LoRA and Its Applications in Multimodal Learning: A Survey

www.surveyx.cn

Abstract

Low-Rank Adaptation (LoRA) has emerged as a transformative technique in the domain of parameter-efficient fine-tuning, particularly for large-scale models like Vision-Language Models (VLMs) and Vision Large Language Models (VLLMs). By introducing trainable low-rank matrices, LoRA significantly reduces computational and storage burdens, facilitating efficient model adaptation across diverse tasks and modalities. This survey explores LoRA's role in enhancing cross-modal integration, crucial for multimodal learning environments. LoRA's versatility is demonstrated through its application in frameworks such as HyperMM and FedEx-LoRA, optimizing performance in scenarios with missing modalities and federated learning settings. Despite challenges in multi-task learning and data heterogeneity, LoRA's integration with pruning techniques, as seen in methods like Focal Pruning (FoPru), further reduces model complexity while maintaining efficacy. The strategic tuning of LoRA hyperparameters plays a critical role in optimizing adaptation processes, enhancing model robustness and efficiency. Case studies across healthcare, e-commerce, and multimedia domains underscore LoRA's potential to enhance model performance and adaptability. Future research should focus on refining hyperparameter optimization, exploring integration with other techniques, and addressing scalability challenges to advance LoRA's application in complex, multimodal learning environments. These efforts will ensure LoRA's continued impact in improving parameter efficiency and model performance across diverse applications.

1 Introduction

1.1 Concept and Importance of LoRA

Low-Rank Adaptation (LoRA) is a pivotal technique in parameter-efficient fine-tuning for large-scale models, significantly alleviating the computational and storage demands associated with traditional fine-tuning methods. By incorporating trainable low-rank matrices, LoRA enables efficient model adaptation to new tasks and modalities, enhancing its applicability across various learning environments. This efficiency is particularly advantageous in federated learning settings, where LoRA optimizes communication and computational resources [1].

In multimodal learning contexts, LoRA's adept management of parameter space is vital for improving cross-modal integration, essential for models processing both text and image data. Despite its advantages, challenges remain in multi-task learning scenarios, where existing LoRA methods may blur task distinctions, leading to performance degradation due to task interference [2]. Nevertheless, LoRA's contribution to enhancing Vision Large Language Models (VLLMs) is well-documented, highlighting its role in maintaining model efficiency while adapting to complex data inputs [3].

Moreover, LoRA's implementation in vision-language models (VLMs) mitigates the high computational costs associated with traditional fine-tuning by modifying model weights at the matrix level rather than adding additional layers [4]. This strategy preserves the structural integrity of the model

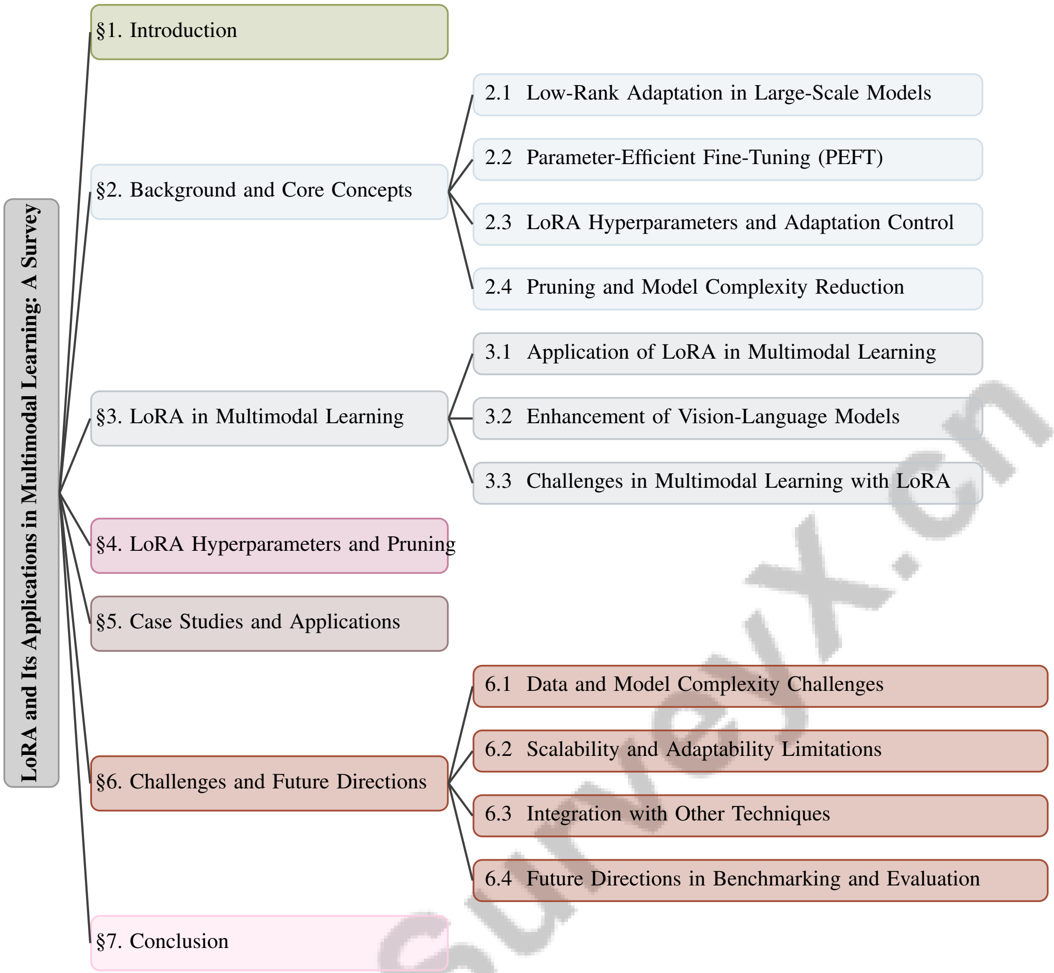

LoRA and Its Applications in Multimodal Learning: A Survey
§1. Introduction
§2. Background and Core Concepts
2.1 Low-Rank Adaptation in Large-Scale Models
2.2 Parameter-Efficient Fine-Tuning (PEFT)
2.3 LoRA Hyperparameters and Adaptation Control
2.4 Pruning and Model Complexity Reduction
§3. LoRA in Multimodal Learning
3.1 Application of LoRA in Multimodal Learning
3.2 Enhancement of Vision-Language Models
3.3 Challenges in Multimodal Learning with LoRA
§4. LoRA Hyperparameters and Pruning
§5. Case Studies and Applications
§6. Challenges and Future Directions
6.1 Data and Model Complexity Challenges
6.2 Scalability and Adaptability Limitations
6.3 Integration with Other Techniques
6.4 Future Directions in Benchmarking and Evaluation
§7. Conclusion

Figure 1: chapter structure

while optimizing performance, making it particularly effective in scenarios with restricted access to model architectures [5]. Through these diverse applications, LoRA establishes itself as a fundamental technique for achieving parameter efficiency and enhancing model adaptability across a wide array of large-scale learning environments, ensuring robustness in handling complex, multimodal data.

1.2 Motivation for Using LoRA in Multimodal Learning

The motivation for employing Low-Rank Adaptation (LoRA) in multimodal learning stems from the necessity to manage the substantial size and computational demands of pre-trained Transformer models, which often pose challenges in traditional fine-tuning scenarios. LoRA effectively addresses these issues by optimizing the adaptation process of large-scale models to specific tasks while minimizing computational costs, thus enhancing model efficiency [5]. This is particularly vital in Vision-Language Models (VLMs), where LoRA's parameter-efficient approach promotes robust cross-modal integration, facilitating the effective utilization of both text and image data.

In situations where modalities may be missing or incomplete, LoRA offers a flexible solution that reduces reliance on imputation, thereby preserving model robustness and performance [6]. This capability is further enhanced through LoRA's soft-prompt tuning, which leverages the intrinsic properties of soft prompts to improve model performance [4]. Additionally, LoRA alleviates inefficiencies caused by redundant visual tokens in LVLMs, enhancing performance without extensive retraining [7].

