Ingénierie des Systèmes d'Information
Vol. 31, No. 8, August, 2026, pp. 2737-2751
Journal homepage: http://iieta.org/journals/isi

Parameter-Efficient Adaptation of Vision-Language Models for Wearable Artificial Intelligence: A Systematic Review of Resource-Constrained Deployment

Folasade Y. Ayankoya[1], Afolashade O. Kuyoro[1], Ogunlere Samson[2], Oluwabukola F. Ajayi[1*], Amanze Ruth[1], Anyaehie Amarachi[3], Fatade Oluwayemisi Boye[1], Oluwole Solanke[1], Oladipo Sunday[1], Akinwunmi Damilare[1], Eweoya Ibukun[4], Adelowo Joshua[5]

[1] Department of Computer Science, Babcock University, Ilishan-Remo 121103, Nigeria
[2] Department of Computer Engineering, Babcock University, Ilishan-Remo 121103, Nigeria
[3] Department of Computer Science, Nile University, Abuja FCT 900001, Nigeria
[4] Department of Software Engineering, Babcock University, Ilishan-Remo 121103, Nigeria
[5] Department of Information Technology, Babcock University, Ilishan-Remo 121103, Nigeria

Corresponding Author Email: ajayioluwa@babcock.edu.ng

https://doi.org/10.18280/isi.310819

Received: 7 June 2026
Revised: 12 August 2026
Accepted: 20 August 2026
Available online: 31 August 2026

Keywords:
vision-language models, parameter-efficient fine-tuning, wearable AI, multimodal learning, edge intelligence, resource-constrained systems

ABSTRACT

Vision-Language Models (VLMs) have advanced multimodal artificial intelligence (AI) by enabling joint understanding and reasoning across visual and textual information. However, the increasing scale of VLMs introduces substantial challenges for deployment in wearable information systems, where computational capacity, memory availability, energy budget, and communication resources are highly constrained. Parameter-Efficient Fine-Tuning (PEFT) has emerged as a a strategy for adapting large pretrained models by updating only a small subset of parameters while preserving most pretrained weights; however, parameter reduction does not by itself imply proportional reductions in inference latency or energy consumption. This review presents a structured analysis of PEFT approaches for large VLMs in resource-constrained wearable AI environments. The study examines major PEFT approaches, including adapter-based tuning, prompt and prefix tuning, Low-Rank Adaptation (LoRA), quantization-aware methods, and hybrid adaptation strategies. Furthermore, on-device, edge-assisted, cloud-based, and federated deployment architectures are examined in terms of computational requirements, privacy, scalability, and deployment feasibility. The review also summarizes current benchmarking practices and identifies limitations related to standardized evaluation, hardware-aware validation, energy measurement, and real-world deployment. Based on the synthesized evidence, future research directions are discussed, including federated multimodal learning, continual adaptation, privacy-preserving optimization, energy-aware computing, and edge-native multimodal intelligence. Overall, the review identifies the conditions under which PEFT may support the adaptation of large VLMs for resource-constrained wearable information systems and highlights the evidence gaps that must be addressed before broader deployment.

1. INTRODUCTION

Recent advances in artificial intelligence (AI) have led to foundation models that learn generalized representations from large-scale datasets. Vision-Language Models (VLMs) represent an important class of multimodal foundation models that integrate visual perception and natural language understanding within unified architectures [1, 2]. Unlike conventional machine learning systems that process visual and textual information independently, VLMs learn shared multimodal representations that facilitate joint reasoning across heterogeneous data modalities [3]. Representative models, including Contrastive Language–Image Pre-training (CLIP), Flamingo, Bootstrapping Language-Image Pre-training 2 (BLIP-2), Gemini, GPT-4V, and Qwen-VL, have demonstrated capabilities across diverse multimodal tasks [4-7]. These tasks include image captioning, visual question answering, image-text retrieval, visual grounding, multimodal dialogue generation, and embodied reasoning. These models leverage Transformer architectures and large-scale image–text pretraining to learn transferable multimodal representations that support a broad range of downstream tasks and application domains [2-8].

In parallel, wearable AI and wearable information systems have attracted increasing research attention as advances in miniaturized sensors, edge computing, wireless communication, and low-power embedded systems have expanded the capabilities of wearable platforms [9]. Modern wearable devices, including smartwatches, smart glasses, augmented reality headsets, healthcare monitoring systems, and body-worn sensors, continuously generate multimodal data that require real-time processing, contextual

interpretation, and intelligent decision support [10]. The integration of VLMs into wearable platforms has been investigated as a potential approach for healthcare monitoring, assistive technologies, industrial safety, contextual navigation, and human–computer interaction [11]. However, the extent to which these capabilities translate into reliable real-world wearable deployment remains dependent on computational resources, energy constraints, latency requirements, privacy considerations, and application-specific validation.

Despite these advances, deploying large VLMs within wearable environments remains challenging because wearable devices operate under stringent constraints in computational capacity, memory availability, battery energy, communication bandwidth, and thermal dissipation [12, 13]. These limitations restrict the practical deployment of fully fine-tuned multimodal foundation models and motivate the need for resource-aware adaptation strategies capable of maintaining high predictive performance while minimizing computational overhead [14]. To address these challenges, Parameter-Efficient Fine-Tuning (PEFT) has emerged as a potentially effective strategy for adapting large pretrained models to wearable multimodal information systems while updating only a small subset of trainable parameters [15]. Rather than modifying the entire parameter space, PEFT techniques preserve most pretrained weights and introduce lightweight adaptation mechanisms such as adapters, prompts, prefixes, or low-rank parameter updates [16]. Studies have shown that PEFT methods can achieve performance comparable to full fine-tuning while reducing trainable parameters by more than 95% and substantially lowering computational requirements [15-17].

Low-Rank Adaptation (LoRA) has become one of the most widely adopted PEFT techniques because of its ability to inject trainable low-rank matrices into transformer layers while maintaining frozen pretrained weights [18]. Subsequent developments such as QLoRA, AdaLoRA, DoRA, and VeRA have further improved memory efficiency and adaptation effectiveness, making them increasingly attractive for wearable and edge AI applications [19-21]. These advances suggest that PEFT may provide an effective adaptation strategy for bringing multimodal foundation models to resource-constrained wearable environments.

Although research in PEFT and multimodal foundation models has expanded rapidly, existing studies remain fragmented across different application domains, architectures, and evaluation frameworks. Furthermore, the unique requirements of wearable AI, including low latency, energy-related implications, continual adaptation, privacy preservation, and real-time multimodal reasoning, have received limited attention within existing surveys [22]. Consequently, a comprehensive review focusing specifically on parameter-efficient adaptation of large VLMs for wearable AI is both timely and necessary.

Recent advances in wearable AI have increasingly positioned PEFT-enabled VLMs within the broader domain of wearable information systems, where intelligent data acquisition, multimodal information processing, and real-time decision support are closely integrated. The overall relationship between pretrained VLMs, PEFT modules, and wearable deployment environments is illustrated in Figure 1. These systems can support distributed processing across wearable devices, edge infrastructure, and cloud platforms, enabling multimodal data to be processed at different computational layers according to application requirements. In this context, PEFT may reduce the computational and memory requirements associated with adapting VLMs; however, its implications for overall system resource consumption depend on factors such as model architecture, hardware configuration, inference workload, and the distribution of processing across device, edge, and cloud environments.

