Spectral Mixture-of-Experts for Continual Learning

Chen Yin[1], Xingbo Dong[1*], Xuelin Shen[2], Zhe Jin[1]
[1]Anhui Provincial Key Laboratory of Secure Artificial Intelligence,
Anhui Provincial International Joint Research Center for Advanced Technology in Medical Imaging,
School of Artificial Intelligence, Anhui University
[2]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
yinchen@stu.ahu.edu.cn {xingbo.dong, jinzhe}@ahu.edu.cn shenxuelin@gml.ac.cn

Abstract

While Parameter-Efficient Fine-Tuning using Mixture-of-Experts (MoE) is a promising solution for continual learning (CL), it suffers from two critical failure modes: structural interference, where expert updates interfere, and compositional forgetting, where the model's routing policy drifts. To address these issues, we introduce Spectral MoE, a novel framework built for CL from three core components. First, Spectral Experts are parameterized using unique, disjoint spectral masks to confine their learnable parameters to distinct frequency subspaces, ensuring a priori orthogonal updates that prevent structural interference. Second, a Dual-Router mechanism decouples online routing that learns new tasks from an offline memory that archives historical expert importance. Finally, this offline memory enables a Dynamic Consistency Projection, a geometric constraint that suppresses router drift and adaptively shields experts based on their past contributions, mitigating compositional forgetting. Validated on a strict cross-domain CL benchmark, our framework significantly outperforms existing methods, demonstrating superior knowledge retention and plasticity for new tasks. Code is available at: https://github.com/ouycc/Spectral_MoE.

1. Introduction

Continual Learning (CL) [4, 13] aims to equip models with the ability to acquire new skills over time while maintaining performance on previously learned tasks. Yet deep neural networks under CL are notoriously susceptible to catastrophic forgetting: optimizing for a new task can sharply degrade performance on older ones [33]. Pre-trained vision-language models (VLMs) such as CLIP [36] have emerged as appealing foundations for CL because they already exhibit strong zero-shot generalization across diverse domains. However, fully retraining such large models for each new task is prohibitively expensive in both computation and storage. This has motivated two main adaptation strategies: full-parameter fine-tuning and Parameter-Efficient Fine-Tuning (PEFT) [10, 21, 27].

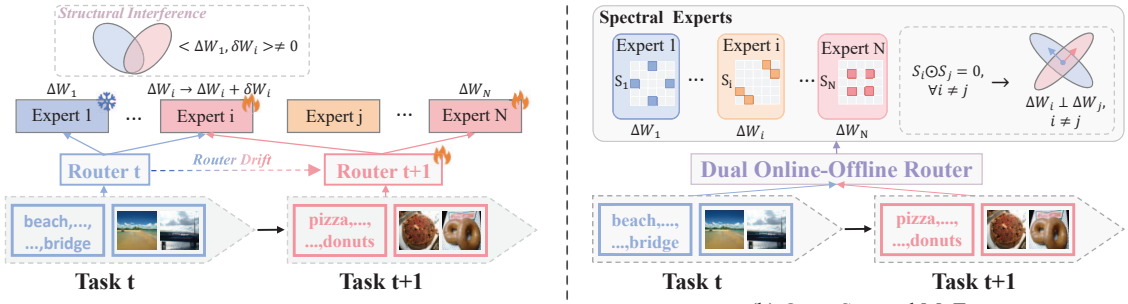
Full-parameter methods aim to preserve the model's general knowledge. For example, ZSCL [57] updates all weights while attempting to retain CLIP's zero-shot behavior through teacher–student distillation and anti-forgetting regularization. These methods can partially preserve prior behavior but incur significant cost and become increasingly difficult to scale across many tasks. The second line focuses on PEFT [41, 54], particularly Low-Rank Adaptation (LoRA) [17], which adapts the model by inserting small learnable modules while keeping the backbone frozen. Building on this concept, approaches like MoE-Adapter [19, 52] organize multiple LoRA modules into sparsely activated experts and use a learned router to dynamically combine them per input. While isolating tasks to different experts is promising for CL, this approach leads to two failures as illustrated in Figure 1a: structural interference and router drift.

The first issue, **Structural Interference**, occurs because all expert adapters operate on the same frozen backbone. Without enforced orthogonality, parameter updates for new tasks tend to occupy overlapping subspaces with those of older tasks [8, 50]. Consequently, when top-k routing activates both old and new experts, their non-orthogonal updates interfere, degrading performance on old tasks even if the expert's weights are untouched (See evidence in supplementary S1).

The second failure, **Router Drift**, occurs because the shared router continuously trains on new tasks, causing its decisions to shift over time [28, 31]. It begins to misroute inputs from old tasks to the wrong mixture of experts. This creates a unique form of forgetting: the model remembers the experts' individual skills but forgets how to compose them for previously learned tasks, i.e, compositional forgetting.

To address structural interference, reactive methods like OGD [8] project new updates onto a gradient subspace computed after previous tasks. In contrast, frequency-domain PEFT enables a priori mechanism [11, 56, 58]. While BiLoRA [58] assigns task-

* Corresponding authors.



(a) LoRA-based MoE (b) Ours: Spectral MoE

Figure 1. Comparison between (a) the standard LoRA-based MoE baseline and (b) our proposed Spectral MoE, which leverages disjoint spectral masks (S_i) to ensure expert orthogonality and employs a dual online-offline router.

level disjoint frequency subspaces, this is insufficient for MoE, where multiple experts can be co-activated for the same input and thus still share a task-level mask, leading to overlapping updates. We therefore propose Spectral Experts, which assign expert-level, pairwise-disjoint spectral masks. Under unitary DFT [11, 58] parameterization and conjugate-symmetry constraints, these masks yield Frobenius-orthogonal update matrices, thereby suppressing cross-expert interference observed under top-k routing.

Second, to manage router drift, prior MoE adapters [52, 53] often adopt task-specific routers, which are incompatible with task-agnostic inference. A more principled alternative is to enforce gating consistency on historical inputs for a shared router. Recent work [31] shows that this consistency can be achieved when updates are orthogonal to historical subspaces. In practice, this raises two challenges: (i) how to access these historical statistics, and (ii) avoiding uniform projection on experts, i.e., applying the same projection strength to all experts, which harms plasticity. We address both with a **Dual Online-Offline Router (O^2 Router)** and **Dynamic Consistency Projection (DCP)**: the O^2 Router uses a shared online router for task-agnostic gating and an offline router that archives per-task gating covariances; DCP projects router gradients onto the null space of past inputs to suppress drift and applies an importance-weighted expert projection that strongly protects high-importance experts while keeping others plastic. Our main contributions can be summarized as follows:

- A new Spectral MoE architecture is proposed where each expert is confined to its own masked frequency subspace, guaranteeing a *priori* orthogonality to mitigate structural interference.
- We introduce the dual O^2 Router that decouples online, task-agnostic inference from an offline memory that archives historical expert statistics, which in turn guides the Dynamic Consistency Projection.
- We propose Dynamic Consistency Projection, a geometric constraint with two components: a router projection to suppress drift, and an importance-weighted expert projection to enable an expert-specific stability-plasticity trade-off, mitigating compositional forgetting.
- Validated under a task-agnostic cross-domain CL setting, our framework improves knowledge retention and zero-shot generalization over strong PEFT-CL baselines while efficiently learning new tasks.

