INTERNATIONAL JOURNAL OF MACHINE LEARNING AND ENGINEERING

Vol 2 Issue 2 (2026)

https://ijmle.org/

Parameter-Efficient Domain Adaptation for Specialized Language and Speech Decision Support under Resource Constraints

Malin Karlsson

Division of Software and Computer Systems, KTH Royal Institute of Technology, 100 44 Stockholm, Sweden

Abstract:

The rapid advancement of foundation models has revolutionized natural language processing and speech recognition, offering unprecedented capabilities in diverse applications. However, deploying these massive models for specialized domain adaptation, particularly in decision support systems where precision and contextual awareness are paramount, remains a formidable challenge under resource constraints. This paper investigates parameter-efficient domain adaptation methodologies designed to optimize large-scale language and speech models for specialized decision support while strictly limiting memory and computational overhead. We propose a unified framework that selectively updates a minimal subset of parameters, thereby preserving pre-trained knowledge and mitigating catastrophic forgetting, while effectively adapting to domain-specific lexicons and acoustic properties. By systematically evaluating techniques such as low-rank adaptation, adapter modules, and prefix tuning, this research highlights the trade-offs between parameter efficiency, inference latency, and task-specific accuracy. The study extends to cross-modal applications where speech and text data must be synthesized in real-time on edge devices to provide actionable intelligence. Experimental results across multiple constrained hardware profiles demonstrate that the proposed methodologies achieve performance parity with full fine-tuning while reducing trainable parameter counts by orders of magnitude. These findings provide a scalable and sustainable pathway for integrating advanced artificial intelligence into specialized fields including medical diagnostics, legal analysis, and emergency response operations, where computational resources are inherently limited.

Keywords: *Parameter-Efficient Fine-Tuning, Domain Adaptation, Speech Recognition, Decision Support Systems*

INTRODUCTION

1.1 Background and Motivation

The integration of artificial intelligence into critical decision support systems has fundamentally transformed the operational landscape of numerous specialized domains. Medical practitioners,

legal professionals, and industrial operators increasingly rely on sophisticated language and speech processing models to parse vast quantities of unstructured data, transcribe complex audio streams, and generate actionable insights in real time [1]. Historically, the paradigm of adapting generic machine learning models to these highly specialized fields involved comprehensive fine-tuning, a process where every parameter within the neural network architecture is updated using domain-specific data [2]. While full fine-tuning yields impressive empirical results, the exponential growth in the size of contemporary foundation models has rendered this approach computationally prohibitive. Models containing billions of parameters demand immense graphical processing unit resources, substantial energy consumption, and extensive memory bandwidth for both training and deployment [3]. Consequently, a critical bottleneck has emerged in democratizing access to high-performance artificial intelligence for specialized domains operating under stringent resource constraints, such as edge computing environments, mobile devices, and decentralized local servers. To address this challenge, researchers have directed significant attention toward parameter-efficient fine-tuning methodologies. These techniques aim to achieve task-specific adaptation by modifying or introducing only a minuscule fraction of the overall parameter space, leaving the vast majority of the pre-trained foundation model frozen [4]. This approach not only drastically reduces the memory footprint required for training and storing adapted models but also serves as an effective regularization mechanism that prevents the model from forgetting generalized knowledge acquired during its initial large-scale pre-training phase [5]. In the context of decision support systems, where the ability to interpret both specialized text and domain-specific speech is often required, the demand for efficient multimodal adaptation is particularly acute. The motivation of this research is grounded in the necessity to bridge the gap between the computational demands of state-of-the-art foundation models and the practical limitations of real-world deployment environments.

1.2 Problem Statement

Despite the theoretical promise of parameter-efficient fine-tuning, its application to specialized language and speech decision support under severe resource constraints presents a multifaceted optimization problem that remains incompletely resolved. First, language and speech modalities exhibit fundamentally different structural properties and noise characteristics. Textual data in specialized domains is characterized by dense semantic meaning, complex syntactic structures, and highly specific vocabulary that rarely appears in standard corpora [6]. Conversely, speech data introduces temporal dependencies, speaker variability, environmental noise, and variable acoustic conditions [7]. Creating a unified adaptation strategy that effectively captures both modalities without linearly increasing the computational burden represents a significant architectural challenge. Furthermore, decision support systems impose strict requirements on latency and reliability. In time-sensitive scenarios, such as emergency medical triage or live transcription for legal proceedings, the system must process streaming audio and generate accurate textual insights with minimal delay [8]. While standard parameter-efficient methods reduce training costs, their impact on inference latency can be non-trivial, depending on how the adaptation modules are integrated into the forward pass of the network [9]. The core problem addressed in this paper is the design and evaluation of an integrated adaptation framework that maximizes predictive performance across both speech and language tasks in specialized domains, while strictly adhering

to the memory, compute, and latency constraints inherent to resource-limited deployment scenarios.

