10.48047/jocaaa.2026.35.01.79

Multimodal Conversational AI: From Perceptual Fusion to Collaborative Intelligence

Panneer Selvam Viswanathan
Tech Mahindra Americas Inc, USA

Abstract

These multimodal conversational systems are defined as systems that allow for natural language processing alongside visual, audio, and structured data on a shared computing architecture. Most multimodal conversational systems use transformer architectures that incorporate modality-specific encoders and shared representational layers, while incorporating methods for cross-modal reasoning, attention, and the alignment of textual information with visual regions, acoustic samples, and structured semantic representations. These systems are supported by tool use, function calling, multi-step planning, and long context windows that allow for extended interaction across multiple steps of complex workflows. They have been applied to a variety of tasks, including software engineering, data analysis, education, healthcare, and creative endeavors. The systems could be viewed as integrated software development partners, analytical assistants, tutors, clinical assistants, and creativity assistants, respectively. There are also a number of technical challenges, such as for latency reduction and uncertainty estimation, context acquisition and modeling, safety across modalities, debiasing and calibration, robustness to out-of-distribution observations and adversarial and safety-critical environments, represented agents, continual learning, causal reasoning, and multi-agent systems. Whether we are able to achieve fairness, privacy, environmental sustainability, and alignment with human values at every stage of the data lifecycle will determine the future direction.

Keywords: Multimodal Conversational AI, Transformer Architectures, Agentic Capabilities, CrossModal Reasoning, Collaborative Intelligence

1. Introduction

Multimodal conversational AI systems represent a recent trend in artificial intelligence, transcending the customary domains of natural language processing, computer vision, and acoustic modeling. Canonical architectures for these multimodal conversational agents take a multimodal approach to information processing, using shared computational substrates that allow representations for text, image, audio, and structured representations to live in the same space. The MMMU benchmark, a set of multi-task benchmarks for large multi-discipline multimodal understanding, shows that college-level knowledge across a variety of topics can now be handled via multi-task modeling and integrated reasoning across both text and visual modalities. Critical reasoning in art and design disciplines, business and finance reasoning, health and medicine reasoning, and scientific reasoning are now possible, albeit limitedly, in a multimodal fashion. This has allowed computers to approach the meaning of information as humans do, where meaning is not based only on each modality separately.

Furthermore, modern multimodal systems have extended beyond simply retrieving information in the form of question-answering systems to also functioning as autonomous agents, being able to reason, plan, and execute tasks, including tool invocation, code generation, and multi-step workflows. Such systems can

10.48047/jocaaa.2026.35.01.79

automatically decompose complex tasks into subtasks, allocate resources, and update their policies based on intermediate results. When used as benchmarking environments for agents in realistic web settings, these tasks have led to both capabilities and challenges in quantifying system capabilities in producing human-like goal-directed behavior involving multi-stage contextual reasoning, navigation in a dynamic web interface, and interacting with the real world via web apps [2]. This agency is an important leap where the human-agent interaction is no longer limited to a transaction-based querying interaction but a long-term collaborative interaction over many interactive steps over extended time periods involving iterative prompting, shared state maintenance and management, concrete goal-directed problem-solving, and long-horizon decision making.

The advent of native multimodal understanding and agentic capabilities has led to a wide range of applications in software engineering, data analysis, education, healthcare, and creative domains, broadly changing the role of artificial intelligence from that of autonomous general intelligence to collaborative intelligence that augments human expertise on cognitively demanding tasks.

Benchmark	Evaluation Focus	Key Characteristics	Performance Indicators
MMMU	College-level multimodal reasoning	Multi-disciplinary knowledge integration across text and visual information	Capabilities in art analysis, business problem-solving, health interpretation, and scientific reasoning
WebArena	Autonomous web interaction	Multi-step task completion with dynamic interface navigation	Goal-directed activities requiring planning, execution, and adaptation

Table 1: Benchmark Evaluation Frameworks for Multimodal Understanding [1,2]

2. Architectural Foundations of Native Multimodal Processing

The underlying technical frameworks for multimodal conversational AI rely mostly on transformer-based networks, which process each modality via encoders feeding into a shared representation. Vision encoders decompose images or videos into patch-level or region-level representations containing spatial and semantic information. Likewise, speech encoders convert raw acoustic features into segmental-level representations, which contain phonetic and prosodic information. ViTs consider an image as a sequence of patches, with the image being split into fixed-size patches and being fed into a linear projection to obtain a latent embedding. Then, attention blocks with positional encodings are stacked on top of those embeddings. The ViTs achieve competitive accuracy with CNNs while requiring less abundant pretraining and being easier to scale to large datasets and models [3]. The patch-based representation allows for ease of integration with language transformers, since both vision and language transformers accept sequence representations that can be attended over by the same attention blocks.

To reduce the high computational cost of pre-training and the subsequent use of large-scale multimodal models in downstream settings, current literature has proposed mixture-of-experts and modular architectures with router networks that dynamically route subgroups of input tokens to specific group(s) of specialized expert modules. The Mixtral model family, for instance, has demonstrated that the sparse mixture-of-experts design model can simultaneously achieve strong performance on many different tasks while reducing the cost of inference by only activating the most relevant experts. This is achieved through a learned gating mechanism, where a different set of experts is chosen for different types of inputs or

10.48047/jocaaa.2026.35.01.79

reasoning tasks [4]. This sparsely activated architecture allows for meaningful expansion of model capacity without increasing inference time, possible specialization of experts for modality- or domainspecific processing, and interpretability through knowledge of which computational pathways are contributing to a particular output. Expertise might map to modalities (as with vision experts for models of image understanding, code experts for programming tasks, and mathematical reasoning experts for quantitative reasoning tasks of a variety of difficulties), as well as to programming languages, scientific domains, or stylistic registers.

A key trend is end-to-end multimodal processing without brittle intermediate symbolic representations. Prior work often decomposed multimodal tasks as a sequence of pipeline stages: optical character recognition (OCR), NLP, and then structured reasoning. The output of each stage would be the input to the next, with errors in one stage propagating downstream. Unlike earlier structures that use symbolic intermediate representation, present architectures use distributed representations for all stages of computation, enabling the model to reason about the relationship between visual elements, textual elements, and their semantics. A native multimodal system, shown a single image of a whiteboard with mathematical equations, diagrams, and human-written notes, can directly reason about the mathematical structure, spatial relations, and semantic content by using an attention mechanism to map visual regions to mathematical concepts and logical relations.

Speech processing shows the value of native architectures especially clearly, since customary speech systems have commonly been based on cascaded architectures with independent data structures for each model. In contrast, native speech-to-speech architectures allow the same representation (with lexical, prosodic, and speaker state information) to be updated for both input and output, allowing the system's response to inherit emotion, speaking style, and conversational context. This design thus supports a more natural experience of interaction, which can include interruptions, repairs, and other paralinguistic signals that carry meaningful information beyond the literal semantic content of interaction.

Due to the real-time demands of many applications, streaming and incremental processing have been studied extensively. In some modern multimodal systems, it is already possible to produce output based on streamed input before the entire input is received. For text and speech processing, incremental decoding allows for early completions and graceful handling of mid-turn interruptions and disfluencies. For video streams and other live input, models process windows of frames, carrying along internal states to model temporal continuity. Several algorithmic techniques are used: efficient attention for long contexts, chunking in time by propagating recurrent states, and speculative decoding, which samples and scores potential continuations in parallel to assess likelihoods. These characteristics, together with system-level optimizations such as hardware acceleration, model quantization, and caching, contribute to conversational agents reaching latencies on the order of human conversations.

Architecture	Processing Method	Efficiency Mechanism	Integration Approach
Vision Transformer	Patch-based sequence processing	Fixed-size patch embedding with positional encoding	Seamless integration with language transformers through unified attention
Mixtral (Mixture-of-Experts)	Sparse expert activation	Router-based token dispatch to specialized modules	Modality-specific and domainspecific expert networks

Table 2: Architectural Innovations in Multimodal Processing [3,4]

10.48047/jocaaa.2026.35.01.79

3. Agentic Capabilities: From Response Generation to Autonomous Collaboration

Recent work has focused on converting multimodal systems from perceptual engines to agents, mainly through the mechanisms of tool use and function calling. To this end, modern multimodal models have been provided with extensible sets of computational tools described by formal schemas. The ToolLLM framework demonstrated that LLMs can effectively leverage large sets of real-world APIs through instruction tuning and decision making, learning to read API documentation, plan multi-step function call sequences, and incorporate the responses into their output. Furthermore, this line of research demonstrated that models can generalize their tool usage and improve their performance on unseen kinds of APIs and new tools not observed in training datasets [5]. Thus, systems can build complex multimodal workflows integrating perception, information retrieval, computation, and synthesis, sequentially using visual understanding to extract relevant information from images, appropriately transforming the images into search queries to retrieve external information, and forming thorough responses.

In addition to their use of individual tools, agentic systems are capable of reasoning and planning to accomplish complex goal-directed tasks by breaking high-level goals into a sequence of logical subtasks in a particular order. The GAIA benchmark includes evaluation protocols for general in-the-wild AI assistants that require reasoning and tool use applications such as web browsing to follow up on questions that require multiple steps of gathering and synthesizing information from different places, revealing promising capabilities appear in simple tasks and that there are large gaps on long-horizon planning, correcting mistakes, and handling information from diverse sources and formats [6]. These systems can work with complex plans, identifying a task from multimodal input, for instance, then generating a sequence of actions to first retrieve data, apply actions and checks, and finally convert the result to the target format. Multimodal planning is closely tied to perceptual processing. Systems might plan actions to acquire further context from diagrams, determine from visual or acoustic input whether further disambiguation is required, or recognize from intermediate outcomes that the intended plan is not applicable.

To achieve safety-strong deployments, an agentic system needs automatic error recovery. When a tool invocation fails, the system extracts the relevant error messages, infers their likely causes, tries to reason about how to correct the situation, and resubmits the invocation with the corrected parameters. When errors occur at runtime, the system can analyze stack traces, compare implementations of a function to the specification or to a reference run to detect likely errors, and generate suitable patches. Perceptually, when multimodal input is ambiguous, contradictory, or of insufficient quality, strong agents can selfcritically output their own uncertainty, ask for information, or output multiple plausible responses to avoid biased decisions that could lead to errors. This self-monitoring and recovery behavior reduces the occurrence of silent failures. Also, it is in line with human problem-solving processes such as reflection, adjusting hypotheses, and seeking help when things go wrong.

Long memory, context windows with large capacity, and more advanced memory mechanisms enable sustained collaborations (across long conversations, multi-artifact environments, and over complex task states). Modern agentic systems maintain a mutable persistence that keeps track of previous goals, decisions, artifacts, and relevant information across evolving states. Agentic systems process a stream of user messages not as a series of question-and-answer pairs, but as project partners that remember previously expressed constraints and preferences, and incrementally contribute to the completion of longrange activities like programming, document writing, or scientific experiment analysis. Some proposed architectures augment the memory of an agent with other knowledge and state representations, for example, knowledge graphs of entities and relations, task lists with states and dependencies, and collections of

10.48047/jocaaa.2026.35.01.79

documents and other artifacts with version histories. Such representations may allow retrieval of more specific information, reduction in the frequency of re-encoding large context windows, more effective reasoning with structured or hierarchical data, and even retention of state among very long (days or weeks) conversations.

Framework	Core Capability	Functional Scope	Challenge Areas
ToolLLM	API mastery and tool orchestration	Multi-step tool usage with generalization to novel APIs	Integration of heterogeneous information sources
GAIA	General assistant evaluation	Real-world question answering with tool use and web browsing	Long-horizon planning and error recovery mechanisms

Table 3: Agentic System Capabilities and Evaluation [5,6]

4. Applications Across Cognitive Domains

The combination of native multimodal tasks with agentic capabilities has enabled a diverse range of applications, including multimodal conversational agents in software engineering, where they can serve as all-in-one software collaborators for developers. Developers provide source code, documentation, screenshots of error messages with stack traces, design mockups, etc., and multimodal conversational agents can refactor, improve, perform framework migrations, and implement new features, providing developers with rapid feedback through code execution and reasoning. These effects can vary widely depending on the tasks' difficulty levels and the developer's skill levels. The models may be further improved by interacting with a user, as they tend to learn project conventions, architectural patterns, and infrastructure limitations over time, becoming more useful through contextualized knowledge of codebases, development workflows, and team preferences.

Data analysis and business intelligence are another area of multimodal agentic systems. Business and analyst users frequently consume heterogeneous data from spreadsheets, database exports, PDF reports, and JPEG images of data visualizations, and they specify a range of queries, such as trend detection, anomaly detection, predictive modeling, and what-if analysis. According to the latest McKinsey State of AI report, surveyed organizations say they have adopted generative AI across a wide range of business functions, and many report substantial productivity gains from generative AI tools for data analysis and perception generation. Successful applications typically combine several AI components and methods, including natural language processing (NLP), computer vision, and automated reasoning, allowing for end-to-end processes from data ingestion to perception delivery [8]. The combination of perception, computation, and generation allows users to iterate, visualize, and analyze data using natural language instructions, explore various statistical assumptions, and produce a combined output in the form of charts, tables, and textual interpretations that are suitable for publication. Since all three tasks are part of the same system and solved by the same system, decisions and their supporting numerical findings can be contextualized, allowing for retrospective assessment and potential scenario exploration and sensitivity analysis across models, variables, and assumptions.

When used in educational contexts, multimodal agents can observe multiple learner artifacts and adaptively regulate instruction according to the student's learning path. The agent can receive studentcreated artifacts like text-based assignments, digitized handwritten solutions and concept maps, or videos or audios of real-world experiments, and provide feedback, hints, or alternative explanations based on the difficulties observed. The system adapts the presentation modality to the learner's explicit preferences or implicitly

10.48047/jocaaa.2026.35.01.79

predicted most effective pedagogical modality (text, speech, static diagram, interactive simulation, video). Agentic capabilities adaptively generate the curriculum by tracking a student's mastery of learning objectives across learning sessions, detecting ongoing incorrect knowledge states through error and hesitation pattern analyses, and prompting practice items in these areas. In addition, this allows students to receive a customized learning path while still achieving curriculum/learning objectives.

Examples include structuring information from different sources of heterogeneous patient data into a clinical summary to support differential diagnosis and care planning. Systems could provide lay explanations that combine visualizations of the diagnosis and treatment options, and that adapt the content and complexity depending on the patient's health literacy. For complex cases, multimodal agents can assist clinicians with reviewing the literature and guidelines and with evidence aggregation. Recommendations include curating reason chains, explicitly documenting uncertainty and confidence, distinguishing evidence-based recommendations from inference-based conjectures, and complying with institutional standards, practices, and regulatory requirements in AI-improved medical systems. AI systems can only be considered for clinical use when safety protocols and processes are enacted and enforced, and when thorough evaluation protocols are in place and maintained.

In creative fields and design, multimodal agents can be employed as ideation partners for the design process or as tools for production. Designers can use multimodal systems to help create mood boards, storyboards, rough compositions, and typographic layouts from a text prompt and/or reference material. Agents can review large data sets from a movement, a brand, or a single designer to detect stylistic patterns such as color, composition, or typography and then suggest new variations aligned with these patterns. Multimodal agents allow for the concurrent modeling of visual generation with narrative reasoning and concept development. This allows the copy, visuals, and interaction patterns to be developed iteratively in an integrated fashion rather than in serial or parallel with different specialists. It allows for rapid prototyping in which different design spaces can be explored before committing resources to the production process.

Application Domain	Implementation Context	Productivity Impact	Integration Features
Software Engineering	Code generation and debugging assistance	Reduced task completion time with improved code quality	Mixed-modality inputs, including code, documentation, and error traces
Business Intelligence	Data analysis and insight generation	Enhanced analytical workflows with faster insight delivery	Natural language processing combined with visual reasoning

Table 4: Domain-Specific Applications and Impact [7,8]

5. Technical Challenges and Research Frontiers

Despite promising progress, many technical difficulties remain to realize multimodal agentic systems that can reason over and respond in real time to high-resolution visual inputs, extended contexts, streaming speech, and perform tool usage. These practical multimodal systems must address the trade-off between model size, architectural sparsity, hardware accelerators, and real-time caching of prompts, tokens, and tool outputs. The inference efficiency of large language models is an active area of research. Several quantization, pruning, knowledge distillation, and model architecture methods can reduce the resource requirements of large language models with minimal loss in quality [9]. Other performance optimizations that focus on reducing the memory bandwidth requirements for transformers include grouped-query

10.48047/jocaaa.2026.35.01.79

attention, repeated speculative decoding of multiple candidate tokens, and attention mechanisms with sublinear time complexity for longer sequences. These improvements enable large language models to process longer context sequences of text and perform other tasks with multimodal data without exceeding the desired amount of latency. This is particularly useful when deploying models on low-computation devices such as mobile phones or edge cloud computing platforms.

A core challenge of multimodal perception in the real world is that the different modalities tend to contain ambiguity, uncertainty, and distribution shift, namely, visual scenes that are low-light, occluded, and taken from unseen viewpoints, and audio signals that are subject to background noise, channel distortion, and speaker overlap, as described by empirical observations. Calibration studies of modern neural networks have observed instances where model confidence is not always well-calibrated. This means that predicted confidence scores do not reflect the model's actual predictive performance: for example, confidence scores often correlate poorly with prediction accuracy on out-of-distribution samples, or samples where distribution shift occurs between the training and deployment environments. In safetycritical applications where system confidence feeds into downstream decisions, uncalibrated high confidence also risks misleading users when the model outputs are incorrect. Strong systems can autonomously identify conditions in which they cannot fully trust their output, and do not behave overconfidently under the pretense of being entirely correct (which would mislead the user and/or cause downstream failure). Best practices are becoming to expose uncertainty to users via alternative hypotheses, saliency markings on low-confidence regions of images or transcripts, and targeted clarifying questions.

Context and memory management pose both theoretical and engineering challenges. For example, naively concatenating all potentially relevant history together, even with an extremely large context window of hundreds of thousands of tokens, quickly becomes intractable and may lead to worse performance as important information is lost amongst the details. The best approaches use a cascade of techniques such as hierarchical summarization that compress away earlier conversation while retaining key entities, decisions, and constraints, external memory such as document or code repositories with retrieval in response to relevance signals, and selective contextual composition of past information via semantic similarity or explicit user reference. They often form a hybrid of the language modeling and information retrieval models, slightly inverting the usual usage of the terms, yielding knowledge-driven agents rather than stateless text generators. This raises open questions such as how to cover as much relevant information as possible while maintaining high precision, a good granularity, and coherence in long dialogues.

Safety in multimodal systems can be challenging due to the threat models, which may include attacks against multimodal interaction, as well as attacks on visual models that try to make them produce harmful, false, or infringing content, and in multimodal prompt-based jailbreaks and instruction following attacks. Multimodal safety constraints require the same policy to be applied to channels uniformly. For example, a command that would be unsafe in text form cannot be satisfied by transforming the command to a different modality, like an image or code. Some cross-modal attacks to supersede prevention mechanisms provide technical guarantees based on multimodal classifiers for adversarial examples, crossmodal consistency checks, and shared safety-aware representations. Some attacks provide procedural defenses like red-teaming exercises to systematically discover weaknesses in modality combinations, cross-modality interaction patterns, and specific compositions of modalities within a given context.

Factors that further complicate bias and fairness in multimodality include demographic disparities in performance (eg, facial recognition across people with different skin tones; pose estimation across people with different body shapes and sizes; accent recognition across speakers of different languages) or imbalances in demographic representation within generated outputs. Multimodal models can also unintentionally infer protected attributes from other modalities, such as visual or acoustic attributes, even

10.48047/jocaaa.2026.35.01.79

in cases where this information is not explicitly presented as input or even asked for as output. For instance, this can manifest in hiring, medical and health applications, and content moderation. Some methods aim to support transparency and explainability, e.g., by visualizing attention on regions of input that are important for output, to gain user understanding and accountability while diagnosing and countering bias over the whole life cycle of the system.

