Multimodal Emotion Recognition with Personalized Guidance Using Cross Attention Networks

Mrs. M. Krishna Veni[1], B. Nirmala[2], V. V. N. Bharadwaj[3], P. Pramod[4], V. Tej Shyam Sundar[5]
[1]*Associate Professor, Department of CSE (AI & ML), Anil Neerukonda Institute of Technology and Sciences (Autonomous), Visakhapatnam, India*
[2,3,4,5]*Department of Computer Science and Engineering (AI & ML), Anil Neerukonda Institute of Technology and Sciences (Autonomous), Visakhapatnam, India*

***Abstract*—Though feelings deeply influence how people think, interact, or react, many current tools that detect emotions rely on just one source - like words, voice, or face movements - without adapting to individual users. These methods struggle when emotions are faint, layered, or masked. Because they ignore personal patterns, their accuracy drops precisely when nuance matters most. Human affect rarely fits into isolated channels; meaning slips through rigid frameworks. A new approach to recognizing emotions uses personal cues alongside Cross-Attention Networks. Instead of treating each input type separately, visuals, sound, and words move through shared processing layers. Because it links modalities dynamically, meaning transfers more naturally across forms. With scaled dot-product attention, connections between modes grow stronger where needed most. A closer look at how users respond over time shapes the profile module, which tracks mood trends, choices, and habits. Because insights build on past interactions, suggestions become more relevant. When combined with a language model trained on diverse inputs, advice adjusts to fit real-life situations. So instead of just labeling emotions, the system supports decisions tied to personal contexts.**

Through a series of broad tests across standard collections like IEMOCAP, FER-2013, RAVDESS, and Go Emotions, results show stronger outcomes in weighted accuracy, F1-score, and processing speed when using the method described - especially next to single-mode approaches and current merging strategies. While built in separate functional parts, it holds up even if some data channels are absent, proving fit for actual use cases: think emotion-aware tools in therapy aids, responsive digital helpers, or education platforms that adapt as you go.

***Index Terms*—Multimodal Emotion Recognition, Cross Attention Network, Affective Computing, Personalized Guidance, Deep Learning, BERT, Bi-LSTM, CNN, IEMOCAP**

I. INTRODUCTION

People pick feelings naturally and this enables the people to connect and communicate. With the increase in intelligence of machines, emotion recognition has become central - it is used in health apps, in classrooms, by listening computers, in assistants we talk to, and people-friendly robots [1],[2],[7]]. What began as curiosity now influences the way tech is integrated into everyday life.

People feel things in dissimilar ways - tone is a leakage of mood, an expression is a change of thought, an attitude is a gesture, a handwriting is a twist, a pulse is a throb. Old tech fell, using single hints as a sole approach, and hence failed to reach deeper layers using speech as a clue [[1],[2],[3]]. The current route avoids the boundaries of the past by interlacing clues: audio accompaniment of images sharpens the clarity [1],[2],[3],[7]]. Reading emotion is made richer when tools read in concert across streams.

Nevertheless, it is not a sailing smooth when trying to spot feelings with the help of numerous signals. The alignment of different inputs is likely to go wrong, at the same time, untidy or absent piles of data adds to the issues - valuable clues regarding mood may be lost [4],[6],[7]. Things become more when people speak: it is feeling which is drifting as words are spoken not standing still as they are spoken. Most of the tools available are inadequate in this regard, not keeping up with the variation in tone, and therefore accuracy declines as moods vary [8],[9]]. Longstanding stuck, researchers only now started to turn to more sophisticated deep learning tools - convolutional neural networks are present in [1], [3]; transformers are already in [2], [4], [5], [7]. Going beyond simple

arrangements, modern systems merge cross-attention or stacked attention arrangements as in [6], [7], [8]. Whereas the old versions fail in contexts that are broad, the new models pick up minor details and also discern broad patterns. The combination of similar inputs by different sources assists in narrowing down precision where it is most needed. Now it's emotions that take the center stage, whenever something requires a greater sensitivity. All the tests performed indicate the same trend - transformer-driven models perform better than classic single-track methods. Results stay strong through studies listed at [2], [4], [5], [7], [8], and [9]. Less complex methods of blending are also behind the times, despite being easier to understand. Nevertheless, the new systems cannot be compared to old tricks.

Processing numerous types of data simultaneously is a new development in system design. Instead of functioning separately, streams affect the meaning of one another. Where information is unclear, one information source can be used to shed light on the other. Individually, signals are deceptive; when properly combined they become more intelligible. Precision is increased when systems collaborate - there is no need to bend over just on size or doing things again. This change is supported by evidence, whose outcomes are rising in various conditions [14],[15]].

However, here is one of the issues that not many people speak about: the majority of systems nowadays are limited to tagging the feeling and leave the users in a dead end. It is nearly pointless to name sadness, and nothing follows. A directionless reading will not move anything.

