LEARNING TO RE-THINK: GATED RECURRENCE WITH LORA FOR EFFICIENT AND EFFECTIVE DOMAIN INCREMENTAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

The adaptation of large-scale foundation models for real-world medical Domain Incremental Learning (DIL) is challenged by data scarcity, significant domain shifts, and severe class imbalance. Current parameter-efficient methods often present a trade-off between knowledge integration, which risks task interference, and parameter isolation, which sacrifices forward transfer. To address this trade-off, we propose a framework that achieves both domain specialization and integrated knowledge transfer. Our two-tiered adaptive paradigm enables a foundation model to learn domain-specific representations while systematically transferring knowledge across a sequence of tasks. For intra-domain specialization, we introduce Recursive LoRA (RecLoRA), a dynamic computation module where a learnable router directs tokens for iterative feature refinement by a shared LoRA block, focusing computation on complex inputs. For inter-domain integration, our Sequential Knowledge Transfer strategy preserves domain-specific expertise by training independent RecLoRA modules for each task, while promoting forward transfer by using the converged weights of a previous task's modules to initialize those of the next. Built upon a frozen foundation model, our framework employs an efficient key-query mechanism for inference-time expert selection. We demonstrate that our approach sets a new state-of-the-art on challenging diabetic retinopathy DIL benchmarks, validating its efficacy for real-world clinical applications.

1 INTRODUCTION

Large-scale medical foundation models have demonstrated significant potential in domains such as ophthalmology (Qiu et al., 2024; Zhou et al., 2023) and radiology (Tiu et al., 2022; Wang et al., 2024; Yan et al., 2022), etc. These models are often trained on data collected from various sources, including different types of machines with unique scanning parameters and from diverse countries. However, their effectiveness is limited by their reliance on static, retrospective datasets. This reliance makes it difficult for the models to adapt to the ever-changing and diverse nature of medical data, which can vary significantly across different institutions, countries, and patient demographics (Queiroz et al., 2025; He et al., 2024; Sheng et al., 2025). Consequently, these models may struggle to generalize well to diverse patient populations, incorporate new medical knowledge, and accurately handle rare conditions.

To address these challenges, domain incremental learning (DIL) has emerged as a crucial approach for large-scale medical foundation models (Chen et al., 2024; Wang et al., 2025b). DIL enables these models to continuously integrate new information while retaining previously acquired knowledge, thereby enhancing their adaptability and clinical utility. However, implementing DIL effectively is challenging due to the substantial number of parameters in these models and the issue of catastrophic forgetting. Several strategies have been proposed to tackle these challenges in DIL. One approach involves regularization techniques, which help the model learn new information while retaining old knowledge (Wang et al., 2022c;b). For example, some methods use a shared pool of prompts with a mechanism to select task-specific subsets, which can lead to task interference as the pool grows. Alternatively, other methods assign independent prompts to each task to prevent forgetting (Wang et al., 2022a), but this limits the potential for forward knowledge transfer. More recent hybrid approaches (Smith et al., 2023; Wang et al., 2025a; Xu et al., 2025) have introduced complex mechanisms to balance shared and task-specific knowledge.

These methods generally overcome catastrophic forgetting across domains but often overlook the issue of class distribution shift, where the distribution of classes varies across datasets. This shift can lead to imbalanced learning, particularly when rare but clinically significant conditions are underrepresented, resulting in suboptimal model performance. This problem is especially significant in the medical field, where underrepresented conditions may lead to missed diagnoses or inadequate treatment recommendations.

Beyond this inter-domain dilemma, we identify a more fundamental limitation inherited by both paradigms: they rely on a static and homogeneous computational graph. Every input token, whether representing healthy tissue or a subtle, rare lesion, undergoes the same fixed-depth processing. This uniform computational approach is profoundly inefficient for imbalanced data. It leads to the underprocessing of diagnostically critical tokens from minority classes, which require more intensive feature extraction, while simultaneously wasting computational resources on abundant, easy-to-classify tokens from majority classes. While methods like Mixture-of-Recursions (Bae et al., 2025) have explored dynamic token processing, their design for training models from scratch is ill-suited for the PEFT context. Such methods rely on the entire network's weights adapting to a dynamic routing policy, a condition not met when operating on a frozen foundation model. Their routers are often designed to simply manage computational budgets rather than explicitly learning to discern token complexity. This necessitates a new approach specifically engineered for the constraints and opportunities of PEFT.