Furthermore, LoRA's effectiveness in multimodal learning is underscored by its capacity to manage task interference and confusion, common challenges when projecting features from different tasks

AI-generated, for reference only.

into a shared low-dimensional space [2]. By addressing these issues, LoRA ensures the preservation of task-specific information, thereby enhancing the overall efficacy of multimodal models. The technique also emphasizes safety alignment in VLLMs, mitigating risks of harmful outputs and adversarial attacks [3]. These multifaceted advantages position LoRA as an essential tool for advancing multimodal learning, fostering efficient cross-modal integration, and ensuring robust model performance in diverse and complex data environments.

1.3 Overview of Paper Structure

This survey systematically explores the multifaceted applications and implications of Low-Rank Adaptation (LoRA) in multimodal learning environments. The introduction elucidates LoRA's significance in parameter-efficient fine-tuning of large-scale models and its critical role in enhancing cross-modal integration in vision-language models. The background section delves into core concepts, including low-rank adaptation, parameter-efficient fine-tuning, and the essential role of LoRA hyperparameters in controlling the adaptation process, alongside introducing pruning techniques to further reduce model complexity.

The third section examines practical applications of LoRA in multimodal learning, highlighting its effectiveness in improving model integration and efficiency within vision-language models while addressing implementation challenges and opportunities. The subsequent section focuses on LoRA hyperparameters and pruning, discussing their impact on fine-tuning processes and optimization strategies, as well as specific pruning techniques that complement LoRA to enhance model performance.

In the fifth section, the paper presents case studies and real-world applications of LoRA, showcasing successful implementations across various domains and the optimization of hyperparameters and pruning strategies. The final section addresses challenges and future directions for LoRA, discussing current limitations and potential advancements in hyperparameter optimization and integration with other techniques. The paper concludes by summarizing key findings and underscoring LoRA's importance in advancing parameter efficiency and model performance in multimodal learning.The following sections are organized as shown in Figure 1.

2 Background and Core Concepts

2.1 Low-Rank Adaptation in Large-Scale Models

Low-Rank Adaptation (LoRA) is pivotal in enhancing the fine-tuning of large-scale models, such as vision-language models (VLMs) and vision large language models (VLLMs), by optimizing parameter efficiency and computational demands. By decomposing parameters into low-rank matrices, LoRA reduces computational overhead while preserving model performance, facilitating effective adaptation across diverse tasks and data environments. As part of Parameter Efficient Fine-Tuning (PEFT) methods, LoRA offers advantages in parameter management through additive, selective, reparameterized, and hybrid strategies [8].

In multimodal learning, LoRA's integration into frameworks like HyperMM demonstrates its robustness against missing modalities by accommodating varying input sizes without imputation [6]. This capability is crucial for seamless visual and textual data integration, enhancing cross-modal learning. Techniques such as normalizing soft-prompt vectors further augment LoRA's adaptability, leveraging the Low-Norm Effect to boost VLM performance [4].

In safety-critical applications, LoRA's role in projects like VLGuard underlines its importance in enhancing VLLM safety and efficiency through targeted fine-tuning [3]. LoRA's framework ensures scalability and adaptability across domains, exemplified by its ability to separate style and content in image stylization tasks, refine visual attributes, and enhance multi-task learning in large language models (LLMs), thus improving performance across varied domains. Its efficiency in parameter reduction significantly lowers resource demands in federated learning and general model adaptation [9, 2, 10, 11, 12].

AI-generated, for reference only.

2.2 Parameter-Efficient Fine-Tuning (PEFT)

Parameter-Efficient Fine-Tuning (PEFT) revolutionizes model adaptation by optimizing computational efficiency while maintaining or enhancing performance. By selectively updating a subset of model parameters, PEFT conserves resources and minimizes retraining needs [8]. LoRA exemplifies PEFT by adaptively modifying model weights with minimal adjustments. PEFT methods include additive approaches with adapter modules and soft prompts, reparameterized methods with low-rank adaptations, selective masking techniques, hybrid approaches, and quantization for reduced precision computations [5].

Innovations in PEFT, such as Context-PEFT, tailor pre-trained language models for multimodal tasks by learning context-specific adaptor parameters, enhancing task-specific performance [13]. LLaMA-Adapter showcases competitive performance with minimal parameters, highlighting PEFT's potential for high efficiency [14]. In zero-shot classification, PEFT uses large language model-generated descriptions and fine-grained image datasets to improve accuracy [15].

Developments like DPD-LoRA enhance adaptability with hierarchical prompt tokens, offering robust model updates [16]. FedEx-LoRA improves LoRA adapter aggregation with a residual error term, enhancing adaptability and performance [1]. These advancements underscore PEFT's evolution, ensuring robustness across tasks and data environments. Context-PEFT, for instance, adapts PEFT for multi-modal and multi-task transfer learning, reducing trainable parameters and GPU memory requirements while retaining performance, making PEFT essential for scalable and sustainable machine learning systems [5, 13, 8, 17].

2.3 LoRA Hyperparameters and Adaptation Control

Strategic hyperparameter selection and optimization in Low-Rank Adaptation (LoRA) are crucial for controlling adaptation processes and ensuring optimal model performance. Key hyperparameters, including learning rates, scaling factors, and rank selection, significantly influence model adaptability to new tasks and modalities. Optimal rank selection across layers is vital, as uniform choices can compromise image quality [18]. This highlights the importance of task-specific methodologies and pretraining techniques in LoRA's adaptation [19].

Context-PEFT exemplifies efficient multimodal fine-tuning by training context-specific adaptor parameters, enhancing performance with reduced computational demands [13]. Dual-path adaptation techniques, involving positive selection and negative exclusion, illustrate effective hyperparameter utilization [20]. Careful hyperparameter tuning is essential for optimal results, as shown in methods requiring precise adjustments to control adaptation dynamics [21]. FLoRA employs modality-specific adapters to incorporate additional modalities into a pre-trained LLM, enabling efficient multimodal data processing through strategic hyperparameter configuration [22]. The Nemesis approach normalizes soft-prompt vectors, with its effectiveness linked to specific norms influencing VLM performance [4].

Hyperparameter complexity in LoRA affects computational resource optimization and memory usage, enhancing adaptability across tasks and maintaining performance in diverse environments. By strategically adjusting these hyperparameters, practitioners can navigate trade-offs between efficiency and accuracy, ensuring LoRA remains competitive for fine-tuning large language models [10, 11, 2, 23]. Their careful tuning is crucial for balancing computational efficiency and model robustness, making them essential for successful LoRA deployment.

2.4 Pruning and Model Complexity Reduction

Pruning techniques are essential for managing the complexity of large-scale models, such as Vision-Language Models (VLMs), while preserving performance. Integrated with Low-Rank Adaptation (LoRA), these methods facilitate efficient parameter management and model adaptation. The Focal Pruning (FoPru) method exemplifies this by selectively eliminating less significant visual tokens based on attention scores, reducing model complexity without sacrificing performance [7]. This aligns with LoRA's goals by retaining critical parameters and optimizing computational resources.

Traditional compression methods, including weight sharing, low-rank factorization, quantization, knowledge distillation, and pruning, often lack efficiency for multimodal models [24]. However, advanced pruning strategies, such as those in the Pruning All-Rounder (PAR) framework, demonstrate potential for reducing model complexity while maintaining efficacy. PAR addresses the computational burden of LVLMs through comprehensive pruning techniques [25].

Adaptive pruning methods like ATP-LLaVA offer dynamic alternatives to fixed strategies, adjusting pruning techniques to sustain optimal performance [26]. This adaptability is beneficial in multimodal learning, where balancing model complexity and performance is crucial. Pruning techniques can also be employed with LoRA to enhance VLLM safety and performance, addressing complexity while ensuring safety evaluations [3].

In aligning visual and textual representations, particularly in long-form video-text tasks, benchmarks like CoSMo highlight the challenge of maintaining alignment across modalities [27]. Through advanced pruning strategies, integration with LoRA offers a robust framework for managing model complexity, optimizing computational efficiency, and maintaining performance in large-scale, multimodal learning environments.

