Representation-Aware Modularity: Efficient Cross-Task Generalization for LLMs

Zheng Gong[1] , **Ying Sun**[2*] , **Chao Wang**[3] , **Xiaohui Huo**[1] ,
Ping Li[4] , **Yi Zheng**[4] and **Zhefeng Wang**[4]
[1]Thrust of Artificial Intelligence, Hong Kong University of Science and Technology (Guangzhou)
[2]The 63rd Research Institute, National University of Defense Technology
[3]School of Artificial Intelligence and Data Science, University of Science and Technology of China
[4]Huawei Cloud
{zgong768, xhuo765}@connect.hkust-gz.edu.cn, sunying@nudt.edu.cn, wangchaoai@ustc.edu.cn,
{liping61, zhengyi29, wangzhefeng}@huawei.com

Abstract

Cross-task generalization (CTG) enables large language models (LLMs) to handle unseen tasks proficiently, enhancing their adaptability in real-world scenarios. However, existing methods relying on per-token dynamic routing to multiple trained LoRA adapters face high computational and GPU memory costs. Recent Representation Fine-Tuning (ReFT) enhances efficiency for single-task adaptation by editing only prefix and suffix token representations. However, the semantic ambiguity of tokens and absence of a self-guided mechanism for parameter selection in unseen tasks limits their application to CTG. To this end, we propose RaMod, a Representation-Aware Modularity framework to extend the ReFT paradigm to CTG through two novel components: (i) Dual-Modular Representation & Parameter Fine-tuning, which manipulates only a strategically chosen subset of hidden representations with modular interventions to guide the model toward solving unseen tasks; and (ii) Asynchronous Orchestrator, which proactively allocates and releases GPU memory for selected interventions, thereby minimizing storage overhead. Extensive experiments demonstrate that RaMod not only achieves superior CTG performance but also substantially reduces the overhead of the latest CTG baseline, achieving 83%, 100%, and 79% reduction in its additional prefill time, generation delays, and memory consumption relative to original LLMs.

1 Introduction

The growing proliferation and accessibility of open-source large language models (LLMs) [Zhang *et al.*, 2025; Gong and Sun, 2024; Xin *et al.*, 2025] has necessitated the development of specialized variants adapted to particular downstream tasks, domain-specific applications, and diverse user requirements [Li and Liang, 2021; Liu *et al.*, 2022b]. Parameter-efficient fine-tuning (PeFT) methods like Low-Rank Adaptation [Hu *et al.*, 2022] have become the standard for adapting LLMs to known downstream tasks. However, a critical limitation remains: real-world scenarios often require LLMs to handle unseen tasks where no task-specific adapter exists. This necessitates a capability called *cross-task generalization* (CTG) [Ye *et al.*, 2021; Mishra *et al.*, 2022][1].

Previous CTG approaches [Xu *et al.*, 2025; Lin *et al.*, 2025; Huang *et al.*, 2024; Belofsky, 2023; Ostapenko *et al.*, 2024] employ a Mixture-of-Experts (MoE) routing mechanism, which leverages a trainable gating mechanism to combine LoRA modules pre-trained on diverse upstream tasks as experts, as illustrated in Figure 1. However, this process presents two significant efficiency issues. (i) **This process requires *per-token* routing in the input to multiple pre-trained LoRA modules**, resulting in time and space complexities that scale linearly with the total number of tokens in the input. This can become particularly challenging in scenarios involving long input sequences and long-text generation. (ii) To meet the demands of rapid computation, all pre-trained LoRA modules must reside in GPU HBM (High-Bandwidth Memory) to maintain inference efficiency, while the pursuit of broader LLM capabilities inevitably increases the number of required upstream tasks, **leading to a heavier memory burden and consequently degraded inference efficiency**.

The inefficiency inherent in per-token adaptation motivates a compelling alternative: if comparable accuracy can be achieved by fine-tuning only a subset of tokens, significant efficiency gains could be realized. Recently proposed Representation Fine-Tuning (ReFT) methods [Wu *et al.*, 2024b; Geiger *et al.*, 2021] offer a promising direction by editing the representation space of only prefix and suffix tokens in the input prompt, thereby improving computational efficiency. However, existing ReFT methods focus on the single-task adaptation, while exhibit two major limitations when applied to CTG: (1) its static selection of prefix and suffix tokens is inflexible, potentially overlooking critical, context-dependent cues in the middle of the sequence that are vital for coherent generation. (2) ReFT assumes a known, single task, thus lacking any internal routing mechanism for unseen tasks. This inherently limits its scalability and necessitates external guidance for each new scenario, contravening the goal of CTG.

*Corresponding Author.

[1]We focus on *zero-shot* cross-task generalization, *i.e.*, without any exemplar data from the unseen tasks.

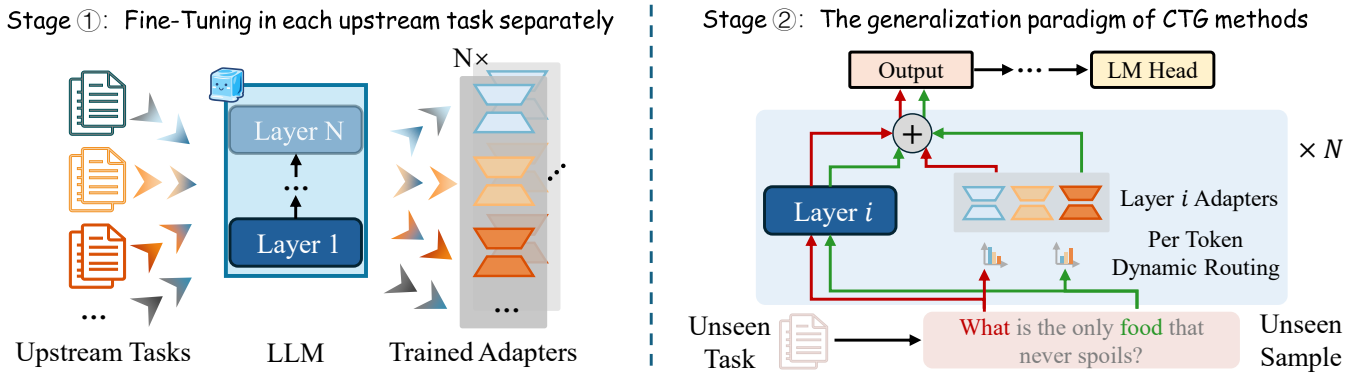

Figure 1: Existing CTG approaches typically begin by fine-tuning LoRA as task-specific adapters for each upstream task. When generalizing to unseen tasks, these methods dynamically compute routing weights for each token across all LoRA adapters, then aggregate the outputs through weighted summation to obtain updated token representations at each layer.

To this end, we propose a Representation-Aware Modularity framework, called RaMod, which extends the ReFT paradigm to CTG while addressing these key limitations. Specifically, RaMod introduces a lightweight token selection mechanism that identifies semantically critical tokens in the input during intervention, instead of relying on fixed positional edits. Moreover, RaMod incorporates a training-free routing strategy termed FluidRoute derived from the Task-Projection Dominance theorem and employs rank-1 approximation, enabling the model to adaptively select modular interventions and address unseen tasks. Furthermore, we develop an Asynchronous Orchestrator to pre-allocate GPU memory to interventions in the subsequent layer that are most likely to be required and release the memory promptly after use, thereby reducing HBM usage. Our experiments prove that RaMod not only achieves superior CTG performance but also drastically trims the overhead of the latest CTG baseline, reducing its excess prefill time, generation delays and HBM usage over the original LLM by 83%, 100%, 79%, respectively.

The principal contributions of this work are summarized as follows:

- To the best of our knowledge, our work is the first to investigate the efficiency of cross-task generalization in LLMs and proposes an effective framework RaMod to address it.
- We introduce (i) a pivot token selection strategy to enhance ReFT for upstream tasks, (ii) Dual-Modular Representation & Parameter Generalization and (iii) Asynchronous Orchestrator to significantly reduce the computational overhead and HBM usage in the process of cross-task generalization.
- Our extensive experiments demonstrate the superior efficiency as well as outstanding generalization performance of RaMod framework[2].