To address these limitations at both the domain and token levels, we propose a novel two-tiered adaptive learning framework. For **token-level adaptation**, we introduce **Recursive LoRA (RecLoRA)**, a dynamic computation module. RecLoRA integrates a lightweight, learnable router that works synergistically with a LoRA block to perform iterative feature refinement. This mechanism dynamically allocates more computation to complex tokens, directly tackling the learning challenges posed by class imbalance. For **domain-level adaptation**, we propose **Sequential Knowledge Transfer**, a new DIL paradigm that reconciles the tension between knowledge isolation and integration. It preserves domain-specific expertise by dedicating an independent set of RecLoRA modules to each task, while promoting forward transfer by initializing the modules for a new domain with the converged weights of its predecessor. This entire framework is built upon a frozen foundation model, leveraging its powerful feature space for an efficient key-query mechanism that selects the appropriate domain expert at inference time. Our main contributions are threefold:

- We propose Recursive LoRA (RecLoRA), a novel adaptive computation module for PEFT, and establish its direct utility in addressing the challenge of intra-domain class imbalance by enabling input-conditional processing depth.
- We design a new DIL paradigm, Sequential Knowledge Transfer, which effectively balances catastrophic forgetting and forward knowledge transfer by combining domain-specific modules with a cross-task weight initialization strategy.
- We demonstrate the effectiveness of our approach on challenging Diabetic Retinopathy DIL benchmarks, where our method sets a new state-of-the-art. To foster further research, we will release our code as a flexible DIL platform.

2 Related Work

Continual Learning. The primary goal of Continual Learning (CL) is to enable models to learn from a sequence of tasks while overcoming the catastrophic forgetting of previously acquired knowledge. CL methodologies are often categorized into three families. Regularization-based approaches, such as Elastic Weight Consolidation (EWC) Kirkpatrick et al. (2017) and Synaptic Intelligence (SI) Boahen (2022), introduce a penalty term to the loss function to protect weights deemed critical for past tasks. Rehearsal-based methods store a small subset of past data (an episodic memory) to be replayed during subsequent training, but this approach is often constrained by data privacy and storage limitations. Architecture-based methods, such as PackNet, dynamically expand the network by allocating new parameters for new tasks. Our work aligns with the principles of architecture-based methods, but instead of significant parameter growth, we leverage PEFT to isolate task-specific knowledge in a highly compact and efficient manner.

Parameter-Efficient Tuning for Continual Learning. With the rise of large-scale pre-trained models, applying PEFT to CL, particularly in the Domain-Incremental setting, has become a prominent research direction. These methods typically freeze the model backbone and manage a small set of trainable parameters across tasks. This has led to a dichotomy between knowledge integration and parameter isolation. Integration-based methods like Learning to Prompt (L2P) Wang et al. (2022c) and DualPrompt Wang et al. (2022b) maintain a global pool of prompts and learn a query mechanism to select a task-adaptive subset for each input, but risk task interference as the pool grows. Conversely, isolation-based methods such as S-Prompts Wang et al. (2022a) dedicate a separate, independent prompt to each task to prevent forgetting, but forgo the opportunity for forward knowledge transfer by training each prompt from scratch. More recent hybrid approaches like CODA-Prompt Smith et al. (2023), HiDe-PET Wang et al. (2025a), and Componential Prompt-Knowledge Alignment Xu et al. (2025) have introduced more complex mechanisms to decompose and align shared and task-specific knowledge. While these methods are effective, our **Sequential Knowledge Transfer** offers a simpler yet potent alternative that facilitates a curriculum-like knowledge progression through weight initialization, avoiding the need for complex architectural modifications.

Adaptive Computation A core challenge in applying deep learning to real-world problems like medical diagnostics is the inherent inefficiency of static computation graphs for imbalanced data. In tasks such as Diabetic Retinopathy (DR) grading, where sight-threatening pathologies are rare, a standard model expends most of its capacity on abundant, simple samples (e.g., healthy tissue), while under-processing the sparse but critical tokens indicating disease. This dilutes the learning signal from minority classes and hinders the learning of fine-grained, diagnostically relevant features.

Our work is inspired by the growing field of **adaptive computation**, which aims to address this by dynamically adjusting the computational graph or depth based on input complexity. Early examples include Mixture-of-Experts (MoE) models Shazeer et al. (2017), which selectively activate different sub-networks. In Transformers, this has been realized through mechanisms like adaptive attention spans or early exiting. More recently, Mixture-of-Recursions Bae et al. (2025) explored learning dynamic recursion depths for different tokens. However, these methods are typically designed for training models from scratch. Our **RecLoRA** module contributes to this line of work but is fundamentally distinct as it is engineered for the PEFT paradigm. It uses a lightweight gating mechanism to enable adaptive processing while preserving the stability of the frozen, pre-trained features, making it uniquely suited to address class imbalance within a foundation model.

