Multimodal Foundation Models for Real-Time Human–AI Interaction in Intelligent Service Systems

Vinay A. Mukherjee
School of Computing, Clemson University, Clemson, SC, USA.
vinay.mukherjee@clemson.edu

Rainer Carpenter
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
rainer.carpenter@ucf.edu

Jose Warner
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
jose766@oregonstate.edu

Roy Dawson
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
roywork@buffalo.edu

Abstract

The rapid integration of multimodal foundation models into intelligent service systems has fundamentally reconfigured the landscape of real-time human–AI interaction. These models, capable of processing and generating information across text, image, audio, and video modalities, offer unprecedented opportunities for creating fluid, context-aware interfaces that respond instantly to human input. However, deploying such models in latency-sensitive service environments introduces profound architectural, infrastructural, and governance challenges. This paper presents a comprehensive systems-level analysis of multimodal foundation models as the computational backbone of real-time interaction within intelligent service systems. We examine the structural trade-offs between model size, multimodal alignment, and inference latency, and explore how modular architectures, caching strategies, and edge–cloud hybrids can balance responsiveness with representational fidelity. The discussion extends to deployment infrastructure, including distributed inference pipelines, model compression techniques, and energy sustainability concerns, as well as to issues of robustness under distributional shift and fairness across diverse user populations. Through case illustrations drawn from healthcare, autonomous mobility, and customer service domains, we highlight the heterogeneity of real-time interaction requirements and the need for domain-specific adaptation strategies. Finally, we address governance frameworks, algorithmic accountability, and policy directions that must accompany the widespread adoption of these systems. The paper argues that realizing the full potential of multimodal foundation models for real-time human–AI interaction demands a holistic approach that merges advances in model design with careful consideration of system-level constraints and societal implications.

Keywords

multimodal foundation models, real-time interaction, intelligent service systems, edge computing, model governance, human–AI interaction, latency-sensitive inference, fairness, robustness.

1. Introduction

Intelligent service systems are increasingly characterized by their ability to engage in natural, multimodal exchanges with human users. From virtual assistants that interpret speech and gestures to autonomous vehicles that fuse camera and LiDAR data with natural language instructions, the demand for real-time, context-aware interaction is accelerating. Multimodal foundation models, which are pretrained on vast corpora spanning multiple data modalities, have emerged as a powerful enabler of such systems because they can compress heterogeneous sensory inputs into shared representational spaces and generate coherent outputs across modalities [1], [2]. Yet the transition from offline research benchmarks to production-ready services that must respond within hundreds of milliseconds raises formidable challenges. The system-level decisions involved in deploying these models for real-time interaction—choices about architecture, infrastructure, compression, and governance—directly affect user experience, operational cost, and societal risk.

This paper adopts a systems engineering perspective to examine the structural trade-offs inherent in using multimodal foundation models for real-time human–AI interaction. We argue that the monolithic vision of a single, giant model serving all interaction needs is both technically and economically untenable; instead, modular and hybrid designs that exploit heterogeneity in latency tolerance, modality composition, and task complexity are required. We analyze the architectural components that underpin real-time performance, including attention mechanisms, modality fusion strategies, and output generation processes, and relate them to deployment constraints such as memory bandwidth, energy consumption, and network latency. Beyond engineering, we consider the socio-technical dimensions of fairness, robustness, and governance, as these systems increasingly mediate critical decisions in domains such as healthcare, transportation, and public service. The paper is organized as follows. Section 2 discusses the architectural foundations of multimodal foundation models, emphasizing the trade-offs between capacity and interactivity. Section 3 examines the specific requirements of real-time interaction and proposes system design patterns that balance responsiveness and quality. Section 4 investigates deployment and infrastructure considerations, including edge computing, model compression, and sustainability. Section 5 addresses robustness, fairness, and governance in the context of real-time multimodal systems. Section 6 presents cross-domain case studies that illustrate how these concepts manifest in practice. Section 7 outlines future directions and policy implications. Section 8 concludes the paper with a synthesis of key findings.

2. Architectural Foundations of Multimodal Foundation Models

Multimodal foundation models are built upon the principle of learning joint representations from multiple data types, typically by aligning encoders for each modality through contrastive or generative objectives and then coupling them with a large language model or a multimodal decoder [3], [4]. The architecture of such a model determines its ability to process streaming input and generate timely responses. A common design choice is to use separate unimodal encoders, such as a vision transformer for images and a text transformer for language, whose outputs are projected into a shared embedding space before being fed into a cross-modal fusion layer [5]. This modularity allows independent optimization of each encoder but introduces latency due to the sequential operation of the encoders and the fusion step.

Alternatively, some architectures integrate modality-specific information earlier, for instance by interleaving patches of visual and textual tokens in a single transformer, which can reduce synchronization overhead at the cost of increased memory footprint [6].

