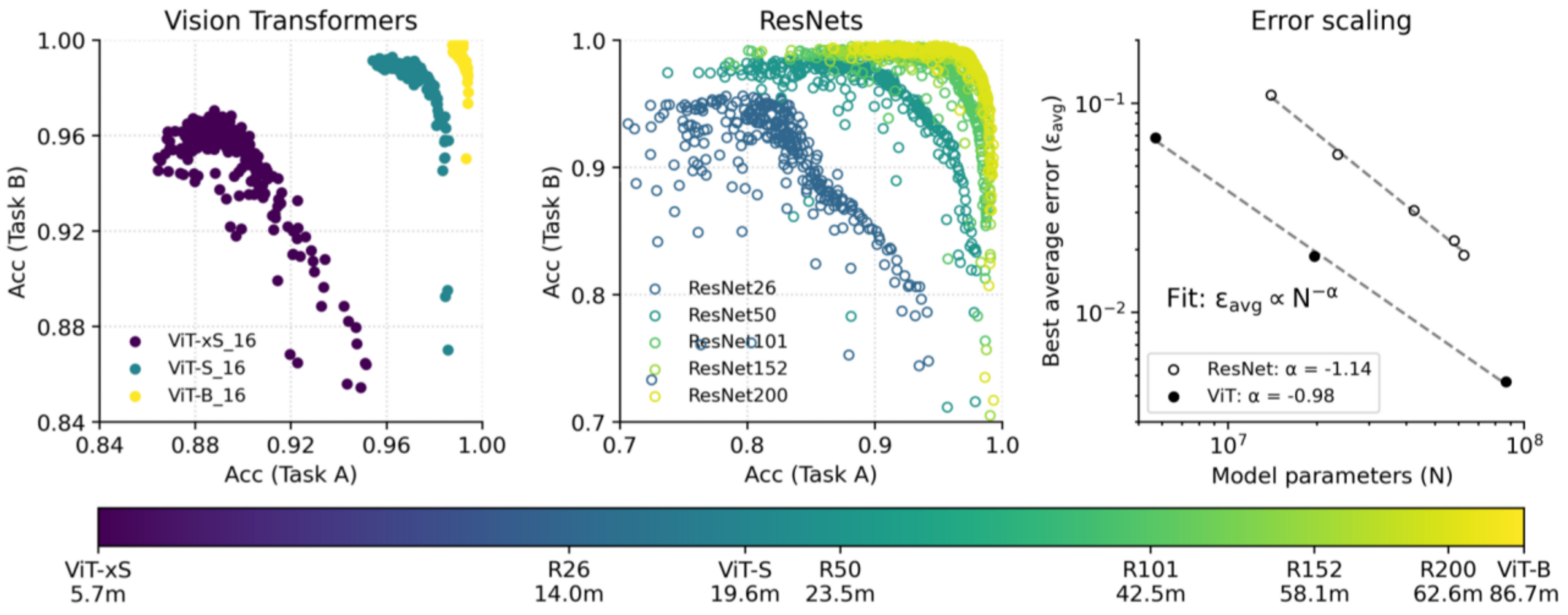
Liyuan Wang, Jingyi Xie, Xingxing Zhang, Mingyi Huang, Hang Su, Jun Zhu

