This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Is Parameter Isolation Better for Prompt-Based Continual Learning?

Jiangyang Li[1] Chenhao Ding[4] Songlin Dong[2,3*] Qiang Wang[1] Jianchao Zhao[1]
Yuhang He[1] Yihong Gong[1]
[1]State Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
National Engineering Research Center for Visual Information and Applications,
and Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
[2]Faculty of Computility Microelectronics, Shenzhen University of Advanced Technology
[3]Guangdong Provincial Key Laboratory of Computility Microelectronics
[4]School of Software Engineering, Xi'an Jiaotong University

Abstract

Prompt-based continual learning methods effectively mitigate catastrophic forgetting. However, most existing methods assign a fixed set of prompts to each task, completely isolating knowledge across tasks and resulting in suboptimal parameter utilization. To address this, we consider the practical needs of continual learning and propose a prompt-sharing framework. This framework constructs a global prompt pool and introduces a task-aware gated routing mechanism that sparsely activates a subset of prompts to achieve collaborative optimization of task-specific feature representations. Furthermore, we introduce a history-aware modulator that leverages cumulative prompt activation statistics to protect frequently used prompts from excessive updates, thereby mitigating inefficient parameter usage and knowledge forgetting. Extensive analysis and empirical results demonstrate that our approach consistently outperforms existing static allocation strategies in effectiveness and efficiency. Code and models will be released.

1. Introduction

Continual learning (CL) [1, 2, 7, 15] aims to endow models with the ability to learn new tasks sequentially without forgetting previously acquired knowledge. A persistent challenge in this setting is catastrophic forgetting [29, 31, 32, 40, 49], where the model's adaptation to new data overwrites past representations. Recently, prompt-based learning has emerged as a promising direction: instead of updating the entire network, it adapts to new tasks by optimizing a small number of task-specific prompts, while keeping the pre-trained backbone frozen [9, 26, 47, 50]. This approach achieves parameter efficiency and naturally supports knowledge isolation, making it attractive for scalable lifelong learning.

Currently, mainstream prompt-based CL methods [24, 39, 43, 44] are built on the assumption that tasks are independent and should be handled separately, and thus typically allocate a fixed set of prompts to each task. However, we rethink the practical demands of continual learning and argue that this paradigm has two main limitations. (i) Although this rigid allocation simplifies task management, it inherently limits parameter sharing between related tasks, struggles to handle the diverse capacity requirements of heterogeneous task sequences, and lacks the flexibility needed for scalable, long-term, continual learning. (ii) Static prompt assignments often result in inefficient use of model capacity, as different tasks may have varying degrees of relatedness or complexity that static prompts cannot easily accommodate.

In this paper, we propose a **h**istory-**a**ware prompt-**sh**aring framework for continual learning (**Hash**). Specifically, we construct a global prompt pool, where each prompt functions as an expert specializing in certain feature patterns or task attributes. To enable efficient knowledge routing, we introduce a lightweight task-aware gating mechanism that dynamically selects a small subset of prompts for each input, facilitating adaptive composition of relevant knowledge. This parameter-sharing mechanism encourages positive transfer across related tasks and enables efficient local updates during continual learning. As shown in Figure 1(a), compared to static prompt allocation strategies, the proposed method demonstrates superior adaptability in learning new classes. Moreover, the mechanism supports flexible activation of prompt subsets based on task demands, providing strong capacity adaptivity. This sparse activation strategy promotes efficient parameter utilization and minimizes redundant computation across tasks. As illustrated in Figure 1(b), the proposed method achieves a better trade-off between accuracy and parameter efficiency, demonstrating

*Corresponding author: Songlin Dong (dongsl@suat-sz.edu.cn)

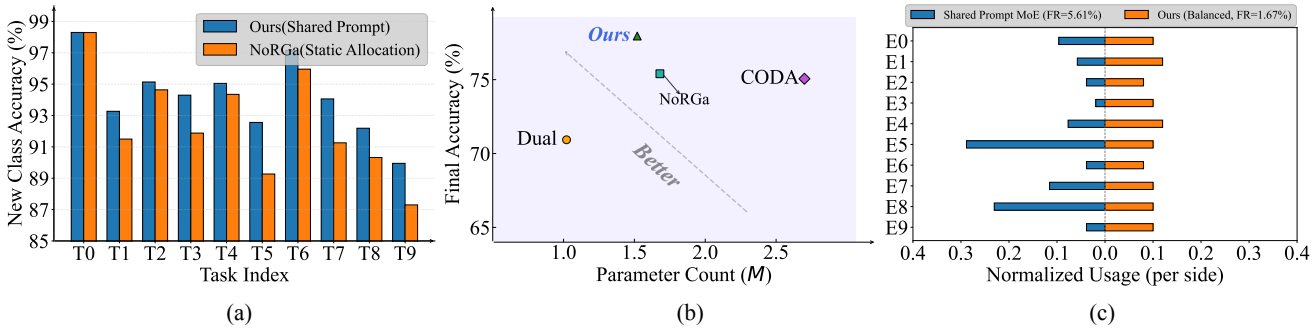


Figure 1. (a) Accuracy of novel classes at each incremental step on the CUB200 dataset. Our method achieves better novel class learning ability compared to static prompt allocation. (b) Comparison of parameter count and accuracy on the Imagenet-R dataset. Our method demonstrates a better accuracy–parameter trade-off than existing approaches. (c) Comparison of expert activation frequency distribution on the Imagenet-R dataset. Our proposed method achieves superior load balancing while effectively reducing knowledge forgetting (FR).

enhanced scalability and computational efficiency.

While this shared prompting mechanism improves generalization on new classes, it may also cause certain prompts to be excessively updated due to repeated reuse, leading to catastrophic forgetting over time [13]. To strike a balance between accurate learning of new tasks and preserving previously acquired knowledge, we propose a history-aware modulator based on cumulative activation statistics. It consists of two components: (i) a routing penalty mechanism that reduces the selection scores of overused prompts to encourage balanced activation; and (ii) a gradient scaling strategy that attenuates the update magnitude for frequently activated prompts during training. These two components optimize the activation strategy of sub-prompt sets without altering the core architecture, enhancing parameter efficiency and effectively mitigating potential forgetting. As shown in Figure 1(c), the proposed history-aware modulator achieves better load balancing, and ablation studies demonstrate its significant reduction in forgetting rate.

