RESEARCH PAPER · AI · FOUNDATION MODELS

Multimodal Representation Learning with Large-Scale Foundation Models

This document presents a research report on the development of a real-time adaptive multimodal fusion system that leverages large-scale foundation models.

17 min read • **20** sections • LISRC R&D practice

R&D

LISRC Research & Development
The applied-research practice of London International Studies & Research Centre

Published by: London International Studies and Research Center · Author: Abdul Ameen Adakkani Veedu, Director - Tech Operations, London INTL · Date: October-2024

01 Research Data

02 Executive Summary

This document presents a research report on the development of a real-time adaptive multimodal fusion system that leverages large-scale foundation models. Through collaborative effort at London International Studies and Research Center, we have created a platform that seamlessly integrates voice, vision, and text inputs to facilitate advanced interactions—particularly in educational contexts. Our fusion system rests on two major components: the Neural Vision Matrix (NVM) for rapid visual processing and the Multimodal Fusion Engine (MFE) for adaptive combination of modalities.

The primary objective is to harness large-scale foundation models—trained on vast, diverse data—to adaptively process and analyze text, audio, and video inputs in real time. The approach addresses strict latency requirements, ensuring near-instantaneous feedback and a natural interactive experience. This paper also details how resource constraints, typically encountered in educational settings or edge deployments, can be managed effectively through techniques like model pruning, quantization, and load balancing. Benchmarks performed on NVIDIA DGX systems demonstrate the viability of our architecture in handling concurrent multimodal interactions at millisecond-scale latency.

This foundational research underpins London INTL's recently released low-latency voice and vision models, intended as AI trainers for students. By offering a robust theoretical and practical framework, we envision these systems revolutionizing e-learning experiences worldwide. Finally, the report concludes with policy recommendations for educational institutions, advocating for strategic infrastructure investments, teacher training, data privacy measures, and ongoing evaluation to maximize the benefits of AI-powered learning tools.

03 Introduction

The term multimodal representation learning refers to AI systems that ingest and interpret various data streams—such as audio, text, and images—to form a unified understanding. Large-scale foundation models (e.g., GPT-like generative models, Vision Transformers, or massive speech recognizers) are increasingly used to handle these complex signals. This paper details the design and deployment of a real-time adaptive fusion approach that operates under resource constraints while meeting tight latency requirements.

This research stems from our work at the London International Studies and Research Center (London INTL). We have deployed advanced voice and vision AI models to create an AI tutor capable of interacting with students by both seeing and hearing them. Students can ask questions verbally, share their screens or show their written work via a camera, and receive immediate responses. This fusion of modalities intensifies the sense of interacting with a human tutor, thus elevating engagement and pedagogical outcomes.

However, the computational cost of simultaneously handling video streams and audio input can be immense, especially if the environment is limited to a single GPU or a lightweight edge device. Moreover, for educational applications, sub-100-millisecond latencies are often desired to create seamless, real-time interactions. Large-scale foundation models, while powerful, can also be resource-intensive; hence, strategies for optimization, adaptive processing, and load-shedding become pivotal.

In the following sections, we outline the institutional context, key technical components, and performance benchmarks of our system. Section 2 explores the background and the impetus behind using large-scale foundation models. Section 3 describes the overall architecture, emphasizing the roles of the Neural Vision Matrix (NVM) and the Multimodal Fusion Engine (MFE). Section 4 showcases empirical benchmarks carried out on NVIDIA DGX systems, highlighting how we achieve sub-50-millisecond latencies. Section 5 details the implementation challenges encountered and the corresponding solutions. Section 6 provides policy recommendations for successful adoption in educational settings. Sections 7, Appendices A and B, and references conclude this report.

04 Background and Context

The London International Studies and Research Center (London INTL) has embarked on a comprehensive program to develop and deploy advanced AI technologies in education. Over the past year, our Research & Development Department has successfully introduced two new foundation models:

- Low-Latency Voice Model : This large-scale speech model excels in real-time transcription and natural language generation. Optimized on GPU servers, it achieves near-instantaneous responses, enabling back-and-forth dialogue with students. The model architecture extends a Transformer-based backbone to accelerate streaming speech-to-text and text-to-speech conversions.
- High-Throughput Vision Model : Built around Vision Transformers (ViT), this model can continuously interpret visual scenes (screen shares, webcam feeds) while maintaining minimal delay. It identifies key elements such as student expressions, text on a whiteboard, or content on a shared screen, thus providing context-aware

assistance. Techniques like patch-based attention and early exits help keep processing overhead in check.

