Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Real-Time Audio–Visual Emotion Detection for Human–AI Interaction Using a Cross-Modal Transformere

Nanda Gopal Malladi
SRM Institute of Science and Technology
Vrisheeka Mulakala
SRM Institute of Science and Technology
Deepa N
deepan3@srmist.edu.in

SRM Institute of Science and Technology

Research Article

Keywords: Multimodal emotion recognition, cross-modal transformer (XMT), multi-head attention (MHA), facial emotion recognition (FER), speech emotion recognition (SER), affective computing, human–AI interaction, virtual learning assistant, Mel frequency cepstral coefficients (MFCC), convolutional neural network (CNN), deep neural network (DNN), real-time inference, spatio-temporal modeling, personalized education

Posted Date: March 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8945109/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Real-Time Audio–Visual Emotion Detection for Human–AI Interaction Using a Cross-Modal Transformere

Nanda Gopal Malladi[1], Vrisheeka Mulakala[2], and Deepa N*[3]

[1] SRM Institute of Science and Technology, Kattankulathur, Chennai - 603203, India
nm2312@srmist.edu.in
[2] SRM Institute of Science and Technology, Kattankulathur, Chennai - 603203, India
vm1424@srmist.edu.in
[3] *SRM Institute of Science and Technology, Kattankulathur, Chennai - 603203, India
deepan3@srmist.edu.in

Abstract. The real-timeness, multi-modality, and actuality of the assessments are aimed at improving the human-to-artificial intelligence interaction in the context of a virtual learning environment using continuous observation and assessment of the learners' facial expressions and speech. A multi-modal XMT architecture is developed so as to process visual features from the CNN-based face video frames and acoustic features from the speech using MFCCs jointly by projecting both into a common embedding space and integrating over multiple heads of self-attention along with the feedforward layers to account for temporal dependencies within a modality as well as the relationships across the modalities. Model training and evaluation was conducted on a merged corpus consisting of the CREMA-D and RAVDESS datasets using the 6 basic emotions of anger, disgust, fear, happiness, neutrality and sadness based on a balanced sample of 240 pairs of recorded samples across different speakers, conditions, and intensities of emotions (240 samples total). The results show that the multimodal XMT outperforms both the audio-only and video-only baseline systems consistently with an average recognition accuracy of 73% and an average macro precision, recall, and F1-score of 76%, 75% and 75%, respectively, while maintaining an end-to-end latency of less than 1.5 seconds, allowing real-time interactivity on standard computing platforms.

The emotion recognition framework contributes to this integrated virtual assistant environment by providing feedback that matches the learner's emotion by changing tone, style of feedback, and manner of instruction. When a learner exhibits frustration or low self-confidence, additional support or encouragement will be provided through alternative means of engagement. Conversely, when learners are fully engaged, they will be provided with additional challenge as well as motivation through the content provided to them. Additionally, because of the robustness of multimodal fusion, should one mode of communication become partially ineffective—for instance, when an audio signal is degraded due to background noise or low lighting conditions—the assistant will still be able

to effectively provide an accurate solution and continue to provide stable affective sensing information during the learner's use. These results indicate not only the versatility of transformer-based multimodal architectures and curated audio-visual corpora in providing technical solutions for developing affect-aware personalized tutoring systems, but also that these technologies can contribute to the development of other emotionally intelligent human-AI interfaces and systems used in education and other interactive areas.

Keywords: Multimodal emotion recognition · cross-modal transformer (XMT) · multi-head attention (MHA) · facial emotion recognition (FER) · speech emotion recognition (SER) · affective computing · human–AI interaction · virtual learning assistant · Mel frequency cepstral coefficients (MFCC) · convolutional neural network (CNN) · deep neural network (DNN) · real-time inference · spatio-temporal modeling · personalized education.

1 Introduction

The increasing popularity of virtual learning platforms, intelligent tutoring systems, and chatbots has resulted in a significant demand for technology that can detect and respond in real time to a learner's emotional state. Emotions greatly affect motivation, focus, engagement, and long-term memory retention, thus making emotion an important piece in determining how technology is effective when used as part of an educational or assistance interaction between humans and AI. Unfortunately, the majority of existing virtual assistants and e-learning systems continue to utilize a "one-size-fits-all" interaction style, where the assistants return predetermined (static) responses regardless of the user's emotional status during the course of the interaction. This misalignment often creates disappointment for the learner and causes disengagement, which negatively impacts the learner's ability to learn when he/she becomes confused, anxious, or bored with the response of the virtual assistant or the e-learning system.

Recent research has demonstrated that recent developments in deep learning and affective computing allow for much more complex analyses of emotional expression from both facial and auditory cues. SER uses features such as prosody, pitch, energy, etc., to identify the emotional state of the speaker through the analysis of spoken language, while FER uses convolutional and deep networks to interpret expressions such as happy/angry/sad that are present in either an image or video stream. Although unimodal systems focused only on audio or video tend to have good performance under controlled conditions, they have had difficulty in real-world environments where the signals may be degraded, incomplete, or ambiguous. For example, if SER relies solely on speech to determine emotion, it will probably not work well if someone says something at a very quiet volume and does not vary their vocal pitch; similarly, FER may not provide valid results if the person's face is partially blocked from view and is in a poorly lit area.

To overcome shortfalls, multimodal emotion recognition systems have been developed to combine information from multiple sources (by incorporating both facial expression and speech features) to create a stronger and more reliable prediction of emotion. Many of the earliest multimodal frameworks were established by concatenating features or utilizing basic fusion methods, which improved the capacity of models to produce correct predictions. However, these models generally have a rigid design that requires configuration specific to a dataset/laboratory setting. Although some of the most recent multimodal pipelines and fusion techniques have demonstrated stronger emotion recognition capabilities, many of them still suffer from high computational burden, dependency on large-scale training datasets, and limited applicability to low-latency, real-time applications. Besides these problems, many emotion-aware teaching tools and chat systems only concentrate on ideas and do not offer complete audio-visual real-time systems to support virtual learning assistants.

One of the main gaps that exists today is the integration of emotion recognition into adaptive conversation behavior. Most emotion classification systems today view emotion as a stand-alone task and do not tightly couple an outcome measure like what the assistant's tone or feedback might be to an outcome measure like an emotional state (e.g., anger). For example, in a virtual learning environment, to effectively respond to a student who is experiencing confusion, frustration, or disengagement, the assistant not only needs to know the student's emotion but may also need to adjust how it responds to the student to address their emotional state appropriately by providing simpler explanations, providing additional hints, or providing motivational messages. To have these types of emotion-aware assistants, it will take more than accurate, real-time, multimodal emotion recognition; it will take developing new architecture that can operate continuously, maintaining a response time that is appropriate for human-to-AI interactivity.

We present a real-time multimodal system for recognizing emotion, based on a XMT architecture that processes both facial video and speech audio in order to identify user emotion for the purpose of human-computer interaction. Our work employs a unified, transformer-based fusion framework that models the spatial and temporal relationships of facial expressions and speech using a dataset of recorded videos of subjects' facial expressions and speech patterns. The audio-visual corpus conveys the relationship between the two modalities and consists of recordings of the same recordings from the CREMA-D and RAVDESS datasets, containing six basic emotions (angry, disgust, frightened, happy, neutral, and sad), at various intensities and across multiple speakers and settings. The cross-modal recognition model achieved an overall accuracy of 73%, with excellent macro-averaged precision, recall, F1, and estimated latency of fewer than 1.5 seconds; thus, it is well-suited for deployment in interactive educational contexts.

This study makes the following main contributions:

1. A novel multimodal transformer architecture that combines audio and visual modalities using sequence embeddings, multi-head attention, and feed-forward networks to learn a joint representation of emotions.

