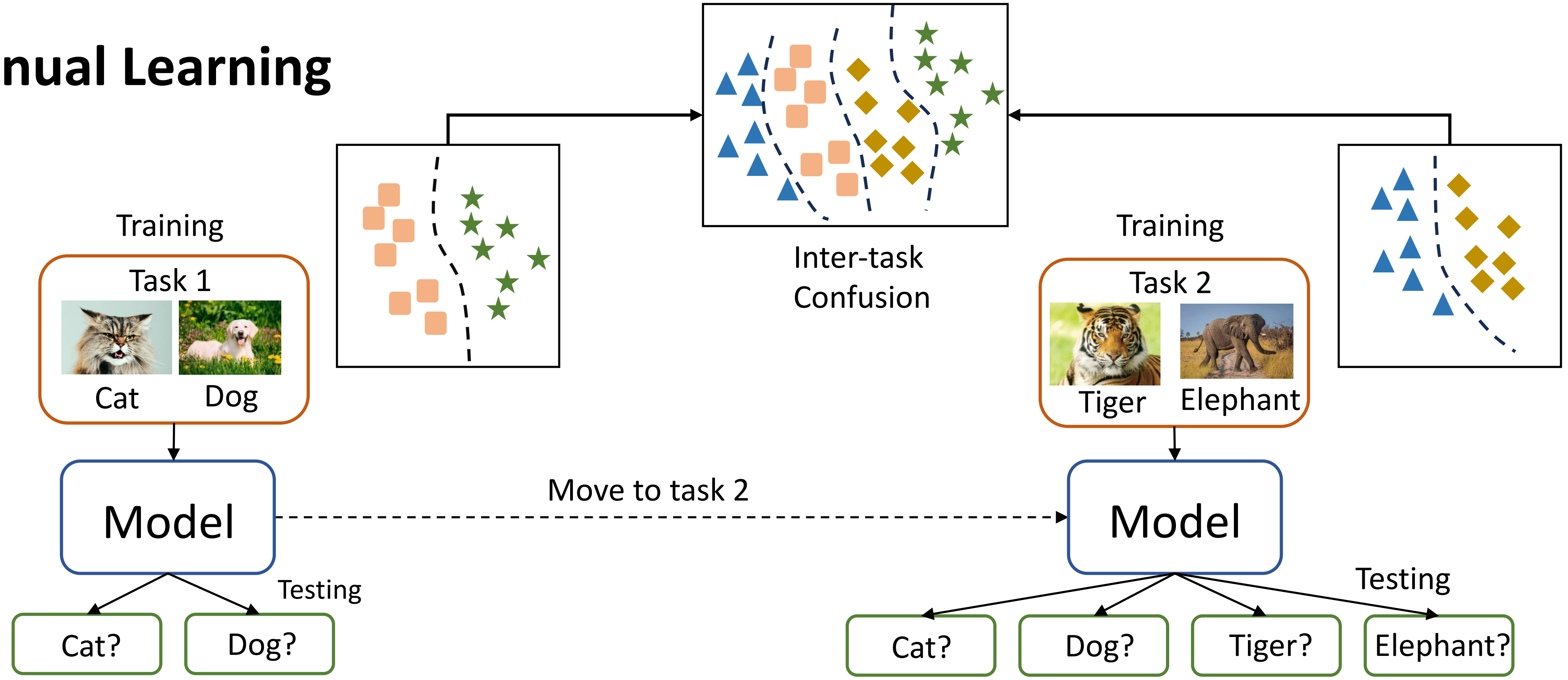
西安交通大學 XI'AN JIAOTONG UNIVERSITY
中国科学院深圳先进技术研究院 SHENZHEN INSTITUTE OF ADVANCED TECHNOLOGY CHINESE ACADEMY OF SCIENCES
The 38th Annual AAAI Conference on Artificial Intelligence
FEBRUARY 20-27, 2024 | VANCOUVER, CANADA

Evolving Parameterized Prompt Memory for Continual Learning

Muhammad Rifki Kurniawan[1], **Xiang Song**[1], **Zhiheng Ma**[3], **Yuhang He**[2], **Yihong Gong**[1,2], **Qi Yang**[4], **Xing Wei**[1]

[1]School of Software Engineering, Xi'an Jiaotong University
[2]College of Artificial Intelligence, Xi'an Jiaotong University
[3]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[4]School of Computer Science and Technology, Xi'an Jiaotong University

Continual Learning



Training 1: These are {cat, dog}

Test-time 1: What is this [img] among {cat, dog}?

Training 2: These are {tiger, elephant}

Test-time 2: What is this [img] among {cat, dog, tiger, elephant}?