2. Related Work

2.1. Continual Learning Settings and Methods

Continual Learning (CL) aims to enable models to learn sequentially without catastrophic forgetting [16, 37]. This challenge spans diverse settings, from Task-Incremental Learning (TIL) [16], which requires task identity at inference, to task-agnostic settings like Class-Incremental (CIL) [37] and Domain-Incremental (DIL) [12]. Our work operates within the highly challenging Cross-domain Task-Agnostic Incremental Learning (X-TAIL) setting [48, 57], which demands robust knowledge retention and generalization across domains without any task identifiers.

To address forgetting, prevailing methods are broadly categorized. Replay-based methods [18, 37, 39], exemplified by iCaRL [37], replay a subset of exemplars alongside new data, though this raises significant memory and scalability concerns. Regularization-based approaches [1, 20, 23] introduce additional loss terms to penalize significant changes to parameters crucial for previous tasks; techniques like EWC [23] constrain parameter shifts, while ZSCL [57] uses reference datasets to preserve robustness without original pre-training data access. Architecture-based methods [42, 51] modify the model's structure, expanding it with task-specific parameters or prompt-based strategies [22, 26, 40, 43, 44]. The critical trade-offs of these traditional strategies, including memory cost, computational overhead, and parameter scalability, motivate the exploration of PEFT as a more efficient foundation for CL.

2.2. Continual Learning with MoE

MoE [19] architectures offer appealing modularity for CL, often leveraging lightweight PEFT experts activated by a router while keeping the large backbone frozen. This formulation aims to reduce inter-

ference by enabling task specialization across experts, sometimes complemented by freezing key experts learned on past tasks. However, applying MoE effectively in challenging task-agnostic settings remains an open problem [48]. Key challenges include mitigating potential interference between experts still operating within a shared parameter space and ensuring router stability to prevent compositional forgetting without relying on task identity. Existing methods often address these via heuristic strategies like partial expert freezing [52, 53], which can offer coarse protection but may limit plasticity, or rely on auxiliary mechanisms incompatible with task-agnostic inference [48]. Our work distinguishes itself by proposing a novel MoE framework that tackles these issues through principled structural decoupling of experts and optimization-level consistency projections.

3. Preliminaries

3.1. Task Settings

We follow a cross-domain, task-agnostic incremental learning setup [48] that requires the model to continually acquire new knowledge while preserving zero-shot generalization to unseen classes. Formally, the model learns a sequence of T tasks $\{\mathcal{T}_1, \ldots, \mathcal{T}_T\}$, where each task $\mathcal{T}_t = (\mathcal{D}_t, \mathcal{C}_t)$ has its own dataset $\mathcal{D}_t$ and label set $\mathcal{C}_t$. Label spaces are mutually disjoint across tasks($\forall i \neq j, \mathcal{C}_i \cap \mathcal{C}_j = \emptyset$) and $\mathcal{D}_t$ becomes inaccessible once training on $\mathcal{T}_t$ finishes. At inference, the model receives no task identity, so it must predict over the combined label space $\mathcal{C} = \mathcal{C}_L^t \cup \mathcal{C}_U$, where $\mathcal{C}_L^t = \bigcup_{\tau=1}^{t} \mathcal{C}_\tau$, and $\mathcal{C}_U$ denotes unseen classes.

3.2. LoRA-based MoE

The MoE [38] architecture enhances model capacity by employing multiple specialized expert subnetworks, which are selectively activated by a router. To adapt this paradigm for parameter-efficient fine-tuning [15, 17, 21, 55], a prevalent strategy is to implement each expert using LoRA [17].

Formally, a LoRA-based MoE layer consists of a router and a set of M experts [7, 30, 46]. The router, typically a linear layer with a softmax activation, takes an intermediate feature $x \in \mathbb{R}^d$ as input and produces a gating vector $a(x) \in \mathbb{R}^M_{\geq 0}$:

$$a(x) = \text{softmax}(xW_g), \quad \|a(x)\|_1 = 1, \tag{1}$$

where $W_g \in \mathbb{R}^{d \times M}$ represents the trainable parameters of the router. Each element $a_m(x), m = [1, ..., M]$ in the gating vector determines the contribution of the corresponding expert $\mathcal{E}_m$. Each expert $\mathcal{E}_m$ is realized as a LoRA module, defined by a pair of low-rank matrices (A_m, B_m), which produces a weight update $\Delta W_m = A_m B_m$. The final output of the entire MoE layer is a combination of the original pre-trained path and the weighted sum of all expert outputs. For a frozen pre-trained weight W_0, the forward pass is defined as:

$$y = W_0 x + \left(\sum_{m=1}^{M} a_m(x)(\Delta W_m) \right) x. \tag{2}$$

While this formulation offers parameter efficiency, its application to continual learning remains challenging due to structural interference between experts and compositional forgetting caused by router drift. These deficiencies necessitate a more principled framework, which we introduce in the following sections.

4. Methodology

4.1. Overview

Our Spectral MoE framework is built upon two strategies that work in concert. First, a **Spectral Experts** (Sec. 4.2) architecture provides structural orthogonality at the parameter level, eliminating structural interference at its source. Second, we introduce an optimization-level **Dynamic Consistency Projection (DCP)** (Sec. 4.4), which applies geometric gradient constraints to suppress router drift and resolve compositional forgetting. This projection mechanism is enabled by our novel dual **Online-Offline Router (O^2 Router)** (Sec. 4.3) mechanism, which decouples online instance routing from an offline memory that archives the historical expert statistics necessary to guide the projection. Together, these components provide a principled solution, illustrated in Figure 2.

4.2. Spectral Experts

In standard LoRA-based MoE (Sec. 3.2), all experts $\{\mathcal{E}_m\}_{m=1}^N$ are trainable in the same parameter space. When learning a new task, gradients flowing into different experts can still collide, because their effective weight updates ΔW_m live in overlapping subspaces. This structural interference becomes more severe as more tasks accumulate. We address the non-orthogonality a priori by parameterizing each expert in a frequency domain so that its updates are Frobenius-orthogonal to all others by design.