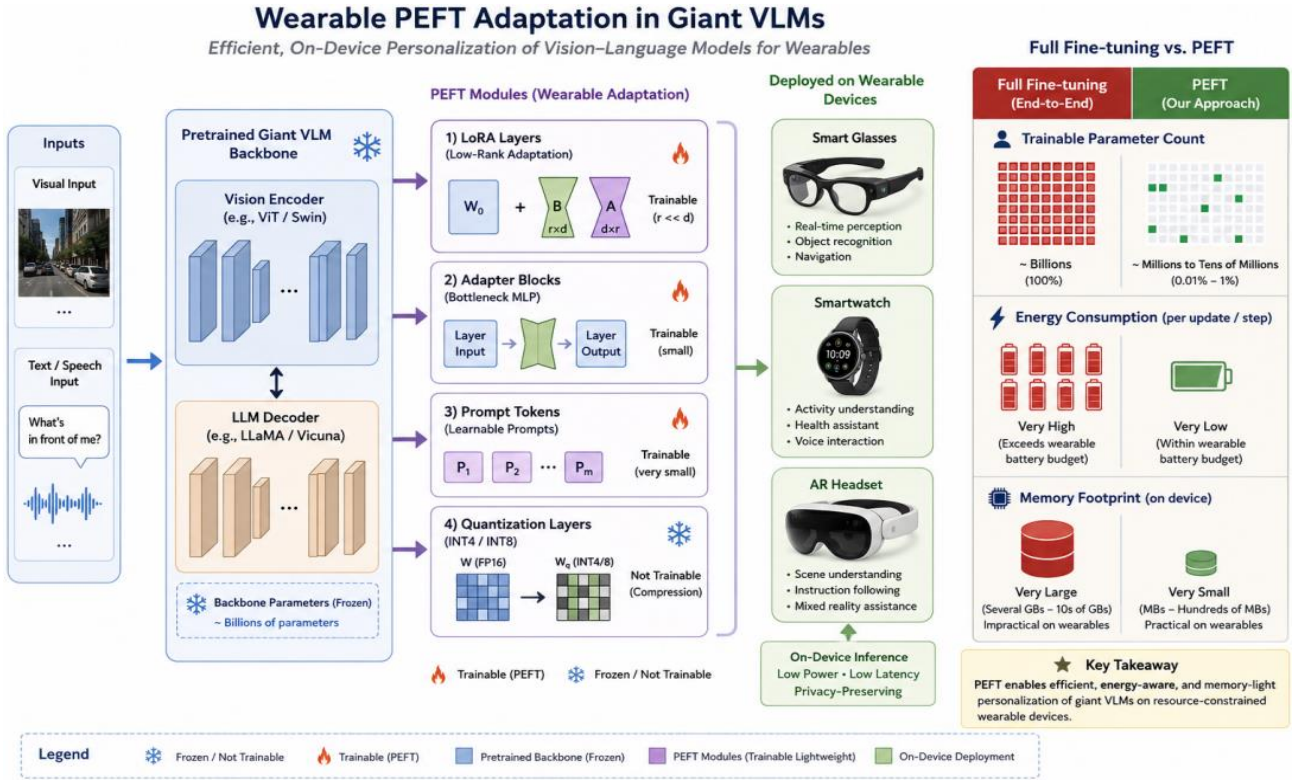

Figure 1. Conceptual architecture of Parameter-Efficient Fine-Tuning (PEFT) modules integrated with a pretrained Vision-Language Model (VLM) for wearable deployment

In these environments, wearable devices continuously acquire heterogeneous sensor data, including images, video streams, physiological measurements, audio signals, and contextual information, which are processed collaboratively across distributed computing infrastructures [14]. Rather than relying solely on centralized cloud intelligence, modern wearable information systems adopt resource-aware processing strategies in which computational tasks are dynamically allocated among on-device processors, edge servers, and cloud platforms according to latency requirements, energy availability, network conditions, and privacy constraints. Within this distributed architecture, PEFT enables large VLMs to be adapted for domain-specific tasks while reducing the number of trainable parameters and associated adaptation costs; the resulting memory and energy benefits depend on the model architecture, hardware platform, and adaptation configuration [16].

From an information systems perspective, PEFT-enabled VLMs support the development of intelligent information services by transforming heterogeneous multimodal data into context-aware knowledge that assists users in real-time decision-making. The resulting information flow extends from wearable sensors through edge computing nodes and cloud-based foundation models to end users, creating an adaptive ecosystem capable of continuous perception, reasoning, and personalized service delivery. Such architectures have important implications for healthcare monitoring, industrial safety, assistive technologies, and augmented reality, where timely interpretation of multimodal information is essential for operational effectiveness [13]. Consequently, the integration of PEFT with distributed wearable information systems represents not only an advancement in efficient model adaptation but also a significant step toward scalable, resource-aware, and intelligent information processing frameworks capable of supporting next-generation wearable AI applications.

Building upon these developments, this paper presents a comprehensive review of PEFT-enabled multimodal information systems for wearable AI environments. Rather than focusing solely on individual PEFT techniques, the review examines how parameter-efficient adaptation supports the design, deployment, and operation of intelligent wearable information systems that integrate multimodal data acquisition, distributed computing, edge intelligence, and cloud-based foundation models. The review synthesizes recent advances in PEFT methodologies, wearable deployment architectures, benchmarking frameworks, and application domains, while critically evaluating their suitability for resource-constrained wearable platforms.

1.1 Terminology and conceptual scope

To ensure terminological consistency, this review distinguishes between related but conceptually different terms used to describe AI-enabled wearable technologies. Wearable AI is used as the broad application domain encompassing the use of artificial intelligence in wearable technologies and applications. Wearable information systems refers specifically to the information-systems perspective, including the acquisition, processing, integration, and delivery of information through wearable technologies. Wearable platform or wearable device refers to the physical hardware through which these capabilities are implemented, whereas a wearable multimodal system denotes a system capable of processing and integrating multiple data modalities, such as vision, audio, language, and sensor data. Accordingly, these terms are used deliberately throughout the manuscript and are not treated as interchangeable.

2. METHODOLOGY

This review adopted a structured narrative review methodology to identify, evaluate, and synthesize research on PEFT for large and giant VLMs and their applications in wearable AI. The methodology focused on PEFT techniques, multimodal foundation models, computational and memory efficiency, edge/on-device deployment, and resource-constrained wearable environments. The review covered studies published between 1 January 2020 and 31 December 2025. Initial database searches were conducted between 6 and 10 January 2026, citation tracking between 12 and 16 January 2026, and an updated database search between 1 and 5 February 2026. The review process comprised literature identification, duplicate removal, screening, eligibility assessment, quality evaluation, data extraction, and thematic synthesis.

2.1 Literature search strategy

The literature search was conducted in IEEE Xplore, Scopus, Web of Science Core Collection, ACM Digital Library, and ScienceDirect. The search strategy combined three conceptual groups: PEFT/adaptation techniques, VLMs/multimodal foundation models, and wearable/edge/resource-constrained AI. Search terms included *"parameter-efficient fine-tuning", PEFT, LoRA, QLoRA, "low-rank adaptation", "adapter tuning", "prompt tuning", "prefix tuning", "vision-language model", VLM, "large vision-language model", "large multimodal model", "multimodal foundation model", "wearable AI", "wearable computing", "edge AI", "edge intelligence", "on-device AI",* and *"resource-constrained AI"*.

Database-specific Boolean search strings were developed to accommodate differences in database search syntax. For example, the Scopus search employed TITLE-ABS-KEY fields, Web of Science used TS, while IEEE Xplore, ACM Digital Library, and ScienceDirect used their respective advanced-search fields. Searches were limited to publications from 2020–2025 and, where supported, English-language publications.

Table 1 shows that the initial database searches retrieved 1,306 records: 214 from IEEE Xplore, 386 from Scopus, 297 from Web of Science, 241 from ACM Digital Library, and 168 from ScienceDirect. Backward and forward citation searching conducted from 12 to 16 January 2026 identified 74 additional records, resulting in 1,380 initially identified records. A database update conducted from 1 to 5 February 2026 retrieved 93 records, of which 41 were new after comparison with the existing review database.

2.2 Eligibility, screening, and study selection

Studies were included when they were peer-reviewed journal articles or recognized international conference papers published between 2020 and 2025; investigated PEFT techniques such as LoRA, QLoRA, adapter tuning, prompt tuning, or related approaches; addressed VLMs, multimodal

foundation models, or related architectures; and provided findings relevant to wearable AI, edge AI, on-device intelligence, resource-constrained computing, or efficient multimodal deployment. Studies focusing primarily on PEFT or VLM methodology without a direct wearable application were also retained when their findings had clear implications for resource-constrained or wearable deployment.

Studies were excluded when they focused exclusively on conventional full-model fine-tuning, lacked a relevant PEFT/VLM or deployment component, were editorials, tutorials, opinion articles, news or magazine publications, lacked sufficient technical information, were duplicates, fell outside the 2020–2025 period, or did not meet the predefined publication and language requirements.

Following consolidation of the retrieved records, 318 duplicates were removed, leaving 1,062 records for title and abstract screening. Of these, 781 were excluded, leaving 281 publications for full-text assessment. A total of 196 full-text publications were excluded because they were outside the review scope (72), lacked a PEFT component (43), lacked a relevant VLM/multimodal component (29), were not relevant to wearable/edge/resource-constrained AI (24), provided insufficient technical or methodological information (13), were editorial/tutorial/non-research publications (8), or represented duplicate/overlapping studies (7). Consequently, 85 studies were included in the initial qualitative synthesis.

The February 2026 update identified 41 new records. Of these, 18 proceeded to full-text assessment, and 7 additional studies met the eligibility criteria. Thus, the final qualitative synthesis comprised 92 studies.

