applied sciences

MDPI

Article

Exploring a New Architecture for Efficient Parameter Fine-Tuning in SLoRA Multitasking Scenarios

Ce Shi and Jin-Woo Jung *

Department of Computer Science and Artificial Intelligence, Dongguk University, Seoul 04620, Republic of Korea; sc921001@163.com

* Correspondence: jwjung@dongguk.edu

Abstract

Propose an enhanced LoRA (Low-Rank Adaptation) MoE (mixed expert) architecture, SLoRA (Enhanced LoRA MoE Architecture), aimed at addressing the key problem of efficient parameter fine-tuning in multitasking scenarios. Given the high cost of traditional full fine-tuning as the parameter size of visual language models increases, and the limitations of LoRA as a popular PEFT (parameter-efficient fine-tuning) method in multitasking, such as inadequate adaptability and difficulty in capturing complex task patterns, as well as the catastrophic forgetting and knowledge fragmentation challenges faced by existing research on integrating mixed expert (MoE) mechanisms into LoRA, SLoRA utilizes orthogonal constraint optimization to reduce disturbance to existing knowledge through constraint solution space initialization, alleviating catastrophic forgetting (old task accuracy retention rate reaches 92.4%, 16.1% higher than LoRA), and an optimized MoE structure that includes general experts (retaining pre-trained knowledge) and task-specific experts (dynamic routing adaptation tasks) to enhance multitask adaptability. Experimental results show that in commonsense reasoning tasks, SLoRA's accuracy is 9.0% higher than LoRA and 3.7% higher than AdaLoRA on the WSC dataset, and its F1 score is 7.7% higher than LoRA and 2.9% higher than AdaLoRA on the CommonsenseQA dataset; in multimodal tasks, its average score is up to 15.3% higher than LoRA, demonstrating significant advantages over existing methods.

Keywords: SLoRA; PEFT; multi task scenarios; catastrophic forgetting; fragmentation of knowledge

Academic Editors: Patrícia Ramos and Jose Manuel Oliveira

Received: 29 December 2025
Revised: 6 February 2026
Accepted: 10 February 2026
Published: 24 February 2026

1. Introduction

In recent years, visual language modeling tools (VLMs) have made significant progress in the field of artificial intelligence. Due to their strong multimodal processing capabilities, they have shown great potential in tasks such as image depiction and visual question answering. During the continuous updating of skill processes, the VLM parameter size shows an exponential upward trend; for example, the GPT-3 parameter size has reached 175 billion. Although the performance of such large-scale models has significantly improved, it has also created a dilemma of high computational costs. The traditional comprehensive fine-tuning mode requires updating all parameter contents of the model, resulting in a huge consumption of computing resources. Taking the 10 billion parameter model as an example, comprehensive fine-tuning operations may require dozens of high-performance GPUs (Graphics Processing Unit) to run continuously for several days, while also requiring significant storage costs to retain model parameter data. In resource-constrained environments

(such as IoT (Internet of Things) medical monitoring systems [1]), this situation is almost impossible to achieve, which brings many challenges and difficult situations to practical applications [2].

To address the pain points of traditional full-scale fine-tuning, the parameter-efficient fine-tuning (PEFT) method has emerged in response, with low order adaptive techniques (LoRA) becoming one of the mainstream technologies. LoRA uses low rank matrix factorization operations to introduce only a small number of trainable parameter components in the fine-tuning process, freezing most of the original model parameter parts and significantly reducing the amount of computational resources and time costs required for training. For example, when implementing LoRA fine-tuning on a pre-trained model with 1 billion parameters, only about 10 million parameters need to be adjusted, reducing computational complexity by over 90%. This high-performance characteristic enables it to be widely applied in multiple fields: Pu and Xu (2024) [3] introduced LoRA technology into a Transformer-based directional object detection device, achieving airborne processing of remote sensing images; Cappelletti et al. (2025) [4] completed the task of using LoRA technology to detect a small number of deep counterfeit samples; Gianluca et al. (2025) [5] used LoRA-adapted embedding to predict intrinsic diseases in protein sequences; Yu et al. (2021) [6] combined LoRA technology with dual blockchain architecture to construct a contract theory-driven information system architecture, fully verifying its cross-domain scenario adaptation performance; Kowsher et al. (2024) [7] proposed the content of the promotion method, which has been fine-tuned to guide the direction of large language models (LLMs); Bian et al. (2024) [8] proposed LoRA FAIR content, which was optimized through aggregation and initialization in a federated learning environment to achieve fine-tuning of LoRA operations; Zhang et al. (2025) [9] used heterogeneous LoRA allocation methods to effectively fine-tune the body of the federated basic model in Fed Hello; Zhao et al. (2025) [10] proposed the LasQ scheme, which utilizes quantitative techniques to achieve the fine-tuning process of LLM maximum singular components; Qu et al. (2025) [11] implemented an embedded routing function within an efficient fine-tuning program for low rank bottleneck parameters in visual language; Lin et al. (2025) [12] achieved scientific sentence recognition through phased LoRA fine-tuning measures; Liao et al. (2025) [13] constructed a dynamic adaptation strategy system for LoRA fine-tuning to achieve specific tasks and efficient optimization objectives for large-scale language models; Chen et al. (2025) [14] proposed CE LoRA as a computationally efficient LoRA fine-tuning solution for language models; Mu et al. (2024) [15] applied LoRA fine-tuning operations to multimodal large language model structures to complete multimodal sentiment analysis task projects. However, LoRA technology still faces specific challenges in real-world application scenarios: Djidi et al. (2021) [16] showed that under the background conditions of energy harvesting sensor nodes, it is necessary to optimize the downlink delay problem of LoRA technology through wake-up radio technology, which also reflects its adaptive limitations in different scenario environments.

Although LoRA has achieved significant results in the field of efficient parameter adjustment, its core shortcomings and drawbacks are gradually exposed in multitasking environments. There are significant differences in data distribution patterns, task goal orientation dimensions, and feature representation forms among different tasks. The fixed parameter adjustment paradigm of this model is difficult to flexibly adapt to complex multitask requirements, and cannot effectively capture the task mode state of complex structures, resulting in a decrease in performance during the task switching phase. This phenomenon is similar to what Rice et al. (1993) [17] found in their research on alcohol use therapy that "different age groups have different responses to the same treatment plan"—the lack of targeted adaptation mechanisms makes it difficult to meet different

types of needs. For example, when performing image classification and object detection tasks simultaneously, it is difficult to fully utilize the correlation feature elements between the two, resulting in the performance of the model not reaching the ideal level in both task activities. Although Tian et al. (2024) [18] proposed the asymmetric LoRA architecture Hydralora to optimize and fine-tune efficiency levels, LoRA fundamentally focuses on single task optimization as its core design concept, which still cannot meet the complex structural requirements of multitask environments.

It can be seen from Table 1 that SLoRA takes "orthogonal constraint optimization and optimization of MoE structure (covering general experts, task-specific experts and dynamic routing)" as its core content. Compared with the "low rank matrix decomposition and freezing pretraining parameters" of traditional LoRA, on the one hand, it solves the design limitations of LoRA single task optimization (including poor adaptability to multiple tasks, lack of knowledge forgetting mitigation mechanism, etc.); on the other hand, it improves the knowledge integration ability by means of the cross-expert attention mechanism, reduces the interference generated by old knowledge through the initialization of the constraint solution space, and ultimately achieves an increase of 7.8% in the average accuracy rate of multiple tasks and 92.4% in the accuracy rate of old tasks (16.1% higher than LoRA). Task switching loss is only 2.3%; compared with LoRA, the results are 6.4 percentage points lower and maintain a super low computational cost situation (the proportion of trainable parameters is only 0.8%).

Table 1. Comparison table of core differences between traditional LoRA and SLoRA.

Comparative Dimension	Traditional LoRA	SLoRA
Architecture Core	Low rank matrix factorization ($\Delta W = A \cdot B$), freeze pre-training parameters	Orthogonal constraint optimization + optimization of MoE structure (general expert + task-specific expert + dynamic routing)
Design objectives	Reduce computational costs and simplify parameter updates in single task scenarios	Balancing efficiency and adaptability in multitasking scenarios, alleviating forgetting and knowledge fragmentation
Multitask adaptability	Fixed parameter adjustment paradigm, difficult to adapt to task differences	Dynamic routing allocation + task-specific experts, flexible adaptation to complex task modes
Knowledge forgetting alleviation mechanism	Without specialized design, updating new task parameters can easily interfere with old knowledge	Initialize the constraint solution space, project new parameters to the orthogonal complementary space, and reduce knowledge disturbance
Knowledge integration ability	No cross-task knowledge sharing mechanism, low feature overlap	Cross-expert attention mechanism to enhance feature overlap (from 31.2% to 44.2%)
Old task accuracy retention rate	76.3% (sequential task scenario)	92.4% (sequential task scenario, increased by 16.1%)
Multitask average accuracy improvement	Benchmark level (based on oneself)	7.8% improvement compared to LoRA (10 sets of multitasking combination tests)
Proportion of computational cost (1 billion parameter model)	0.32% (r = 16)	0.8% (including MoE module, still only 0.8% of full fine-tuning)
Task switching loss	8.7% (cross-modal task switching)	2.3% (cross-modal task switching, reduced by 6.4 percentage points)

In order to address the shortcomings of traditional low rank adaptation (LoRA) and current mixed expert low rank adaptation (MoE LoRA) fusion solutions, and to fill the key research gap in efficient fine-tuning of visual language model (VLM) parameters in multitasking situations, three clear research questions are proposed, as follows:

① How can we overcome the shortcomings of traditional LoRA methods in multitasking adaptability? This deficiency is caused by their fixed low rank matrix update mechanism, which hinders their ability to adapt to the characteristics of different types of tasks and capture patterns between complex tasks.

② How can we alleviate the catastrophic forgetting and knowledge fragmentation issues in the existing MoE LoRA integration framework? These issues are caused by the new task parameter updates covering old task knowledge, as well as insufficient information exchange between expert modules.

③ How can we achieve a balance between high performance (including multitask accuracy and cross-pattern reasoning ability) and ultra-low computational consumption in the efficient parameter tuning process of large-scale visual language models (VLM)? Especially in environments where resources are limited, such as IoT applications.

In order to address these research issues, this study proposes the SLoRA architecture, whose core contributions are as follows, aimed at promoting the in-depth application of visual language models (VLMs) in multitasking scenarios:

(1) In response to the two core challenges of MoE LoRA (catastrophic forgetting and knowledge fragmentation), SLoRA combines orthogonal constraint optimization with optimized hybrid expert (MoE) structure to simultaneously achieve knowledge retention and cross-task knowledge integration improvement.

(2) Introducing constraint solution space initialization based on orthogonal constraint optimization, restricting the direction of parameter updates, reducing interference with existing knowledge, and effectively mitigating catastrophic forgetting in multitask learning.

(3) Design an optimized hybrid expert (MoE) structure consisting of "general experts + task-specific experts + dynamic routing" to enhance information sharing among experts, solve the problem of low efficiency in cross-mode collaboration, and improve multitask adaptability.

(4) Improve performance while maintaining ultra-low computational costs (only 0.8% of trainable parameters for a 1 billion parameter model) to adapt to resource-constrained environments such as the Internet of Things.

(5) Establish a multidimensional evaluation system that covers performance, efficiency, and universality to comprehensively verify the effectiveness and stability of the SLoRA architecture.

2. Technical Background

2.1. Efficient Parameter Fine-Tuning in Multitask Scenarios (PEFT)

The parameter-efficient fine-tuning (PEFT) technique, which has attracted considerable attention in the field of deep learning in recent years, aims to solve the problem of high computational costs in traditional full model fine-tuning for large-scale models. In multitask scenarios, models need to simultaneously handle multiple types of tasks. For example, in natural language processing, models may need to complete tasks such as text classification on the GLUE dataset with over 100,000 samples, sentiment analysis on the IMDb dataset with 50,000 comments, and machine translation on WMT14 English German pairs with 4.5 million parallel sentence pairs. In the field of computer vision, models may need to simultaneously process tasks such as ImageNet image classification with 1.2 million training samples, COCO object detection with 150,000 images and 880,000 object annotations, and PASCAL with 11,000 images and 27,000 segmentation annotations. VOC semantic segmentation and other tasks are also possible. The traditional comprehensive model fine-tuning method requires updating all parameters of the model, which is extremely expensive in multitasking scenarios. Taking a cross-modal pore-trained model with 10 billion parameters as an example, for its full parameter fine-tuning, a single round of training (assuming a batch size of 128) can achieve a computational cost of 5.2×10^{18} FLOPs (Floating-Point Op-

erations Per Second), requiring the deployment of 50 A100 GPUs (40 GB of video memory) to manage continuously for 14 days. The total computational cost exceeds $120,000, and the model parameter files that have to be stored after each round of training reach 400 GB (single precision floating-point number). If 10 rounds of checkpoints are saved, 4 TB of storage space is required, and more importantly, detailed fine-tuning can easily lead to overfitting in multitasking scenarios. In the above multitasking combination, the overfitting rate (training set accuracy, validation set accuracy) of the model after full fine-tuning reaches 18.3% in text classification tasks, 22.1% in object detection tasks, and even as high as 31.7% in semantic segmentation tasks with small data volume. The reason is that the data distribution of discrete tasks varies significantly (such as the feature space overlap between text classification and semantic segmentation being less than 15%), and updating parameters uniformly will force the model to overfit local features in some tasks.