Definition 1 (Spectral Expert)**.** *A Spectral Expert E_m is defined by its weight update $\Delta W_m \in \mathbb{R}^{d_o \times d_i}$, synthesized from learnable complex spectral coefficients $\Theta_m \in \mathbb{C}^{d_o \times d_i}$ and a fixed, unique binary mask $S_m \in \{0, 1\}^{d_o \times d_i}$:*

$$\Delta W_m = F_o(S_m \odot \Theta_m)F_i^H. \tag{3}$$

Let $F_o \in \mathbb{C}^{d_o \times d_o}$ and $F_i \in \mathbb{C}^{d_i \times d_i}$ be unitary DFT matrices, and $\odot$ denotes the Hadamard product. We adopt the convention that the inverse 2D DFT [58] is $X = F_o \hat{X} F_i^H$, hence Eq. (3) synthesizes the spatial-domain update from masked spectral coefficients. To ensure ΔW_m is real-valued, we enforce conjugate symmetry. This requires both projecting Θ_m onto the

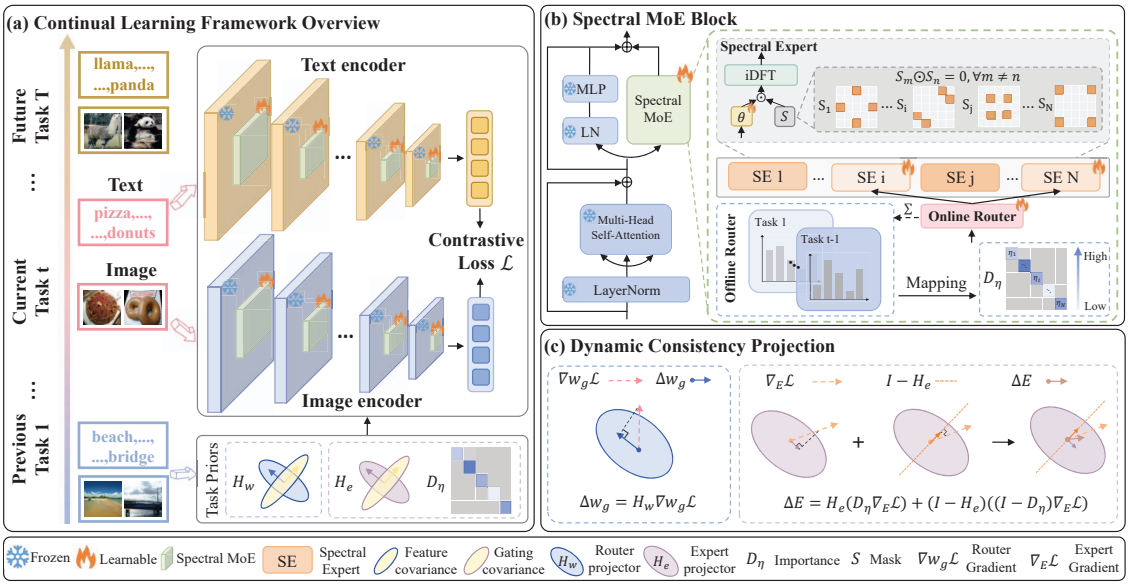


Figure 2. Overview of our Scalable Continual Learning Framework. **(a) Continual Learning Framework Overview:** A dual-encoder Parameter-Efficient Finetuning backbone where frozen blocks are augmented with trainable Spectral MoE modules. The model adapts to a continuous stream of tasks via a contrastive loss $\mathcal{L}$. **(b) Spectral MoE Block:** Each block contains dual Online-Offline Router and Spectral Experts (SEs). SEs are parameterized in the frequency domain with disjoint masks S and learnable frequency components Θ, ensuring innate orthogonality. After task t, offline Priors are archived: router covariance H_w, expert covariance H_e, and average expert usage. **(c) Dynamic Consistency Projection:** Priors are used to protect past tasks. The router gradient is projected: $\Delta w_g = H_w \nabla_{w_g} \mathcal{L}$. The expert gradient undergoes a dynamic two-branch update based on importance D_η: $\Delta E = H_e(D_\eta \nabla_E \mathcal{L}) + (I - H_e)((I - D_\eta)\nabla_E \mathcal{L})$, splitting the update into a stability branch and a plasticity branch.

conjugate-symmetric subspace after each update and constructing S_m to respect this symmetry.

The orthogonality hinges on mask construction. We generate masks $\{S_m\}_{m=1}^N$ to be pairwise disjoint via a layer-wise, without-replacement allocator. Let $\mathcal{P}$ be the set of unique conjugate frequency pairs and $\mathcal{B}$ the subset already assigned. For a new expert E_m with bandwidth budget K_m (number of complex pairs), we sample K_m pairs from $\mathcal{P} \setminus \mathcal{B}$. These locations are set to 1 in S_m, and $\mathcal{B}$ is updated: $\mathcal{B} \leftarrow \mathcal{B} \cup \text{supp}(S_m)$.

Proposition 1 (A Priori Orthogonality of Spectral Experts)**.** *Under unitary DFT parameterization and conjugate-symmetric, pairwise-disjoint masks $\{S_m\}$ satisfying $S_m \odot S_n = 0$ for all $m \neq n$, the updates are Frobenius-orthogonal, i.e.,*

$$\langle \Delta W_m, \Delta W_n \rangle_F = 0. \tag{4}$$

***then** their effective weight updates, ΔW_m and ΔW_n, are strictly orthogonal in the spatial domain. Full proof is in the supplementary S2.*

Unlike task-level frequency masking, we assign a unique, expert-level mask S_m to each expert. Consequently, even when multiple experts are co-activated for the same input, their updates remain orthogonal under the stated conditions; because masks are fixed, this property is training-invariant, confining modifications to dedicated bands and substantially reducing cross-expert interference.

4.3. Dual Online-Offline Router

However, ensuring the integrity of expert parameters is not enough; the model must also remember how to access the correct expert for a given task. To resolve this router drift and enable task-agnostic, instance-level routing, our O^2 Router decouples online inference from offline analysis. It consists of a shared online router for real-time gating and an offline memory that archives historical expert statistics to guide our dynamic consistency projection.

Online Shared Instance Router At the core of our routing system lies a single shared instance router, G^I_{shared}, parameterized by W_g, which operates online during both training and inference. For any given intermediate feature $x \in \mathbb{R}^d$, it performs real-time gating, producing a sparse gating vector $a(x) \in \mathbb{R}^N_{\geq 0}$:

$$a(x) = G^I_{\text{shared}}(x; W_g) = \sigma(Topk(xW_g, k)), \tag{5}$$

where σ is the softmax function, the $Topk(\cdot, k)$ function ensures sparse activation of the top k experts. By sharing W_g across all tasks, this router inherently facilitates the learning and reuse of common expert compositions, thereby promoting knowledge transfer. Its dependence on the specific instance x allows for fine-grained expert selection tailored to the input's characteristics, overcoming the granularity limitations of task-level routers.