2 Preliminary

Problem Statement. Assume we have N distinct *upstream* tasks $\mathbb{T} = \{\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_N\}$ available for LLM fine-tuning, where each task $\mathcal{T}_i$ is associated with a dataset containing a set of samples $\mathcal{D}_i = \{(x_1, y_1), ..., (x_n, y_n)\}$. The union of the training sets comprise our multi-task dataset $\mathcal{D} = \{D_i\}_{i=1}^{N}$. For an unseen task or sample, *cross-task generalization* aims at adapting the model to accommodate the new task $\mathcal{T}'$ or sample $s' \in \mathcal{T}'$.

LoRA *v.s.* ReFT. Parameter-efficient Fine-Tuning (PeFT) reduces the number of trainable parameters for fine-tuning in downstream tasks. As a common choice, LoRA achieves competitive trade-offs between parameter efficiency and model performance [Hu *et al.*, 2022]. Given the input hidden state $\boldsymbol{x} \in \mathbb{R}^{d_1 \times 1}$, LoRA modifies the output $\boldsymbol{o} \in \mathbb{R}^{d_2 \times 1}$ of a linear layer as follows:

$$\boldsymbol{o} = \boldsymbol{W}\boldsymbol{x} + \Delta\boldsymbol{W}\boldsymbol{x} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{A}\boldsymbol{x}, \qquad \text{(LoRA)}$$

where $\boldsymbol{W}$ are the (frozen) weights of the linear layer, and $\boldsymbol{A} \in \mathbb{R}^{r \times d_1}, \boldsymbol{B} \in \mathbb{R}^{d_2 \times r}$ are two learnable low-rank parameters. LoRA achieves parameter efficiency since the rank $r \ll \{d_1, d_2\}$. Representation Fine-Tuning (ReFT) [Wu *et al.*, 2024b] learns task-specific interventions on hidden representations[3]. As an instance of ReFT, DiReFT can be thought of as LoRA applied directly to hidden representation $\boldsymbol{h} = \boldsymbol{W}\boldsymbol{x} \in \mathbb{R}^{d_2 \times 1}$ at certain position set $\mathcal{P}$ in a prompt:

$$\boldsymbol{o} = \begin{cases} \boldsymbol{h} + \boldsymbol{B}(\boldsymbol{A}\boldsymbol{h}), & x \in \mathcal{P}, \\ \boldsymbol{h}, & x \notin \mathcal{P}, \end{cases} \qquad \text{(DiReFT)}$$

where $\boldsymbol{A} \in \mathbb{R}^{r \times d_2}, \boldsymbol{B} \in \mathbb{R}^{d_2 \times r}$. In the implementation of DiReFT, the position set $\mathcal{P} = \mathcal{P}_{\text{Fix}}$ consists of a fixed number of p prefix $\{1, ..., p\}$ and s suffix $\{n - s + 1, ..., n\}$ positions in the prompt.

3 RaMod: Efficient Cross-Task Generalization via Modular ReFT

In this section, we illustrate RaMod, our cross-task generalization (CTG) framework extensively. RaMod comprises two

[2]Our code is available at github.com/KellyGong/RaMod.

[3]For clarity, we will refer to the LoRA modules as adapters and the changed parameters in DiReFT as interventions. Together, these two will be collectively described as experts.

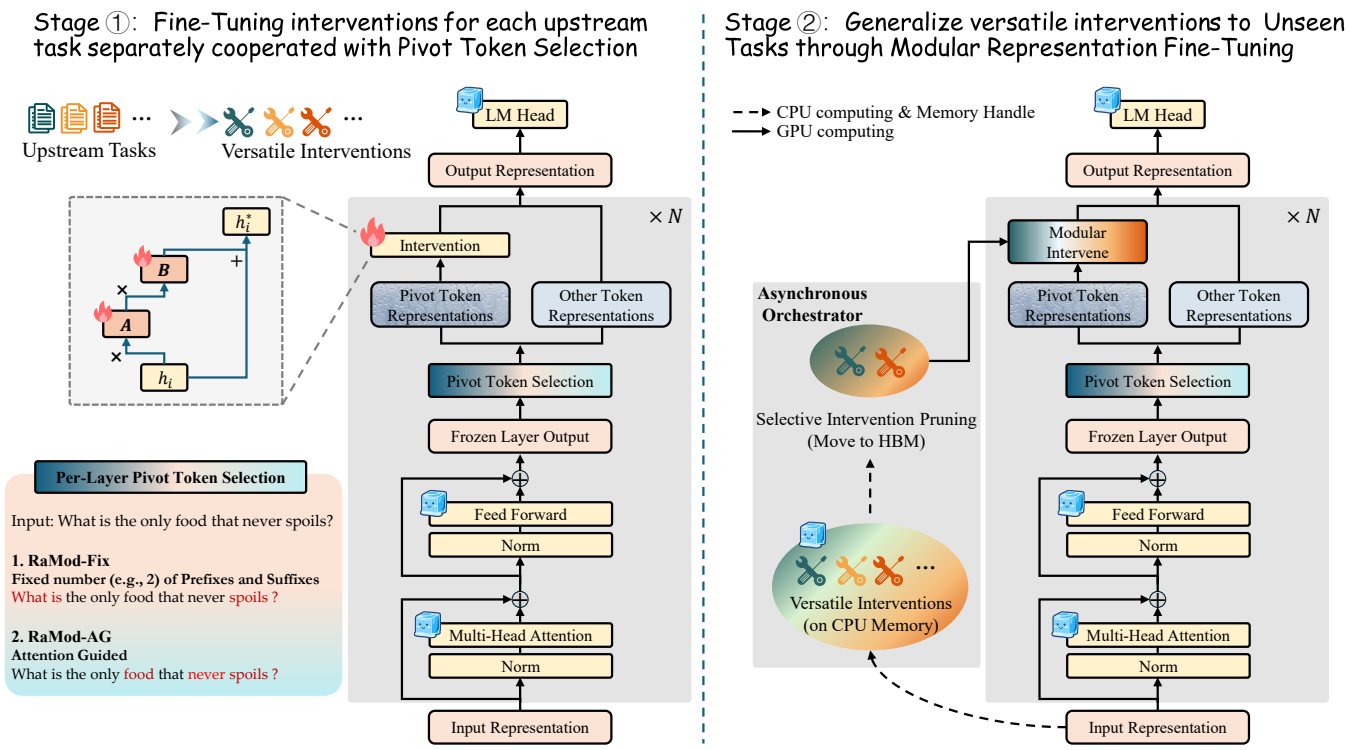

Figure 2: Our CTG framework RaMod consists of two stages. (i) In the first stage, we perform fine-tuning interventions separately for each upstream task. The intervention is applied at each layer to adapt the representations of pivot tokens. (ii) In the second stage, we generalize LLM with well-trained intervenes to unseen tasks. We propose a zero-shot intervention method for modular representation fine-tuning, enabling the model to adapt to unseen tasks effectively. Furthermore, we design an asynchronous orchestrator to dynamically allocate HBM as needed and proactively prune unnecessary interventions, thereby enhancing memory efficiency.

stage, (i) pivot-representation fine-tuning in each upstream task and (ii) dual-modular parameter & representation generalization to unseen tasks, detailed in Sec. 3.1 and Sec. 3.2. At last, we introduce an asynchronous orchestrator for memory efficiency in Sec. 3.3. The overall architecture of our framework is shown in Fig. 2.

3.1 Stage 1: Pivot-Representation Fine-tuning

LoRA's efficiency in single-task adaptation stems from the post-training merger of its parameters ($\boldsymbol{W} + \boldsymbol{BA}$). However, when generalizing to cross-task scenarios, LoRA-based approaches employ MoE routing over multiple upstream adapters, and the required per-token adaptive computation precludes this merger, leading to inference inefficiency [Xu *et al.*, 2025; Wu *et al.*, 2024a; Ostapenko *et al.*, 2024; Huang *et al.*, 2024]. To mitigate this issue, we apply Representation Fine-tuning (ReFT, [Wu *et al.*, 2024b]) to CTG.