The scaling laws of these models pose a direct tension with real-time requirements. Larger models with more parameters generally achieve higher accuracy on complex multimodal tasks, but their inference latency grows superlinearly with model size, especially when the sequence length expands due to multiple input modalities [7]. For intelligent service systems that must respond in less than a second—such as conversational agents in call centers or augmented reality overlays—the latency budget is extremely tight. Consequently, system architects must often sacrifice some representational fidelity for speed by employing distilled or quantized versions of larger models, or by partitioning the model across the cloud and edge to exploit local computational resources [8]. The choice of attention mechanism also matters: standard full self-attention is quadratic in sequence length, whereas linear approximations or sparse attention patterns can reduce complexity and accelerate inference for multimodal streams that incorporate long video or audio sequences [9].

Another architectural consideration is the output modality. Many multimodal foundation models are designed to generate text only, but real-time interaction often requires synthesizing images, speech, or control commands. Text generation is inherently autoregressive and sequential, creating a fundamental latency floor that is hard to compress. For applications that demand immediate visual feedback, such as a robot responding to a human gesture, the model may need to output a low-dimensional action vector rather than a full image, which shifts the burden to the controller tier. Thus, the architecture must be tailored not only to the input modalities but also to the output format and the permissible delay. In summary, the architectural decisions for multimodal foundation models in real-time service systems are governed by a multidimensional trade-off space involving model capacity, modality count, sequence length, output type, and acceptable latency, all of which must be resolved through careful system design.

3. Real-Time Interaction Requirements and System Design

Real-time human–AI interaction imposes a set of performance constraints that go beyond simple throughput. The perceived quality of an interaction depends on the end-to-end latency from user input to system response, the temporal coherence of multimodal outputs (e.g., synchronizing spoken words with gestures), and the ability to handle interruptions, topic shifts, and ambiguous inputs gracefully [10]. In intelligent service systems, these interactions often occur in mission-critical contexts where a delayed or incorrect response can have serious consequences, as in medical triage or autonomous driving. Therefore, the system design must prioritize predictability and bounded latency rather than average-case performance.

A common approach to achieving bounded latency is to decompose the multimodal inference pipeline into stages that can be pipelined, parallelized, or cached. For example, an image encoder may run on a dedicated edge accelerator while the language model processes previous text turns, allowing overlap of computation [11]. Another technique is to employ early exits or cascade architectures where a lightweight model produces an initial response while a larger model refines it if time permits. These designs introduce a trade-off between responsiveness and accuracy: the early exit may yield lower quality but meets the deadline, whereas the cascade can improve quality at the cost of potential deadline violations. The choice of which strategy to use depends on the service-level objective, which may be

expressed as a probabilistic guarantee, such as responding within 200 milliseconds in 99 percent of cases.

The system design also must manage the variable computational load induced by different input modalities. Processing a single image is generally faster than processing a full video frame sequence, and audio speech recognition can be especially heavy when accent or noise is high. Adaptive modality selection is a promising direction: the system can decide, based on past interactions or context, to skip or downsample certain modalities to meet latency constraints [12]. For instance, a virtual assistant might only process visual input when it detects a user's gaze or a gesture, otherwise relying solely on voice. Such dynamic resource allocation requires predictive models of user behavior and computational cost, which themselves must be lightweight enough not to offset the savings. Overall, the system design for real-time multimodal interaction must embrace heterogeneity and adaptivity, moving away from one-size-fits-all pipelines toward configurable, context-aware architectures.

4. Deployment and Infrastructure Considerations

Deploying multimodal foundation models at scale for real-time services necessitates a robust infrastructure that spans from centralized cloud clusters to edge devices. The computational demands of these models, especially during inference, are immense: a single forward pass of a model with hundreds of billions of parameters can consume multiple gigabytes of memory and require specialized hardware such as tensor processing units or high-end GPUs [7], [13]. For services that serve millions of concurrent users, the cost of inference can quickly dominate operational budgets. Consequently, system architects must optimize the deployment topology to minimize cost while meeting latency targets.

A common deployment pattern is the edge–cloud hybrid, where a lightweight model runs on the user's device or a nearby edge server to handle routine, low-latency interactions, while more complex requests are forwarded to larger models in the cloud [14]. This split reduces the average latency and network load but introduces complexity in model synchronization, data privacy, and consistency between the edge and cloud versions. For instance, if the edge model is a distilled version of the cloud model, the two may diverge in their predictions, leading to user confusion when the same input yields different responses. Techniques such as federated learning can help keep edge models up-to-date without transmitting raw user data, but they add overhead and may not converge quickly enough for fast-changing interaction patterns [15].