While each model has demonstrated utility, their combined deployment (i.e., multimodal fusion) represents the next frontier. A student might hold up a paper with a question and simultaneously articulate a query. The vision model identifies the text, while the voice model captures the student's intonation. Without fusion, these insights remain siloed. Together, they enrich each other—leading to more contextually relevant, human-like responses.

Large-scale foundation models often rely on HPC clusters or specialized hardware such as NVIDIA DGX systems , well-known for parallelism and computational density. These systems allow large-batch processing of data for training and inference. However, the practical challenge arises when we must adapt these models for real-time operation under limited resources , as might be the case in a classroom or a remote learning scenario. This tension between high-performance HPC-based models and the constraints of local hardware forms the crux of our research.

The impetus behind the present study is twofold:

- Educational Efficacy: Evidence suggests that immediate feedback loops in learning boost retention and motivation. Instant acknowledgment or corrective guidance from an AI tutor can be invaluable for a student struggling with a concept.
- Technical Feasibility: Emerging GPU technologies, combined with advanced model compression and adaptive processing, now allow for millisecond-scale latencies in multimodal tasks. This real-time synergy was previously unattainable on modest hardware; we are now on the threshold of making it a reality.

Moving forward, we delve into the system architecture—how these large-scale foundation models are orchestrated to yield real-time performance—while managing resource constraints. Our approach offers a reference blueprint for others looking to implement similar multimodal frameworks in educational or analogous domains.

05 System Architecture Overview

At the heart of our solution are two principal components: the Neural Vision Matrix (NVM) and the Multimodal Fusion Engine (MFE) . These modules integrate seamlessly with our large-scale foundation models for voice and vision, ensuring that data from each modality is processed swiftly and coherently.

06 3.1 Neural Vision Matrix (NVM)

The Neural Vision Matrix is a dedicated pipeline for processing continuous video streams under stringent time constraints. Developed using a Vision Transformer backbone, NVM ingests frames at up to 30 frames per second (FPS), extracting meaningful features like:

- Detected objects (e.g., textbooks, calculators, whiteboard content)
- Text recognition (handwritten or typed)
- Facial expressions and approximate emotional cues

A unique characteristic of NVM is its adaptive resolution strategy , where resolution or frame rates may be scaled down if GPU resources become saturated. This allows NVM to handle bursts of visual complexity without exceeding latency budgets. Additionally, early-exit classifiers embedded at intermediate network layers produce results swiftly when confidence thresholds are met, skipping deeper (and more time-consuming) layers for straightforward frames. This design ensures that the average latency per frame remains low, even if occasionally a particularly complex frame needs deeper processing.

Model Architecture Highlights:

- Patch Embedding: Images are divided into fixed-size patches (e.g., 16×16). Each patch is linearly projected into an embedding, analogous to tokens in natural language.
- Multi-Head Self-Attention: The Vision Transformer includes self-attention layers that discern spatial relationships among patches. This global perspective is crucial for tasks like reading text from different regions of a frame.
- Adaptive Attention: We introduced a gating mechanism that cross-references the voice model’s partial transcription. If the student’s speech references “the diagram in

the corner," the attention mechanism re-weights relevant patches in the top-left region of the frame.

- Early-Exit Heads: At multiple transformer blocks, a lightweight classification or detection head checks if it can produce a confident result. If yes, it aborts further deep analysis—trimming the latency overhead.

Together, these architectural details facilitate real-time vision processing. When tested in isolation on a single GPU, NVM can consistently operate above 30 FPS for 1080p input, with latencies around 30–40 milliseconds per frame.

07 3.2 Multimodal Fusion Engine (MFE)

While NVM interprets the visual stream, the voice foundation model processes student speech in near real-time. Outputs from the two streams—visual descriptors and transcribed text—are then combined within the Multimodal Fusion Engine. MFE effectively merges these separate inferences, generating a singular context that can drive the AI tutor's behavior.