Table 1. Database-specific search strategy and retrieval results

Database	Search Date	Search Field	Exact Search String	Records Retrieved
IEEE Xplore	6–10 Jan. 2026	Metadata/Abstract	("Parameter-Efficient Fine-Tuning" OR PEFT OR LoRA OR QLoRA OR "Low-Rank Adaptation" OR "Adapter Tuning" OR "Prompt Tuning" OR "Prefix Tuning") AND ("Vision-Language Model" OR VLM OR "Large Vision-Language Model" OR "Multimodal Foundation Model" OR "Large Multimodal Model") AND ("Wearable Artificial Intelligence" OR "Wearable AI" OR "Wearable Computing" OR "Edge AI" OR "Edge Intelligence" OR "On-Device AI" OR "Resource-Constrained AI")	214
Scopus	6–10 Jan. 2026	TITLE-ABS-KEY	TITLE-ABS-KEY(("parameter-efficient fine-tuning" OR PEFT OR LoRA OR QLoRA OR "low-rank adaptation" OR "adapter tuning" OR "prompt tuning" OR "prefix tuning") AND ("vision-language model*" OR VLM OR "large vision-language model*" OR "multimodal foundation model*" OR "large multimodal model*") AND ("wearable artificial intelligence" OR "wearable AI" OR "wearable computing" OR "edge AI" OR "edge intelligence" OR "on-device AI" OR "resource-constrained AI")) AND PUBYEAR > 2019 AND PUBYEAR < 2026	386
Web of Science Core Collection	6–10 Jan. 2026	Topic (TS)	TS=(("parameter-efficient fine-tuning" OR PEFT OR LoRA OR QLoRA OR "low-rank adaptation" OR "adapter tuning" OR "prompt tuning" OR "prefix tuning") AND ("vision-language model*" OR VLM OR "large vision-language model*" OR "multimodal foundation model*" OR "large multimodal model*") AND ("wearable artificial intelligence" OR "wearable AI" OR "wearable computing" OR "edge AI" OR "edge intelligence" OR "on-device AI" OR "resource-constrained AI")) AND PY=(2020-2025)	297
ACM Digital Library	6–10 Jan. 2026	Title/Abstract	(("parameter-efficient fine-tuning" OR PEFT OR LoRA OR QLoRA OR "low-rank adaptation" OR "adapter tuning" OR "prompt tuning" OR "prefix tuning") AND ("vision-language model" OR VLM OR "large vision-language model" OR "multimodal foundation model" OR "large multimodal model") AND ("wearable AI" OR "wearable artificial intelligence" OR "wearable computing" OR "edge AI" OR "edge intelligence" OR "on-device AI" OR "resource-constrained AI"))	241
ScienceDirect	6–10 Jan. 2026	Title/Abstract/ Keywords	("parameter-efficient fine-tuning" OR PEFT OR LoRA OR QLoRA OR "low-rank adaptation" OR "adapter tuning" OR "prompt tuning" OR "prefix tuning") AND ("vision-language model" OR VLM OR "large vision-language model" OR "multimodal foundation model" OR "large multimodal model") AND ("wearable AI" OR "wearable artificial intelligence" OR "wearable computing" OR "edge AI" OR "edge intelligence" OR "on-device AI" OR "resource-constrained AI") AND pubyear:2020-2025	168
Total	—	—	—	1,306

Source: Authors' database search records, January–February 2026.

2.3 Quality assessment and data extraction

The methodological quality of the included studies was assessed qualitatively using predefined criteria covering methodological rigor, technical reporting completeness, experimental validation, technical contribution, and relevance to wearable or resource-constrained deployment. Particular attention was given to the clarity of dataset descriptions, model architectures, PEFT configurations, training procedures, computational requirements, memory consumption, evaluation metrics, baseline comparisons, and reported limitations.

A structured data-extraction framework was applied consistently across the selected studies. Extracted information included publication characteristics, VLM architecture, PEFT technique, trainable-parameter proportion, datasets, application domain, adaptation strategy, computational and memory requirements, hardware platform, deployment architecture, edge/cloud/on-device strategy, inference characteristics, evaluation metrics, benchmark results, advantages, limitations, scalability considerations, and future research directions.

2.4 Citation tracking, search update, and reproducibility

Backward and forward citation tracking was conducted between 12 and 16 January 2026 to identify relevant studies not captured through the electronic database searches. References of eligible studies and relevant review papers were examined, while forward citation tracking was used to identify subsequent research citing influential publications. The 74 citation-derived records were subjected to the same screening criteria as database-retrieved records.

Because PEFT and VLM research is evolving rapidly, a supplementary database search was conducted between 1 and 5 February 2026 using the same search strategy and eligibility criteria. The update retrieved 93 records, of which 41 were new after comparison with the existing review database. This update resulted in seven additional eligible studies.

To enhance reproducibility, an audit trail was maintained containing the database name, search dates, update dates, exact database-specific search strings, search fields, publication-period restrictions, language restrictions, retrieved record numbers, duplicate-removal counts, screening decisions, full-text exclusion reasons, citation-search results, and final inclusion numbers. The exact search strings and search log should be provided as supplementary material where permitted by the target journal.

2.5 Data synthesis and analytical framework

The evidence was synthesized qualitatively because of substantial heterogeneity among the included studies in terms of VLM architectures, PEFT techniques, datasets, application domains, hardware platforms, experimental protocols, and evaluation metrics. Consequently, statistical meta-analysis was not considered appropriate.

The selected literature was organized into five thematic areas: (1) PEFT methodologies for VLM adaptation, (2) computational, parameter, and memory efficiency, (3) wearable, edge, and on-device deployment architectures, (4) multimodal wearable AI applications, and (5) open challenges and future research directions. Comparative analysis focused on parameter reduction, computational cost, memory requirements, adaptation performance, inference latency, deployment feasibility, scalability, and robustness.

Particular consideration was given to the practical constraints of wearable AI, including limited processing resources, memory capacity, energy consumption, inference latency, connectivity requirements, privacy, and the feasibility of local or edge-based model adaptation. This thematic synthesis enabled the review to move beyond a descriptive classification of PEFT methods by identifying how parameter-efficient adaptation can support the practical deployment of increasingly large multimodal models in wearable and resource-constrained intelligent systems.

2.6 Foundations of Parameter-Efficient Fine-Tuning

PEFT methods are designed to minimize the number of trainable parameters required during adaptation while preserving the representational capabilities of large pretrained models. Instead of modifying the entire parameter space, PEFT selectively updates compact parameter subsets or inserts lightweight trainable structures into existing architectures.

The theoretical foundation of PEFT stems from the observation that downstream adaptation often occupies a low-dimensional subspace within the original parameter manifold [15]. Early parameter-efficient adaptation strategies include adapter modules [23], prefix tuning [16], and prompt tuning [17]. These approaches demonstrated that competitive downstream performance can often be achieved by updating only a small fraction of model parameters, thereby motivating subsequent developments such as LoRA and QLoRA. Recent empirical evidence indicates that PEFT behavior in multimodal large language models is method- and architecture-dependent. Zhou et al. [24] evaluated multiple PEFT strategies across seven multimodal datasets and showed that the relative effectiveness of PEFT methods varies according to model architecture, adaptation-module location, training-data scale, and evaluation task. Their findings further indicate that adapter-based approaches can provide strong performance across several evaluated settings, highlighting the importance of empirical benchmarking rather than assuming that a single PEFT strategy is universally optimal.

2.6.1 Low-Rank Adaptation

PEFT has emerged as an important approach for adapting large VLMs by updating only a small subset of model parameters while preserving most pretrained weights. Among the proposed PEFT strategies are adapter-based methods, prompt tuning, prefix tuning, and LoRA techniques, which collectively aim to reduce the number of trainable parameters and the computational and memory requirements associated with model adaptation. This review examines these approaches from the perspective of wearable AI, where resource efficiency is a primary design consideration. LoRA has been widely adopted for adapting large language models and VLMs to downstream tasks, including multimodal question answering, industrial inspection, embodied AI, and edge intelligence. Within the reviewed literature, LoRA presents characteristics that may be relevant to smart glasses, augmented-reality headsets, and portable healthcare monitoring scenarios; however, suitability for specific wearable platforms requires validation on the target hardware.

The mathematical formulation of LoRA can be expressed as:

$$W = W_0 + BA \tag{1}$$

where, W_0 denotes the original pretrained weight matrix, while B and A are trainable low-rank matrices. The output can therefore be represented as:

$$h = (W_0 + BA)x \tag{2}$$

where, x denotes the input vector, and h represents the transformed output.

Recent studies demonstrate that LoRA-based adaptation achieves highly competitive performance in multimodal tasks, including image-text retrieval and visual reasoning [25], while

typically reducing the number of trainable parameters by approximately 95% to over 99.9%, depending on the target architecture and selected rank configuration [15-18]. Variants such as QLoRA, AdaLoRA, and MoRA further optimize parameter allocation and quantization efficiency.

2.6.2 Adapter-based fine-tuning

Adapter methods insert lightweight neural modules between transformer layers. During training, only these adapter modules are updated while the base model remains frozen. Adapters offer modularity and facilitate rapid switching between downstream tasks.

Adapter-based tuning has potential relevance to personalized healthcare monitoring scenarios because task-specific modules can be trained while the backbone remains frozen; however, direct validation on wearable healthcare hardware remains limited. Adapters may reduce communication overhead in distributed settings because only compact modules require transmission between devices and servers.

Recent VLM-specific research also indicates that parameter-efficient adaptation need not always rely on additional trainable modules. Li et al. [26] demonstrated that selectively updating bias and normalization parameters can improve CLIP adaptation without introducing additional trainable modules. This finding broadens the PEFT design space and suggests that adaptation efficiency can also be achieved through selective modification of existing model parameters.

2.6.3 Prompt and prefix tuning

Prompt tuning adapts VLM behavior through learnable soft prompt embeddings prepended to input sequences. Prefix tuning extends this idea by optimizing trainable vectors inserted into transformer attention mechanisms.

These methods may be relevant to resource-constrained wearable systems. Multimodal prompt-tuning approaches extend conventional prompt-based adaptation by incorporating modality-specific information during parameter-efficient adaptation. For example, M²PT introduces multimodal prompt tuning for zero-shot instruction learning, addressing the limitation of approaches that treat modalities independently [27]. Such methods may be relevant to wearable multimodal systems because they provide a mechanism for adapting multimodal representations without updating the full model parameter space.

Multimodal PEFT has also been explored through query-based adaptation. Liang et al. [28] proposed Querying as Prompt, in which modality-specific queries are introduced as soft prompts to a frozen language model, providing an alternative mechanism for integrating multimodal information while limiting parameter updates.

2.6.4 Quantization-aware Parameter-Efficient Fine-Tuning

Quantization-aware PEFT combines parameter-efficient tuning with low-bit precision representations. QLoRA introduced a memory-efficient fine-tuning framework based on 4-bit quantized weights and reported competitive task performance [18].