A key advantage of ReFT over LoRA is its ability to achieve comparable performance by intervening on the representations of only a small subset of tokens. However, the original ReFT framework is constrained in its expressive power as it rigidly selects a fixed number of prefix and suffix tokens for this intervention. To enhance ReFT, we propose a pivot token selection strategy alternative.

Attention-Guided (AG) Pivot Token Selection. The attention score a token receives can be interpreted as the aggregated focus directed towards its representation by all other token representations in the sequence. A higher attention score implies that a token's representation is more influential in the context of the entire sequence [Ghaeini *et al.*, 2018; Luo *et al.*, 2018; Huang *et al.*, 2025a]. Consequently, intervening on the representations of such high-attention tokens is likely to have a more significant impact on the representations of other tokens. Based on this premise, we introduce the following metric to identify pivot tokens in the l-th layer:

$$\mathcal{P}_{\text{AG}}^{(l)} = \underset{j\in\{1,\dots,|\mathcal{S}|\}}{\arg\,\text{top-}k_{\text{AG}}}\Big(\sum_{i=j+1}^{|\mathcal{S}|} \mathcal{A}_{ij}^{(l)}/(|\mathcal{S}|-j)\Big), \qquad (1)$$

where $\mathcal{A}_{ij}^{(l)}$ denotes the attention score token t_i pays to token t_j in the l-th layer, and $|\mathcal{S}|$ represents the length of the sequence $\mathcal{S}$.

Optimize task-specific interventions. For each upstream task $\mathcal{T}_i \in \mathbb{T}$, we intervene on the per-layer pivot token representations at positions $\mathcal{P} = \mathcal{P}_{\text{AG}}^{(l)}$, which is an instance of our framework we refer to as **RaMod-AG**. Another instance of our framework, **RaMod-Fix**, performs interventions at fixed positions $\mathcal{P} = \mathcal{P}_{\text{fix}}$. The intervention param-

eters $\{\boldsymbol{A}_{\mathcal{T}_i}^{(l)}, \boldsymbol{B}_{\mathcal{T}_i}^{(l)}\}_{\mathcal{T}_i \in \mathbb{T}}^{l \in \mathbb{L}}$ for each upstream task are optimized by maximizing the likelihood of the target sequence following a language modeling objective.

3.2 Stage 2: Dual-Modular Parameter & Representation Generalization

After fine-tuning pivot representations for each upstream task, a critical challenge emerges: *how can these versatile, modular interventions be effectively applied to unseen tasks?* Although prior work has leveraged exemplar data [Huang *et al.*, 2024] or task instructions [Charakorn *et al.*, 2025] to fine-tune routing weights, the zero-shot setting, where no information about the unseen task is available, is still a nascent area of research. Existing approaches in this domain are hindered by inefficiencies, requiring the per-token dynamic routing to all trained experts. A comprehensive overview of these methods can be found in Sec. 5.

Based on the modular and versatile interventions trained in the first stage, we propose a dual-modular parameter & representation generalization method, as shown in the right of Fig. 2. First, we employ the same pivot token selection approach as in Stage 1 to identify tokens with significant influence at the l-th layer of the LLM. The representation for each of these pivot tokens is denoted as $\boldsymbol{h}_*$, where we drop the dependency on l for legibility.

FluidRoute: Training-Free & Efficient Routing. In the CTG setting, where no task-specific information is available for unseen tasks, a training-free routing method is required. This method must efficiently and automatically select the most appropriate intervention based on intermediate hidden states to adjust the pivot token representations $\boldsymbol{h}_*$, thus minimizing the impact on LLM inference speed. We first propose to quantify the relevance of each task to the given token using the l_2-norm of the task-parameterized transformation of $\boldsymbol{h}$:

$$\mathcal{R}_{\mathcal{T}_i}(\boldsymbol{h}) = \|\boldsymbol{B}_{\mathcal{T}_i}\boldsymbol{A}_{\mathcal{T}_i}\boldsymbol{h}\|_2. \tag{2}$$

We provide the theoretical guarantee following:

Theorem 1 (Task-Projection Dominance)**.** *Let $P_{\mathcal{T}_i}$ denote the hidden representation distributions in task $\mathcal{T}_i$, the projection magnitude in expectation is maximized for the corresponding task $\mathcal{T}_i$. That is*

$$\mathbb{E}_{\boldsymbol{h}\sim P_{\mathcal{T}_i}}\left(\mathcal{R}_{\mathcal{T}_i}(\boldsymbol{h})\right) > \mathbb{E}_{\boldsymbol{h}\sim P_{\mathcal{T}_i}}\left(\mathcal{R}_{\mathcal{T}_j}(\boldsymbol{h})\right). \tag{3}$$

Proof. The proof is detailed in Section A. □

However, since computing $\mathcal{R}_{\mathcal{T}_i}(\boldsymbol{h})$ requires passing $\boldsymbol{h}$ through the parameters of each task to obtain the final result entailing substantial computational overhead, we propose an efficient alternative, which we term FluidRoute. FluidRoute begins by applying singular value decomposition (SVD) to the low-rank parameterization of each expert's intervention $\{\boldsymbol{A}_{\mathcal{T}_i}, \boldsymbol{B}_{\mathcal{T}_i}\}$:

$$\boldsymbol{U}_{\mathcal{T}_i}, \boldsymbol{\Sigma}_{\mathcal{T}_i}, \boldsymbol{V}_{\mathcal{T}_i} = \mathrm{SVD}(\boldsymbol{B}_{\mathcal{T}_i}\boldsymbol{A}_{\mathcal{T}_i}). \tag{4}$$

From this decomposition, we extract the principal singular triplet for each expert's intervention: the first right singular vector $\boldsymbol{v}_{1,\mathcal{T}_i} \in \mathbb{R}^{d_2}$, the first left singular vector $\boldsymbol{u}_{1,\mathcal{T}_i} \in \mathbb{R}^{d_2}$ and the largest singular value $\lambda_{1,\mathcal{T}_i}$. These components possess distinct semantic interpretations: $\boldsymbol{v}_{1,\mathcal{T}_i}$ captures the principal input direction to which the expert is most sensitive, $\boldsymbol{u}_{1,\mathcal{T}_i}$ represents the principal output direction of the expert's transformation, and $\lambda_{1,\mathcal{T}_i}$ quantifies the magnitude of this transformation.

Cateogry	**MoE Routing & LoRA Adaptation**	**RaMod** (Ours)
Fine-tuning token	All	A small portion
Modularity	θ	θ & $\boldsymbol{h}$
Affected stage	Prefill & Generation	Prefill
Expert loading	Synchronous	Asynchronous

Table 1: The comparison of RaMod with recent CTG methods via MoE routing and LoRA adaptation. θ denotes model parameters, and $\boldsymbol{h}$ denotes hidden representations.