Key Functions of the MFE:

- Temporal Alignment: Audio is continuous; video is discrete. The MFE tags each input with a timestamp, buffering them in short intervals (e.g., 50 ms windows). It then aligns them so that relevant frames are matched to the corresponding audio excerpt.
- Cross-Modal Attention: A specialized attention mechanism ensures that references to particular objects or texts in the student's utterance link to the correct visual features. For instance, if the student says, "I don't understand this part," the MFE attempts to identify "this part" visually and highlight that region in the fused representation.
- Adaptive Modality Weighting: If the audio signal is poor (noise or unclear speech), the MFE partially discounts the voice model's input. Conversely, if the camera feed is temporarily disrupted, it relies more heavily on the voice transcript. This dynamic weighting preserves system stability under changing conditions.
- Output Representation: The MFE produces a consolidated representation of the student's state—for example: Student's question: "What does this formula mean?"

Visual context: “Formula E=mc^2 on the whiteboard, student pointing at it.” This fused context can be fed to a response generation module or further policy logic.

Internally, the MFE is itself a transformer-based model that conditions on both textual embeddings (from the voice model) and visual embeddings (from NVM). The result is a robust, context-rich feature vector capturing the “who, what, and where” of each interaction. An optional dialogue policy or a large language model (like GPT-based architectures) can then craft a response or take an action, leveraging the fused context. This design offers a blueprint for embedding real-time concurrency and synergy between large-scale vision and language models.

08 Performance Evaluation and Benchmarks

We subjected our system to a comprehensive performance evaluation, measuring end-to-end latency, throughput, and scalability across various hardware configurations. The highlight was the test on NVIDIA DGX systems , chosen for their unmatched GPU density and high-speed interconnects.

09 4.1 Testing Methodology

Latency is the total time from receiving new input (audio frame, video frame) to generating a fused representation and an AI response. We break down latency into:

- NVM Processing Time per Frame
- Voice Model Processing for Audio
- Fusion Overhead in the MFE
- Any Additional Policy/Response Time

Throughput measures how many frames per second (FPS) or how many simultaneous interaction streams the system can handle. A single stream corresponds to one student. We tested scaling from 1 to 8 parallel student interactions on a DGX A100 (8 GPUs), distributing tasks across the available GPUs.

We additionally performed tests on a Jetson Xavier NX to reflect edge deployment scenarios with limited memory and compute.

10 4.2 Quantitative Benchmarks

The benchmarks indicate:

- High GPU Utilization: On the DGX, distributing tasks (NVM, voice model, MFE) across separate GPUs yields near-linear scaling for up to 8 concurrent streams.
- Single-GPU Baseline: Achieving ~50 ms latency on a single A100 is sufficient for real-time interactions. The system can handle around 20 FPS (1 stream) at that latency, which is adequate for a single student or low concurrency environment.
- Edge Deployment Feasibility: On Jetson Xavier NX, we see a higher latency (~90 ms) but still near real-time. This demonstrates the system's adaptability, albeit with scaled-down, quantized models.
- Low Jitter: Standard deviation of latency was modest, indicating consistent, predictable performance vital for interactive educational scenarios.

11 Implementation Challenges and Solutions

Building a real-time adaptive fusion solution around large-scale foundation models surfaced numerous obstacles. We summarize the most notable challenges here, along with the strategies used to mitigate them.

12 5.1 Meeting Strict Latency Requirements

Challenge: Large-scale transformers can be computationally heavy, making sub-50-millisecond latencies difficult—especially when both video frames and audio streams are processed in parallel.

Solution: We employed model compression (pruning, quantization), early exits (for easy frames), and dynamic load management . For example, if the system detects a processing spike (multiple complex frames or a flurry of speech), it briefly scales back

the vision resolution or the voice model's beam search width. These micro-adjustments keep the processing time under tight bounds without drastically sacrificing accuracy. Pipeline parallelism on multi-GPU setups further helps by distributing tasks to specialized GPUs in real time.

13 5.2 Operating Under Resource Constraints

Challenge: Despite the convenience of DGX systems for development, many schools or educational platforms may only have modest hardware.

Solution: We created lighter variants of each foundation model—replacing standard FP16 with INT8 quantization and pruning unimportant layers. This shrinks memory usage and speeds up inference. We also gave the MFE modular logic to operate in a "vision-only" or "audio-only" mode if resources were insufficient to run both modalities simultaneously. The synergy is lost in such cases, but the system remains operational and can scale up again if additional compute resources become available.

14 5.3 Synchronization of Multimodal Data

Challenge: Aligning audio and video timelines is tricky. A typical camera runs at 30 FPS, while audio is streamed continuously. Discrepancies in arrival times can cause misalignment.