For wearable systems with constrained memory and computational resources, quantized PEFT may improve adaptation and deployment feasibility by reducing model memory requirements; however, latency, thermal behavior, and energy consumption depend on the underlying hardware and quantization. Figure 2 compares representative PEFT methods in terms of trainable-parameter proportion, memory requirements, computational characteristics, and reported relevance to wearable deployment.

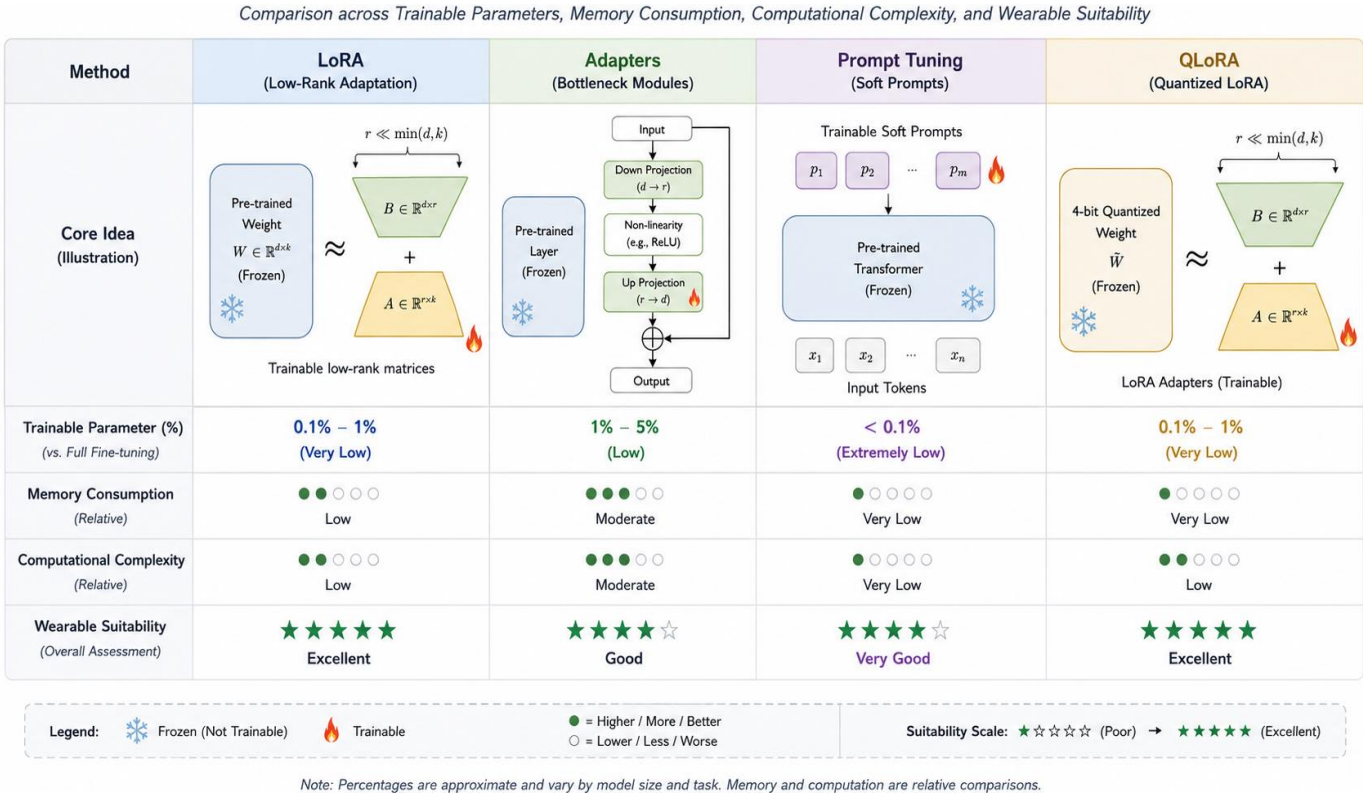

Figure 2. Comparison of Low-Rank Adaptation (LoRA), adapter tuning, prompt tuning, and QLoRA based on trainable parameters, memory requirements, computational complexity, and wearable deployment suitability

Table 2. Comparative characteristics of major Parameter-Efficient Fine-Tuning (PEFT) methods

PEFT Method	Trainable Parameter Ratio	Parameter Reduction	Memory Evidence	Latency/Computation Evidence	Energy Evidence	Wearable Relevance
Full Fine-Tuning	100%	Baseline	Highest adaptation-state requirement	Highest training/update burden	Direct energy depends on hardware	Generally unsuitable for highly constrained devices
Adapter Tuning	1–5% [23]	~95–99% [23]	Reduced trainable state; adapter overhead remains	Adapter computation is architecture-dependent	Direct wearable measurements limited	Potentially relevant where modular task adaptation is required
Prompt Tuning	<0.1% [17]	>99.9% [17]	Very small trainable state	Prompt-length dependent	Direct wearable measurements limited	Potentially relevant to lightweight personalization
Prefix Tuning	<0.1% [16]	>99.9% [16]	Very small trainable state	Prefix-length dependent attention overhead	Direct wearable measurements limited	Potentially relevant to contextual adaptation
LoRA	<1% [15]	~95%–>99.9% [15]	Reduced adaptation-state requirement	Rank- and architecture-dependent	No standardized wearable ranking	Potentially relevant to memory-constrained edge/wearable scenarios
QLoRA	<1% [18]	~95%–>99.9% [18]	4-bit quantization reduces frozen-backbone memory	Hardware-dependent	Direct wearable measurements limited	Potentially relevant where low-bit hardware is supported

Note: N/B Reported ranges are representative values derived from cited studies and are configuration- and architecture-dependent; they should not be interpreted as universal performance guarantees.

Table 3. The principal technical terms

Term	Definition
Adaptation Efficiency	The capability of a PEFT method to adapt a pretrained foundation model to a downstream task while updating only a small proportion of model parameters, thereby minimizing computational cost, memory consumption, and training time without substantially compromising predictive performance.
Deployment Efficiency	The effectiveness with which a trained AI model can be deployed and executed within operational environments by satisfying constraints related to latency, memory usage, computational resources, communication bandwidth, scalability, and energy consumption.
Resource Optimization	The process of maximizing model performance while minimizing the utilization of computational resources, including processor usage, memory footprint, storage requirements, communication overhead, and battery energy, particularly in resource-constrained wearable devices.
Wearable Intelligence	The capability of wearable devices to continuously perceive, interpret, and respond to multimodal sensory information through embedded AI models that support context-aware reasoning and personalized decision support.
Edge Intelligence	The execution of AI models at or near the data source, such as wearable devices or edge servers, to reduce communication latency, preserve user privacy, and enable real-time inference without relying exclusively on centralized cloud infrastructure.
Parameter Efficiency	The degree to which a learning algorithm minimizes the number of trainable parameters required for downstream adaptation while maintaining competitive predictive performance relative to conventional full fine-tuning approaches.
Inference Efficiency	The ability of a deployed AI model to generate predictions within acceptable latency and resource constraints. Energy consumption is treated as a separate empirical deployment metric because it depends on hardware, workload, model architecture, and implementation.

As shown in Table 2, several low-rank and prompt-based PEFT methods use less than one percent of model parameters for adaptation, although the ratio varies substantially across adaptation strategies and model architectures. Such reductions in trainable parameters can be advantageous for memory-constrained adaptation; however, their implications for energy consumption depend on the hardware, workload, memory access patterns, and deployment configuration.

The quantitative comparison demonstrates that PEFT methods differ substantially in the degree and mechanism of resource reduction. Reported studies indicate that LoRA can reduce the number of trainable parameters by approximately 95% to more than 99.9%, depending on model architecture and rank configuration [15-18]. Prompt and prefix tuning can reduce the trainable parameter ratio to below 0.1%, making them attractive when adaptation-state size is the primary constraint. Adapter-based approaches generally require a larger trainable parameter fraction of approximately 1–5% but provide modularity that can be advantageous when multiple wearable tasks or personalized models must be supported.

QLoRA provides an additional efficiency mechanism by combining LoRA with 4-bit quantization. Its advantage therefore extends beyond the number of trainable parameters to the memory required for representing the frozen backbone [18]. However, parameter reduction should not be interpreted as a direct equivalent of inference-speed or energy reduction. Actual latency and energy consumption depend on hardware architecture, quantization support, memory bandwidth, model size, sequence length, and workload characteristics. Consequently, the reviewed evidence supports PEFT as a mechanism for reducing adaptation and memory costs, but it does not yet establish a universal ranking of PEFT methods in terms of wearable inference latency or battery consumption.
This distinction reveals an important limitation in the current literature. Existing studies frequently report parameter efficiency and model performance but provide fewer standardized measurements of battery consumption, thermal behavior, end-to-end latency, and sustained inference

efficiency on actual wearable hardware. The absence of standardized hardware-aware measurements limits direct comparison among PEFT approaches and reinforces the need for wearable-specific benchmarking protocols that evaluate both model-level and system-level performance.

2.7 Key terminology

To ensure consistency throughout this review, the principal technical terms are defined in Table 3.

The definitions presented in the Table 3 establish a consistent conceptual framework for the terminology adopted throughout this review. These terms are subsequently used to evaluate and compare PEFT techniques according to their computational characteristics, deployment evidence, and potential relevance to wearable multimodal information systems.