Offline Task Specific Router Complementing the online router, we introduce an offline task specific

router, G_t^T. This is not a learnable network but rather a static summary vector $\bar{a}_t$ computed after training on task $\mathcal{T}_t$ is completed. It captures the average expert usage pattern for that task:

$$\bar{a}_t = G_t^T(x; null) \;=\; \frac{1}{|\mathcal{D}_t|} \sum_{x \in \mathcal{D}_t} a(x) \in \mathbb{R}^N, \qquad (6)$$

where $a(x)$ is the gating vector produced by the shared instance router G^I_{shared} during task t' s training. Crucially, G_t^T does not participate in online gating decisions. Instead, it is archived as historical meta-information, providing essential context about expert importance that guides the dynamic consistency projection mechanism in subsequent learning stages.

4.4. Dynamic Consistency Projection

Building upon the dual-router architecture, we introduce Dynamic Consistency Projection, a core mechanism that addresses router drift and expert composition forgetting by applying dynamic gradient projections to balance stability and plasticity.

Consistency Objective and Conditions Our fundamental goal is to ensure that for any input x^t from a past task, the model's output remains unchanged even after training on new tasks. The detailed derivation (Supp.S3) shows that this objective is guaranteed if two sufficient conditions are met simultaneously: i) *Router Consistency*: The router update ΔW_g must be orthogonal to historical inputs x^t: $x^t \Delta W_g = \mathbf{0}$.

ii) *Expert Composition Consistency*: The weighted sum of sparse spectral expert updates $\Delta\Theta'_m = S_m \odot \Delta\Theta_m$ must be zero under historical gating vectors a^t: $\sum_{m=1}^{N} a_m^t \Delta\Theta'_m = \mathbf{0}$.

Router Consistency Projection To enforce router consistency, we project the router gradient onto the null space of historical inputs. Let $X_t = [x^{t,1}, \ldots, x^{t,|D_t|}]^\top \in \mathbb{R}^{|D_t| \times d}$ be the stacked features from task t. Define the uncentered covariance $\bar{X}_t = X_t^\top X_t$. We compute its SVD and extract right singular vectors associated with near-zero singular values to form a null-space basis $\tilde{V}_x \in \mathbb{R}^{d \times R_x}$. The projector is

$$H_w = \tilde{V}_x \tilde{V}_x^\top \in \mathbb{R}^{d \times d}. \qquad (7)$$

Given the loss gradient $\nabla_{W_g} L$ for the online router $W_g \in \mathbb{R}^{d \times N}$, we update

$$\Delta W_g \;=\; H_w \, \nabla_{W_g} \mathcal{L}. \qquad (8)$$

This ensures that the resulting update ΔW_g is orthogonal to the past inputs.

Dynamic Expert Projection Instead of rigidly enforcing that the weighted sum of expert updates is zero, we decompose the gradient into a stability component to preserve past knowledge and a plasticity component to acquire new knowledge. We first vectorize the masked parameters per expert $e_m = \text{vec}(S \odot \Theta_m) \in \mathbb{R}^K$, where K denotes the number of frequency parameters, and stack rows:

$$E = \begin{bmatrix} e_1^\top \\ \vdots \\ e_N^\top \end{bmatrix} \in \mathbb{R}^{N \times K}. \qquad (9)$$

Let $A_t = [a^{t,1}, \ldots, a^{t,|D_t|}]^\top \in \mathbb{R}^{|D_t| \times N}$ be historical gates from the shared instance router, with $\bar{A}_t = A_t^\top A_t$. As above, SVD of $\bar{A}_t = A_t^\top A_t$ yields a null-space basis $\tilde{V}_a \in \mathbb{R}^{N \times R_a}$ and

$$H_e = \tilde{V}_a \tilde{V}_a^\top \in \mathbb{R}^{N \times N}. \qquad (10)$$

This projection matrix defines a stable subspace that is orthogonal to past expert compositions, directly enforcing our Expert Composition Consistency objective. We employ a two-branch decomposition update rule, which strictly separates the gradient into a null space for stability and a residual space for plasticity:

$$\Delta E \;=\; \underbrace{H_e \left(D_\eta \, \nabla_E \mathcal{L}\right)}_{\text{stability}} + \underbrace{\left(I_N - H_e\right)\left(\left(I_N - D_\eta\right) \nabla_E \mathcal{L}\right)}_{\text{plasticity}}, \qquad (11)$$

where $I_N \in \mathbb{R}^{N \times N}$ is the identity matrix, $D_\eta = \text{diag}(\eta_1, \ldots, \eta_N) \in \mathbb{R}^{N \times N}$ is a diagonal relaxation matrix, where each stability coefficient η_m is determined by expert importance derived from the offline task-specific router via a monotonic mapping:

$$\eta_m \;=\; \eta_{\min} + (\eta_{\max} - \eta_{\min}) \, \bar{a}^{\gamma}_{\{t,m\}}, \qquad (12)$$

where $0 < \eta_{\min} \leq \eta_{\max} \leq 1$ and $\gamma \geq 1$, $\eta_{\min}, \eta_{\max}, \gamma$ are hyperparameter, $\bar{a}_{\{t,m\}}$ is the normalized value from Eq. (6) for task t and expert m. This mapping ensures that more important experts receive a stability coefficient η_m closer to $\eta_{\max}$, thus channeling their gradient updates more strongly into the stable null space defined by H_e. This dynamically balances the need for consistency on past tasks with plasticity for new ones.

5. Experiments

5.1. Experimental Setting

Benchmarks Our experimental evaluation follows the setting in [52] to assess our methods under full-shot,few-shot MTIL and CIL. We use the 11 task benchmark sequence, comprising: Aircraft [32], Caltech101 [9], CIFAR100 [25], DTD [3], EuroSAT [14], Flowers [34], Food [2], MNIST [5], OxfordPet [35], StanfordCars [24], and SUN397 [47]. For CIL, we experiment on TinyImageNet [49]. The 100 classes from TinyImageNet are divided into 5, 10, 20 subsets to evaluate class distribution adaptability. We use three metrics from [57]: "Last", the average accuracy on previously seen tasks to measure retention; "Transfer", performance on unseen tasks to evaluate generalization; and "Average", the mean of both.