3 LoRA in Multimodal Learning

3.1 Application of LoRA in Multimodal Learning

Low-Rank Adaptation (LoRA) significantly advances multimodal learning by efficiently integrating parameters across diverse data modalities. This is accomplished through strategic parameter decomposition, enhancing cross-modal integration vital for processing complex datasets. In Vision-Language Models (VLMs), LoRA improves adaptability and performance, particularly in handling Out Of Distribution (OOD) samples and supporting robust online adaptation [5]. LoRA's versatility spans NLP and computer vision tasks, demonstrating adaptability across various model architectures [11]. For instance, the FedEx-LoRA framework enhances federated learning by ensuring precise updates and effective cross-modal integration [1]. Additionally, PC-LoRA leverages low-rank adapters for efficient fine-tuning of large-scale models [28].

LoRA's transformative impact is further illustrated in frameworks like HyperMM, which employs a universal feature extractor and a permutation-invariant network to enhance model integration [6]. Nemesis addresses the Low-Norm Effect in multimodal environments, improving model integration and efficiency [4]. Furthermore, LoRA reduces computational costs and enhances efficiency, as demonstrated by methods like FoPru, which improves inference efficiency through visual token pruning [7]. This strategic application underscores LoRA's critical role in advancing multimodal learning, ensuring models effectively integrate and adapt information from multiple modalities.

3.2 Enhancement of Vision-Language Models

LoRA is crucial for enhancing Vision-Language Models (VLMs) by facilitating efficient parameter adaptation necessary for integrating visual and textual data. By enabling models to learn from both positive and negative examples, LoRA improves classification accuracy, particularly in few-shot adaptation contexts [20]. This approach enhances generalization capabilities across diverse datasets and tasks. LoRA supports adaptive aggregation of client models and the use of multimodal prototypes, essential for making predictions on unseen classes, thereby extending applicability in open-vocabulary settings [29]. The ReCoVERR framework exemplifies this adaptability by allowing VLMs to recover relevant visual evidence related to low-confidence predictions, improving output reliability [30].

Moreover, LoRA enhances multimodal models by estimating missing modality tokens based on available data, as demonstrated in the MMP framework, significantly boosting performance in scenarios with absent modalities [31]. The XMAdapter showcases LoRA's benefits by integrating cross-modal information, improving accuracy and generalization across multiple benchmark datasets [32]. Under black-box conditions, LoRA achieves substantial improvements in adapting VLMs, attaining state-of-the-art performance across various benchmarks, highlighting its robustness and versatility [33]. The application of LoRA within the VLGuard dataset provides a framework for safety evaluation, ensuring VLMs operate within safe and efficient parameters [3].

Strategic implementation of LoRA underscores its critical role in enhancing parameter adaptation efficiency, facilitating superior model performance, and ensuring seamless integration within complex multimodal learning environments. LoRA diverges from traditional approaches, such as the Adapter method, by expanding model width rather than depth, mitigating training instability and inference latency. In multi-task learning contexts, MTL-LoRA enhances LoRA's capabilities by incorporating task-adaptive parameters, effectively reducing task interference and improving overall performance across domains, reinforcing LoRA's position as a leading technique for efficient adaptation in large language models [11, 2].

3.3 Challenges in Multimodal Learning with LoRA

Implementing LoRA in multimodal learning environments presents challenges that can hinder effectiveness and scalability. A significant issue is data heterogeneity across clients, complicating model adaptation to new classes not represented in training data [29]. This challenge is exacerbated by interference between model parameters from different tasks, potentially leading to performance degradation [34]. Addressing representation and association biases inherent in multimodal data is another core challenge, as existing methods often struggle to counter these biases effectively, compromising model integrity and generalization capabilities [35]. Distortion in multimodal representation learning due to heterogeneous client data hampers the generalization ability of federated models, posing substantial obstacles to achieving ideal cross-modal alignment [36].

LoRA also faces technical constraints, such as high memory demands for large model architectures, exemplified by methods like C OND P-D IFF [37]. This challenge is compounded by limited support for heterogeneous models and reliance on data-unaware model structures in frameworks like FedAMoLE [38]. Additionally, converting dense models to sparse models while maintaining effective training and inference capabilities often results in performance degradation, highlighting the complexities involved in model adaptation [39]. The dependency on the availability of paired data for effective alignment presents another challenge, as this may not always be feasible in real-world scenarios [40]. Moreover, the lack of clarity regarding whether larger encoders directly correlate with better performance in multimodal tasks complicates model architecture optimization [41]. The combinatorial nature of the modality selection problem further complicates the process, making it difficult to efficiently find the optimal subset as the number of modalities increases [42].

Achieving few-shot adaptability in VLMs without explicit fine-tuning remains a challenge due to latency and instability with limited data, often rendering fine-tuning infeasible [43]. The excessive abstention of vision-language models when uncertain, resulting in low coverage of correct predictions, is another significant challenge addressed by methods like ReCoVERR [30]. The challenges associated with LoRA's implementation in multimodal learning environments underscore the need for innovative solutions that enhance both effectiveness and flexibility. While LoRA has proven efficient for adapting Large Language Models (LLMs), issues such as task interference in multi-task learning and the integration of structured and unstructured data remain significant hurdles. Addressing these challenges may involve developing advanced techniques like MTL-LoRA, which introduces task-adaptive parameters to mitigate task confusion, and frameworks like LANISTR, which effectively learn from diverse data types while maintaining performance even in the presence of missing modalities [11, 2, 44]. A multifaceted approach incorporating advancements in model architecture, data alignment, and parameter management is essential for ensuring robust and scalable multimodal learning systems.

4 LoRA Hyperparameters and Pruning

4.1 Impact of LoRA Hyperparameters on Fine-tuning

Method Name	Hyperparameter Configuration	Model Robustness	Resource Efficiency
MMP[31]	Low-rank Decomposition	Missing Modality Scenario	Computationally Expensive
PC-LoRA[28]	Decay Factor Scheduler	Competitive Performance Maintenance	Computational Demands Reductions
FEL[1]	Residual Error Term	Exact Updates	Minimal Communication Overhead
HMM[6]	Conditional Hypernetwork	Missing Modalities Handling	Computational Efficiency
FM[45]	Sparse Moe Layer	Missing Modality Bank	Sparse Mixture-of-Experts
LoRA[11]	Matrix Level Adaptation	Training Stability Enhancement	Reduced Training Overhead
GMS[42]	Utility Function Optimization	Informative Modality Selection	Reduced Computational Cost

Table 1: Comparison of various multimodal learning methods highlighting their hyperparameter configurations, model robustness, and resource efficiency. The table provides insights into how different approaches manage missing modalities, computational demands, and training stability, emphasizing the role of hyperparameter tuning in optimizing model performance.

The configuration of hyperparameters in Low-Rank Adaptation (LoRA) is crucial for optimizing fine-tuning, significantly affecting model performance across various tasks and data environments. Hyperparameters, especially those related to low-rank decomposition, shape the adaptation process, as evidenced by MMP's ability to learn compensatory strategies for missing information, highlighting the importance of hyperparameter selection for model robustness and adaptability [31]. Performance improvements through hyperparameter adjustments are demonstrated by PC-LoRA, which uses low-rank adapters to counteract phased-out pre-trained weights while maintaining model performance [28]. This underscores hyperparameters' role in resource efficiency and robustness. Additionally, FedEx's use of an error residual term allows precise updates without compromising LoRA's low-rank efficiency, further illustrating hyperparameters' critical role in effective model adaptation [1]. Table 1 presents a comparative analysis of several multimodal learning methods, illustrating the impact of hyperparameter configurations on model robustness and resource efficiency.