The product $\lambda_{1,\mathcal{T}_i}\boldsymbol{u}_{1,\mathcal{T}_i}\boldsymbol{v}_{1,\mathcal{T}_i}^T$ constitutes the best rank-1 approximation of the matrix product $\boldsymbol{B}_{\mathcal{T}_i}\boldsymbol{A}_{\mathcal{T}_i}$. According to the Eckart-Young-Mirsky theorem [Golub *et al.*, 1987], this approximation minimizes the Frobenius norm of the error:

$$\min_{\mathrm{rank}(\hat{\boldsymbol{M}})=1} \left\|\boldsymbol{B}_{\mathcal{T}_i}\boldsymbol{A}_{\mathcal{T}_i} - \hat{\boldsymbol{M}}\right\|_F^2 = \sum_{i=2}^{r} \lambda_{i,\mathcal{T}_i}^2, \tag{5}$$

where the minimum is achieved when $\hat{\boldsymbol{M}} = \lambda_{1,\mathcal{T}_i}\boldsymbol{u}_{1,\mathcal{T}_i}\boldsymbol{v}_{1,\mathcal{T}_i}^T$. To promise the precision and efficiency, we formulate the routing logit of FluidRoute for expert i given input token representation $\boldsymbol{h} \in \mathbb{R}^{d_2 \times 1}$ via $\boldsymbol{u}_{1,\mathcal{T}_i}$ and $\boldsymbol{v}_{1,\mathcal{T}_i}$:

$$\begin{aligned}\hat{\mathcal{R}}_{\mathcal{T}_i}(\boldsymbol{h}) &= \left\|\lambda_{1,\mathcal{T}_i}\boldsymbol{u}_{1,\mathcal{T}_i}\boldsymbol{v}_{1,\mathcal{T}_i}^T\boldsymbol{h}\right\|_2 = |\lambda_{1,\mathcal{T}_i}| \cdot \left\|\boldsymbol{u}_{1,\mathcal{T}_i}\boldsymbol{v}_{1,\mathcal{T}_i}^T\boldsymbol{h}\right\|_2 \\ &\overset{(i)}{=} |\lambda_{1,\mathcal{T}_i}| \cdot \left|\boldsymbol{v}_{1,\mathcal{T}_i}^T\boldsymbol{h}\right| \cdot \|\boldsymbol{u}_{1,\mathcal{T}_i}\|_2 \overset{(ii)}{=} |\lambda_{1,\mathcal{T}_i}| \cdot \left|\boldsymbol{v}_{1,\mathcal{T}_i}^T\boldsymbol{h}\right|,\end{aligned} \tag{6}$$

where step (i) is due to the $\boldsymbol{v}_{1,\mathcal{T}_i}^T\boldsymbol{h}$ being a scalar, and step (ii) is derived from each singular vector $\|\boldsymbol{u}\|_F = 1$. The routing logit of FluidRoute effectively estimates the alignment of the hidden state $\boldsymbol{h}_*$ with the directions specific to task $\mathcal{T}_i$.

Through Theorem 1, we can directly estimate the intervention of the most relevant task from the hidden representations during the inference process of an LLM. We employ a top-k routing strategy from the routing logits of all intervenes $\hat{\mathcal{R}}(\boldsymbol{h}_*) = \{\hat{\mathcal{R}}_{\mathcal{T}_i}(\boldsymbol{h}_*)\}_{\mathcal{T}_i \in \mathbb{T}}$ and apply a softmax composition of target intervention parameters $\{\boldsymbol{B}_*, \boldsymbol{A}_*\}$ for each pivot token representation $\boldsymbol{h}_*$:

$$w_{\mathcal{T}_i}(\boldsymbol{h}_*) = \begin{cases} -\infty, & \hat{\mathcal{R}}_{\mathcal{T}_i}(\boldsymbol{h}_*) \notin \arg\text{top-}k_\theta\, \hat{\mathcal{R}}(\boldsymbol{h}_*), \\ \hat{\mathcal{R}}_{\mathcal{T}_i}(\boldsymbol{h}_*), & \text{otherwise}, \end{cases} \tag{7}$$

$$\boldsymbol{A}_* = \sum\nolimits_{\mathcal{T}_i \in \mathbb{T}} \mathrm{Softmax}(w_{\mathcal{T}_i}(\boldsymbol{h}_*))\boldsymbol{A}_{\mathcal{T}_i}. \tag{8}$$

The composition of $\boldsymbol{B}_*$ and $\boldsymbol{b}_*$ is defined similarly and omitted for brevity. Building on modular representation fine-tuning and modular intervention parameters, our method improves the efficiency of generalizing to unseen tasks.

Prefill-only Intervention. A key advantage is that our approach **only** requires modifications on the token representations during the *prefill* stage, *i.e.,* processing the input prompt. This avoids any latency impact on the *generation* process, *i.e.,*

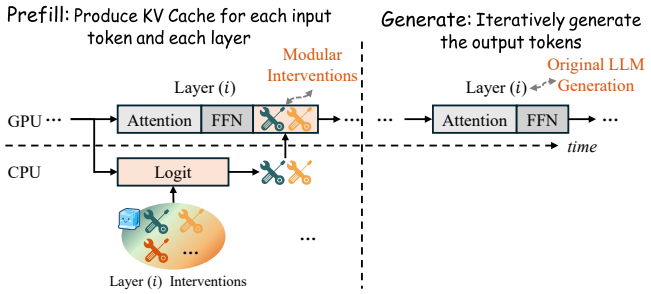

Figure 3: RaMod only intervenes during the prefill phase and leverages our Asynchronous Orchestrator to significantly reduce the memory footprint of modular parameters. The generation process utilizes the original LLM procedure, ensuring efficient generation.

producing one token at a time. This is a significant efficiency gain over prior work that required routing for every single token in both prefill and generation phase. As demonstrated by the baseline `Arrow` in Table 2, the routing adaptation applied to each token during the generation process leads to a significant increase in latency, causing a per-token generation delay that is approximately 62 times greater (1.482s / 0.024s $\approx$ 62).

3.3 Asynchronous Orchestrator

Previous works on CTG generally require all experts to be stored in High Bandwidth Memory (HBM). This allows for the immediate determination of requisite experts based on input representations at each layer during inference, enabling subsequent computations. However, as the number of tasks increases, the linear relationship between experts and tasks leads to a significant HBM storage overhead. This, in turn, limits the maximum achievable batch size and adversely affects LLM inference efficiency metrics such as throughput. To mitigate this issue, some methods employ clustering to reduce the number of experts and facilitate parameter sharing [Gabrielsson *et al.*, 2024; Ostapenko *et al.*, 2024].

We propose an orthogonal approach to this problem, called the Asynchronous Orchestrator. The core principle of this approach is to store the parameters for the vast majority of experts in low-cost, high-storage CPU memory. By anticipating which experts will likely be required for upcoming inputs, it proactively transfers the relevant parameters into HBM just before they are needed for computation. This strategy significantly mitigates the storage burden that expert parameters place on the HBM. This process is depicted in Figure 3.

Based on the common observation that hidden representations between adjacent layers exhibit a high degree of similarity [Lee *et al.*, 2024; Liu *et al.*, 2023b], we hypothesize that the experts required by these layers are also closely related. Therefore, we leverage the input pivot token representations $\boldsymbol{h}_*^{(l-1)}$ from the preceding layer instead of $\boldsymbol{h}_*^{(l)}$ to coarsely choose expert interventions with the top-K_θ ($K_\theta > k_\theta$) frequency ranked by FluidRoute:

$$\left\{\{\boldsymbol{A}_{\mathcal{T}_i}^{(l)}, \boldsymbol{B}_{\mathcal{T}_i}^{(l)}\} | \hat{\mathcal{R}}_{\mathcal{T}_i}(\boldsymbol{h}_*^{(l-1)}) \in \text{top-}K_\theta\ \hat{\mathcal{R}}(\boldsymbol{h}_*^{(l-1)})\right\}. \quad (9)$$

Concurrently with the current layer's computation, the parameters of these selected experts are asynchronously loaded into HBM for use in Stage 2. Upon completion of Stage 2, these parameters are evicted from HBM. This allows the data transfer to overlap with the model's dense computations (e.g., FFNs, attention), thereby reducing the substantial storage overhead of numerous expert parameters in CTG without compromising inference speed.

4 Experiments

In this section, we conduct empirical experiments to demonstrate the efficacy of our RaMod framework.