Tsinghua University

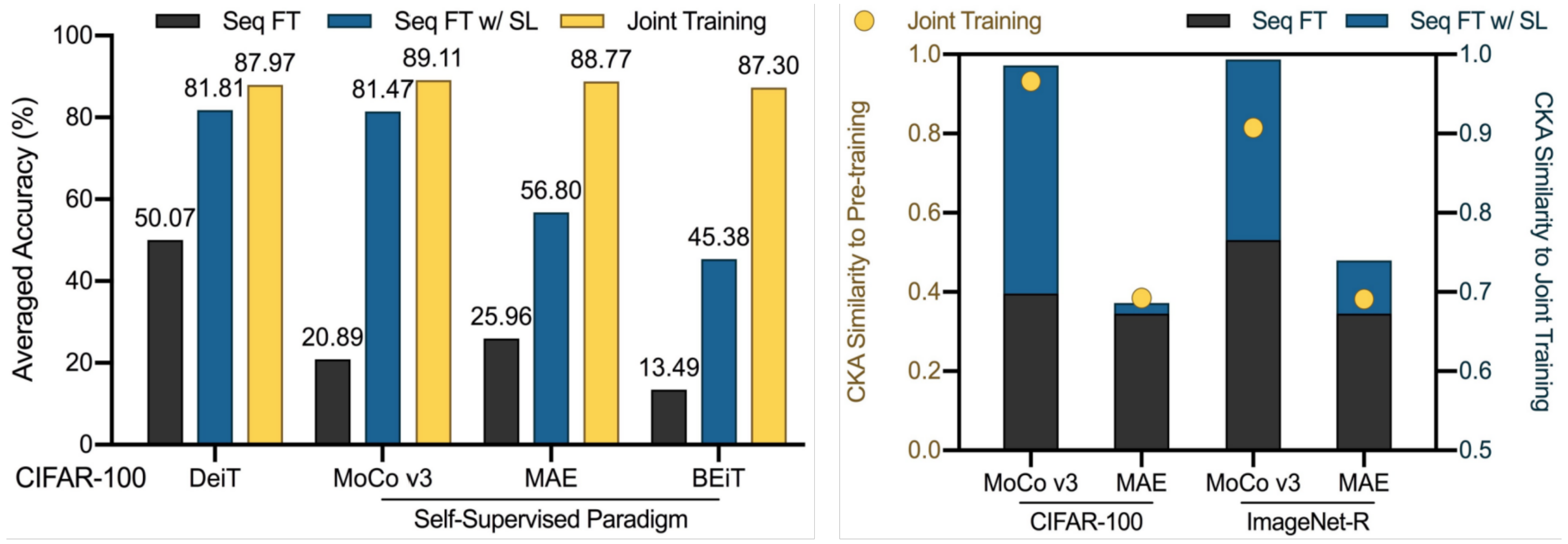
NeurIPS 2023, Spotlight

Continual Learning & the Impact of Pre-training



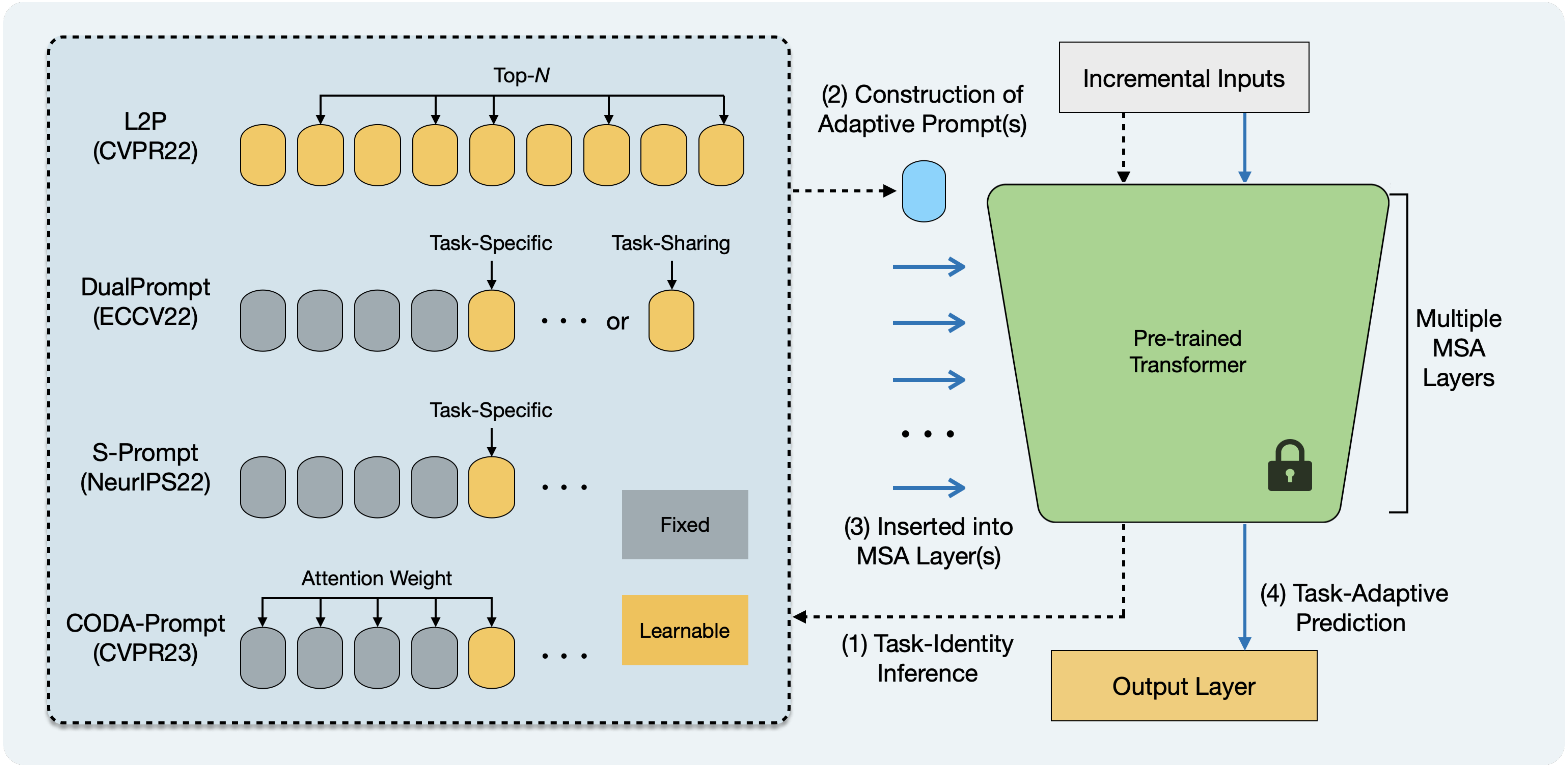


Ramasesh et al., ICLR 2022

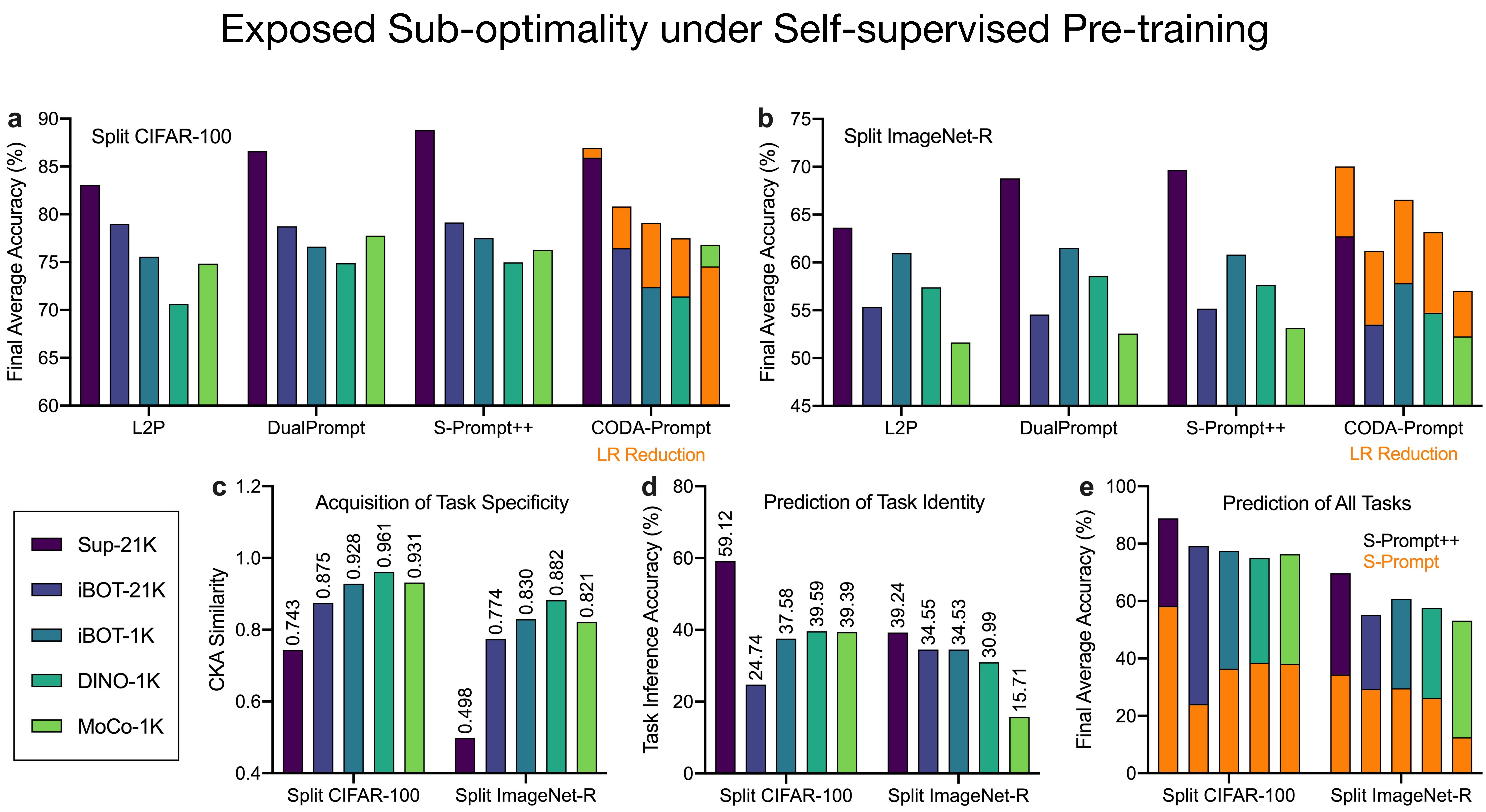


Zhang et al., ICCV 2023

State-of-the-Art Prompt-Based Approaches



Construction and Inference of Appropriate Prompts for Each Training / Testing Data.


Exposed Sub-optimality under Self-supervised Pre-training