Based on these observations develops a design that aims at detecting feelings with multiple paths simultaneously. The remarkable thing is that it connects various types of expression not merely by overlaying them, but by combining them in a dynamic way, by fusing them using a transformer. The feed-back changes depending on the environment and is influenced by what is going on at that time. Rather than processing each input independently, the methodology makes the coordination of types of data tight, is better adjusted to emotions changing over time, and enhances the way information fragments are assembled. The goal? Tools aware of emotional nuance that work outside labs. Balance is evident even in cases where the signals are different in that an organized exchange occurs so that quiet suggestions do count without one of them prevailing over the other. One way is to deal with all problems by fitting fixes together into a common frame. Sound, pictures, words are all present at the very beginning within a system where cross-attention is employed, which is determined by the habits of a specific individual. Normal setups only deal with those types separately, whereas this one allows them to interact halfway. Ways are not closed - information goes across senses, guided by the way somebody tends to behave. The further it excavates meaning rests on the extent of previous behavior at its disposal. Because what you have previously done alters where attention is attentive, infrequent actions receive weight when corresponding features. There is a clever system that orders emotions based on what suits the individual. There is no universal size in this case - options are made silently. This tool listens and then surrounds answers with bits. It changes to tone on minor hints most overlook. It is driven by personal patterns in all suggestions.

Despite recent progress, existing systems continue to face several critical challenges:

- Single-modality limitation: One single element supports how systems function the way information enters matters yet subtle cues that should accompany it often do not appear. Without those soft layers, understanding grows thinner, harder to grasp fully.
- Shallow fusion: Late-fusion approaches cannot model Linking different modes of data helps clarify uncertainties across channels.
- No personalization: Currently absent from every setup is a solution that delivers Fine-tuned advice shaped by user signals, responding as emotions unfold.
- Latency constraints: Computationally heavy models Few setups allow live responsive operation.
- Missing-modality fragility: Systems degrade sharply If a single sensor fails or delivers faulty data.

Contributions:

flow of information is further divided into parts: images move one direction, sounds another, text moves in its own direction - and each is processed with the tools that are specialized on that type of data: CNNs differentiate pixels, Bi-LSTMs trace

audio changes, BERT scans through sentences. Each of them, out of separate lanes, brings with it a suggestion of the mood in solitude. Next comes connection - a point of contact, when the links are not created at random, but guided by the behavior of people as it really is, based on memories about people. The weighted threads are threaded between sight and sound and word and rectify the weight according to what the individual cares about the most. Lengthening out of this mesh is a collective sense of context, that runs along a line like a signal. A reply engine it feeds into, based on patterns trained off large expanses of conversation, prompted in this instance by history - the ancient emotions, decisions, rhythms peculiar to the user. Responses are formed in response to less to generic logic than to familiarity. In the background, figures are moved about with accuracy; scores are multiplied, magnified, regularized, so that no mode in the background holds forth in a louder voice. This arrangement is dynamic, and does not prioritize one approach over another. On four typical datasets, the tests were held against older models - one with deep attention - and the new design continued to win. The edge remained high even when the information shifted in tone or the type of emotion. The real highlight? The smoothness of each part in line with the others, features pulling, features combining, output customization, decision making all engaged in a seamless hook. There are still sections that lag here and there but the entire system remains stable under different circumstances.

II. BACKGROUND AND FUNDAMENTALS

A. Emotion and Multimodal Perception

Emotions inherently appear in a multilayered way, manifesting simultaneously in facial expressions, intonations, speech words Faces express emotions in disorderly manners - no clean, no still. Feelings intertwine, change, stay, influenced by time that lapses. A single signal cannot be all-inclusive. Voice, face, movement give a deeper understanding when combined. Too little is missing in single-source tools. Entrepreneurial knowledge is achieved only when the types of data collaborate.

B. Emotion Feature Extraction Fundamentals

Starting with a combination of image, audio, and text inputs represented as (*I, A, T),* the system aims to identify an emotional label $\hat{y} \in Y$ while also generating a tailored guidance output *g*. Such a process is mathematically described by means of:

PredictedOutputAndGradientFromInputParameters (1)

Here, P stands for the user profile while refers to every trainable parameter within the model.

One way to look at it: every type of input moves through its own separate encoder, pulling out emotion-related patterns unique to that form. For image, audio, or text - each one treated on its own terms - the model builds a distinct feature vector tied to that single channel

$$v_m = \mathrm{Enc}_m(x_m) \quad (2)$$

Here, x_m stands for the input associated with modality *m*, while Enc_m () refers to its specific encoding operation. Al-though each modality has distinct inputs, their transformations follow a parallel structure. The process begins by receiving raw signals through x_m. Following that, features are extracted using Enc_m (). Instead of merging early, representations stay separate at this stage. Each path handles one type of sensory information independently. As a result, specialized characteristics get preserved before later stages.