3 METHOD

In this section, we first define the Domain-Incremental Learning problem and its key challenge, **Class Imbalance Shift**. We then detail our groundbreaking two-level adaptive framework for addressing class imbalance shift (intra-domain) and domain shift (inter-domain). Specifically, for class imbalance shift, we design **Recursive LoRA (RecLoRA)**, that provides adaptive, token-level computational depth to mitigate class imbalance. In parallel, we confront the domain shift problem with our pioneering **Sequential Knowledge Transfer** strategy, which leverages domain-specific modules to preserve specialized knowledge while using cross-domain weight initialization to foster positive transfer. We first formalize the DIL problem setting (§3.1), then detail the RecLoRA architecture (§3.2) and its subsequent integration into our DIL framework (§3.3).

3.1 PRELIMINARIES

Incremental Learning Scenarios. In a general Continual Learning (CL) setting, a model learns from a sequence of tasks $\mathcal{T} = \{T_1, \dots, T_N\}$. This field is primarily divided into two scenarios. In **Class-Incremental Learning (CIL)**, tasks introduce disjoint sets of classes, causing the label space to expand over time (i.e., $\mathcal{Y}_t \cap \mathcal{Y}_{t'} = \emptyset$ for $t \neq t'$). The model must learn new classes while preserving knowledge of old ones. In contrast, **Domain-Incremental Learning (DIL)**, the focus of this work, assumes a fixed and shared label space across all tasks ($\mathcal{Y}_t = \mathcal{Y}$). The primary challenge in DIL, known as domain shift, stems from the varying distribution of the input data, i.e., $P_t(X) \neq P_{t'}(X)$.

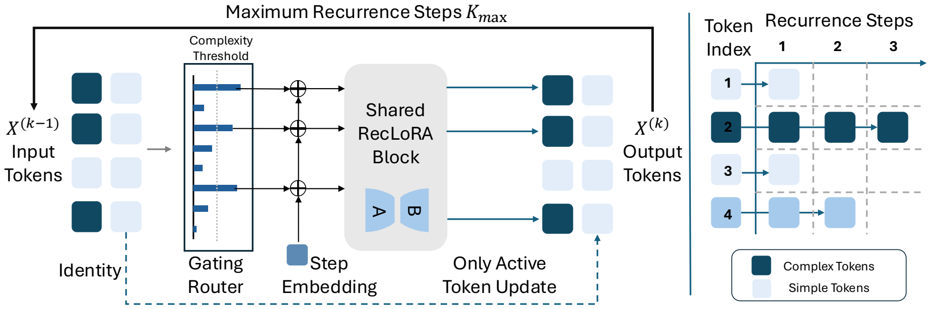

Figure 1: **The Architecture of the Recursive LoRA (RecLoRA) Module.** At each recursion step k, a Step Embedding is added to the input token representations $X^{(k-1)}$. The Gating Router $\mathcal{R}$ then computes a binary gate $g_{i,k}$ for each token, selecting a subset of ”active” tokens. To ensure coherent progression, only tokens selected at step $k-1$ are re-evaluated at step k. These active tokens are processed by the shared LoRA block. The resulting update (delta) is then gated and sparsely applied back to the representations to produce the output $X^{(k)}$. This process repeats for a maximum of $K_{\max}$ steps.

Class Imbalance Shift (CIS). In DIL setting, we are the first to identify Class Imbalance Shift, a challenge beyond domain shift. CIS occurs when the class prior probabilities change across domains, i.e., $P_t(Y) \neq P_{t'}(Y)$. Unlike CIL, which deals with an *expanding* label space, CIS involves adapting to varying class prevalences within a *fixed* label space. This dual challenge requires models to adjust to shifts in both input feature distribution $P(X)$ and class prior distribution $P(Y)$.

3.2 RecLoRA: Recursive LoRA for Adaptive Computation

To effectively tackle the challenge of class imbalance within the PEFT framework, our approach is grounded in the principle that tokens representing complex or rare samples require more intensive computational processing than those from simpler, majority classes. To achieve this, we introduce **Recursive LoRA (RecLoRA)**, a novel module that enables recursive and input-adaptive adaptation of foundation models. Unlike approaches that necessitate training a dynamic computation policy from scratch, RecLoRA is crafted as a lightweight, parameter-efficient solution. It incorporates a gating mechanism that routes tokens for recursive processing, enabling the model to dynamically allocate computational depth to simpler or complex classes. The following sections provide a detailed exploration of the architectural components of this mechanism and its impact on handling class imbalance.