	Method	Aircraft [32]	Caltech101 [9]	CIFAR100 [25]	DTD [3]	EuroSAT [14]	Flowers [34]	Food [2]	MNIST [5]	OxfordPet [35]	Cars [24]	SUN397 [47]	*Average*
CLIP	Zero-shot	24.3	88.4	68.2	44.6	54.9	71.0	88.5	59.4	89.0	64.7	65.2	65.3
	Full Fine-tune	62.0	95.1	89.6	79.5	98.9	97.5	92.7	99.6	94.7	89.6	81.8	89.2
	Fine-tune Adapter	56.8	92.6	89.4	79.0	98.4	97.0	92.9	99.2	94.1	89.1	82.7	88.3
Transfer	Continual-FT	–	67.1	46.0	32.1	35.6	35.0	57.7	44.1	60.8	20.5	46.6	44.6
	LwF [29]	–	74.5	56.9	39.1	51.1	52.6	72.8	60.6	75.1	30.3	55.9	58.9
	iCaRL [37]	–	56.5	44.6	32.7	39.3	46.6	68.0	46.0	77.4	31.9	60.5	50.4
	LwF-VR [6]	–	77.1	61.0	40.5	45.3	54.4	74.6	47.9	76.7	36.3	58.6	57.2
	WiSE-FT [45]	–	73.5	55.6	35.6	41.5	47.0	68.3	53.9	69.3	26.8	51.9	52.3
	ZSCL [57]	–	86.0	67.4	**45.4**	50.4	69.1	87.6	61.8	86.8	60.1	**66.8**	68.1
	MoE-Adapters [52]	–	87.9	68.2	44.4	49.9	70.7	88.7	59.7	89.1	64.5	65.5	68.9
	MoE-Adapters++[53]	–	87.9	68.2	45.1	54.6	**71.2**	88.8	59.5	88.6	63.3	63.1	69.0
	Ours	–	**88.1**	**68.4**	44.3	**55.4**	**71.2**	**89.0**	**64.4**	**89.3**	**64.7**	66.4	**70.1**
Average	Continual-FT	25.5	81.5	59.1	53.2	64.7	51.8	63.2	64.3	69.7	31.8	49.7	55.9
	LwF [29]	36.3	86.9	72.0	59.0	73.7	60.0	73.6	74.8	80.0	37.3	58.1	64.7
	iCaRL [37]	35.5	89.2	72.2	60.6	68.8	70.0	78.2	62.3	81.8	41.2	62.5	65.7
	LwF-VR [6]	29.6	87.7	74.4	59.5	72.4	63.6	77.0	66.7	81.2	43.7	60.7	65.1
	WiSE-FT [45]	26.7	86.5	64.3	57.1	65.7	58.7	71.1	70.5	75.8	36.9	54.6	60.7
	ZSCL [57]	45.1	92.0	80.1	64.3	79.5	81.6	89.6	75.2	88.9	64.7	**68.0**	75.4
	MoE-Adapters [52]	50.2	91.9	83.1	69.4	78.9	84.0	89.1	73.7	89.3	67.7	66.9	76.7
	MoE-Adapters++[53]	**55.8**	**94.6**	81.4	**70.4**	**82.4**	83.6	90.0	73.5	90.1	66.0	64.2	77.5
	Ours	51.3	**94.6**	**85.0**	69.1	80.5	**84.9**	**90.3**	**76.9**	**90.7**	**68.2**	67.5	**78.1**
Last	Continual-FT	31.0	89.3	65.8	67.3	88.9	71.1	85.6	**99.6**	92.9	77.3	81.1	77.3
	LwF [29]	26.3	87.5	71.9	66.6	79.9	66.9	83.8	**99.6**	92.1	66.1	80.4	74.6
	iCaRL [37]	35.8	93.0	77.0	70.2	83.3	88.5	90.4	86.7	93.2	81.2	**81.9**	80.1
	LwF-VR [6]	20.5	89.8	72.3	67.6	85.5	73.8	85.7	**99.6**	93.1	73.3	80.9	76.6
	WiSE-FT [45]	27.2	90.8	68.0	68.9	86.9	74.0	87.6	**99.6**	92.6	77.8	81.3	77.7
	ZSCL [57]	40.6	92.2	81.3	70.5	94.8	90.5	91.9	98.7	93.9	**85.3**	80.2	83.6
	MoE-Adapters [52]	49.8	92.2	86.1	78.1	95.7	94.3	89.5	98.1	89.9	81.6	80.0	85.0
	MoE-Adapters++[53]	**55.8**	95.2	84.3	**79.9**	**98.2**	**96.3**	91.4	98.1	94.2	78.3	76.2	86.2
	Ours	50.3	**95.4**	**88.6**	78.5	95.0	95.3	**92.0**	98.7	**94.3**	83.9	78.0	**86.3**

Table 1. Comparison with state-of-the-art methods on MTIL benchmark in terms of "Transfer", "Average", and "Last" scores (%). "Ours" indicates our method trained for 1k iterations. We mark the best and second-best methods with bold and underline styles, respectively. The same applies to the subsequent tables.

Implementation Details The training process is carried out using the AdamW optimizer, with a learning rate of 0.001 and a batch size of 64 across all tasks. Our Spectral MoE architecture employs a total of $N = 32$ experts, with the router selecting the $topk = 4$ experts per input. For each Spectral Expert, we set the number of non-zero frequency components following [58]. For the Eq. (12), we set the stability coefficients to $\eta_{\text{min}} = 0.95$, $\eta_{\text{max}} = 1.0$, and the focusing strength to $\gamma = 2$.

5.2. Experimental Analysis

Full-shot MTIL As shown in Table 1, our method achieves SOTA performance on the 11-task MTIL benchmark across all three key metrics. Our Average score of 78.1% surpasses the strongest competitor, MoE-Adapters++, which achieved 77.5%. This overall victory stems from our superior balance of stability and plasticity, which addresses forgetting from two different angles. For stability, our Last score of 86.3% is the highest. This is enabled by our dual-pronged solution: Spectral Experts enforce a priori orthogonality to remove structural interference, while our DCP protects against compositional forgetting. For plasticity, our SOTA Transfer score of 70.1% demonstrates that our protection is not rigid. The DCP, driven by Dual O^2 Router memory, uses importance-weighted projections to protect only critical experts, leaving others plastic and avoiding the coarse freezing of prior methods. This superior balance of retention and generalization yields the best overall performance.

Few-shot MTIL As shown in Table 2, our method maintains its SOTA performance on the 11-task Few-shot MTIL benchmark, demonstrating our framework's robustness and sample efficiency in the data-scarce regime. Our Average score of 72.1% surpasses the strongest competitor, MoE-Adapters++, which achieved 71.7%. For stability, our Last score of 76.5% is the highest, as Spectral Experts enforce a priori orthogonality, ensuring the limited new data is learned in a clean, non-interfering subspace. Moreover, our SOTA Transfer score of 69.7% proves our protection is not a rigid "freeze". The DCP, guided by dual online-offline router statistics, keeps non-critical experts plastic. This allows the model to generalize from its existing knowledge to learn the new task from a few samples, rather than requiring data-hungry training from scratch. This superior balance of retention and ge neralization yields the best overall performance.