4.1 Setup

Benchmarks. We train our method and CTG baselines on 256 tasks from the Flan v2 dataset [Longpre *et al.*, 2023]. Following [Ostapenko *et al.*, 2024] and [Wang *et al.*, 2023], we sample 10,000 instances per task and use 1,000 for validation to ensure computational feasibility. For cross-task generalization, we evaluate these methods on 10 widely-used tasks in the zero-shot manner, including: (1) common-sense reasoning: WinoGrande [Sakaguchi *et al.*, 2021], HellaSwag [Zellers *et al.*, 2019], PIQA [Bisk *et al.*, 2020]; (2) question answering: BoolQ [Clark *et al.*, 2019], OpenbookQA [Mihaylov *et al.*, 2018], ARC-Easy and ARC-Hard [Clark *et al.*, 2018]; (3) coding: HumanEval [Liu *et al.*, 2023a] and MBPP [Austin *et al.*, 2021]; and (4) general-purpose reasoning: BHH [Suzgun *et al.*, 2023]. To ensure a valid evaluation of the LLM's zero-shot generalization capabilities, we remove any training data from the Flan v2 dataset that overlapped with these 10 unseen tasks.

Baselines. We verify the effectiveness of RaMod compared with the following zero-shot baselines: (1) `Base`: the base LLM without any adaptation; (2) `Shared`: a single LoRA expert finetuned on the joint training set of all 256 tasks; (3) `FullFT`: the base LLM is fully fine-tuned like `Shared`; (4) `Arrow` [Ostapenko *et al.*, 2024]: A zero-shot dynamic routing mechanism that directs each token to the most relevant parameters among a set of well-trained LoRA modules. Besides these baselines, we compare our method with `MeteoRA` [Xu *et al.*, 2025], a method based on a MoE architecture that adaptively selects LoRA parameters, in order to evaluate the impact of an additional learnable gating mechanism. The gating parameters in MeteoRA are trained using a held-out set of 500 instances from each dataset.

Implementations. Following [Ostapenko *et al.*, 2024], we use the Phi-2 and Mistral-7B as the base LLMs. Unless stated otherwise, we set the rank of 4, dropout of 0.05 and learning rate of 1e-4 for all expert modules, *i.e.,* interventions and LoRAs. For our RaMod method, we select the tokens from the calculation in Eqn. (1) to serve as pivot tokens, with k_{AG}=14. The original ReFT method uses a fixed set of 7 prefix and 7 suffix tokens. These two token selection strategies are configured as two variants of our method: **RaMod-AG** and **RaMod-Fix**, respectively. For selecting expert interventions (Eqn. (7)), we choose k_θ=4 expert parameters for each pivot token. Our asynchronous orchestrator uses the CPU to calculate the logits from Eqn. (9) and screens for the top-K_θ most likely to be utilized by the current layer, where K_θ=10. Results are averaged over three runs with distinct random seeds. All experiments were conducted on an Ubuntu 20.04.5 server,

Method \ Eval	Efficiency (s / GB)			Performance (%)										
	TTFT	TPOT	Mem.	WG	HSwag	PIQA	BoolQ	OQA	ARC-E	ARC-C	HE	MBPP	BHH	Avg.
Phi-2 (2.7B)	0.141	0.024	9.0	75.7	72.5	79.2	82.7	49.8	77.5	52.9	45.1	56.0	48.0	63.8
`Shared`	0.141	0.024	9.0	76.6	73.4	80.4	82.4	50.4	83.2	55.8	46.3	50.4	48.4	65.5
`FullFT`	0.141	0.024	9.0	77.0	73.2	80.3	80.8	48.0	83.5	57.9	50.0	57.2	47.7	65.6
`Arrow`	1.925	1.482	19.8	77.6	72.6	80.2	84.3	52.2	84.2	56.4	50.6	59.9	47.7	66.6
`MeteoRA`	2.169	1.253	20.7	76.1	73.8	79.6	83.2	**52.7**	83.7	56.0	49.3	57.9	48.1	66.0
RaMod-Fix	0.381	0.024	11.6	78.1	**75.2**	80.2	84.1	51.8	84.6	57.9	51.3	60.3	**48.9**	67.2
RaMod-AG	0.474	0.024	11.6	**78.4**	74.9	**80.9**	**84.6**	52.3	**85.3**	**58.3**	**51.7**	**60.8**	48.6	**67.6**
Mistral (7B)	0.162	0.032	14.9	66.5	78.8	81.1	82.2	44.6	68.9	49.6	28.0	47.5	47.9	59.5
`Shared`	0.162	0.032	14.9	**68.6**	79.5	50.4	84.6	50.4	84.8	60.0	24.4	47.5	49.2	63.1
`Arrow`	2.637	1.861	32.6	66.6	81.1	82.8	86.6	50.6	85.7	60.8	30.5	**49.4**	49.5	64.4
`MeteoRA`	3.096	1.803	33.8	67.3	81.2	81.9	86.3	48.7	85.3	60.1	29.1	48.5	49.1	63.8
RaMod-Fix	0.840	0.032	18.5	68.1	81.7	81.6	86.8	50.4	85.9	60.3	29.3	49.1	**50.0**	64.3
RaMod-AG	1.014	0.032	18.5	68.5	**81.9**	**82.9**	**87.4**	**50.9**	**86.1**	**61.3**	**31.4**	48.8	49.8	**64.9**

Table 2: The results of cross-task generalization performance and efficiency for Phi-2 and Mistral backbones. All experiments are conducted using the BF16 precision format. We evaluate efficiency by measuring time (s) and peak GPU memory (GB) with a batch size of 8, processing 2048 input tokens and generating up to 1024 tokens. Prefill is measured by Time-to-First-Token (TTFT), and generation by Time-Per-Output-Token (TPOT).

equipped with a 2.20GHz Intel Xeon Gold 5220R CPU with 1024GB memory, an NVIDIA A100 (80G) GPU.

4.2 Performance Comparison

The cross-task generalization results are enumerated in Table 2. Overall, our proposed RaMod-AG demonstrates superior performance in CTG across most datasets, achieving optimal average performance. When using Phi-2 as the backbone, RaMod-AG shows an average improvement of 1.0% compared to the second-best baseline, Arrow. With Mistral as the backbone, RaMod-AG improves by an average of 0.5% over Arrow and 1.1% over MeteoRA. This indicates that the additional parameters introduced by learnable gating mechanisms pose a higher risk of overfitting. In contrast, our proposed zero-shot routing only intervenes on the representations of a few tokens, which enhances the generalization performance of LLMs. Furthermore, RaMod-AG outperforms RaMod-Fix with a 0.5% improvement on the Phi-2 backbone and a 0.6% improvement on the Mistral backbone. This suggests that intervening on the representations of pivot tokens based on attention scores is more effective at reinforcing sample-specific task features, which improves generalization to unseen tasks.

4.3 Efficiency Gains

The efficiency evaluation of our method and the baselines is also included in Table 2. The evaluation primarily focuses on the maximum GPU memory usage and the time cost of the prefill and generation stages. Shared and FullFT incur identical storage overhead and maintain the same prefill and generation speeds as the original LLM, since each employs only a single expert that can be directly merged into the original model weights or fine-tuned in-place. In contrast, Arrow requires storing all expert parameters in HBM, and MeteoRA not only stores all experts but also trains an additional gating network. Moreover, computing logits for all experts introduces considerable computational overhead. For instance of Phi-2, MeteoRA increases TTFT by 15.4× and TPOT by 52.2× compared with the original LLM. Our approach, however, avoids interference during the generation phase, thus keeping TPOT consistent with the original model. During prefill, we intervene only on a small number of token representations, substantially reducing the memory usage associated with dynamic routing. Furthermore, by leveraging the *Asynchronous Orchestrator* to keep non-essential experts in CPU memory, RaMod-Fix drastically trims the overhead of MeteoRA, reducing its excess TTFT, TPOT and HBM usage over Phi-2 by 88.2%, 100%, 77.8%, and Mistral by 76.9%, 100%, 81.1%, respectively.

Method \ Eval	Efficiency (s / GB)			Avg. Perf.
	TTFT	TPOT	Mem.	(%)
RaMod-AG	0.474	0.024	11.6	67.6
w/o AG (Sec. 3.1)	0.381	0.024	11.6	67.2
w/o A.O. (Sec. 3.3)	0.849	0.024	20.1	67.7

Table 3: Ablation studies evaluating the contribution of key components of our method on Phi-2. A.O. denotes the Asynchronous Orchestrator in Sec. 3.3.