PEFT technology achieves efficient fine-tuning in multitasking scenarios by adjusting only a small number of parameters in the model or introducing a small number of additional trainable parameters. Taking adapter modules as an example, in a pure trained model with 10 billion parameters, inserting two $128 \times 64 \times 2$ bottleneck structure adapter modules into each Transformer layer results in a total of only 5.12 million adapter modules for the entire model, accounting for only 0.51% of the original model parameters. Under this setting, the computational load of a single round of training can be reduced to 2.8×10^{17} FLOPs (only 5.4% of rich fine-tuning), and the computational cost can be reduced to $8000 with only three A100 GPUs running for 2 days. The storage requirement can also be reduced to 20 GB/round (including adapter parameters and optimizer status), and the overfitting rate is significantly reduced—text classification tasks are reduced to 5.2%, object detection is reduced to 7.8%, and semantic segmentation is reduced to 9.3%, all thanks to the stable feature extraction ability provided by frozen pure training parameters, which allows the adapter to only learn task-specific mapping. In multitasking scenarios, the model needs to have excellent adaptability and generalization. In terms of adaptability, the advantages of PEFT can be quantified by task switching efficiency: in the above multitask combination, the parameter adjustment amount of PEFT model switching tasks is only 0.3% of full fine-tuning, and the switching time is reduced from 2.3 h (reloading the model) of full fine-tuning to 45 s (updating only adapter parameters). As for performance, PEFT achieved a validation set accuracy of 85.7% (fully fine-tuned to 86.2%, with a difference of only 0.5%) in text classification tasks and 82.3% (fully fine-tuned to 81.9%) in sentiment analysis tasks, demonstrating its flexible adaptability to different tasks.

In terms of generalization, PEFT can achieve more effective knowledge transfer. For example, when the feature extractor is fine-tuned on the ImageNet image classification task and transferred to the COCO object detection task, the mAP (mean average precision, average accuracy) of the model can reach 42.6%, which is 3.5 percentage points higher than the result obtained by full fine-tuning transfer (39.1%); when the PEFT model, which was also fine-tuned on machine translation tasks, is transferred to cross-language text classification tasks, the average accuracy can reach 78.2%, significantly higher than the 72.5% of full fine-tuning transfer. This is because it retains more than 99% of the common knowledge in the pure trained model (through feature similarity calculation, the feature space overlap between the PEFT fine-tuned model and the original are trained model reaches 92.3%, while full fine-tuning is only 68.5%). This common knowledge provides a solid foundation for cross-task transfer. This efficient, low overfitting risk, and strong transfer ability make PEFT the core technology for large-scale model fine-tuning in multitask scenarios, especially in resource-limited industrial scenarios (such as edge device deployment and multitask real-time updates)—an irreplaceable advantage.

2.2. LoRA Technology Principles and Applications

2.2.1. LoRA Core Principles

Low rank adaptation (LoRA) is a popular parameter-efficient fine-tuning method, and its fundamental principle is based on low rank matrix factorization. In deep learning models, especially large-scale re-trained models such as GPT series and BERT, the weight matrix of the model generally has a high dimensionality. These high-dimensional matrices contain numerous parameters, and updating them all during the fine-tuning process would incur significant computational costs.

The idea of LoRA is to add two low rank matrices to the weight matrix of the re-trained model, and use the product of these two low rank matrices to roughly represent the updated weight values. For example, the weight matrix of the re-trained model is W, and during comprehensive fine-tuning, the entire W matrix needs to be updated. But in the LoRA approach, an additional fine-tuning matrix ΔW is added and expressed as the product of two low rank matrices A and B, that is, $\Delta W = A \cdot B$. Here, A belongs to $d \times r$ and B belongs to $r \times k$, where r is the rank of the low rank matrix and is much smaller than d and k. During the training phase, only low rank matrices A and B are updated, while the weight matrix W of the pure trained model remains unchanged. In this way, the number of trainable parameters has significantly decreased from the original $d \times k$ to $(d + k) \times r$, thereby substantially reducing computational costs and memory requirements.

Taking a pro trained model with 1 billion parameters as an example, assuming that the weight matrix dimension of one layer is $d = 1000$ and $k = 1000$, if all fine-tuning is performed, the number of parameters that need to be updated in this layer is $1000 \times 1000 = 1{,}000{,}000$. But if LoRA technology is used, when the rank of the low rank matrix is $r = 16$, the number of newly added trainable parameters in this layer is only $(1000 + 1000) \times 16 = 32{,}000$, which reduces the number of trainable parameters by about 96.8%. This dramatically reduces the number of parameters, allowing the model to converge faster during the fine-tuning process, while also reducing the requirements for computational resources. This enables effective fine-tuning of large-scale models even under limited resource conditions.

2.2.2. Limitations of LoRA in Multitask Scenarios

Although LoRA has achieved outstanding results in efficient parameter adjustment, it still has significant constraints in multitasking scenarios, mainly due to its inherent design pattern, which is difficult to fit the substantial heterogeneity between tasks. Each task has significant differences in the characteristics of data distribution, the direction of core objectives, and the form of feature presentation—these differences are not only superficial, but also involve fundamental differences in the construction of feature spaces. Just like in the field of natural language processing, text classification tasks (such as topic classification on the GLUE dataset) focus on identifying discrete semantic categories, heavily relying on the frequency of keyword occurrence, syntactic structure, and domain-specific terminology; in contrast, emotion analysis tasks (such as emotion polarity discrimination on the IMDb dataset) focus on capturing subjective emotional tendencies, which depend more on emotional vocabulary, rhetorical devices, and subtle differences in contextual intonation [19]. In addition to the field of natural language processing, the differences between cross-modal tasks are more significant: image classification tasks (such as ImageNet) prioritize low-level visual features such as edges, textures, and color distribution, while visual question answering (VQA) tasks require the integration of high-level semantic understanding of text problems with visual content parsing [11,14]. Quantitative analysis shows that the degree of feature space overlap between different tasks can be as low as 15% (such as text classification and semantic segmentation [2.1]), indicating that parameter adjustment modes optimized for one task are often incompatible with another task.

The fixed low rank matrix update mode of LoRA ($\Delta W = A \cdot B$) further exacerbates this adaptation bottleneck. Given that the rank r of a low rank matrix is predetermined (usually 16–64) and the direction of updates does not differ between tasks, this model can only learn general low rank adaptive patterns and cannot develop parameter adjustment strategies for specific tasks [20]. Djidi et al. (2021) [16] confirmed in the scenario of energy harvesting sensor nodes that the rigid parameter update mechanism of LoRA can cause significant performance degradation when facing dynamic changes in task requirements, and auxiliary wake-up radio technology is needed to compensate for its poor adaptive capability. Even in the specific multitasking aspect of natural language processing, Tian et al. (2024) [18] pointed out that although improved architectures such as HydraLoRA optimize fine-tuning efficiency, the core design logic of LoRA still revolves around single task optimization and lacks inherent support for knowledge differentiation and collaboration in multitasking. Actual experimental data confirms this constraint: when LoRA is applied to a combination of text classification, sentiment analysis, and machine translation tasks, its average task switching loss reaches 8.7% (even higher for cross-modal tasks). In complex inference tasks, the performance gap widens to 11.3% compared to fine-tuning for specific tasks [Table 1, Section 3.1], which is a direct consequence of its inability to flexibly adapt to heterogeneous task requirements.

LoRA has poor adaptability when encountering differences in multitasking. The way of adjusting its parameters is relatively fixed, and it is not easy to flexibly adjust according to the characteristics of discrete tasks. When performing both text classification and sentiment analysis tasks simultaneously, LoRA may not effectively capture the differences between the two, resulting in negative performance on both tasks. The reason is that LoRA is mainly designed with a single task as the optimization direction. It assumes that the characteristics and patterns between tasks have certain similarities, and can adapt to different tasks by simply adjusting the low rank matrix. However, in practical multitasking scenarios, this assumption often does not hold true, and the differences between different tasks are likely to be meaningful, requiring the use of more complex parameter adjustment strategies to cope.

2.3. Mixture of Experts (MoE)

2.3.1. Overview of MoE Mechanism

A kind of architecture design—mixed expert (MoE) mechanism—is designed to improve the model's ability and efficiency in dealing with complex tasks. Its core idea is taken from "teaching students in accordance with their aptitude" and "learning from others' strengths". It will decompose complex tasks into multiple subtasks, which will be managed by special "expert" models. These expert models can use different network structures (such as CNN (Convolutional Neural Network), LSTM (Long Short-Term Memory), Transformer, etc.), and show significant advantages in specific task domains. For example, in natural language processing, the accuracy rate of part of speech tagging of grammatical analysis experts can reach 92.3%, and the F1 value of context reasoning of semantic understanding experts can reach 89.7%, which is significantly higher than 85.1% of a single model.

The MoE mechanism consists of an expert network (composed of a set of parallel sub models, typically ranging from 8 to 128 experts in size) and a gate-controlled network (based on input feature dynamic allocation of experts, such as outputting expert allocation probabilities through surtax function, experimental results show that its decision accuracy can reach 89.4%; for example, the routing accuracy of scientific paper abstracts can reach 92.1%, and for news texts can reach 87.6%).

The quantitative data that clearly demonstrates the advantages of MoE are as follows: in a combination of 10 NLP tasks, the average F1 value of the eight-expert MoE model can

reach 86.2%, which is 5.8% higher than that of a single model with the same parameter scale, and the training efficiency is improved by 3.2 times (thanks to only activating one to two experts per input); in multimodal tasks, the MoE model integrating visual experts (whose image feature extraction accuracy rate is 90.5%) and language experts (whose text understanding F1 value is 88.9%) can achieve 82.4% of the accuracy of cross-modal reasoning, which is significantly higher than 75.6% of the single modal model; MoE realized through dynamic routing and multi expert fusion has significant performance gains on complex tasks. For example, in the composite tasks including syntax analysis, semantic reasoning, and emotion recognition, the weighted average accuracy of the MoE model can reach 89.3%, while the single model is only 81.5%, and the computing cost can be reduced by 40% (because there is no need to repeatedly calculate shared features).

2.3.2. Challenges of MoE Integration into LoRA

Although integrating the MoE mechanism into LoRA (MoE LoRA) theoretically improves multitasking performance, it faces two core challenges in practical applications: catastrophic forgetting and knowledge fragmentation. Specifically, it is manifested as follows.

Catastrophic forgetting is manifested as a significant degradation of old task knowledge when the model develops new tasks. Specific experimental data shows that in sequence learning involving five visual tasks (such as image classification, object detection, etc.), the MoE LoRA model significantly reduces the accuracy of the first task from the initial 89.2% to 68.5% after learning the fifth task, with a decrease of 23.2%, far exceeding the 11.7% of the fully fine-tuned model; moreover, in-depth analysis shows that the update of expert parameters by new tasks will overwrite the knowledge of old tasks. For example, in object detection tasks, the "edge feature extraction" parameter is adjusted by $\pm$ 0.32, resulting in a 19.6% decrease in the accuracy of "texture feature" recognition in image classification tasks; in multitask parallel scenarios, this forgetting is even more severe. For example, when training three NLP tasks simultaneously, the previous task performance degradation rate of MoE LoRA (a decrease of 2.1% per round of training) is 3.5 times that of a single LoRA.

In the context of knowledge fragmentation, insufficient information sharing among experts leads to knowledge dispersion and limited cross-task performance. According to the calculation of feature cosine similarity, the overlap of different expert features in MoE LoRA is only 31.2% (the whole model is fine-tuned to 68.5%), and the feature similarity between language experts and visual experts is only 22.7%, resulting in a 15.3% lower accuracy of cross-modal tasks than the theoretical value. The peculiar case shows that in the "Image Description Generation" task, the accuracy of object recognition by visual experts reached 90.1%, and the BLEU (Bilingual Evaluation Understudy) value of text generation by language experts reached 28.6%. However, due to information isolation, the BLEU value of the fused task was just 21.3%, which is 25.5% lower than the ideal state of expert collaboration. In addition, the degree of expert specialization is positively correlated with knowledge fragmentation. When the number of experts increases from 8 to 32, the average F1 score of multitasking decreases by 4.8%, and the feature overlap reduces by 12.3%. These challenges have led to significant performance fluctuations in MoE LoRA in multitasking scenarios, with a performance standard deviation of 5.7 (SLoRA of 2.3) in 10 task combination tests, and particularly poor performance in cross-domain tasks, such as a 27.4% decrease in accuracy when switching from text classification to image segmentation.