In multimodal learning contexts, frameworks like HyperMM, which prioritize input data integrity without imputation, are significantly influenced by hyperparameter tuning, impacting fine-tuning and overall model performance [6]. Similarly, Flex-MoE's capacity to generalize knowledge across various modality combinations while specializing in specific ones relies heavily on strategic hyperparameter configuration, enabling learning from both complete and incomplete datasets [45]. Challenges associated with increased model depth, such as training instability and convergence issues in large models like GPT-3, underscore the necessity for precise hyperparameter tuning to mitigate these obstacles and enhance performance [11]. The iterative selection of modalities based on marginal utility, as seen in greedy modality selection methods, further illustrates hyperparameters' influence on optimizing resource usage and ensuring efficient model adaptation [42].

Strategic hyperparameter tuning is essential for optimizing resource usage, enhancing model adaptability, and maintaining robust performance across diverse applications. Findings emphasize hyperparameters' crucial role in improving the fine-tuning process for Multimodal Large Language Models (MLLM), ensuring resilience and adaptability in complex learning environments characterized by diverse and often incomplete data modalities. This necessity for careful hyperparameter tuning is evident in frameworks like HyperMM, which adeptly manage missing modalities without prior imputation, and strategies that balance parameter importance during fine-tuning to prevent knowledge degradation [46, 47, 48, 35, 6].

4.2 Strategies for Optimizing LoRA Hyperparameters

Method Name	Optimization Strategies	Computational Efficiency	Adaptation Techniques
SVL[49]	Hyperparameter Selection Automation	Reduce Computational Expense	Self-supervised Encoder
SpIEL[50]	Iterative Updating Deltas	Memory-efficient Methods	Adaptive Growth Criteria
CPEFT[13]	Hyperparameter Optimization	Reduced Computational Requirements	Context-specific Adaptations
RAFFT[21]	Riemannian Procrustes Analysis	Reducing Computational Costs	Riemannian Parameter Matching
ROSITA[51]	Contrastive Learning Objective	Reduce Computational Overhead	Dynamically Updated Feature
Nemesis[4]	Systematic Approach	Reduce Computational Overhead	Selective Normalization
BA-LoRA[52]	Regularization Strategies	Minimal Computational Overhead	Optimized Hyperparameter Configurations
PMR[53]	Prototypical Cross-entropy	Entropy Regularization Term	Prototypical Entropy Regularization

Table 2: Overview of optimization strategies, computational efficiency, and adaptation techniques employed by various methods for optimizing Low-Rank Adaptation (LoRA) hyperparameters. The table highlights the diverse approaches and innovations in hyperparameter optimization, showcasing methods such as SVL, SpIEL, and CPEFT, among others, to enhance model adaptability and performance across different computational environments.

Optimizing hyperparameters in Low-Rank Adaptation (LoRA) is vital for enhancing model adaptation and achieving efficient performance across various tasks and data environments. Numerous strategies have been proposed to automate and refine this process, reducing reliance on traditional hyperparameter tuning methods. For example, SVL-Adapter automates hyperparameter selection without requiring held-out labeled validation data, addressing a significant limitation of conventional adaptation methods [49].

Research into adaptive growth criteria for optimizing hyperparameters, as highlighted in SpIEL, shows potential for improving backward pass efficiency [50]. This approach aligns with the broader goal of enhancing LoRA's adaptability and efficiency, ensuring models can be fine-tuned with minimal computational overhead. The Context-PEFT framework illustrates how hyperparameter optimization can significantly enhance model capabilities, suggesting future research should focus on refining these techniques for more efficient adaptations [13]. Moreover, exploring Riemannian optimization techniques in federated learning environments presents a promising avenue for improving hyperparameter tuning, particularly under resource constraints [21].

In vision-language models, strategies in ROSITA, such as utilizing an out-of-distribution (OOD) detection module and dynamically updated feature banks, highlight the importance of strategic hyperparameter configuration for optimizing model performance in complex data environments [51]. Additionally, the normalization process in Nemesis offers opportunities for further optimization, with future research potentially investigating the effects of varying normalization strengths and strategies across different tasks [4].

Despite advancements, challenges persist, such as the reliance on hyperparameter tuning for regularization terms in methods like BA-LoRA, necessitating careful optimization for optimal results [52]. Moreover, computational efficiency in prototype calculation processes, as seen in PMR, could be further enhanced to improve overall model efficiency [53]. Continuous refinement and automation of hyperparameter optimization strategies are crucial for advancing LoRA and other parameter-efficient fine-tuning techniques. These efforts enhance model adaptability and performance, paving the way for broader applications of LoRA across diverse domains and learning environments [8]. Table 2 presents a comprehensive comparison of various methods and their respective strategies for optimizing Low-Rank Adaptation (LoRA) hyperparameters, focusing on optimization strategies, computational efficiency, and adaptation techniques.

4.3 Pruning Techniques in Conjunction with LoRA

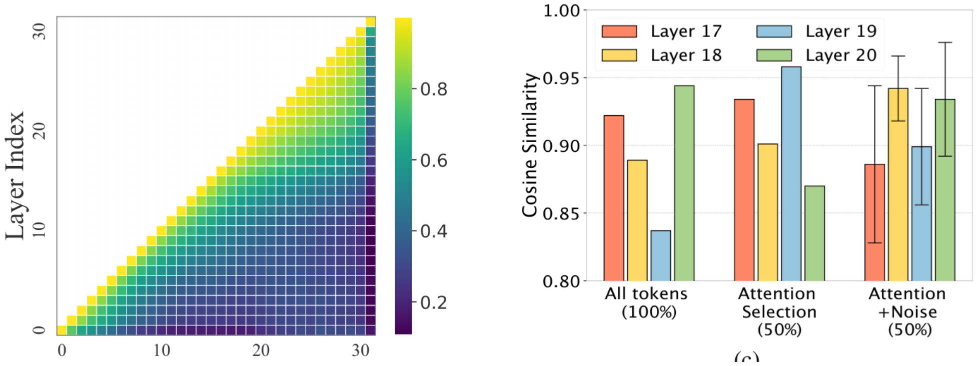
Pruning techniques, when integrated with Low-Rank Adaptation (LoRA), can significantly enhance model performance by efficiently managing complexity and optimizing parameter usage. The combination of pruning strategies with LoRA facilitates the selective removal of less important components, maintaining high performance while reducing computational overhead. A notable approach, the Pruning All-Rounder (PAR) framework, optimizes the pruning of both parameters and tokens in a self-supervised manner, enhancing model performance through simultaneous parameter and token optimization [25].

In Vision-Language Models (VLMs), rethinking pruning techniques involves integrating them with specialized fine-tuning approaches that preserve the sparse patterns of pruned models. This method optimizes LoRA weights using task-specific objectives, ensuring that sparse configurations do not compromise model adaptability and performance [54]. Such integration emphasizes the potential for pruning strategies to enhance LoRA's effectiveness in managing model complexity. The Unified Progressive Pruning and Optimization (UPop) framework exemplifies the synergy between pruning and LoRA. UPop combines unified searching for compression ratios with a progressive pruning approach, effectively reducing model complexity while maintaining performance. This method highlights the importance of adaptive pruning strategies alongside LoRA for achieving optimal model configurations [24].

Focal Pruning (FoPru) employs two token pruning strategies—Rank Pruning and Row Pruning—focusing on global and local token significance, respectively. These strategies aim to optimize the fine-tuning process by prioritizing critical tokens, thereby enhancing model efficiency and performance in conjunction with LoRA [7]. The integration of pruning techniques with LoRA provides a robust framework for managing model complexity, optimizing computational resources, and maintaining strong performance in large-scale, multimodal learning environments. Advanced pruning strategies, when aligned with LoRA's parameter-efficient adaptation capabilities, significantly reduce trainable parameters by leveraging redundancies in pre-trained weights, ultimately enhancing performance across various applications and data environments while addressing challenges in fine-tuning large language models (LLMs) within resource-constrained or distributed settings [1, 11, 28, 23].

As shown in Figure 2, exploring LoRA (Low-Rank Adaptation) hyperparameters and pruning techniques necessitates considering their effective integration to optimize model performance. The accompanying figures illustrate two crucial aspects of this integration. The first image is a heatmap visually representing the layer index and corresponding values within a dataset, providing insights into the distribution and magnitude of these values, with color intensity indicating magnitude on a scale from 0 to 1. Such representation aids in understanding different layers' contributions to the

AI-generated, for reference only.