CIL To validate our framework beyond multi-domain MTIL, we evaluated it on the single-domain CIL benchmark using TinyImageNet, as shown in Table 3. Our method demonstrates superior scalability by consistently achieving the SOTA Average and Last scores across all 5, 10, and 20-step settings. Our top Average score of 81.95% surpasses the strong MoE-Adapters baseline. This robust performance validates the effectiveness of our Spectral Experts in a CIL context. Unlike the diverse MTIL benchmarks, this single-domain setting forces the model to separate many closely related classes. The priori orthogonal frequency subspaces are critical here, as they prevent structural interference between these fine-grained representations. This confirms that our frequency-based separation is a highly effective and scalable solution for both single- and multi-domain continual learning.

	Method	Aircraft [32]	Caltech101 [9]	CIFAR100 [25]	DTD [3]	EuroSAT [14]	Flowers [34]	Food [2]	MNIST [5]	OxfordPet [35]	Cars [24]	SUN397 [47]	*Average*
CLIP	Zero-shot	24.3	88.4	68.2	44.6	54.9	71.0	88.5	59.4	89.0	64.7	65.2	65.3
	5-shot Full Fine-tune	30.6	93.5	76.8	65.1	91.7	92.9	83.3	96.6	84.9	65.4	71.3	77.5
	5-shot Fine-tune Adapter	29.7	90.0	75.3	63.9	81.1	94.2	87.8	90.4	89.0	68.2	72.5	76.6
Transfer	Continual-FT	–	72.8	53.0	36.4	35.4	43.3	68.4	47.4	72.6	30.0	52.7	51.2
	LwF [29]	–	72.1	49.2	35.9	44.5	41.1	66.6	50.5	69.0	19.0	51.7	50.0
	LwF-VR [6]	–	82.2	62.5	40.1	40.1	56.3	80.0	60.9	77.6	40.5	60.8	60.1
	WiSE-FT [45]	–	77.6	60.0	41.3	39.4	53.0	76.6	58.1	75.5	37.3	58.2	57.7
	ZSCL [57]	–	84.0	68.1	44.8	46.8	63.6	84.9	61.4	81.4	55.5	62.2	65.3
	MoE-Adapters [52]	–	87.9	68.2	44.1	48.1	64.7	88.8	**69.0**	89.1	64.5	65.1	68.9
	MoE-Adapters++ [53]	–	87.9	68.2	**45.1**	54.6	**71.3**	88.7	59.5	89.1	63.9	64.9	69.3
	Ours	–	**88.1**	**68.4**	44.1	**55.3**	71.2	**89.0**	59.9	**89.3**	**64.7**	**66.4**	**69.7**
Average	Continual-FT	28.1	86.4	59.1	52.8	55.8	62.0	70.2	64.7	75.5	35.0	54.0	58.5
	LwF [29]	23.5	77.4	43.5	41.7	43.5	52.2	54.6	63.4	68.0	21.3	52.6	49.2
	LwF-VR [6]	24.9	89.1	64.2	53.4	54.3	70.8	79.2	66.5	79.2	44.1	61.6	62.5
	WiSE-FT [45]	32.0	87.7	61.0	55.8	68.1	69.3	76.8	71.5	77.6	42.0	59.3	63.7
	ZSCL [57]	28.2	88.6	66.5	53.5	56.3	73.4	83.1	56.4	82.4	57.5	62.9	64.4
	MoE-Adapters [52]	30.0	89.6	**73.9**	58.7	69.3	79.3	88.1	**76.5**	**89.1**	**65.3**	65.8	71.4
	MoE-Adapters++ [53]	32.7	**93.9**	70.8	**59.0**	73.9	80.1	**88.3**	72.3	89.0	63.9	65.2	71.7
	Ours	**33.0**	93.1	**73.9**	58.2	**74.3**	**82.6**	86.7	71.5	88.4	65.0	**66.9**	**72.1**
Last	Continual-FT	27.8	86.9	60.1	58.4	56.6	75.7	73.8	93.1	82.5	57.0	66.8	67.1
	LwF [29]	22.1	58.2	17.9	32.1	28.1	66.7	46.0	84.3	64.1	31.5	60.1	46.5
	LwF-VR [6]	22.9	89.8	59.3	57.1	57.6	79.2	78.3	77.7	83.6	60.1	69.8	66.9
	WiSE-FT [45]	30.8	88.9	59.6	60.3	80.9	81.7	77.1	**94.9**	83.2	62.8	70.0	71.9
	ZSCL [57]	26.8	88.5	63.7	55.7	60.2	82.1	82.6	58.6	85.9	66.7	70.4	67.4
	MoE-Adapters [52]	30.1	89.3	74.9	64.0	82.3	89.4	87.1	89.0	**89.1**	**69.5**	**72.5**	76.1
	MoE-Adapters++ [53]	**32.7**	**94.5**	71.3	**65.2**	84.9	87.7	**87.8**	94.8	89.0	63.3	68.0	76.3
	Ours	**32.7**	93.6	**75.19**	63.5	**85.2**	**92.2**	84.0	91.6	85.77	65.9	71.4	**76.5**

Table 2. Comparison with state-of-the-art methods on few-shot MTIL benchmark (%).

Method	5 step		10 step		20 step	
	Avg.	Last	Avg.	Last	Avg.	Last
CLIP Zero-shot	69.62	65.30	69.55	65.59	69.49	65.30
Fine-tune	61.54	46.66	57.05	41.54	54.62	44.55
LwF [29]	60.97	48.77	57.60	44.00	54.79	42.26
iCaRL [37]	77.02	70.39	73.48	65.97	69.65	64.68
LwF-VR [6]	77.56	70.89	74.12	67.05	69.94	63.89
ZSCL [57]	80.27	73.57	78.61	71.62	77.18	68.30
MoE-Adapters[52]	81.12	76.81	80.23	76.35	79.96	75.77
Ours	**81.95**	**77.12**	**81.85**	**76.52**	**80.52**	**76.09**

Table 3. Comparison of different methods on TinyImageNet in class-incremental settings with 100 base classes.

Method	Train Params ↓	GPU ↓	Times ↓
LWF [42]	149.6M	32172MiB	1.54s/it
LWF-VR [15]	149.6M	32236MiB	1.51s/it
ZSCL [78]	149.6M	26290MiB	3.94s/it
MoE-Adapters	59.8M	22358MiB	1.58s/it
Ours	23.5 M	21580MiB	1.24s/it

Table 4. A comparison of computational costs during the training process between our method and other methods, in terms of trainable parameters, GPU memory usage, and training time per iteration.