Continual Learning from Pre-trained

Existing CL From Pre-trained

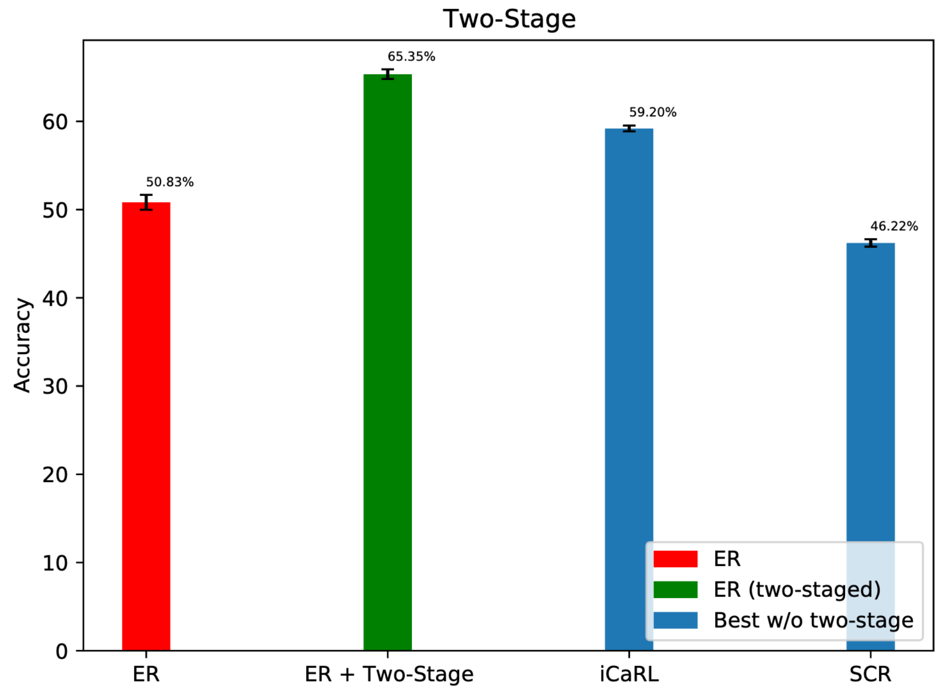
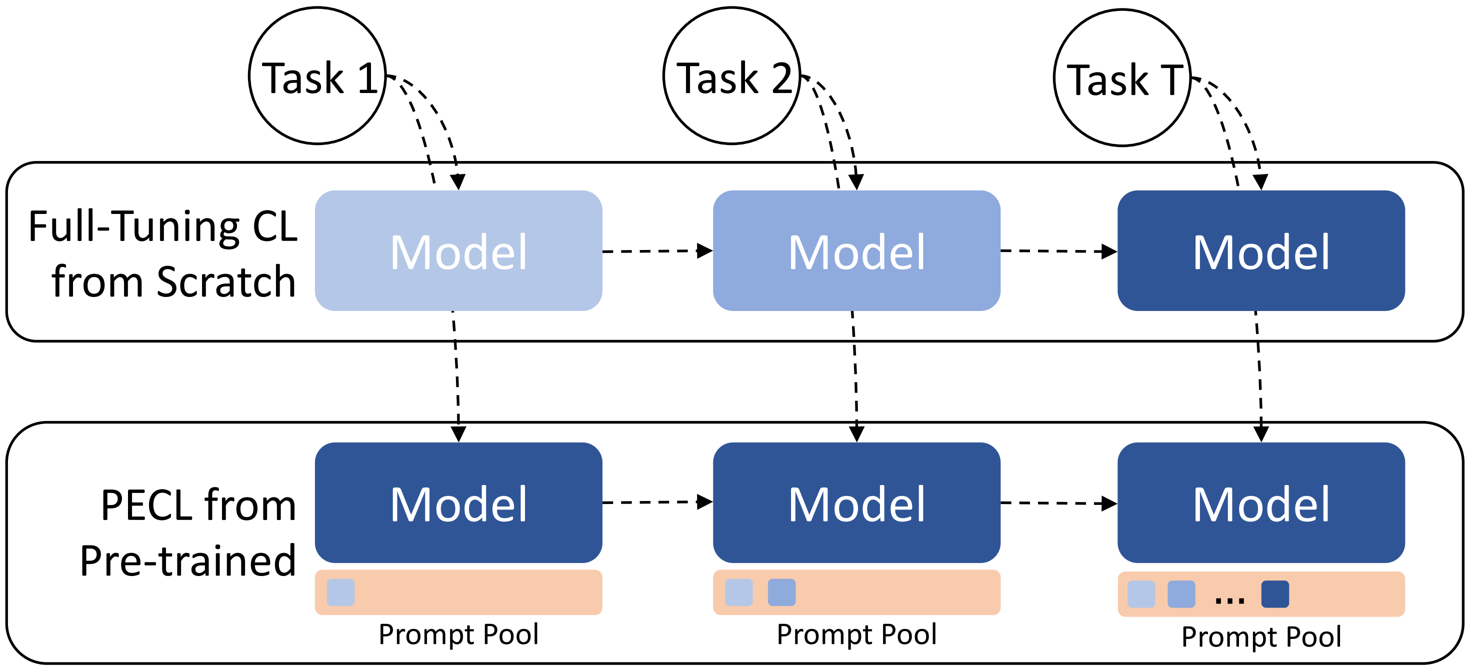


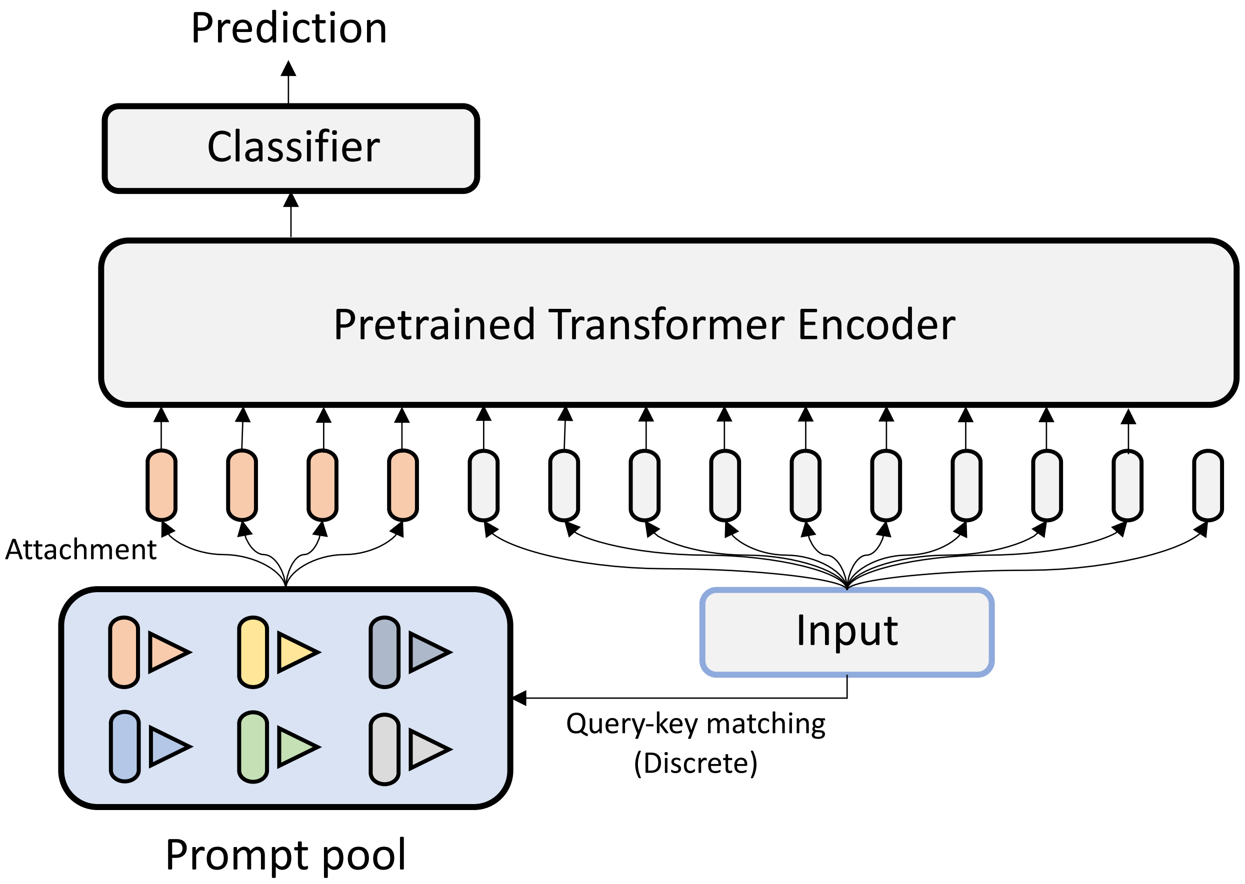
Image from [1].