(a) The image is a heatmap representing the layer index and layer index values for a certain dataset.[23]

(b) Comparison of Cosine Similarity Scores Across Different Layers and Selection Rates[25]

Figure 2: Examples of Pruning Techniques in Conjunction with LoRA

overall model architecture. The second image presents a comparative analysis of cosine similarity scores across various layers and selection rates, depicted through a bar chart. This chart highlights how layers (specifically Layer 17 and Layer 19) perform under varying selection rates (50

In recent years, the exploration of Low-Rank Adaptation (LoRA) has gained significant traction across various domains, showcasing its versatility and effectiveness in improving model performance. As illustrated in Figure **??**, this figure depicts the hierarchical categorization of case studies and applications of LoRA, highlighting key frameworks, datasets, and domain-specific implementations across sectors such as healthcare, e-commerce, natural language processing (NLP), and multimedia. The visual representation emphasizes the transformative potential of hyperparameter tuning and pruning optimizations, which are critical in enhancing both the performance and efficiency of models. By dissecting these applications, we can better understand the broader implications of LoRA in advancing technological capabilities across diverse fields.

Figure 3: This figure illustrates the hierarchical categorization of case studies and applications of Low-Rank Adaptation (LoRA), highlighting key frameworks, datasets, and domain-specific implementations across sectors such as healthcare, e-commerce, NLP, and multimedia. It emphasizes the transformative potential of hyperparameter tuning and pruning optimizations in enhancing model performance and efficiency.

5 Case Studies and Applications

5.1 Case Studies on Hyperparameter and Pruning Optimization

Case studies on hyperparameter and pruning optimization in Low-Rank Adaptation (LoRA) demonstrate significant improvements in model efficiency across diverse domains. The Multimodal Missing Modality Prediction (MMP) framework, evaluated using datasets like MCubeS and NYUDv2, showcases LoRA's capability in addressing missing modalities through strategic hyperparameter tuning, enhancing model robustness and adaptability [31].

The Unified Progressive Pruning and Optimization (UPop) framework exemplifies successful pruning in multimodal tasks such as Visual Question Answering, Image Captioning, and Image-Text Retrieval, maintaining performance while substantially reducing model complexity, highlighting the integration of advanced pruning with LoRA [24].

Studies on datasets like Patch-MNIST, PEMS-SF, and CMU-MOSI reveal the efficacy of greedy modality selection methods in optimizing resource utilization and performance, underscoring the critical role of hyperparameter optimization in surpassing baseline results [42].

AI-generated, for reference only.

The VLGuard dataset provides a robust evaluation framework for Vision Large Language Models (VLLMs), demonstrating successful hyperparameter optimization and pruning techniques alongside LoRA, thereby enhancing safety and performance in large-scale models [3].

Moreover, the Focal Pruning (FoPru) method achieves up to 2.52x speedup in inference time while maintaining high accuracy across multimodal datasets, exemplifying the benefits of pruning strategies in optimizing model performance [7].

These case studies highlight the significant impact of hyperparameter tuning and pruning optimization on LoRA applications, particularly techniques like SHARE LoRA, which optimize trainable parameters by leveraging redundancies in pre-trained weights while maintaining or improving performance compared to traditional fine-tuning methods, showcasing the transformative potential of these optimizations in large language models (LLMs) [11, 23].

5.2 Real-World Scenario Implementations

LoRA has been effectively implemented in various real-world scenarios, demonstrating its capacity to enhance model outcomes across different sectors. The Bunny framework, for example, has shown improved performance on eleven popular benchmarks, highlighting LoRA's efficacy in optimizing model efficiency and accuracy in practical applications [55]. Similarly, the LORS framework has achieved significant parameter reductions of 50

In the commercial sector, the OCLEAR framework has been successfully deployed in a real-world e-commerce system, illustrating LoRA's potential to enhance outcomes in industry-specific applications [56]. This use case exemplifies how LoRA can effectively manage large-scale data.

The MMStar benchmark further exemplifies LoRA's application in evaluating multi-modal capabilities, utilizing a dataset of 1,500 carefully selected samples to improve model outcomes [57].

In healthcare, experiments on datasets like ADNI and MIMIC-IV demonstrate LoRA's effectiveness in handling complex data types, such as imaging and clinical data, particularly in binary mortality prediction tasks [45]. These implementations highlight LoRA's adaptability and its role in enhancing predictive accuracy in critical applications.

Additionally, tasks such as facial age estimation, historical image dating, image quality assessment, and object count sorting have shown that LoRA can outperform baseline methods, reinforcing its applicability in various real-world scenarios [58]. Collectively, these implementations underscore LoRA's transformative potential in enhancing model performance and efficiency across multiple sectors.

5.3 Domain-Specific Applications

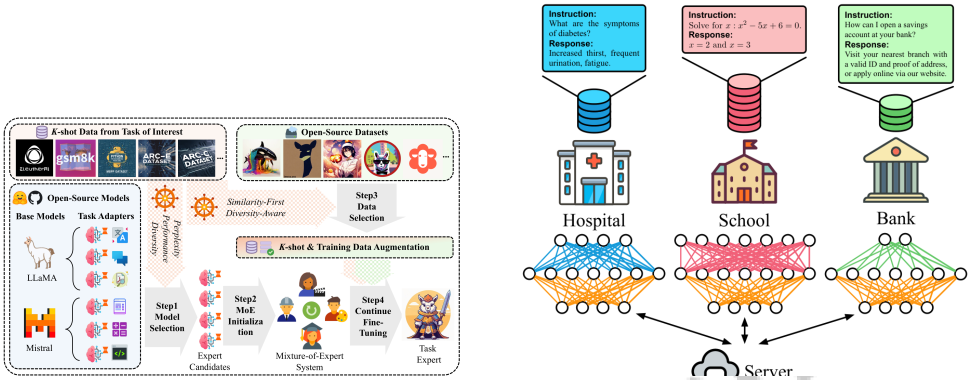
LoRA has been effectively utilized across various domains, enhancing model performance through efficient parameter tuning and adaptation. In healthcare, for example, the Flex-MoE framework has demonstrated the ability to model arbitrary modality combinations, significantly improving binary mortality prediction accuracy on datasets like ADNI and MIMIC-IV [45].

In e-commerce, the OCLEAR framework has optimized recommendation systems, achieving notable improvements in category-oriented representation learning and enhancing efficiency and accuracy in real-world applications [56]. This application highlights LoRA's potential to streamline operations and improve customer experiences.

LoRA's versatility is further exemplified in natural language processing (NLP) and computer vision tasks. The Bunny framework has achieved significant performance enhancements across multiple benchmarks, demonstrating LoRA's effectiveness in optimizing efficiency and accuracy in practical NLP applications [55]. Furthermore, LoRA's integration into frameworks like LORS has resulted in substantial parameter reductions while maintaining or enhancing performance, particularly beneficial in resource-constrained environments [59].

In multimedia and entertainment, LoRA has been employed to enhance capabilities in tasks such as facial age estimation, historical image dating, and image quality assessment, consistently outperforming baseline methods and reinforcing its applicability in improving model accuracy and efficiency across various multimedia tasks [58].

LoRA has shown transformative potential in enhancing model performance across sectors, including healthcare, e-commerce, NLP, and multimedia, by facilitating efficient parameter adaptation and optimization. This technique enables fine-tuning of large language models (LLMs) with minimal computational resources, particularly advantageous in resource-constrained or distributed environments. Recent advancements, such as FedEx-LoRA, further demonstrate its applicability in federated learning scenarios, where precise parameter updates are crucial for maintaining model accuracy across diverse client datasets. LoRA's innovative approach to model adaptation streamlines the fine-tuning process while significantly boosting performance in specialized applications [11, 1].