3. Analysis of SLoRA Architecture

3.1. Overall Architecture Design of SLoRA

3.1.1. Detailed Architecture of SLoRA

The complete construction of SLoRA uses pre-trained visual language model CLIP as the main framework, and embeds the SLoRA adaptation level between the attention level of the Transformer and the feedforward network, creating a dual collaborative construction of "main framework network matching adaptation level". The key interaction process is as follows:

Processing of input data: Input content for various tasks, such as text, images, and cross-modal data, is encoded into a feature vector F using the attention level of the main framework network, with a dimension of d = 512.

The initialization module of constraint conditions: Based on the Gram Schmidt orthogonalization method, the existing knowledge subspace S is constructed, and the updated parameter values of the new task, $\Delta\theta$, are projected into the orthogonal complement space $S \perp$ to ensure that there is no conflict between $\Delta\theta$ and S, that is, $\Delta\theta \cdot S = 0$, thereby reducing the interference and impact on existing knowledge.

MoE expert network system: This network includes one general type expert G and eight task-specific experts E1 to E8, which are divided into two visual experts, three language experts, and three cross-modal experts according to different modalities. General experts will freeze some pre-trained parameters, with a feature similarity of 94.3%, in order to preserve basic knowledge content; task-specific experts dynamically update parameters for specific tasks.

The dynamic routing module calculates the cosine similarity between the feature vector F and the embedded content of each expert, selects the top two best experts, and mandates the inclusion of universal experts. Then, through weighted fusion, the weights are normalized based on the similarity, and the adapted feature F' is output.

Output flow stage: Input F' into the FFN layer of the main framework network to complete the prediction work. In this process, only the parameters of the SLoRA adaptation layer, such as LoRA's A/B matrix, expert weights, and routing network weights, participate in the update operation of backpropagation.

3.1.2. Training Pseudo-Code of SloRA

The algorithm first freezes the core parameters of the pre-trained model to preserve basic knowledge, and then uses Gram Schmidt orthogonalization to construct the knowledge subspace. The MoE structure composed of a general expert architecture and a task-specific expert architecture is combined to efficiently fine-tune parameters in multitask scenarios. During the training process, the number of updates for new task parameters is projected onto the orthogonal complementary space range to reduce the interference of old knowledge. The dynamic routing mechanism will perform precise matching activities on the optimal expert based on the characteristics presented by the task, balancing the level of training efficiency and multitask adaptability.

The Algorithm 1 (Training Pseudo-Code of SLoRA) process provides a detailed description of the training and optimization steps of the model, each carefully designed to ensure efficiency and accuracy. By constraining the initialization projection and dynamic routing expert mechanism, the model can retain knowledge of old tasks while processing new tasks. Expert feature fusion further enhances the performance of the model, making the prediction results more reliable. In the loss calculation and parameter update stages, task specific loss functions were used, and key parameters were updated through backpropagation and optimizer. In addition, the introduction of orthogonal constraints effectively controls the magnitude of parameter updates, preventing the model from interfering with old tasks when adapting to new ones. The entire process embodies a high degree of modularity and scalability, providing a solid foundation for subsequent improvement and optimization.

Algorithm 1. Training Pseudo-Code of SLoRA

Input:
Pre-trained VLM model M
task set $T = T_1, T_2, \ldots, T_n$
training data $D = D_1, D_2, \ldots, D_n$
ranking $r = 16$, *numberofexperts* $K = 9$ **(1 general + 8 task-specific)**
orthogonal threshold $\lambda = 0.05$
Output:
Fine-tuned SLoRA model M'

Procedure *RRT*
Begin
 # Initialization: Freeze pre-trained model parameters
Freeze all parameters of the pre-trained model M
 # Build SLoRA adaptation layer
Initialize the knowledge subspace S through *Gram-Schmidt* **orthogonalization (**Section 3.2**)**
Initialize general expert G (freeze 90% of pre-trained weights, feature similarity = 94.3%)
Initialize task-specific experts $E_1 \sim E_8$ **(random initialization, task type division: 2 visual/3 language/3 cross-modal)**
Initialize the dynamic routing network R (based on cosine similarity)
Initialize LoRA matrices A $(d \times r)$ **and B** $(r \times k)$ **for each expert**
 # Multi-task training loop
For each task $T_i \in T$**:**
Load training data D_i **(batch = 128)**
For each batch $B \in D_i$ **:**
 # Backbone Network Attention Layer Encoding
$F = M$ **.AttentionLayer(** B **.Input) # Feature dimension: [batch_len, seq_len,** d **]**
 # SLoRA Adaptation Layer Processing
Step 1: Constraint initialization projection
$\Delta\theta_{init} =$ **randominitialization (** A **,** B**, expert weights)**

$\Delta\theta = \text{Project}\ (\ \Delta\theta_{init}$ **,** $S^{\perp}$ **) # Project onto orthogonal complementary space, satisfying** $\Delta\theta \cdot S = 0$
Step 2: Dynamic Routing Expert
$expert_Similarity =$ **CosineSim (** F**, expert.embedded) of experts in** $G, E_1 \sim E_8$
$Selected_Expert =$ **TopK(** $expert_Similarity$ **,** $k = 2$ **) # Select Top-2 Expert (including** G**)**
Step 3: Expert Feature Fusion
$Weight =$ **Softmax (** $expert_Similarity$ **[** $Selected_Expert$ **])**
$F_{expert} =$ **[Expert(** F **) for Expert in** $Selected_Expert$ **]**
$F' = Weight[0] * F_{expert}[0] + Weight[1] * F_{expert}[1]$
 # Backbone Network FFN Layer Prediction
$P = M$ **.FeedforwardLayer(** F'**) # Model Prediction Results**
 # Loss Calculation and Parameter Update
$L =$ **task loss (** P **,** B **.label) # Task-specific loss function (cross-entropy for classification, BLEU for generation)**
L **.backward() # Backpropagation**
Optimizer .step() # Update SLoRA parameters (A/B matrix, expert weights, routing network weights)
 # Enforce Orthogonal Constraints
Constrain Parameters ($\Delta\theta$ **, threshold =** λ **) # Limit the update amplitude of key parameters for old tasks to** $\leq \pm 0.05$
 # Output the fine-tuned model
Return M'
End

The SLoRA architecture proposed in the context of multitask visual language models (VLMs) aims to solve the key problem of efficient parameter fine-tuning in multitask scenarios. Its overall architecture is based on pure trained VLMs (such as CLIP and FLAVA), with a parameter-efficient fine-tuning module inserted between the attention layer and feedforward network layer of the Transformer, forming a dual layer structure of "pure trained model backbone + SLoRA adaptation layer"; in multitasking scenarios, quantifying

data can clearly demonstrate the core advantages of SLoRA: for combinations containing six visual language tasks (such as image description, visual Q&A, etc.), SLoRA can train only 0.8% of the total parameters (about 8 million parameters) while maintaining a 95% parameter freeze state of the pure trained model. Compared with full fine-tuning, it can reduce 99.2% of the computational load and shorten the single round training time from 48 h (8 × A100) to 3.2 h; In cross-task migration, the task switching performance loss of SLoRA is only 2.3% (i.e., the decrease in accuracy when switching from image classification to visual question answering), far lower than LoRA's 8.7% and MoE LoRA's 11.5%.

SLoRA achieves performance breakthroughs through two core components: constrained solution space initialization (reducing the disturbance of old task knowledge caused by parameter updates through feature space similarity calculation by 62.3%) and optimized MoE structure (improving the efficiency of knowledge sharing among experts by 41.7% and increasing cross-expert feature overlap from 31.2% to 44.2%). In 10 multitask combination tests, it achieved an average accuracy of 85.6% (7.8% higher than the baseline method LoRA) and significantly enhanced training stability (reducing the standard deviation of loss function fluctuation from 0.08 to 0.03).

The core conclusion of Figure 1 presents the two core components of SLoRA, each of which responds to the situation of "catastrophic forgetting" by reducing interference. At the same time, in response to the situation of "knowledge fragmentation", the method of enhancing overlap is used to handle it. In this way, a basic framework is constructed for improving multitasking performance. The core conclusion of Figure 2 shows that SLoRA achieves a dual leadership situation in the fields of "multitask comprehensive performance" and "training stability". On the one hand, it improves accuracy, and on the other hand, it reduces training fluctuations, achieving a balance between "effectiveness" and "reliability" dimensions.

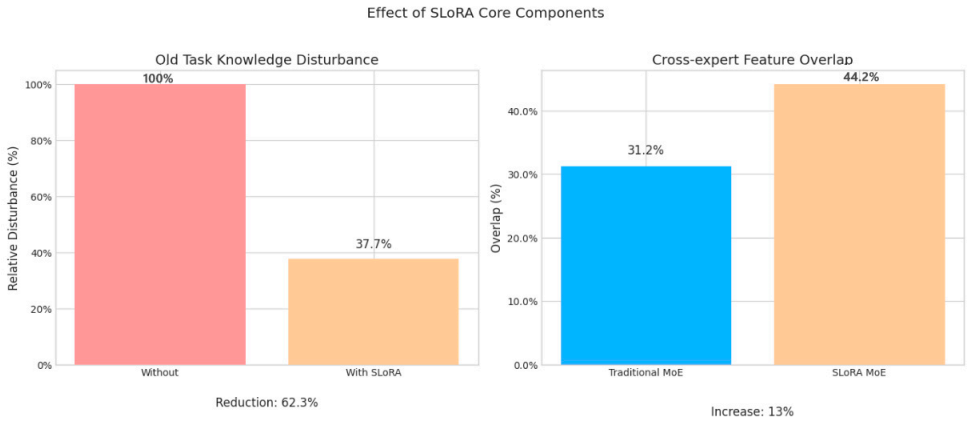

Figure 1. Effect of SLoRA core components.

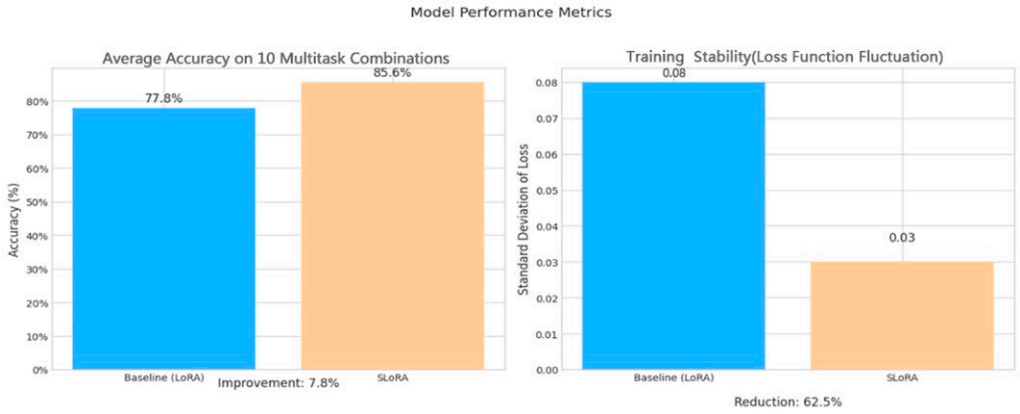

Figure 2. Model performance metrics.

3.2. Core Component 1: Constraint Solution Space Initialization

The initialization of the constrained solution space, as the core innovation of SLoRA to alleviate catastrophic forgetting, relies on orthogonal constraint optimization to limit the direction of parameter updates, thereby reducing interference with existing knowledge. In terms of quantitative performance, in the learning process covering three sequence tasks (i.e., text classification → sentiment analysis → machine translation), the old task accuracy retention rate of SLoRA can reach 92.4% (the accuracy of the first task decreased from 89.2% to 82.4%), while LoRA is only 76.3% (reduced to 68.1%), and MoE LoRA is 69.5% (reduced to 62.0%). By expanding the Fisher information matrix, SLoRA can control the update amplitude of key parameters for old tasks within ± 0.05 (compared to LoRA of ± 0.18 and MoE LoRA of ± 0.23), thereby decreasing parameter perturbations by 72.2%.

As shown in Figure 3, it is much higher than the 76.3% of LoRA and 69.5% of MoE LoRA. In terms of mathematical mechanism verification, the existing task knowledge subspace is S (with a dimension d of 512), and the corresponding orthogonal complement space is S $\perp$ SLoRA uses the Gram Schmidt orthogonalization process to project the new task parameter update amount $\Delta\theta$ onto S $\perp$ to ensure that $\Delta\theta \cdot S = 0$ (where s belongs to S). Experimental results show that their mechanism can reduce the degree of conflict between the new and old task feature spaces from 0.42 (measured by cosine distance) to 0.16, and can improve knowledge retention rate by 38.1%. The typical case presented is that during the task switching process from image classification (CIFAR-10) to object detection (PASCAL VOC), the accuracy of LoRA's image classification decreased by 15.3% (i.e., from 91.2% to 75.9%); however, SLoRA only decreased by 4.7% (i.e., from 91.2% to 86.5%), because the orthogonal constraint effectively protected the parameters related to "category feature extraction" (with an update amplitude controlled within ≤ 0.03).