Solution: We implemented a robust timestamp-based buffering strategy within the MFE. Every chunk of audio or frame of video is time-stamped. The MFE merges only the data with the closest timestamps, allowing small delays (50–100 ms) to ensure a correct "match" of spoken words and relevant frames. This ensures that references like "this diagram" are indeed associated with the frame containing that diagram, rather than an older or a future frame.

15 5.4 Accuracy vs. Speed Trade-offs

Challenge: Any optimization to reduce latency can degrade model accuracy. Pruning and quantization might hamper the AI's ability to recognize subtle features. Similarly, skipping

deeper network layers for “easy frames” can risk missing detail.

Solution: We integrated confidence-based gating. If the model’s confidence in a certain detection or transcript is below a threshold, it proceeds to a more thorough path. This ensures performance is only reduced for straightforward tasks, preserving near-baseline accuracy for more complicated content. We also used a tiered approach for speech recognition: a fast pass for normal sentences, and a fallback to a more robust pass for words flagged as possibly misrecognized.

16 Policy Implications and Recommendations

The combination of large-scale foundation models with real-time multimodal fusion has transformative potential in education. Nonetheless, deploying such technology responsibly demands supportive policies, infrastructure, and oversight. Our key recommendations include:

- Infrastructure Investment: Government bodies and educational institutions should consider financing GPU clusters or HPC resources—whether on-premise or cloud-based—to support advanced AI. Public-private partnerships can help schools obtain the necessary hardware or establish shared data centers.
- Teacher Training and Collaboration: Educators must be trained to integrate AI tools effectively. Professional development programs can illustrate best practices, limitations, and ethical considerations of AI-driven tutoring.
- Data Privacy and Ethics: Real-time audio and video capture in a classroom requires stringent data governance. Clear guidelines on storage, anonymization, and usage of student data are imperative. Students and parents should be made aware of how AI processes their information.
- Shared AI Model Repositories: We advocate for open (or semi-open) distribution of specialized educational AI models. This fosters innovation, allowing others to customize or improve upon them while adhering to data privacy protocols.
- Evaluation and Accountability: Institutions deploying these systems should conduct regular evaluations of the AI’s effectiveness (e.g., improvements in student learning outcomes) and ensure the technology remains aligned with curriculum objectives.

Policy should mandate transparency around AI-driven decisions and an avenue for addressing errors or biases.

- Continuous Model Updates: Educational curricula evolve, and so must the AI. Policies can promote routine updates or expansions of the knowledge base, ensuring that the AI's content stays relevant and correct over time.

By enacting thoughtful policies and building supportive ecosystems, governments and educational stakeholders can harness the full potential of large-scale multimodal AI systems to revolutionize teaching and learning experiences.

17 Conclusion

This research report outlined the practical steps, challenges, and solutions in building a real-time multimodal fusion system around large-scale foundation models in voice and vision. Our experiences at London INTL underscore the feasibility of achieving millisecond-scale latencies while handling complex tasks (speech recognition, visual understanding, and textual analysis) all in parallel.

We validated the system's performance using NVIDIA DGX hardware, demonstrating near-linear scaling as additional GPUs are introduced, and tested edge deployment on Jetson devices to confirm viability in constrained environments. The results indicate that the proposed approach can be adapted for a range of educational scenarios, from well-equipped computer labs to smaller schools with minimal resources. Adaptive load management, early-exit strategies, and dynamic modality weighting proved critical in maintaining low-latency, high-accuracy performance under variable or constrained computational budgets.

Looking ahead, the synergy of foundation models with real-time streaming data presents numerous exciting possibilities—advanced tutoring systems, robust telemedicine solutions, immersive AR/VR experiences, and more. As these models scale further, forging robust policy frameworks around data privacy, ethical deployment, and teacher training will be paramount. We remain optimistic that, through continued collaboration among policymakers, educators, and technologists, these intelligent systems can significantly enhance student engagement and success worldwide.

18 Appendix A: Technical Specifications

This appendix provides technical details about the core hardware, software stacks, and model configurations used in our benchmarks and development process.