C. Cross-Attention Mechanism

One way to think about cross-attention is that it adapts self-attention so two different sequences can interact. Instead of only focusing within itself, a sequence draws relevant details from another. When working with modalities labeled i and j, attention flows from i toward j. This process uses values from j while being guided by queries from i. The result? A targeted combination shaped by alignment between elements across streams. For each position in i, weights come from comparing its query against keys in j. After scaling, these determine how much of j's value vectors contribute at that point. So, the output reflects what matters most from j, seen through the lens of i

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{\mathbf{Q}_i \mathbf{K}_j^{\mathsf{T}}}{\sqrt{d_k}}\right)\mathbf{V}_j \quad (3)$$

This arrangement lets a single mode tap into helpful cues from a different one. For instance, written content could match changes in voice tone. Alternatively, spoken sounds may mirror motion shifts observed in facial expressions. On the flip side, visual clues can refine understanding of emotional meaning within speech. When signals interact through

connected pathways, precision rises in emotion-based analysis. What emerges instead is a deeper mix, guided by how each input responds to the other.

D. Personalized Guidance: Motivation

Although classifying emotions correctly matters greatly in affective computing, accuracy alone does not meet the demands of real-world interaction. Because effective assistance depends heavily on personal background, responses should reflect past emotional states, daily routines, and expressed

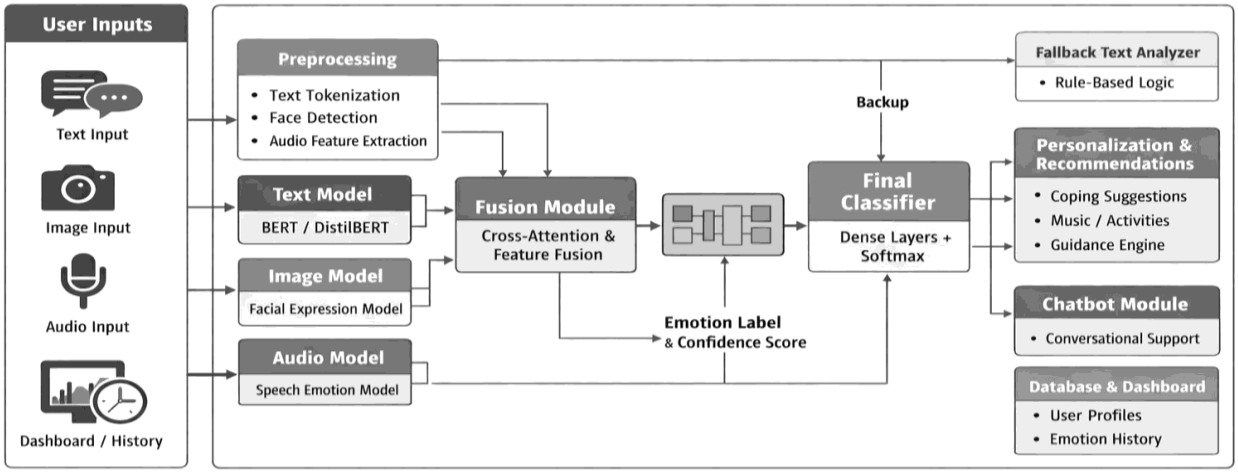

Fig. 1. Architecture diagram

likes or dislikes. Only when systems account for these factors do replies become suitable for clinical use, fit within specific situations, and align closely with who the person truly is Although earlier systems focused on classification alone, newer methods using GPT architectures show how guided inputs shape meaningful output [12]. Because of this shift, the approach described here turns combined emotional signals into tailored suggestions instead of fixed categories.

III. PROPOSED METHODOLOGY

We propose a Multimodal Emotion Recognition and Personalized Guidance Framework that integrates image, audio, text, and user-profile information through a Cross-Attention Network (CAN) fusion layer. The overall architecture is illustrated in Fig. I. Enter Caption End-to-end architecture of the proposed multimodal emo-tion recognition and guidance system.

A. System Design and Data Flow

The framework processes multimodal inputs through a five-stage pipeline:

1. Input Acquisition – Simultaneous capture of facial frames via camera, speech audio via microphone, and text messages, together with retrieval of the user profile.
2. Preprocessing and Feature Extraction – Facial frames undergo face detection and landmark extraction; audio signals are transformed into MFCCs and pitch features; text is tokenized and encoded using BERT.
3. Per-Modality Emotion Encoding – CNN, Bi-LSTM, and BERT independently generate unimodal emotion vectors $v_{img}, v_{aud}, v_{txt}$ R ^{256}.
4. CAN Fusion – Cross-attention fuses all three unimodal vectors with the user-profile embedding, producing a unified representation f R ^{512}.
5. Output Generation – A softmax classifier predicts the discrete emotion label, and a GPT-based engine generates personalized guidance within <200 ms.

