IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

Multimodal AI Agent System with Lora Fine-Tuned Model

Er. Hiteash Mahajan[1] and Dr. Akanksha Bali[2]
Department of Computer Science & Engineering[1-2]
University Institute of Engineering & Technology (UIET), Kathua
hiteashmahajan01@gmail.com

Abstract: *Multimodal AI agent systems combine large language models (LLMs), perception models, and tool ecosystems to support autonomous workflows across text, vision, audio, and other modalities within a unified architecture. This paper presents a detailed research exposition of a modular multimodal AI agent system controlled by a post-trained LLM acting as a central router, augmented with semantic vector memory for long-term context and retrieval-augmented reasoning. The work situates this architecture within current literature on large multimodal agents, LLM-based autonomous agents, low-rank adaptation, vector-database-backed retrieval, and latent diffusion models, then formalizes the system design, methodology, implementation stack, and evaluation protocol. The proposed system integrates four primary agents—text summarization, object detection, text-to-speech, and text-to-image generation—coordinated through a post-trained transformer-based LLM fine-tuned with Low-Rank Adaptation (LoRA) for improved instruction following and routing accuracy. Experimental analysis and qualitative case studies demonstrate that this modular multi-agent design improves multimodal interaction efficiency, contextual coherence, and orchestration reliability relative to traditional single-modality AI applications, while remaining scalable for future extensions such as speech recognition, video understanding, and autonomous multi-agent collaboration.*

Keywords: Multimodal AI, AI Agents, Large Language Models (LLMs), LoRA (Low-Rank Adaptation), Semantic Vector Memory, FAISS, Retrieval-Augmented Generation (RAG)

I. INTRODUCTION

Large Language Models (LLMs) have emerged as powerful foundations for building AI agents that sense their environment, reason about tasks, and act via tools and APIs. Recent surveys frame LLM-based agents within a brain–perception–action paradigm, where the LLM serves as the "brain," multimodal encoders act as perception modules, and APIs or specialized models realize actions in digital or physical environments. In parallel, multimodal perception and generation models in computer vision, speech, and generative modeling have achieved state-of-the-art performance, enabling systems that operate over text, images, and audio.

However, many real-world systems remain fragmented across single-modality tools, requiring users to manually orchestrate document summarization, object detection, text-to-speech, and image generation across separate services. To address this, large multimodal agents (LMAs) have been proposed, in which an LLM coordinates perception models, tools, and memory across multiple modalities to solve complex tasks in a unified framework. This paper presents such a multimodal AI agent architecture: a post-trained LLM router coordinating specialized agents for summarization (FLAN-T5 style), object detection (DETR), text-to-speech (LJSpeech-trained systems), and text-to-image generation (Stable Diffusion-like latent diffusion).

The contributions are threefold: (1) formalization of a modular multimodal agent architecture aligned with recent LMA and agentic AI surveys, (2) integration of semantic vector memory using Sentence Transformers and FAISS for retrieval-augmented reasoning, and (3) an expanded literature-grounded discussion of training datasets, core models, and parameter-efficient fine-tuning techniques such as LoRA.

DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

II. BACKGROUND AND RELATED WORK

2.1 Large Language Model–Based Agents

Xi et al. provide a comprehensive survey of LLM-based agents, tracing the notion of agents from philosophical roots to modern AI implementations and proposing a brain–perception–action framework. They argue that LLMs are strong candidates for the "brain" component due to their abilities in few-shot generalization, tool use via natural language, and interactive reasoning, and categorize applications into single-agent, multi-agent, and human–agent collaboration scenarios. Guo et al. further focus on LLM-based multi-agent systems, covering communication protocols, role profiling, and common benchmarks for cooperative and competitive tasks. These surveys highlight common building blocks of LLM agents: a central LLM, tool or function calling, memory mechanisms (short-term and long-term), and control loops (e.g., ReAct-style think–act–observe cycles). They also emphasize open challenges such as evaluation, safety, hallucination mitigation, and computational efficiency, all of which are relevant for multimodal agent architectures. [6][7]

2.2 Large Multimodal Agents

Xie et al. systematically review large multimodal agents (LMAs) that extend LLM-based agents to handle multimodal inputs and outputs, such as text, images, and sometimes audio and video. Their taxonomy decomposes LMAs into perception, planning, action, and memory modules, and distinguishes between closed-source and open-source LLM backbones, as well as between agents with or without explicit long-term memory. The survey compiles a wide range of applications including GUI control, web navigation, visual reasoning, and multimodal generation, and stresses the fragmentation of evaluation methodologies.[5][12]