2. A new multi-corpus dataset that merges the CREMA-D and RAVDESS databases into a balanced corpus of six emotion classes. This corpus enables training and evaluation of emotion recognition algorithms in a realistic multimodal environment with variations in speakers, recording conditions, and emotional intensity.
3. A real-time multimodal emotion recognition pipeline with an inference latency of less than 1.5 seconds, making it suitable for deployment on mobile devices and for displaying dynamic instructional content.
4. An emotion recognition module integrated into a virtual assistant conceptual framework that personalizes and adapts tone of voice, response style, and level of instructional support based on the user's current emotional state, thereby enhancing the naturalness, personalization, and emotional intelligence of human–AI interaction.

2 Related Work

This section reviews and summarizes the literature and contributions made in the area of emotion recognition for human and AI interaction through speech-based, vision-based, multimodal, and emotion-aware pedagogical and conversational agents. In addition, we will identify limitations in these areas, specifically in real-time adaptive multimodal learning environments.

2.1 Speech Emotion Recognition

SER is often accomplished by utilizing the acoustic characteristics (e.g., pitch, energy, intensity, and frequency) contained within speech for the purpose of estimating emotional states from spoken utterances. The two main types of machine learning techniques that have been employed in SER systems are DNN and CNN models, which both have demonstrated their ability to model complex relationships between speech features and the corresponding discrete emotion categories, as evidenced by their success on standard SER databases. For example, Pancharia et al. created a DNN-based SER that establishes a good baseline for the framework based on relative computational speed. Furthermore, hybrid ConvLSTM architectures (e.g., Hasan et al.) also provide increased SER performance by modeling temporal dependencies and identifying dynamic features of speech over time, which improves the robustness of emotion recognition for emotions that develop over a period.

However, although the above-mentioned advances have been made in SER, when audio-only SER systems are used in real-world, interactive situations, they have limitations. For example, the quality of the speech signal may be drastically altered due to background noise, variability in microphonics, variability in channels, etc. Specifically, background noise, channel variances, and microphone quality deplete recognition performance. Emotions that are milder (e.g., mild confusion and muted sadness) generally will provide less of an acoustic cue versus the same emotion that can be demonstrated via a facial expression;

therefore, systems that only utilize speech for estimating emotion will not perform well when individuals display only a small amount of speech or are silent. These limitations in the use of solely audio-based SER systems create the need for incorporating visual-based additional modalities to improve the accuracy of emotion estimates.

2.2 Facial Emotion Recognition

FER is the use of facial images or video streams to extract emotions. Many of the current methods of FER rely primarily on the use of CNN and other deep learning architectures to extract emotion-encoding features from images or videos. CNN-based FER methods can learn to identify facial muscle movement patterns and configurations associated with basic emotions such as happiness, anger, sadness, and fear (Deshmukh et al., 2023). The authors discuss the most recent FER systems in the literature and highlight their significance as a foundation for real-time emotion recognition in videos.

However, FER provides a narrow angle from which to analyze an emotional expression. Within real-time video streams, facial expressions can be occluded, viewed from different angles (i.e., rotations), viewed in low light, and viewed when the subject is moving (i.e., motion blur). As a result of these factors, determining whether an individual expressed their emotions through a facial expression with a high degree of certainty can be difficult. Additionally, some individuals do not show a typical facial expression or may show a culturally specific facial expression that does not match the traditional databases used for developing FER systems. The use of only visual information to analyze an emotional expression limits the ability to use prosodic cues to assist with disambiguating facial expression when the expression is ambiguous (e.g., what is the difference between a facial expression for being happy and an expression for being anxious?). Given these limitations, there is a need for the use of multimodal systems that combine both facial expression and prosody when analyzing emotion.

2.3 Multimodal Emotion Recognition and Fusion

FER is the use of facial images or video streams to extract emotions. Many of the current methods of FER rely primarily on the use of CNN and other deep learning architectures to extract emotion-encoding features from images or videos. CNN-based FER methods can learn to identify facial muscle movement patterns and configurations associated with basic emotions such as happiness, anger, sadness, and fear (Deshmukh et al., 2023). The authors discuss the most recent FER systems in the literature and highlight their significance as a foundation for real-time emotion recognition in videos.

However, FER provides a narrow angle from which to analyze an emotional expression. Within real-time video streams, facial expressions can be occluded, viewed from different angles (i.e., rotations), viewed in low light, and viewed when the subject is moving (i.e., motion blur). As a result of these factors,

determining whether an individual expressed their emotions through a facial expression with a high degree of certainty can be difficult. Additionally, some individuals do not show a typical facial expression or may show a culturally specific facial expression that does not match the traditional databases used for developing FER systems. The use of only visual information to analyze an emotional expression limits the ability to use prosodic cues to assist with disambiguating facial expression when the expression is ambiguous (e.g., what is the difference between a facial expression for being happy and an expression for being anxious?). Given these limitations, there is a need for the use of multimodal systems that combine both facial expression and prosody when analyzing emotion.

2.4 Emotion-Aware Pedagogical and Conversational Agents

In addition to foundational recognition models, there has been extensive research into emotion-aware pedagogical agents and conversational AI that adapt to users' affective states. In particular, there have been systematic literature reviews (Septiana et al.) of emotion-related pedagogical agents that indicate that being able to adapt to learners' emotions is critical for engagement, motivation, and success in learning to summarize some design principles on how to integrate emotion into instruction. Huang et al. introduced Emotional RAG (an emotional retrieval-augmented generation framework), which demonstrates that enhancing role-playing agents through the use of emotional memory increases perceived naturalness and feelings of empathy when interacting with text-based dialogue by adding emotional context to that interaction. Other literature continues to explore AI-driven learning platforms, generative AI to create personalized content, and monitoring learners' attention while participating in e-learning environments, indicating that there is an increase in the number of affect-aware educational technologies available.

Many emotion-aware agents have been developed in the research environment; however, they have not yet been fully realized in practical use. Some ways in which emotion-aware agents fall short are: a. There is a lack of a direct relationship between emotion-aware agents and high-fidelity audio-visual measures of emotion recognition. B. Evaluation of their effectiveness in highly controlled and limited generalizability environments. Furthermore, existing emotion-aware technologies primarily focus on either recognition or adaptation, but they are not integrated into a continuous experience pipeline that both senses emotions and immediately alters interaction strategies during a virtual learning experience.

2.5 Summary and Positioning of the Present Work

First, unimodal SER and FER systems work well in a controlled environment, but they are insufficiently robust in real-world, uncontrolled environments where an audio-vocal signal may have degraded or ambiguous information that cannot be used to determine human emotion. Second, many multimodal fusion approaches provide accuracy at a significantly greater computational expense than

unimodal methodologies and may not work well under low latency and real-time constraints for mobile or interactive technologies. Third, many emotion-aware pedagogical and conversational agents do not provide an end-to-end multimodal emotion recognition pipeline that considers facial and vocal signals together in real-time.

This paper proposes a XMT architecture that jointly learns audio and visual signals using sequential embeddings, multi-head self-attention, and feed-forward networks to capture temporal and cross-modal relationships. The XMT model was developed using a combined CREMA-D and RAVDESS corpus and is capable of accurately classifying six different emotion categories from multiple speakers and recording conditions and maintaining inference latency below 1.5 seconds (real-time). In addition, the XMT architecture was designed to assist with the development of emotion-aware virtual assistants that modify their tone, feedback, and instructional materials based on the emotional state of the user detected in virtual learning environments.

3 Methodology

This section describes the overarching design of the overall architecture of the proposed multimodal emotion recognition system, the creation and preprocessing of the audio–visual dataset, the feature extraction pipelines used for speech and facial expressions, and the design and training of the XMT used to classify emotions.

3.1 System Overview

A live multimodally connected pipeline calling on audiovisual user input to produce output of six distinct emotions is outlined in detail here. The pipeline can be broken down into four basic components: data acquisition, modality-specific preprocessing, creation of tokens, and finally fusion using a transformer followed by emotion classification. The architecture of this pipeline is represented in Figure 1 and illustrates how the raw samples of each dataset will be transformed to their unified sequential representation. This unified representation is then classified into one of six emotion categories by way of the Cross Modal Transformer and classification head.