Model compression is another linchpin of real-time deployment. Pruning, quantization (e.g., from 32-bit floats to 8-bit integers), and knowledge distillation can reduce model size and inference latency by several factors while retaining most of the accuracy [8], [16]. However, aggressive compression can amplify biases present in the training data or cause fragile behavior on out-of-distribution inputs, which is especially dangerous in safety-critical service systems. Therefore, validation pipelines that test compressed models on diverse interaction scenarios are essential before deployment. Energy sustainability also becomes a concern as the number of inference requests grows. The carbon footprint of serving large multimodal models is non-trivial, and service providers are increasingly pressured to report and reduce their environmental impact [17]. Techniques such as model sparsity, dynamic voltage and frequency scaling, and using renewable energy sources for data centers are actively explored but remain areas of ongoing research.

5. Robustness, Fairness, and Governance

Multimodal foundation models inherit robustness and fairness challenges from their unimodal predecessors, but these challenges are compounded by the complex interactions between modalities. A model may perform well on images with clean text overlays but fail when the image is occluded or the speech is whispered, leading to inconsistent behavior across interaction contexts [18]. Real-time systems are especially vulnerable to distributional shift: a virtual assistant trained on controlled laboratory data may degrade severely when deployed in a noisy cafeteria or a moving vehicle. Ensuring robustness requires adversarial testing across diverse environmental conditions and the incorporation of uncertainty estimation mechanisms that allow the system to gracefully degrade or fall back to simpler modes when confidence is low [19].

Fairness in multimodal interaction is a multifaceted issue. Models may exhibit biases that disproportionately affect certain demographic groups based on accent, dialect, skin tone, or cultural gestures. For example, speech recognition components may have higher error rates for non-native speakers, while vision encoders may misclassify individuals with certain facial features [20]. When these biases are amplified by the multimodal fusion process, the resulting interaction experience can be systematically inferior for marginalized populations. Addressing fairness demands not only diverse training data but also careful auditing of the entire pipeline at the intersection of modalities. Moreover, real-time systems often lack the transparency needed for users to understand why an erroneous response occurred, making it difficult to hold providers accountable.

Governance frameworks for multimodal foundation models in real-time services are still nascent. Existing regulations such as the European Union's AI Act classify high-risk applications and require human oversight, but applying these requirements to systems that operate at millisecond timescales is challenging [21]. Developers must implement mechanisms for logging and post-hoc analysis of interactions, as well as real-time intervention capabilities for safety-critical services. The allocation of responsibility—whether the model provider, the service integrator, or the end user bears liability for harms—remains an open legal question. As these systems become more embedded in everyday life, policy makers will need to establish standards for latency, accuracy, and fairness that are enforceable and adaptive to technological progress.

6. Case Studies and Cross-Domain Analysis

The diversity of intelligent service systems reveals that real-time multimodal interaction requirements are far from uniform. In healthcare, a diagnostic support system that analyzes medical images and patient history to advise clinicians must balance thoroughness with speed: a delayed recommendation during a surgery could be fatal, yet a premature, inaccurate suggestion could lead to misdiagnosis. Here, cascade architectures with early exits are often used, where a fast screening model flags critical cases for immediate attention while a deeper model provides detailed analysis for less urgent cases [22]. Modality fusion is particularly delicate because imaging data (e.g., X-rays, MRIs) must be aligned with textual clinical notes and sometimes audio recordings of patient interviews. The system must maintain high accuracy across all modalities while respecting the clinician's workflow, which requires tight integration with existing hospital information systems.

In autonomous mobility, multimodal foundation models process streams from cameras, LiDAR, radar, and microphones to perceive the environment and interact with passengers or pedestrians. Real-time constraints are extremely tight, on the order of tens of milliseconds for control loops, yet the model must also understand natural language commands from

passengers, such as requesting a detour [23]. This dual requirement has led to the development of shared representations that can serve both perception and interaction tasks, reducing the need for separate pipelines. However, the safety-critical nature of driving demands high robustness to sensor failure and adversarial attacks, necessitating redundancy and model ensembling that can increase computational cost. Edge computing is essential here because reliable cloud connectivity cannot be assumed.

In customer service, chatbots and virtual agents must handle millions of concurrent conversations across text, voice, and, increasingly, video. Latency tolerance is somewhat higher (a few seconds) but the cost pressure is immense, driving adoption of compressed models and aggressive caching of common interaction patterns [24]. Multimodal capabilities allow agents to analyze images sent by customers (e.g., a photo of a damaged product) and to synthesize empathetic speech with appropriate tone. Fairness issues arise from dialect and accent biases in speech recognition, and from visual biases in product recognition. Service providers are investing in continual learning and human-in-the-loop auditing to mitigate these risks. Cross-domain comparison shows that the degree of latency criticality, the number of modalities, and the acceptability of errors vary significantly, implying that no single deployment architecture can serve all intelligent service systems.