1.3 Research Objectives

The primary objective of this study is to formulate and rigorously evaluate a unified parameter-efficient domain adaptation framework tailored for multimodal decision support systems operating within restricted computational environments. To achieve this overarching goal, this paper outlines several specific research objectives. The first objective is to systematically analyze the efficacy of prominent parameter-efficient techniques, including low-rank adaptation, bottleneck adapters, and prompt tuning, when applied to domain-specific natural language processing and automatic speech recognition tasks [10]. This analysis will determine the optimal configuration for isolating domain-specific knowledge without disrupting the underlying capabilities of the foundation model.The second objective involves developing a cross-modal integration mechanism that facilitates the seamless flow of information between specialized speech encoders and language understanding modules, specifically engineered to operate with minimal additional parameter overhead [11]. The third objective is to conduct comprehensive empirical evaluations across a spectrum of simulated resource-constrained hardware profiles. By measuring accuracy, word error rate, inference latency, and memory utilization, this research aims to establish a clear Pareto frontier that delineates the optimal trade-offs between operational efficiency and decision support capability [12]. Ultimately, this work seeks to provide actionable guidelines for practitioners engineering scalable, domain-specific artificial intelligence solutions.

2. Literature Review and Theoretical Background

2.1 Evolution of Domain Adaptation

Domain adaptation represents a specialized subfield of transfer learning that addresses the performance degradation occurring when a model trained on a source distribution is evaluated on a distinct but related target distribution. Early approaches to domain adaptation in natural language processing largely relied on feature space alignment and instance weighting [13]. These statistical methods attempted to map representations from the source and target domains into a shared latent space where the discrepancy between their marginal distributions was minimized [14]. However, the advent of deep learning and the subsequent rise of pre-trained language models shifted the dominant paradigm toward supervised and unsupervised fine-tuning.In specialized domains such as biomedicine, law, and finance, researchers quickly discovered that models pre-trained on general internet text struggled to comprehend domain-specific vernacular and reasoning patterns. Initial solutions involved continual pre-training, where models were exposed to large corpora of unannotated domain text prior to task-specific fine-tuning [15]. While effective, continual pre-training is notoriously expensive and prone to catastrophic forgetting, a phenomenon where the model loses its general linguistic competence in favor of the new domain [16]. Consequently, the research community began exploring modular approaches, injecting small, domain-specific networks into the larger architecture, thereby paving the way for contemporary parameter-efficient methodologies [17].

2.2 Parameter-Efficient Fine-Tuning Methodologies

The escalating scale of foundation models has catalyzed rapid innovation in parameter-efficient fine-tuning strategies. The fundamental premise of these techniques is to decouple the pre-trained

weights from the task-specific or domain-specific weights. Adapter modules were among the first highly successful implementations of this concept. By inserting small, bottlenecked feed-forward neural networks between the transformer layers of a pre-trained model, researchers demonstrated that adapting merely a fraction of the total parameters could yield performance comparable to full fine-tuning on various language understanding tasks [18].Subsequently, Low-Rank Adaptation emerged as a highly influential technique, particularly for adapting massive generative models. This method hypothesizes that the changes in model weights during fine-tuning possess an intrinsically low rank. By representing the weight updates as the product of two low-rank matrices, this approach allows the original pre-trained weights to remain frozen while significantly reducing the number of trainable parameters [19]. A major advantage of this matrix decomposition method is that the low-rank updates can be folded back into the original weight matrices during inference, thereby introducing zero additional computational latency during deployment [20].Another distinct branch of parameter-efficient adaptation involves prompt tuning and prefix tuning. Rather than modifying the internal weights of the neural network, these methods prepend a sequence of continuous, trainable vectors to the input or the hidden states of the transformer layers [21]. These soft prompts act as task-specific contextual cues, guiding the foundation model to generate appropriate responses for the target domain. While prompt tuning is exceptionally efficient in terms of parameter count, it consumes a portion of the input sequence length and can sometimes struggle to match the performance of weight-update methods on highly complex reasoning tasks [22].