Computational Cost As shown in Table 4, we analyze the efficiency of our method against key baselines, comparing trainable parameters, GPU memory footprint, and training speed. Our framework is the most efficient solution across all three metrics. For Trainable Parameters, our method requires only 23.49 M, which is approximately 2.5 times fewer than MoE-Adapters and 6.4 times fewer than the full-parameter method ZSCL. This significant efficiency stems from our Spectral Expert design, which replaces parameter-heavy LoRA modules with a small, fixed set of k spectral coefficients. This reduced parameter count directly translates to the lowest GPU memory footprint at 21580 MiB. Furthermore, our method is the fastest in terms of training speed, achieving 1.24s/it. This is notably faster than MoE-Adapters and over 3 times faster than ZSCL. This validates that our DCP is a highly efficient protection mechanism, avoiding the significant computational overhead associated with regularization or distillation based methods.

5.3. Ablation Study

Analysis of Spectral Experts To isolate the contribution of our Spectral Experts (SE), we conduct an ablation that replaces the expert modules in our full Spectral MoE with either standard LoRA experts or SE, and report the Last Forgetting Rate (FR) under two regimes: (a) full-shot and (b) few-shot. We summarize performance using the Area Under the Forgetting–Time Curve (AUC); detailed definitions and computation procedures for FR and AUC are provided in the supplementary material (Supp. S4). As shown in Figure 4, in the full-shot setting (a) simply swapping LoRA for SE reduces the total forgetting by about half, with AUC 0.91% vs. 1.81% for LoRA. This supports our claim that SE's a priori orthogonality better mitigates structural interference arising from overlapping subspaces in LoRA. The trend persists in the few-shot setting (b): the framework paired with LoRA shows AUC 1.68%, whereas integrating SE lowers forgetting to 0.61%. These results indicate that SE is a key factor in reducing structural interference across data regimes, and that its synergy with DCP further contributes to the overall gains.

Analysis of DCP We now ablate the components of our DCP, with results in Table 5. DCP is designed to solve "compositional forgetting" by enforcing geometric constraints on both the router and the experts.

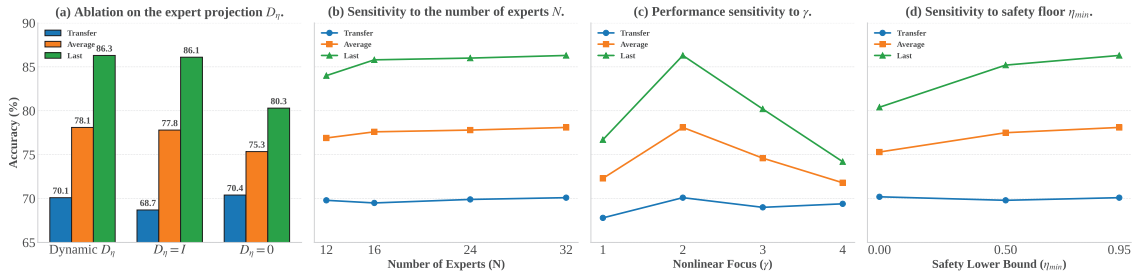


Figure 3. Ablation studies and sensitivity analysis. (a) Ablation on the expert projection D_η. We compare our "Dynamic D_η" against two rigid baselines: (a) $D_\eta = I$, a full stability setting, and (b) $D_\eta = 0$, a full plasticity setting. Our dynamic approach yields the best overall performance. (b-d) Sensitivity analysis of key hyperparameters: (b) the number of experts N, (c) the non-linear focus γ, and (d) the safety floor η_{min}. Our method shows stable performance across a range of N values, peaking at $N = 32$. The optimal performance is achieved with $\gamma = 2$ and $\eta_{min} = 0.95$. All experiments report Transfer (blue, circle), Average (orange, square), and Last (green, triangle) accuracy (%).

Method	Transfer	Average	Last
Full	**70.1**	**78.1**	**86.3**
(a) w/o H_w	66.2	75.1	84.0
(b) w/o $H_e(D_\eta)$	69.4	74.6	79.8
(c) w/o H_w & $H_e(D_\eta)$	69.8	75.0	80.2

Table 5. Component ablations on our method. "Full" is our complete method. We measure the impact of removing the router projection (a: w/o H_w), the dynamic expert projection (b: w/o $H_e(D_\eta)$), or both (c).

Removing the Router Consistency Projection (H_w) alone in (a) significantly drops the Average score to 75.1%, confirming H_w is essential for stabilizing the router. However, removing the Dynamic Expert Projection ($H_e(D_\eta)$) in (b) causes a more severe performance collapse. This results in the lowest Average score of 74.6% and the largest drop in the Last score to 79.8%. This demonstrates that our fine-grained, importance-weighted protection is the most critical component for maintaining stability. Finally, row (c), which removes both DCP components, confirms that our SE alone are insufficient. While SE solves structural interference, it is the synergy of SE and DCP that achieves our "Full" SOTA performance.

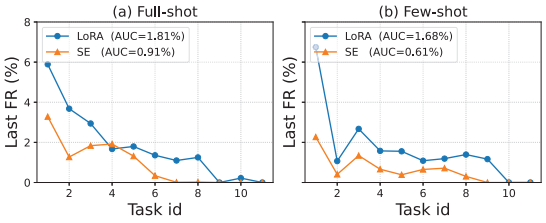


Figure 4. Analysis of Spectral Experts (SEs) vs. LoRA. We plot the Last Forgetting Rate for standard LoRA versus our SEs on our full Spectral MoE framework under (a) full-shot and (b) few-shot scenarios. The Area Under Curve (AUC) measures total forgetting.

Figure 3a analyzes our Dynamic D_η by comparing it to two rigid extremes. The "full stability" setting ($D_\eta = I$) is too rigid, dropping the Transfer score to 68.7%. Conversely, the "full plasticity" setting ($D_\eta = 0$) causes catastrophic forgetting, as the Last score plummets to 80.3%. Our Dynamic D_η method avoids both pitfalls, achieving the highest Last score (86.3%) while maintaining a high Transfer score (70.1%). This proves our dynamic, importance-weighted balance is the key to the best Average.

5.4. Hyperparameter Sensitivity

We analyze the sensitivity of our key hyperparameters in Figure 3 (b-d). We provide a brief summary here; full details are deferred to the supplementary material (supp.S5). Our framework benefits from a finer-grained spectral partition, as performance improves with more specialized experts, peaking at our chosen $N = 32$. The focusing strength γ is critical for balancing stability and plasticity; a linear mapping ($\gamma = 1$) is too blunt to protect important experts, while higher values ($\gamma \geq 3$) are too rigid, making our setting of $\gamma = 2$ optimal. Finally, a high safety floor η_{min} is essential. Setting no floor ($\eta_{\text{min}} = 0$) causes catastrophic forgetting in less-used experts, and performance increases monotonically with this value, validating our choice of $\eta_{\text{min}} = 0.95$ to ensure robust knowledge retention across all experts.