Gating Router with Differentiable Decision. The Gating Router, $\mathcal{R}$, determines whether a token requires additional processing by making a binary decision (”yes” or ”no”). It is implemented as a simple linear layer that maps a token embedding $x_i \in \mathbb{R}^D$ to a scalar logit l_i: $l_i = \mathcal{R}(x_i) = W_r x_i + b_r$, where $W_r \in \mathbb{R}^{1\times D}$ and $b_r \in \mathbb{R}$ are its trainable parameters. To make the discrete routing decision differentiable for end-to-end training, we employ the Straight-Through Estimator (STE). The logits are first normalized into probabilities using a sigmoid function, modulated by a temperature τ that is annealed from a starting value τ_{start} to an ending value τ_{end} over the course of training: $p_i = \sigma(l_i/\tau)$.

During the forward pass in training, a binary gate $g_i \in \{0, 1\}$ is generated via stochastic sampling:

$$g_i = \mathbf{1}_{(\mathrm{u_i<p_i})}, \quad \text{where} \quad u_i \sim \mathcal{U}(0,1). \tag{1}$$

For inference, the gate becomes deterministic: $g_i = \mathbf{1}(p_i > 0.5)$. While the indicator function $\mathbf{1}(\cdot)$ is non-differentiable, the STE creates a shortcut for gradients. In the backward pass, the gradient is passed directly through the continuous probability p_i, allowing the router to learn from the main task loss. The gradient with respect to the logit is thus approximated as:

$$\frac{\partial \mathcal{L}}{\partial l_i} \approx \frac{\partial \mathcal{L}}{\partial g_i}\frac{\partial p_i}{\partial l_i}. \tag{2}$$

This mechanism empowers the router to learn an effective, data-driven policy for token selection directly from the downstream task signal.

Gated Recursive Update. We introduce gated update mechanism to manage the recursive updating of tokens representation within a single Transformer block. Let $X^{(k-1)}$ represent the set of token representations at the beginning of recursion step k. The router first selects a subset of these tokens for processing, indexed by the active set $\mathcal{A}_k$. To ensure a coherent refinement process, we implement a **hierarchical filtering** strategy: only those tokens that were selected at step $k-1$ are eligible for selection in step k.

The representations for the active tokens are updated by passing them through the LoRA-augmented Transformer block, $f_\ell(\cdot)$. This block is adapted using Low-Rank Adaptation (LoRA) (Hu et al., 2022), a parameter-efficient technique that injects small, trainable matrices into a frozen model. Specifically, for a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, such as the query ($W_q$) and key ($W_k$) projection matrices in the self-attention layer, its forward pass is modified as

$$h = W_0 x + BAx, \tag{3}$$

where the original weights W_0 remain frozen. The trainable update is represented by two low-rank matrices, $A \in \mathbb{R}^{r\times k}$ and $B \in \mathbb{R}^{d\times r}$, where the rank $r \ll \min(d,k)$. These matrices are the only parameters updated during training within the block. Tokens not selected by the router remain unchanged. This process continues until a termination condition is met, either no tokens are selected ($\mathcal{A}_k = \emptyset$), or the maximum number of recursions, $K_{\max}$, is reached. Upon termination, the final representations of all processed tokens are used to substitute their counterparts in the block's original input sequence. This fully updated sequence then serves as the input for the subsequent Transformer block or the classification head.

Step Embedding. A single, shared router and LoRA block would be permutation-invariant across recursion steps, limiting their ability to learn step-dependent strategies. To break this symmetry, we introduce learnable Step Embeddings. We define a set of unique embedding vectors, $\{E^k_{\text{step}}\}_{k=1}^{K_{\max}}$, where each $E^k_{\text{step}} \in \mathbb{R}^D$. At the beginning of each recursion step k, the corresponding embedding is broadcast and added to all token representations being considered:

$$X'_{k-1} = X_{k-1} + E^k_{\text{step}}. \tag{4}$$

The router and LoRA block then operate on this conditioned input, X'_{k-1}. This explicitly informs the modules of the current recursion depth, enabling them to learn step-dependent strategies, such as performing coarse-grained analysis in early steps and refining intricate features in later ones.