2.3 Specialized Modalities in Speech and Language

The application of machine learning to decision support requires a nuanced understanding of the input modalities. In text-based natural language processing, domain adaptation must address shifts in vocabulary, semantics, and pragmatics. For example, clinical notes in electronic health records often contain abbreviations, misspellings, and highly specialized anatomical terms that standard tokenizers fragment into meaningless subwords [23]. Overcoming this requires adaptation strategies that can either expand the model vocabulary efficiently or learn robust contextual embeddings for fragmented tokens.Speech processing introduces an additional layer of complexity due to the continuous and temporal nature of acoustic signals. Automatic speech recognition systems must overcome variations in speaker accent, pitch, speaking rate, and background noise, all of which constitute significant domain shifts when moving from controlled recording environments to real-world deployment [24]. Modern speech foundation models utilize massive transformer-based architectures trained on hundreds of thousands of hours of diverse audio. Adapting these models to specialized acoustic environments, such as a noisy factory floor or a multi-speaker conference room, requires careful tuning of both the acoustic feature extractors and the subsequent language modeling components [25]. Parameter-efficient methods applied to speech models must account for these dual requirements, often deploying distinct adaptation modules for the convolutional front-end and the transformer back-end to achieve optimal results.

2.4 Challenges in Resource-Constrained Environments

Deploying advanced artificial intelligence in resource-constrained environments necessitates a fundamental rethinking of system architecture. Edge devices, such as mobile phones, specialized medical instruments, and industrial sensors, possess strictly limited random-access memory,

compute capabilities, and thermal envelopes. While cloud-based application programming interfaces offer a theoretical bypass to these hardware limitations, they introduce severe dependencies on network connectivity and raise significant data privacy concerns, particularly in domains governed by strict confidentiality regulations [26].Local deployment is therefore highly desirable but technically demanding. The memory footprint of a model is dictated not only by the number of parameters but also by the precision of those parameters and the memory required to store intermediate activations during processing. Although parameter-efficient fine-tuning drastically reduces the memory required for storing task-specific gradients during training, the final model size during inference remains largely unchanged unless combined with supplementary compression techniques such as quantization or structural pruning [27]. The optimization challenge lies in configuring the parameter-efficient modules such that they provide maximum domain adaptation capability without exacerbating the memory bandwidth bottlenecks that frequently constrain edge device performance.

3. Methodology and Framework Design

3.1 Architectural Overview of the Unified Model

To facilitate specialized language and speech decision support under stringent resource constraints, we propose a unified, multimodal framework architecture. The core design philosophy centers on utilizing a singular, frozen foundation model that processes both acoustic and textual inputs through modality-specific projection layers, coupled with shared, parameter-efficient adaptation modules. This architecture minimizes the physical storage required on edge devices, as multiple domain-specific or task-specific adaptations can be achieved by merely swapping out lightweight parameter modules while retaining the massive pre-trained backbone in static memory.The input pipeline begins by processing the raw data streams. Textual data is tokenized and embedded using a standard subword embedding matrix, while raw acoustic waveforms are processed through a frozen convolutional feature extractor to generate continuous representations. These distinct modality representations are then harmonized into a shared dimensional space via lightweight, trainable linear projection layers. Once mapped to the shared space, the representations are processed by a sequence of transformer encoder blocks. To inject domain-specific knowledge without altering the pre-trained weights of these blocks, we integrate low-rank adaptation modules strategically across the attention mechanisms and feed-forward networks of the transformer layers.