a
Split CIFAR-100
Final Average Accuracy (%)
L2P
DualPrompt
S-Prompt++
CODA-Prompt
LR Reduction
b
Split ImageNet-R
Final Average Accuracy (%)
L2P
DualPrompt
S-Prompt++
CODA-Prompt
LR Reduction
Sup-21K
iBOT-21K
iBOT-1K
DINO-1K
MoCo-1K
c
Acquisition of Task Specificity
CKA Similarity
0.743
0.875
0.928
0.961
0.931
0.498
0.774
0.830
0.882
0.821
Split CIFAR-100
Split ImageNet-R
d
Prediction of Task Identity
Task Inference Accuracy (%)
59.12
24.74
37.58
39.59
39.39
39.24
34.55
34.53
30.99
15.71
Split CIFAR-100
Split ImageNet-R
e
Prediction of All Tasks
Final Average Accuracy (%)
S-Prompt++
S-Prompt
Split CIFAR-100
Split ImageNet-R

Hierarchical Decomposition of Continual Learning Objective

$$P(\boldsymbol{x} \in \mathcal{X}_{\bar{i},\bar{j}} | \mathcal{D}, \theta) \longrightarrow \max[P(\boldsymbol{x} \in \mathcal{X}_{\bar{i},\bar{j}} | \mathcal{D}, \theta), P(\boldsymbol{x} \in \mathcal{X}^{y} | \mathcal{D}, \theta)]$$