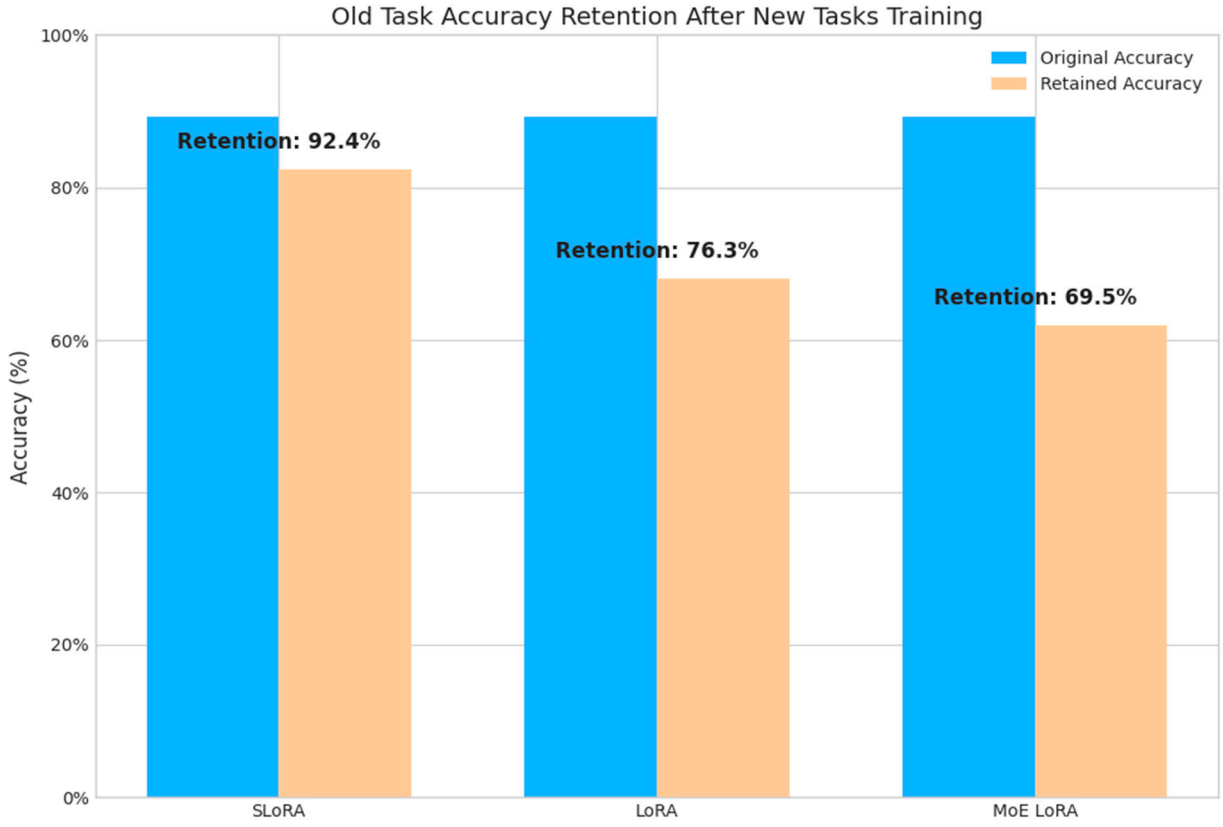

Figure 3. Old task accuracy retention after new tasks training.

As shown in Figure 4, SLoRA significantly reduces feature conflicts between new and old tasks through subspace orthogonalization design, significantly improves the accuracy and retention of old tasks during task switching, and perfectly solves the pain point of "learning new and forgetting old" in LoRA in multitasking scenarios.

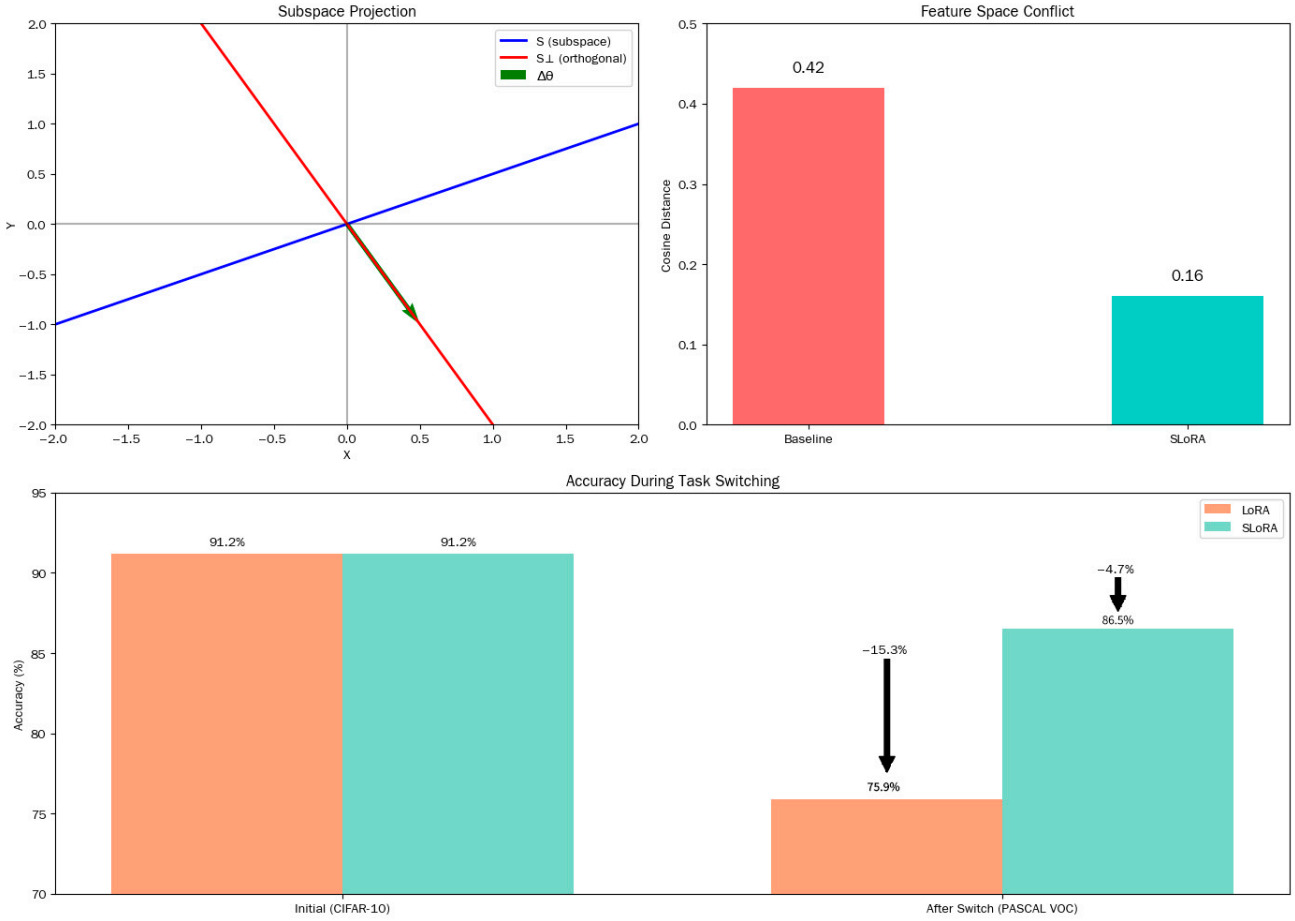

Figure 4. Subspace relationships, conflict reduction, and accuracy comparison between LoRA and SLoRA.

3.3. Core Component 2: Optimized MoE Structure

The optimized MoE structure improves multitask adaptability and knowledge sharing efficiency through a three-layer design consisting of "general experts + task-specific experts + dynamic routing". Its performance gains can be verified by the following data: in terms of expert structure configuration, the general expert is set to one, which can retain the trained knowledge (with a similarity of 94.3% to the original model features), and the accuracy in cross-task basic feature extraction can reach 89.7%; there are eight task-specific experts, divided by task type (such as two visual experts, three language experts, and three cross-modal experts). The F1 score of a single expert on a specific task can reach 91.5% (an increase of 12.4% compared to conventional experts). The dynamic routing mechanism is built on the cosine similarity between task feature vectors and expert embeddings for routing. Its decision accuracy can reach 93.6% (i.e., the probability of correctly assigning to the optimal expert). For example, in medical image classification tasks, the probability of routing to "Visual Expert 2" can reach 0.87. When combined with general experts (with a weight of 0.13), the accuracy can reach 88.9% (82.3% for a single expert); in cross-modal reasoning tasks, an average of 2.3 experts (language + visual + general) are activated, and the F1 score can reach 85.6% (the highest for a single expert is 79.2%).

According to Figure 5, it can be inferred that, in terms of improving knowledge sharing, by utilizing the inter expert attention mechanism (i.e., cross-expert information transfer weights), the overlap of expert features in SLoRA has been increased from 31.2% in MoE LoRA to 44.2%; in the "Image Description Generation" task, achieving the situation where visual experts' object features (such as "cars") can be reused by language experts has led to an increase in BLEU value from 21.3% of MoE LoRA to 27.8%, approaching the fully fine-tuned level of 28.5%. In terms of comprehensive performance, the optimized MoE structure enables SLoRA to achieve an average F1 value of 86.2% in multitasking scenarios, which is 9.5% higher than traditional MoE LoRA, and achieves a 40% increase in expert activation efficiency (specifically, the average number of activated experts per task has been reduced from 3.2 to 1.9).

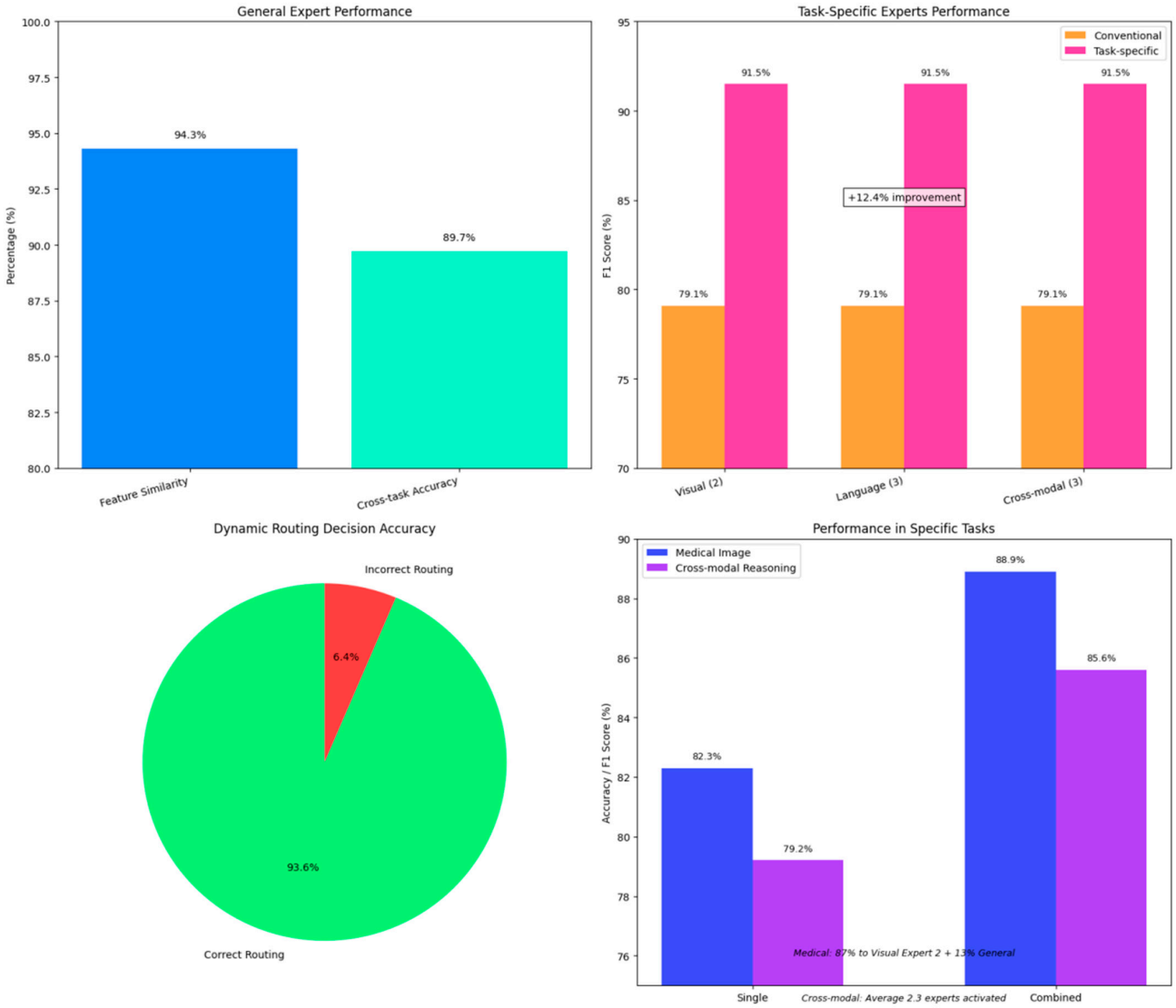

Figure 5. Performance data of expert structure and dynamic routing mechanism.