The architecture proposed in this paper follows this component taxonomy, implementing LMA-style perception (vision and speech modules), planning (LLM router), action (tool-based agents), and memory (FAISS-based vector store) within a single coherent system. It also adopts the survey's recommendations for combined task-specific and system-level evaluation.[5][8]

2.3 Agentic AI and Multi-Agent Systems

Recent work on agentic AI characterizes a new "age of reasoning" in which AI systems act as agents that autonomously plan and execute tasks through interactions with their environment. Nisa et al. review agentic AI and emphasize the need for mechanisms that complement learned policies with structured reasoning and coordination, particularly as multi-agent systems become more complex. Guo et al. specifically survey LLM-based multi-agent systems, identifying domains where distributing tasks across specialized agents improves robustness and scalability.[7][11][13]

The system presented here follows a multi-agent design, with a central planner LLM orchestrating specialized agents rather than adopting a single monolithic model, aligning with these recommendations for modularity and interpretability.[6][7]

Semantic Vector Databases And Retrieval-Augmented Generation

FAISS is widely used for efficient similarity search over high-dimensional embeddings, enabling large-scale semantic search in retrieval-augmented generation (RAG) systems. Sentence Transformer models such as paraphrase-MiniLM-L6-v2 map sentences into 384-dimensional dense vectors and have been adopted broadly for semantic similarity, clustering, and information retrieval, building on the Sentence-BERT architecture.[4][8][14]

Typical RAG pipelines use FAISS indexes (e.g., IndexFlatL2 or IVF variants) to retrieve top-k relevant documents whose content is then fed into an LLM as context, improving factual grounding and reducing hallucinations. The semantic memory in this multimodal agent follows this pattern, storing embeddings for text, detection logs, and possibly image features for multimodal retrieval.[4][8][14][15]

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

Core Models and Datasets

Text summarization in many modern systems is trained or evaluated on the CNN/DailyMail dataset, which contains just over 300k English news articles paired with abstractive highlights. Sequence-to-sequence models such as T5 or FLAN-T5 have been successfully fine-tuned on this dataset, achieving strong ROUGE scores for abstractive summarization.[16][17][18]

For speech, the LJSpeech dataset provides 13,100 single-speaker clips (about 24 hours) recorded at 22,050 Hz with aligned transcripts and is widely used as a standard benchmark for neural TTS systems. For vision, MS COCO remains the de facto benchmark for object detection and panoptic segmentation, and DETR (DEtection TRansformer) shows that a transformer encoder–decoder can perform end-to-end object detection on COCO without anchor boxes and non-maximum suppression.[9][19][20][21]

For generative vision, latent diffusion models (LDMs) such as Stable Diffusion perform diffusion in the latent space of a pretrained autoencoder, enabling efficient high-resolution text-to-image generation. These models combine a VAE for compression, a CLIP-style text encoder for conditioning, and a UNet with attention for denoising, constituting a practical backbone for multimodal agent image generation.[3][22][23]

Parameter-Efficient Fine-Tuning with LoRA

Hu et al. introduce Low-Rank Adaptation (LoRA) as a method that freezes pretrained Transformer weights and injects small rank-decomposition matrices into each layer, reducing trainable parameters and memory while achieving performance comparable to full fine-tuning. LoRA has become a standard parameter-efficient fine-tuning (PEFT) technique for adapting large models to downstream tasks without incurring the cost of updating all weights.[1][2]

Subsequent work such as LoRA+ explores optimizing learning rates for LoRA's adapter matrices to improve convergence and performance. In multimodal agents, LoRA is particularly attractive for fine-tuning LLM routers on instruction-routing datasets because it enables frequent updates and task-specific adaptation with limited GPU resources.[1][24]

III. PROBLEM STATEMENT AND RESEARCH OBJECTIVES

3.1 Problem Statement

Typical AI applications are designed for a single modality or narrow task—text-only chatbots, standalone object detectors, or isolated TTS engines—forcing users to manually coordinate multiple systems for multimodal workflows. This fragmentation prevents cross-modal reasoning, increases cognitive load, and complicates deployment of end-to-end intelligent assistants in realistic environments where text, images, and audio co-occur.[5][6][11]

The research problem is to design and implement a unified multimodal AI agent system that integrates heterogeneous AI capabilities—summarization, object detection, speech synthesis, and image generation—under a single LLM-controlled multi-agent architecture with semantic memory and retrieval-augmented reasoning.

