THIA: A Modular Therapeutic Human-like Intelligent Agent Combining Sentiment Analysis, Conversational AI, and 3D Avatar Embodiment

Sanat Samal, Kaijie Zhang, Eli Liang, Aaditya Khanuja, Gaurav Nair, Sathwika Peechara, Tania Jain, Shruti Senthilram

***Abstract*—This paper presents the development of THIA (Therapeutic Human-like Intelligent Agent), a modular, multi-component AI system designed to simulate therapeutic dialogue with human-like emotional and visual fidelity. THIA integrates a Retrieval-Augmented Generation (RAG) conversational backend powered by Mistral-7B, a fine-tuned sentiment recognition model using Wav2Vec2, an end-to-end text-to-speech and speech-to-text pipeline, and an expressive 3D avatar system using generative AI and blendshape-based facial animation. By combining these systems, THIA aims to provide a responsive, emotionally aware, and visually engaging virtual therapy experience. We describe the technical design of each module and discuss the challenges and future directions for building intelligent, emotionally attuned virtual agents.**

I. Introduction

Conversational AI systems have progressed significantly with the rise of large language models (LLMs). However, traditional chatbots often fall short in emotionally sensitive domains such as mental health. THIA (Therapeutic Human-like Intelligent Agent) aims to address this gap by integrating multiple AI technologies to provide emotionally intelligent and visually embodied interactions. THIA's architecture comprises four key components: sentiment analysis via audio emotion recognition, a Retrieval-Augmented Generation backend for dialogue generation, an audio pipeline for speech synthesis and recognition, and a 3D avatar engine for realistic facial animation.

II. Sentiment Analysis

Recognizing user emotions is essential in therapeutic settings. We reviewed emotion recognition literature, particularly MELD and other multimodal datasets, to identify the importance of incorporating both vocal and contextual cues in mental health conversations. Our final pipeline uses a custom PyTorch model trained on multimodal data to predict emotions.

A. Dataset Strategy

The most important aspect of training this model was the dataset, as it would be able to accurately represent human emotions. For this purpose, we chose the MELD (Multimodal EmotionLines Dataset) due to its high-quality data and direct relevance to our goals. MELD provides synchronized audio and video at the utterance level, covering a range of emotions from anger to disgust, which is more representative of human emotion than polarity-based scales.

We initially started with a 10% split of the dataset, working our way up to a 60% split (approximately 29,768 samples) once the model design stabilized. This curriculum ramp-up allowed us to iterate quickly, test the training process, and make adjustments for extremely low latency.

We also chose CREMA-D [1] and RAVDESS [2] for their expressive diversity and clear labeling across six core emotions: happiness, sadness, anger, neutral, disgust, and fear. The CREMA-D dataset was chosen as it provided a large and balanced set of six emotions. RAVDESS was used to test the model's accuracy and generalize across different speakers.

B. Model Architecture and Training

The model itself is a custom-built PyTorch network designed for fusing streams of video and audio into a single emotion prediction (angry, sad, happy, neutral, afraid, disgusted). The architecture was constrained to meet three key targets: true emotion recognition, robustness on a medium-scale dataset, and prevention of overfitting.

1) Data Preprocessing and Augmentation: Audio enters the model as a 40-dim COVAREP-like feature (pitch, spectral disrupters, energy), which generalizes better and is more cost-effective than raw audio. For video, we used 512-dim pre-extracted feature embeddings that were pre-trained on faces rather than raw frames. This avoided retraining a heavy video CNN on a limited dataset and kept the compute bounded.

Standardization was performed for each modality by computing the mean and variance on the training set and then z-score normalizing the audio (40-d) and video (512-d) features. On-the-fly augmentation was applied during training to increase diversity and reduce overfitting, including light pitch shifts, additive noise, and SpecAugment-style masking for audio. For video, we used random temporal dropping and averaging of the embedding sequence to simulate missed detections.

2) Model Structure: The model uses separate encoders for each modality before a late fusion step.

- **Audio Encoder:** Takes a 40-dim input and passes it through a small MLP (e.g., 40→128→128) with BN/ReLU/Dropout.
- **Video Encoder:** Takes a 512-dim input and passes it through a small MLP (e.g., 512→256→128) with BN/ReLU/Dropout.