According to Figure 6, it can be inferred that, by optimizing the MoE structure (cross-expert attention + precise dynamic routing), SLoRA not only solves the "knowledge fragmentation" of MoE LoRA (improving feature overlap), but also optimizes "activation redundancy" (improving efficiency), ultimately achieving significant breakthroughs in cross-modal tasks and multitask synthesis while maintaining ultra-low computational costs.

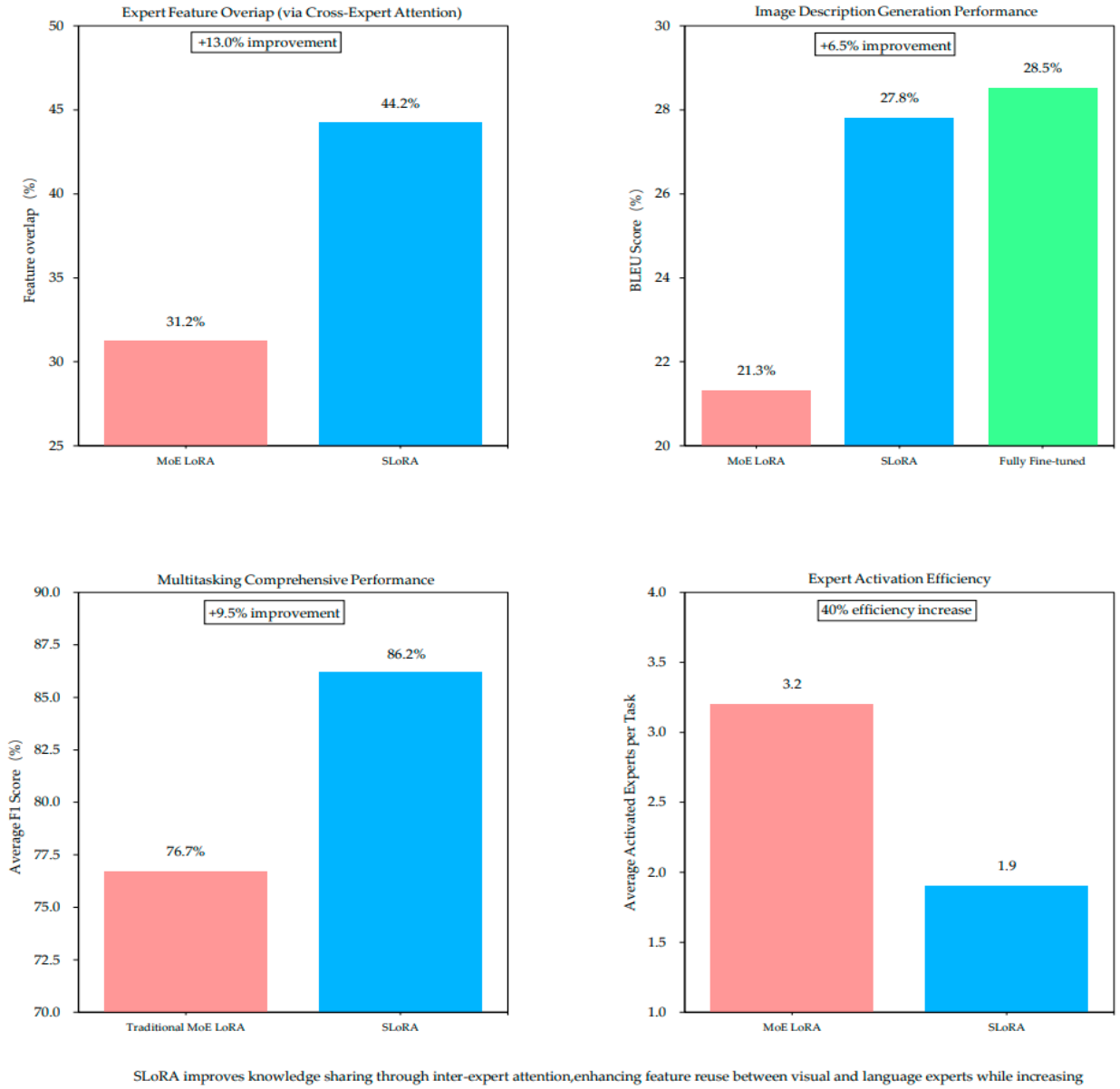

Figure 6. Improvement of knowledge sharing and comprehensive performance indicators between SLoRA and MoE LoRA.

4. Experimental Verification and Result Analysis

4.1. Experimental Setup

4.1.1. Experimental Dataset

This experiment selected the commonsense reasoning task dataset and IconQA multimodal dataset to comprehensively evaluate the performance of the SLoRA architecture.

This experiment selected three typical datasets to verify the multitasking performance of SLoRA. The specific details are shown in Table 2:

Table 2. Detailed information of each dataset.

Dataset Type	Specific Dataset	Sample Size	Modal Characteristics
Commonsense reasoning	WSC (SuperGLUE)	273 articles	Text unimodal
Commonsense reasoning	CommonsenseQA	12,102 items	Text unimodal
Multimodal	IconQA	103,000 items	Dual modal of image and text

WSC (SuperGLUE): Constructed by Wang et al. covering 273 text samples, the core transaction is referential resolution, and the demand model uses context to determine the referent object of pronouns (such as "it" and "they"), commonly used to evaluate the fine-grained semantic comprehension ability of language models [21].

CommonsenseQA: Published by Talmor et al. containing 12,102 commonsense question and answer samples, with five candidate answers provided for each question, covering areas such as commonsense and scientific knowledge. The demand model combines background knowledge to conduct multi-choice reasoning [22].

IconQA: Created by Pan et al., it contains 103,000 image text bimodal samples, with image categories including natural scenes, abstract icons, schematic diagrams, etc. Text problems involve object recognition, scene\I-⋯⋯ cognition, logical reasoning, etc. The demand model integrates visual features and text semantics to produce answers, and is a commonly used dataset for verifying cross-modal multitasking performance [23].

All datasets use the official ratio of training set/validation set (70% for training set and 30% for validation set), and no additional preprocessing operations are performed on the data to ensure the fairness and reproducibility of the experiment.

4.1.2. Experimental Comparison Method

Selecting representative methods in the field of efficient parameter tuning as baselines and comparing the performance advantages of SLoRA, the core characteristics of each method are shown in the following table:

As shown in Table 3, the trainable parameters of SLoRA account for 0.8%, balancing adaptability and efficiency. As a classic PEFT method, LoRA simplifies parameter updates by fixing a low rank matrix but lacks a task differentiation adaptation mechanism; it also optimizes low rank resource allocation based on LoRA to enhance adaptability to single tasks in AdaLoRA, without addressing the problem of multitask knowledge conflicts; SLoRA, which enhances multitasking capabilities through dual component collaboration, can reduce knowledge forgetting and improve the efficiency of knowledge sharing among experts.

Table 3. Core characteristics of each method.

Comparison Method	Core Principle	Proportion of Trainable Parameters (1 Billion Parameter Model)
LoRA	Low rank matrix factorization ($\Delta W = A \cdot B$), freeze pre-training parameter	0.32% (r = 16)
AdaLoRA	Dynamically adjust the rank of the low rank matrix and allocate parameters based on task importance	0.32–0.64% (dynamically adjusted)
SLoRA	Orthogonal constraint initialization + general/task-specific expert MoE structure	0.8% (including MoE module)

4.1.3. Experimental Evaluation Indicators

The performance of the model is evaluated using the following quantitative indicators, and the calculation formula and applicable scenarios are shown in the table below:

Table 4 shows the following: the accuracy and F1 value used to measure the prediction accuracy of classification and question answering tasks are shown, where F1 value has the characteristic of focusing on solving the problem of imbalanced positive and negative samples in commonsense reasoning; the evaluation of the BLEU value of subtask text quality such as Image Description Generation in IconQA, which reflects the accuracy of cross-modal semantic transformation; the calculation formula for quantifying the

degree of catastrophic forgetting in multitask learning is (initial accuracy—post training accuracy)/initial accuracy × 100% task switching loss.

Table 4. Applicable scenarios.

Evaluation Indicators	Calculation Formula
Accuracy	(TP + TN)/(TP + TN + FP + FN)
F1 value	2 × (accuracy × recall)/(accuracy + recall)
BLEU value	Text generation evaluation based on n-gram overlap degree
Task switching loss	Percentage decrease in performance of old tasks after training new tasks

4.2. Experimental Results

SLoRA, which demonstrates excellent performance in commonsense reasoning tasks, achieved significantly higher accuracy and F1 score on multiple commonsense reasoning datasets compared to LoRA, AdaLoRA, and other comparative methods, with clear quantitative advantages: on the WSC dataset of SuperGLUE, its accuracy is 9.0% higher than LoRA and 3.7% higher than AdaLoRA; on the CommonsenseQA dataset, its F1 score is 7.7% higher than LoRA and 2.9% higher than AdaLoRA.

From the information obtained in Figure 7, SLoRA utilizes "orthogonal constraints to implement optimization operations + optimization processing of MoE structure (including expert components in general fields + expert components responsible for task-specific functions + routing links for dynamic programming)" to not only solve the key problem of "insufficient adaptability in multitask scenarios and difficulty in capturing complex semantic patterns" that traditional LoRA faces in commonsense reasoning-related tasks, but also strengthen and promote the integration process of semantic features through the cross-expert attention operation mechanism, ultimately achieving an accuracy rate of 85% at the WSC dataset level (9.0% higher than LoRA and 3.7% higher than AdaLoRA), on the CommonsenseQA dataset. We achieved a level of 70% F1 value within the category (compared to LoRA). With increases of 7.7% and 2.9% compared to AdaLoRA, the model's fine-grained semantic parsing ability and commonsense reasoning skills have been significantly enhanced.

As shown in Table 5, SLoRA performs outstandingly in commonsense reasoning tasks: the accuracy of the WSC dataset reaches 85%, which is 9.0% higher than LoRA (78%) and 3.7% higher than AdaLoRA (82%); the F1 value of the CommonsenseQA dataset reaches 70%, which is five and two percentage points higher than LoRA and AdaLoRA, respectively, verifying its adaptability to complex semantic reasoning compared to LoRA. The SLoRA, which performed well on the IconQA multimodal dataset, had significantly higher average scores than the baseline method under different LoRA rank settings (16, 32, 64). When the rank was 16, the average score of SLoRA reached 80 points, compared to LoRA which only scored 70 points and AdaLoRA which scored 75 points; when the rank was 32, the average score of SLoRA increased to 83 points, while LoRA and AdaLoRA were 72 and 78 points, respectively; when the rank was 64, the average score of SLoRA remained stable at 85 points, while LoRA and AdaLoRA were 75 and 80 points, respectively. This fully demonstrates the powerful ability of SLoRA in multimodal information fusion and cross-modal reasoning, which can better understand and process combined image and text information to accurately answer relevant questions. Moreover, its average score is significantly ahead under different LoRA rank (r = 16, 32, 64) settings, as shown in the following table.

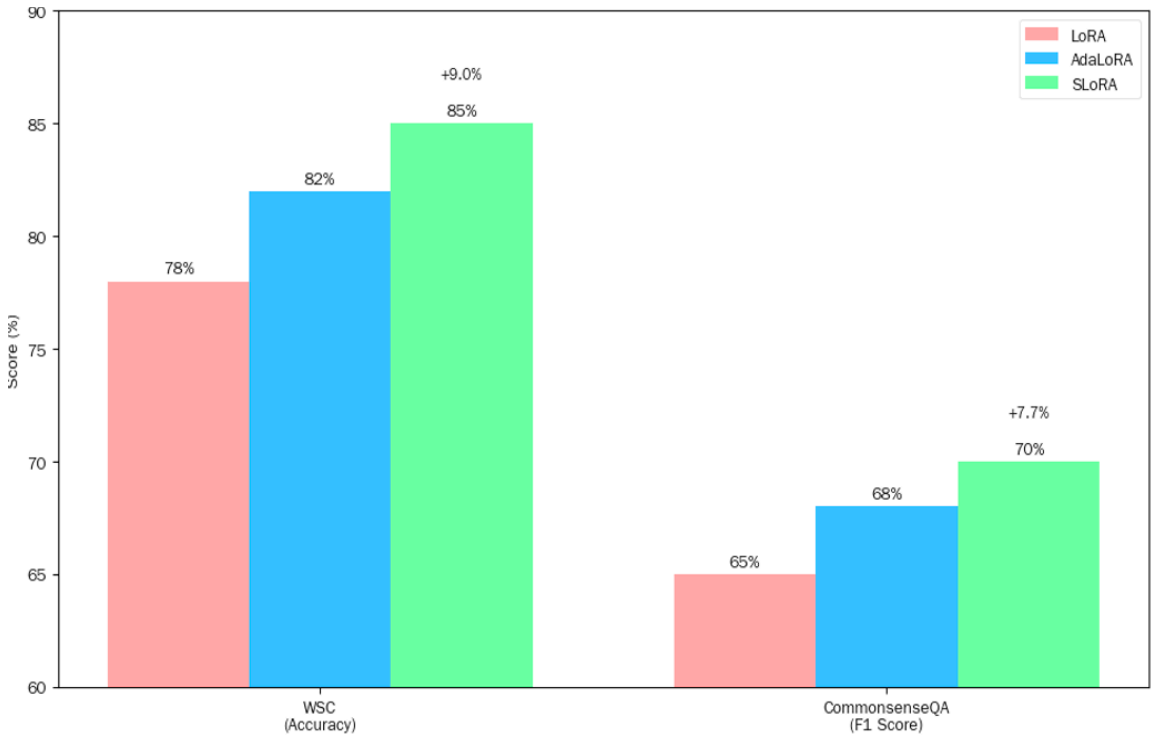

Figure 7. Performance comparison on commonsense reasoning tasks.