Throughout this review, 'technical feasibility' refers to evidence that a method can theoretically or computationally satisfy wearable resource constraints; 'experimental validation' refers to evaluation on simulated, benchmark, edge, or wearable hardware; and 'real-world deployment' refers to sustained operational use in an actual wearable application environment. These categories are not treated as equivalent.

3. GIANT VISION-LANGUAGE MODELS AND WEARABLE AI

VLMs represent a significant advancement in the evolution of multimodal AI. Their primary objective is to establish meaningful relationships between visual and textual representations, enabling machines to understand, generate, and reason across multiple modalities simultaneously. Unlike traditional computer vision systems that focus solely on image analysis or natural language models that process textual information independently, VLMs create shared representation spaces that facilitate cross-modal interaction and knowledge transfer [1, 2]. Through large-scale multimodal pretraining, these models learn generalized representations that can be transferred across diverse downstream tasks, including image captioning, visual question answering, image-text retrieval, visual grounding, and multimodal dialogue generation [3, 12].

The emergence of contrastive learning techniques marked a turning point in the development of modern VLMs. The introduction of CLIP demonstrated that large-scale image-text alignment could produce highly transferable visual and linguistic representations capable of supporting a wide range of downstream applications [1]. Subsequent architectures such as Flamingo, BLIP-2, Kosmos-2, LLaVA, Gemini, and GPT-4V introduced increasingly sophisticated mechanisms for multimodal fusion, cross-attention reasoning, instruction tuning, and conversational intelligence [4-7]. These developments culminated in the creation of giant multimodal foundation models that exhibit competitive generalization capabilities across previously unseen tasks and domains. Recent studies suggest that these models are approaching a level of multimodal reasoning that enables them to perform complex cognitive tasks involving perception, contextual understanding, and knowledge integration [14].

At the same time, wearable AI has evolved from simple activity-monitoring systems into sophisticated platforms capable of continuous sensing, contextual reasoning, and personalized decision support. Advances in miniaturized sensors, low-power embedded processors, wireless communication technologies, and edge computing have accelerated the development of intelligent wearable devices [10]. Modern wearable systems integrate cameras, microphones, inertial measurement units, physiological sensors, environmental monitors, and communication modules to continuously collect and process heterogeneous data streams [11]. These data streams are inherently multimodal, creating a natural synergy between wearable computing and vision-language intelligence.

The convergence of wearable computing and VLMs has created opportunities for multimodal information processing across several application domains. In healthcare, wearable multimodal systems may combine physiological measurements, visual information, and contextual data to support health monitoring and decision-support tasks. In augmented-reality environments, VLMs may assist with scene interpretation, object recognition, and natural-language interaction. Assistive technologies may use multimodal reasoning to provide context-aware information to users with different accessibility needs. Similarly, industrial wearable systems may support research into hazard recognition, equipment monitoring, and context-aware decision support. Nevertheless, these application scenarios should not be interpreted as evidence of established real-world deployment. The reviewed literature provides stronger evidence for technical feasibility and experimental evaluation than for long-term operation on physical wearable platforms. Validation under realistic workloads, hardware constraints, environmental conditions, privacy requirements, and user-specific requirements remains necessary [29]. These application scenarios illustrate the potential of VLMs to support multimodal interaction and context-aware decision-making in wearable information systems, including contextual scene understanding and natural language interaction in augmented reality environments [30], multimodal guidance for individuals with visual, auditory, or cognitive impairments [31], and real-time hazard detection and context-aware decision support in industrial wearable systems [13]. However, the extent of their practical deployment remains dependent on computational, architectural, and application-specific constraints.

Despite these opportunities, realizing the full potential of wearable multimodal intelligence requires overcoming significant computational barriers. Giant VLMs were originally designed for cloud-scale computing environments equipped with extensive hardware resources, including high-performance graphical processing units and large memory capacities [31]. Wearable devices, by contrast, operate under strict constraints involving energy consumption, latency, memory availability, storage capacity, thermal dissipation, and communication bandwidth [32]. These limitations can make conventional full fine-tuning computationally impractical on many resource-constrained wearable platforms. Furthermore, cloud-based processing introduces additional concerns related to privacy preservation, network reliability, communication overhead, and operational costs [15].

Bridging this gap requires adaptation strategies that preserve the powerful capabilities of large VLMs while reducing computational overhead. PEFT has therefore emerged as a critical enabling technology for wearable multimodal intelligence. By updating only a small subset of

model parameters or introducing lightweight trainable modules while keeping most pretrained weights frozen, PEFT can reduce the parameter and computational requirements of model adaptation. However, its effects on inference latency, memory use, and energy consumption depend on the model architecture, hardware platform, workload, and adaptation configuration [33]. Consequently, PEFT may provide a viable adaptation strategy for investigating the integration of advanced multimodal models within resource-constrained wearable AI systems.

4. PARAMETER-EFFICIENT FINE-TUNING ARCHITECTURES FOR VISION-LANGUAGE MODELS

The increasing scale of contemporary VLMs has created significant challenges for model adaptation and deployment. While full fine-tuning remains the conventional approach for transferring pretrained knowledge to downstream tasks, it becomes increasingly impractical as model sizes grow into the billions of parameters. Full fine-tuning generally imposes greater computational and memory requirements and may consequently increase energy demand, although the actual energy impact depends on the hardware and workload. These challenges are particularly problematic in wearable AI environments where computational resources are inherently limited. Consequently, PEFT has emerged as a promising paradigm for adapting large multimodal foundation models while reducing the number of trainable parameters and, in many settings, the associated adaptation memory and computational requirements.

Although recent studies have demonstrated parameter-efficient adaptation of multimodal models and increasingly efficient multimodal inference on mobile or edge hardware, direct evidence combining PEFT-enabled VLM adaptation with long-term evaluation on physical wearable platforms remains limited. In particular, standardized measurements of energy consumption, thermal behavior, latency, memory use, and adaptation performance across representative wearable hardware are not yet consistently reported. Consequently, the reviewed evidence supports the technical relevance of PEFT to resource-constrained multimodal systems but does not establish universal real-world deployment suitability.

Figure 3 presents a hierarchical taxonomy of PEFT techniques for VLMs, including adapter-based methods, prompt-based methods, LoRA methods, quantization-aware methods, and hybrid PEFT architectures. These approaches can be further grouped into additive-based techniques, selective parameter-updating methods, and reparameterization approaches. PEFT methods can be broadly categorized into additive-based techniques (adapters, prompt tuning, prefix tuning), selective parameter updating methods (BitFit, layer freezing, partial fine-tuning), and reparameterization approaches (LoRA, QLoRA, DoRA). These techniques can support efficient adaptation of vision encoders, language encoders, and cross-modal fusion modules while significantly reducing trainable parameters, computational cost, and memory requirements for deployment in resource-constrained environments such as wearable AI systems and edge devices. The taxonomy further highlights emerging hybrid PEFT architectures for multimodal VLMs and their deployment in wearable AI applications, including activity recognition, health monitoring, assistive systems, and edge intelligence.

The deployment of large VLMs in wearable environments requires computational architectures that balance model performance, latency, privacy, and resource constraints. Several deployment frameworks have emerged to support wearable multimodal intelligence.

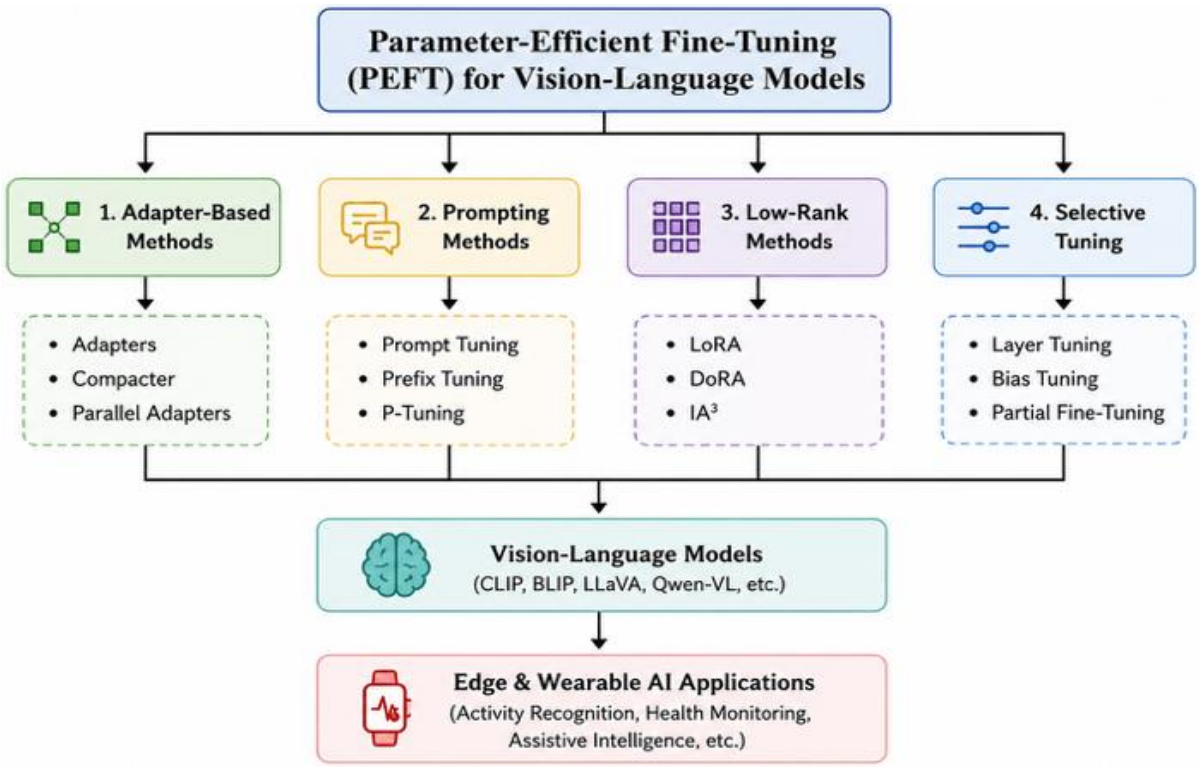

Figure 3. Taxonomy of Parameter-Efficient Fine-Tuning (PEFT) techniques for Vision-Language Models (VLMs)