6. Conclusion

This paper addresses the core challenges of Structural Interference and Router Drift when applying LoRA-MoE to Continual Learning by proposing a unified framework combining spectral orthogonality with geometric projection. We introduced Spectral Experts to guarantee a priori orthogonality, eliminating parameter overlap at its source, and designed a Dynamic Consistency Projection mechanism based on a Dual Online-Offline Router to resolve compositional forgetting in a principled manner. Experiments on MTIL and CIL benchmarks demonstrate our method achieves SOTA performance. Ablation studies further confirm that Spectral Experts are critical for eliminating structural interference and that DCP is essential for preventing compositional forgetting.

7. Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant No. 62306003), the Open Research Fund of Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) (Grant No. GML-KF-24-29), and the Anhui Provincial Key Research and Development Program under Grant NO.202304a05020047.

References

[1] Hongjoon Ahn, Sungmin Cha, Donggyu Lee, and Taesup Moon. Uncertainty-based continual learning with adaptive regularization. *Advances in neural information processing systems*, 32, 2019. 2

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014. 5, 6, 7

[3] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 5, 6, 7

[4] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021. 1

[5] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012. 5, 6, 7

[6] Yuxuan Ding, Lingqiao Liu, Chunna Tian, Jingyuan Yang, and Haoxuan Ding. Don't stop learning: Towards continual learning for the clip model. *arXiv preprint arXiv:2207.09248*, 2022. 6, 7

[7] Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhiheng Xi, Xiaoran Fan, et al. Loramoe: Alleviating world knowledge forgetting in large language models via moe-style plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1932–1945, 2024. 3

[8] Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *International conference on artificial intelligence and statistics*, pages 3762–3773. PMLR, 2020. 1

[9] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 5, 6, 7

[10] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International journal of computer vision*, 132(2):581–595, 2024. 1

[11] Ziqi Gao, Qichao Wang, Aochuan Chen, Zijing Liu, Bingzhe Wu, Liang Chen, and Jia Li. Parameter-efficient fine-tuning with discrete fourier transform. *arXiv preprint arXiv:2405.03003*, 2024. 1, 2

[12] Prachi Garg, Rohit Saluja, Vineeth N Balasubramanian, Chetan Arora, Anbumani Subramanian, and CV Jawahar. Multi-domain incremental learning for semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 761–771, 2022. 2

[13] Raia Hadsell, Dushyant Rao, Andrei A Rusu, and Razvan Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in cognitive sciences*, 24(12):1028–1040, 2020. 1

[14] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 5, 6, 7

[15] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019. 3

[16] Yen-Chang Hsu, Yen-Cheng Liu, Anita Ramasamy, and Zsolt Kira. Re-evaluating continual learning scenarios: A categorization and case for strong baselines. *arXiv preprint arXiv:1810.12488*, 2018. 2

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022. 1, 3

[18] David Isele and Akansel Cosgun. Selective experience replay for lifelong learning. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2

[19] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. 1, 2

[20] Saurav Jha, Dong Gong, He Zhao, and Lina Yao. Npcl: Neural processes for uncertainty-aware continual learning. *Advances in Neural Information Processing Systems*, 36:34329–34353, 2023. 2

[21] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pages 709–727. Springer, 2022. 1, 3

[22] Dahuin Jung, Dongyoon Han, Jihwan Bang, and Hwanjun Song. Generating instance-level prompts for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11847–11857, 2023. 2

[23] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 2

[24] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained cate-

gorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 5, 6, 7

[25] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5, 6, 7

[26] Muhammad Rifki Kurniawan, Xiang Song, Zhiheng Ma, Yuhang He, Yihong Gong, Yang Qi, and Xing Wei. Evolving parameterized prompt memory for continual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13301–13309, 2024. 2

[27] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 3045–3059, 2021. 1

[28] Hongbo Li, Sen Lin, Lingjie Duan, Yingbin Liang, and Ness Shroff. Theory on mixture-of-experts in continual learning. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[29] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 6, 7

[30] Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. When moe meets llms: Parameter efficient fine-tuning for multi-task medical applications. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 1104–1114, 2024. 3

[31] 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. 1, 2

[32] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5, 6, 7

[33] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, pages 109–165. Elsevier, 1989. 1

[34] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 5, 6, 7

[35] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 5, 6, 7

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1

[37] 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, 6, 7

[38] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 3

[39] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017. 2

[40] 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

[41] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5227–5237, 2022. 1

[42] 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. 2

[43] 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. 2

[44] 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

[45] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971, 2022. 6, 7

[46] Xun Wu, Shaohan Huang, and Furu Wei. Mixture of loRA experts. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[47] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 5, 6, 7

[48] Yicheng Xu, Yuxin Chen, Jiahao Nie, Yusong Wang, Huiping Zhuang, and Manabu Okumura. Advancing cross-domain discriminability in continual learning of vision-language models. *Advances in Neural Informa-*

tion Processing Systems, 37:51552–51576, 2024. 2, 3

[49] Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3014–3023, 2021. 5

[50] Shuo Yang, Kun-Peng Ning, Yu-Yang Liu, Jia-Yu Yao, Yong-Hong Tian, Yi-Bing Song, and Li Yuan. Is parameter collision hindering continual learning in llms? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4243–4259, 2025. 1

[51] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. In *International Conference on Learning Representations*, 2018. 2

[52] Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23219–23230, 2024. 1, 2, 3, 5, 6, 7

[53] Jiazuo Yu, Zichen Huang, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Moe-adapters++: Towards more efficient continual learning of vision-language models via dynamic mixture-of-experts adapters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2, 3, 6, 7

[54] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, 2022. 1

[55] Jeffrey O Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, and Jitendra Malik. Side-tuning: a baseline for network adaptation via additive side networks. In *European conference on computer vision*, pages 698–714. Springer, 2020. 3

[56] Yifei Zhang, Hao Zhu, Alysa Ziying Tan, Dianzhi Yu, Longtao Huang, and Han Yu. pfedmxf: Personalized federated class-incremental learning with mixture of frequency aggregation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30640–30650, 2025. 1

[57] Zangwei Zheng, Mingyuan Ma, Kai Wang, Ziheng Qin, Xiangyu Yue, and Yang You. Preventing zero-shot transfer degradation in continual learning of vision-language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19125–19136, 2023. 1, 2, 5, 6, 7

[58] Hao Zhu, Yifei Zhang, Junhao Dong, and Piotr Koniusz. Bilora: almost-orthogonal parameter spaces for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25613–25622, 2025. 1, 2, 3, 6