From Full-tuning CL to Parameter Efficient CL

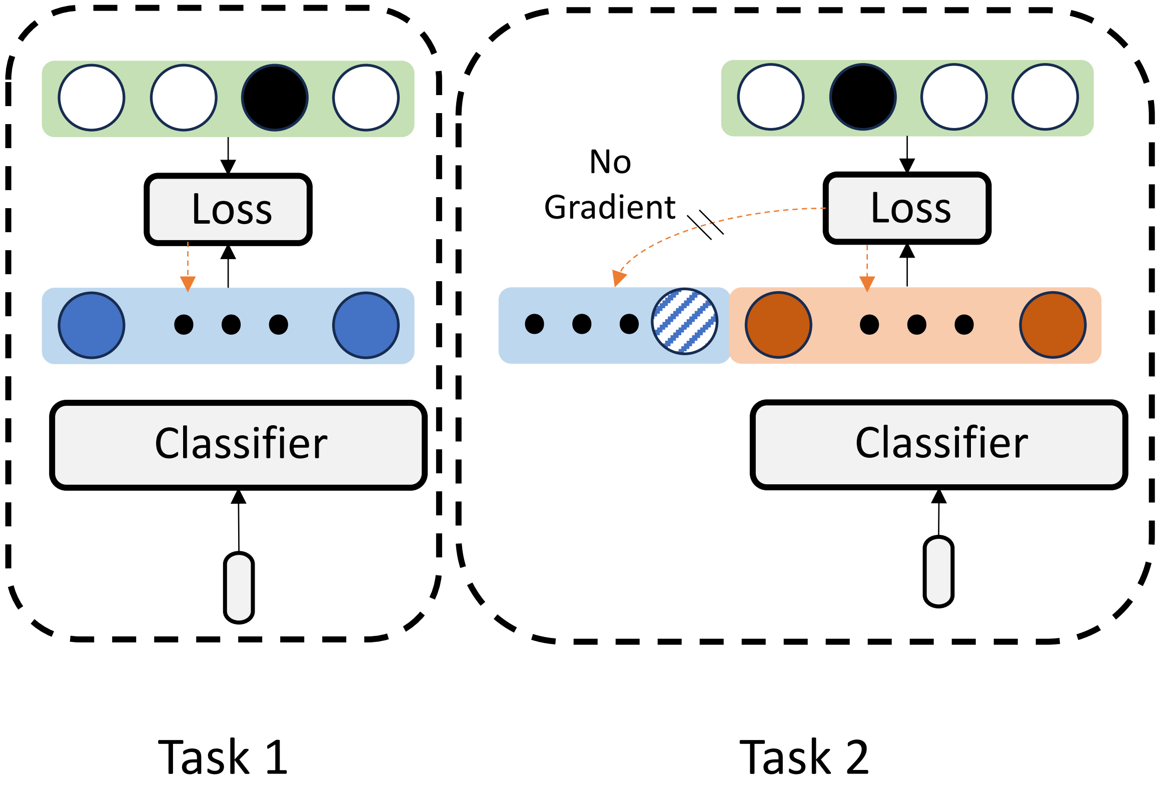


[1] K. Y. Lee, Y. Zhong, and Y. X. Wang, "Do Pre-Trained Models Benefit Equally in Continual Learning?" in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), January 2023, pp. 6485-6493.

Learning To Prompt In Continual Learning



Architecture of Common
Prompt-based CL
(L2P, DualPrompt, S-Prompt)



Learning Classifier in Isolation
(Not backward Compatible)

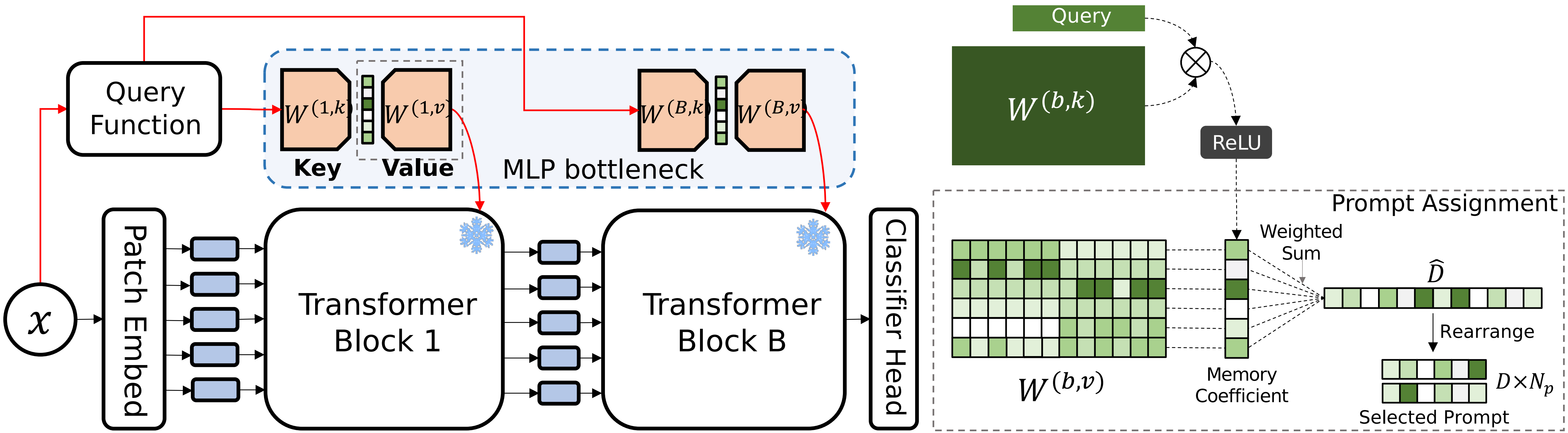
Continual Learning from Pre-trained

Method	**Prompt Selection**	**Dynamically Expand?**	**Backward Compatible?**	**Additional Params M**	**Additional Params %**
L2P	Discrete	✓	✗	0.89	1.21
DualPrompt	Discrete	✓	✗	0.95	1.28
CODA-P	Continuous	✓	✗	3.84	1.25
EvoPrompt	**Continuous**	✗	✓	**0.29**	**0.69**

Ours, Evolving Prompt (EvoPrompt):

- **Continuous** prompting
- **Non-expanded** prompt memory
- **Backward-compatible** classifier