Within-Task Prediction (WTP) $H_{\mathrm{WTP}}(\boldsymbol{x}) = \mathcal{H}(\mathbf{1}_{\bar{j}}, \{P(\boldsymbol{x} \in \mathcal{X}_{\bar{i},j} | \boldsymbol{x} \in \mathcal{X}_{\bar{i}}, \mathcal{D}, \theta)\}_{j}),$

Task-Identity Inference (TII) $H_{\mathrm{TII}}(\boldsymbol{x}) = \mathcal{H}(\mathbf{1}_{\bar{i}}, \{P(\boldsymbol{x} \in \mathcal{X}_{i} | \mathcal{D}, \theta)\}_{i}),$

Task-Adaptive Prediction (TAP) $H_{\mathrm{TAP}}(\boldsymbol{x}) = \mathcal{H}(\mathbf{1}_{\bar{c}}, \{P(\boldsymbol{x} \in \mathcal{X}^{c} | \mathcal{D}, \theta)\}_{c}),$

Applicable to TIL / DIL / CIL

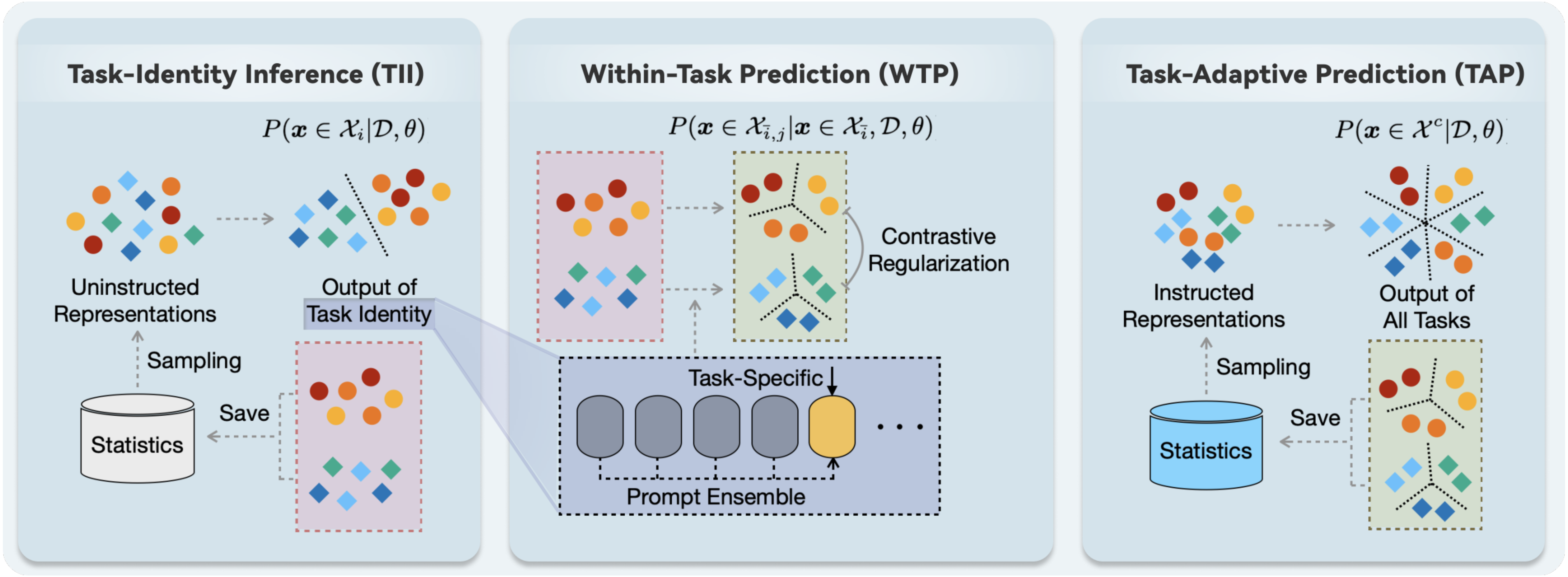
Sufficient Conditions

Theorem 1 *For continual learning with pre-training, if* $\mathbb{E}_{\boldsymbol{x}}[H_{\mathrm{WTP}}(\boldsymbol{x})] \leq \delta$, $\mathbb{E}_{\boldsymbol{x}}[H_{\mathrm{TII}}(\boldsymbol{x})] \leq \epsilon$, *and* $\mathbb{E}_{\boldsymbol{x}}[H_{\mathrm{TAP}}(\boldsymbol{x})] \leq \eta$, *we have the loss error* $\mathcal{L} \in [0, \max\{\delta + \epsilon, \eta\}]$, *regardless whether WTP, TII and TAP are trained together or separately.*

Necessary Conditions

Theorem 2 *For continual learning with pre-training, if the loss error* $\mathcal{L} \leq \xi$, *then there always exist (1) a WTP, s.t.* $H_{\mathrm{WTP}} \leq \xi$; *(2) a TII, s.t.* $H_{\mathrm{TII}} \leq \xi$; *and (3) a TAP, s.t.* $H_{\mathrm{TAP}} \leq \xi$.