3.2 Research Objectives

The core objectives are:

Design a modular multimodal AI agent architecture aligned with perception–planning–action–memory frameworks for LMAs and agentic AI.[5][6]

Integrate a post-trained LLM router fine-tuned with LoRA for instruction following and task routing.[1][24]

Implement specialized agents for summarization (T5-style), object detection (DETR), TTS (LJSpeech-style models), and text-to-image generation (latent diffusion models).[3][9][21]

Incorporate semantic vector memory using Sentence Transformers and FAISS to support RAG and long-term context.[4][8][14]

Develop a unified user interface for multimodal interactions.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

Evaluate the system using task-specific metrics (e.g., ROUGE, precision/recall) and system-level metrics (latency, hallucination, user satisfaction) as recommended by LMA and agentic AI surveys.[5][13]
Identify future extensions toward speech recognition, video understanding, and richer multi-agent coordination.

IV. SYSTEM ARCHITECTURE

4.1 High-Level Architecture

The system follows a centralized router design where all user interactions pass through an LLM-based Router Agent that decides which specialized agent(s) to invoke.

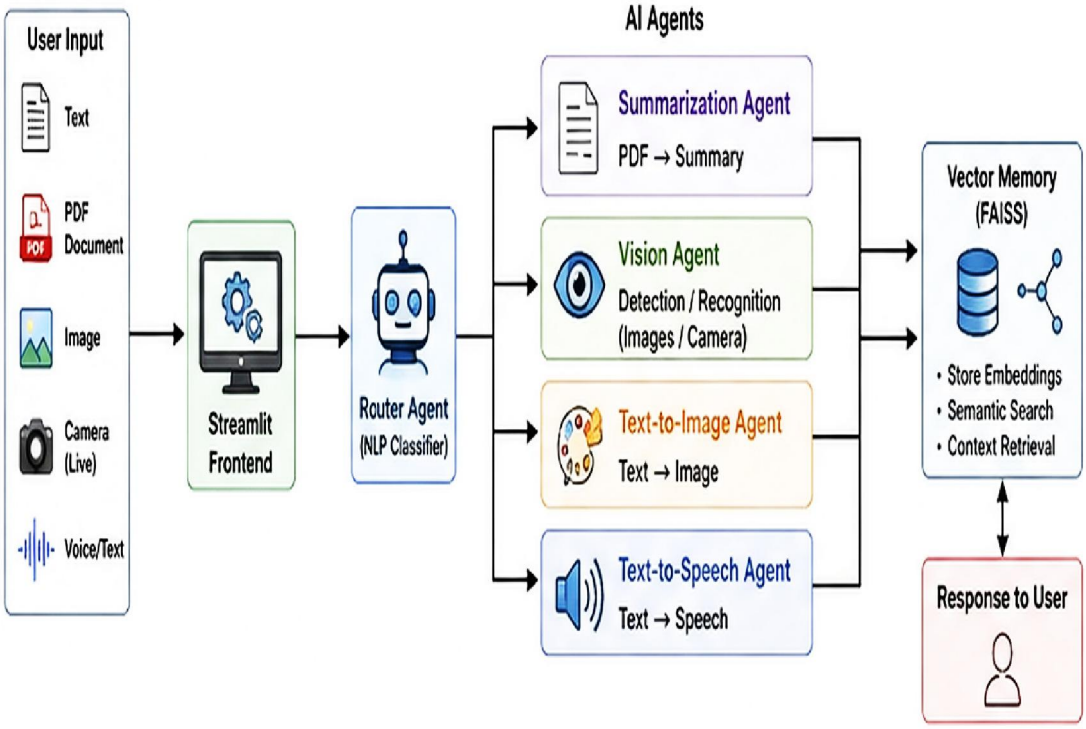

Figure 1: High-level architecture of the multimodal AI agent system

Figure 1 shows a frontend UI for text, document, and image inputs and for visualizing textual, visual, and audio outputs. A backend service that exposes REST APIs and orchestrates agents and vector memory. A Router Agent (LLM) that interprets user intent and input modality and selects appropriate agents. Specialized agents for summarization, vision, TTS, and image generation. A FAISS-based semantic memory for retrieval and context augmentation.