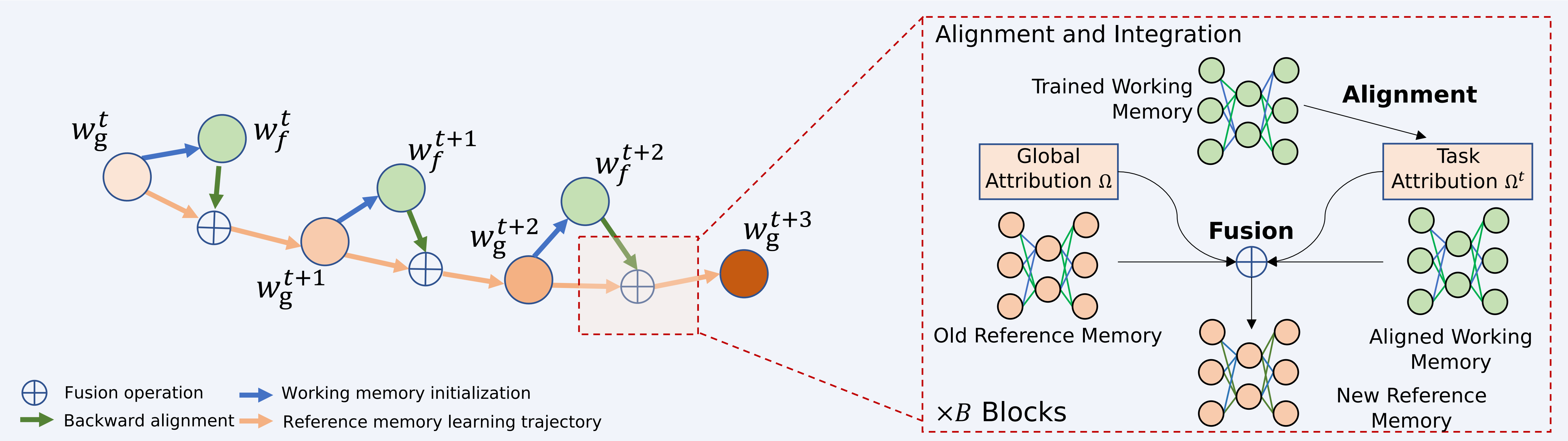
Parameterized Prompt Memory



$$
\begin{aligned}
\boldsymbol{p}_b &= f_{\mathbf{W}^{(b)}}\left(q(x)\,;\mathbf{W}^{(b)}\right),\\
&= \underbrace{\mathrm{ReLU}\left(q(x)\cdot\mathbf{W}^{(b,k)}\right)}_{\text{Query-key matching} \rightarrow \text{Prompt Coefficient}}\cdot\mathbf{W}^{(b,v)},
\end{aligned}
$$

How to learn this memory without **catastrophic forgetting**?

Evolving Prompt Memory via Incremental Fusion



Alignment:

$$\hat{\mathbf{W}}^t_{f,\ell} \leftarrow \mathrm{diag}\left(\frac{1}{\beta_\ell}\right)\mathbf{P}_\ell \mathbf{W}^t_{f,\ell}\mathbf{P}_{\ell-1}\mathrm{diag}\left(\frac{1}{\beta_{\ell-1}}\right)$$

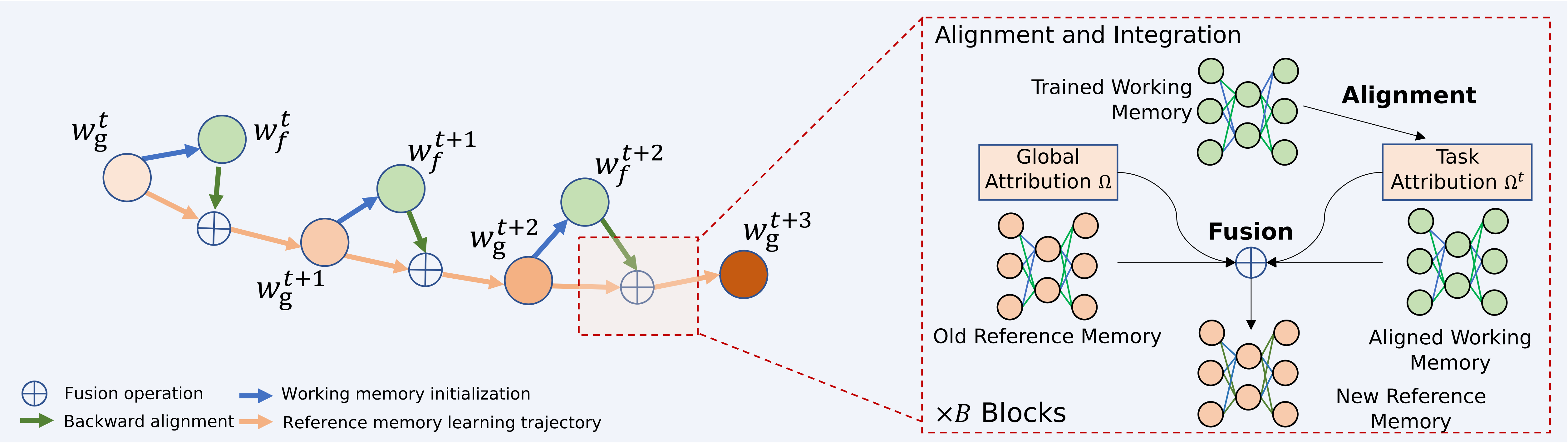
Fusion:

$$\mathbf{W}^{t+1}_{g,\ell} \leftarrow \lambda_\ell \hat{\mathbf{W}}^t_{f,\ell} + (1-\lambda_\ell)\mathbf{W}^t_{g,\ell},$$

Solving optimal transport (OT) problem:

$$\mathbf{P}_\ell = \min_{\mathbf{P}_\ell \in \mathbb{R}_+^{N_\ell \times N_\ell}} \mathrm{tr}\left(\mathbf{P}_\ell^T \mathbf{D}_\ell\right) = \mathrm{OT}\left(\alpha_\ell, \beta_\ell, \mathbf{D}_\ell\right),$$
$$\text{s.t. } \mathbf{P}_\ell \mathbf{1}_\ell = \alpha_\ell,\ \mathbf{P}_\ell^T \mathbf{1}_n = \beta_\ell,$$

Evolving Prompt Memory via Incremental Fusion



Attribution-aware aggregation momentum:

Set to 0.5

$\Omega^t_{\ell(j)} - \Omega_{\ell(j)}$

$$\lambda_\ell = \lambda_{\text{base},\ell} + \Delta\hat{\Omega}_\ell \lambda_{\text{adj},\ell}, \quad \lambda \in [0,1],$$

$$\lambda_{\text{adj},\ell} = \begin{cases} \lambda_{\text{base},\ell}, & \text{if } \hat{\Omega}^t_\ell - \hat{\Omega}_\ell < 0 \\ 1 - \lambda_{\text{base},\ell}, & \text{if } \hat{\Omega}^t_\ell - \hat{\Omega}_\ell \geq 0. \end{cases}$$