Hierarchical Decomposition (HiDe-)Prompt



Explicit Optimization of the Hierarchical Components:

Task-Specific Prompt, Representation Statistics, Contrastive Regularization

Performance of Downstream Continual Learning

PTM	Method	Split CIFAR-100 FAA (↑)	Split CIFAR-100 CAA (↑)	Split CIFAR-100 FFM (↓)	Split ImageNet-R FAA (↑)	Split ImageNet-R CAA (↑)	Split ImageNet-R FFM (↓)	
Sup-21K	L2P [41]	83.06 ±0.17	88.25 ±0.01	6.58 ±0.40	63.65 ±0.12	67.25 ±0.02	7.51 ±0.17	
	DualPrompt [40]	86.60 ±0.19	90.64 ±0.01	4.45 ±0.16	68.79 ±0.31	71.96 ±0.04	4.49 ±0.14	
	S-Prompt++ [39]	88.81 ±0.18	92.25 ±0.03	3.87 ±0.05	69.68 ±0.12	72.50 ±0.04	3.29 ±0.05	
	CODA-Prompt [30]*	86.94 ±0.63	91.57 ±0.75	4.04 ±0.18	70.03 ±0.47	74.26 ±0.24	5.17 ±0.22	
	HiDe-Prompt (Ours)	**92.61** ±0.28	**94.03** ±0.01	**3.16** ±0.10	**75.06** ±0.12	**76.60** ±0.01	**2.17** ±0.19	+ 3.80% / 5.03% FAA
iBOT-21K	L2P [41]	79.00 ±0.28	85.13 ±0.05	5.55 ±0.36	55.35 ±0.28	58.62 ±0.05	3.73 ±0.53	
	DualPrompt [40]	78.76 ±0.23	86.16 ±0.02	9.84 ±0.24	54.55 ±0.53	58.69 ±0.01	5.38 ±0.70	
	S-Prompt++ [39]	79.14 ±0.65	85.85 ±0.17	9.17 ±1.33	55.16 ±0.83	58.48 ±0.18	4.07 ±0.16	
	CODA-Prompt [30]	80.83 ±0.27	87.02 ±0.20	7.50 ±0.25	61.22 ±0.35	66.76 ±0.37	9.66 ±0.20	
	HiDe-Prompt (Ours)	**93.02** ±0.15	**94.56** ±0.05	**1.33** ±0.24	**70.83** ±0.17	**73.23** ±0.08	**2.46** ±0.21	+ 12.19% / 9.61% FAA
iBOT-1K	L2P [41]	75.57 ±0.41	82.69 ±0.06	7.23 ±0.93	60.97 ±0.26	65.95 ±0.02	4.07 ±0.66	