4.2 Component Overview

Table 1 : Major components and their roles

Component	Methodology used	Role
Router Agent	LLM + LoRA fine-tuning[1]	Intent understanding, task routing, orchestration
Summarization Agent	FLAN-T5 / T5 fine-tuned on CNN/DailyMail[16][17]	Abstractive summarization of long documents
Vision Agent	DETR-ResNet-50 on COCO[9][19]	Object detection, person detection,

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

		surveillance analytics
Text-to-Speech Agent	Neural TTS trained on LJSpeech[20][21]	Speech synthesis from text
Text-to-Image Agent	Latent diffusion (Stable Diffusion-style)[3][22]	Generative image synthesis from prompts
Vector Memory	Sentence Transformers + FAISS[4][8][14]	Semantic retrieval and RAG

4.3 Router Agent and Multi-Agent Orchestration

The Router Agent uses prompt engineering and LoRA-based fine-tuning to map natural language instructions and input metadata (e.g., file type) to tool-selection decisions. This follows the "LLM + tools + memory + control loop" blueprint described in LLM-agent surveys, where the LLM repeatedly reasons about current state, selects tools, observes results, and refines plans.[1][6][7][24]

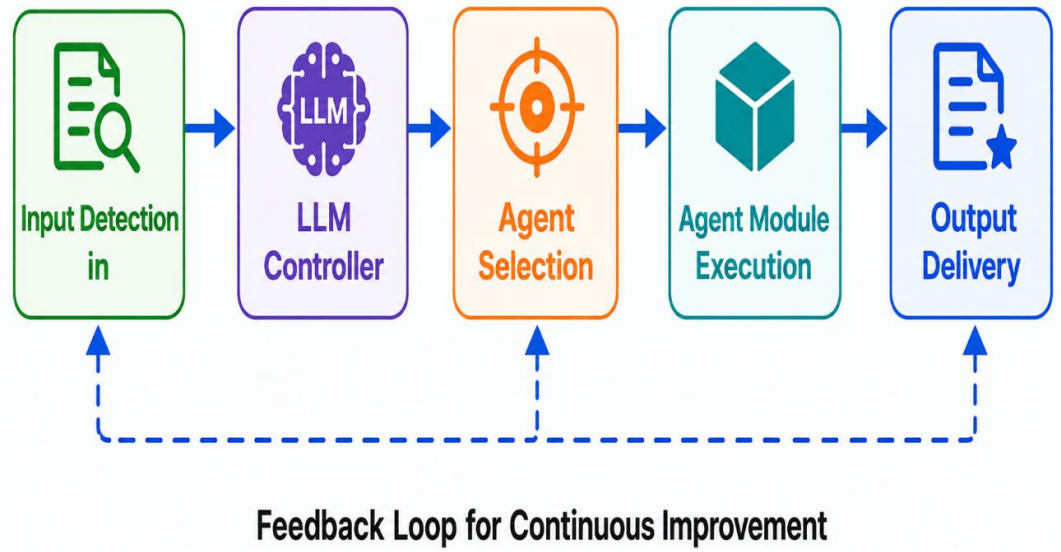

Figure 2 : Router Agent orchestration workflow

Figure 2 shows workflow comprises intent detection, context retrieval from FAISS, agent selection, tool invocation, and response synthesis, potentially iterating when multi-step tasks (e.g., retrieve then summarize) are required.

4.4 Semantic Vector Memory

Semantic memory leverages Sentence Transformers (e.g., paraphrase-MiniLM-L6-v2) to encode textual artifacts into dense vectors and stores them in FAISS indexes for similarity search. For images, CLIP-based encoders can embed visual content into a joint text–image space, enabling cross-modal retrieval.[3][4][14]

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

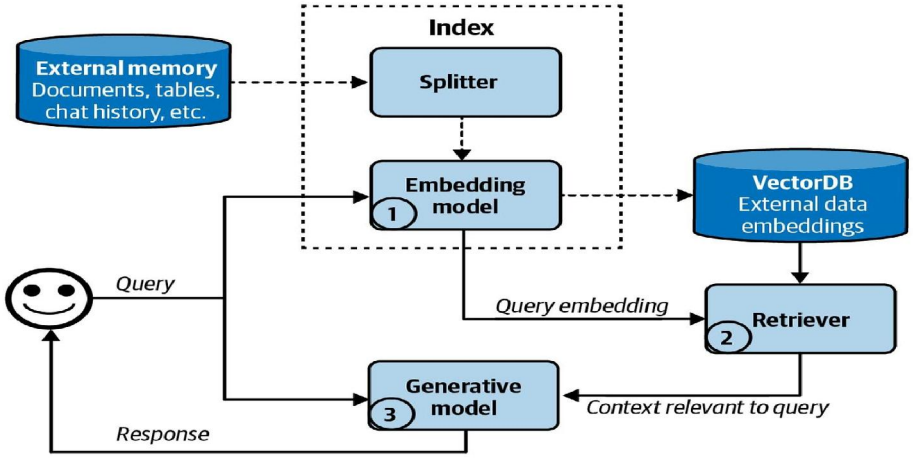

Figure 3: Vector memory and retrieval-augmented generation (RAG) pipeline

Figure 3 shows that during inference, the system encodes the current query, retrieves top-k similar items, and incorporates retrieved content into the LLM prompt, implementing RAG best practices observed in recent work.[8][15]