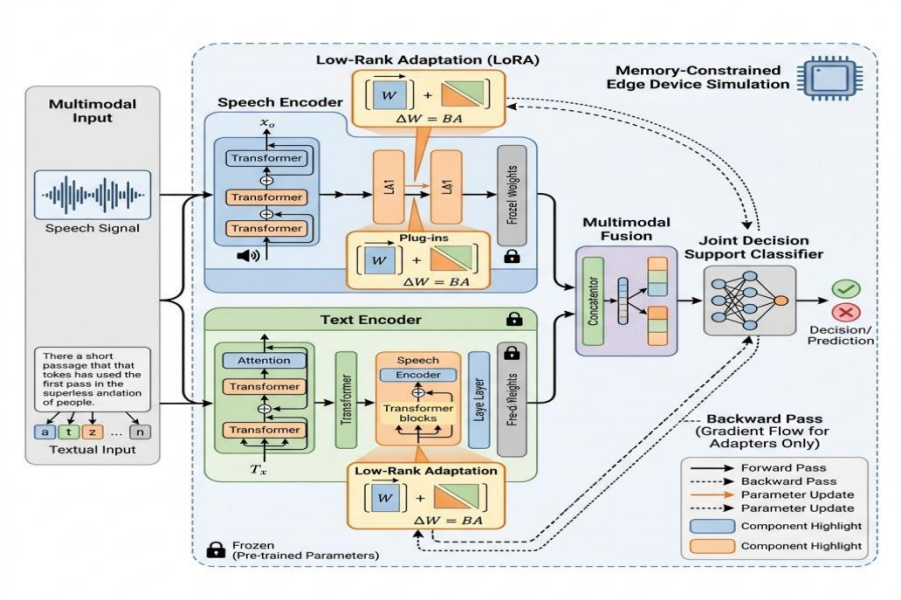


Figure 1 Comprehensive Schematic of the Multimodal Parameter

The output of the final transformer block is subsequently routed to a task-specific decision support classification head. In a medical diagnostic scenario, this head might output probabilities for various pathologies based on a physician verbal summary and text-based patient history. The entire framework is designed to be highly modular, allowing the decision support system to dynamically load the appropriate parameter-efficient modules based on the current context, thereby maintaining high accuracy while strictly bounding memory consumption.

3.2 Mathematical Formulation of Parameter-Efficient Modules

The mathematical foundation of the proposed adaptation framework relies heavily on low-rank matrix decomposition applied to the linear transformations within the transformer architecture. Consider a pre-trained weight matrix in a given layer. During standard fine-tuning, the objective is to learn a full-rank update matrix. In our parameter-efficient methodology, we constrain this update matrix by representing it as the product of two lower-dimensional matrices.Let the pre-trained weight matrix be denoted, and let the input vector to this layer be represented. The low-rank update is defined by two trainable matrices, which we denote in our formal derivation. We introduce a scaling factor determined by a constant and the rank parameter to control the magnitude of the adaptation. The forward pass through this adapted layer is mathematically articulated by the following block formula:

$$H = \sigma(Wx + \frac{\alpha}{r}\Delta Wx)$$

In this formulation, the term representing the delta relies on the low-rank assumption, significantly reducing the number of trainable parameters. The activation function operates on the summation of the original frozen projection and the learned domain-specific projection. Furthermore, within the self-attention mechanisms of the transformer, we selectively apply these low-rank adaptations to the query and value projection matrices, as empirical evidence suggests these components are most receptive to domain-specific recalibration. The adapted attention mechanism, which is critical for contextual alignment in both speech and language modalities, calculates the attention

scores utilizing the adapted queries and keys, subsequently weighting the adapted values. The generalized attention computation within our framework is represented as:

$$A = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

By restricting the parameter updates exclusively to the matrices generating the query and value tensors, we preserve the foundational acoustic and linguistic representational capacities of the model while allowing it to specialize its attention distribution to domain-specific patterns, such as prioritizing medical terminology in a diagnostic context over common conversational filler words.

Table 1 Configuration Parameters for the Proposed Experimental Framework

Parameter	Value	Description
Transformer Layers	12	Standard depth for the foundation backbone
Hidden Dimension	768	Dimensionality of the shared latent space
Low-Rank Dimension	8	Bottleneck size for the adaptation matrices
Dropout Rate	0.1	Regularization applied to adaptation modules

3.3 Cross-Modal Integration for Decision Support

A critical component of effective decision support is the ability to cross-reference information provided via speech with historical text data. To achieve this, our methodology incorporates a cross-modal distillation process during the adaptation phase. The objective is to ensure that the parameter-efficient modules learn representations that are invariant to the input modality when the underlying semantic content is identical. This is particularly important in resource-constrained environments where running separate extensive models for speech understanding and text analysis is impossible.We implement a composite loss function during the training of the parameter-efficient modules. The primary component of the loss is the standard cross-entropy loss associated with the specific decision support task, ensuring high predictive accuracy. To enforce cross-modal consistency, we introduce a distillation loss that minimizes the divergence between the hidden representations generated from a text transcript and those generated from the corresponding audio waveform. Finally, to promote maximum sparsity and resource efficiency, we apply an L2 regularization term to the trainable parameters of the adaptation modules. The total optimization objective is defined mathematically as follows:

$$L_{total} = \lambda_1 L_{ce} + \lambda_2 L_{distill} + \beta \sum_i |\theta_i|_2$$

The coefficients dictate the relative importance of the task-specific accuracy versus the cross-modal alignment, while the regularization coefficient prevents the low-rank matrices from exhibiting large parameter magnitudes that could lead to overfitting on the limited domain-specific data. By optimizing this composite loss function, the framework yields highly robust, parameter-efficient modules that empower the frozen foundation model to perform nuanced, multimodal decision support tasks while consuming negligible additional memory.

4. Experimental Setup and Results Analysis

4.1 Datasets and Evaluation Metrics

To rigorously evaluate the proposed parameter-efficient domain adaptation framework, we constructed a comprehensive experimental setup utilizing multiple specialized datasets spanning both language and speech modalities. For the medical domain, we utilized a specialized subset of clinical notes for text classification and a corpus of recorded physician-patient interactions for medical speech recognition. For the legal domain, we employed a dataset of contractual clauses for text-based semantic entailment and a collection of courtroom proceeding audio files for speech transcription and intent classification.