4.4 Ablation Study

In this subsection, we attempt to understand how the attention-guided (AG) pivot token selection and the asynchronous orchestrator (A.O.) contribute to the performance and efficiency of CTG. As shown in Table 3, RaMod-AG yields a notable improvement in CTG performance, achieving a 0.4% gain while reducing TTFT latency by less than 0.1 seconds compared to the variant without AG (i.e., RaMod-

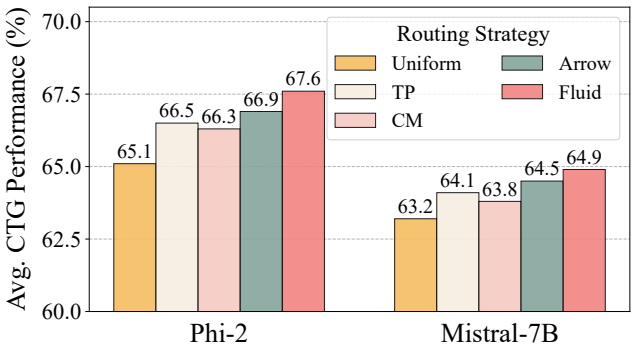

Figure 4: Comparison of routing strategies for average CTG performance across different backbones.

Fix in Table 2). Furthermore, by comparing RaMod-AG with the variant where the A.O. is removed, it can be observed that A.O. significantly reduces HBM usage. Benefiting from asynchronous scheduling and the involvement of fewer experts parameters in token representation intervention, the TTFT latency is also substantially reduced. This process filters out less critical experts while maintaining the model's generalization capability without significant loss.

4.5 Routing Strategy Comparison

The effectiveness of the FluidRoute logit proposed in Sec 3.2 is evaluated in Figure 4 by comparing CTG performance across different LLM backbones within the RaMod-AG framework using several routing strategies: 1) μ-routing, which assigns uniform weights to all experts; 2) TP-routing, which learns a classifier based on categorical cross-entropy loss that depends on the entire input prompt rather than individual token representations; 3) CM-routing, which preserves a prototype for each expert by averaging the hidden representations from the LLM forward pass on each upstream task and applies softmax aggregation based on the cosine similarity between these expert prototypes and the token representations; and 4) Arrow, which ignores the effect of singular values and utilizes only the first right singular vector. Figure 4 demonstrates the superior CTG performance of our FluidRoute method, which outperforms the second-best baseline (Arrow) by an average of 0.7%. The key advantage of FluidRoute lies in its efficient and precise estimation of expert interventions per token, achieved via an optimal rank-1 SVD approximation and theoretical guarantee of Theorem 1.

5 Related Work

Parameter-efficient Fine-Tuning (PeFT). PeFT has emerged as a prominent approach to mitigate the high computational overhead associated with full fine-tuning of LLMs for downstream tasks [Houlsby *et al.*, 2019; Liu *et al.*, 2022a; Zhang *et al.*, 2023; Huang *et al.*, 2025b]. LoRA [Hu *et al.*, 2022] stands out as a prominent example of PeFT. It offers efficient adaptation by updating only injected low-rank matrices while keeping the parameters of the base LLM frozen. Recently, Representation Fine-tuning [Wu *et al.*, 2024b] has been proposed as an alternative to LoRA, achieving high computational efficiency by directly editing a small subset of hidden representations from the LLM.

Cross-task generalization (CTG). The goal of CTG is for LLMs to exhibit strong performance on tasks they have not been explicitly trained on, also known as unseen tasks. CTG is typically divided into two settings depending on the requirement for exemplar data from unseen tasks: few-shot learning [Huang *et al.*, 2024; Ponti *et al.*, 2023; Vu *et al.*, 2022], which uses a small number of labeled examples for alignment, and zero-shot learning [Sanh *et al.*, 2022], which uses none. [Wu *et al.*, 2024a] treat each trained LoRA as an expert and integrate learnable gating function to learn composition weights for *each token* independently in the MoE's manner. [Xu *et al.*, 2025] further develop a GPU kernel operator for enhancing the efficiency of the naive MoE implementations. [Ostapenko *et al.*, 2024] decompose LoRA parameters through SVD and design a training-free routing mechanism based on the correlation between the SVD decomposition results and the token representations. [Charakorn *et al.*, 2025] use instruction tuning to generate LoRA parameters conditioned on task description embeddings from a separate, small sentence model. However, the efficiency of existing works is hampered by the need to activate different adapters on a token-by-token basis, the requirement to keep all trained adapters in GPU memory for immediate use, or the dependency on an additional sentence embedding model.

6 Conclusion

This study investigated a crucial research question: efficient cross-task generalization for LLMs. The aim is to not only improve the generalization capabilities of LLMs but also reduce the memory usages of fine-tuned parameters and enhance the inference speed. To tackle this issue, we proposed a Representation-Aware Modularity framework, called RaMod. Within this framework, we developed a modular representation fine-tuning strategy to manipulate a minimal subset of token representations through a series of highly targeted interventions. Furthermore, we implemented an asynchronous orchestrator to allocate GPU memory for required interventions and promptly release memory upon intervention completion, thereby minimizing memory overhead. Our comprehensive experimental results demonstrated the effectiveness of RaMod for LLM cross-task generalization.

A Theoretical Proof

As established in Sec. 3.2, Theorem 1 guarantees that given an input representation for a specific positional task, our proposed routing mechanism exhibits a higher probability of selecting the intervention corresponding to the most analogous task. We proceed to demonstrate the derivation of this theorem through mathematical expectation. Let a pretrained weight matrix be denoted by $W \in \mathbb{R}^{d\times m}$. For a task $\mathcal{T}_k$, its adaptation is governed by a low-rank matrix $\Delta \boldsymbol{W}_{\mathcal{T}_k} = \boldsymbol{B}_{\mathcal{T}_k}\boldsymbol{A}_{\mathcal{T}_k}$, where $\boldsymbol{B}_{\mathcal{T}_k} \in \mathbb{R}^{d\times r}$, $\boldsymbol{A}_{\mathcal{T}_k} \in \mathbb{R}^{r\times m}$, and $r \ll \min(d, m)$. The adapted forward pass for an input $\boldsymbol{h} \in \mathbb{R}^m$ is $(1 + \boldsymbol{B}_{\mathcal{T}_k}\boldsymbol{A}_{\mathcal{T}_k})\boldsymbol{W}\boldsymbol{h}$. We consider a cross-task generalization setting with upstream tasks $\mathbb{T}$. Each task k has

a data distribution $\mathcal{D}_k$. The parameters are optimized by minimizing the following objective:

$$\mathcal{L}_{\mathcal{T}_k} = \mathbb{E}_{\boldsymbol{h}\sim\mathcal{D}_{\mathcal{T}_k}}\left[\|(W + B_{\mathcal{T}_k}A_{\mathcal{T}_k})\,h - y_{\mathcal{T}_k}\|_2^2\right],$$

where y_k is the target output. We aim to demonstrate that for a set of experts $\{\Delta W_1, ..., \Delta W_N\}$, where each expert i is trained on a specific task $\mathcal{T}_i$, the proposed static routing score will, in expectation, be maximized for samples drawn from its corresponding dataset. Given a hidden state $\boldsymbol{h}$ from task $\mathcal{T}_i$, we seek to prove that the routing logit for expert i, defined as:

$$\mathcal{R}_{\mathcal{T}_i}(\boldsymbol{h}) = \|\boldsymbol{B}_{\mathcal{T}_i}\boldsymbol{A}_{\mathcal{T}_i}\boldsymbol{h}\|_F, \quad \boldsymbol{h} \sim \mathcal{D}_k \tag{10}$$

is statistically greater than the logit for any other expert $j \neq i$. To better illustrate the proof, we first illustrate two key definitions.