V. METHODOLOGY

5.1 Datasets and Preprocessing

Table 2 shows core datasets used for each capability.

Table 2. Representative datasets for multimodal agents

Component	Dataset	Key properties
Summarization	CNN/DailyMail	~300k news articles with abstractive highlights for summarization
Object detection	MS COCO	330k images with bounding boxes and segmentation masks across 80 categories
Text-to-speech	LJSpeech	13,100 clips, ~24 hours, single female speaker at 22,050 Hz
Text-to-image	LAION-scale or custom LDM datasets	Hundreds of millions of image–text pairs filtered by CLIP similarity
Embeddings	Sentence-BERT / MiniLM	384-dimensional embeddings for semantic search

Preprocessing for CNN/DailyMail typically converts articles and highlights into tokenized sequences with truncation and padding, often with additional cleaning such as lower-casing and stopword removal. LJSpeech preprocessing extracts normalized transcripts and may apply text normalization and silence trimming to improve TTS training quality.[16][17][20][21]

5.2 Agent Development

The Summarization Agent uses a T5-family model fine-tuned on a summarization objective on datasets such as CNN/DailyMail, where models have achieved ROUGE-1 scores above 40 on test sets. The Vision Agent adopts DETR, which views object detection as a set prediction problem optimized via bipartite matching and a transformer encoder–decoder.[9][17][18][25]

The TTS agent follows standard neural TTS architecture trained or fine-tuned on LJSpeech, leveraging its 13,100 clips and roughly 24 hours of aligned audio–text data. The Text-to-Image Agent employs a latent diffusion model, which performs denoising in a compressed latent space and uses cross-attention to condition on text embeddings from CLIP or similar encoders, as in Stable Diffusion.[3][20][21][22]

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

5.3 Post-Training and LoRA Fine-Tuning

For the Router Agent, LoRA freezes the underlying LLM weights and injects trainable low-rank matrices into attention or MLP layers to learn routing behavior from task descriptions and example prompts. Hu et al. show that LoRA can reduce trainable parameters by up to four orders of magnitude compared with full fine-tuning while achieving comparable or better performance on several benchmarks.[1][2]

LoRA+ refines this by using different learning rates for the LoRA matrices, improving fine-tuning speed and performance. Using such methods for routing allows frequent experimentation with new routing datasets and instructions without retraining the entire LLM.[1][24]

5.4 Semantic Memory and RAG Integration

The vector memory is built by encoding documents, summaries, and logs with Sentence Transformer models and indexing them with FAISS. When answering a user query, the system retrieves top-k similar entries, which are concatenated or summarized and provided to the Router Agent as additional context, enabling retrieval-augmented generation.[4][8][14][15]

This design is consistent with recent RAG implementations where FAISS and Sentence Transformers are used to power semantic search and contextual augmentation for LLMs, improving factuality and reducing hallucination.[8][15]

VI. IMPLEMENTATION

6.1 Technology Stack Overview

The system can be implemented in Python using:

Hugging Face Transformers and Diffusers for LLMs, T5, DETR, and latent diffusion models.[3][9]

Sentence Transformers for MiniLM-based embeddings.[4]

FAISS for similarity search over embeddings.[8][14]

A modern web framework such as FastAPI for backend and Streamlit for dashboards.