3.2.1 Recursive Forward Pass Algorithm

We formalize the dynamic computation within a RecLoRA-augmented block in Algorithm 1. The algorithm iteratively refines token representations, employing the hierarchical filtering strategy previously described. Based on the router's decision at each step, an `UPDATE` function sparsely modifies the token representations from the previous state to generate the input for the next iteration. This recursive process is self-contained within a single block. After the loop terminates, the final token representations are passed as a complete sequence to the next layer.

3.3 Domain Incremental Learning Framework

The RecLoRA module addresses intra-domain class imbalance by enabling adaptive computation at the token level. We now integrate this module into a broader framework to tackle the full DIL problem. While existing DIL methods focus on adapting to **Domain Shift** ($P(X)$), our work addresses the dual challenge of concurrently handling both Domain Shift and the critical, often-overlooked **Class Imbalance Shift** ($P(Y)$). This section details the training and inference pipeline of our two-level adaptive system designed for this compound challenge.

3.3.1 Training Procedure

Our training procedure sequentially adapts to new domains by training dedicated modules with a knowledge transfer mechanism.

Algorithm 1 Recursive Forward Pass within a RecLoRA-augmented Block

```
Input: Initial token representations $X^{(0)}$
Initialize active index set $\mathcal{A}_0 \leftarrow \{0, 1, \ldots, M-1\}$        ▷ $M$ is the sequence length
for $k = 1, \ldots, K_{\max}$ do
    if $\mathcal{A}_{k-1}$ is empty then break
    end if
    $X' \leftarrow X^{(k-1)} + E^k_{\text{step}}$        ▷ Condition on current recursion depth
    Let $X'_{\mathcal{A}_{k-1}} = \{x'_i \mid i \in \mathcal{A}_{k-1}\}$
    $\mathcal{A}_k \leftarrow \mathcal{R}(X'_{\mathcal{A}_{k-1}})$        ▷ Router selects new active indices
    $X^{(k)} \leftarrow$ UPDATE$(X^{(k-1)}, \mathcal{A}_k)$        ▷ Sparsely update active tokens
end for
Return: Final token representations $X^{(k)}$
```

Domain-Specific Modules. To preserve domain-specific knowledge and adapt to each domain's unique class distribution (i.e., the Class Imbalance Shift), we allocate an independent set of trainable modules θ_t for each domain $\mathcal{D}_t$. This set includes the Gating Router, LoRA matrices, Step Embeddings, and a domain-specific classification head:

$$\theta_t = \{\mathcal{R}_t, \Lambda_t, E_{\text{step},t}, h_t\}. \tag{5}$$

When training on domain $\mathcal{D}_t$, only the parameters in θ_t are optimized; the foundation model backbone and all previously trained modules $\{\theta_j\}_{j=1}^{t-1}$ remain frozen.

Cross-Domain Weight Initialization. To foster positive knowledge transfer, we employ a sequential initialization strategy. For any subsequent domain $t > 1$, its module set θ_t is initialized with the converged weights of the preceding module set θ_{t-1}:

$$\theta_t^{(\text{init})} \leftarrow \theta_{t-1}^{(\text{final})}. \tag{6}$$

This provides a strong starting point for optimization on the new domain.

Optimization Objective. For each domain $\mathcal{D}_t$, the module set θ_t is trained to minimize the Focal LossLin et al. (2017), which is effective at mitigating the effects of severe class imbalance by focusing on hard-to-classify samples.

3.3.2 Inference Procedure

At inference time, the domain identity of a given sample is unknown. We use an efficient key-query mechanism to select the appropriate domain-specific module.

Key-Query Domain Selection. We first pre-compute a ”domain key” k_t for each domain by averaging the [CLS] token features of its training data, as extracted by the frozen backbone F_{pre}. When a new sample x_{new} arrives, we compute its ”query” vector q in the same fashion. The domain $\hat{t}$ is then identified by finding the key with the highest cosine similarity to the query:

$$\hat{t} = \underset{j \in \{1, \ldots, T\}}{\arg\max} \frac{q \cdot k_j}{\|q\|_2 \|k_j\|_2}. \tag{7}$$

Once the domain expert $\hat{t}$ is selected, its corresponding module set $\theta_{\hat{t}}$ is loaded. The final prediction for x_{new} is then made by a forward pass through the backbone augmented with these expert RecLoRA modules.