(a) Image Description[60]

(b) A diagram illustrating the interaction between a hospital, school, and bank through neural networks and databases.[38]

Figure 4: Examples of Domain-Specific Applications

As shown in Figure 4, the exploration of domain-specific applications through case studies provides valuable insights into the effective implementation of tailored solutions across various sectors. The first example, "Image Description," presents a flowchart outlining the process of selecting and training a task expert system using K-shot data from datasets such as EleutherAI, gsm8k, and MBPP. This approach emphasizes the significance of utilizing specialized data to enhance model training for specific tasks. The second example visualizes a network of interconnected nodes representing entities like hospitals, schools, and banks, showcasing the potential of neural networks and databases in facilitating efficient data flow and interaction across diverse sectors. Collectively, these examples demonstrate the practical applications of domain-specific solutions in optimizing processes and fostering innovation in various fields [60, 38].

6 Challenges and Future Directions

Understanding the challenges of Low-Rank Adaptation (LoRA) is vital for its effective use in large-scale machine learning models. This section delves into the complexities of implementing LoRA, focusing on data and model complexity issues that affect scalability and performance. By identifying these challenges, we can highlight areas for innovation that aid the integration of LoRA in diverse modeling scenarios. The following subsection discusses specific issues related to data and model complexity, emphasizing their implications for scalability and performance in machine learning systems.

6.1 Data and Model Complexity Challenges

Implementing LoRA in large-scale models faces challenges related to data and model complexity, impacting scalability and performance. A major concern is increased inference latency in multi-task learning scenarios, as shown by MTL-LoRA, which may underperform compared to merged LoRA methods [2]. This latency highlights the need for more efficient adaptation strategies to mitigate performance trade-offs.

The scalability of Parameter Efficient Fine-Tuning (PEFT) methods, including LoRA, across diverse model architectures is crucial. Current studies often face performance-computational efficiency trade-

AI-generated, for reference only.

offs, limiting adaptability across domains. The performance degradation in some implementations, such as a -3.56% accuracy drop compared to LoRA Fine-Tuning, illustrates the tension between compression and accuracy [28].

In multimodal learning, the assumption of approximate conditional independence among modalities, as required by methods like greedy modality selection, may not consistently hold, leading to suboptimal performance [42]. Furthermore, the inconsistent manifestation of the Low-Norm Effect when using soft prompts in Vision-Language Models (VLMs) across datasets poses challenges to model robustness [4].

The robustness of multimodal models is further compromised by the absence of input modalities, resulting in significant performance drops with traditional methods [6]. This issue is exacerbated by difficulties in determining optimal pruning ratios for various tasks, as retaining more tokens may be necessary for tasks reliant on positional information [7].

Addressing challenges related to data management, model architecture flexibility, and adaptation strategies is essential for effectively implementing LoRA in large-scale model adaptations across diverse domains. Developing innovative approaches to enhance these processes, particularly in resource-constrained or federated learning environments, is necessary. By leveraging LoRA's capacity to minimize trainable parameters and computational overhead while maintaining performance, new strategies can ensure precise adaptation and deployment of large language models (LLMs) tailored to specific tasks [61, 10, 11, 12, 1].

6.2 Scalability and Adaptability Limitations

LoRA's scalability and adaptability in diverse model environments face significant limitations. A key challenge is the dependency on accurate parameter sensitivity estimation, which can be particularly difficult in complex models [62]. Inaccuracies in this estimation may lead to suboptimal model performance and hinder efficient adaptation across different tasks and domains.

Integrating LoRA into large-scale models often encounters scalability issues due to inherent trade-offs between model complexity and computational efficiency. While LoRA aims to reduce the computational burden of fine-tuning, its scalability is constrained by the necessity to balance parameter efficiency with model accuracy, especially in environments with large data volumes and diverse modalities. Maintaining an optimal balance in LoRA's design is essential for effective application across various model architectures, enabling seamless adaptation without sacrificing performance or incurring additional computational costs, thus addressing challenges posed by traditional methods like Adapter, which often lead to increased inference latency and training instability [61, 11, 28, 12, 1].

Additionally, LoRA's adaptability is limited by its reliance on specific model architectures and training paradigms, which may not universally apply across all model types. This limitation is particularly pronounced in scenarios requiring rapid adaptation to new data or tasks, necessitating more flexible and dynamic strategies. The rigidity in LoRA's adaptation processes can hinder model performance in dynamic learning environments, where its reliance on a fixed number of parameters may compromise scalability and responsiveness to evolving data [11, 28].

Future research should focus on developing robust parameter sensitivity estimation techniques and exploring alternative adaptation strategies that enhance LoRA's flexibility and efficiency across diverse environments. Innovations like Progressive Compression LoRA (PC-LoRA) and FedEx-LoRA illustrate LoRA's potential in model compression and federated learning contexts, reinforcing its impact across multiple fields [10, 12, 11, 28].

6.3 Integration with Other Techniques

Integrating LoRA with complementary techniques presents a promising pathway for enhancing model performance and efficiency across various applications. The Context-PEFT method exemplifies the potential for integration with other parameter-efficient fine-tuning techniques, indicating substantial improvements in model performance and efficiency through strategic synergies [13]. Future research should refine the integration of prompts with LoRA, as demonstrated by DPD-LoRA, to bolster model robustness and efficiency in diverse learning environments [16].

AI-generated, for reference only.

In vision-language models, combining LoRA with techniques that enhance zero-shot capabilities may significantly improve adaptability and performance, particularly in vision-language few-shot adaptation scenarios [20]. Additionally, exploring adaptive learning strategies and incorporating additional modalities could broaden the applicability of LoRA-enhanced models, as suggested by advancements in cross-modal adapter frameworks [32].

Further investigation into discrete prompt tuning techniques could enhance LoRA's applicability to various black-box models, facilitating broader adoption and improved adaptation performance [33]. Moreover, integrating LoRA with compression techniques like UPop could optimize model performance and efficiency, especially in resource-constrained environments [24].

In privacy-sensitive contexts, the application of FedEx-LoRA presents opportunities for enhancing model adaptability and performance, with potential adaptations for other model types [1]. Additionally, integrating FoPru with other inference optimization techniques could further enhance efficiency in Vision Large Language Models (LVLMs), paving the way for more streamlined model implementations [7].

Lastly, refining the Distribution Discriminative Auto-Selector (DDAS) for better management of multiple tasks and enhancing the zero-shot transfer ability of models like CLIP could significantly improve continual learning capabilities in vision-language contexts [63]. These strategic integrations hold significant potential for advancing LoRA-enhanced models, ensuring robust performance and efficiency across a wide range of applications and domains.

6.4 Future Directions in Benchmarking and Evaluation

Benchmark	Size	Domain	Task Format	Metric
LHRS-Bench[64]	690	Remote Sensing	Single-choice Questions	Accuracy
MMRB[65]	60,763	Multimodal Learning	Classification	Accuracy, AUC
ROBUST-VL[66]	1,000,000	Visual Question Answering	Visual Question Answering	Relative Robustness
VL-Benchmark[67]	1,000,000	Safety Evaluation	Safety Assessment	ASR, RR
LLaVA[68]	8,000	Language Reasoning	Question Answering	Accuracy, F1-score
MML[46]	7,216	Crisis Information Detection	Multiclass Classification	F1-score, Accuracy
MMStar[57]	1,500	Visual Question Answering	Multi-modal Evaluation	Accuracy, MG
GeoLLaVA[69]	100,000	Geographical Information Systems	Temporal Change Detection	BERT, ROUGE-1

Table 3: This table provides a comprehensive overview of various benchmarks used in evaluating LoRA-enhanced models across different domains and tasks. It includes information on the benchmark name, dataset size, domain, task format, and the metrics used for evaluation, highlighting the diversity and complexity of the evaluation landscape.

Future research in benchmarking and evaluation of LoRA-enhanced models should emphasize optimizing model architectures and exploring new methodologies to ensure robust performance across diverse applications. Optimizing FLoRA's architecture, particularly in limited data scenarios, and extending it to additional modalities represents a promising direction for enhancing adaptability and efficiency [22]. Similarly, exploring layer optimization in LLaMA-Adapter and generalizing zero-initialized attention mechanisms to other architectures could significantly enhance performance across various tasks [14].