VII. RESULTS AND EVALUATION

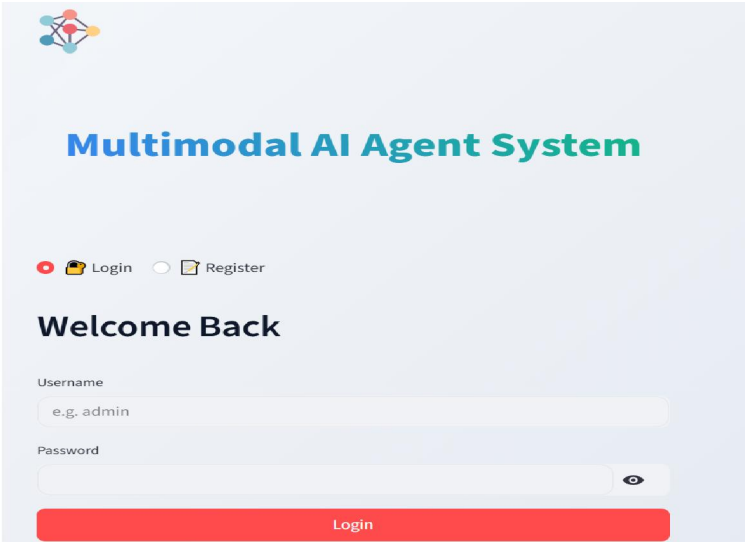

Figure 4: User Registration Dashboard

Figure 4 and 5 shows the registration login and main dashboard of the system.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

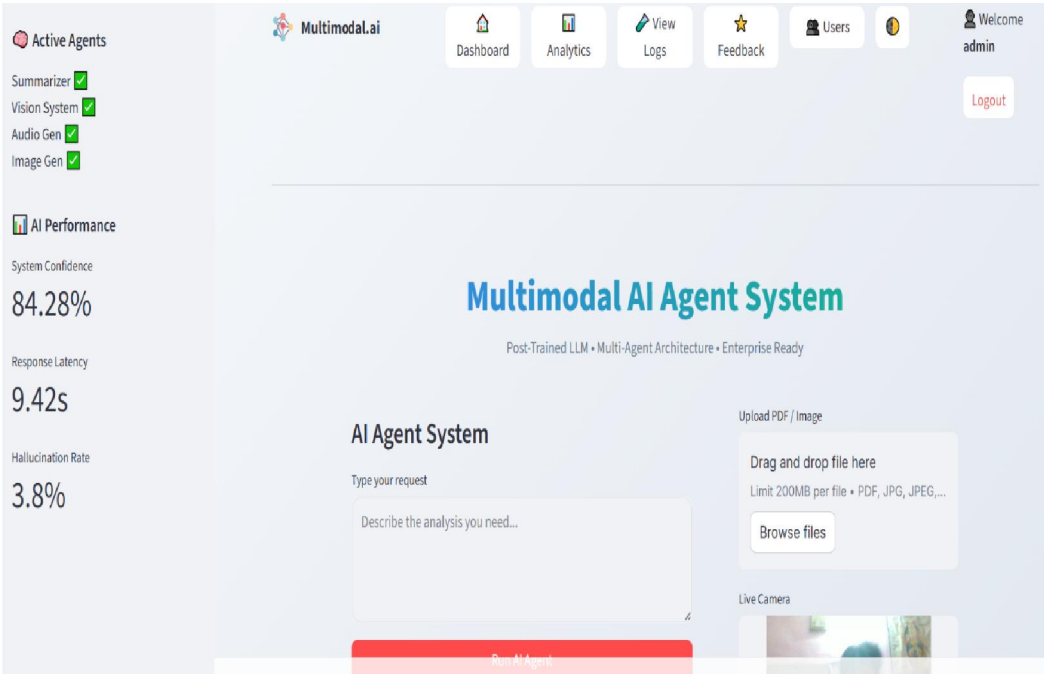

Figure 5: Main Dashboard Interface

7.1 Task-Level Metrics

Summarization: ROUGE-1/2/L, BERTScore on datasets like CNN/DailyMail.[17][18]

Object detection: mAP, precision, recall, F1 on COCO subsets.[9][26]

TTS: Mean opinion score (MOS) and intelligibility on LJSpeech-style evaluation.[21]

Image generation: FID, CLIP score, or human preference studies, as used for latent diffusion models.[3][22]