Momentum offset

Computing local task-specific attribution and global attribution:

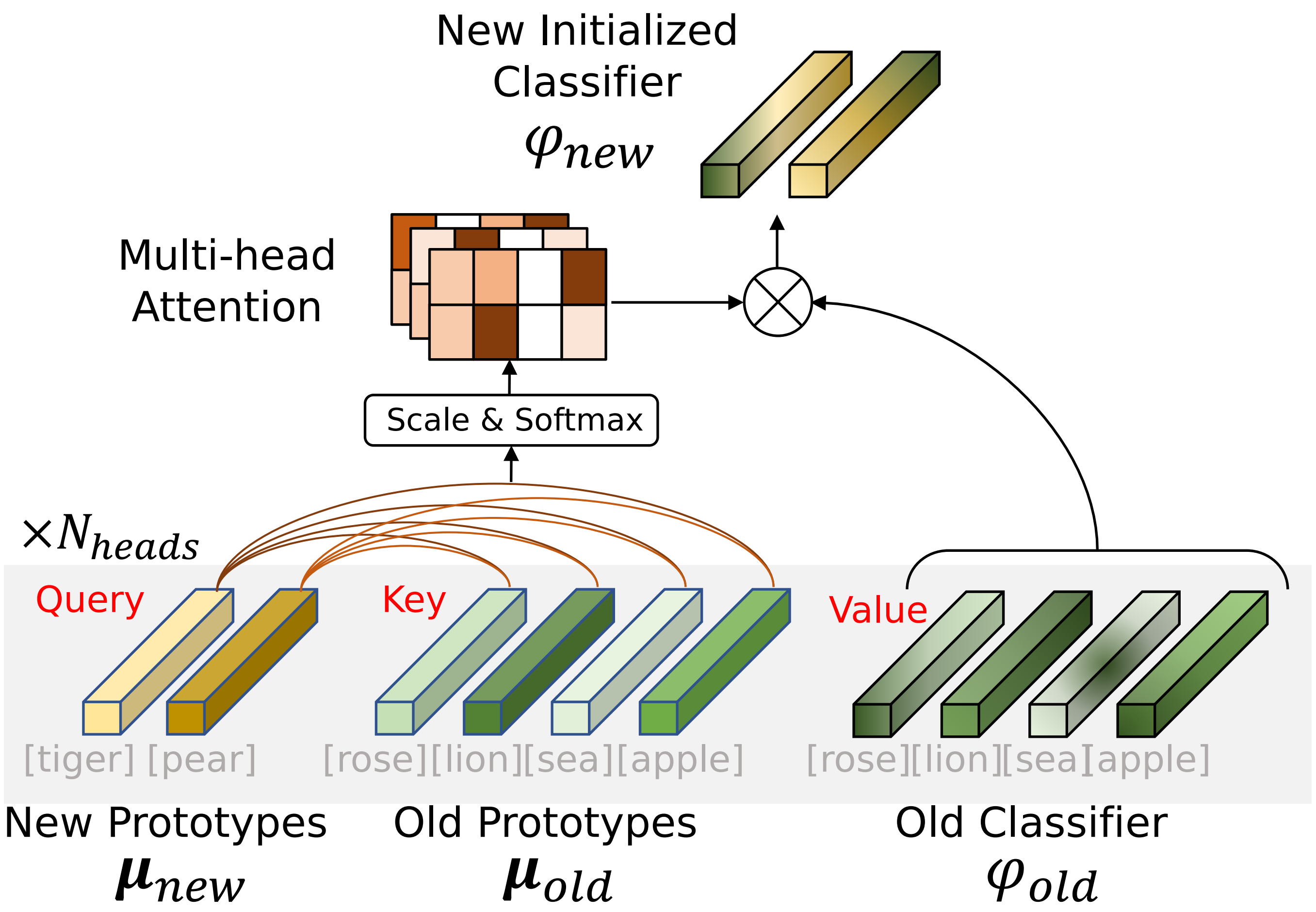
$$\Omega^t_{\ell(j)} = \frac{1}{|\mathcal{X}^t|} \sum_{x^t_i \in \mathcal{X}^t} \text{ReLU}\left(f_{n_{\ell(j)}}\left(x^t_i\right)\right), \forall i, \hat{y}_i = y_i,$$

Task-specific attribution

$$\Omega_{\ell(j)} = \max\left(\Omega_{\ell(j)}, \Omega^t_{\ell(j)}\right),$$

Global attribution

Compositional Classifier Initialization (CCI)



Prototypical attention:

$$\varphi_{new} = \mathrm{PA} = \mathrm{Softmax}\left(\frac{d\left(\boldsymbol{\mu}_{new}, \boldsymbol{\mu}_{old}\right)}{\tau}\right)\boldsymbol{\varphi}_{old},$$

Extending PA into multi-head:

$$\begin{aligned}\varphi_{new} &= \mathrm{MPA}\left(\boldsymbol{\mu}_{new}, \boldsymbol{\mu}_{old}, \boldsymbol{\varphi}_{old}\right) \\ &= \mathrm{Concat}\left(\mathrm{PA}_1\left(\boldsymbol{\mu}_{new,1}, \boldsymbol{\mu}_{old,1}, \boldsymbol{\varphi}_{old,1}\right),\right. \\ &\qquad \left.\ldots, \mathrm{PA}_k\left(\boldsymbol{\mu}_{new,k}, \boldsymbol{\mu}_{old,k}, \boldsymbol{\varphi}_{old,k}\right)\right).\end{aligned}$$

Empirical Results on Class Incremental Learning

- Split CIFAR-100: 50,000 training images, 1,000 testing images, 100 classes.