4 Experiments

To validate the effectiveness of our proposed framework, RecLoRA we conduct a comprehensive set of experiments on challenging, real-world medical imaging benchmarks for Domain Incremental Learning. Our evaluation is designed to answer the following key research questions:

- Does RecLoRA outperform state-of-the-art DIL techniques in terms of both accuracy and knowledge retention?
- What is the contribution of each key design choice within RecLoRA including the DIL framework and the adaptive module?
- Does the adaptive computation mechanism behave as intended, focusing more resources on complex, diagnostically relevant samples?

4.1 Experimental Setup

Datasets. We evaluate RecLoRA on two distinct DIL benchmarks constructed from publicly available retinal imaging datasets.

- **Diabetic Retinopathy (DR) Grading:** A 5-class classification task. We construct a domain sequence from three datasets known for significant domain shifts and class imbalance: **APTOS 2019**, **Messidor-2**, and **IDRiD**.
- **Age-Related Macular Degeneration (AMD) Grading:** A multi-class classification task. The domain sequence is formed by three datasets: **ADAM**, **ARIA**, and **ODIR-5K** (AMD subset).

Baselines. We compare RecLoRA against a comprehensive suite of baselines, including: (1) **Full Fine-tuning (FT-Seq)**, which sequentially fine-tunes the entire backbone and represents a lower bound for forgetting. (2) Traditional CL methods: **EWC** (Kirkpatrick et al., 2017) and **LwF** (Li & Hoiem, 2017). (3) State-of-the-art PEFT-based DIL methods, covering both integration-based approaches (**L2P** (Wang et al., 2022c), **DualPrompt** (Wang et al., 2022b)) and isolation/hybrid approaches (**S-Prompts** (Wang et al., 2022a), **CODA-Prompt** (Smith et al., 2023), **HiDe-PET** (Wang et al., 2025a)).

Evaluation Metrics. Following standard DIL evaluation protocols, we use two primary metrics:

- **Average Accuracy (ACC):** The average accuracy across all seen tasks after the final task is learned. Higher is better.
- **Average Forgetting (AF):** The average drop in performance on a task after training, as new tasks are introduced. Lower is better. It is calculated as:

$$\text{AF} = \frac{1}{T-1}\sum_{i=1}^{T-1}(\max_{t\in[1,i]} A_{t,i} - A_{T,i})$$

where $A_{t,i}$ is the accuracy of the task i after learning task t.

Implementation Details. Our framework is implemented in PyTorch. We use the pre-trained **RETFound** Zhou et al. (2023) as our frozen backbone. For all PEFT-based methods, we use LoRA with a rank of $r = 8$. All models are trained using the AdamW optimizer with a learning rate of 2×10^{-4} and a batch size of 32. We set the maximum recursion depth $K_{\max} = 3$ for RecLoRA.

4.2 Main Results

The main results on the DR and AMD benchmarks are presented in Table 1, respectively. The findings clearly demonstrate that RecLoRA establishes a new state-of-the-art in both challenging medical DIL scenarios.

On the more complex DR benchmark (Table 1), which is characterized by significant class imbalance, RecLoRA achieves an average accuracy of 68.50%, surpassing the next best method by a large margin of 5.15%. Crucially, it reduces average forgetting to a near-zero 0.19%. This powerful combination of high accuracy and low forgetting validates our core design philosophy. The performance gain over pure-isolation methods like S-Prompts highlights the substantial benefit of our **Sequential Knowledge Transfer** strategy, which promotes positive forward transfer. Simultaneously, the drastic reduction in forgetting compared to integration-based methods like L2P and DualPrompt confirms the effectiveness of using domain-specific modules to prevent catastrophic forgetting.

This effectiveness is further underscored on the AMD benchmark (Table 1), where RecLoRA not only achieves the highest accuracy (94.93%) but does so with a perfect forgetting score of 0.02%.

Table 1: Overall performance comparison on the **DR Grading** and **AMD Grading** benchmarks. RecLoRA demonstrates superior performance in average accuracy and knowledge retention across both challenging medical imaging datasets. Best results are in **bold**.

METHODS	DR GRADING		AMD GRADING	
	AVG. ACC (%) ↑	FORGETTING (%) ↓	AVG. ACC (%) ↑	FORGETTING (%) ↓
FT-SEQ	61.89	11.61	85.22	0.20
EWC	61.95	9.82	91.73	1.10
LWF	62.32	6.97	54.24	45.53
L2P	62.34	6.11	91.17	3.80
DUALPROMPT	63.35	4.87	80.52	0.20
S-PROMPTS	49.39	15.78	82.56	0.20
CODA-PROMPT	62.18	6.91	76.45	2.30
HIDE-PROMPT	55.60	9.52	71.69	0.11
HIDE-LORA	58.85	1.99	75.35	0.63
RECLORA (OURS)	**68.50**	**0.19**	**94.93**	**0.02**
UPPERBOUND	69.60	-	94.53	-

This result demonstrates the framework's robustness and its capability to fully preserve domain-specific expertise while adapting to new data distributions.