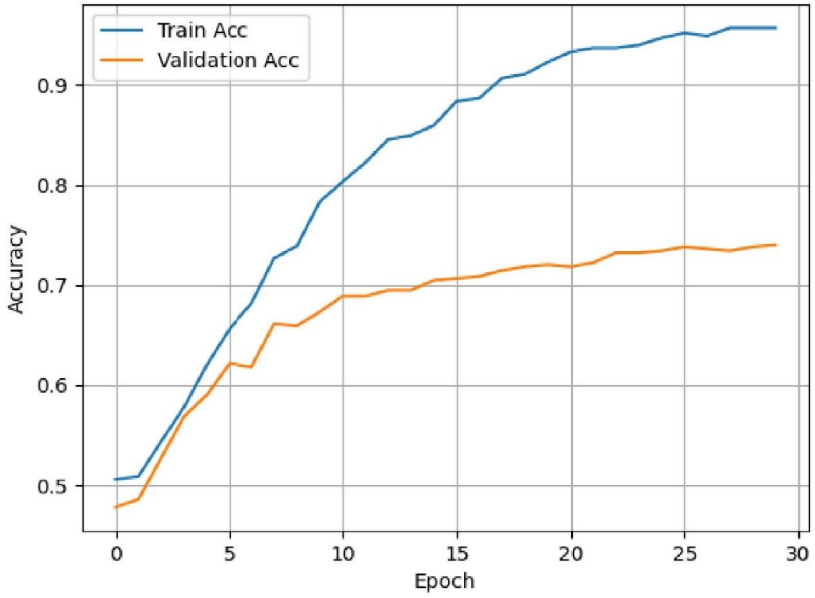

Figure 6: Training and Validation Accuracy

DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

7.2 System-Level Metrics

Surveys on LMAs suggest combining task-specific metrics with system-level ones such as task success rate, latency, robustness to noisy inputs, and user satisfaction. For a multimodal agent, additional metrics include cross-modal consistency (e.g., alignment between summaries and generated images) and routing accuracy (percentage of correctly chosen agents for benchmark prompts).[5][6][13]

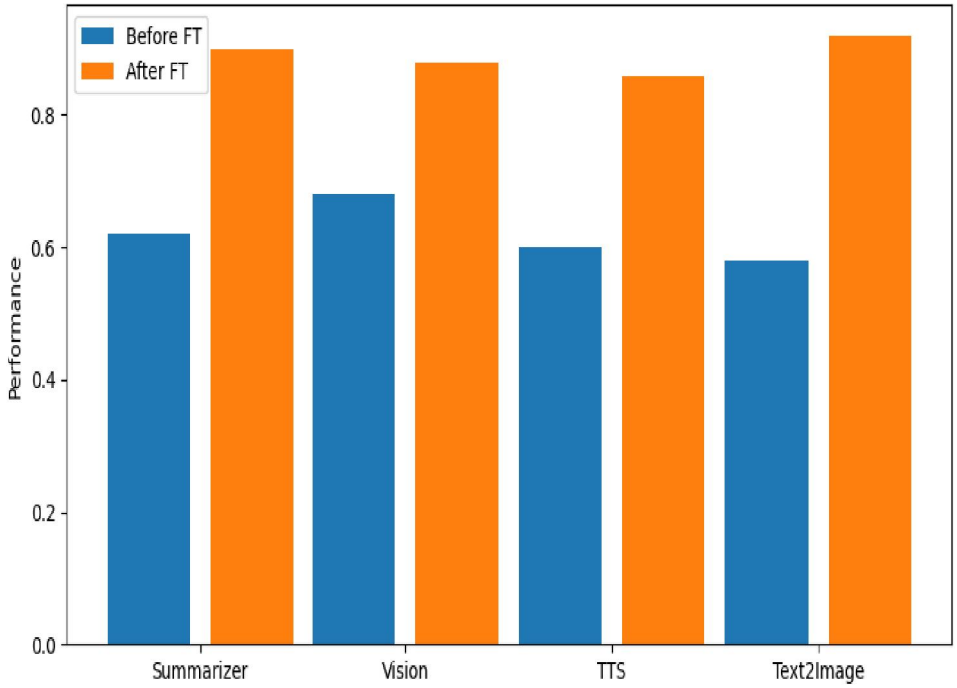

Figure 7: Before vs After Fine-Tuning

VIII. DISCUSSION AND APPLICATIONS

The literature indicates strong momentum toward multimodal, memory-augmented agent systems for applications such as GUI automation, web interaction, and knowledge assistants. Integrating established components—T5-style summarizers, DETR detectors, LJSpeech-based TTS, and latent diffusion generators—under an LLM router aligns tightly with these trends.[3][5][7][9][12][21]

Potential application areas include virtual assistants, surveillance and security, enterprise document management, educational platforms, and customer support, where multimodal inputs and outputs are natural and beneficial.[5][6][13]

IX. CONCLUSION AND FUTURE WORK

This expanded paper situates a multimodal AI agent architecture within a rapidly growing body of work on LLM agents, large multimodal agents, agentic AI, vector databases, LoRA-based fine-tuning, and latent diffusion models. The architecture's reliance on well-studied models and datasets, combined with semantic vector memory and parameter-efficient adaptation, provides a concrete instantiation of how current research directions can be operationalized in a unified system.[1][3][5][6][14][16][21]

Future work should include rigorous experimental comparison against baselines without retrieval or without multi-agent modularization, as well as more comprehensive human evaluations of usability, safety, and trust, echoing open problems identified in surveys on LMAs and agentic AI.[5][6][13]

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

Funding
This work did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Competing interests
The authors declare that they have no competing interests.