For late fusion, the encoded audio and video vectors are concatenated and then processed by a fusion MLP (e.g., 256→128) before the final classifier head. The classifier head uses a 6-way softmax to predict one of the six emotions. A dropout of 0.4 was used in the encoders and fusion to prevent overfitting, which provided a slowed convergence and lower calibration to ensure the model was not over-analyzing. We also used Focal Loss to down-weight easy, majority-class examples and force the model to learn on hard, minority cases (e.g., afraid, disgusted).

With a total of approximately 990K parameters, the model proved expressive enough while maintaining low latency.

C. Results

Our model achieved approximately 49.5

III. LLM Backend: Retrieval-Augmented Generation

To power THIA's dialogue, we designed a lightweight Retrieval-Augmented Generation (RAG) system using Mistral-7B. Traditional LLMs often hallucinate or lose context in multi-turn conversations—critical flaws for therapeutic use. Our RAG architecture enhances factual grounding and memory by retrieving relevant QA pairs before generation.

A. Background

The RAG paradigm [3] combines the strengths of information retrieval with generative models. This architecture reduces hallucinations and improves factual consistency. Mistral-7B [4] is a powerful open-weight transformer-based model. In our use case, it is paired with a vector database that indexes 3.5k QA pairs to provide memory and knowledge context. Tools such as LangChain [5] assist in chaining retrieval and generation steps together, while also enabling conversational memory modules.

B. System Overview

We use a dense vector index of 3.5k curated QA pairs, embedded using models like E5 or BGE. Retrieval is conducted via Chroma or FAISS vector databases, selecting the top-k similar entries. These entries are prepended to the user's input and passed to the Mistral-7B model via a HuggingFace inference server. LangChain [5] supports chaining and memory, enabling THIA to retain context and simulate consistent therapeutic personas.

C. Future Work

Planned enhancements include prompt engineering, RLHF-based alignment for therapeutic tone, and migration to more capable LLMs such as GPT-4 or Claude. We will also investigate prompt engineering strategies and persona conditioning techniques. This includes experimenting with modular prompt structures, predefined therapist personas, and instruction-tuning techniques.

IV. Speech Processing: TTS and STT

In therapy-oriented dialogue, natural speech synthesis and recognition are vital for building rapport. Our speech-to-text module uses FasterWhisper, modified to transcribe audio word-by-word in real time. For text-to-speech (TTS), we adopted a modular pipeline for low-latency generation.

A. Speech-to-Text (STT)

For the speech-to-text (STT) component, we evaluated and used OpenAI's Whisper model. Whisper is trained on 680,000 hours of diverse and noisy audio-transcript pairs. It performs well in noisy, real-world conditions, supports 96 languages, and includes both transcription and translation functions. A key limitation is that Whisper processes audio in fixed 30-second segments, which introduces latency.

We used the large-v2 version of Whisper. Initially, the real-time factor was around 0.8 using a CPU, meaning the processing time was 80% of the audio length. We modified the existing STT model implementation to improve real-time performance. Originally, the model accumulated all transcribed text into a list and only output the results once the entire audio had been processed. To address this, we adjusted the logic so that the first audio chunk is transcribed and printed separately. For all subsequent chunks, each transcription result is printed immediately using print(..., flush=True). After these adjustments, the real-time factor for STT was reduced to 0.1, making it 8 times faster.

B. Text-to-Speech (TTS)

For the text-to-speech (TTS) component, we chose a two-stage pipeline for its low latency, following the approach described in "A Survey on Neural Speech Synthesis" by Xu Tan et al. The two-stage approach uses an acoustic model to convert text to mel-spectrograms and a vocoder to generate waveforms. For the acoustic model, we chose FastSpeech 2 due to its speed, robustness, and high quality. For the vocoder, we used HiFi-GAN, which is designed to produce natural, human-like audio with very low latency. Together, these models form a pipeline that processes text into waveform output. In testing, the system produced audio with a real-time factor of about 0.4 using a CPU, meaning that the audio is generated more than twice as fast as its duration, which is suitable for live conversational scenarios.

V. Avatar Embodiment: Audio-to-Face Animation

To humanize the interaction, THIA includes a realistic 3D avatar pipeline. This component leverages research in neural rendering and generative modeling to animate facial expressions directly from voice.

A. Background

We drew inspiration from Karras et al.'s audio-to-face paper [6] and Saito et al.'s PIFuHD [7] for 3D digitization. Our pipeline begins with a concept image generated by Google Gemini, then uses Meshy AI and Cube by CSM AI to generate high-resolution 3D meshes. These are cleaned in Blender, then rigged via Mixamo for body movement.

B. Methodology

Our pipeline begins with a character concept image from Google Gemini, which serves as the visual foundation for our 3D avatar. This image is then processed through Meshy AI and Cube by CSM AI to create high-resolution, full-body 3D meshes. These meshes are imported into Blender, where they are cleaned and prepped for rigging. We use Adobe Mixamo to automatically rig the full-body mesh. Since Mixamo does not support face-only rigging, we are implementing a custom facial rigging pipeline. This involves creating blendshapes (shape keys) manually in Blender to model expressions and phonemes such as smile, frown, jaw open, and eyebrow raise. These shape keys will later be animated in Unity using audio-driven weights.

C. Results

Facial rigging is handled manually using blendshapes. Expressions are animated in Unity based on voice emotion vectors, which will be tuned in parallel with TTS emotion conditioning. We generated high-quality 3D avatars using Meshy AI and Cube by CSM AI, based on concept images from Gemini. The meshes were cleaned in Blender and successfully rigged for full-body animation using Mixamo. Facial rigging is underway using blendshapes, with expressions like smile, blink, and jaw movement manually sculpted. Early tests show good compatibility between generated meshes and both body and facial animation workflows.

VI. System Integration and Modularity

Each component of THIA is designed as a microservice, allowing parallel development and scalable deployment. The sentiment model outputs are used to modulate avatar expression and TTS tone. The LLM backend is queried through a FastAPI server that supports contextual chat and session memory.

This modularity supports rapid experimentation—components like the LLM or TTS engine can be swapped without affecting the rest of the system. A structured proof-of-concept is underway to evaluate performance, latency, and user experience across all modules.

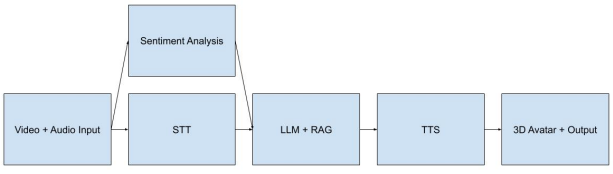

Fig. 1. System Design

VII. Conclusion and Future Work

THIA combines the latest in speech processing, emotion recognition, conversational AI, and 3D avatar animation to create an emotionally intelligent therapeutic assistant. Our current focus is on optimizing latency and alignment across modules. Future plans include RLHF tuning, dynamic prompt templates, and expanded multilingual support. We also aim to enhance cross-module coherence by integrating sentiment signals more deeply into LLM prompting and avatar control. This holistic integration will support richer, more human-like therapeutic engagement. We built a functional pipeline for generating and animating expressive 3D avatars using AI tools and open-source software. With full-body rigging complete and facial rigging in progress, the next steps involve audio-driven animation and emotion tuning. Our workflow shows strong potential for creating realistic AI assistants with personalized, human-like interaction.

Acknowledgment

We thank the THIA development team and faculty mentors for their contributions and support throughout this interdisciplinary project.

References

[1] Cheyney Computer Science and et al., "CREMA-D: Crowd-sourced emotional multimodal actors dataset," https://github.com/CheyneyComputerScience/CREMA-D, 2014.

[2] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, p. e0196391, 2018.

[3] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[4] Mistral AI, "Mistral-7B: A High-Performance Language Model," https://mistral.ai/news/introducing-mistral-7b, 2023.

[5] LangChain Inc., "LangChain Documentation," https://python.langchain.com/docs/introduction/, 2025.

[6] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-Driven Facial Animation by Joint End-to-End Learning of Pose and Emotion," in *ACM SIGGRAPH 2017 Talks*, 2017, pp. 1–2.

[7] S. Saito, T. Simon, J. Saragih, and H. Joo, "PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1–10.