Refining the MMP framework and extending its applicability to other multimodal tasks and datasets presents ongoing research opportunities, underscoring the need for continuous innovation in model adaptability and robustness [31]. Additionally, enhancing LoRA's adaptability for various tasks, improving training efficiency, and addressing challenges related to model updates are critical areas for future exploration [11].

In multi-task learning, optimizing MTL-LoRA's design to reduce inference time while maintaining effectiveness is essential, highlighting the importance of efficient adaptation strategies [2]. Furthermore, investigating emerging trends in Parameter Efficient Fine-Tuning (PEFT) and exploring its integration with novel model architectures are vital for advancing real-world applications [5].

Future directions also include refining decay factor schedulers and enhancing initialization techniques for low-rank adapters, crucial for improving adaptation and performance [28]. Extending frameworks like HyperMM to other domains aligns with the broader goal of enhancing benchmarking and evaluation methods for LoRA-enhanced models, ensuring their applicability across diverse environments

[6]. Table 3 presents a detailed summary of representative benchmarks employed in the assessment of LoRA-enhanced models, illustrating their applicability across a range of domains and tasks.

These research directions will significantly improve the benchmarking and evaluation of LoRA-enhanced models, addressing challenges such as task interference in multi-task learning and inefficiencies in federated learning environments. By focusing on robustness and efficiency across diverse applications and domains, these advancements will facilitate the development of more scalable and adaptable machine learning systems, ultimately enhancing performance in resource-constrained and distributed settings [2, 11, 12, 1, 23].

7 Conclusion

Low-Rank Adaptation (LoRA) emerges as a pivotal technique in enhancing parameter efficiency and performance in multimodal learning environments. By optimizing parameter space and facilitating effective cross-modal integration, LoRA addresses the computational demands of large-scale models, ensuring consistent performance across diverse tasks. Techniques like DualAdapter highlight LoRA's capacity to excel in few-shot learning and domain generalization, marking significant progress in multimodal applications.

Innovative approaches such as BA-LoRA demonstrate substantial improvements over conventional fine-tuning by reducing pre-training biases and boosting the robustness and generalization of language models. In specialized domains, the SeLoRA method showcases LoRA's potential in generating high-fidelity medical images, emphasizing its adaptability and precision. Furthermore, the integration of LoRA with pruning techniques in the PAR framework markedly enhances inference efficiency for Vision Large Language Models (LVLMs), offering versatile solutions for various applications.

The application of LoRA in recommender systems, especially under conditions of incomplete data, underscores its superior performance and robustness, providing valuable insights for multimodal learning strategies. Collectively, these advancements underscore LoRA's transformative role in improving model adaptability and efficiency.

Future research directions, such as exploring redundancy reduction techniques like TIVE, will be crucial for further refining model performance and efficiency. These developments underscore the ongoing importance of advancing LoRA methodologies to maintain their relevance and effectiveness in the evolving landscape of multimodal learning.

AI-generated, for reference only.

References

[1] Raghav Singhal, Kaustubh Ponkshe, and Praneeth Vepakomma. Fedex-lora: Exact aggregation for federated parameter-efficient fine-tuning of foundation models. In *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*.

[2] Yaming Yang, Dilxat Muhtar, Yelong Shen, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Denvy Deng, Feng Sun, Qi Zhang, Weizhu Chen, and Yunhai Tong. Mtl-lora: Low-rank adaptation for multi-task learning, 2024.

[3] Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. *arXiv preprint arXiv:2402.02207*, 2024.

[4] Shuai Fu, Xiequn Wang, Qiushi Huang, and Yu Zhang. Nemesis: Normalizing the soft-prompt vectors of vision-language models, 2024.

[5] Luping Wang, Sheng Chen, Linnan Jiang, Shu Pan, Runze Cai, Sen Yang, and Fei Yang. Parameter-efficient fine-tuning in large models: A survey of methodologies. *arXiv preprint arXiv:2410.19878*, 2024.

[6] Hava Chaptoukaev, Vincenzo Marcianó, Francesco Galati, and Maria A. Zuluaga. Hypermm : Robust multimodal learning with varying-sized inputs, 2024.

[7] Lei Jiang, Weizhe Huang, Tongxuan Liu, Yuting Zeng, Jing Li, Lechao Cheng, and Xiaohua Xu. Fopru: Focal pruning for efficient large vision-language models, 2024.

[8] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024.

[9] Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-lora, 2024.

[10] 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, 2021.

[11] Vlad Fomenko, Han Yu, Jongho Lee, Stanley Hsieh, and Weizhu Chen. A note on lora, 2024.

[12] Raghav Singhal, Kaustubh Ponkshe, and Praneeth Vepakomma. Fedex-lora: Exact aggregation for federated and efficient fine-tuning of foundation models.

[13] Avelina Asada Hadji-Kyriacou and Ognjen Arandjelovic. Context-peft: Efficient multi-modal, multi-task fine-tuning. *arXiv preprint arXiv:2312.08900*, 2023.

[14] Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. Llama-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*, 2024.

[15] Oindrila Saha, Grant Van Horn, and Subhransu Maji. Improved zero-shot classification by adapting vlms with text descriptions, 2024.

[16] Chushan Zhang, Ruihan Lu, Zeeshan Hayder, and Hongdong Li. Dpd-lora: Dynamic prompt-driven low-rank adaptation for improved generalization.

[17] Chengyu Wang, Junbing Yan, Wei Zhang, and Jun Huang. Towards better parameter-efficient fine-tuning for large language models: A position paper. *arXiv preprint arXiv:2311.13126*, 2023.

[18] Yuchen Mao, Hongwei Li, Wei Pang, Giorgos Papanastasiou, Guang Yang, and Chengjia Wang. Selora: Self-expanding low-rank adaptation of latent diffusion model for medical image synthesis. *arXiv preprint arXiv:2408.07196*, 2024.

[19] Muhammad Arslan Manzoor, Sarah Albarri, Ziting Xian, Zaiqiao Meng, Preslav Nakov, and Shangsong Liang. Multimodality representation learning: A survey on evolution, pretraining and its applications, 2024.

AI-generated, for reference only.

[20] Ce Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Enhancing vision-language few-shot adaptation with negative learning, 2024.

[21] Zihan Zhou, Yang Zhou, Tianshi Che, Zeru Zhang, Jiaxiang Ren, Da Yan, Zhe Jiang, Ruoming Jin, Jianfeng Gao, et al. Riemannian low-rank adaptation for federated fine-tuning of foundation models.

[22] Shruti Palaskar, Oggi Rudovic, Sameer Dharur, Florian Pesce, Gautam Krishna, Aswin Sivaraman, Jack Berkowitz, Ahmed Hussen Abdelaziz, Saurabh Adya, and Ahmed Tewfik. Multimodal large language models with fusion low rank adaptation for device directed speech detection, 2024.

[23] Zheyu Shen, Guoheng Sun, Yexiao He, Ziyao Wang, Yuning Zhang, Souvik Kundu, Eric P Xing, Hongyi Wang, and Ang Li. Sharelora: Less tuning, more performance for lora fine-tuning of llms.

[24] Dachuan Shi, Chaofan Tao, Ying Jin, Zhendong Yang, Chun Yuan, and Jiaqi Wang. Upop: Unified and progressive pruning for compressing vision-language transformers. In *International Conference on Machine Learning*, pages 31292–31311. PMLR, 2023.

[25] Wei Suo, Ji Ma, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. Pruning all-rounder: Rethinking and improving inference efficiency for large vision language models, 2024.

[26] Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. Atp-llava: Adaptive token pruning for large vision language models, 2024.

[27] Alex Jinpeng Wang, Linjie Li, Kevin Qinghong Lin, Jianfeng Wang, Kevin Lin, Zhengyuan Yang, Lijuan Wang, and Mike Zheng Shou. Cosmo: Contrastive streamlined multimodal model with interleaved pre-training, 2024.

[28] Injoon Hwang, Haewon Park, Youngwan Lee, Jooyoung Yang, and SunJae Maeng. Pc-lora: Low-rank adaptation for progressive model compression with knowledge distillation, 2024.

[29] Huimin Zeng, Zhenrui Yue, and Dong Wang. Open-vocabulary federated learning with multimodal prototyping, 2024.

[30] Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Raghavi Chandu. Selective "selective prediction": Reducing unnecessary abstention in vision-language reasoning, 2024.

[31] Niki Nezakati, Md Kaykobad Reza, Ameya Patil, Mashhour Solh, and M. Salman Asif. Mmp: Towards robust multi-modal learning with masked modality projection, 2024.

[32] Juncheng Yang, Zuchao Li, Shuai Xie, Weiping Zhu, Wei Yu, and Shijun Li. Cross-modal adapter: Parameter-efficient transfer learning approach for vision-language models, 2024.

[33] Zixian Guo, Yuxiang Wei, Ming Liu, Zhilong Ji, Jinfeng Bai, Yiwen Guo, and Wangmeng Zuo. Black-box tuning of vision-language models with effective gradient approximation, 2023.

[34] Guoqing Zhao, Qi Zhang, Shaopeng Zhai, Dazhong Shen, Yu Qiao, Tong Xu, et al. I-lora: Iterative merging of routing-tuned low-rank adapters for multi-task learning.

[35] Ibrahim Alabdulmohsin, Xiao Wang, Andreas Steiner, Priya Goyal, Alexander D'Amour, and Xiaohua Zhai. Clip the bias: How useful is balancing data in multimodal learning?, 2024.

[36] Zitao Shuai and Liyue Shen. Align as ideal: Cross-modal alignment binding for federated medical vision-language pre-training, 2024.

[37] Xiaolong Jin, Kai Wang, Dongwen Tang, Wangbo Zhao, Yukun Zhou, Junshu Tang, and Yang You. Conditional lora parameter generation, 2024.

[38] Yicheng Zhang, Zhen Qin, Zhaomin Wu, and Shuiguang Deng. Personalized federated fine-tuning for llms via data-driven heterogeneous model architectures. *arXiv preprint arXiv:2411.19128*, 2024.

[39] Bin Lin, Zhenyu Tang, Yang Ye, Jinfa Huang, Junwu Zhang, Yatian Pang, Peng Jin, Munan Ning, Jiebo Luo, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models, 2024.

[40] Paul Pu Liang, Peter Wu, Liu Ziyin, Louis-Philippe Morency, and Ruslan Salakhutdinov. Cross-modal generalization: Learning in low resource modalities via meta-alignment, 2020.

[41] Bozhou Li, Hao Liang, Zimo Meng, and Wentao Zhang. Are bigger encoders always better in vision large models?, 2024.

[42] Runxiang Cheng, Gargi Balasubramaniam, Yifei He, Yao-Hung Hubert Tsai, and Han Zhao. Greedy modality selection via approximate submodular maximization, 2022.

[43] Zixuan Hu, Yongxian Wei, Li Shen, Chun Yuan, and Dacheng Tao. Unlocking tuning-free few-shot adaptability in visual foundation models by recycling pre-tuned loras. *arXiv preprint arXiv:2412.02220*, 2024.

[44] Sayna Ebrahimi, Sercan O. Arik, Yihe Dong, and Tomas Pfister. Lanistr: Multimodal learning from structured and unstructured data, 2024.

[45] Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts, 2024.

[46] Gaurav Verma, Rohit Mujumdar, Zijie J. Wang, Munmun De Choudhury, and Srijan Kumar. Overcoming language disparity in online content classification with multimodal learning, 2022.

[47] Chongjie Si, Xuehui Wang, Xue Yang, Zhengqin Xu, Qingyun Li, Jifeng Dai, Yu Qiao, Xiaokang Yang, and Wei Shen. Maintaining structural integrity in parameter spaces for parameter efficient fine-tuning. In *The Thirteenth International Conference on Learning Representations*.

[48] Wenke Huang, Jian Liang, Zekun Shi, Didi Zhu, Guancheng Wan, He Li, Bo Du, Dacheng Tao, and Mang Ye. Learn from downstream and be yourself in multimodal large language model fine-tuning. *arXiv preprint arXiv:2411.10928*, 2024.

[49] Omiros Pantazis, Gabriel Brostow, Kate Jones, and Oisin Mac Aodha. Svl-adapter: Self-supervised adapter for vision-language pretrained models, 2022.

[50] Alan Ansell, Ivan Vulić, Hannah Sterz, Anna Korhonen, and Edoardo M. Ponti. Scaling sparse fine-tuning to large language models, 2024.

[51] Manogna Sreenivas and Soma Biswas. Effectiveness of vision language models for open-world single image test time adaptation, 2024.

[52] Yupeng Chang, Yi Chang, and Yuan Wu. Ba-lora: Bias-alleviating low-rank adaptation to mitigate catastrophic inheritance in large language models. *arXiv preprint arXiv:2408.04556*, 2024.

[53] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multimodal learning, 2022.

[54] Shwai He, Ang Li, and Tianlong Chen. Rethinking pruning for vision-language models: Strategies for effective sparsity and performance restoration, 2024.

[55] Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*, 2024.

[56] Zida Cheng, Chen Ju, Shuai Xiao, Xu Chen, Zhonghua Zhai, Xiaoyi Zeng, Weilin Huang, and Junchi Yan. Category-oriented representation learning for image to multi-modal retrieval, 2024.

[57] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024.

[58] Wei-Hsiang Yu, Yen-Yu Lin, Ming-Hsuan Yang, and Yi-Hsuan Tsai. Ranking-aware adapter for text-driven image ordering with clip, 2025.

[59] Jialin Li, Qiang Nie, Weifu Fu, Yuhuan Lin, Guangpin Tao, Yong Liu, and Chengjie Wang. Lors: Low-rank residual structure for parameter-efficient network stacking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15866–15876, 2024.

[60] Yuncheng Yang, Yulei Qin, Tong Wu, Zihan Xu, Gang Li, Pengcheng Guo, Hang Shao, Yuchen Shi, Ke Li, Xing Sun, et al. Leveraging open knowledge for advancing task expertise in large language models. *arXiv preprint arXiv:2408.15915*, 2024.

[61] Raghav Singhal, Kaustubh Ponkshe, and Praneeth Vepakomma. Exact aggregation for federated and efficient fine-tuning of foundation models. *arXiv preprint arXiv:2410.09432*, 2024.

[62] Tom Pégeot. *Efficient Transfer Learning Towards Constrained Environments*. PhD thesis, Université Paris-Saclay, 2024.

[63] Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters, 2024.

[64] Dilxat Muhtar, Zhenshi Li, Feng Gu, Xueliang Zhang, and Pengfeng Xiao. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. In *European Conference on Computer Vision*, pages 440–457. Springer, 2024.

[65] Maciej Pawłowski, Anna Wróblewska, and Sylwia Sysko-Romańczuk. Does a technique for building multimodal representation matter? – comparative analysis, 2022.

[66] Shuo Chen, Jindong Gu, Zhen Han, Yunpu Ma, Philip Torr, and Volker Tresp. Benchmarking robustness of adaptation methods on pre-trained vision-language models. *Advances in Neural Information Processing Systems*, 36:51758–51777, 2023.

[67] Seongyun Lee, Geewook Kim, Jiyeon Kim, Hyunji Lee, Hoyeon Chang, Sue Hyun Park, and Minjoon Seo. How does vision-language adaptation impact the safety of vision language models?, 2024.

[68] Neale Ratzlaff, Man Luo, Xin Su, Vasudev Lal, and Phillip Howard. Training-free mitigation of language reasoning degradation after multimodal instruction tuning, 2024.

[69] Hosam Elgendy, Ahmed Sharshar, Ahmed Aboeitta, Yasser Ashraf, and Mohsen Guizani. Geollava: Efficient fine-tuned vision-language models for temporal change detection in remote sensing, 2024.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.