During the first stage, the audio–video pairs from the merged CREMA-D and RAVDESS corpus are temporally aligned so that both speech and facial expression share the same emotional segment of time. The audio stream within the audio preprocessing block undergoes operations such as resampling, normalization, and segmenting into fixed length windows, then MFCC features are extracted to capture the spectral and prosodic properties of the speech signal. Meanwhile, the video stream undergoes processing by the video preprocessing block where frames from each video are extracted, detected/cropped, resized, and temporally uniform sampled to obtain a sequence of frames of user's face at various times to represent how user's face evolved with emotion through time.

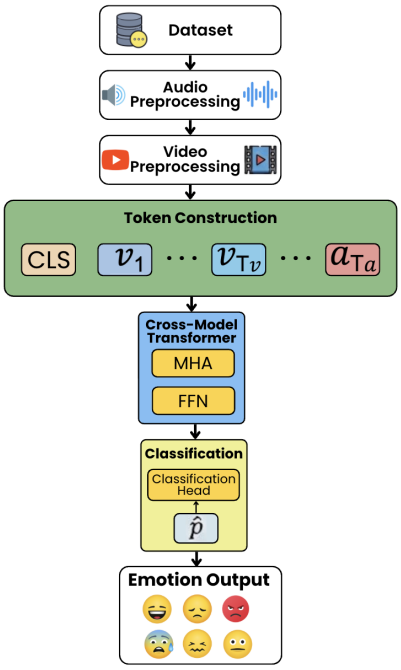


Fig. 1. Multimodal audio video pipeline for emotion recognition.

In the next phase, which is represented by the token creation module as seen in Fig. 1, the pre-processed feature sets are combined into one sequence that can be processed by a transformer. The CNN-generated embedding vectors for the face frames of each sample are transformed into visual feature vectors. The MFCC generated embedding vectors of each audio window are transformed to become audio feature vectors, and then both sets of vectors are projected into a common embedding dimensionality. A unique CLS token is added to these combined visual and audio tokens before the two sets of tokens are concatenated to generate one sequence; each token has position and modality/type embedding so as to track temporal position and provide information about the modality or type of source used. This final combined sequence is input into a Cross-Modal Transformer that models both intra and inter-modal relationships over time through the use of MHA and FFN layers.

After traversing through several layers of the transformer architecture, the CLS token's output is transmitted through a classification head which produces a probability distribution over 6 different classes representing emotions. After this distribution is generated, the application will select the class with the highest probability as the final emotion output for the virtual assistant to adjust its real-time responses. Overall, the architecture has been shown in figure 1 to provide an end-to-end method of going from raw A/V (audio/visual) data to emotion

labels by incorporating modality specific preprocessing, cross modal fusion and classification into one cohesive framework.

3.2 Dataset Construction and Preprocessing

The system uses the merged CREMA-D and RAVDESS datasets as a training base to allow for strong multimodal learning with respect to multiple speakers and recording environments. The CREMA-D dataset contains 7,442 audio recordings from 91 actors, and the RAVDESS dataset contains 2,880 audio-visual recordings of 24 trained actors with superior quality. Both datasets contain six basic emotions, including anger, disgust, fear, happiness, neutral, and sadness, thus providing a common label space for multimodal training.

Only the recordings that contain the six target emotions will be selected from each of the datasets and combined to create one multimodal corpus. The combined multimodal corpus will be split into a training set and a held-out test set of 240 balanced samples with 40 samples for each of the different emotions to ensure that there is a fair assessment across all the classes (emotions). The training set will contain recordings from a diverse range of speakers with various levels of emotion intensity and multiple recording environments (e.g., within different locations and under various lighting conditions), which will support the model's increased generalization to other users. When processing videos visually, they are divided into their individual frames, with all faces within those frames detected and cropped prior to resizing them so that they're in a fixed size for processing by convolutional neural networks. The frames are then sampled uniformly over the times of the video clips, creating a manageable sequence length of frames that will sequentially show the temporal progression of expressions. Finally, the audio portion of the video will also be inspected to determine if there is a resampling requirement, thus allowing for proper normalization and segmentation to match the video's time window.

3.3 Audio Feature Extraction

Converting speech signals into sequences of feature vectors, frequently used in speech processing, is done using a technique known as MFCC, which simulates how the human ear perceives sound. This includes dividing each audio segment (speech) into overlapping windows and measuring MFCCs for the resulting signals through use of appropriate frame size and hop length to yield a time-frequency representation of speech. The final output MFCC matrices from this process are then either flattened or aggregated along the temporal axis to create a sequence of audio-frame feature vectors.

The linear projection layer maps raw audio features to the model's embedding D so that each audio/video token will have a common latent representation. Bias terms are added to the linear transformed vector, and training can be stabilized using optional normalization on the resulting vector.This process produces a sequence $\{a_u\}_{u=1}^{T_a}$ of audio tokens, where T_a is the number of audio frames considered per sample during training.

To extract features from the visual modality of the video, every preprocessed image of the face is passed through a CNN-based feature extractor that produces a compact representation of the facial expression within that image. A pre-trained CNN backbone (for example, a standard image classification network) is used to produce frame-level feature maps that will ultimately be pooled or flattened into a single fixed-length feature vector for each frame.This step converts the input video into a sequence of visual feature vectors $\{f_t\}_{t=1}^{T_v}$, where T_v is the number of sampled frames.

These visual feature vectors are then passed through a linear projection layer to map them into the same embedding dimension D used for audio, yielding visual tokens $\{v_t\}_{t=1}^{T_v}$.As with the audio tokens, bias terms and optional normalization are applied, ensuring that both modalities are compatible for joint processing in the transformer.

3.4 Cross-Modal Transformer Architecture

The methodology includes the use of a XMT to enable joint temporal and cross-modal modelling between audio and visual tokens. To represent an overall summary for emotion classification purposes, a special CLS token is added as the first token in a sequence of tokens. The sequence of tokens created for the XMT is built by concatenating a CLS token with the sequence of visual tokens v_t and the sequence of audio tokens a_u.

To ensure the ordering of tokens and their modality information, each token is assigned a positional embedding indicating its position within the sequence, in addition to a representation of its modality/type, i.e., CLS, video stream, or audio stream. The resultant embedding sequence is sent into L transformer layers, where each transformer layer contains a multi-head self-attention block followed by a position-wise feedforward network (with both sublayers using residual connections and layer normalization).

With multi-head self-attention, all tokens in a sequence can attend to one another, enabling the framework to establish how the model forms long-range temporal correlations between modalities and how they interact (i.e., the relationship between audio and video). Specifically, through audio and video modality interaction, the transformer is able to identify how specific facial expressions are matched with prosodic and spectral speech characteristics associated with each emotion class. Additionally, the token representations will undergo further nonlinear transformations using position-wise feed-forward networks, thereby enhancing the model's expressiveness.

At the end of the last transformer layer, the CLS token contains an ideal summary of the audio and video parts of the input. This representation is passed to a fully connected programmable classification (with or without nonlinear activation) to provide a probability distribution over the 6 emotion classes via the softmax layer. Regularizers (e.g., dropout, weight decay) are also utilized throughout the transformer layers to limit overfitting.

3.5 Cross-Modal Transformer Algorithm

The detailed processing steps of the proposed Cross-Modal Transformer (XMT) for audio–visual emotion recognition are summarized below.

Algorithm 1: Cross-Modal Transformer (XMT) for Audio–Visual Emotion Recognition

1) Data & Token Construction

Extract per-frame visual features and audio-frame features and project them to the model dimension D:

$$v_t = W_v \phi_v(\text{frame}_t) + b_v, \quad t = 1, \ldots, T_v \tag{1}$$

$$a_u = W_a \phi_a(\text{mfcc}_u) + b_a, \quad u = 1, \ldots, T_a \tag{2}$$

where $\phi_v(\cdot)$ is the CNN frame extractor and $\phi_a(\cdot)$ is the audio feature extractor.