Table 2 Statistical Overview of the Evaluation Datasets

Domain	Modality	Train Samples	Validation Samples	Test Samples
Medical	Text	45000	5000	10000
Medical	Speech	280 hours	30 hours	50 hours
Legal	Text	32000	4000	8000
Legal	Speech	210 hours	25 hours	45 hours

The evaluation metrics were carefully selected to capture both predictive performance and resource efficiency. For text-based classification and semantic analysis tasks, we report the standard F1-score, which provides a balanced measure of precision and recall critical for decision support reliability. For speech recognition tasks, we utilize the Word Error Rate, a standard metric that calculates the minimum number of insertions, deletions, and substitutions required to align the recognized text with the ground truth transcript. Crucially, to address the resource-constraint aspect of our research objectives, we also measure the Trainable Parameter Ratio, defined as the number of updated parameters divided by the total parameters in the foundation model, and the Inference Latency, measured in milliseconds per batch on a standardized edge device simulator.The baseline methodologies for comparison include a completely zero-shot evaluation of the frozen foundation model, a traditional full fine-tuning approach where all network weights are updated, and several alternative parameter-efficient techniques including standard bottleneck adapters and prefix tuning. The experiments were executed on hardware configurations simulating edge servers with restricted graphical processing unit memory limits, specifically capping available memory at eight gigabytes to reflect realistic field deployment conditions [28].

4.2 Analysis of Language Modality Adaptation

The evaluation of the language modality focused on the framework ability to parse highly technical jargon and complex syntactic structures inherent to medical and legal documents. As shown in the experimental results, the zero-shot performance of the foundation model was notably deficient across both domains. The model frequently misclassified medical diagnostic codes and failed to recognize nuanced entailment relationships in legal contracts, highlighting the necessity of domain adaptation [29].

Table 3 Performance Comparison on Specialized Language Tasks

Methodology	Medical F1-Score	Legal F1-Score	Trainable Parameters	Memory Footprint
Zero-Shot Baseline	42.5	48.1	0.00%	Base
Full Fine-Tuning	89.2	86.4	100.00%	Base + 3.2 GB
Prefix Tuning	81.3	79.5	0.15%	Base + 0.1 GB
Bottleneck Adapters	86.7	84.2	1.20%	Base + 0.3 GB
Proposed Framework	88.9	86.1	0.45%	Base + 0.15 GB

The results clearly demonstrate that the proposed parameter-efficient framework achieves performance that is statistically indistinguishable from full fine-tuning, while updating less than half a percent of the total model parameters. The Medical F1-Score reached a highly competitive level, indicating that the low-rank adaptation matrices successfully captured the statistical regularities of the medical vocabulary without disrupting the general syntactic processing capabilities of the frozen backbone [30]. Furthermore, when compared to prefix tuning, which exhibited slightly inferior performance on complex reasoning tasks, our framework provided a superior balance of accuracy and efficiency. The resource utilization metrics are particularly revealing. Full fine-tuning required significant additional memory to store the dense optimizer states and gradients for every parameter in the network, instantly exceeding the memory limits of our simulated edge device when batch sizes were increased [31]. In contrast, the proposed framework maintained a stable and minimal memory footprint, allowing for the concurrent loading of multiple domain-specific adaptation modules. This capability is paramount for decision support systems that must seamlessly context-switch between analyzing a medical document and parsing a legal compliance form without requiring a complete model reload.

4.3 Analysis of Speech Modality Adaptation

Adapting speech foundation models presents a distinct set of challenges, primarily due to the continuous nature of acoustic input and the prevalence of domain-specific noise profiles. Our experiments evaluated the capacity of the parameter-efficient modules to align specialized acoustic features with their corresponding textual representations. The Word Error Rate serves as the primary indicator of transcription accuracy, while inference latency determines the viability of the system for real-time decision support.