Definition 1 (Task Subspace)**.** *For a task k, the task subspace U_k is defined as the column space of A_k^T, or equivalently, the row space of A_k: $U_k = span(\text{rows of } A_k) \subset \mathbb{R}^m$.*

Definition 2 (Subspace Similarity)**.** *The similarity between two task subspaces U_i and U_j is defined as the cosine of the first principal angle θ_1 between them:*

$$S(U_i, U_j) = \cos(\theta_1) = \max_{u\in U_i,\, v\in U_j} \frac{|u^T v|}{\|u\|_2\|v\|_2}.$$

Note that $0 \leq S(U_i, U_j) \leq 1$, where 0 implies orthogonality and 1 implies identical subspaces.

The proof relies on two foundational results from matrix analysis and subspace theory.

Lemma 1 (Eckart–Young–Mirsky Theorem, [Golub *et al.*, 1987])**.** *Let $X \in \mathbb{R}^{d\times n}$ be a matrix of rank p, and let $r < p$. The optimal rank-r approximation X_r that minimizes $\|X - \widehat{X}\|_F$ is given by the truncated Singular Value Decomposition (SVD) of X, retaining the r largest singular values and their corresponding singular vectors. Furthermore, the approximation error is $\|X - X_r\|_F^2 = \sum_{i=r+1}^{p} \lambda \boldsymbol{A}_{\mathcal{T}_i}^2$.*

Lemma 2 (Projection Bound via Principal Angles, [Golub and Van Loan, 2013])**.** *Let U_i and U_j be subspaces of $\mathbb{R}^m$ with first principal angle θ_1. Let P_{U_j} be the orthogonal projection matrix onto U_j. For any vector $h \in U_i$, the norm of its projection onto U_j is bounded by:*

$$\|P_{U_j}(h)\|_2 \leq \cos(\theta_1)\|h\|_2 = S(U_i, U_j)\|h\|_2.$$

For a general vector $h = h_{U_i} + h_\perp$ where $h_{U_i} \in U_i$ and $h_\perp \perp U_i$, the bound holds approximately:

$$\|P_{U_j}(h)\|_2 \lesssim S(U_i, U_j)\|h_{U_i}\|_2 + \delta,$$

where $\delta = \|P_{U_j}(h_\perp)\|_2$ is a small term under the assumption that h is predominantly in U_i.

Our proof is built upon the following three assumptions:

Assumption 1 (Bounded Parameters)**.** *The operator norms of the parameter matrices are bounded, i.e., $\exists\, C_B, C_A > 0$ such that $\|B_k\|_{op} \leq C_B$ and $\|A_k\|_{op} \leq C_A$ for all tasks k.*

Assumption 2 (Optimization Efficiency)**.** *The task-specific residual $r_k = y_k - Wh$ is non-zero and correlates with the input, i.e., $\mathbb{E}[\|r_k\|_2] \geq K\mathbb{E}[\|h\|_2]$ for some $K > 0$ and $h \sim \mathcal{D}_k$.*

Assumption 3 (Distinct Subspace)**.** *The regularizer $\mathcal{R}$ ensures the similarity between any two distinct task subspaces is bounded above by a small constant $\delta > 0$, i.e., $S(U_i, U_j) \leq \delta$ for all $i \neq j$.*

Proof of Theorem 1. For task i, the learned matrix $\boldsymbol{B}_{\mathcal{T}_i}\boldsymbol{A}_{\mathcal{T}_i}$ is a low-rank approximation of the residual mapping $r_i = y_i - \boldsymbol{W}\boldsymbol{h}_i$. By Lemma 1 (Eckart-Young), this approximation is optimal. Therefore, for an input $\boldsymbol{h}_i$, we have:

$$\boldsymbol{B}_{\mathcal{T}_i}\boldsymbol{A}_{\mathcal{T}_i}\boldsymbol{h}_i \approx r_i.$$

This implies that the output norm is approximately the norm of the residual:

$$\|\boldsymbol{B}_{\mathcal{T}_i}\boldsymbol{A}_{\mathcal{T}_i}\boldsymbol{h}_i\|_2 \approx \|r_i\|_2. \tag{11}$$

From Assumption 2, $\|r_i\|_2$ is significant relative to $\|\boldsymbol{h}_i\|_2$.

For a different task $j \neq i$, we analyze $\|\boldsymbol{B}_{\mathcal{T}_j}\boldsymbol{A}_{\mathcal{T}_j}\boldsymbol{h}_i\|_2$. First, we bound this using the operator norms:

$$\begin{aligned}\|\boldsymbol{B}_{\mathcal{T}_j}\boldsymbol{A}_{\mathcal{T}_j}\boldsymbol{h}_i\|_2 &= \|\boldsymbol{B}_{\mathcal{T}_j}(\boldsymbol{A}_{\mathcal{T}_j}\boldsymbol{h}_i)\|_2 \\ &\leq \|\boldsymbol{B}_{\mathcal{T}_j}\|_{\text{op}} \cdot \|\boldsymbol{A}_{\mathcal{T}_j}\boldsymbol{h}_i\|_2 \leq C_B \cdot \|\boldsymbol{A}_{\mathcal{T}_j}\boldsymbol{h}_i\|_2.\end{aligned} \tag{12}$$

The term $\|\boldsymbol{A}_{\mathcal{T}_j}\boldsymbol{h}_i\|_2$ measures the projection of $\boldsymbol{h}_i$ onto the row space of $\boldsymbol{A}_{\mathcal{T}_j}$, i.e., the task subspace U_j. Applying Lemma 2, and noting that $\boldsymbol{h}_i$ is predominantly in its own task subspace U_i, we get:

$$\begin{aligned}\|\boldsymbol{A}_{\mathcal{T}_j}\boldsymbol{h}_i\|_2 &\lesssim \|\boldsymbol{A}_{\mathcal{T}_j}\|_{\text{op}} \cdot S(U_i, U_j) \cdot \|\boldsymbol{h}_i\|_2. \\ &\leq C_A \cdot \delta \cdot \|\boldsymbol{h}_i\|_2.\end{aligned} \tag{13}$$

Substituting Eqn. (13) into Eqn. (12) yields the final bound for the cross-task output:

$$\|\boldsymbol{B}_{\mathcal{T}_j}\boldsymbol{A}_{\mathcal{T}_j}\boldsymbol{h}_i\|_2 \leq C \cdot \delta \cdot \|\boldsymbol{h}_i\|_2, \tag{14}$$

where $C = C_B \cdot C_A$.

We now compare the task-specific output (Eqn. (11)) and the cross-task output (Eqn. (14)):

$$\|\boldsymbol{B}_{\mathcal{T}_i}\boldsymbol{A}_{\mathcal{T}_i}\boldsymbol{h}_i\|_2 - \|\boldsymbol{B}_{\mathcal{T}_j}\boldsymbol{A}_{\mathcal{T}_j}\boldsymbol{h}_i\|_2 \gtrsim \|r_i\|_2 - C\delta\|\boldsymbol{h}_i\|_2.$$

From Assumption 2 ($\|r_i\|_2 \geq K\|\boldsymbol{h}_i\|_2$), it follows that:

$$\|\boldsymbol{B}_{\mathcal{T}_i}\boldsymbol{A}_{\mathcal{T}_i}\boldsymbol{h}_i\|_2 - \|\boldsymbol{B}_{\mathcal{T}_j}\boldsymbol{A}_{\mathcal{T}_j}\boldsymbol{h}_i\|_2 \gtrsim (K - C\delta)\|\boldsymbol{h}_i\|_2.$$

From Assumption 3, the similarity δ is small. By ensuring that $\delta < K/C$, the term $(K - C\delta)$ is positive. Therefore,

$$\|\boldsymbol{B}_{\mathcal{T}_i}\boldsymbol{A}_{\mathcal{T}_i}\boldsymbol{h}_i\|_2 > \|\boldsymbol{B}_{\mathcal{T}_j}\boldsymbol{A}_{\mathcal{T}_j}\boldsymbol{h}_i\|_2 \quad \text{for all} \quad j \neq i,$$

which completes the proof. □

Acknowledgements

This work is partly supported by the National Natural Science Foundation of China (No. 62306255, No. 62506348), the Natural Science Foundation of Anhui Province (Grant No. 2508085QF211), New Generation Artificial Intelligence-National Science and Technology Major Project (Grant No. 2025ZD0122601), the Opening Foundation of State Key Laboratory of Cognitive Intelligence, iFLYTEK (COGOS-2025HE02).