2) Transformer Processing

For each layer $\ell = 1$ to L:

a) Build token sequence and add positional and modality/type embeddings:

$$z^{(0)} = [\text{CLS}, v_1, \ldots, v_{T_v}, a_1, \ldots, a_{T_a}] \tag{3}$$

$$z_i^{(0)} = z_i^{(0)} + P(i) + E_{\text{type}}(m(i)) \tag{4}$$

where $m(i) \in \{\text{cls}, \text{video}, \text{audio}\}$.

b) Multi-head self-attention and feed-forward update:

$$y^{(\ell)} = \text{LayerNorm}\left(z^{(\ell-1)} + \text{MHA}\big(z^{(\ell-1)}, z^{(\ell-1)}, z^{(\ell-1)}\big)\right) \tag{5}$$

$$z^{(\ell)} = \text{LayerNorm}\left(y^{(\ell)} + \text{FFN}\big(y^{(\ell)}\big)\right) \tag{6}$$

c) Optional regularization (dropout / weight decay) applied during training.

3) Readout / Output Layer

Compute the prediction using the final CLS token representation:

$$\hat{p} = \text{softmax}\left(W_o \sigma\left(W_h z_{\text{CLS}}^{(L)} + b_h\right) + b_o\right) \tag{7}$$

where $\sigma(\cdot)$ is a nonlinear activation function and $\hat{p}$ is the predicted class-probability vector.

4) Final Prediction

$$\hat{y} = \arg\max_k \hat{p}_k \tag{8}$$

$$\hat{p} = f_{\text{XMT}}\left(X^{(v)}, X^{(a)}\right) \tag{9}$$

5) Training Objective

Training is performed using cross-entropy loss with optional class weighting and regularization:

$$\mathcal{L} = -\sum_{i=1}^{N} w_{y_i} \log\left(\hat{p}_{i,y_i}\right) + \lambda R(\Theta) \tag{10}$$

where w_{y_i} are class weights, $R(\Theta)$ is a regularization term (e.g., weight decay), and Θ denotes all trainable parameters.

To ensure the ordering of tokens and their modality information, each token is assigned a positional embedding indicating its position within the sequence, in addition to a representation of its modality/type, i.e., CLS, video stream, or audio stream. The resultant embedding sequence is sent into L transformer layers, where each transformer layer contains a multi-head self-attention block followed by a position-wise feedforward network (with both sublayers using residual connections and layer normalization).

With multi-head self-attention, all tokens in a sequence can attend to one another, enabling the framework to establish how the model forms long-range temporal correlations between modalities and how they interact (i.e., the relationship between audio and video). Specifically, through audio and video modality interaction, the transformer is able to identify how specific facial expressions are matched with prosodic and spectral speech characteristics associated with each emotion class. Additionally, the token representations will undergo further nonlinear transformations using position-wise feed-forward networks, thereby enhancing the model's expressiveness.

At the end of the last transformer layer, the CLS token contains an ideal summary of the audio and video parts of the input. This representation is passed to a fully connected programmable classification (with or without nonlinear activation) to provide a probability distribution over the 6 emotion classes via the softmax layer. Regularizers (e.g., dropout, weight decay) are also utilized throughout the transformer layers to limit overfitting.

3.6 Training Objective and Optimization

Training the model was done by using the labeled multi-modal datasets taken from two sources. The two datasets, named the CREMA-D and RAVDESS groups of datasets, were combined into one dataset where all of the labels for a specific target emotion were represented as a one-hot vector for that target emotion, and then the outputs of the network are one predictive probability (vector) over the six classes. The training objective of cross-entropy losses between the true labels of these datasets versus the predicted probabilities (e.g., using a class weighting to address possible residual imbalances of the compound dataset).

Using the Adam optimizer at the right learning rate schedule, we optimize jointly the parameters of the CNN feature extractor (if fine-tuned), projection layer, transformer blocks, and classification head. Hyperparameters specify the number of transformer layers L, the size of the embedding dimension D, the number of attention heads, the size of the batch, the learning rate, the dropout rate, and the weight decay coefficient. To avoid overfitting and select which trained model should be evaluated, either validation-based model selection or early stopping techniques are employed.

3.7 Real-Time Inference Considerations

The objective of the entire pipeline is to provide real-time capacity using an interactive avatar (or digital assistant). The design goal of this overall system is to ensure that the average latency for inference (or prediction) will be less than 1.5 seconds for every input segment to achieve this objective. This is accomplished by limiting both the average number of frames and the number of audio windows that will be processed for each sample, therefore limiting the total amount of processing needed to perform inference for the entire set of samples. For example, the backbone of the CNNs will have to be efficient (e.g., low number of parameters, etc.), as will the overall configuration of transformers used to process the data collected by the digital assistant. Additionally, the preprocessing steps (i.e., face detection and extraction of MFCCs) will be optimized for speed of execution on standard hardware, allowing for the base XMT to be integrated into mobile or web-based learning systems.

The methodology presented in this paper will use both speech and facial cues in order to produce accurate and timely emotion recognition, which provides the basis for creating a context-aware adaptive emotion function for an educational digital assistant using the proposed XMT-based system.

4 Experimental Results

This part discusses the experimental setup, evaluation protocol, quantitative results, and detailed information about the XMT to determine if sound and visual recognition will provide accurate recognition of emotions when presented together with an ablation study, qualitative observations, and real-time characteristics when testing them in this manner.

4.1 Evaluation Protocol

We conduct our experiments with a merged audio-visual corpus constructed from CREMA-D and RAVDESS across six emotion classes (angry, disgust, fear, happy, neutral, and sad) after filtering out recordings containing one of the six core emotion classes. All recordings filtered for containing the six core emotion classes were combined together into one data set, partitioned into an experiment-training sample set, a validation set for hyperparameter tuning, and a fixed

test set containing 240 sample clips. The test set is balanced between each of the six classes (i.e., there are exactly 40 samples from each of the 6 classes), thereby allowing for fair comparison between emotion classes and ensuring that all emotion classes will be treated without bias toward majority categories.

The splitting of the dataset into training and test/validation sets is done at the speaker level, where applicable, meaning that no speaker in the test set will be overrepresented in the training set. This allows for generalizing the model across speakers and recording conditions. For the primary set of experiments, a fixed train/valid/test split is maintained, and for verification purposes, other validation runs with random splits are also run to show that the model is finding replicable results and that performance is not specific to an individual dataset.

To evaluate the model's performance, we will report overall accuracy, macro-averaged precision, recall, and F1-score (for emotional detection). In addition to the above performance metrics, we will also report metrics by class, including confusion matrices. Overall accuracy corresponds to the percentage of correctly classified samples divided by the number of total test samples. Precision and recall reflect the confidence with which the predicted label (i.e., emotion) corresponds to a correctly classified sample; specifically, precision reflects the extent to which all predicted instances correspond to a successful classification of that sample as an instance of the emotion being predicted (precision), and recall reflects how many instances of the emotion are retrieved by the model from the total number of actual instances of the emotion (recall). The F1-score is a single summary statistic per class that is particularly useful in cases where class distribution is not uniform or there is a significant trade-off between precision and recall. To visualize common misclassifications, we will also provide confusion matrix plots for emotional category misclassifications.

4.2 Quantitative Results

The swaddled researchers worked with their highly complex analysis of the proposed multimodal XMT model's per-class performance summaries from the held-out testing set, reporting precision, recall, F1 score, and support levels for each of the six defined emotion classes. On 240 testing samples, the model achieved an overall accuracy rate of 73%, meaning that nearly three out of every four audio-visual clips were correctly classified to their identified emotional class. Another measure of performance, macro-averaging across all of the classes on precision (76%), recall (75%), and F1 score (75%), also suggested relatively balanced performance across all classes to ensure performance was not dominated by one easy-to-classify class.