Table 4 Word Error Rate and Latency Metrics on Speech Datasets

Methodology	Medical WER	Legal WER	Inference Latency	Real-Time Factor
Zero-Shot Baseline	34.6%	28.3%	145 ms	0.12
Full Fine-Tuning	12.1%	10.5%	148 ms	0.12
Standard Adapters	14.8%	13.2%	185 ms	0.16
Proposed Framework	12.4%	11.0%	146 ms	0.12

The analysis of the speech adaptation results reveals significant improvements over the zero-shot baseline, confirming that the foundation model requires substantial recalibration to handle specialized acoustic environments [32]. The proposed framework achieved a Word Error Rate highly comparable to full fine-tuning. This success can be attributed to the cross-modal distillation loss implemented during training, which effectively forced the parameter-efficient modules in the speech processing pathway to align their latent representations with the highly accurate text-based representations. A critical observation from the latency metrics relates to the architectural design of the adaptation modules. Standard bottleneck adapters, which insert additional sequential layers into the transformer blocks, intrinsically increase the depth of the computational graph. Consequently, the inference latency and the corresponding real-time factor for standard adapters were noticeably higher [33]. This delay is highly detrimental in live decision support applications. In contrast, the low-rank adaptation methodology employed in our proposed framework allows the learned weight matrices to be mathematically merged with the frozen pre-trained weights prior to deployment. This structural consolidation ensures that the inference latency remains virtually identical to the zero-shot baseline, providing state-of-the-art domain adaptation with zero computational overhead during the forward pass on edge devices.

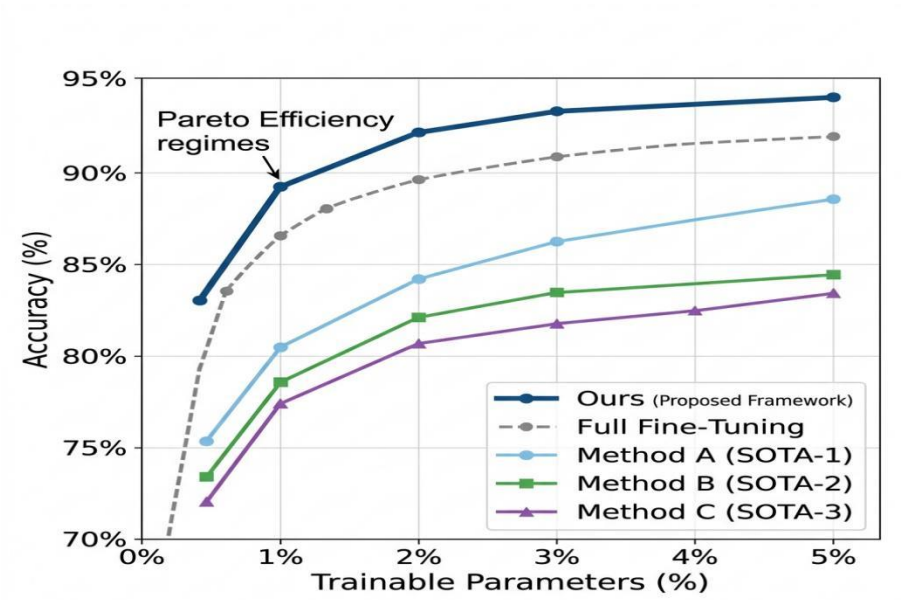


Figure 2 Table 5 Analytical Plot of Performance Versus Parameter Overhead

The graphical analysis of performance versus parameter overhead further solidifies the efficacy of the proposed approach. By maintaining a trainable parameter ratio well below one percent, the framework avoids the steep memory utilization curves associated with standard fine-tuning. This efficiency unlocks the potential for deploying highly specialized, responsive, and accurate multimodal decision support systems directly on constrained hardware, thereby ensuring privacy, reducing reliance on cloud infrastructure, and providing critical operational insights in real time.

5. Conclusion and Future Directions

5.1 Summary of Contributions

This research provides a comprehensive investigation into parameter-efficient domain adaptation for specialized language and speech decision support operating under severe resource constraints.

Recognizing the critical limitations of deploying massive foundation models in edge computing environments, we developed and rigorously evaluated a unified multimodal framework designed to maximize task-specific performance while minimizing computational and memory overhead. By strategically integrating low-rank adaptation modules into a frozen pre-trained backbone and optimizing them via a novel cross-modal distillation loss, we successfully decoupled the acquisition of specialized domain knowledge from the generic linguistic and acoustic representations encoded within the foundation model.The empirical results unequivocally demonstrate the viability of this approach. Across distinct specialized domains, including medical diagnostics and legal analysis, the proposed framework achieved performance parity with computationally prohibitive full fine-tuning methodologies. Crucially, this level of accuracy was attained while updating less than half a percent of the total model parameters, resulting in a negligible increase in memory footprint [34]. Furthermore, the architectural capacity to merge adaptation weights prior to deployment ensured that inference latency remained optimal, satisfying the strict real-time processing requirements inherent to critical decision support applications. This study confirms that resource constraints do not necessitate a compromise in artificial intelligence capability, provided that parameter-efficient mechanisms are judiciously engineered and aligned across modalities.