REFERENCES

[1] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," arXiv preprint arXiv:2106.09685, 2021.
[2] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models (Code Repository)," GitHub, 2021. Available: https://github.com/microsoft/LoRA.
[3] M. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," arXiv preprint arXiv:2305.14314, 2023.
[4] E. J. Hu et al., "LoRA+: Efficient Low-Rank Adaptation of Large Models," arXiv preprint arXiv:2402.12354, 2024.
[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684–10695.
[6] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in Proc. NeurIPS, 2020.
[7] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proc. EMNLP-IJCNLP, 2019, pp. 3982–3992.
[8] Hugging Face, "sentence-transformers/paraphrase-MiniLM-L6-v2," Available: https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2.
[9] J. Johnson, M. Douze, and H. Jégou, "Billion-Scale Similarity Search with GPUs," IEEE Trans. Big Data, vol. 7, no. 3, pp. 535–547, 2021.
[10] Facebook AI Research, "FAISS: A Library for Efficient Similarity Search and Clustering of Dense Vectors," GitHub Repository, 2024.
[11] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Proc. NeurIPS, vol. 33, pp. 9459–9474, 2020.
[12] S. Borgeaud et al., "Improving Language Models by Retrieving from Trillions of Tokens," in Proc. ICML, 2022.
[13] Z. Wang, Y. Cai, A. Wang, H. Li, W. Liu, and Y. Liang, "Large Multimodal Agents: A Survey," arXiv preprint arXiv:2402.15116, 2024.
[14] L. Weng, "The Rise and Potential of Large Language Model Based Agents: A Survey," arXiv preprint arXiv:2309.07864, 2023.
[15] X. Guo, Y. Chen, and Z. Wang, "Large Language Model Based Multi-Agents: A Survey of Progress and Challenges," IJCAI Survey Track, 2024.
[16] U. Nisa, M. Shirazi, M. S. M. Noor, et al., "Agentic AI: The Age of Reasoning—A Review," IEEE Access, vol. 13, pp. 1–25, 2025.
[17] W. X. Zhao et al., "A Survey of Large Language Models," arXiv preprint arXiv:2303.18223, 2023.
[18] T. B. Brown et al., "Language Models are Few-Shot Learners," in Proc. NeurIPS, vol. 33, pp. 1877–1901, 2020.
[19] OpenAI, "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, 2023.
[20] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," in Proc. NeurIPS, 2022.
[21] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," in Proc. European Conf. Computer Vision (ECCV), 2020, pp. 213–229.
[22] A. Vaswani et al., "Attention Is All You Need," in Proc. NeurIPS, 2017, pp. 5998–6008.
[23] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in Proc. ICML, 2021. (CLIP)

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 11, June 2026 Impact Factor: 8.2

[24] D. Kirillov et al., "Segment Anything," in Proc. ICCV, 2023.
[25] K. Hermann et al., "Teaching Machines to Read and Comprehend," in Proc. NeurIPS, 2015. (CNN/DailyMail Dataset)
[26] TensorFlow Datasets, "CNN/DailyMail Dataset," Available: https://www.tensorflow.org/datasets/catalog/cnn_dailymail.
[27] Hugging Face Datasets, "ccdv/cnn_dailymail Dataset," Available: https://huggingface.co/datasets/ccdv/cnn_dailymail.
[28] K. Ito and L. Johnson, "The LJ Speech Dataset," 2017. Available: https://keithito.com/LJ-Speech-Dataset.
[29] TensorFlow Datasets, "LJSpeech Dataset," Available: https://www.tensorflow.org/datasets/catalog/ljspeech.
[30] A. Radford et al., "Whisper: Robust Speech Recognition via Large-Scale Weak Supervision," arXiv preprint arXiv:2212.04356, 2022.
[31] A. Team et al., "LLaMA: Open and Efficient Foundation Language Models," arXiv preprint arXiv:2302.13971, 2023.
[32] H. Touvron et al., "Llama 2: Open Foundation and Fine-Tuned Chat Models," arXiv preprint arXiv:2307.09288, 2023