A more in-depth look at Table I indicates that there are significant differences between the emotions represented. Neutral expressions achieve perfect scores across the various performance measures (100% precision, recall, and F1-score), which means that this model has successfully created features that reliably distinguish the neutral emotional state from the emotionally charged ones. The respective F1-scores (80% and 74%) of happy and disgust also indicate that these two emotional states are composed of distinct facial and vocal features

Table 1. Classification Report for Multimodal Emotion Recognition

Emotion	**Precision**	**Recall**	**F1-score**	**Support**
Angry	64%	75%	69%	44
Disgust	85%	65%	74%	44
Fear	64%	54%	59%	44
Happy	78%	81%	80%	44
Neutral	100%	100%	100%	20
Sad	65%	77%	70%	44
Accuracy		73%		240
Macro Avg	76%	75%	75%	240
Weighted Avg	74%	73%	73%	240

within the combined CREMA-D and RAVDESS datasets. However, the lower F1-scores of angriness (69%), fear (59%), and sadness (70%) suggest that these three negative emotions have a greater potential of being confused with each other (which aligns with the confusion matrix analysis discussed in Section 4.3).

To evaluate how much improvement is attributable to multimodal fusion, we conducted two unimodal baselines using our shared train/validation/test splits. The first was an audio-based SER model that inputted only MFCC feature vectors using a transformer encoder specifically designed to process single-stream temporal sequences. The second baseline consisted of a video-based FER model that processed only the CNN-based frame embeddings without any audio content. These models were made to be as complex as the XMT and trained under the same conditions so that any performance differences are due to the different modalities and not the models' sizes.

Fig. 1 shows a visual representation of the relationship between overall accuracy and macro F1-score for the audio-only, video-only, and multimodal XMT models. The audio-only model has overall low scores due to its susceptibility to background noise and lack of visual data when voice prosody is not easily identified. The video-only model typically provides an improvement over the audio-only baseline; however, it encounters similar problems to the audio-based models when facial data is obstructed, there is a lack of lighting, or subtle facial expressions are presented. The multimodal XMT presented in Fig. 1 outperformed both of the unimodal systems in both accuracy and macro F1-score, as such, supporting the conclusion that combining both complementary speech signals and facial signals improves the robustness and accuracy of recognizing emotions regardless of recording conditions.

Having multiple sources of information gives a quantitative advantage to a virtual assistant that is sensitive to emotions. Having multiple modalities allows for more consistency in responses because even if a user's emotional state cannot be inferred from one source of data alone, if there is another source available to infer the same emotional state, that source can be taken into account. For instance, if someone is using a quiet voice when speaking but is also providing good emotional cues from their body language, or vice versa, where the user is using exaggerated body language but is not visible because they are off camera,

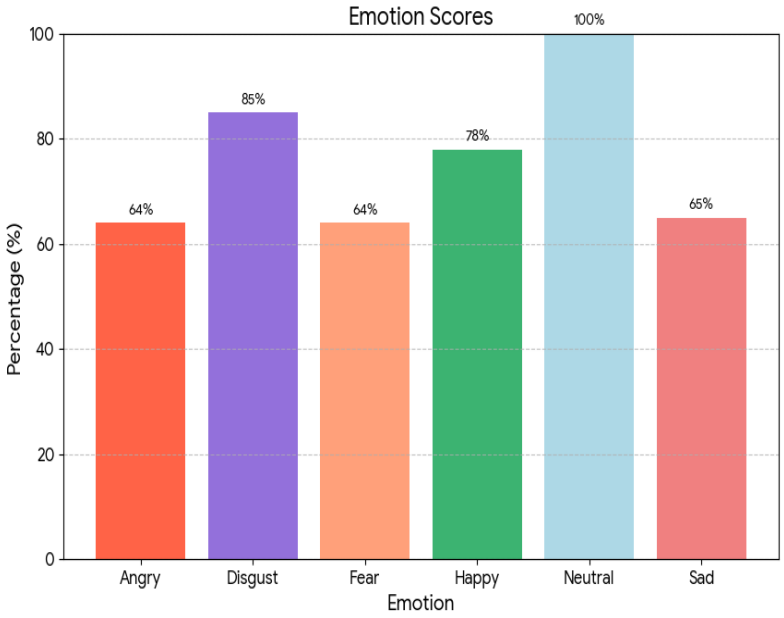


Fig. 2. Comparison matrix for multimodal emotion recognition.

these characteristics will allow for the XMT to leverage whichever gives them enough information to make a decision. The data shown in Table I and Figure 1 provide strong empirical evidence in support of the use of XMT in human/AI interaction in a real-time environment.

4.3 Confusion Matrix and Emotion-Wise Behavior

The confusion matrix from a multimodal XMT model is illustrated in Fig. 2 for the 240 sample test set. Confirmed by the three diagonal entries for neutral, happy, and disgust, the model has correctly identified each emotion with high accuracy; specifically, the emotion neutral has been classified with accuracy. According to the data from this test, these results indicate that it is easier for the model to record neutral and very expressive positive or negative emotions using unique facial and voice characteristics. Neutral speech usually has little or no variation in prosody (pitch) and little or no movement of facial muscles, with both sound and visual components being similar throughout. Happy emotions are identified using eye and mouth smiles with high pitch; disgust emotions are recognized when certain parts of the mouth and nose move at the same time as distinct sounds are made.

Unlike happiness and disgust, fear and sadness have lower F1-scores and recall measures, indicating that either of these emotional states will be confused

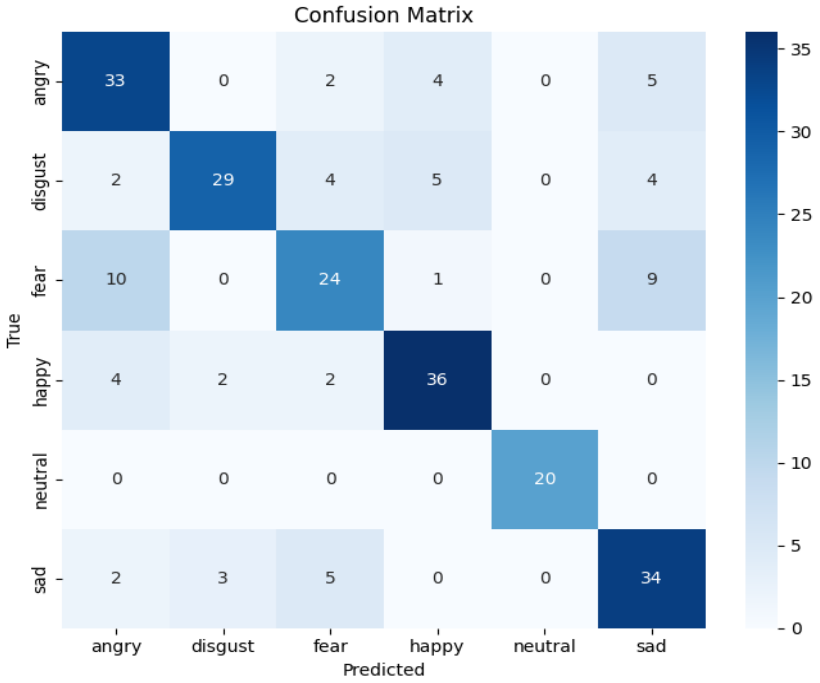


Fig. 3. Confusion matrix for multimodal emotion recognition.

with other negative emotional states, such as anger, or even with a neutral state. This misidentification may occur because fear or sadness can be presented visually with overlapping or subtle activation of facial muscles, creating facial expressions that can resemble one another or the expressions of other mildly negative emotional states when expressed at moderate levels. On the audio side of the analysis, the prosody of the fear, sadness, and anger emotion expressions may not be as apparent in acted-out corpora as they are naturally (e.g., real), creating a much higher likelihood of being misclassified. Additionally, the distribution of classes and the recordings themselves (usage biases) may create an overabundance of happy and disgusted emotion expressions, giving clear cues for the model to learn from yet providing few distinct examples for the model to learn from in the case of subtle fear and sadness. The information from the misclassifications in the confusion matrix reveals where structure exists within the model's errors and indicates where new data collection and modeling improvements should occur.

Looking at these emotion results from a design standpoint, these results indicate that there is good reliance on system models when detecting neutral, happy, and disgust states; however, they will take more care before using system models for predicting fear and sadness states (e.g., requiring greater confidence

thresholds and by smoothing predictions temporally over a number of successive predictions).

4.4 Ablation Studies

The contributions of various architectural components are examined using ablation experiments where important components of XMT are modified (e.g., deleted or changed). One experiment removed modality/type embeddings; thus, the transformer received tokens without knowledge of their source modality (audio or video). As would be expected from previous ablation studies, this manipulation typically produced a decrease in the macro F1-score, indicating that there is value in providing explicit modality cues to aid learning appropriate patterns of attention between and within streams.

The CLS token was removed for a second variant, and all of the token representations at the last transformer layer were averaged together through simple average pooling to create a global embedding. This second variant still provided support for classification, but classification performance decreased compared with the full XMT, indicating that a dedicated classification token directs the model's focus to creating a single representation of global, cross-modal information and is more efficient than pooling uniformly.

Ablation studies were also performed to assess the effect of varying the number of transformer layers and attention heads. In general, reducing the number of heads or layers results in lower performance, as it reduces the model's ability to capture complex temporal and cross-modal relationships, while excessive increases in these parameters result in diminishing returns in performance and could even hurt generalization by overfitting, as well as increasing computational cost. Even relatively small ablation studies summarized in a table provide evidence that the configuration of the XMT used to perform the main experiments provides a practical tradeoff between accuracy and efficiency.

4.5 Case Studies and Qualitative Examples

Besides quantitative analyses, we also utilized qualitative case reviews from the testing samples, as we found many similar examples from our test dataset across sampled modalities as we examined our cases. Success has included evidence of strong emotional expression across modalities (i.e., an example of happy speech that exhibits facial movement along with voice) that can be classified with certainty (happy). An additional example of a sample classified consistently as neutral across the frames is an example of neutral emotion that shows no facial movement and has perceived prosody with little energy.

Challenging cases of failure give us information about the limitations of our current ability to recognize human emotion. For example, in some of the fear samples, the facial expression was low in intensity while the voice produced only mild tension. The combination of both created a situation where our model misclassified the sample(s) as sad or neutral. Conversely, some samples of anger (e.g., angry speech) had very little facial expression but high levels of voice volume

and tension, resulting in the model misclassifying them (e.g., as disgust or fear) due to the occlusion or poor visibility of the video stream. Qualitative examples we have seen provide further evidence of these conclusions and demonstrate that we need more diverse, spontaneous data and better modeling of subtle emotional cues to make better classifications. The value of aggregated temporal recognition of emotion across longer durations of interactions and the incorporation of emotion recognition with contextual information (i.e., task performance or dialogue history) into practical applications has been demonstrated.

4.6 Real-Time Performance

While measuring the time taken by an emotion-based video, we will assess the performance of the system in terms of how long it takes to perform end-to-end inference between the time an audio/video stream has been sampled until the emotion label has been produced. When using our target hardware platform (i.e., a modern machine equipped with either a laptop or a graphics processing unit), the average inference time per video file was less than or equal to 1.50 seconds. In other words, the computation time required to perform the emotion-based classification of an audio/video file was less than or equal to 1.50 seconds. The latency budget was dominated by the CNN portion of the computation, which involves processing the RGB frames from capture until the CNN produces its output (i.e., the predicted emotion label), along with the processing cost required to generate MFCC features from the audio. The transformer model produces a small but predictable fraction of the overall latency budget.

The outcomes validate that our XMT-based pipeline meets the requirement of being responsive enough for interactive 'emotional virtual learning assistants.' Timely, appropriate responses to a user's emotional state are important to keep them engaged. By quickly recognizing emotions from different sources, the pipeline can be used in situations where people and AI interact in real-time, like in tutoring conversations, coaching sessions, or education platforms that understand emotions.

5 Discussion

The results of the experiment strongly suggest that multimodal (XMT) models outperform both audio-only and video-only baselines and provide strong evidence of the importance of combining speech and facial expression into one model to enhance emotion recognition capability. A primary source of this advantage is the use of cross-modal attention, which provides a means for the transformer to align and combine information over time from both audio and video streams; therefore, these modalities are not treated independently. For example, in the presence of a significant rise in pitch, when allowing each token to attend to all other tokens, the model will be able to connect that rise in pitch and its associated facial expression with an emotion, resulting in a more identifiable joint representation than either type of cue could individually provide. The

transformer's capability to model temporally also allows the model to capture dynamic patterns of increasing anger or decreasing sadness over multiple frames, which is critical to accurately recognizing emotional expression in time.

An additional advantage of the XMT's superiority is that it can counterbalance individual modality limitations under realistic conditions. For example, in situations where there is a noisy, reverberant, or poorly recorded audio signal, visual facial expressions can provide the model with sufficient information. Conversely, when there is an occlusion of the face, low lighting conditions, and minimal changes in the face, prosodic and spectral features in the speech stream will assist the model in determining the type of emotion being expressed. The ability of the model to use cross-modal attention allows it to reduce the number of unreliable tokens and then rely on that amount of information in a given context, leading to more accurate performance even when there is noise or variability present.

Although the proposed system has various positive qualities, there are some limitations to be aware of. Creative Emotion Acts (CREMA-D) and the Ryerson Audio-Visual Database of Emotional Speech (RAVDESS) have been developed using actors (both professional and non-professional) who were acting in a controlled environment (and, therefore, all emotive displays have been exaggerated and clean). Therefore, any expressions recorded may be less authentic than the spontaneous, subtle, and culturally conditioned emotions that one would typically encounter in a real-world virtual education environment; thus, they may limit the ability to generalize the model accurately to truly natural behavior. Additionally, the datasets contain very limited representation of accents, languages, and demographic groups, and, therefore, the performance of the model may not transfer well to individuals from underrepresented cultures or those whose respective prosodic characteristics are significantly different from what has been seen during training. Finally, the difficulty the model has with sadness and fear suggests that the subtlety and complexity associated with expressing negative emotions (and of identifying subtle differences through perceptual cues and vocal/visual cues) need much more diverse and higher-quality data to support the development of this type of model.

The results have implications for the application of emotion-aware virtual learning assistants. The robustness of modality XMT to moderate levels of background noise, changes in light, and types of microphones makes it a powerful candidate for deployment on typical students' laptops or mobile computing devices. The use of both modalities will help stabilize predictions and reduce the chance that the affective state of a learner will be coded incorrectly due to a single noisy channel in environments with home/classroom illumination that is not studio quality and/or where there will be other environmental sounds affecting the audio. Some limitations to the prediction have to do with the reliance on frontal facial views and reasonable quality audio; therefore, in the severe occlusion, very low light, and highly distorted speech scenarios, performance will be degraded, and the design of the interface will provide users an incentive to keep their faces visible and speak in their normal voice whenever possible.

As demonstrated by the model's error characteristics, there are possible methods for utilizing it in virtual learning environments. Predictions regarding neutral, happy, and disgusting affective states have a high degree of reliability, enabling the assistant to act directly on these predictions to adaptively use behaviors (such as using the current performance strategy if the learner is neutral or developing an increased range of challenges if the learner appears to be happy or engaged). Performance predictions for other affective states (e.g., fear or sadness) are less reliable. Therefore, the assistant will need to combine affective state predictions with contextual information (e.g., activity failures, long response times, and explicit user feedback) before making decisions for those affective states that require significant adjustment based on their predictions. By combining external influences with imprecise affective state predictions, the impact of incorrect predictions can be reduced. Overall, results presented indicate that multimodal transformers form an effective foundation for practical real-time affect-aware tutoring, but further training on additional spontaneous, diverse, and in-the-wild data is needed to achieve effective and CIF (Culturally Inclusive Framework) virtual learning assistants for all learners.

6 Application to Emotion-Aware Virtual Assistant

The Multimodal XMT Model acts as an emotion recognition tool for a virtual learning assistant and can detect video-based facial expressions and audio-based speech and use this information to provide accurate estimates of the learner's emotional state in real time. Then, based on the estimated emotional state, the system applies different adaptation policies to adjust the assistant's tone of voice, the content of feedback, and the instructional strategy during interactions with the learner. Rather than using emotion recognition passively as an output, the system uses it as a dynamic control signal to develop how to provide explanations, hints, and prompts.

As a neutral learner, the assistant will use a baseline, informative tone, and will generally follow the planned instructional pathway with minimal deviation. Explanations will be brief and task-oriented, and the pace of the instruction will remain consistent with an assumption that the learner is in a steady state, ready to learn (stable). If the learner has demonstrated happiness through successful completion of a task or understanding of a concept, the assistant will provide positive reinforcement and can increase the level of difficulty for the next task by gently increasing the difficulty of the activity or providing enrichment activities in order to continue to stimulate and engage the learner. The response style to the learner will increase in enthusiasm and will give recognition for achievement, as well as encourage the learner to continue building off their previous successes.

When someone is angry or disgusted (which can be a sign of their frustration with the materials they are working with, their user interface, or how they are being instructed), the assistant's response is to use a calming and supportive tone and to try to reduce how much cognitive load the learner experiences when possible. To do this, the assistant will usually slow down their speech rate when

providing explanations, use simpler language in their explanation, and provide step-by-step instructions rather than expecting the learner to connect many different areas of reasoning into one idea. The assistant may also ask specific questions to the learner to give them control over how they will continue, e.g., "Would you like to see a simpler example?" or "Would you like to review what we did before?" By using these types of inquiries, learners can feel more in control and experience less frustration. By only introducing new material to a learner after they have resolved their misunderstanding of the previous material, the assistant is trying to prevent the accumulation of unresolved confusion.

The assistant looks for low confidence, anxiety, and withdrawal from emotion when it sees sadness or fear as well as difficulty achieving tasks. If it detects any of these three feelings, it uses encouragement and reassurance by telling the individual that although the learning materials are difficult, they are also normal, so there is a reason for their struggle. Instructional strategies that the assistant can employ to promote re-engagement include focal hints, breaking problems up into manageable subtasks, and providing a demonstration that the student can refer back to when they attempt other similar tasks alone. The assistant will also ask simple questions such as, "Would you like to try an easier problem first?" or "Can I review the main concept with you?" as a means of actively re-engaging the student without creating additional stress.

To account for all feelings expressed by the Assistant, the Assistant aggregates its predictions across continuous time segments to not make a sudden change in behavior due to a transient expression. For example, one frame of an individual being frustrated will not result in an immediate change to the Assistant's Strategy, but a series of negative affect patterns with repeated mistakes over time would result in the Assistant switching to remediation mode. By applying temporal smoothing through use of the Sequential XMT model, the Assistant will not overreact to brief facial or vocal expression artifacts and will be able to identify consistent emotional trends.

The virtual assistant creates individualized, emotion-based learning experiences by combining real-time, multi-modal emotion detection with adaptive dialogue and teaching strategies. When students are engaged and confident, they are presented with increasingly difficult lessons and positive feedback; when they are confused or frustrated, they are provided with additional assistance and with encouraging responses that are specifically tailored to their current needs. This continuous, closed-loop process—sensing emotion, interpreting emotion, and providing instruction based on interpretation—gives the virtual assistant the ability to mimic a human tutor who would observe nonverbal indicators of a student's emotional state and provide altered support accordingly, increasing both engagement and learning achievements in all human–AI interactions.

7 Conclusion

We proposed a XMT multimodal architecture for real-time, audio-visual emotion recognition within a human-AI interaction environment. There is one unified

transformer that jointly models facial expressions and speech signals from fused video and audio tokens by using temporal dynamics and cross-modal dependencies. The fused corpus, derived from the CREMA-D and RAVDESS—limited to six basic emotions—was used to train the model. The overall accuracy of the model was found to be 73% on balanced test sets of 240 total samples with macro-averaging across three metrics: Precision, Recall, and F1, yielding approximately 76%, 75%, and 75%, respectively. Our multimodal XMT consistently outperformed the unimodal audio-only and video-only baselines at under 1.5 seconds of inference latency, making it a viable candidate for potential use in interactive applications.

The emotion recognition component of this model is included in the design of a conceptual virtual education assistant that will modify the tone, style of feedback, and teaching strategy according to the emotions detected in learners. By making effective use of the combination of multiple sensor modalities, this assistant will also provide learners with supportive responses during periods of frustration or anxiety, while also providing learners with more challenging and motivational forms of interaction when the learners are engaged and confident. This framework demonstrates how affective computing can be used to build real-time emotion-sensitive systems through multimodal data and cross-modal transformer networks. This type of personalized education system can increase learner engagement and facilitate improved educational outcomes by enabling virtual tutors to respond more in real-time to their students in a way that is analogous to human teachers.

8 Future Work

There are many ways that the multimodal emotion-aware assistant can be made stronger and more broadly applicable through future work. The first priority would be to gather and include actual, spontaneous, emotional data that is collected during real virtual learning sessions in the wild rather than only through acting corpora. This data would include natural variations in expression intensity, mixed or ambiguous emotions, and the impact of realistic environmental conditions (background sound, inconsistencies in light levels, and casual camera positioning) on how people express feelings. The ability to curate and annotate this type of data would provide a better basis for the model to use when trying to recognize subtle, context-dependent affective signals and increase the chance that the model will perform well in everyday educational settings.

The second way is to increase the emotional and demographic coverage of the system. Future work can aim to include additional emotions beyond the basic six (e.g., boredom, confusion, curiosity, and flow) that will provide helpful information about how to make instructional decisions. Additionally, collecting speaker data from a diverse set of languages, accents, ages, and cultures will be important to minimize bias and ensure that both vocal and facial cues are interpreted equally and accurately across different populations. Strategies for modeling with

an emphasis on cultural awareness or domain adaptation techniques may be required to explicitly consider cross-cultural differences in emotional expression.

Ultimately, the efficiency of the models and the evaluation of real-world effectiveness will be critical for the large-scale implementation of emotion recognition technology. By creating smaller versions of the XMT (e.g., using model distillation, pruning, or low-rank factorization methods), we will be able to perform emotion recognition in real time on mobile and edge devices with less computing and energy processing capacity than traditional desktop computers. In addition to conducting rigorous user studies in actual classrooms and online courses to evaluate not only how accurately we can recognize the emotions of students through the emotion-aware assistant and how well these assistants enhance engagement, persistence, and learning results, we will obtain empirical data regarding how adaptive affective feedback affects students over longer periods of time, allowing further refinement of both the recognition model and instructional methods controlled by the recognition model.

Declarations

Funding

This research received no external funding.

Conflict of Interest / Competing Interests

The authors declare that they have no conflict of interest.

Ethics Approval and Consent to Participate

Not applicable. The study used publicly available datasets (CREMA-D and RAVDESS), and no human subjects were directly involved.

Consent for Publication

Not applicable.

Data Availability

The datasets used in this study are publicly available:

- CREMA-D dataset
- RAVDESS dataset

Further processed data are available from the corresponding author upon reasonable request.

Materials Availability

Not applicable.

Code Availability

The code used to develop and evaluate the proposed model is available from the corresponding author upon reasonable request.