On-device deployment represents the most privacy-preserving architecture because all adaptation and inference processes occur locally on wearable hardware. This approach minimizes communication latency and eliminates dependence on external infrastructure. However, the severe computational constraints of wearable devices often limit the complexity of models that can be deployed. PEFT methods can improve the feasibility of adapting large models in resource-constrained environments by reducing the number of trainable parameters and associated adaptation costs. However, actual on-device deployment depends on the computational capability, memory capacity, energy budget, thermal characteristics, and software support of the target wearable platform.

5. WEARABLE DEPLOYMENT FRAMEWORKS FOR PARAMETER-EFFICIENT FINE-TUNING-ENABLED VISION-LANGUAGE MODELS

Edge-assisted deployment distributes computation between wearable devices and nearby edge servers. In this architecture, computationally intensive operations are offloaded to edge infrastructure while latency-sensitive tasks remain local. This framework provides an attractive compromise between performance and responsiveness. The integration of PEFT with edge computing may reduce adaptation and communication costs; however, latency and energy performance depend on the characteristics of the edge hardware, communication link, model architecture, and workload.

Federated PEFT represents an emerging approach for privacy-sensitive wearable and distributed multimodal systems. In federated learning environments, wearable devices collaboratively train shared models without exchanging raw data, while PEFT can reduce communication costs associated with parameter transmission. Table 4 summarizes the relationship between representative wearable scenarios, dominant constraints, and potentially relevant PEFT strategies.

Table 4. Evidence-based mapping of Parameter-Efficient Fine-Tuning (PEFT) strategies to resource-constrained wearable AI scenarios

Wearable Scenario	Dominant Constraint	Potentially Relevant PEFT Strategy	Evidence-Based Rationale	Evidence Status
Smart glasses	Latency and memory	LoRA	Low trainable-parameter requirement may reduce adaptation overhead	Emerging
Smartwatches	Memory and energy	QLoRA	Combines low-rank adaptation with low-bit quantization	Emerging
Medical wearables	Privacy and personalization	Federated PEFT	Reduces the need for centralized model adaptation data	Research-stage
AR headsets	Latency and task adaptation	Adapters / LoRA	Supports modular or task-specific adaptation	Emerging
Industrial wearables	Edge resources and robustness	LoRA / Hybrid PEFT	Potentially reduces adaptation resource requirements	Research-stage

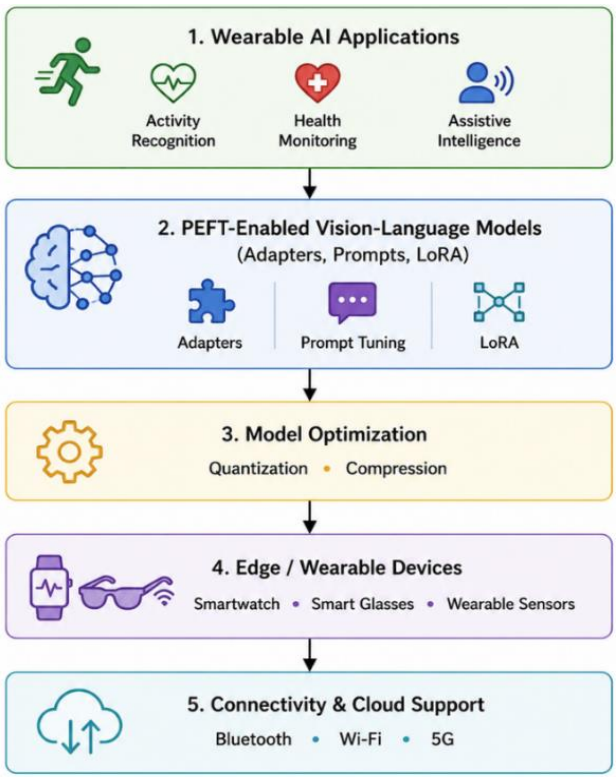

Figure 4. Wearable deployment framework for parameter-efficient fine-tuned Vision-Language Models (VLMs)

Figure 4 illustrates the deployment pipeline of PEFT-enabled VLMs for wearable AI applications. Lightweight adaptation methods such as adapters, prompt tuning, and LoRA enable efficient model customization, while optimization techniques facilitate deployment on resource-constrained wearable devices. Connectivity services provide optional cloud support for synchronization, updates, and advanced analytics.

5.1 Benchmarking and evaluation frameworks

The evaluation of PEFT methods requires a multidimensional benchmarking framework that considers both effectiveness and efficiency. Traditional machine learning evaluations focus primarily on predictive performance metrics such as accuracy, precision, recall, and F1-score. While these metrics remain important, wearable AI introduces additional evaluation criteria that are equally critical.

Multimodal benchmarks commonly employ datasets such as VQAv2, COCO Captions, Flickr30K, ScienceQA, TextVQA, and ImageNet to evaluate visual reasoning and language understanding capabilities. These datasets provide valuable insights into model effectiveness but do not adequately capture the operational constraints of wearable environments.

A benchmarking framework for wearable VLMs should therefore consider multiple dimensions, including model

performance, multimodal reasoning capability, computational efficiency, deployment feasibility, privacy, and sustainability, as summarized in Table 5.

Table 5. Benchmarking metrics for wearable Vision-Language Models (VLMs)

Category	Metric
Performance	Accuracy, Precision, Recall, F1-Score
Multimodal Reasoning	BLEU, CIDEr, ROUGE, VQA Accuracy
Efficiency	FLOPs, Throughput, Memory Usage
Wearable Deployment	Battery Consumption, Latency, Thermal Footprint
Privacy	Information Leakage Risk
Sustainability	Carbon Emissions, Energy-related implications

Consequently, wearable benchmarking frameworks increasingly incorporate metrics related to computational efficiency. Inference latency measures the responsiveness of deployed systems and is particularly important for real-time applications such as augmented reality and assistive navigation. Memory footprint quantifies storage requirements and directly influences deployability on resource-constrained devices. Energy consumption evaluates battery utilization and remains one of the most critical metrics for wearable applications. Thermal efficiency measures heat generation during operation, while communication overhead quantifies data transmission requirements in distributed architectures.

Recent studies have demonstrated that LoRA and QLoRA achieve competitive performance relative to full fine-tuning while significantly reducing memory consumption and computational cost. The reviewed evidence indicates that PEFT methods can provide a favorable balance between model adaptation and resource efficiency. These characteristics suggest potential suitability for wearable AI environments, particularly where memory and computational resources are constrained. However, the extent to which these benefits translate into reliable real-world wearable deployment remains dependent on hardware characteristics, workload requirements, energy availability, and application-specific validation.

To avoid implying a standardized ranking where comparable wearable measurements are unavailable, the PEFT methods are compared using evidence-based qualitative descriptions rather than symbolic scores. The comparison considers trainable-parameter requirements, memory implications, latency and computational characteristics, personalization potential, federated-learning compatibility, and the availability of evidence for wearable deployment. The comparative characteristics of representative PEFT methods under these criteria are summarized in Table 6. These characteristics are configuration-dependent and should therefore not be interpreted as universal performance rankings. Energy consumption is treated separately because reductions in trainable parameters or memory requirements do not necessarily translate into proportional reductions in energy consumption.