Method	Split CIFAR-100 5 Steps Acc.(↑)	5 Steps Forget.(↓)	10 Steps Acc.(↑)	10 Steps Forget.(↓)	20 Steps Acc.(↑)	20 Steps Forget.(↓)	Avg Acc.(↑)	Avg Forget.(↓)
FT-seq	73.17 ±0.75	2.95 ±0.56	62.77 ±2.30	20.73 ±2.05	55.97 ±2.95	32.74 ±2.97	63.97 (+0.00)	18.81 (−0.00)
LP-seq	71.69 ±0.61	**1.36 ±0.27**	66.90 ±0.53	13.08 ±0.32	60.98 ±0.74	21.27 ±1.20	66.52 (+2.55)	11.90 (−6.91)
NME-seq	78.30	7.70	78.33	**1.14**	78.33	**2.68**	78.32 (+14.35)	**3.84 (−14.97)**
L2P	86.53 ±0.14	7.67 ±0.20	84.97 ±8.21	8.21 ±0.22	83.39 ±0.41	10.18 ±0.24	84.96 (+20.99)	8.69 −10.12)
DualPrompt	88.26 ±0.33	5.72 ±0.43	86.83 ±0.37	6.21 ±0.35	84.11 ±0.45	8.75 ±0.38	86.40 (+22.43)	6.89 (−11.92)
ESN	88.09 ±0.21	5.18 ±0.13	85.96 ±0.14	4.54 ±0.35	82.71 ±0.51	6.44 ±0.31	85.59 (+21.62)	5.39 (−13.42)
CODA-P-S	88.90 ±0.26	6.29 ±0.27	86.33 ±0.25	6.29 ±0.52	81.71 ±0.47	9.41 ±0.22	85.65 (+21.68)	7.33 (−11.48)
CODA-P	**89.16 ±0.26**	6.08 ±0.33	87.31 ±0.14	5.95 ±0.41	81.69 ±0.38	9.85 ±0.58	86.05 (+22.08)	7.29 (−11.52)
EvoPrompt-S	88.69 ±0.16	9.93 ±0.22	87.95 ±0.13	2.38 ±0.14	**84.98 ±0.36**	3.42 ±0.39	**87.20 (+23.23)**	5.24 (−13.57)
EvoPrompt	88.97 ±0.41	10.12 ±0.35	**87.97 ±0.30**	2.60 ±0.42	84.64 ±0.14	3.98 ±0.24	87.19 (+23.22)	5.57 (−13.24)
Upper-bound†	90.85 ±0.12	-	90.85 ±0.12	-	90.85 ±0.12	-	90.85	-

+1.15% Acc
-2.05% Forget

- Split ImageNet-R: 24,000 training images, 6,000 testing images, 200 classes.

Method	Split ImageNet-R 5 Steps Acc.(↑)	5 Steps Forget.(↓)	10 Steps Acc.(↑)	10 Steps Forget.(↓)	20 Steps Acc.(↑)	20 Steps Forget.(↓)	Avg Acc.(↑)	Avg Forget.(↓)
FT-seq	61.41 ±0.38	5.76 ±0.48	50.28 ±2.29	24.28 ±1.73	39.25 ±0.90	40.38 ±0.77	50.31 (+0.00)	23.48 (−0.00)
LP-seq	59.83 ±0.33	**1.50 ±0.41**	55.30 ±0.12	7.85 ±0.10	51.97 ±0.34	13.87 ±0.21	53.64 (+3.33)	7.74 (−15.74)
NME-seq	61.06	6.64	61.40	**0.76**	61.76	2.89	61.41 (+11.10)	**3.43 (−20.05)**
L2P	66.63 ±0.33	6.65 ±0.38	64.05 ±0.39	10.05 ±0.26	60.34 ±0.17	14.44 ±0.61	63.67 (+13.36)	10.38 (−13.10)
DualPrompt	71.06 ±0.35	4.19 ±0.25	69.71 ±0.25	5.44 ±0.12	66.26 ±0.46	8.74 ±0.33	69.01 (+18.70)	6.12 (−17.36)
ESN	73.42 ±0.40	3.79 ±0.55	71.07 ±0.29	4.99 ±0.49	64.77 ±0.71	6.65 ±1.24	69.75 (+19.44)	5.14 (−18.34)
CODA-P-S	73.80 ±0.40	5.56 ±0.64	71.95 ±0.41	5.92 ±0.35	69.67 ±0.35	6.23 ±0.40	71.81 (+21.50)	5.90 (−17.58)
CODA-P	73.77 ±0.48	6.60 ±0.52	72.42 ±0.40	6.26 ±0.61	70.18 ±0.43	5.53 ±0.21	72.12 (+21.81)	6.13 (−17.35)
EvoPrompt-S	76.79 ±0.23	9.84 ±0.15	76.22 ±0.16	2.33 ±0.24	**74.68 ±0.51**	2.70 ±0.19	75.90 (+25.59)	4.96 (−18.52)
EvoPrompt	**77.16 ±0.18**	9.89 ±0.30	**76.83 ±0.08**	2.78 ±0.06	74.41 ±0.23	**2.56 ±0.22**	**76.13 (+25.82)**	5.08 (−18.40)
Upper-bound†	79.13 ±0.18	-	79.13 ±0.18	-	79.13 ±0.18	-	79.13	-

+4.01% Acc
-1.05% Forget

Empirical Results on Domain Incremental Learning

- Total:
  - fixed 50 classes
  - 11 domains
  - 120,000 images
- Training: 8 domains
- Testing: 3 unseen domains
- Metrics using final accuracy.

Method	Test Acc. (↑)	Δ Acc. (↑)
NME-seq	78.20	+00.00
EWC †	74.82 ±0.60	−3.38
LwF †	75.45 ±0.40	−2.75
L2P†	78.33 ±0.06	+0.13
S-iPrompts‡	83.13 ±0.51	+4.93
S-liPrompts‡	89.06 ±0.86	+10.86
ESN‡	91.80 ±0.31	+13.60
EvoPrompt-S	94.77 ±0.50	+16.57
EvoPrompt	**95.27** ±0.15	**+17.07**
Upper-bound	91.32 ±0.23	-

+3.47% Acc

AAAI 2024

Empirical Results on Online Continual Incremental

- The model encounters the samples in a single pass, technically with epoch set to 1.