- Hardware:
- NVIDIA DGX A100 Server: 8× A100 GPUs (80GB VRAM each), NVSwitch Interconnect, Dual AMD EPYC CPUs, 1 TB RAM.
- NVIDIA Jetson Xavier NX: 384-core Volta GPU, 48 Tensor Cores, 8GB LPDDR4 memory.
- Software:
- PyTorch 1.12 with Mixed Precision enabled (AMP).
- NVIDIA TensorRT 8.x for optimized inference.
- NVIDIA NeMo / Riva for speech recognition and text-to-speech acceleration.
- Custom alignment and fusion modules written in Python, employing Torch-based transformers for MFE.
- Model Configurations:
- Vision Transformer (NVM): 12–24 transformer blocks, patch size 16×16, 768 hidden dimensions, adaptive resolution down to 720p if needed.
- Voice Model: 1–2 second window streaming, Transformer-based acoustic model, advanced beam search, fallback to a robust decoding path for uncertain transcriptions.
- Fusion Engine: Cross-attention, multi-head alignment with gating for dynamic weighting. Early-exit triggers if a frame or transcript is deemed “simple.”
- Data Pipeline: Training used a mix of in-house collected educational video data (student–teacher interactions), public speech datasets, and general object detection corpora (e.g., COCO) for broad coverage. Vision pretraining leveraged a standard large-scale dataset (e.g., ImageNet or LAION subsets) before fine-tuning on educational contexts.

19 Appendix B: System Workflow Example

Consider a scenario in which a student attempts a math question, says “I’m getting 42. Am I correct?” , and holds their notebook to the camera showing “42.” The system processes the video feed through NVM, recognizes the written “42,” and detects the student’s uncertain facial expression. Simultaneously, the voice model transcribes the student’s query in real time. The MFE fuses these elements, concluding that the student is referencing the number “42” on paper and seeking confirmation. The dialogue policy then fetches a known correct solution (e.g., “The correct answer is 45”), and the voice model quickly generates a spoken explanation. This entire loop typically completes within 200–300 ms from the end of the student’s query, preserving the natural flow of conversation.

In more complex examples, the student may show multiple steps of algebra, and the camera feed changes frequently. The system might rely on the voice transcript to focus on the relevant step. If the voice model’s confidence is high about the phrase “look at step two,” the NVM applies an adaptive attention mechanism to the middle portion of the page. Early-exit heads can skip unneeded details, thereby optimizing performance. Throughout the process, the MFE ensures temporal alignment so that any reference to “step two” is matched with the correct frame region, culminating in a single coherent understanding. The fluid synergy of the voice model’s transcripts and NVM’s visual processing characterizes the system’s main advantage over unimodal solutions.

20 References

- X. Zhang, Y. Cui, C. Fu, W. Wu, Z. Li, and F. Liu, “Transtreaming: Adaptive Delay-aware Transformer for Real-time Streaming Perception,” arXiv preprint arXiv:2409.06584, 2024. [Online].
- M. Pollach, F. Schiegg, A. Knoll, “Low Latency and Low-Level Sensor Fusion for Automotive Use-Cases,” in Proc. IEEE Int. Conf. on Robotics and Automation (ICRA), 2020.
- TRG Datacenters, “NVIDIA DGX Components, Pricing, and other FAQs,” TRG Blog, 2023.

- Viso.ai, “YOLOv7: The Fastest Object Detection Algorithm (2024),” Viso AI Tech Blog, 2024.
- Reddit - r/LocalLLaMA, “Hertz-Dev: An Open-Source 8.5B Audio Model for Real-Time Conversational AI,” Aug. 2023.
- NVIDIA, “Get Started with NVIDIA Riva – Real-time Speech AI,” NVIDIA Developer Guide, 2023.
- J. K. Singh and S. K. Pandey, “Handling Simultaneous Multimodal Inputs with Multi-threaded Architecture,” Int. J. of Research and Analytical Reviews (IJRAR), vol. 6, no. 2, 2019.
- P. J. Perera, R. Karunatillake, et al., “A Joint Cross-Attention Model for Audio-Visual Fusion in Emotion Recognition,” in Proc. CVPR Workshops (ABAW), 2022.
- W. Huang, Y. Li, Y. Wang, “Empowering Adaptive Early-Exit Inference with Latency Awareness,” in Proc. AAAI Conf. Artificial Intelligence, 2023.
- U.S. Department of Education, “Artificial Intelligence (AI) and the Future of Teaching and Learning: Insights and Recommendations,” Government Report, May 2023.
- National Education Association, “Teaching in the Age of AI – Policy Statement,” NEA Publications, 2023.
- D. Erzin, L. Besacier, etc., “Multimodal Speaker Identification using Adaptive Decision Fusion,” in Proc. Workshop on Robust Methods for Speech Recognition, 2004.