Author Contributions

Nanda Gopal Malladi and Vrisheeka Mulakala contributed equally to conceptualization, methodology, software development, data curation, formal analysis, validation, investigation, and visualization. Nanda Gopal Malladi and Vrisheeka Mulakala wrote the main manuscript text. Deepa N contributed to supervision, project administration, resources, and critical review and editing of the manuscript. All authors read and approved the final manuscript.

References

1. S. G. Rajesh, S. V. Madangarli, G. S. Pisharady, and R. Subrahmanyam, "Enhancement of Virtual Assistants Through Multimodal AI for Emotion Recognition," IEEE Access, vol. 13, no. 1, 2025. https://doi.org/10.1109/ACCESS.2025.3577664
2. J. Chen, Z. Shen, Y. Zhu, G. Xiong, and X. Liu, "A Deep Multi-task Learning Method for E-learning Behavior Analysis," CSIS-IAC, vol. 1, no. 1, 2025. https://doi.org/10.1109/CSIS-IAC65538.2025.11160758
3. J. Pancharia, H. Meena, and V. K. Singh, "Emotion-Aware Speech Recognition System Using Deep Neural Networks," ICCRTEE 2025, vol. 1, no. 1, 2025. https://doi.org/10.1109/ICCRTEE64519.2025.11053055
4. A. Cremonesi, "The Evolution of Edge AI: Contextual Awareness and Generative Intelligence," JSAP Symposium on VLSI and Circuits, vol. 1, no. 1, 2025.
5. A. Madireddy, P. Mathilakath, K. Jayakrishnan, and S. D. Saveetha, "NLP-Driven Audio Transcription and Contextual Summary Generation," GINOTECH 2025, vol. 1, no. 1, 2025. https://doi.org/10.1109/GINOTECH63460.2025.11076801
6. M. Hasan, M. Islam, M. M. Hassan, and M. S. Mia, "Enhanced Speech Emotion Recognition Using Hybrid ConvLSTM Model," QPAIN 2025, vol. 1, no. 1, 2025. https://doi.org/10.1109/QPAIN66474.2025.11171908
7. L. A. Vatsal, B. Tiwari, L. Upadhyay, and T. Sharma, "Fusion Techniques and Contextual Integration in Multimodal Sentiment Analysis," ICCTDC 2025, vol. 1, no. 1, 2025. https://doi.org/10.1109/ICCTDC64446.2025.11158728
8. S. Kuzhaloli et al., "Generative Adversarial Approaches for Cross-Lingual Emotion-Aware Speech Synthesis," ICIMIA 2025, vol. 1, no. 1, 2025. https://doi.org/10.1109/ICIMIA67127.2025.11200684
9. Q. Luo, "Development of College English Teaching System based on Artificial Intelligence Algorithm," ICDCECE 2025, vol. 1, no. 1, 2025. https://doi.org/10.1109/ICDCECE65353.2025.11035986

10. S. R. S. Subitsha, K. Saravanan, M. Vetrivel, and N. Shrimaulik, "NLP-Driven Learning Platform for Personalized Education," ICVADV 2025, vol. 1, no. 1, 2025. https://doi.org/10.1109/ICVADV63329.2025.10961687
11. S. Dhepe, D. Lothe, P. Gebise, and S. Pawar, "A Real-Time Monitoring System for User Attentivity on e-learning platforms," ICOECA 2025, 2025. https://doi.org/10.1109/ICOECA66273.2025.00149
12. Y. V. B. Reddy, K. S. Gopi, G. N. Sri, and B. U. Sravanthi, "Real-time Speech Emotion Detection in Employees using Multimodal Emotion Recognition," ISAC3 2025, 2025. https://doi.org/10.1109/ISAC364032.2025.11156447
13. N. B., P. V. R. Chintalapudi, S. Vekkot, and S. Kochuvila, "Emotion-Driven Conversational AI: Speech Recognition and Response with Emotional Intonation," ICICI 2025, vol. 1, no. 1, 2025. https://doi.org/10.1109/ICICI65870.2025.11069581
14. B. Singh, K. Kaur, and G. Kaur, "Optimizing Emotion Detection: An NLP-Driven Deep Learning Approach to Sentiment Encoding," ICDSBS 2025, vol. 1, no. 1, 2025. https://doi.org/10.1109/ICDSBS63635.2025.11031712
15. A. I. Septiana, K. Mutijarsa, B. L. Putro, and Y. Rosmansyah, "Emotion-Related Pedagogical Agent: A Systematic Literature Review," IEEE Access, vol. 12, no. 1, 2024. https://doi.org/10.1109/ACCESS.2024.3374376
16. L. Huang, H. Lan, Z. Sun, C. Shi, and T. Bai, "Emotional RAG: Enhancing Role-Playing Agents Through Emotional Retrieval," IEEE International Conference on Knowledge Graph (ICKG), vol. 1, no. 1, 2024. https://doi.org/10.1109/ICKG63256.2024.00023
17. B. H. Rao, M. Rengarajan, R. Changala, K. Santosh, and M. Aarif, "Hybrid AI Approach Combining Decision Trees and SVM for Intelligent Tutoring Systems in STEM Education," ICACCS 2024, vol. 1, no. 1, 2024. https://doi.org/10.1109/ICACCS60874.2024.10716888
18. S. Deshmukh, S. Chaudhary, M. Gayakwad, K. Kadam, N. S. More, and A. Bhosale, "Advances in Facial Emotion Recognition: Deep Learning Approaches and Future Prospects," MITADTSoCiCon 2024, 2024. https://doi.org/10.1109/MITADTSoCiCon60330.2024.10574908
19. P. Madan, D. Dixit, and S. J. Shobana, "Speech Emotion Recognition System Using MLP Classifier Algorithm," 2024 International Conference on Communication, Computing and Internet of Things (IC3IoT), 2024. https://doi.org/10.1109/IC3IoT60841.2024.10550312
20. B. Kumar, P. S. V. Reddy, P. N. Reddy, A. K. Varma, C. S. Reddy, and V. R. Kumbhare, "Performance Analysis of CNN-Based Speech Emotion Recognition System," 2024 First International Conference on Electronics, Communication and Signal Processing (ICECSP), 2024.
21. C. D. Kloos, P. Callejo-Pinardo, P. M. Moreno-Marcos, C. Alario-Hoyos, M. A. Hombrados-Herrera, M. Muñoz-Organero, I. Estévez-Ayres, P. J. Muñoz-Merino, and M. B. Ibáñez, "How Can Generative AI Support Education?," 2024 IEEE Global Engineering Education Conference (EDUCON), vol. 1, no. 1, 2024. https://doi.org/10.1109/EDUCON60312.2024.10578716
22. G. Sivasathiya, K. R., A. Kumar D., and H. R. A. R., "Emotion-Aware Multimedia Synthesis: A Generative AI Framework for Personalized Content Generation Based on User Sentiment Analysis," 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), vol. 1, no. 1, 2024. https://doi.org/10.1109/IDCIOT59759.2024.10467761
23. S. Kamada and T. Ichimura, "Multimodal Adaptive Structural Deep Belief Network for Emotion Recognition on RAVDESS Dataset," 2024 16th IIAI Interna-

tional Congress on Advanced Applied Informatics (IIAI-AAI), vol. 1, no. 1, 2024. https://doi.org/10.1109/IIAI-AAI63651.2024.00062

24. I. R. Ulgen, Z. Du, C. Busso, and B. Sisman, "Revealing Emotional Clusters in Speaker Embeddings: A Contrastive Learning Strategy for Speech Emotion Recognition," ICASSP 2024 – IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, no. 1, 2024. https://doi.org/10.1109/ICASSP48485.2024.10447060
25. J. Hu, Y. Huang, X. Hu, and Y. Xu, "The Acoustically Emotion-Aware Conversational Agent With Speech Emotion Recognition and Empathetic Responses," IEEE Transactions on Affective Computing, vol. 14, no. 1, 2023. https://doi.org/10.1109/TAFFC.2022.3205919