References

[Austin *et al.*, 2021] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

[Belofsky, 2023] Joshua Belofsky. Token-level adaptation of lora adapters for downstream task generalization. In *Proceedings of the 2023 6th Artificial Intelligence and Cloud Computing Conference*, pages 168–172, 2023.

[Bisk *et al.*, 2020] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, number 05, pages 7432–7439, 2020.

[Charakorn *et al.*, 2025] Rujikorn Charakorn, Edoardo Cetin, Yujin Tang, and Robert Tjarko Lange. Text-to-loRA: Instant transformer adaption. In *Forty-second International Conference on Machine Learning*, 2025.

[Clark *et al.*, 2018] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[Clark *et al.*, 2019] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of ACL: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, 2019.

[Gabrielsson *et al.*, 2024] Rickard Brüel Gabrielsson, Jiacheng Zhu, Onkar Bhardwaj, Leshem Choshen, Kristjan Greenewald, Mikhail Yurochkin, and Justin Solomon. Compress then serve: Serving thousands of lora adapters with little overhead. In *Forty-second International Conference on Machine Learning*, 2024.

[Geiger *et al.*, 2021] Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Neurips*, 34:9574–9586, 2021.

[Ghaeini *et al.*, 2018] Reza Ghaeini, Xiaoli Fern, and Prasad Tadepalli. Interpreting recurrent and attention-based neural models: a case study on natural language inference. In *Proceedings of the 2018 Conference on EMNLP*, pages 4952–4957, 2018.

[Golub and Van Loan, 2013] Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013.

[Golub *et al.*, 1987] Gene H Golub, Alan Hoffman, and Gilbert W Stewart. A generalization of the eckart-young-mirsky matrix approximation theorem. *Linear Algebra and its applications*, 88:317–327, 1987.

[Gong and Sun, 2024] Zheng Gong and Ying Sun. Graph reasoning enhanced language models for text-to-sql. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2447–2451, 2024.

[Houlsby *et al.*, 2019] 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.

[Hu *et al.*, 2022] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[Huang *et al.*, 2024] Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. Lorahub: Efficient cross-task generalization via dynamic lora composition. In *First Conference on Language Modeling*, 2024.

[Huang *et al.*, 2025a] Chenxi Huang, Shaotian Yan, Liang Xie, Binbin Lin, Sinan Fan, Yue Xin, Deng Cai, Chen Shen, and Jieping Ye. Enhancing chain-of-thought reasoning with critical representation fine-tuning. In *Proceedings of the 63rd Annual Meeting of ACL (Volume 1: Long Papers)*, pages 23173–23195, 2025.

[Huang *et al.*, 2025b] Weizhong Huang, Yuxin Zhang, Xiawu Zheng, Jing Lin, Yiwu Yao, Rongrong Ji, et al. Dynamic low-rank sparse adaptation for large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Lee *et al.*, 2024] Wonbeom Lee, Jungi Lee, Junghwan Seo, and Jaewoong Sim. {InfiniGen}: Efficient generative inference of large language models with dynamic {KV} cache management. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pages 155–172, 2024.

[Li and Liang, 2021] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of ACL and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, 2021.

[Lin *et al.*, 2025] Tianwei Lin, Jiang Liu, Wenqiao Zhang, Yang Dai, Haoyuan Li, Zhelun Yu, Wanggui He, Juncheng Li, Jiannan Guo, Hao Jiang, Siliang Tang, and Yueting Zhuang. Teamlora: Boosting low-rank adaptation with expert collaboration and competition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 13622–13637. Association for Computational Linguistics, 2025.

[Liu *et al.*, 2022a] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Neurips*, 35:1950–1965, 2022.

[Liu *et al.*, 2022b] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of ACL (Volume 2: Short Papers)*, pages 61–68, 2022.

[Liu *et al.*, 2023a] Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *Neurips*, 36:21558–21572, 2023.

[Liu *et al.*, 2023b] Zichang Liu, Jue Wang, Tri Dao, Tianyi Zhou, Binhang Yuan, Zhao Song, Anshumali Shrivastava, Ce Zhang, Yuandong Tian, Christopher Re, et al. Deja vu: Contextual sparsity for efficient llms at inference time. In *International Conference on Machine Learning*, pages 22137–22176. PMLR, 2023.

[Longpre *et al.*, 2023] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR, 2023.

[Luo *et al.*, 2018] Ling Luo, Xiang Ao, Feiyang Pan, Jin Wang, Tong Zhao, Ningzi Yu, and Qing He. Beyond polarity: Interpretable financial sentiment analysis with hierarchical query-driven attention. In *IJCAI*, pages 4244–4250, 2018.

[Mihaylov *et al.*, 2018] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on EMNLP*, pages 2381–2391, 2018.

[Mishra *et al.*, 2022] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of ACL (Volume 1: Long Papers)*, pages 3470–3487, 2022.

[Ostapenko *et al.*, 2024] Oleksiy Ostapenko, Zhan Su, Edoardo Ponti, Laurent Charlin, Nicolas Le Roux, Lucas Caccia, and Alessandro Sordoni. Towards modular llms by building and reusing a library of loras. In *International Conference on Machine Learning*, pages 38885–38904. PMLR, 2024.

[Ponti *et al.*, 2023] Edoardo Maria Ponti, Alessandro Sordoni, Yoshua Bengio, and Siva Reddy. Combining parameter-efficient modules for task-level generalisation. In *Proceedings of the 17th Conference of the European Chapter of ACL*, pages 687–702, 2023.

[Sakaguchi *et al.*, 2021] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

[Sanh *et al.*, 2022] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, et al. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2022.

[Suzgun *et al.*, 2023] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. In *ACL (Findings)*, 2023.

[Vu *et al.*, 2022] Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Cer. Spot: Better frozen model adaptation through soft prompt transfer. In *Proceedings of the 60th Annual Meeting of ACL (Volume 1: Long Papers)*, pages 5039–5059, 2022.

[Wang *et al.*, 2023] Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. How far can camels go? exploring the state of instruction tuning on open resources. *Neurips*, 36:74764–74786, 2023.

[Wu *et al.*, 2024a] Xun Wu, Shaohan Huang, and Furu Wei. Mixture of lora experts. In *The Twelfth International Conference on Learning Representations*, 2024.

[Wu *et al.*, 2024b] Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Reft: Representation finetuning for language models. *Neurips*, 37:63908–63962, 2024.

[Xin *et al.*, 2025] Haoran Xin, Ying Sun, Chao Wang, and Hui Xiong. Llmcdsr: Enhancing cross-domain sequential recommendation with large language models. *ACM Transactions on Information Systems*, 43(5):1–33, 2025.

[Xu *et al.*, 2025] Jingwei Xu, Junyu Lai, and Yunpeng Huang. Meteora: Multiple-tasks embedded lora for large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Ye *et al.*, 2021] Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. Crossfit: A few-shot learning challenge for cross-task generalization in nlp. In *Proceedings of the 2021 Conference on EMNLP*, pages 7163–7189, 2021.

[Zellers *et al.*, 2019] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of ACL*, pages 4791–4800, 2019.

[Zhang *et al.*, 2023] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations*, 2023.

[Zhang *et al.*, 2025] Mingxu Zhang, Dazhong Shen, Qi Zhang, and Ying Sun. Chematp: A training-free chemical reasoning framework for large language models. *arXiv preprint arXiv:2512.19240*, 2025.