	DualPrompt [40]	76.63 ±0.05	85.08 ±0.12	8.41 ±0.40	61.51 ±1.05	67.11 ±0.08	5.02 ±0.52	
	S-Prompt++ [39]	77.53 ±0.56	85.66 ±0.16	8.07 ±0.97	60.82 ±0.68	66.03 ±0.91	4.16 ±0.14	
	CODA-Prompt [30]	79.11 ±1.02	86.21 ±0.49	7.69 ±1.57	66.56 ±0.68	73.14 ±0.57	7.22 ±0.38	
	HiDe-Prompt (Ours)	**93.48** ±0.11	**95.02** ±0.01	**1.00** ±0.24	**71.33** ±0.21	**73.62** ±0.13	**2.79** ±0.26	+ 14.37% / 4.77% FAA
DINO-1K	L2P [41]	70.65 ±0.57	79.02 ±0.01	9.46 ±1.68	57.40 ±0.23	62.56 ±0.20	3.58 ±0.28	
	DualPrompt [40]	74.90 ±0.21	83.98 ±0.16	10.26 ±0.62	58.57 ±0.45	64.89 ±0.15	5.80 ±0.21	
	S-Prompt++ [39]	74.97 ±0.46	83.82 ±0.39	7.78 ±0.66	57.64 ±0.16	63.79 ±0.05	5.08 ±0.31	
	CODA-Prompt [30]	77.50 ±0.64	84.81 ±0.30	8.10 ±0.01	63.15 ±0.39	69.73 ±0.25	6.86 ±0.11	
	HiDe-Prompt (Ours)	**92.51** ±0.11	**94.25** ±0.01	**0.99** ±0.21	**68.11** ±0.18	**71.70** ±0.01	**3.11** ±0.17	+ 15.01% / 4.96% FAA
MoCo-1K	L2P [41]	74.85 ±0.28	83.14 ±0.03	6.51 ±0.95	51.64 ±0.19	58.87 ±0.24	**2.37** ±0.59	
	DualPrompt [40]	77.77 ±0.68	85.31 ±0.07	6.61 ±1.08	52.57 ±0.82	60.65 ±0.16	2.73 ±0.49	
	S-Prompt++ [39]	76.30 ±0.54	83.88 ±0.12	14.67 ±0.64	53.15 ±1.10	60.03 ±0.95	4.11 ±1.84	
	CODA-Prompt [30]	76.83 ±0.34	84.97 ±0.23	12.60 ±0.02	55.75 ±0.26	65.49 ±0.36	10.46 ±0.04	
	HiDe-Prompt (Ours)	**91.57** ±0.20	**93.70** ±0.01	**1.19** ±0.18	**63.77** ±0.49	**68.26** ±0.01	3.57 ±0.96	+ 13.80% / 8.02% FAA