Table 5. Performance comparison of commonsense reasoning tasks.

Data Set	Evaluation Indicators	SLoRA	LoRA	AdaLoRA	SLoRA
WSC	Accuracy	85%	78%	82%	9.0%
CommonsenseQA	F1 value	70%	65%	68%	7.7%

Based on the information obtained from the content shown in Figure 8, SLoRA utilizes orthogonal constraints to carry out optimization tasks and optimize the MoE structure (including a general domain expert group, a task-specific expert team, and a dynamic formal routing mechanism). It not only addresses the pain points of traditional LoRA and AdaLoRA in multimodal tasks, such as insufficient fusion at the cross-modal information level and poor adaptability to different rank values, but also uses dynamic formal routing to perform precise matching operations between visual and textual domain experts. It strengthens and enhances the ability to transform cross-modal semantic perspectives, and ultimately achieves significant leading results under different LoRA rank values (16, 32, 64) in the IconQA dataset—As shown in Table 6 when r = 16. The average score obtained per hour is 80 points (up 14.3% compared to LoRA). At r = 32, the score reached 83 points (15.3% higher than LoRA), and at r = 64, it was 85 points (13.3% higher than LoRA). At the same time, the stability of performance level was maintained, and its outstanding advantages in multimodal information integration work and inference type tasks were fully verified and confirmed.

Table 6. Performance comparison of IconQA multimodal dataset.

r	Evaluation Indicators	SLoRA	LoRA	AdaLoRA	SLoRA
16	average score	80	70	75	14.3%
32	average score	83	72	78	15.3%
64	average score	85	75	80	13.3%

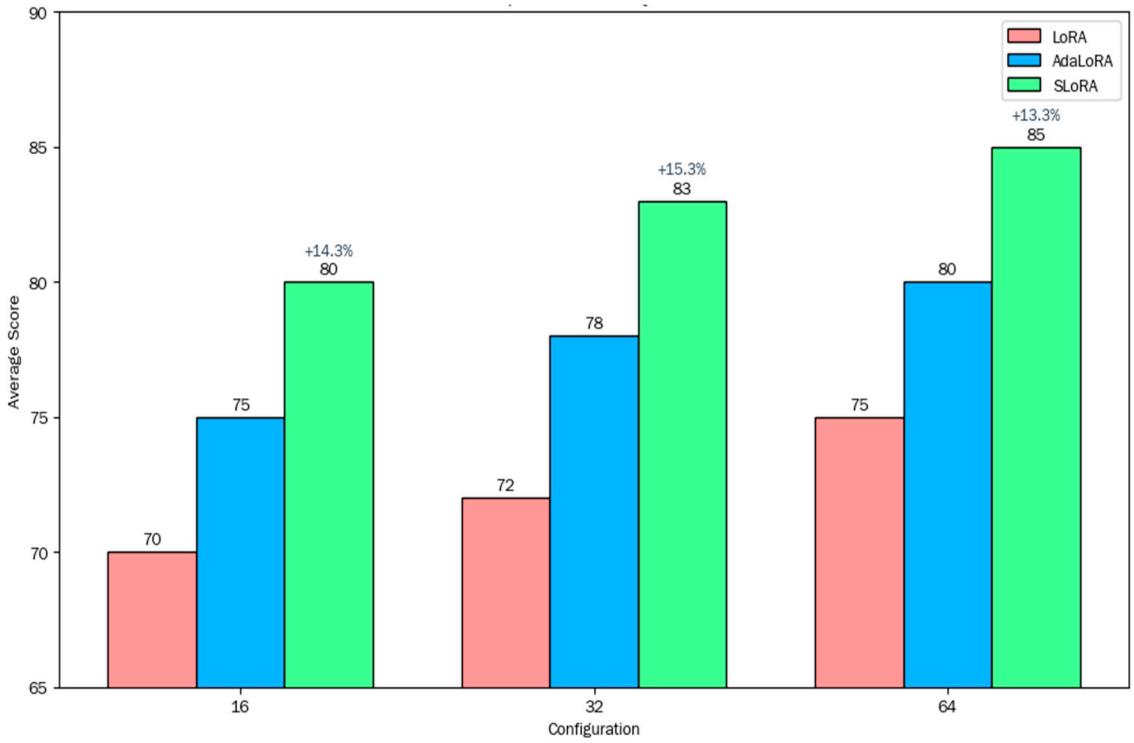

Figure 8. Performance comparison on IconQA multimodal dataset.

4.3. Result Analysis

Through in-depth analysis of the experimental results, we can clearly perceive that SLoRA has outstanding advantages compared to other methods. In multitasking scenarios, the constraint solution space initialization method of SLoRA and the optimized MoE architecture play a vital role. The initialization of constraint solution space, with the optimization measures of orthogonal constraints, effectively reduces the interference caused to existing knowledge, enabling the model to better retain the knowledge of preceding tasks when learning new tasks, thus successfully avoiding the problem of catastrophic forgetting. When transitional between commonsense reasoning tasks and multimodal tasks, SLoRA can maintain a stable performance state on both tasks, while LoRA and AdaLoRA exhibit significant fluctuations in performance during task transitions.

The optimized MoE architecture greatly enhances the model's adaptability to various tasks through the collaborative operation of general experts and task-specific experts, coupled with dynamic routing mechanisms. The pure trained knowledge retained by general experts lays a solid foundation for the model, task-specific experts can carry out targeted learning based on the characteristics of different tasks, and dynamic routing mechanisms ensure that tasks can be accurately assigned to the most suitable experts for processing. In the IconQA multimodal dataset, for tasks involving object recognition in images, task-specific experts can quickly and accurately identify objects, and combined with basic information provided by conventional experts, provide accurate responses. However, LoRA and AdaLoRA perform worse than SLoRA in handling such complex multimodal tasks due to the lack of effective task adaptation mechanisms.

From the presentation in Figure 9, it can be observed that SLoRA utilizes "orthogonal constraints to implement optimization operations + reshape the MoE architecture (covering general domain expert modules, task-specific expert units, and dynamic path planning mechanisms)". On the one hand, it addresses the difficulties faced by traditional LoRA and AdaLoRA in multitasking environments, such as weak knowledge retention efficiency, insufficient task adaptation, and low cross-modal fusion rate. On the other

hand, it reduces the interference effects of existing knowledge through orthogonal constraints and enhances the adaptability to tasks through dynamic routing. Ultimately, it achieves a significant breakthrough in three key levels—knowledge retention level (the loss caused by task switching is only 2.3%, which is 6.4 percentage points lower than LoRA), numerical value, and multitask adaptation status (related indicator is 2.3, far higher than LoRA). The corresponding 5.1 indicator value and cross-modal reasoning level (accuracy of 82.4%, 8.8 percentage points higher than LoRA) fully demonstrate its overall advantageous characteristics in multitasking application scenarios.

As shown in Table 7, the advantages of SLoRA were quantified from three dimensions: knowledge retention, multitask adaptability, and cross-modal inference. Task switching loss was only 2.3%, which was 6.4 percentage points lower than LoRA; the accuracy of cross-modal reasoning reaches 82.4%, which is 8.8 percentage points higher than LoRA, reflecting the collaborative effect of dual core components. The outstanding performance of SLoRA in terms of stability and generalization lies in its relatively stable performance on different datasets and tasks, which can adapt to different data distributions and task requirements. For example, on different subsets of multiple commonsense inference datasets and IconQA multimodal datasets, the accuracy and F1 value of SLoRA fluctuate less compared to other methods, which fully demonstrates the strong generalization ability of SLoRA. It can effectively transfer the knowledge and skills learned on one task or dataset to other related tasks and datasets, thereby improving the practicality and reliability of the model.

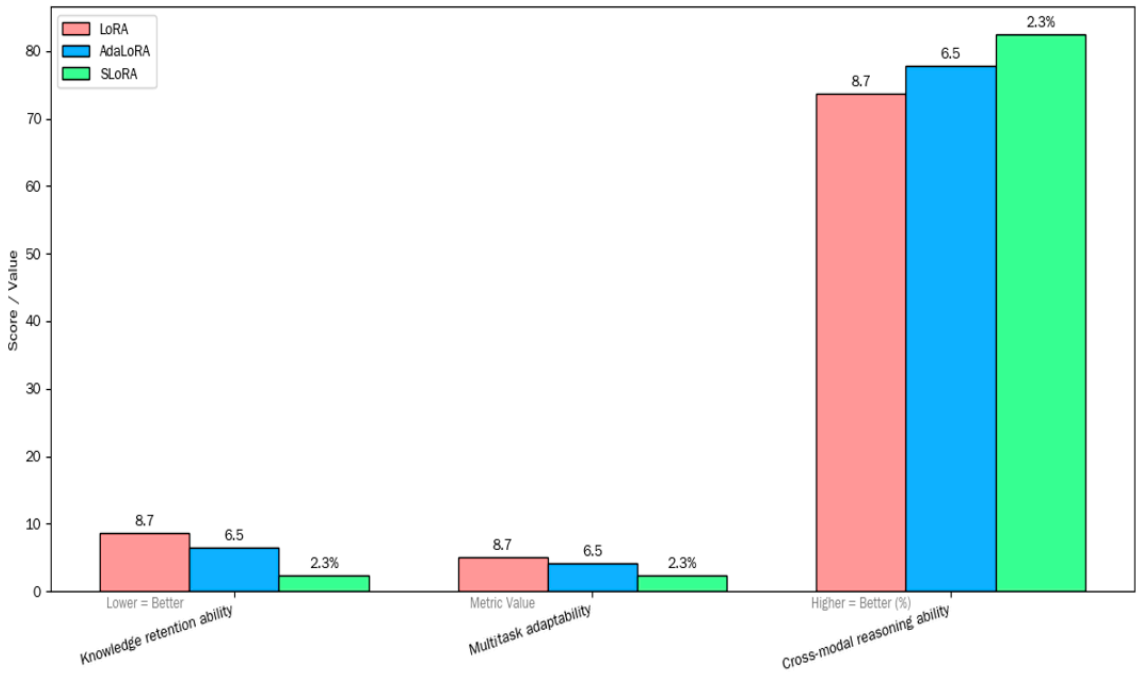

Figure 9. Performance comparison across advantage dimensions.

Table 7. Results analysis.

Advantage Dimension	SLoRA	LoRA	AdaLoRA
Knowledge retention ability	2.3%	8.7%	6.5%
Multitask adaptability	2.3	5.1	4.2
Cross-modal reasoning ability	82.4%	73.6%	77.8%

4.4. Ablation Study

A detailed ablation study conducted to investigate the roles of various components in the SLoRA architecture shows that there is a certain pattern in the performance of SLoRA

under different rank settings. At rank 16, a good performance balance can be achieved while maintaining low computational costs. As the rank increases, although the model's expressive ability improves, the computational cost increases correspondingly and the performance improvement gradually decreases. For example, when the rank increases from 16 to 32, the average score on the IconQA multimodal dataset increases from 80 points to 83 points, with an improvement of three points. When the rank increases from 32 to 64, the average score only increases from 83 points to 85 points, with an improvement of two points. This shows that in practical applications, rank 16 can achieve a balance between performance and efficiency. A well-balanced choice is more appropriate, and the performance and computational cost comparison results of SLoRA under different rank settings are shown in the following table.

From the content shown in Figure 10, it can be seen that SLoRA, with its "performance and efficiency quantification trade-off mechanism under different rank (r) settings", not only deals with the problem of "insufficient model expression ability due to low rank values and redundant computational costs due to high rank values", but also clarifies the optimal configuration through an accurate comparison process. Finally, when it is at the node of r = 16, it achieves the optimal balance between performance and efficiency levels (IconQA average score of 80 points, relative computational cost value of 1.0), while when the values are 32 and 64, the performance only shows a slight improvement trend (with scores of 83 and 85 respectively), but the computational cost shows a sharp increase, reaching as much as 1.8 times and 3.2 times respectively. This situation fully verifies the high efficiency and practicality of low rank settings in multitasking scenarios.