Our contributions can be summarized as follows: (1) We propose a continual learning framework based on a shared prompt pool with dynamic, sparse prompt selection, enabling flexible knowledge composition and improved parameter efficiency. (2) We introduce a history-aware modulator that adaptively adjusts prompt selection and protects important prompts based on cumulative prompt activation statistics. (3) We provide analytical and empirical evidence showing that shared prompt learning outperforms static prompt allocation in continual learning settings. (4) **We conduct extensive experiments across diverse benchmarks, with comprehensive ablations and visualizations, consistently achieving superior performance over existing methods.**

2. Related Work

Continual Learning: Continual learning aims to enable models to learn sequential tasks without suffering from catastrophic forgetting. Rehearsal-based methods mitigate forgetting by storing and replaying representative samples from previous tasks [8, 18, 33, 38], but raise concerns regarding memory efficiency and data privacy. A widely studied alternative is non-exemplar class-incremental learning, where no data from prior tasks is retained and task identifiers are not available during inference. To address this challenging setting, prior works have explored regularization-based constraints that penalize deviations from old parameters or impose knowledge distillation losses to preserve learned representations [19, 27, 48], augmentation strategies that synthesize pseudo-features or generate class prototypes from limited statistics to approximate old class distributions [10, 21, 52, 53], and model rectification methods that adjust classifier outputs through bias correction or score calibration to mitigate prediction bias toward new classes [42]. Recently, approaches leveraging pre-trained models [11] have shown promising results for non-exemplar CIL, exploiting rich transferable representations to achieve strong generalization without maintaining old-task memory. Among them, many adopt prompt-based tuning and employ instance-level prompt selection to dynamically retrieve task-relevant knowledge, effectively disentangling task-specific representations while reusing shared backbone knowledge [14, 24, 35, 36, 39, 44, 45].

Prompt Tuning: Prompt tuning is a parameter-efficient paradigm for adapting pre-trained models to downstream tasks. Existing approaches can be broadly categorized into two groups: task-specific prompt allocation and shared prompt pooling. The first group assigns dedicated prompts to each task. DualPrompt [44], CPrompt [14], and HiDe-Prompt [39] statically allocate task-specific prompts to encode task identity, often combining them with shared components or contrastive regularization to improve generalization. NoRGa [24] extends this idea by introducing a non-linear residual gating mechanism, viewing prompts as implicit

experts to enhance flexibility. The second group explores shared prompt pools with dynamic selection. L2P [45], CODA-Prompt [36], and SMoP [6] maintain a global pool of prompts and select subsets based on input features using learned routing mechanisms. CPG [28] also adopts a shared MoE-based prompt pool, but enforces consistency via orthogonal projection constraints on both router and experts to prevent representational drift. OVOR [20] takes a minimalist approach by using only a *single shared prompt* across all tasks, relying instead on virtual outlier regularization applied to the classifier head to tighten decision boundaries and reduce inter-task confusion. ConvPrompt [35] similarly uses a shared prompt pool but incorporates external knowledge by retrieving class-level attributes from GPT-3 during training, introducing additional supervision beyond standard continual learning settings and thus excluded from direct comparison. Building upon HiDe-Prompt [39], which achieves high accuracy in task identification through hierarchical key-prompt matching, we introduce MoE into prompt tuning for enhanced specialization capabilities. However, this raises a critical challenge in continual learning: balancing forgetting and generalization [16]. We propose history-aware modulation to address this, dynamically regulating expert contributions via cumulative usage statistics to maintain both backward compatibility and forward adaptability.

3. Methods

3.1. Problem Definition

We consider the task of *non-exemplar class-incremental learning* (CIL), where a model must continually learn a sequence of classification tasks $\{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_T\}$ without storing any data from previously encountered tasks. Each task $\mathcal{T}_t$ introduces a disjoint set of classes $\mathcal{C}_t$, and training data is only accessible during its corresponding learning phase. At test time, the model must jointly classify over the union of all seen classes $\mathcal{C}_{1:t} = \bigcup_{i=1}^{t} \mathcal{C}_i$ without access to the explicit task identifier. At stage t, the model receives a dataset $\mathcal{D}_t = \{(x_i^{(t)}, y_i^{(t)})\}_{i=1}^{N_t}$, where $x_i^{(t)} \in \mathbb{R}^{L\times d}$ of sequence length L and embedding dimension d denotes the input and $y_i^{(t)} \in \mathcal{C}_t$ the corresponding label. The model is optimized to minimize a supervised loss $\mathcal{L}_{\text{cls}}$ over $\mathcal{D}_t$ while maintaining performance on all prior tasks $\mathcal{T}_{1:t-1}$.

3.2. Shared Prompt Pool with Dynamic Allocation

To promote knowledge sharing across tasks, we introduce the shared prompt pool architecture based on Mixture-of-Experts, which dynamically selects prompts for each input to enhance generalization and reuse.

The prompt pool $\mathcal{P} = \{p_1, p_2, \ldots, p_K\}$ contains K prompts, each $p_k \in \mathbb{R}^{L_p\times d}$ of sequence length L_p acting as an expert. To enable task-specific selection, we assign each of the T tasks a dedicated router for dynamically selecting relevant prompts.