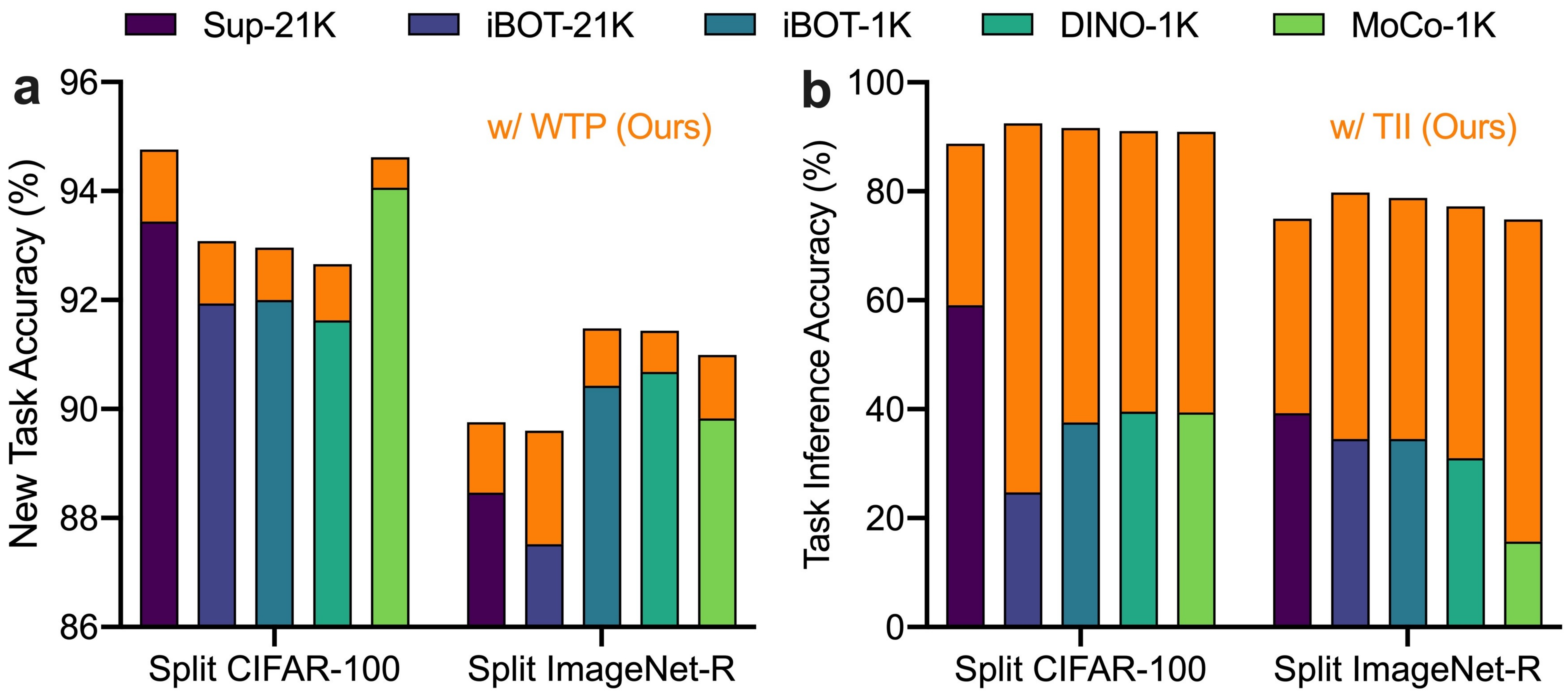
Self-supervised Pre-training

Performance of Downstream Continual Learning

Baseline	Split CIFAR-100					Split ImageNet-R				
	Sup-21K	iBOT-21K	iBOT-1K	DINO-1K	MoCo-1K	Sup-21K	iBOT-21K	iBOT-1K	DINO-1K	MoCo-1K
Naive Architecture	85.11	73.05	72.20	73.74	75.67	60.22	48.00	53.68	54.33	48.77
WTP	87.86	78.86	75.93	75.15	77.15	71.57	55.16	60.86	57.61	53.21
WTP+TII	88.05	80.77	78.90	76.27	77.78	73.76	55.19	61.22	58.41	53.08
WTP+TAP	89.85	84.23	86.04	84.76	85.17	72.57	60.01	67.13	64.26	58.36
WTP+TII+TAP	92.50	90.21	90.52	88.93	89.28	74.89	70.44	70.66	66.78	63.59
WTP+TII+TAP w/ CR	**92.61**	**93.02**	**93.48**	**92.51**	**91.57**	**75.06**	**70.83**	**71.33**	**68.11**	**63.77**

Ablation Study:
All components are effective.

Detailed Analysis:
WTP and TII are improved.



Discussion and Conclusion

1. Sub-optimality of current prompt-based approaches is exposed under the more realistic self-supervised pre-training.

2. Our theoretical analysis decomposes the objective of continual learning with pre-training into three hierarchical components.

3. We propose HiDe-Prompt to optimize the hierarchical components explicitly, which achieves outstanding performance.

4. The proposed framework can be generalized to other parameter-efficient fine-tuning techniques (Adapter, LoRA, FiLM…)

5. The proposed framework is potentially related to biological learning in selective activation of memory and non-memory cells.

Thank You!

Code: https://github.com/thu-ml/HiDe-Prompt

Paper Link: https://arxiv.org/abs/2310.07234