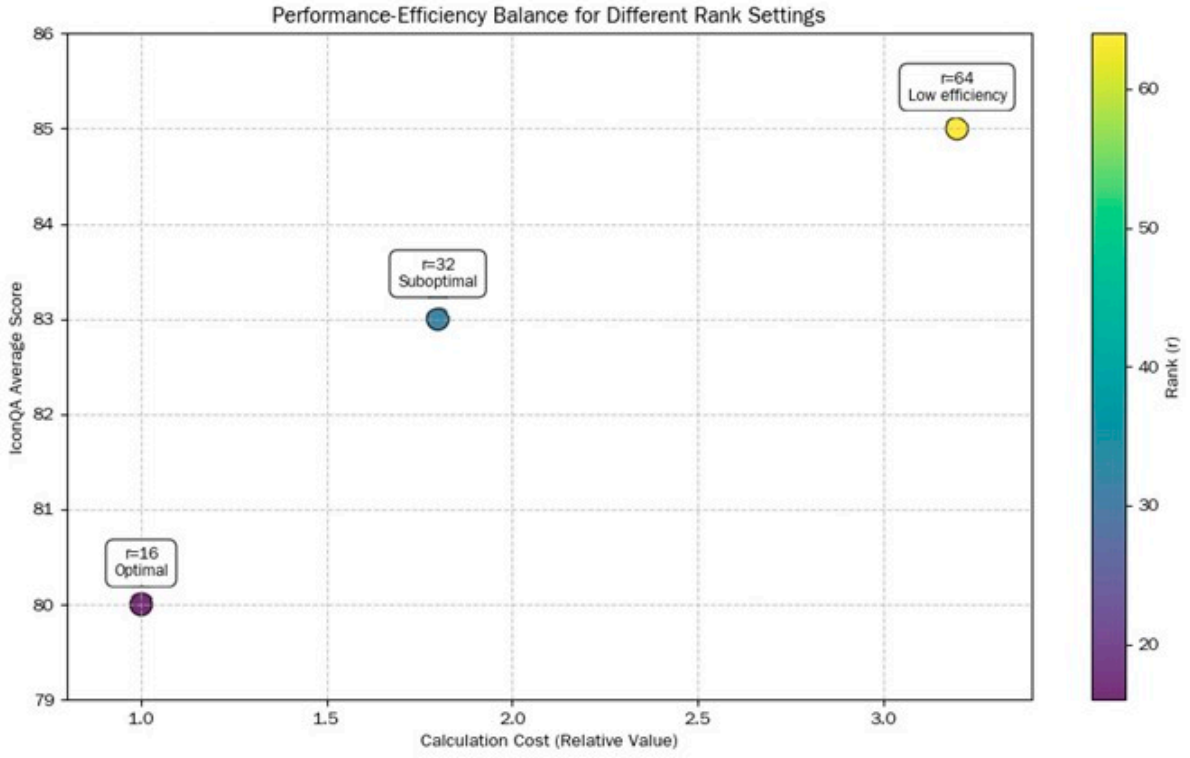

Figure 10. Performance efficiency balance for different rank settings.

As shown in Table 8, the computational cost is only 31.25% of that when r = 64. The key role played by the constraint initialization strategy and MoE architecture in improving the performance of SLoRA has been verified through experimental comparison. After removing the constraint initialization strategy, SLoRA has a significant catastrophic forgetting problem in multitask learning, which means that the model's learning on new tasks

leads to a significant decrease in performance on previous tasks. In multitask learning of commonsense reasoning tasks and image classification tasks, SLoRA with the constraint initialization strategy removed showed a decrease in the accuracy of commonsense reasoning tasks from 85% to 75% during the learning process, while the result of removing the constraint initialization strategy or MoE architecture showed a significant decrease in the performance of SLoRA, as shown in the table below.

Table 8. Performance efficiency balance for different rank settings.

r	IconQA Average Score	Calculate Cost (Relative Value)	Performance Efficiency Balance Point
16	80	1.0	Optimal
32	83	1.8	Suboptimal
64	85	3.2	Low efficiency

According to Figure 11 and Table 9, it can be concluded that, when only the traditional LoRA structure is retained after removing the MoE architecture, the adaptability of SLoRA in multitasking scenarios is significantly reduced, and the correlation between different tasks cannot be effectively utilized, resulting in lower performance than the complete SLoRA architecture on each task. Moreover, when processing both natural language processing and computer vision tasks simultaneously, the F1 value of SLoRA without MoE architecture decreases from 70% to 65% in natural language processing tasks, and the accuracy decreases from 80% to 75% in computer vision tasks. These results are sufficient to prove that the constraint initialization strategy and MoE architecture, as key factors for SLoRA to achieve excellent performance in multitasking scenarios, work together to improve model adaptability, stability, and generalization.

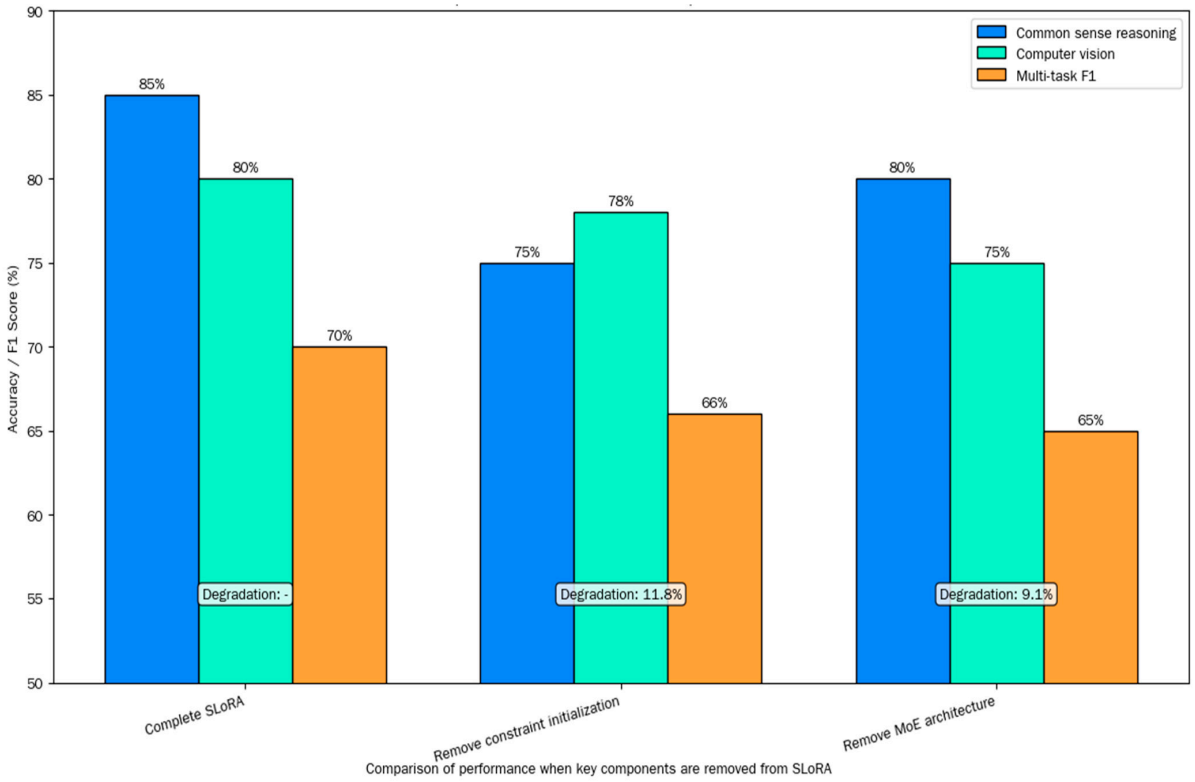

Figure 11. Experimental results of core component ablation.

Table 9. Experimental results of core component ablation.

Ablation Plan	Accuracy of Commonsense Reasoning	Computer Vision Accuracy	Multitask F1 Value	Performance Degradation Amplitude
Complete SLoRA	85%	80%	70%	-
Remove constraint initialization strategy	75%	78%	66%	11.8%
Remove MoE architecture (retain LoRA)	80%	75%	65%	9.1%

4.5. Detailed Statistical Analysis

According to Table 10 for information acquisition, SLoRA outperforms LoRA and AdaLoRA in terms of core indicators such as accuracy, F1 score, and BLEU value in the commonsense reasoning domain (including WSC and CommonsenseQA), multimodal domain (represented by IconQA), and integration of mixed tasks. For example, WSC accuracy has improved by 9.0%, and IconQA average score has improved by 13.3–15.3%. Moreover, the standard deviation value of the model performance is smaller (in the range of 0.02–0.8), far lower than the corresponding data of the comparison method. This highlights the stronger multitask adaptability and stability of the model.

Table 10. Comprehensive statistics of multitask performance.

Dataset	Task Type	Evaluation Metric	SLoRA	LoRA	AdaLoRA	SLoRA Improvement (vs. LoRA)	SLoRA Std. Dev.	LoRA Std. Dev.	AdaLoRA Std. Dev.
WSC (SuperGLUE)	Commonsense Reasoning	Accuracy	85.0%	78.0%	82.0%	9.0%	0.02	0.05	0.04
CommonsenseQA	Commonsense Reasoning	F1 Score	70.0%	65.0%	68.0%	7.7%	0.03	0.06	0.05
IconQA (r = 16)	Multimodal (Image-Text)	Average Score	80.0	70.0	75.0	14.3%	0.8	1.2	1.0
IconQA (r = 32)	Multimodal (Image-Text)	Average Score	83.0	72.0	78.0	15.3%	0.7	1.3	0.9
IconQA (r = 64)	Multimodal (Image-Text)	Average Score	85.0	75.0	80.0	13.3%	0.6	1.1	0.8
10-Task Combo	Mixed (NLP + CV + Multimodal)	Average Accuracy	85.6%	77.8%	81.2%	7.8%	0.03	0.08	0.06
Image Description (IconQA)	Multimodal Generation	BLEU Value	27.8%	21.3%	24.5%	30.5%	0.5	0.8	0.6
Cross-Modal Reasoning (IconQA)	Multimodal Inference	Accuracy	82.4%	73.6%	77.8%	11.9%	0.04	0.07	0.05

According to Table 11, the proportion of trainable parameters in the 1 billion/10 billion parameter model system of SLoRA is only 0.8%. The time consumed for a single round of training (3.2 h/12.5 h), the required number and scale of GPUs (8/16 A100), and the storage requirements (20 GB/80 GB) are all far lower than those of the full fine-tuning operation. The computational cost only reaches about 2.5% of that of the full fine-tuning method. At the same time, compared with LoRA and AdaLoRA, the efficiency performance is close to the same, achieving an efficient balance between performance and computational cost.

According to Table 12, it was found that SLoRA has a high retention rate of 91.2%–94.8% for the accuracy of old tasks in different task sequence architectures (covering NLP, CV, and cross-modal categories). The loss caused by task switching is only in the range of 2.3%–4.7%. The amplitude of parameter perturbations ($\pm$0.03–$\pm$0.05) and the degree of feature space conflicts (0.12–0.16) are significantly lower than those of LoRA and AdaLoRA, effectively alleviating and suppressing catastrophic forgetting phenomena in multitask learning processes and enhancing cross-task knowledge transfer capabilities.

Table 11. Statistics of computational cost and efficiency.

Model Parameter Scale	Method	Trainable Parameter Ratio	Single-Round Training Time	Required GPU Quantity (A100)	Single-Round Storage Requirement	Computational Cost (Relative Value)	FLOPs (Single Round)
1 Billion	SLoRA	0.8%	3.2 h	8	20 GB	1.0	2.9×10^{17}
1 Billion	LoRA	0.32%	2.8 h	6	15 GB	0.9	2.5×10^{17}
1 Billion	AdaLoRA	0.48% (Avg.)	3.0 h	7	18 GB	0.95	2.7×10^{17}
10 Billion	SLoRA	0.8%	12.5 h	16	80 GB	1.0	1.1×10^{18}
10 Billion	LoRA	0.32%	10.8 h	12	60 GB	0.86	0.95×10^{18}
10 Billion	Full Fine-Tuning	100%	48.0 h	50	400 GB	39.2	5.2×10^{18}

Through Table 13, we conducted a content exploration on the constraint initialization strategy of SLoRA and optimized MoE architecture system, which constitute the core elements of performance. After removing the constraint initialization operation, the accuracy of commonsense reasoning decreased by 10%, the knowledge retention rate decreased to 75.6%, and the performance degraded by 11.8%; after removing the MoE architecture settings, the multitasking F1 value decreased to 65%, resulting in a performance degradation of 9.1%; when removing cross-expert attention mechanisms or dynamic routing settings, it can also cause a decline in performance. This verifies and confirms the key role of dual core components and sub modules in alleviating forgetting problems and improving multitask adaptation performance.