Table 6. Comparative characteristics of representative Parameter-Efficient Fine-Tuning (PEFT) methods for resource-constrained wearable AI

Constraint / Criterion	LoRA	QLoRA	Adapters	Prompt Tuning	Hybrid PEFT
Low latency	Rank- and architecture-dependent; inference overhead depends on implementation	Hardware- and quantization-dependent; low-bit execution may reduce memory traffic where supported	Additional adapter computation may introduce modest inference overhead	Prompt-length and architecture-dependent	Depends on combination and implementation
Low memory	Reduced trainable-state memory; frozen backbone remains	Strong memory-reduction potential through low-bit quantization	Reduced trainable parameters, but adapter modules add parameters	Very small trainable state	Depends on constituent methods
Energy-related implications	Direct wearable energy evidence remains limited	Potential memory-related energy benefits, but no universal wearable energy advantage established	Direct wearable energy evidence remains limited	Direct wearable energy evidence remains limited	Highly configuration- and hardware-dependent
Personalization	Suitable for task-specific adaptation; evidence depends on configuration	Suitable for parameter-efficient task adaptation	Strong modular personalization potential	Suitable for lightweight task/domain adaptation	Can support multi-objective or multimodal personalization
Federated learning	Potentially suitable because only a small update state may need to be communicated	Potentially suitable, although quantization and aggregation introduce additional considerations	Potentially suitable for modular local adaptation	Potentially suitable because of small trainable state	Potentially suitable but communication and coordination costs depend on the combination
Evidence strength for wearable deployment	Primarily technical/experimental evidence; physical wearable validation remains limited	Primarily technical/experimental evidence; wearable hardware validation remains limited	Primarily technical/experimental evidence	Primarily technical/experimental evidence	Evidence is comparatively heterogeneous
Key limitation	Rank selection and target-module choice affect performance and efficiency	Quantization can affect accuracy and requires suitable hardware/software support	Additional modules may increase model complexity	Performance may depend strongly on prompt design and length	Greater implementation and optimization complexity

Table 6 highlights the trade-offs among representative PEFT techniques for wearable AI applications. Full fine-tuning provides maximum adaptation flexibility but generally requires substantially more trainable parameters and adaptation resources than PEFT, which can make it difficult to support on resource-constrained wearable platforms. In contrast, Prompt tuning and prefix tuning update only a very small fraction of model parameters and can provide high adaptation-state efficiency; however, their inference overhead depends on prompt or prefix length and the underlying model architecture. Adapter-based methods provide a balance between adaptability and computational efficiency but introduce additional architectural components.

LoRA and its variants, including QLoRA, AdaLoRA, and DoRA, have demonstrated favorable trade-offs between adaptation performance and parameter or memory efficiency in several reported settings.. Their ability to maintain competitive downstream performance while updating only a small subset of parameters makes them potentially relevant to wearable environments with stringent memory and computational constraints; however, suitability with respect to battery consumption requires hardware-specific validation. Among these methods, QLoRA may be particularly attractive for memory-constrained edge and wearable scenarios because low-bit quantization can substantially reduce model memory requirements while maintaining competitive task performance in the settings reported by the literature. Hybrid PEFT approaches further improve flexibility by combining complementary adaptation mechanisms, although they require more careful optimization and system design.

Importantly, performance on conventional multimodal benchmarks should not be interpreted as equivalent to successful wearable deployment. Benchmark datasets primarily assess model-level capabilities, whereas real-world wearable operation additionally depends on latency, memory availability, battery consumption, thermal behavior, communication reliability, privacy, and sustained operation. Therefore, application claims concerning healthcare, industrial monitoring, and assistive systems should be interpreted in relation to the level of deployment evidence reported by each study.

6. GAPS AND OPEN CHALLENGES

6.1 Multimodal alignment

Recent studies have demonstrated that PEFT techniques, particularly LoRA, QLoRA, and adapter-based methods, can effectively adapt large VLMs while updating only a small fraction of model parameters. These approaches preserve most pretrained knowledge and achieve competitive performance across a wide range of multimodal tasks. However, the reviewed literature also indicates that parameter-efficient adaptation may not consistently preserve cross-modal alignment when models are deployed in potentially specialized wearable scenarios, such as personalized healthcare monitoring and industrial safety. Misalignment between visual and textual representations can reduce contextual reasoning accuracy under dynamic operating conditions. Future research should therefore investigate adaptive multimodal alignment mechanisms that continuously preserve semantic consistency while accommodating evolving user contexts and resource constraints.

6.2 Continual learning and model adaptation

Existing studies have shown that PEFT substantially reduces the computational cost of adapting large foundation models and supports incremental model updates without retraining the entire network. Nevertheless, most current approaches assume relatively static downstream tasks and are not specifically designed for continuously evolving wearable environments. As wearable devices generate non-stationary data streams and user behaviors change over time, PEFT-enabled models remain susceptible to catastrophic forgetting and performance degradation during continual adaptation. Future research should focus on lifelong multimodal learning frameworks that integrate PEFT with continual learning strategies capable of preserving previously acquired knowledge while adapting efficiently to new tasks and environmental changes.

6.3 Privacy and secure federated adaptation

Federated PEFT has emerged as a promising approach for enabling collaborative model adaptation without requiring centralized collection of sensitive user data. Although this paradigm significantly reduces direct data sharing, recent studies continue to report challenges related to communication overhead, gradient leakage attacks, model inversion risks, and limited scalability across heterogeneous wearable devices. These issues are particularly important for healthcare and other privacy-sensitive applications. Future research should therefore develop communication-efficient, privacy-preserving PEFT frameworks that integrate secure aggregation, differential privacy, and robust defense mechanisms while maintaining acceptable adaptation performance in distributed wearable ecosystems.

6.4 Explainability and trustworthy multimodal intelligence

Several studies have explored attention visualization and feature attribution techniques to improve the interpretability of large multimodal models. However, the decision-making processes of PEFT-enabled VLMs remain difficult to explain, particularly when deployed in safety-critical wearable applications. Limited interpretability may reduce user trust and hinder regulatory acceptance in domains such as healthcare and industrial monitoring. Future investigations should therefore focus on explainable PEFT architectures capable of providing transparent multimodal reasoning, uncertainty estimation, and human-interpretable explanations without significantly increasing computational complexity.

6.5 Energy-aware and hardware-constrained optimization

Recent advances, including QLoRA and other quantization-aware PEFT techniques, have significantly reduced memory requirements and computational complexity, thereby improving the feasibility of deploying large VLMs on resource-constrained devices. Despite these improvements, relatively few studies explicitly optimize adaptation strategies for battery consumption, thermal efficiency, hardware heterogeneity, and long-term operational sustainability in wearable systems. Differences in processor architectures, memory capacity, and communication capabilities continue to affect deployment performance across wearable platforms. Future research should investigate hardware-aware and energy-aware PEFT algorithms that jointly optimize model

accuracy, computational efficiency, power consumption, and deployment scalability across diverse wearable devices.

6.6 Standardized benchmarking for wearable AI

Current benchmarking studies primarily evaluate PEFT techniques using conventional metrics such as accuracy, precision, recall, memory usage, and computational efficiency. While these metrics provide valuable insights, they do not fully capture the operational characteristics of wearable AI, including battery endurance, thermal behavior, communication latency, user personalization, and long-term adaptation performance. The absence of standardized evaluation protocols makes objective comparison among competing PEFT approaches difficult. Future research should establish comprehensive benchmarking frameworks specifically designed for wearable AI, incorporating both predictive performance and system-level deployment metrics to facilitate fair and reproducible evaluation. The major research gaps and corresponding future opportunities in PEFT-enabled VLMs for wearable AI are summarized in Table 7.

Table 7. Research gap matrix

Research Area	Existing Solutions	Limitation	Research Opportunity
Multimodal Alignment	LoRA, Adapters	Reduced flexibility	Adaptive alignment modules
Continual Learning	Incremental PEFT	Forgetting	Lifelong PEFT frameworks
Privacy	Federated Learning	Communication overhead	Communication-efficient federated PEFT
Explainability	Attention Visualization	Limited interpretability	Explainable multimodal adaptation
Energy-related implications	QLoRA	Limited optimization	Battery-aware PEFT

A notable limitation of the current literature is the limited availability of long-term, real-world evaluations of PEFT-enabled VLMs on physical wearable platforms. Much of the available evidence concerns model-level performance, parameter efficiency, computational requirements, or conceptual deployment architectures rather than sustained operation under realistic wearable conditions. Consequently, reported benefits should not be interpreted as evidence that PEFT-enabled VLMs are already suitable for all healthcare, industrial, or assistive applications. Future studies should evaluate these systems on representative wearable hardware using realistic workloads and should report end-to-end latency, energy consumption, thermal behavior, communication overhead, reliability, privacy, and user-centered performance. Such evaluations would provide stronger evidence for translating PEFT-based multimodal models from experimental settings into operational wearable information systems.

7. FUTURE DIRECTIONS

The future of wearable multimodal intelligence will likely be shaped by the convergence of parameter-efficient adaptation, edge computing, federated learning, and next-generation hardware architectures. Adaptive PEFT systems capable of dynamically selecting optimal fine-tuning strategies based on task complexity, hardware availability, and energy budgets represent a particularly promising direction.

Neuromorphic computing offers another compelling opportunity. Event-driven architectures inspired by biological neural systems may enable highly efficient multimodal processing while reducing energy consumption. The integration of PEFT techniques with neuromorphic hardware could provide an alternative pathway for improving energy-efficient multimodal processing in wearable systems, although its practical benefits require further empirical validation.

Federated multimodal learning is expected to become increasingly important as privacy regulations continue to evolve. By enabling collaborative adaptation across distributed wearable ecosystems without centralized data collection, federated PEFT frameworks may provide a scalable pathway for privacy-preserving multimodal intelligence.

Retrieval-augmented multimodal adaptation represents another emerging area. Future wearable systems may dynamically access external knowledge repositories, visual memories, and contextual databases to enhance reasoning capabilities without increasing model size. Such architectures could provide sophisticated contextual intelligence while maintaining computational efficiency.

Finally, future VLMs may be designed specifically for wearable environments rather than adapted from cloud-centric architectures. Edge-native multimodal foundation models optimized for low-power hardware could reduce some of the deployment challenges associated with large VLMs, although substantial computational, thermal, energy, and communication constraints would remain.