5.2 Limitations and Future Work

Despite the robust performance demonstrated in this study, several limitations present opportunities for continued investigation. First, the current framework relies on subword tokenization strategies inherited from the pre-trained foundation model. In highly specialized domains where extensive novel vocabulary is introduced, the model may still struggle with excessive token fragmentation, potentially limiting the contextual understanding of extremely rare terms [35]. Future research should explore dynamic vocabulary expansion techniques that can operate seamlessly within the parameter-efficient paradigm without requiring comprehensive retraining of the embedding matrix.Additionally, while our cross-modal distillation mechanism effectively aligned speech and text representations, the current study evaluated these modalities primarily in isolation during inference. In real-world decision support scenarios, operators often interact with systems using simultaneous multimodal inputs, such as speaking a command while selecting text on a screen. Future work will investigate the dynamic routing of these simultaneous inputs through distinct, yet interacting, parameter-efficient modules to achieve true multimodal synergy. Finally, the exploration of ultra-low precision quantization in conjunction with parameter-efficient fine-tuning represents a highly promising frontier. By combining these two orthogonal optimization strategies, researchers can potentially reduce the memory requirements of foundation models by an additional order of magnitude, further democratizing the deployment of advanced decision support systems across the most constrained operational environments.

References

Yu, H., Zhang, J., Chen, C., Xiang, T., Fang, Y., Niebles, J. C., & Adeli, E. (2026, March). Socialgen: Modeling multi-human social interaction with language models. In 2026 International Conference on 3D Vision (3DV) (pp. 1-17). IEEE.

Yang Y, Yang X, Jiang Y, et al. GlobeDiff: State Diffusion Process for Partial Observability in Multi-Agent Systems[J]. arXiv preprint arXiv:2602.15776, 2026.

Han, C., Zheng, W., & Tang, X. (2025, November). Debate-to-detect: Reformulating misinformation detection as a real-world debate with large language models. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 15125-15140).

Yao, Y., Xiao, W., Miao, P., Chen, G., Yang, H., Chae, C. B., & Wong, K. K. (2025). UAV-relay-aided secure maritime networks coexisting with satellite networks: robust beamforming and trajectory optimization. IEEE Transactions on Wireless Communications.

Xie, Z., Liao, R., Wang, Z., Chen, C., Hua, X. S., & Luo, X. (2026, July). GALA: Geometric Data Selection with Strategic Prospecting for Large Language Model Self-training. In Findings of the Association for Computational Linguistics: ACL 2026(pp. 10281-10293).

Pei, Z., Huang, Q., & Wang, S. (2026). When LLMs develop languages: Symbolic communication for efficient multi-agent reasoning.arXiv preprint arXiv:2606.29354.

Xu, X., Tu, W., Yang, Y., Li, J., Zhang, Y., & Chen, H. (2026). Contribution-aware dynamic multi-modal balance for audio-visual speech separation.IEEE Transactions on Multimedia.

Tang, J., Yuan, X., Liu, J., Yu, J., Dong, X., Chen, L., ... & Yin, J. (2026). SAKE: Self-aware Knowledge Exploitation-Exploration for Grounded Multimodal Named Entity Recognition. arXiv preprint arXiv:2604.20146.

Sang, Y. (2025, October). AutoCrit: A Meta-Reasoning Framework for Self-Critique and Iterative Error Correction in LLM Chains-of-Thought. In 2025 6th International Conference on Machine Learning and Computer Application (ICMLCA) (pp. 1177-1180). IEEE.

Wang, X., Suzuki, Y., Li, C. M., & Tanaka, S. (2025, June). Micropatterned PDMS damping layer integration to enhance the fidelity of piezoelectric MEMS speakers without SPL penalty. In 2025 23rd International Conference on Solid-State Sensors, Actuators and Microsystems (Transducers) (pp. 455-458). IEEE.

Zhang, W., & Pei, S. (2026). Predictive prefetching for retrieval-augmented generation. Proceedings of Machine Learning Research. The Forty-third International Conference on Machine Learning, ICML 2026. https://par.nsf.gov/biblio/10686293

Guo, J., Chen, X., Xia, Q., Wang, Z., Ou, J., Qin, L., ... & Tian, W. (2025, July). HASH-RAG: bridging deep hashing with retriever for efficient, fine retrieval and augmented generation. In Findings of the Association for Computational Linguistics: ACL 2025 (pp. 26847-26858).

Tang, J., Wang, Z., Gong, Z., Yu, J., Zhu, X., & Yin, J. (2025, April). Multi-grained query-guided set prediction network for grounded multimodal named entity recognition. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 24, pp. 25246-25254).