For a sample $x_i^{(t)}$ from task $\mathcal{T}_t$, its corresponding router $\mathcal{R}_t$ computes the relevance score between the sample and each prompt p_k in the prompt pool using a weight matrix $W_r \in \mathbb{R}^{d\times K}$ as follows:

$$\tilde{s}^{(t)} = \mathcal{R}_t(x_i^{(t)}) = \frac{x_i^{(t)} W_r}{\sqrt{d}}, \tag{1}$$

where $\sqrt{d}$ is a normalization factor. The score matrix $\tilde{s} \in \mathbb{R}^{L\times K}$ is averaged over the sequence length dimension L to obtain a one-dimensional score vector $s = [s_1, s_2, \ldots, s_K]$ of length K. The top-k prompts are then selected based on s, denoted as $\mathcal{K}(s) = \text{TopK}(s)$. The normalized weights α_k for the selected prompts are computed as:

$$\alpha_k = \frac{\exp(s_k)}{\sum_{j\in\mathcal{K}(s)} \exp(s_j)}, \quad k \in \mathcal{K}(s). \tag{2}$$

The instance-specific prompt $\tilde{p}$ is then computed as the weighted sum of the selected top-k prompts:

$$\tilde{p} = \sum_{k\in\mathcal{K}(s)} \alpha_k p_k. \tag{3}$$

To integrate the task-specific prompt $\tilde{p}$ into the attention mechanism, we divide it into two parts: $\tilde{p}^K, \tilde{p}^V \in \mathbb{R}^{\frac{L_p}{2}\times d}$, which are added to the key and value matrices, respectively. Although we omit the layer index for notational simplicity, a distinct instance-specific prompt $\tilde{p}^{(l)}$ is generated and injected into each attention layer l. Here, the splitting process is denoted as $\tilde{p}^K, \tilde{p}^V = \text{split}(\tilde{p})$.

Subsequently, we prepend $\tilde{p}^K$ and $\tilde{p}^V$ to the key and value matrices in the attention computation, respectively, resulting in the augmented matrices: $\mathbf{K}'^{(l)} = [\tilde{p}^K; \mathbf{K}^{(l)}]$ and $\mathbf{V}'^{(l)} = [\tilde{p}^V; \mathbf{V}^{(l)}]$. Accordingly, the final attention is computed as follows:

$$\text{Attn}(\mathbf{Q}^{(l)}, \mathbf{K}'^{(l)}, \mathbf{V}'^{(l)}) = \text{softmax}\left(\frac{\mathbf{Q}^{(l)}(\mathbf{K}'^{(l)})^\top}{\sqrt{d_H}}\right)\mathbf{V}'^{(l)}, \tag{4}$$

where $\mathbf{Q}^{(l)}$, $\mathbf{K}^{(l)}$, and $\mathbf{V}^{(l)}$ represent the query, key, and value matrices for layer l, respectively, and d_H is the dimensionality of the attention heads. By integrating dynamically composed prompts into the attention backbone, our framework unifies specialization and generalization in a single, efficient mechanism.

Benefiting from our shared prompt pool and dynamic allocation mechanism, our method achieves a lower total parameter count than Hide-Prompt, even when including our router.

Analysis: Benefits of Prompt Sharing. Unlike static methods that allocate disjoint prompt subsets to each task, our

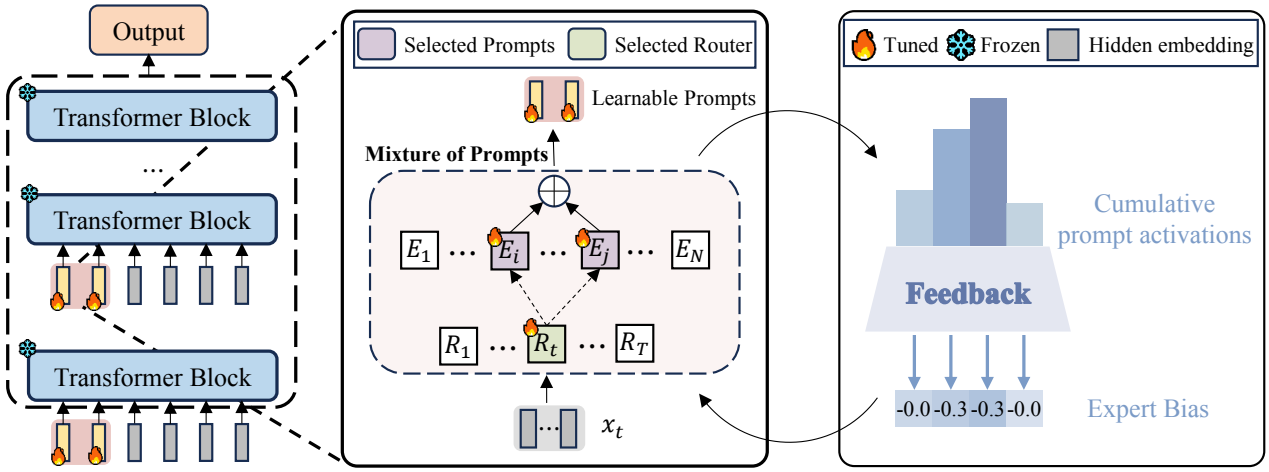


Figure 2. Network architecture overview: Our method adopts a MoE-based shared prompt strategy. During training, the historical activation count of each prompt is recorded and incorporated as a bias term into the router's scoring mechanism, enabling dynamic and informed expert selection based on historical usage.

approach constructs a continuous and expressive representation space by dynamically composing prompts from a global pool. Formally, for static allocation where each task $\mathcal{T}_t$ uses a fixed subset $\mathcal{P}_t = \{p_{t,1}, \ldots, p_{t,k}\} \subset \mathcal{P}$, the optimization objective is:

$$\mathcal{L}_t^{\text{static}} = \mathcal{L}_{\text{cls}}(f(x; \theta, \mathcal{P}_t)), \tag{5}$$

yielding isolated updates with no inter-task parameter sharing. In contrast, our method leverages a shared pool $\mathcal{P} = \{p_1, \ldots, p_K\}$ and constructs input-adaptive prompts $\tilde{p}(x)$ via relevance-weighted composition:

$$\tilde{p}(x) = \sum_{k \in \mathcal{K}(s)} \alpha_k p_k, \quad \mathcal{L}_t^{\text{moe}} = \mathcal{L}_{\text{cls}}(f(x; \theta, \tilde{p}(x))), \tag{6}$$

This design induces a smooth and compositional prompt space that facilitates prompt reuse, expands representational capacity, and enhances generalization to novel classes.

The entropy $H(\alpha) = -\sum_k \alpha_k \log \alpha_k$ quantifies the diversity of prompt activation for each input. A higher $H(\alpha)$ implies that the model distributes attention across multiple prompts to accommodate input complexity, while a lower $H(\alpha)$ reflects selective activation for simpler prompts. This input-aware modulation supports a flexible allocation of capacity, allowing the model to generalize across tasks while retaining sufficient specialization where needed.

3.3. History-Aware Modulator

To further address the issue of prompt overuse in traditional shared prompt pool methods [13], we introduce a history-aware mechanism which dynamically adjusts expert routing scores and gradient update magnitudes based on historical usage statistics, thereby mitigating the overuse of prompt and alleviating catastrophic forgetting.

History-Aware Dynamic Routing. In the **h**istory-aware **d**ynamic **r**outing (HDR) mechanism, we record the cumulative activation count $\mathcal{H}_e^t$ for the e-th prompt up to task t, which is then used for load balancing within the prompt pool. Specifically, for any input, each prompt is assigned an initial relevance score s_e. To prevent the overuse of certain prompts, we apply a penalization function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ to their activation history:

$$\tilde{s}_e = s_e - \psi(\mathcal{H}_e^t), \tag{7}$$

where $\tilde{s}_e$ is the updated score. The penalty function ψ may be logarithmic (e.g., $\log(1+\mathcal{H}_e^t)$), polynomial (e.g., $(\mathcal{H}_e^t)^\beta$), or even task-specific. In our default implementation, we adopt a simple stepwise penalization applied to top-k activated experts:

$$\tilde{s}_e = \begin{cases} s_e - \delta, & \text{if } e \in \mathcal{A}_t, \\ s_e, & \text{otherwise}, \end{cases} \tag{8}$$

where $\delta > 0$ is a fixed deduction and $\mathcal{A}_t$ denotes the top-k experts with the highest cumulative activation count $\mathcal{H}_e^t$ up to task t. This can be seen as a discrete approximation of a continuous penalization function, enforcing sparse regularization in a greedy fashion.

History-Aware Gradient Modulation. To further preserve the historical knowledge of prompts and prevent its disruption in subsequent learning, we propose **h**istory-aware **g**radient **m**odulation (HGM) to protect prompts that have been activated in the past. Specifically, we use a monotonic

Table 1. Overall performance comparison on Split CIFAR-100 and Split ImageNet-R. We present Final Average Accuracy (FAA), Cumulative Average Accuracy (CAA), and Forgetting Measure (FM) of all methods. "Shared" indicates whether the method adopts a shared prompt strategy.

Method	Shared	Split CIFAR-100			Split Imagenet-R		
		FAA (↑)	**CAA**(↑)	FM(↓)	**FAA** (↑)	**CAA**(↑)	FM(↓)
DualPrompt[44]	×	87.30 ± 0.27	91.23 ± 0.65	3.87 ± 0.43	70.93 ± 0.08	75.67 ± 0.52	5.47 ± 0.19
CPrompt [14]	×	87.82 ± 0.21	92.53 ± 0.23	5.06 ± 0.50	77.15 ± 0.11	82.92 ± 0.70	5.97 ± 0.68
HiDe-Prompt [39]	×	92.61 ± 0.28	94.03 ± 0.01	1.50 ± 0.28	75.06 ± 0.12	76.60 ± 0.01	4.09 ± 0.13
NoRGa [24]	×	94.48 ± 0.13	95.83 ± 0.37	**1.44** ± 0.27	75.40 ± 0.39	79.52 ± 0.07	4.59 ± 0.07
EvoPrompt [23]	×	87.97 ± 0.30	89.12 ± 0.21	2.60 ± 0.42	76.83 ± 0.08	78.43 ± 0.42	2.78 ± 0.06
L2P [45]	✓	83.06 ± 0.17	88.27 ± 0.71	5.61 ± 0.32	67.53 ± 0.44	71.98 ± 0.52	5.84 ± 0.38
CODA-Prompt [36]	✓	86.94 ± 0.63	91.57 ± 0.75	4.04 ± 0.18	70.03 ± 0.47	74.26 ± 0.24	5.17 ± 0.22
OVOR [20]	✓	85.99 ± 0.21	90.26 ± 0.35	6.42 ± 0.17	76.11 ± 0.33	79.12 ± 0.28	7.16 ± 0.28
CPG [28]	✓	90.63 ± 0.44	93.24 ± 0.33	3.98 ± 0.65	78.63 ± 0.52	81.04 ± 0.23	7.18 ± 0.62
Hash (Ours)	✓	**95.02** ± 0.34	**95.97** ± 0.28	1.67 ± 0.23	**79.02** ± 0.36	**82.96** ± 0.17	**2.63** ± 0.15

decay function $\gamma : \mathbb{R}_+ \to (0, 1]$ to modulate the gradients.

$$\tilde{g}_e = \gamma(\mathcal{H}_e^t) \cdot g_e, \tag{9}$$

where g_e is the raw gradient for expert e. The modulation function can take various forms, such as inverse scaling $\gamma(h) = 1/(1 + \beta h)$ or exponential decay $\exp(-\beta h)$, with $\beta > 0$. In practice, we use a piecewise constant modulation for the top-k experts:

$$\gamma(\mathcal{H}_e^t) = \begin{cases} \alpha, & \text{if } e \in \mathcal{A}_t, \\ 1, & \text{otherwise}, \end{cases} \tag{10}$$

with $0 < \alpha < 1$. This reduces the learning rate for frequently activated experts, implicitly enforcing parameter stability.

History-Aware Modulator from a Regularization Perspective. We aim to reconsider the problem from a regularization perspective, formalizing the trade-off between maximizing expert scores and penalizing their overuse. This unified framework not only explains our dynamic expert routing and history-aware gradient modulation but also validates our design intuition.

The overall mechanism can be interpreted through the lens of regularized optimization. Expert routing aims to maximize the score while penalizing overused experts:

$$\min_{\mathbf{p} \in \Delta^E} \mathcal{L}_{\text{route}}(\mathbf{p}) = \sum_{e=1}^{E} p_e \cdot (-s_e) + \sum_{e=1}^{E} p_e \cdot \psi(\mathcal{H}_e^t), \tag{11}$$

where Δ^E is the probability simplex. While the full minimization corresponds to soft attention over all experts, our top-k routing can be viewed as a sparse approximation under entropic regularization.

For gradient modulation, our update step approximates the minimization of a stability-aware regularized loss:

$$\min_{\theta} \mathcal{L}(\theta) + \sum_{e=1}^{E} \underbrace{R_e(\theta_e; \mathcal{H}_e^t)}_{\text{history-aware regularization}}, \tag{12}$$

where R_e penalizes deviation from prior parameters θ_e^t according to historical activation:

$$R_e(\theta_e; \mathcal{H}_e^t) = \frac{1}{2} \cdot \mathcal{H}_e^t \cdot \|\theta_e - \theta_e^t\|^2. \tag{13}$$

This induces a stability constraint that grows with expert usage, aligning with the intuition of selectively freezing experts that are heavily relied upon.

By embedding historical usage into both expert routing and gradient modulation, our mechanism strikes a dynamic balance between adaptability and stability. This approach enables continuous control over routing selection and gradient flow, while inherently maintaining robustness in response to diverse inputs.

4. Experiments

4.1. Experemental Details

Benchmark. We evaluate our method on a suite of diverse class-incremental learning (CIL) benchmarks to assess its adaptability across various scenarios. *Split CIFAR-100* comprises 100 classes of natural images [22], partitioned into 10 sequential tasks with disjoint class labels. *Split ImageNet-R* is constructed from 200 classes sampled from ImageNet-R [17], including renditions, sketches, and abstract styles. These are divided into 10 tasks and present a significant domain shift. *Split CUB-200* is derived from the CUB-200-2011 dataset [37], a fine-grained bird classification benchmark split into 10 tasks. 5-Datasets [12] comprises CIFAR-10 [22], MNIST [25], Fashion-MNIST [46], SVHN [30],

Table 2. The results on Split CUB-200 and 5-Datasets. Here we present FAA and CAA for all approaches.

Method	Split CUB-200		5-Datasets	
	FAA	CAA	FAA	CAA
L2P	75.46	81.59	81.84	85.36
DualPrompt	77.56	83.24	77.91	84.92
CODA-Prompt	74.34	79.52	64.18	77.62
CPrompt	80.35	87.66	84.96	88.67
HiDe-Prompt	86.56	89.34	93.83	94.32
NoRGa	90.90	92.53	94.16	94.55
Hash (Ours)	**91.34**	**92.98**	**95.12**	**95.67**

Table 3. Long sequence incremental analysis: Experimental results with 20 incremental tasks.

Method	Split CIFAR-100		Split ImageNet-R	
	FAA	FM	FAA	FM
L2P	79.95	5.69	69.72	5.92
DualPrompt	81.34	4.02	66.72	5.63
CODA-Prompt	81.94	4.92	69.96	6.79
CPrompt	84.57	6.01	74.79	7.34
HiDe-Prompt	91.68	2.34	73.98	4.98
NoRGa	93.12	2.12	74.32	5.14
Hash (Ours)	**94.07**	**1.98**	**76.57**	**4.62**

and notMNIST [3], each serving as a distinct incremental task to evaluate the model's robustness under significant inter-task distributional shifts.

Baselines. We compare our method with a suite of representative prompt-based continual learning approaches, including L2P [45], DualPrompt [44], CODA-Prompt [36], C-Prompt [14], HiDe-Prompt [39], and NoRGa [24], covering diverse designs for prompt integration and selection. All methods are evaluated under consistent experimental settings using the same backbone and training pipeline. The main paper reports results under the supervised ImageNet-21K (Sup-21K) [34] pre-training, while results under additional pre-training paradigms (iBOT [51], DINO [4], MoCo [5]) are deferred to the Appendix.

Evaluation Metrics. Following standard practice, we report three widely-used metrics to evaluate continual learning performance: final average accuracy (FAA), cumulative average accuracy (CAA), and final forgetting measure (FM) [40]. FAA reflects the average performance on all tasks after learning the final task, while CAA summarizes the historical performance across all seen tasks. FM quantifies the extent of knowledge forgetting between tasks.

Implementation Details. All experiments are conducted on a single NVIDIA A100 GPU. We adopt the pre-trained ViT-B/16 as the backbone across all methods. Training is performed using the Adam optimizer with default parameters

Table 4. Results on CORe50. A_N gives the accuracy averaged over tasks and buffer size represents the number of old exemplars stored in the memory buffer.

Method	Buffer size	A_N (↑)
ER	**50/class**	80.10
GDumb		74.92
BiC		79.28
DER++		79.70
Co^2L		79.75
DyTox		79.21
L2P		81.07
EWC	**0/class**	74.82
LwF		75.45
L2P		78.33
CODA-P-S		85.41
S-Prompts		83.13
PINA		86.74
Hash (Ours)		**87.94**

($\beta_1 = 0.9$, $\beta_2 = 0.999$), a batch size of 128. We adopt the same epoch configuration for all methods as provided by HiDe-Prompt [39]. Prompt selection is performed at the instance level during both training and inference. We empirically verify that the test results are invariant to batch size. All results are averaged over three random seeds.

4.2. Comparison Results

Class-IL Results. Table 1 reports results on Split CIFAR-100 and Split ImageNet-R under the Sup-21K pre-trained vision transformer. Our method achieves the highest FAA and CAA across both benchmarks, while maintaining competitive FM, demonstrating strong performance and stability in class-incremental learning. Specifically, we obtain 95.02% FAA on CIFAR-100 and 79.02% on ImageNet-R, outperforming NoRGa by 0.54% and 3.66%, respectively. Notably, while regularization-based methods typically achieve low FM, our design strikes a better balance between final accuracy and forgetting, validating the effectiveness of expert routing and gradient modulation. Table 2 further shows performance on Split CUB-200 and 5-Datasets benchmark. Again, our method achieves the best FAA of 91.34% on CUB-200 and 95.12% on 5-Datasets, outperforming the strongest baseline by 0.44% and 0.96%, respectively. These improvements are particularly meaningful in settings prone to overfitting or under-generalization. Overall, the consistent performance gains across different datasets and pre-training types validate the effectiveness and adaptability of our MoE-enhanced dynamic prompt learning framework.

Long Sequence Incremental Analysis. We assess the performance of all methods under a more challenging long-sequence incremental learning setup, where each dataset is uniformly divided into 20 sessions. All methods are evaluated using the ViT-B/16 backbone pre-trained on ImageNet-

Table 5. Ablation study on key components of the proposed method.

MoE	HDR	HGM	Split CIFAR-100		Split Imagenet-R	
			FAA	CAA	FAA	CAA
✓			88.56	90.12	76.99	79.53
✓	✓		94.15	95.03	78.10	81.21
✓		✓	93.86	94.77	77.56	80.82
✓	✓	✓	**95.02**	**95.97**	**79.02**	**82.96**

Table 6. Performance comparison with different prompt positions.

Hash		Split CIFAR-100		Split ImageNet-R	
		FAA	CAA	FAA	CAA
(Default)	1–4	**95.02**	**95.97**	**79.02**	**82.96**
Deep	5–12	94.67	95.59	78.61	82.48
	1–12	94.81	95.76	78.75	82.65
Shallow	1	93.28	94.12	77.00	80.94
	12	93.01	93.85	76.82	80.70

21K. Compared to short-session setups, this setting imposes greater pressure on knowledge retention and prompt scalability due to finer task granularity, increased semantic overlap, and extended task sessions. These challenges tend to amplify catastrophic forgetting in existing methods. As shown in Table 3, our method consistently maintains high accuracy and low forgetting across both CIFAR-100 and ImageNet-R. Notably, the performance gap widens with increasing task length, indicating that our approach scales more effectively under prolonged incremental learning. This further highlights the advantage of our shared, input-adaptive prompt mechanism, which promotes efficient knowledge reuse and mitigates catastrophic forgetting even in the presence of long and complex task sequences.

Domain-IL Results. To complement our evaluation under class-incremental settings, we further assess the performance of our approach in the domain-incremental learning (DIL) scenario, where tasks share the same label space but differ in domain-specific input distributions. We compare our method with three recent DIL baselines: CODA [36], S-Prompts [43], and PINA [41]. CODA employs domain-specific adapters with contrastive objectives to separate domain-invariant and domain-specific features, while S-Prompts maintains a per-domain prompt pool with a learned scoring mechanism for selection. PINA, in contrast, introduces a cross-domain concept integration framework that effectively aggregates knowledge from multiple domains, demonstrating strong performance on several DIL benchmarks. We report results on CORe50 in Table 4, showing that our method outperforms CODA by +2.53% and surpasses S-Prompt, which was specifically tailored for DIL. Notably, our approach also exceeds the performance of PINA by a margin of 1.2%, further highlighting its effectiveness. All methods are evaluated using the same pre-trained backbone (ViT-B/16), and exemplar-based approaches store 50 exemplars per class following standard protocol. These findings highlight the robustness and generality of our approach across diverse continual learning settings.

Table 7. Sensitivity of the number of router experts.

N_r	Split CIFAR-100		Split Imagenet-R	
	FAA	CAA	FAA	CAA
10	94.89	95.32	78.76	82.54
15	95.02	95.97	79.02	82.96
20	95.08	96.14	78.95	82.72
25	94.54	95.08	78.28	81.92

4.3. Ablation Study

Impact of History-Aware Modulator. To evaluate the effectiveness of our proposed history-aware routing mechanism, we conduct a comparative analysis of forgetting rates across multiple datasets. Figure 3 illustrates the impact of our approach on catastrophic forgetting mitigation. Figure 3(a) shows that history-aware routing consistently reduces forgetting rates across all datasets, with absolute decreases of 2.65%, 1.65%, 2.39%, and 2.74% on CIFAR-100, Imagenet-R, CUB-200, and 5-Datasets, respectively. This stems from two complementary synergistic effects: routing penalties prevent over-reliance on dominant prompts, while gradient scaling protects their learned representations from excessive updates. Figure 3(b) shows that without the modulator, expert usage is highly skewed, with a strong bias toward a few dominant experts. After applying the history-aware modulator (Figure 3(c)), expert utilization becomes significantly more balanced without sacrificing cross-task sharing, thus improving the overall efficiency of the shared prompt pool. The activation pattern shifts from highly concentrated to broadly distributed, ensuring more diverse prompt combinations across tasks.

Module-Wise Ablation. To disentangle the contributions of our two key components, namely history-aware dynamic routing (HDR) and history-aware gradient modulation (HGM), we perform an ablation study summarized in Table 5. The study reports the FAA and CAA. Beginning with a naive MoE-based prompt tuning baseline with uniform expert assignment, we first introduce HDR alone, which improves accuracy from 88.56% to 94.15% on Split CIFAR-100 and from 76.99% to 78.10% on Split Imagenet-R under the Sup-21K backbone, yielding relative gains of 5.59% and 1.11% respectively. This confirms that leveraging historical expert activations enhances expert selection efficacy. Applying HGM in isolation achieves similar improvements—an increase of 0.57% on Imagenet-R (Sup-21K), reaching 77.56%,

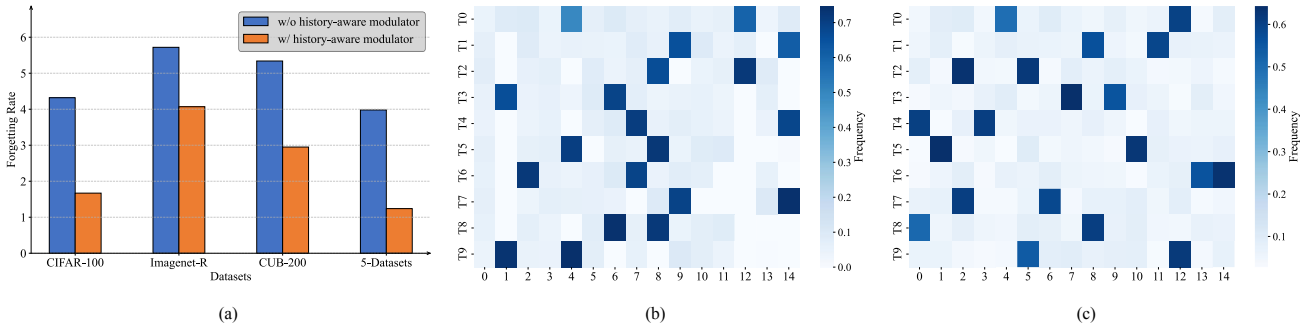


Figure 3. (a) Forgetting rate reduction with history-aware modulator. (b) Cross-task expert utilization frequency distribution without history-aware modulator. (c) Cross-task expert utilization frequency distribution with history-aware modulator.

Table 8. Sensitivity of prompt lengths.

L	Split CIFAR-100		Split Imagenet-R	
	FAA	CAA	FAA	CAA
5	94.12	94.72	77.44	81.22
10	94.95	95.72	78.55	82.44
15	95.02	95.97	79.02	82.96
20	95.11	96.03	78.39	82.21
25	94.79	95.88	78.32	82.04

and a gain of 5.30% on CIFAR-100. Finally, combining HDR and HGM yields the best results across all configurations, achieving 95.02% on CIFAR-100 and 79.02% on Imagenet-R, which corresponds to total gains of 6.46% and 2.03% over the naive MoE variant. These improvements underscore the complementary nature of the two components: HDR optimizes *which* experts to activate, while HGM regulates *how* they are updated, jointly enhancing both knowledge acquisition and retention.

Layer-wise Prompt Deployment Analysis. To investigate the impact of layer-wise prompt deployment, we systematically placed the prompt modules at different layers of the backbone and evaluated their effect. As shown in Table 6, deploying prompts at the lower layers consistently yields the best results across benchmarks, achieving 95.02% FAA on CIFAR-100 and 79.02% on ImageNet-R. Notably, using deeper layers or all layers degrades performance by 0.21–0.41%, while single-layer deployment performs significantly worse with drops of 2.02–2.20%. This indicates that prompt-based adaptation is most effective when applied to early feature representations, where multi-layer coordination enables better balance between plasticity and stability. Lower layers encode more task-agnostic patterns that are easier to modulate without disrupting prior knowledge, whereas deeper features are more task-specific and less adaptable to continual learning scenarios.

Hyperparameter Sensitivity Analysis. We analyze the sensitivity of our method to the number of experts N_r and the prompt length L, as reported in Table 7 and Table 8. The performance remains relatively stable across different configurations, indicating the robustness of the method. For the number of experts, increasing N_r generally leads to performance improvement up to a certain point, with optimal results achieved at $N_r = 15 - 20$. When N_r becomes too large (e.g., 25), the overall accuracy slightly declines, possibly due to insufficient training samples per expert or increased optimization difficulty. This highlights the importance of balancing model capacity and expert coverage. In terms of prompt length, longer prompts improve performance initially, with $L = 15$ emerging as the optimal configuration across both benchmarks. Beyond this point, the benefit quickly saturates and may even slightly decline, suggesting that moderate lengths suffice for adaptation without incurring unnecessary memory cost.

5. Conclusion

This study presents a prompt-sharing framework for continual learning, which adopts a sparse and dynamic prompt allocation strategy. By enabling prompt sharing, our approach inherently promotes generalization and reduces redundancy compared to static prompt allocation designs. To address catastrophic forgetting, we further introduce a history-aware modulator which jointly balance knowledge transfer and task stability. Empirically, we show that it strikes an effective balance between generalization and forgetting, outperforming parameter-isolation prompt-based approaches.

Limitations. Building on the current work, which is constrained by computational resources, future efforts will explore larger-scale backbone models. Additionally, we aim to further enhance the performance of the network framework, particularly in maintaining its advantages under conditions of extremely low task similarity.

Acknowledgement

This work is supported by the National Natural Science Foundation of China under Grant No.U21B2048, No.62302382, and No.62576224, the Shenzhen Key Technical Projects under Grant CJGJZD20220517141605013, the China Postdoctoral Science Foundation No.2024M752584, and Guangdong Provincial Key Laboratory of Computility Microelectronics 2024B1212010007.

References

[1] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3366–3375, 2017. 1

[2] Eden Belouadah, Adrian Popescu, and Ioannis Kanellos. A comprehensive study of class incremental learning algorithms for visual tasks. *Neural Networks*, 135:38–54, 2021. 1

[3] Yaroslav Bulatov. Notmnist dataset. http://yaroslavvb.blogspot.com/2011/09/notmnist-dataset.html, 2011. Google (Books/OCR), Tech. Rep. 6

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 6

[5] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021. 6

[6] Joon-Young Choi, Junho Kim, Jun-Hyung Park, Wing-Lam Mok, and SangKeun Lee. Smop: Towards efficient and effective prompt tuning with sparse mixture-of-prompts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14306–14316, 2023. 3

[7] Songlin Dong, Xiaopeng Hong, Xiaoyu Tao, Xinyuan Chang, Xing Wei, and Yihong Gong. Few-shot class-incremental learning via relation knowledge distillation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1255–1263, 2021. 1

[8] Songlin Dong, Haoyu Luo, Yuhang He, Xing Wei, Jie Cheng, and Yihong Gong. Knowledge restore and transfer for multi-label class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18711–18720, 2023. 2

[9] Songlin Dong, Xinyuan Gao, Yuhang He, Zhengdong Zhou, Alex C Kot, and Yihong Gong. Ceat: Continual expansion and absorption transformer for non-exemplar class-incremental learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(4):3146–3159, 2024. 1

[10] Songlin Dong, Yingjie Chen, Yuhang He, Yuhan Jin, Alex C Kot, and Yihong Gong. Analogical augmentation and significance analysis for online task-free continual learning. *IEEE Transactions on Multimedia*, 27:3370–3382, 2025. 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 2

[12] Sayna Ebrahimi, Franziska Meier, Roberto Calandra, Trevor Darrell, and Marcus Rohrbach. Adversarial continual learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 386–402. Springer, 2020. 5

[13] Xinyuan Gao, Songlin Dong, Yuhang He, Qiang Wang, and Yihong Gong. Beyond prompt learning: Continual adapter for efficient rehearsal-free continual learning. In *European Conference on Computer Vision*, pages 89–106. Springer, 2024. 2, 4

[14] Zhanxin Gao, Jun Cen, and Xiaobin Chang. Consistent prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28463–28473, 2024. 2, 5, 6

[15] Jizhou Han, Chenhao Ding, Yuhang He, Songlin Dong, Qiang Wang, Xinyuan Gao, and Yihong Gong. Learn by reasoning: Analogical weight generation for few-shot class-incremental learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 1

[16] Jizhou Han, Shaokun Wang, Yuhang He, Chenhao Ding, Qiang Wang, Xinyuan Gao, SongLin Dong, and Yihong Gong. Consistent supervised-unsupervised alignment for generalized category discovery. *arXiv preprint arXiv:2507.04725*, 2025. 3

[17] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021. 5

[18] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 831–839, 2019. 2

[19] Libo Huang, Yan Zeng, Chuanguang Yang, Zhulin An, Boyu Diao, and Yongjun Xu. etag: Class-incremental learning via embedding distillation and task-oriented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12591–12599, 2024. 2

[20] Wei-Cheng Huang, Chun-Fu Chen, and Hsiang Hsu. Ovor: Oneprompt with virtual outlier regularization for rehearsal-free class-incremental learning. *arXiv preprint arXiv:2402.04129*, 2024. 3, 5

[21] Taehoon Kim, Jaeyoo Park, and Bohyung Han. Cross-class feature augmentation for class incremental learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13168–13176, 2024. 2

[22] Alex Krizhevsky. Learning multiple layers of features from tiny images. *Citeseer*, 2009. 5

[23] Muhammad Rifki Kurniawan, Xiang Song, Zhiheng Ma, Yuhang He, Yihong Gong, Yang Qi, and Xing Wei. Evolving parameterized prompt memory for continual learning. In *Pro-*

ceedings of the AAAI Conference on Artificial Intelligence, pages 13301–13309, 2024. 5

[24] Minh Le, Huy Nguyen, Trang Nguyen, Trang Pham, Linh Ngo, Nhat Ho, et al. Mixture of experts meets prompt-based continual learning. *Advances in Neural Information Processing Systems*, 37:119025–119062, 2024. 1, 2, 5, 6

[25] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324, 1998. 5

[26] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 1

[27] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 2

[28] Yue Lu, Shizhou Zhang, De Cheng, Guoqiang Liang, Yinghui Xing, Nannan Wang, and Yanning Zhang. Training consistent mixture-of-experts-based prompt generator for continual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19152–19160, 2025. 3, 5

[29] Sanket Vaibhav Mehta, Darshan Patil, Sarath Chandar, and Emma Strubell. An empirical investigation of the role of pre-training in lifelong learning. *Journal of Machine Learning Research*, 24(214):1–50, 2023. 1

[30] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, page 4. Granada, 2011. 5

[31] Aristeidis Panos, Yuriko Kobe, Daniel Olmeda Reino, Rahaf Aljundi, and Richard E Turner. First session adaptation: A strong replay-free baseline for class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18820–18830, 2023. 1

[32] Vinay Venkatesh Ramasesh, Aitor Lewkowycz, and Ethan Dyer. Effect of scale on catastrophic forgetting in neural networks. In *International conference on learning representations*, 2021. 1

[33] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017. 2

[34] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 6

[35] Anurag Roy, Riddhiman Moulick, Vinay K Verma, Saptarshi Ghosh, and Abir Das. Convolutional prompting meets language models for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23616–23626, 2024. 2, 3

[36] James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11909–11919, 2023. 2, 3, 5, 6, 7

[37] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *California Institute of Technology*, 2011. 5

[38] Liyuan Wang, Xingxing Zhang, Kuo Yang, Longhui Yu, Chongxuan Li, Lanqing Hong, Shifeng Zhang, Zhenguo Li, Yi Zhong, and Jun Zhu. Memory replay with data compression for continual learning. *arXiv preprint arXiv:2202.06592*, 2022. 2

[39] Liyuan Wang, Jingyi Xie, Xingxing Zhang, Mingyi Huang, Hang Su, and Jun Zhu. Hierarchical decomposition of prompt-based continual learning: Rethinking obscured sub-optimality. *Advances in Neural Information Processing Systems*, 36: 69054–69076, 2023. 1, 2, 3, 5, 6

[40] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 6

[41] Qiang Wang, Yuhang He, Songlin Dong, Xinyuan Gao, Shaokun Wang, and Yihong Gong. Non-exemplar domain incremental learning via cross-domain concept integration. In *European Conference on Computer Vision*, pages 144–162. Springer, 2024. 7

[42] Shaokun Wang, Weiwei Shi, Yuhang He, Yifan Yu, and Yihong Gong. Non-exemplar class-incremental learning via adaptive old class reconstruction. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4524–4534, 2023. 2

[43] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. *Advances in Neural Information Processing Systems*, 35:5682–5695, 2022. 1, 7

[44] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European conference on computer vision*, pages 631–648. Springer, 2022. 1, 2, 5, 6

[45] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 139–149, 2022. 2, 3, 5, 6

[46] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 5

[47] Yi Xin, Siqi Luo, Haodi Zhou, Junlong Du, Xiaohong Liu, Yue Fan, Qing Li, and Yuntao Du. Parameter-efficient fine-tuning for pre-trained vision models: A survey. *arXiv preprint arXiv:2402.02242*, 2024. 1

[48] Lu Yu, Bartlomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6982–6991, 2020. 2

[49] Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. Slca: Slow learner with classifier alignment

for continual learning on a pre-trained model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19148–19158, 2023. 1

[50] Da-Wei Zhou, Hai-Long Sun, Jingyi Ning, Han-Jia Ye, and De-Chuan Zhan. Continual learning with pre-trained models: a survey. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8363–8371, 2024. 1

[51] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image bert pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022. 6

[52] Fei Zhu, Zhen Cheng, Xu-yao Zhang, and Cheng-lin Liu. Class-incremental learning via dual augmentation. *Advances in neural information processing systems*, 34:14306–14318, 2021. 2

[53] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5871–5880, 2021. 2