8. CONCLUSION

Overall, the reviewed evidence indicates that PEFT offers a practical approach for adapting large VLMs to wearable AI applications by reducing the number of parameters that require updating and facilitating deployment under constrained computational conditions. The findings further suggest that the effectiveness of PEFT depends on the adaptation strategy, model architecture, hardware configuration, and deployment setting, rather than on parameter reduction alone. PEFT therefore represents an important enabling technology for wearable multimodal information processing and edge-oriented AI, while its implications for computational and energy efficiency should be evaluated according to specific implementation and workload conditions.

The literature indicates that PEFT can substantially reduce the number of trainable parameters and adaptation-related computational requirements. Memory and energy benefits are promising but remain dependent on model architecture, hardware implementation, quantization support, and workload characteristics.

These improvements indicate that PEFT is relevant to resource-constrained wearable application scenarios, including healthcare monitoring, augmented and mixed reality, assistive technologies, and industrial information systems. However, the reviewed evidence does not establish

uniform real-world deployment across these domains. Long-term validation on physical wearable hardware, including evaluation of energy consumption, thermal behavior, latency, privacy, reliability, and user-specific requirements, remains necessary before widespread operational deployment can be established. Accordingly, these domains should be regarded as promising application areas rather than universally validated deployment environments.

Furthermore, recent advances in LoRA, QLoRA, AdaLoRA, DoRA, and hybrid PEFT approaches have expanded the range of adaptation strategies available for wearable platforms with varying computational capabilities and deployment constraints. The review also highlights that selecting an appropriate PEFT strategy depends on application-specific requirements, including latency, memory availability, energy-related implications, privacy, and communication constraints.

Despite these advances, the reviewed literature identifies several important challenges that continue to limit the widespread adoption of PEFT-enabled wearable VLMs. Maintaining multimodal alignment during adaptation, mitigating catastrophic forgetting in continual learning environments, preserving privacy in distributed and federated deployments, improving model explainability, and accommodating hardware heterogeneity remain open research problems. In addition, aggressive parameter reduction and low-bit quantization may introduce performance degradation for complex multimodal reasoning tasks, while communication overhead and resource limitations continue to constrain large-scale deployment in wearable ecosystems. These findings indicate that further methodological and system-level research is required before PEFT-enabled VLMs can be broadly deployed in real-world wearable applications.

Based on the reviewed evidence, future research should focus on developing adaptive and context-aware PEFT strategies that jointly optimize model performance, computational efficiency, energy consumption, and deployment scalability. Greater attention should also be directed toward communication-efficient federated PEFT, continual multimodal learning, privacy-preserving adaptation, and hardware-aware optimization for next-generation wearable devices. Furthermore, standardized benchmarking frameworks and real-world deployment studies are needed to facilitate objective comparison among emerging PEFT techniques and accelerate their translation into practical wearable information systems. Continued advances in these areas are expected to strengthen the integration of giant VLMs into intelligent, resource-aware wearable platforms while addressing the technical challenges identified throughout this review.

REFERENCES

[1] Radford, A., Kim, J.W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020. https://doi.org/10.48550/arXiv.2103.00020

[2] Alayrac, J.B., Donahue, J., Luc, P., et al. (2022). Flamingo: A visual language model for few-shot learning. Advances in Neural Information Processing Systems, 35: 23716-23736. https://doi.org/10.52202/068431-1723

[3] Li, J., Li, D., Savarese, S., Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International Conference on Machine Learning, pp. 19730-19742.

[4] Peng, Z., Wang, W., Dong, L., et al. (2023). Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824. https://doi.org/10.48550/arXiv.2306.14824

[5] Anil, R., Borgeaud, S., Alayrac, J.B., et al. (2023). Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805. https://doi.org/10.48550/arXiv.2312.11805

[6] OpenAI. (2023). GPT-4V(ision) System Card. https://openai.com/index/gpt-4v-system-card/.

[7] Lee, Y.J., Li, C., Liu, H., Wu, Q. (2023). Visual instruction tuning. Advances in Neural Information Processing Systems, 36: 34892-34916. https://doi.org/10.52202/075280-1516

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929

[9] Chen, M., Hao, Y., Hwang, K., Wang, L., Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. IEEE Access, 5: 8869-8879. https://doi.org/10.1109/ACCESS.2017.2694446

[10] Majumder, S., Mondal, T., Deen, M.J. (2017). Wearable sensors for remote health monitoring. Sensors, 17(1): 130. https://doi.org/10.3390/s17010130

[11] Hu, C.Y., Hu, L.S., Yuan, L., Lu, D.J., Lyu, L., Chen, Y.Q. (2023). FedIERF: Federated incremental extremely random forest for wearable health monitoring. Journal of Computer Science and Technology, 38(5): 970-984. https://doi.org/10.1007/s11390-023-3009-0

[12] Hao, Y., Song, H., Dong, L., et al. (2022). Language models are general-purpose interfaces. arXiv preprint arXiv:2206.06336. https://doi.org/10.48550/arXiv.2206.06336

[13] Ding, N., Qin, Y., Yang, G., et al. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. Nature Machine Intelligence, 5(3): 220-235. https://doi.org/10.1038/s42256-023-00626-4

[14] Deng, S., Zhao, H., Fang, W., Yin, J., Dustdar, S., Zomaya, A.Y. (2020). Edge intelligence: The confluence of edge computing and artificial intelligence. IEEE Internet of Things Journal, 7(8): 7457-7469. https://doi.org/10.1109/JIOT.2020.2984887

[15] Hu, E.J., Shen, Y., Wallis, P., et al. (2021). LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685. https://doi.org/10.48550/arXiv.2106.09685

[16] Li, X.L., Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp. 4582-4597. https://doi.org/10.18653/v1/2021.acl-long.353

[17] Lester, B., Al-Rfou, R., Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 3045-3059. https://doi.org/10.18653/v1/2021.emnlp-main.243

[18] Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMS. Advances in Neural Information Processing Systems, 36: 10088-10115. https://doi.org/10.52202/075280-0441

[19] Zhang, Q., Chen, M., Bukharin, A., et al. (2023). AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. arXiv preprint arXiv:2303.10512. https://doi.org/10.48550/arXiv.2303.10512

[20] Liu, S.Y., Wang, C.Y., Yin, H., et al. (2024). DoRA: Weight-decomposed low-rank adaptation. arXiv preprint arXiv:2402.09353. https://doi.org/10.48550/arXiv.2402.09353

[21] Kopiczko, D., Blankevoort, T., Asano, Y. (2024). VeRA: Vector-based random matrix adaptation. In International Conference on Learning Representations, pp. 6815-6835.

[22] Han, Z., Gao, C., Liu, J., Zhang, J., Zhang, S.Q. (2024). Parameter-efficient fine-tuning for large models: A comprehensive survey. arXiv preprint arXiv:2403.14608. https://doi.org/10.48550/arXiv.2403.14608

[23] Houlsby, N., Giurgiu, A., Jastrzebski, S., et al. (2019). Parameter-efficient transfer learning for NLP. arXiv preprint arXiv:1902.00751. https://doi.org/10.48550/arXiv.1902.00751

[24] Zhou, X., He, J., Ke, Y., Zhu, G., Gutiérrez-Basulto, V., Pan, J. (2024). An empirical study on parameter-efficient fine-tuning for multimodal large language models. In Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, pp. 10057-10084. https://doi.org/10.18653/v1/2024.findings-acl.598

[25] Hoque, M., Chowdhury, R.N., Hasan, M.R., Peter, O.O.E., Khalifa, F., Rahman, M.M. (2025). An empirical evaluation of low-rank adapted vision–language models for radiology image captioning. Bioengineering, 12(12): 1330. https://doi.org/10.3390/bioengineering12121330

[26] Li, M., Zhong, J., Li, C., Li, L., Lin, N., Sugiyama, M. (2024). Vision-language model fine-tuning via simple parameter-efficient modification. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, Florida, USA, pp. 14394-14410. https://doi.org/10.18653/v1/2024.emnlp-main.797

[27] Wang, T., Liu, Y., Liang, J.C., et al. (2024). M2PT: Multimodal prompt tuning for zero-shot instruction learning. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, USA, pp. 3723-3740. https://doi.org/10.18653/v1/2024.emnlp-main.218

[28] Liang, T., Huang, J., Kong, M., Chen, L., Zhu, Q. (2024). Querying as prompt: Parameter-efficient learning for multimodal language model. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 26845-26855. https://doi.org/10.1109/CVPR52733.2024.02536

[29] Speicher, M., Hall, B.D., Nebeling, M. (2019). What is mixed reality? In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, New York, USA, pp. 1-15. https://doi.org/10.1145/3290605.3300767

[30] Kaur, K., Batth, J.S. (2024). Human activity recognition using machine learning. In 2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0, Raigarh, India, pp. 1-5. https://doi.org/10.1109/OTCON60325.2024.10688067

[31] Boyes, H., Hallaq, B., Cunningham, J., Watson, T. (2018). The Industrial Internet of Things (IIoT): An analysis framework. Computers in Industry, 101: 1-12. https://doi.org/10.1016/j.compind.2018.04.015

[32] Satyanarayanan, M. (2017). The emergence of edge computing. Computer, 50(1): 30-39. https://doi.org/10.1109/MC.2017.9

[33] Bonawitz, K., Eichner, H., Grieskamp, W., et al. (2019). Towards federated learning at scale: System design. arXiv preprint arXiv:1902.01046. https://doi.org/10.48550/arXiv.1902.01046