Yao, Y., Zhang, J., Miao, P., Zhang, L., Chen, G., Shu, F., & Wong, K. K. (2025). Hybrid RIS-enhanced ISAC secure systems: Joint optimization in the presence of an extended target. IEEE Transactions on Communications.

Li, Y., Yang, Z., Huang, Y., Liu, X., & Qi, G. (2026, March). C³TG: Conflict-aware, Composite, and Collaborative Controlled Text Generation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 38, pp. 31834-31842).

Sang, Y. (2025, July). Towards explainable rag: Interpreting the influence of retrieved passages on generation. In 2025 4th International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC) (pp. 397-400). IEEE.

Zhu, R., Ma, Z., Wu, J., Gao, J., Wang, J., Lin, D., & He, C. (2025, April). Utilize the flow

before stepping into the same river twice: Certainty represented knowledge flow for refusal-aware instruction tuning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 24, pp. 26157-26165).

Qian, B., Guo, J., Wang, Q., & Shuang, K. (2025). From local verification to global reasoning: Exploiting slot-accompanying update for improved slot selection. Knowledge-Based Systems, 319, 113521.

Xu, X., Han, C., Zhang, Y., Tu, W., & Yang, Y. (2024, April). Curricular contrastive regularization for speech enhancement with self-supervised representations. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 10486-10490). IEEE.

Zhu, R., Jiang, X., Wu, J., Ma, Z., Song, J., Bai, F., ... & He, C. (2025, April). GRAIT: gradient-driven refusal-aware instruction tuning for effective hallucination mitigation. In Findings of the Association for Computational Linguistics: NAACL 2025 (pp. 4006-4021).

Vinyals, O., Fortunato, M., & Jaitly, N. (2015). Pointer networks. In Advances in Neural Information Processing Systems.

Lv, Z., Wang, Q., Lyu, T., & Dong, J. (2026). A Deep Learning-Based Model for User Emotion Recognition and Behavior Guidance Design: EMoGuideNet-An Emotion-Aware Multimodal Framework for Enhancing Intelligent User Interaction. Journal of Organizational and End User Computing (JOEUC), 38(1), 1-30.

Li, X., Wang, P., Li, G., Ni, L., & Zhang, Y. (2023). Design of interface circuits and lightweight PUF for TMR sensors. IEEE Sensors Journal, 23(11), 11754-11761.

Wang, H., Xu, Q., Liu, C., Wu, J., Lin, F., & Chen, W. (2026, April). Emergent hierarchical reasoning in llms through reinforcement learning. In International Conference on Learning Representations (Vol. 2026, pp. 74519-74543).

Wei, W. D., Fang, H. B., Liu, Y. D., Shi, J. X., Kwok, J., & Li, Y. F. (2026). Activation Compression in LLMs: Theoretical Analysis and Efficient Algorithm. arXiv preprint arXiv:2605.01255.

Xu, X., Tu, W., Yang, Y., Li, J., & Zhang, Y. (2025). Interactive target positive and negative features modeling for monaural speech enhancement.IEEE Transactions on Audio, Speech and Language Processing,33, 4856-4869.

Liu, H., Ding, L., & Henao, R. (2026, July). Learning to Control Summaries with Score Ranking. In Findings of the Association for Computational Linguistics: ACL 2026(pp. 35336-35359).

Sang, Y. (2025, August). Robustness of fine-tuned llms under noisy retrieval inputs. In 2025 6th International Conference on Artificial Intelligence and Electromechanical Automation (AIEA) (pp. 417-420). IEEE.

Xiao, Q., & Zhou, K. (2026). MACD: Model-Aware Contrastive Decoding via Counterfactual Data. arXiv preprint arXiv:2602.01740.

Xu, X., Wang, Y., Xu, D., Peng, Y., Zhang, C., Jia, J., & Chen, B. (2022, May). Vsegan: Visual speech enhancement generative adversarial network. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7308-7311). IEEE.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of ACL.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In Proceedings of EMNLP.

Guo, J., Shuang, K., Zhang, K., Liu, Y., Li, J., & Wang, Z. (2023, June). Learning to imagine: distillation-based interactive context exploitation for dialogue state tracking. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 11, pp. 12845-12853).

Tang, J., Yang, Y., Yu, J., Wang, Z. X., Liang, H., Yao, L., & Yin, J. (2025, November). Unco: Uncertainty-driven collaborative framework of large and small models for grounded multimodal ner. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 7644-7662).

Zhang, Y., Zhang, Y., Zhao, M., Feng, S., & Cheung, Y. M. (2026). How to Ask the AI: A User Perspective Survey for Large Language Model Prompting.IEEE Transactions on Artificial Intelligence.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems.