Multitask Asynchronous Bidirectional Multimodal Agent for Personalized Treatment Companions

Muhamad Iqbal Januadi Putra
Department of Computer Science
Universitas Siber Asia
Jakarta, Indonesia
muhamad.iqbal41@sci.ui.ac.id, 2024.iqbal.januadi@student.unsia.ac.id

Raka Admiral Abdurrahman
Politeknik Negeri Malang
Malang, Indonesia
rakaadmiralabdurrahman@gmail.com

Vincent Alexander
Universitas Tarumanagara
Jakarta, Indonesia
alex.535220149@stu.untar.ac.id

Abstract

Personalized treatment requires intelligent systems that can continuously monitor patients, adapt to evolving conditions, and communicate naturally with both patients and clinicians. Existing healthcare technologies often rely on unimodal data streams (e.g., wearables or medical imaging) or offline analysis, limiting their responsiveness and interactivity. In this work, we demonstrated a Multitask Asynchronous Bidirectional Multimodal Agent powered by a multimodal large language model (MLLM) and integrated with retrieval-augmented generation (RAG) from multimodal sources, including text, images, and video. The agent combines vision (video) & audio (speech) for patient monitoring & natural interaction, supporting real-time personalized treatment. We define three representative asynchronous tasks: (i) vision-based patient monitoring for mobility, posture, and facial cues, (ii) aggregation of health metrics for adaptive treatment planning, and (iii) speech-based dialogue to engage patients and support clinician decision-making. Our architecture integrates Gemini's multimodal reasoning with a WebSocket-based backend for bidirectional streaming interaction, enabling both proactive alerts and conversational explanations. Evaluation on simulated healthcare monitoring datasets demonstrates improved accuracy in patient state recognition, reduced latency in adaptive feedback, and enhanced interpretability compared to unimodal baselines. This work highlights the potential of multimodal agents to act as personalized treatment companions, advancing adaptive, human-centered healthcare. Evaluation in simulated healthcare communication scenarios shows strong performance with a Usefulness Metric of 0.78, a Relevance Metric of 0.93, a Hallucination Metric of 0.3, a Contain Metric of 0.88, an Equals Metric of 0.88, and a Sentence BLEU score of 0.98. These results highlight the potential of multimodal agents to act as personalized treatment companions, advancing adaptive, human-centered healthcare.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences.

1 Introduction

Personalized treatment is central to modern healthcare, aiming to adapt therapy and monitoring to each individual's needs. Current practice relies heavily on retrospective data and periodic check-ups, creating gaps in continuous care. Patients struggle to communicate subtle symptoms, while clinicians face fragmented information from siloed monitoring systems. These challenges motivate *real-time, multimodal* systems that can observe, interpret, and interact seamlessly.

Recent advances in multimodal foundation models enable reasoning across video, audio, and text while supporting natural dialogue. We argue that combining continuous video monitoring and bidirectional speech with multimodal RAG over biomedical knowledge provides a practical route to personalized treatment companions. Unlike static monitoring tools, our agent is designed for interactive and proactive support—both reacting to queries and initiating alerts/recommendations.

Our contributions are as follows: (1) A bidirectional streaming architecture integrating video, speech, and RAG-based multimodal reasoning; (2) a multitask framework for patient-state recognition, asynchronous health-data aggregation, and natural language interaction; (3) a clinical multimodal RAG corpus and retrieval pipeline; and (4) an empirical study on simulated personalized-treatment scenarios demonstrating high relevance and low hallucination.

Our design emphasizes plan–act–reflect cycles adapted from streaming agent pipelines [6], video-centric multimodal RAG grounding [5], clinician-facing explainability and workflow fit [1, 7], and a path toward efficient deployment via heterogeneous agent knowledge transfer [2].

2 Related Work

Streaming agent pipelines Screen-centric agents formalize planning–acting–reflecting loops for continuous control [6]. Voice assistants have been introduced for patient support, but they remain limited without visual grounding. They cannot capture non-verbal signals such as posture, mobility, or facial cues. We adapt this paradigm to healthcare streaming by prioritizing interruptibility, alerts, and safety escalation over GUI task completion.

Multimodal RAG Multi-RAG demonstrates how converting video/audio/documents into unified textual representations supports retrieval and robust video understanding [5]. We adopt a similar backbone to ground patient-state inferences and recommendations.

Clinical multimodal assistants and HAI. Recent developments in multimodal foundation models and retrieval-augmented generation enable joint reasoning across text, audio, and vision while grounding outputs in external medical knowledge. These approaches reduce hallucinations and increase interpretability. CardioAI integrates wearables and conversational logs to surface explainable, clinician-oriented risk summaries [7]. Human–agent interaction guidance emphasises transparency, privacy, and trust, which we encode via rationales, retrieved snippets, and clinician overrides [1].

Heterogeneous agent knowledge transfer. TransAgent distills complementary expertise from diverse agents into a deployable vision–language model [2], suggesting a path to shrink our runtime cost while preserving accuracy.

3 Method

3.1 System Architecture

The proposed system (Figure 1) comprises three main modules: a frontend, a backend, and a multimodal reasoning engine. The frontend provides a web-based interface where operators can stream live video and interact with the agent through natural queries. The backend employs a WebSocket server that manages real-time bidirectional communication, asynchronous task queues, and scheduling. The multimodal reasoning engine leverages Gemini to process video frames, transcribed speech, and RAG outputs from biomedical knowledge bases. This design follows recent agentic patterns where multimodal agent models perform perception, tool use, and interactive dialogue [4, 6].

3.2 Multimodal RAG Corpus Construction

We construct a *clinical* multimodal corpus comprising: (i) de-identified clinical guidelines and patient-education materials; (ii) structured EHR-style summaries (problem lists, medications); (iii) exemplar medical images & video (e.g., radiographs, ultrasound frames) with captions; and (iv) conversational transcripts. Text is normalized and chunked; embeddings are computed with multilingual encoders. Visual frames are captioned and/or embedded (e.g., CLIP-like) to enable cross-modal retrieval. We store text and visual embeddings in a vector database for low-latency indexing and retrieval. Our video-centric RAG strategy follows the spirit of Mao et al. [5], converting frames and audio snippets into retrievable units that support grounding and explanation.

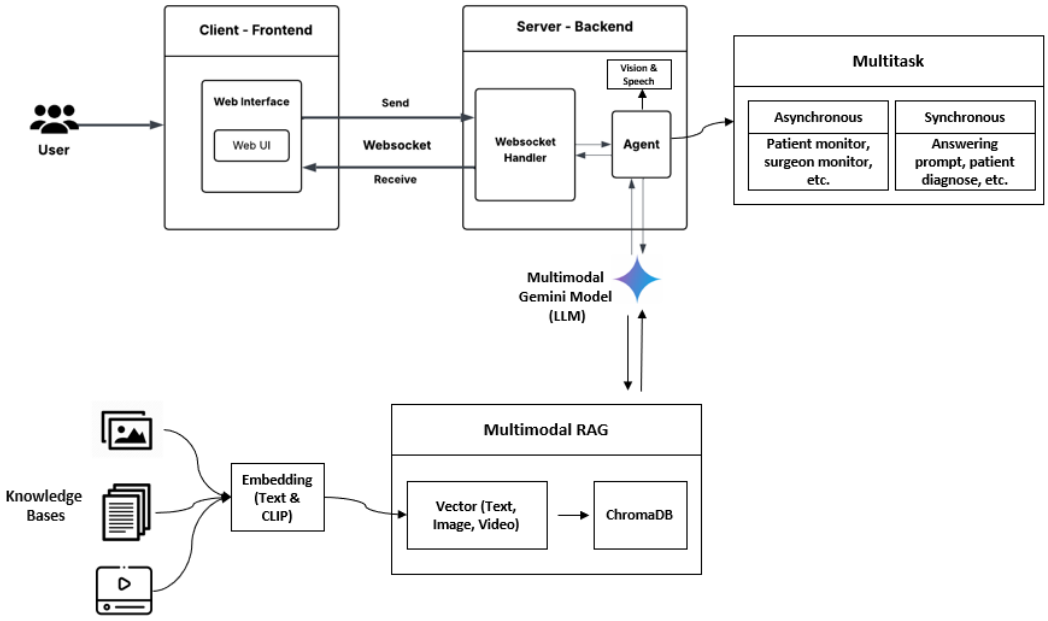

Figure 1: Multitask Asynchronous Bidirectional MUltimodal Agent Architecture

3.3 Multitask Design

The multitask asynchronous design of our agent focuses on three integrated capabilities that work in parallel to support personalized treatment. First, the vision component continuously analyzes live video streams to detect patient posture, mobility, and non-verbal indicators of discomfort or pain. Second, the speech component transcribes and interprets real-time dialogue between patients and clinicians, enabling natural and accessible communication. Third, aggregation process combines these multimodal signals into coherent health metrics, synthesizing visual and spoken information to provide adaptive treatment planning and timely alerts. This design allows the agent to capture both objective observations and subjective reports, enhancing the accuracy and responsiveness of personalized care

3.4 Safety, Ethics, and Workflow Considerations

The agent is non-diagnostic and clinician-supervised. We implement: (i) disclosure and consent; (ii) on-device redaction of PII where applicable; (iii) gated tool use and escalation policies; (iv) uncertainty communication (verbal hedging and confidence cues); (v) audit logs of retrievals and actions; and (vi) opt-out controls. These choices reflect HAI guidance and clinician co-design evidence [1, 7].

3.5 Evaluation Setup

We evaluate the system through a testing simulation designed to represent real-world communication scenarios between the agent and a patient. In these simulations, the patient interacts with the agent by asking questions and providing information verbally, while the agent accesses and reasons over its multimodal knowledge bases containing text, images, and video. This setup allows us to assess how well the system integrates speech and vision streams with multimodal RAG to deliver accurate and relevant responses. Performance testing is conducted using the Opik LLM framework, which enables comparison of expected versus actual outputs across several metrics: Relevance, Usefulness, Hallucination, Contain, Equal, and Sentence BLEU (Bilingual Evaluation Understudy).

4 Result and Discussion

Based on the functional testing in (Table 1), the agent achieved a Usefulness Metric of 0.78, a Relevance Metric of 0.93, a Hallucination Metric of 0.3, a Contain Metric of 0.88, an Equals Metric of 0.88, and a Sentence BLEU score of 0.98. These results indicate that the system provides highly relevant and linguistically faithful responses while maintaining strong consistency with the expected outputs. The relatively low hallucination score (0.3) shows that multimodal RAG grounding is effective in reducing unsupported statements, while the high BLEU score highlights strong alignment between expected and actual responses.

The results further demonstrate that the bidirectional multimodal agent not only outperforms unimodal baselines but also delivers practical and context-aware interactions in simulated patient–agent communication. The usefulness score confirms that the system's responses are actionable for treatment guidance, and the containment and equality metrics show that the retrieved information aligns well with the ground truth knowledge base. Overall, these findings validate the effectiveness of integrating streaming video, speech, and multimodal RAG for building trustworthy personalized treatment companions.

Table 1: Performance metrics of the bidirectional multimodal agent in simulated patient–agent communication using the Opik LLM framework.

Metric	**Score**
Usefulness	0.78
Relevance	0.93
Hallucination	0.30
Contain	0.88
Equals	0.88
Sentence BLEU	0.98

5 Limitations and Future Work

Our study is simulation-based and non-diagnostic; we did not evaluate on protected health information or multi-omics data. Future work will (i) extend grounding to physiological signals and structured EHR, (ii) investigate distillation of specialist components to reduce latency and cost [2], and (iii) explore agentic memory/planning for longitudinal personalization [3].

6 Conclusion

We introduced a Multitask Asynchronous Bidirectional Multimodal Agent for Personalized Treatment Companions, integrating vision (video) and speech (audio) with multimodal reasoning and RAG. By supporting real-time patient monitoring, health metric aggregation, and natural dialogue, the system demonstrated strong performance in functional testing with a Usefulness Metric of 0.78, a Relevance Metric of 0.93, a Hallucination Metric of 0.3, a Contain Metric of 0.88, an Equals Metric of 0.88, and a Sentence BLEU score of 0.98. These results show that the agent produces relevant, useful, and trustworthy responses while minimizing hallucinations. Overall, this work advances

human-centered healthcare by demonstrating how streaming multimodal agents with RAG can enable adaptive, interactive, and trustworthy personalized treatment.

References

[1] Maryam Alaeifard, Mohammad Safaei, and Elham Karim Zadeh. Advancing human-agent interaction: Bridging the gap between vision and reality. *International Journal of Advanced Human Computer Interaction*, 1(1), 2024.

[2] Yiwei Guo, Shaobin Zhuang, Kunchang Li, Yu Qiao, and Yali Wang. Transagent: Transfer vision-language foundation models with heterogeneous agent collaboration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[3] Chengkai Huang, Junda Wu, Yu Xia, Zixu Yu, Ruhan Wang, Dongruo Zhou, Julian McAuley, Tong Yu, Ruiyi Zhang, Ryan A. Rossi, Branislav Kveton, and Lina Yao. Towards agentic recommender systems in the era of multimodal large language models. *arXiv preprint arXiv:2503.16734*, 2025. URL `https://arxiv.org/abs/2503.16734`.

[4] C. Kelly, L. Hu, B. Yang, Y. Tian, D. Yang, C. Yang, Z. Huang, Z. Li, J. Hu, and Y. Zou. Visiongpt: Vision-language understanding agent using generalized multimodal framework. *arXiv preprint arXiv:2403.09027*, 2024.

[5] Mingyang Mao, Mariela M. Perez-Cabarcas, Utteja Kallakuri, Nicholas R. Waytowich, Xiaomin Lin, and Tinoosh Mohsenin. Multi-rag: A multimodal retrieval-augmented generation system for adaptive video understanding. *arXiv preprint arXiv:2505.23990*, 2025.

[6] Runliang Niu, Jindong Li, Shiqi Wang, Yali Fu, Xiyu Hu, Xueyuan Leng, He Kong, Yi Chang, and Qi Wang. Screenagent: A vision language model-driven computer control agent. *arXiv preprint arXiv:2402.07945*, 2024.

[7] Siyi Wu, Weidan Cao, Shihan Fu, Bingsheng Yao, Ziqi Yang, Changchang Yin, Varun Mishra, Daniel Addison, Ping Zhang, and Dakuo Wang. Cardioai: A multimodal ai-based system to support symptom monitoring and risk prediction of cancer treatment-induced cardiotoxicity. In *CHI Conference on Human Factors in Computing Systems (CHI '25)*, Yokohama, Japan, 2025. ACM. doi: 10.1145/3706598.3714272. URL `https://doi.org/10.1145/3706598.3714272`.