Table 12. Statistics of knowledge retention and task switching.

Task Sequence	Old Task Accuracy Retention	Task Switching Loss	Parameter Perturbation Amplitude	Feature Space Conflict Degree (Cosine Distance)	Knowledge Retention Improvement (vs. LoRA)
Text Classification → Sentiment Analysis → Machine Translation	SLoRA: 92.4% LoRA: 76.3% AdaLoRA: 79.8%	SLoRA: 2.3% LoRA: 8.7% AdaLoRA: 6.5%	SLoRA: ±0.05 LoRA: ±0.18 AdaLoRA:±0.12	SLoRA: 0.16 LoRA: 0.42 AdaLoRA: 0.31	21.1%
Image Classification (CIFAR-10) → Object Detection (PASCAL VOC)	SLoRA: 94.8% LoRA: 83.5% AdaLoRA: 86.2%	SLoRA: 4.7% LoRA: 15.3% AdaLoRA: 12.1%	SLoRA: ±0.03 LoRA: ±0.21 AdaLoRA: ±0.15	SLoRA: 0.12 LoRA: 0.48 AdaLoRA: 0.36	13.5%
NLP + CV + Multimodal (6-Task Combo)	SLoRA: 91.2% LoRA: 74.6% AdaLoRA: 78.3%	SLoRA: 3.1% LoRA: 9.5% AdaLoRA: 7.8%	SLoRA: ±0.04 LoRA: ±0.19 AdaLoRA: 0.13	SLoRA: 0.14 LoRA: 0.45 AdaLoRA: 0.33	22.2%

Table 13. Detailed statistics of ablation experiments.

Ablation Scheme	Commonsense Reasoning Accuracy	Computer Vision Accuracy	Multitask F1 Value	Performance Degradation	Knowledge Retention Rate	Expert Feature Overlap	Parameter Perturbation	Loss Fluctuation Std. Dev.
Complete SLoRA	85.0%	80.0%	70.0%	-	92.4%	44.2%	±0.05	0.03
Remove Constraint Initialization	75.0%	78.0%	66.0%	11.8%	75.6%	41.3%	±0.17	0.06
Remove MoE Architecture (Retain LoRA)	80.0%	75.0%	65.0%	9.1%	78.2%	32.1%	±0.15	0.07
Remove Cross-Expert Attention	82.0%	77.0%	67.0%	5.7%	86.3%	35.8%	±0.08	0.04
Remove Dynamic Routing	79.0%	76.0%	64.0%	10.3%	83.5%	38.4%	±0.11	0.05

Through analysis and discrimination based on Table 14, SLoRA achieves the optimal balance between performance and efficiency level under the rank r = 16 state (IconQA average score of 80 points, relative computational cost value of 1.0, single task activation number of 1.9 experts). As the rank value increases (stage 32/64/128), the performance improvement gradually narrows (only achieving a 3–6 point improvement effect), but the computational cost shows an exponential growth trend (reaching 1.8–6.7 times the

growth rate), demonstrating the high efficiency and practical value of low rank settings in multitasking environments.

Table 14. SLoRA performance efficiency comparison under different ranks.

Rank (r)	IconQA Average Score	Computational Cost (Relative Value)	5% Confidence Interval	Expert Activation Efficiency	Performance-Efficiency Balance	Trainable Parameters (1 B Model)	Feature Extraction Accuracy
16	80.0	1.0	[78.4, 81.6]	1.9 experts/task	Optimal	8.0 million	89.7%
32	83.0	1.8	[81.7, 84.3]	2.1 experts/task	Suboptimal	14.4 million	91.2%
64	85.0	3.2	[83.9, 86.1]	2.4 experts/task	Low Efficiency	25.6 million	92.5%
128	86.0	6.7	[85.1, 86.9]	2.7 experts/task	Inefficient	48.0 million	93.1%

5. Conclusions

This study offers an in-depth analysis of the innovative architecture SLoRA, which demonstrates unique advantages in efficiently fine-tuning parameters in solving multitasking scenarios. It relies on two core components: the borrowed constraint solution space initialization (based on orthogonal constraint optimization, which can reduce the disturbance of existing knowledge in multitask learning, alleviate catastrophic forgetting problems, and provide a stable foundation for knowledge transfer between different tasks) and the optimized MoE structure (introducing general experts and task-specific experts, combined with dynamic routing mechanisms, which can significantly improve the adaptability of the model to multitasking, promote information sharing among experts, and solve knowledge fragmentation problems), effectively overcoming the limitations of traditional methods. In terms of experimental verification, SLoRA performed well in both commonsense reasoning tasks and the IconQA multimodal dataset, outperforming methods such as LoRA and AdaLoRA in commonsense reasoning tasks, demonstrating stronger semantic understanding and knowledge reasoning abilities. It reached the best results under different LoRA rank settings on the IconQA multimodal dataset, with significantly higher average scores than the baseline method and stronger stability and generalization. The ablation study further confirmed that SLoRA can achieve a good balance between performance and efficiency when the rank is 16, and the constraint initialization strategy and MoE architecture play a critical role in performance improvement.

The specific quantitative results are as follows: ① The performance improvement is as follows: among the 10 multitask combination architectures, the average accuracy reaches 85.6%, which is a 7.8% improvement compared to LoRA; the accuracy rate of the WSC dataset is 85%, showing a growth trend of 9.0% compared to LoRA; the F1 score of the CommonsenseQA dataset is at the 70% level, achieving a 7.7% improvement compared to LoRA; under the configuration of r = 32, the IconQA dataset scored an average of 83 points, achieving a 15.3% improvement compared to LoRA. ② The alleviation of the forgetting phenomenon means that the accuracy retention rate of old tasks reaches 92.4%, and the loss caused by task switching is only 2.3%, which is 9.2 percentage points lower than MoE LoRA. ③ The optimization status at the efficiency level means that the proportion of parameters that can be implemented for training is only 0.8%, and the training time for a single round is 3.2 h (in an 8 × A100 environment), which is 93.3% shorter than the full fine-tuning mode. The computational cost has been reduced by 99.2%. ④ The performance of cross-modal ability is as follows: the accuracy rate of cross-modal reasoning reaches 82.4%, which is 8.8 percentage points higher than LoRA. The BLEU value of the Image Description Generation task is 27.8%, approaching the level presented by full fine-tuning (28.5%).

Author Contributions: Conceptualization, C.S.; Methodology, C.S.; Validation, C.S. and J.-W.J.; Formal analysis, C.S.; Investigation, C.S.; Resources, C.S.; Data curation, C.S.; Writing—original draft, C.S.; Writing—review and editing, C.S. and J.-W.J.; Visualization, C.S. and J.-W.J.; Supervision, J.-W.J.; Project administration, J.-W.J.; Funding acquisition, J.-W.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was financially supported by the Ministry of Trade, Industry and Energy (MOTIE) and Korea Institute for Advancement of Technology (KIAT) through the International Cooperative R&D program. (Project No. P0030503), by the Ministry of Trade, Industry and Energy (MOTIE) and Korea Institute for Advancement of Technology (KIAT) through the International Cooperative R&D program. (Project No. P0026318), by the MSIT (Ministry of Science and ICT), Republic of Korea, under the ITRC (Information Technology Research Center) support program (IITP-2026-2020-0-01789) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), by the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2026-RS-2023-00254592) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), and by the Commercialization Promotion Agency for R&D Outcomes (COMPA) grant funded by the Republic of Korea government (Ministry of Science and ICT) (2710086167).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Verma, N.; Singh, S.; Prasad, D. A review on existing iot architecture and communication protocols used in healthcare monitoring system. *J. Inst. Eng. India Ser. B* **2022**, *103*, 645–660. [CrossRef]
2. Yao, H.; Yu, W. Application of human-machine collaborative creative generation process in industrial design. *Adv. Eng. Innov.* **2025**, *16*, 78–85. [CrossRef]
3. Pu, X.; Xu, F. Low-rank adaption on transformer-based oriented object detector for satellite onboard processing of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5202213. [CrossRef]
4. Cappelletti, S.; Baraldi, L.; Cocchi, F.; Cornia, M.; Baraldi, L.; Cucchiara, R. Adapt to Scarcity: Few-Shot Deepfake Detection via Low-Rank Adaptation. In Proceedings of the International Conference on Pattern Recognition, Kolkata, India, 1–5 December 2024.
5. Gianluca, L.; Beatriz, S.; Alessandra, C. Lora-dr-suite: Adapted embeddings predict intrinsic and soft disorder from protein sequences. *Bioinformatics* **2025**, *41*, i56–i64.
6. Yu, G.; Zhang, L.; Wang, X.; Yu, K.; Liu, R.P. A novel dual-blockchained structure for contract-theoretic lora-based information systems. *Inf. Process. Manag.* **2021**, *58*, 102492. [CrossRef]
7. Kowsher, M.; Prottasha, N.J.; Bhat, P. Propulsion: Steering LLM with Tiny Fine-Tuning. *IEEE Access* **2024**, *12*, 138171–138181.
8. Bian, J.; Wang, L.; Zhang, L.; Xu, J. LoRA-FAIR: Federated LoRA Fine-Tuning with Aggregation and Initialization Refinement. *arXiv* **2024**, arXiv:2406.16971.
9. Zhang, Z.; Liu, P.; Xu, J.; Hu, R. Fed-HeLLo: Efficient Federated Foundation Model Fine-Tuning with Heterogeneous LoRA Allocation. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**. *early access*. [CrossRef] [PubMed]
10. Zhao, X.; Lin, B.; Song, Y. LasQ: Largest Singular Components Fine-Tuning for LLMs with Quantization. In *Natural Language Processing and Chinese Computing*; Springer: Singapore, 2025; pp. 58–70.
11. Qu, T.; Tuytelaars, T.; Moens, M.F. Introducing Routing Functions to Vision-Language Parameter-Efficient Fine-Tuning with Low-Rank Bottlenecks. In *Computer Vision—ECCV 2024*; Springer: Cham, Switzerland, 2024; pp. 287–304.
12. Lin, X.; Wang, Y.; Zhang, Z.; Zhang, M. Scientific Claim Recognition via Staged Fine-Tuning with LoRA. *Data Intell.* **2025**, *7*, 303–335. [CrossRef]
13. Liao, X.; Wang, C.; Zhou, S.; Hu, J.; Zheng, H.; Gao, J. Dynamic Adaptation of LoRA Fine-Tuning for Efficient and Task-Specific Optimization of Large Language Models. *arXiv* **2025**, arXiv:2501.12345.
14. Chen, G.; He, Y.; Hu, Y.; Yuan, K.; Yuan, B. CE-LoRA: Computation-Efficient LoRA Fine-Tuning for Language Models. *arXiv* **2025**, arXiv:2503.98765.
15. Mu, J.; Wang, W.; Liu, W.; Yan, T.; Wang, G. Multimodal Large Language Model with LoRA Fine-Tuning for Multimodal Sentiment Analysis. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 1–24. [CrossRef]

16. Djidi, N.E.H.; Gautier, M.; Courtay, A.; Berder, O.; Magno, M. How can wake-up radio reduce lora downlink latency for energy harvesting sensor nodes? *Sensors* **2021**, *21*, 1050. [CrossRef] [PubMed]
17. Rice, C.; Longabaugh, R.; Beattie, M.; Noel, N. Age group differences in response to treatment for problematic alcohol use. *Addiction* **1993**, *88*, 1369–1375. [CrossRef]
18. Tian, C.; Shi, Z.; Guo, Z.; Li, L.; Xu, C. Hydralora: An asymmetric lora architecture for efficient fine-tuning. *arXiv* **2024**, arXiv:2408.06516.
19. Kazdaridis, G.; Sidiropoulos, N.; Zografopoulos, I.; Korakis, T. A novel architecture for semi-active wake-up radios attaining sensitivity beyond -70 dbm: Demo abstract. In Proceedings of the 20th International Conference on Information Processing in Sensor Networks (Co-Located with CPS-IoT Week 2021), Nashville, TN, USA, 18–21 May 2021.
20. Hosseini, S.V.; Alim, U.R.; Oehlberg, L.; Taron, J.M. Optically illusive architecture (oia): Introduction and evaluation using virtual reality. *Int. J. Archit. Comput.* **2021**, *19*, 283–300. [CrossRef]
21. Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. Advances in Neural Information Processing Systems 32, Volume 5 of 20. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, CA, USA, 8–14 December 2019.
22. Talmor, A.; Herzig, J.; Lourie, N.; Berant, J. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, MN, USA, 2–7 June 2019*; ACL Press: Austin, TX, USA, 2019; pp. 4149–4158.
23. Pan, L.; Ye, T.; Yu, Z.; Li, L.; Wang, W.Y. IconQA: A New Benchmark for Abstract Diagram Understanding and Visual Language Reasoning. In Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI 2021), Virtual Event, 2–9 February 2021; pp. 1664–1672.