7. Future Directions and Policy Implications

Looking ahead, several trends will shape the evolution of multimodal foundation models for real-time human–AI interaction. The development of more efficient architectures, such as mixture-of-experts models and linear attention mechanisms, promises to reduce the latency gap between large models and small ones [25]. Simultaneously, advances in hardware, including neuromorphic chips and in-memory computing, could enable real-time inference on power-constrained devices. However, these innovations must be paired with software frameworks that support dynamic model selection and resource orchestration across heterogeneous compute nodes.

Policy and governance will need to evolve in tandem. The growing reliance on real-time multimodal AI in services like emergency response, public transit, and education raises questions about accessibility, digital divide, and data sovereignty. For instance, edge deployment can improve latency but may limit the ability to audit models centrally, potentially making biased decisions harder to detect. Regulators may require that critical services maintain a human-in-the-loop capability, even when the AI can respond faster than a human. Standards for evaluating real-time multimodal systems—covering latency, accuracy, fairness, and interpretability—should be developed through multi-stakeholder processes. Finally, the environmental cost of serving large models must be factored into policy incentives, encouraging the use of green computing practices and the development of benchmark suites that include energy consumption as a key metric.

8. Conclusion

Multimodal foundation models offer a transformative capability for real-time human–AI interaction in intelligent service systems, but their deployment is fraught with system-level complexities that span architecture, infrastructure, robustness, fairness, and governance. This paper has provided a holistic analysis of these challenges, emphasizing the trade-offs between model capacity and interaction latency, the need for adaptable deployment topologies, and the importance of addressing biases and safety concerns in multimodal contexts. We have shown through case studies that domain-specific requirements dictate the choice of pipeline design,

compression strategies, and human intervention mechanisms. As these systems become more pervasive, an interdisciplinary approach that integrates systems engineering, AI research, ethics, and policy is essential to ensure that real-time multimodal AI serves human needs equitably and reliably. Future work should focus on building standardized evaluation frameworks and scalable governance mechanisms that can keep pace with rapid technological change.

References

1. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

2. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning, 139, 8748–8763.

3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations.

4. Tsimpoukelli, M., Menick, J., Cabi, S., Eslami, S. M. A., Vinyals, O., & Hill, F. (2021). Multimodal few-shot learning with frozen language models. Advances in Neural Information Processing Systems, 34, 200–212.

5. Li, J., Li, D., Savarese, S., & Hoi, S. C. H. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. Proceedings of the 40th International Conference on Machine Learning, 202, 19730–19742.

6. Jaegle, A., Borgeaud, S., Alayrac, J.-B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., Henaff, O. J., Botvinick, M., Vinyals, O., Zisserman, A., & Carreira, J. (2022). Perceiver IO: A general architecture for structured inputs & outputs. International Conference on Learning Representations.

7. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

8. Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. International Conference on Learning Representations.

9. Katharopoulos, A., Vyas, A., Pappas, N., & Fleuret, F. (2020). Transformers are RNNs: Fast autoregressive transformers with linear attention. Proceedings of the 37th International Conference on Machine Learning, 119, 5156–5165.

10. Huang, C.-M., & Mutlu, B. (2016). Anticipatory robot control for efficient human-robot collaboration. Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction, 83–90.

11. Crankshaw, D., Wang, X., Zhou, G., Franklin, M. J., Gonzalez, J. E., & Stoica, I. (2017). Clipper: A low-latency online prediction serving system. Proceedings of the 14th USENIX Symposium on Networked Systems Design and Implementation, 613–627.

12. Ren, Z., Yeh, M. C., & Schwing, A. G. (2020). Adaptive inference for video recognition. European Conference on Computer Vision, 12355, 153–169.

13. Jouppi, N. P., Yoon, D. H., Kurian, G., Li, S., Patil, N., Laudon, J., Young, C., & Patterson, D. (2021). A domain-specific supercomputer for training deep neural networks. Communications of the ACM, 64(7), 67–78.

14. Satyanarayanan, M. (2017). The emergence of edge computing. Computer, 50(1), 30–39.

15. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 54, 1273–1282.

16. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

17. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650.

18. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. Proceedings of the 35th International Conference on Machine Learning, 80, 60–69.

19. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems, 30, 6402–6413.

20. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 81, 77–91.

21. European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM(2021) 206 final.

22. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.

23. Bansal, M., Krizhevsky, A., & Ogale, A. (2019). ChauffeurNet: Learning to drive by imitating the best and synthesizing the worst. Robotics: Science and Systems.

24. Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. Mobile Networks and Applications, 19(2), 171–209.

25. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q. V., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. International Conference on Learning Representations.