4.3 ABLATION STUDY

To thoroughly investigate the sources of RecLoRA's performance gains and validate its key design choices, we conduct a series of detailed ablation studies on the more challenging DR benchmark. We first evaluate the effectiveness of the overall DIL framework, then dissect the core components of the RecLoRAmodule.

4.3.1 IMPACT OF OUR DIL FRAMEWORK

We first validate the **Sequential Knowledge Transfer** strategy, which is critical to RecLoRA. As shown in Table 2, we compare the full model against a variant that removes cross-domain weight initialization (i.e., modules for each new domain are randomly initialized). The results show that without knowledge transfer, the model's average accuracy drops significantly by 9.04%, while forgetting increases by nearly nine-fold. This provides strong evidence that leveraging knowledge from prior tasks as a starting point for new ones is crucial for improving performance and accelerating convergence.

4.3.2 COMPONENT-WISE ANALYSIS OF THE ADAPTIVE MODULE

Next, we delve into the core adaptive module of RecLoRAto quantify the contribution of each internal component. As presented in Table 2: (1) Removing the entire recursive mechanism (**w/o Recursion**) by setting $K_{\max} = 1$ leads to the most significant drop in performance, directly confirming the fundamental benefit of iterative refinement for complex features. (2) Replacing the learned gating router with a random policy (**w/o Learned Router**) also markedly degrades performance, indicating that an intelligent routing strategy is key to the effectiveness of the recursion, not just the recursive operation itself. (3) Removing the step embedding (**w/o Step Embedding**) causes a slight performance drop, which validates its role in helping the model learn a hierarchical, step-aware refinement strategy.

4.4 ANALYSIS OF DESIGN CHOICES

Finally, we conduct analyses to validate two key design choices within our framework: the mechanism for inference-time domain identification and the sensitivity to the maximum recursion depth hyperparameter.

Table 2: Comprehensive ablation studies on the DR benchmark. We analyze the contributions of both the high-level DIL framework and the core components of our adaptive module. The performance degradation across all variants validates our key design choices.

Category	**Ablation Variant**	**ACC (%)** $\uparrow$	**AF (%)** $\downarrow$
Full Model (Ours)		**68.50**	**0.19**
DIL Framework	w/o Sequential Knowledge Transfer	59.46	1.66
Adaptive Module	w/o Recursion ($K_{\max} = 1$)	61.14	2.95
	w/o Learned Router (Random)	61.01	1.01
	w/o Step Embedding	66.62	0.57

Impact of Maximum Recursion Depth. We analyze the sensitivity of RecLoRAto the core hyperparameter, $K_{\max}$, which controls the maximum depth of token-level processing. As shown in Figure 2, performance on the DR benchmark steadily improves as $K_{\max}$ increases from 1 to 3, validating the benefit of deeper, iterative refinement. The performance saturates at $K_{\max} = 3$ and slightly declines at $K_{\max} = 4$, likely due to overfitting on the training data. This result justifies our choice of $K_{\max} = 3$ as an effective trade-off between representational power and overfitting risk.

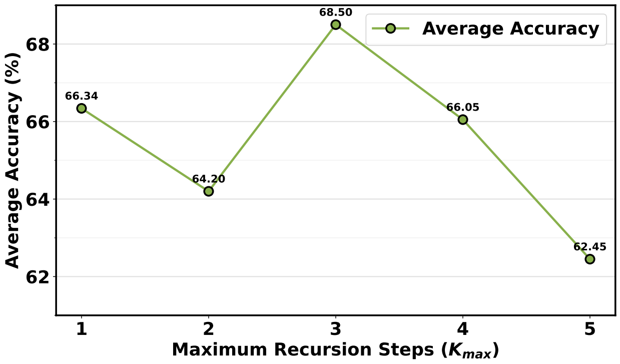

Figure 2: **Impact of maximum recursion depth** ($K_{\max}$). Performance peaks at $K_{\max} = 3$ on the DR benchmark, justifying our default setting.