Conclusion and Future Trajectories

The observation of an architectural model shift from multimodal conversational AI systems leveraging discrete perceptual modules to native multimodal systems operating on the same representational substrate, enabling the emergence of agentic capabilities, including autonomous tool use, multi-step planning, code execution, and long-term memory, and allowing systems to act more like active agents than passive participants in a dialogue. These dimensions have coalesced into systems capable of participating in detailed cognitive processes in software development, data analysis, education, healthcare, and creative production, marking a qualitative advance in the role AI can play in knowledge work and creative production. Open difficulties remain in achieving acceptable latencies and throughputs, representing context and long-term memory, ensuring fairness and robustness across modalities and demographic groups, protecting privacy when processing rich sensory data, and reducing the environmental impact of scale in multimodal capabilities. Future directions of multimodal conversational agents will focus on additional modalities like haptics, communication in 3D space, biosignals, or openended augmented or virtual reality spaces, which will be particularly important for represented agents that operate in a shared, physical environment. These agents will mix learned neural representations with classical models of physics, affordances, and safety. There are also critical alignment, supervision, and corrigibility questions for environments with more powerful agents controlling the actuators. With continuous training under well-controlled privacy and safety, agents would display greater proactiveness and personalization by better understanding the individual user, the organization they represent, and the domain knowledge, raising issues, for example, in personalization, bias, and autonomy. Better causal inference, counterfactual understanding, and neuro-symbolic integration could enable more verifiable and reliable reasoning in the domains where it is essential. Multi-agent systems, especially with specialized agents orchestrated with particular protocols, could lead to complementary capabilities, task distribution, and interpretable reasoning processes. However, appropriate coordination, incentives, and safety measures for such systems have yet to be fully addressed and represent areas requiring further research. The main opportunity and responsibility of developing multimodal conversational AI systems is to augment human capacities while respecting human values, constraints, and agency. As such systems become more capable, ubiquitous, and integrated into socio-technical infrastructures in education, healthcare, the creative industries, scientific research, and governance, alignment and accountability, along with transparency and participatory design, become increasingly important. This means that technical innovation must be accompanied by governance structures, assessment methods that measure the social impact of this technology beyond its extrinsic performance, as well as collaborative co-design. Therefore, the development of collaborative intelligence through perceptual fusion is not simply an engineering challenge; it is a sociotechnical challenge. It is determined by the values, priorities, and design choices in the field. The role of multimodal agentic intelligence as a trusted collaborator for knowledge-creation, creativity, and social benefit will depend on the continued commitment to the responsible development of advanced systems that strike a reasonable balance between optimizing benefits, risks, equity, and human flourishing.

10.48047/jocaaa.2026.35.01.79

References

[1] Xiang Yue, et al., "MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI," arxiv, 2023. [Online]. Available: https://arxiv.org/abs/2311.16502 [2] Shuyan Zhou et al., "WebArena: A Realistic Web Environment for Building Autonomous Agents," arxiv, 2023. [Online]. Available: https://arxiv.org/abs/2307.13854

[3] Alexey Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," ResearchGate, Oct. 2020. [Online]. Available: https://www.researchgate.net/publication/344828174_An_Image_is_Worth_16x16_Words_Transformers_for_Image_Recognition_at_Scale

[4] Albert Q. Jiang et al., "Mixtral of Experts," arxiv, 2024. [Online]. Available: https://arxiv.org/abs/2401.04088

[5] Yujia Qin, et al., "ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs," arxiv, 2023. [Online]. Available: https://arxiv.org/abs/2307.16789

[6] Grégoire Mialon, et al., "GAIA: A Benchmark for General AI Assistants," arxiv, 2023. [Online]. Available: https://arxiv.org/abs/2311.12983

[7] Sida Peng, et al., "The Impact of AI on Developer Productivity: Evidence from GitHub Copilot," ResearchGate, 2023. [Online]. Available: https://www.researchgate.net/publication/368473822_The_Impact_of_AI_on_Developer_Productivity_Evidence_from_GitHub_Copilot

[8] Alex Singla, "The state of AI in early 2024: Gen AI adoption spikes and starts to generate value," McKinsey & Company, 2024. [Online]. Available: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-2024

[9] Shulin Zeng et al., "FlightLLM: Efficient Large Language Model Inference with a Complete Mapping Flow on FPGAs," ACM Digital Library, 2024. [Online]. Available: https://dl.acm.org/doi/10.1145/3626202.3637562

[10] Matthias Minderer, et al., "Revisiting the Calibration of Modern Neural Networks," ResearchGate, 2021. [Online]. Available: https://www.researchgate.net/publication/352430876_Revisiting_the_Calibration_of_Modern_Neural_Networks