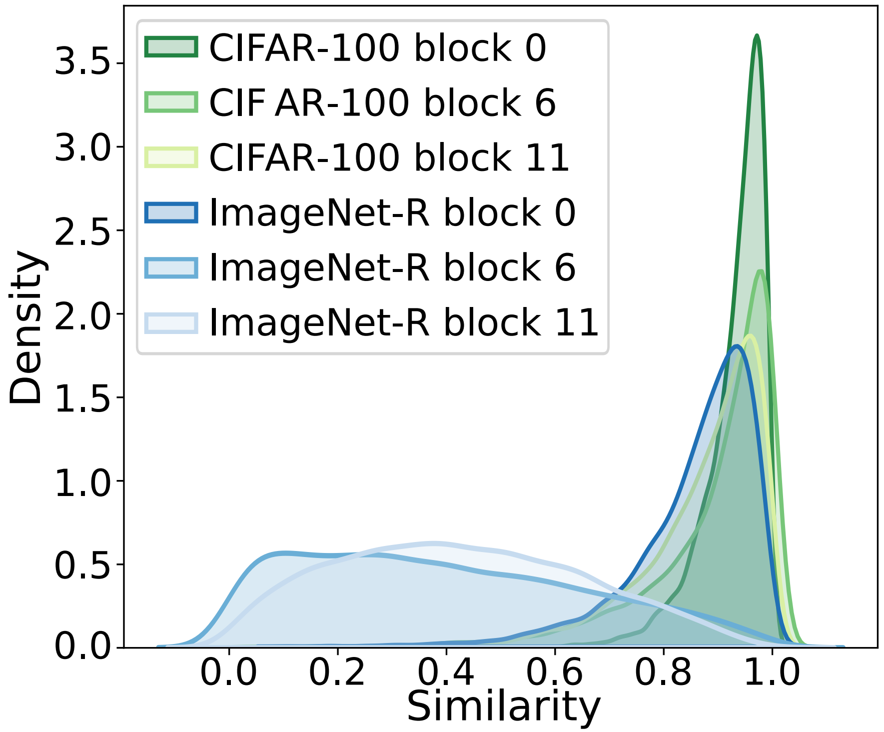
Method	Split CIFAR-100		Split ImageNet-R	
	Acc.(↑)	Forget.(↓)	Acc.(↑)	Forget.(↓)
FT-seq	35.39 ± 1.00	32.98 ± 1.53	7.51 ± 3.73	11.22 ± 0.71
L2P	80.49 ± 0.28	8.74 ± 0.44	57.52 ± 0.18	6.54 ± 0.34
DualPrompt	82.17 ± 0.34	7.52 ± 0.21	61.09 ± 0.18	4.40 ± 0.62
ESN	74.17 ± 1.14	10.59 ± 1.39	-	-
CODA-P-S	79.46 ± 0.06	11.92 ± 1.27	64.60 ± 0.99	6.09 ± 1.18
CODA-P	81.07 ± 0.38	10.10 ± 0.84	66.47 ± 0.33	5.42 ± 0.87
EvoPrompt-S	84.23 ± 0.57	1.64 ± 0.29	73.56 ± 0.42	3.82 ± 0.24
EvoPrompt	$\mathbf{84.72 \pm 0.94}$	$\mathbf{0.89 \pm 0.72}$	$\mathbf{74.05 \pm 0.48}$	$\mathbf{3.66 \pm 0.36}$

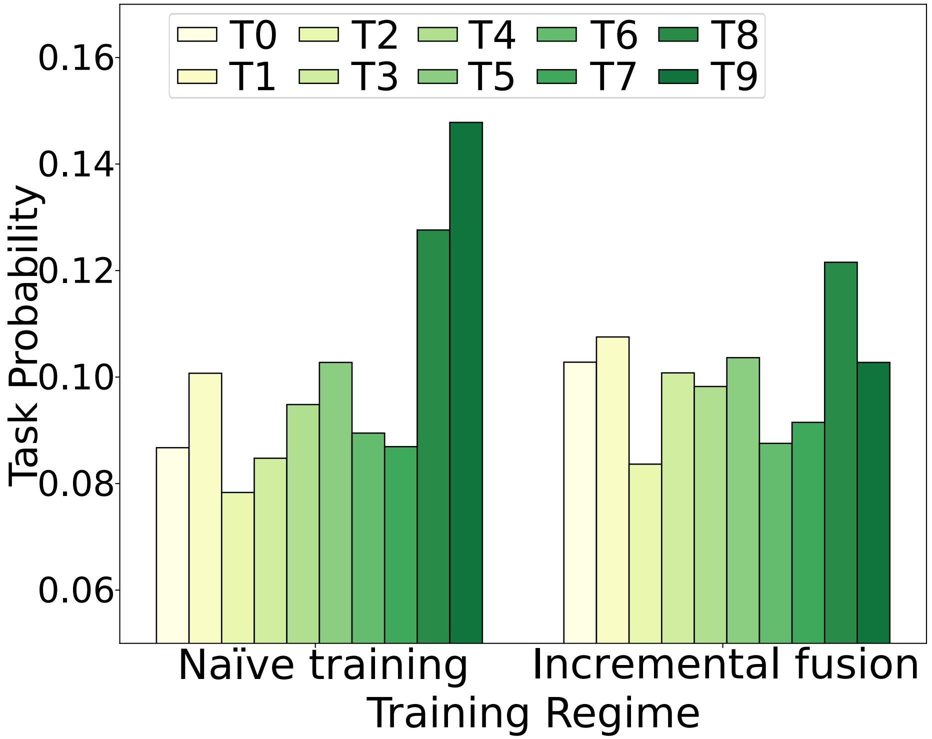
+3.65% Acc
-9.21% Forget

+7.58% Acc
-1.76% Forget

- Ours is better on **knowledge transfer** thus **improve consolidation**
- Key components for knowledge transfer: 1) WPM init from RPM 2) Compositional classifier initialization.

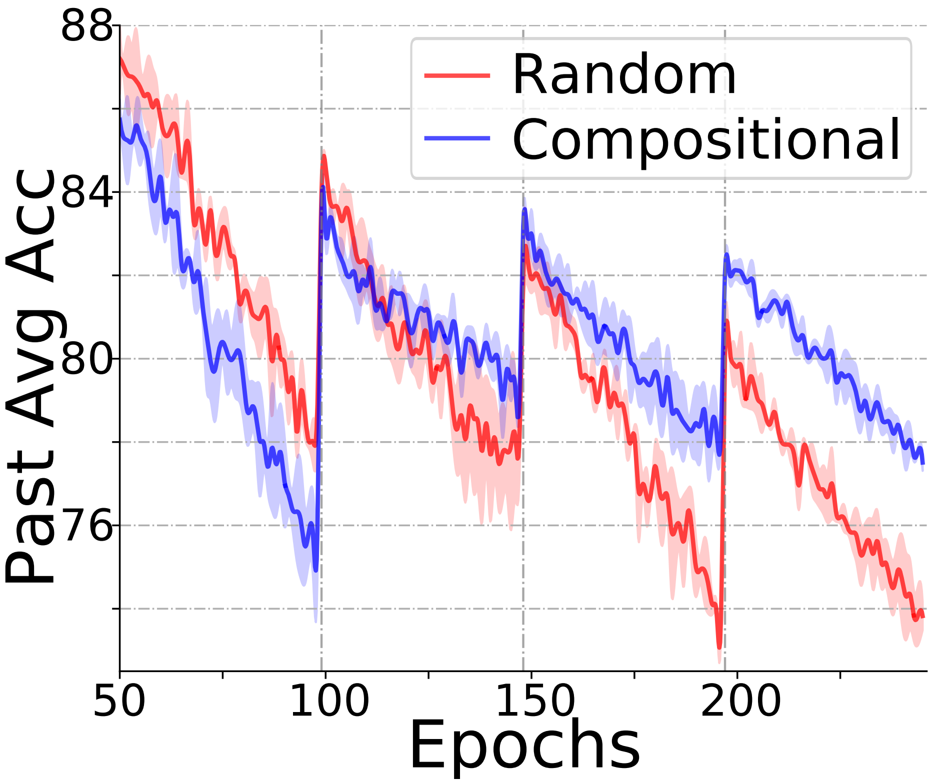
Further Analysis – Assignment Diversity and Recency Bias