5 Conclusion

In this work, we addressed the dual challenges of continual domain adaptation and intra-domain class imbalance for large-scale medical foundation models. We proposed a novel two-tiered adaptive framework featuring **Recursive LoRA (RecLoRA)**, which performs token-level adaptive computation to iteratively refine features of rare and complex samples. At the domain level, our **Sequential Knowledge Transfer** paradigm utilizes domain-specific modules to prevent catastrophic forgetting while employing cross-domain weight initialization to promote forward knowledge transfer. Our experiments on challenging Diabetic Retinopathy benchmarks demonstrate that this integrated approach sets a new state-of-the-art, paving the way for more robust, efficient, and continually adapting AI systems suitable for real-world clinical deployment.

Ethics Statement

This work adheres to the ICLR Code of Ethics. In this study, no human subjects or animal experimentation was involved. All datasets used, including ADAM, ODIR, ARIA, APTOS, MESSIDOR2 and IDRID, were sourced in compliance with relevant usage guidelines, ensuring no violation of privacy. We have taken care to avoid any biases or discriminatory outcomes in our research process. No personally identifiable information was used, and no experiments were conducted that could raise privacy or security concerns. We are committed to maintaining transparency and integrity throughout the research process.

Reproducibility Statement

We have made every effort to ensure that the results presented in this paper are reproducible. All code has been made publicly available in an anonymous repository to facilitate replication and verification. The experimental setup, including training steps, model configurations, and hardware details, is described in detail in the paper. We have also provided a full description of our method to assist others in reproducing our experiments. Additionally, all the datasets are publicly available, ensuring consistent and reproducible evaluation results. We believe these measures will enable other researchers to reproduce our work and further advance the field.

References

Sangmin Bae, Yujin Kim, Reza Bayat, Sungnyun Kim, Jiyoun Ha, Tal Schuster, Adam Fisch, Hrayr Harutyunyan, Ziwei Ji, Aaron Courville, et al. Mixture-of-recursions: Learning dynamic recursive depths for adaptive token-level computation. *arXiv preprint arXiv:2507.10524*, 2025.

Kwabena Boahen. Dendrocentric learning for synthetic intelligence. *Nature*, 612(7938):43–50, 2022.

Xiaoyang Chen, Hao Zheng, Yifang Xie, Yuncong Ma, and Tengfei Li. A classifier-free incremental learning framework for scalable medical image segmentation. *arXiv preprint arXiv:2405.16328*, 2024.

Yuting He, Fuxiang Huang, Xinrui Jiang, Yuxiang Nie, Minghao Wang, Jiguang Wang, and Hao Chen. Foundation model for advancing healthcare: Challenges, opportunities and future directions. *IEEE Reviews in Biomedical Engineering*, 2024.

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

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.

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

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Jianing Qiu, Jian Wu, Hao Wei, Peilun Shi, Minqing Zhang, Yunyun Sun, Lin Li, Hanruo Liu, Hongyi Liu, Simeng Hou, et al. Development and validation of a multimodal multitask vision foundation model for generalist ophthalmic artificial intelligence. *NEJM AI*, 1(12):AIoa2300221, 2024.

Dilermando Queiroz, Anderson Carlos, André Anjos, and Lilian Berton. Fair foundation models for medical image analysis: Challenges and perspectives. *arXiv preprint arXiv:2502.16841*, 2025.

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.

Bin Sheng, Pearse A Keane, Yih-Chung Tham, and Tien Yin Wong. Synthetic data boosts medical foundation models. *Nature Biomedical Engineering*, 9(4):443–444, 2025.

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*, pp. 11909–11919, 2023.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature biomedical engineering*, 6(12):1399–1406, 2022.

Liyuan Wang, Jingyi Xie, Xingxing Zhang, Hang Su, and Jun Zhu. Hide-pet: continual learning via hierarchical decomposition of parameter-efficient tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Qiang Wang, Xiang Song, Yuhang He, Jizhou Han, Chenhao Ding, Xinyuan Gao, and Yihong Gong. Boosting domain incremental learning: Selecting the optimal parameters is all you need. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4839–4849, 2025b.

Sheng Wang, Zihao Zhao, Xi Ouyang, Tianming Liu, Qian Wang, and Dinggang Shen. Interactive computer-aided diagnosis on medical image using large language models. *Communications Engineering*, 3(1):133, 2024.

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, 2022a.

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*, pp. 631–648. Springer, 2022b.

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*, pp. 139–149, 2022c.

Kunlun Xu, Xu Zou, Gang Hua, and Jiahuan Zhou. Componential prompt-knowledge alignment for domain incremental learning. *arXiv preprint arXiv:2505.04575*, 2025.

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. Radbert: adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258, 2022.

Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023.