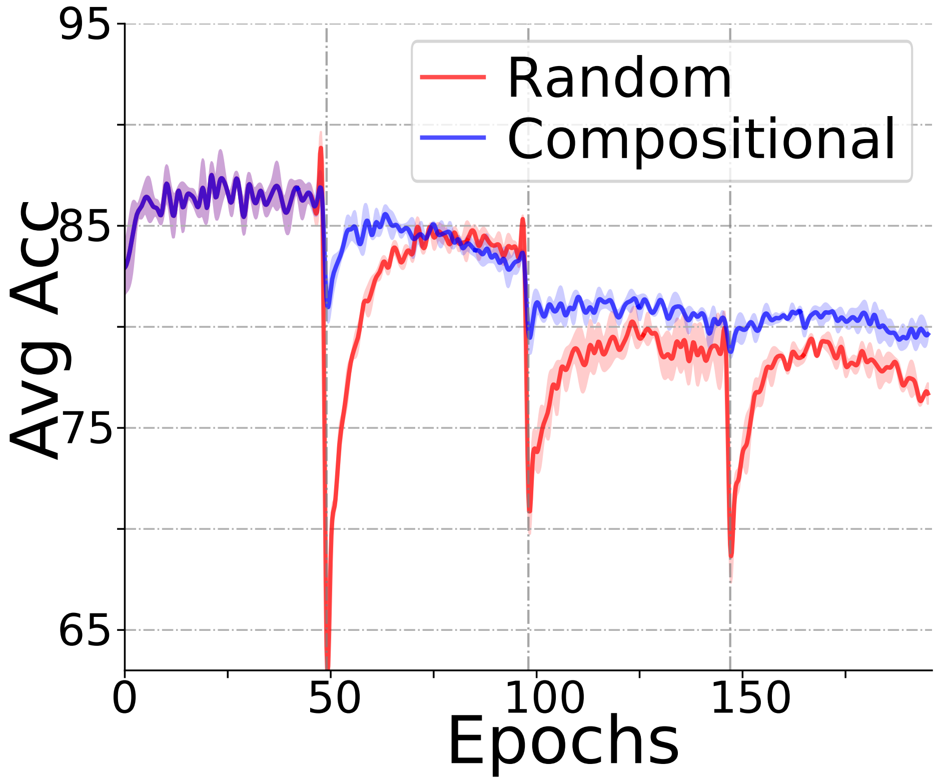




- Coefficient variability is **dataset-dependent**.
- Incremental fusion maintains **parameter balance**, ensuring **general solutions**, and **mitigating recency bias**.

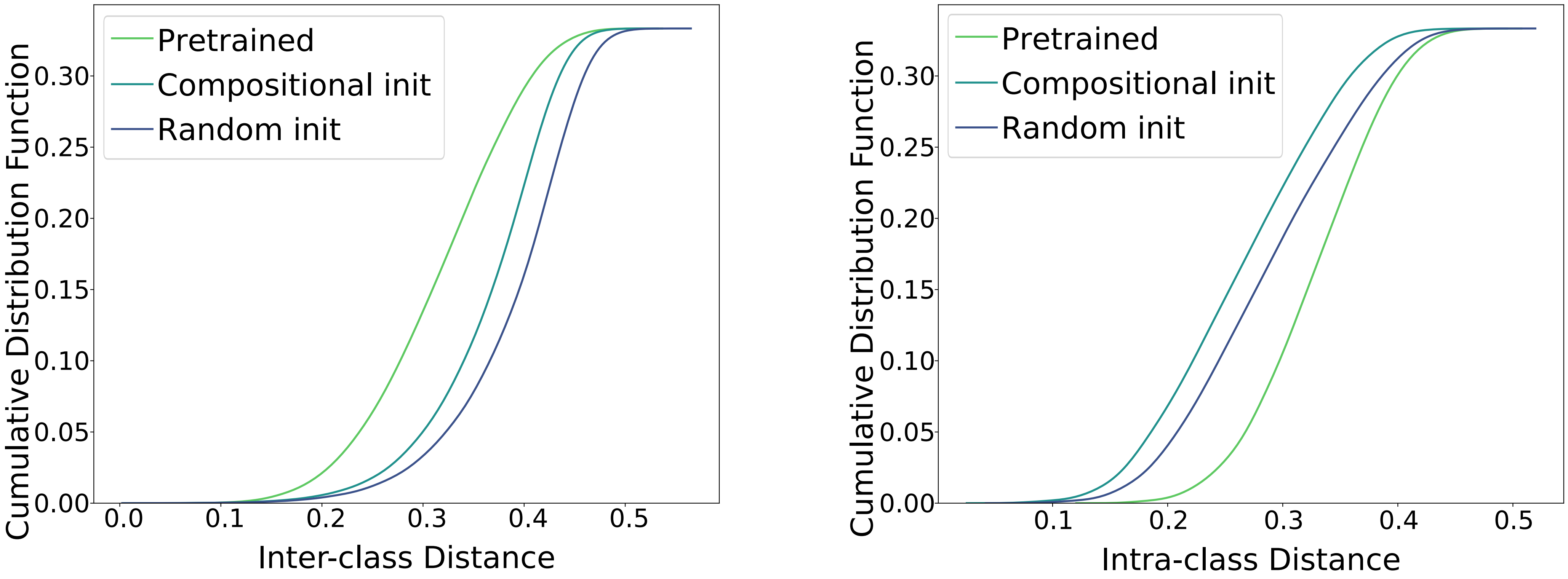
Further Analysis – Stability Gap





- **No stability gap is observed** in either random or CCI.
- CCI exhibits stable performance, **smooth transitions** between tasks, and an **accelerated assimilation** of current knowledge

Further Analysis – Representation Compactness and Discriminativeness



- CCI producing more **compact intra-class structure**.
- CCI maintains **equilibrium** in inter-class margins, resulting in **smaller inter-class distances**, thereby improving **backward compatibility**.

The 38th Annual AAAI Conference on Artificial Intelligence

FEBRUARY 20-27, 2024 | VANCOUVER, CANADA

Thank You

github.com/MIV-XJTU/EvoPrompt

Code

Blog

Paper

西安交通大學

XI'AN JIAOTONG UNIVERSITY

中国科学院深圳先进技术研究院

SHENZHEN INSTITUTE OF ADVANCED TECHNOLOGY

CHINESE ACADEMY OF SCIENCES