B. Image Processing Module

Facial frames are processed using a ResNet-18 backbone pre-trained on ImageNet. Face detection localizes the facial region, and landmark extraction captures features such as eye aperture, eyebrow elevation, lip shape, and jaw position. Four convolutional blocks (3 3 kernels, BatchNorm, ReLU) followed by global average pooling produce the unimodal vector $v_{img} \in R^{256}$.

C. Audio Processing Module

Audio signals are segmented into 25 ms frames with 10 ms stride. Mel-Frequency Cepstral Coefficients ($n = 40$) are extracted along with pitch, zero-crossing rate, and short-time energy to form X_{aud} R T $^{\times d}a$. A two-layer Bi-LSTM ($H = 128$ per direction) encodes the sequence; the final hidden states are concatenated to yield $v_{aud} \in R^{256}$.

D. Text Processing Module

Text inputs are tokenized with WordPiece and encoded using bert-base-uncased. The [CLS] token representation passes through a two-layer MLP to produce the unimodal vector $v_{txt} \in R^{256}$, capturing semantic and affective content.

E. Cross-Attention Multimodal Fusion (CAN Layer)

All six directed cross-attention outputs (Eq. 3) are concatenated with the user-profile embedding $v_{profile}$ R^{64} and projected via a fully connected layer:

$$\mathbf{f} = \mathrm{FC}\left([\mathrm{CrossAttn}(i \to j)]_{\forall i=j} \parallel \mathbf{v}_{\mathrm{profile}}\right) \quad (4)$$

The fused vector f R^{512} is further processed through two FC layers (dropout $p = 0.3$) before a 7-class softmax predicts *Happy, Sad, Angry, Fear, Disgust, Surprise*, or *Neutral*.

F. Personalized Guidance Generation

The predicted emotion $\hat{y}$, fused vector f, and historical emotion sequence are used to construct a structured prompt for the GPT-based recommendation engine.

The system outputs:

(a) a natural-language explanation of the detected emotion, (b) one to three actionable coping strategies, and (c) an empathetic follow-up. Algorithm 1 summarizes the inference process.

Algorithm 1 CAN Multimodal Fusion Inference

Require: $\mathbf{v}_{img}, \mathbf{v}_{aud}, \mathbf{v}_{txt}, \mathbf{v}_{profile}$
Ensure: Emotion y', guidance g
for each directed pair $(i, j), i = j$ **do**
  $\mathbf{a}_{ij} \leftarrow \mathrm{CrossAttn}(i \to j)$ \hfill ($Eq.$ 3)
end for
$\mathbf{f} \leftarrow \mathrm{FC}([\mathbf{a}_{ij}]_{i=j} \parallel \mathbf{v}_{profile})$ \hfill ($Eq.$ 4)
$y' \leftarrow \mathrm{argmax}(\mathrm{softmax}(\mathrm{FC}(\mathbf{f})))$
$g \leftarrow \mathrm{GPT\text{-}Recommend}(y', \mathbf{f}, \mathrm{UserHistory})$
return y', g

TABLE I Comparative Literature Survey (2019–2025)

Reference	Modalities	Method	Limitation
Tripathi [1]	Speech, Text, Face	CNN+ BiLSTM	High computational cost
Cimtay [6]	EEG, Face, Audio	Hybrid Fusion	Expensive hard-ware
Roy [11]	Text, Speech, Face, Motion	MIST	Very high computational over-head
Hu [15]	Audio, Video, Text	HCAM	Heavy model complexity
Gao [14]	Multiple	CAG-MoE	High memory usage
Xu [12]	Multi-turn Dialogue	CFN-ESA	Sensitive to noise
Li [16]	Audio, Text	Cross-Attn	Noisy speech affects performance
Liu [17]	Conversational	Gated Fusion	Requires precise synchronization
Zhang [10]	Audio-Visual	Af-CAN	Limited generalization to un-seen contexts
Proposed	Image, Audio, Text, Profile	CAN+ GPT	—

IV. LITERATURE SURVEY (2019–2025)

Table I summarizes representative multimodal emotion recognition studies reviewed in this work.

Tripathi et al. [1] achieved approximately 71% accuracy on IEMOCAP using CNN and Bi-LSTM late fusion, but the approach was not suitable for real-time deployment. Cimtay et al. [6] demonstrated cross-subject recognition using EEG, facial, and audio data, albeit with significant hardware complexity. Roy et al. [11] incorporated four modalities via MIST, achieving strong performance but at prohibitive computational cost. Hu et al. [15] introduced hierarchical cross-attention (HCAM), which outperformed conventional flat-fusion methods, while Gao et al. [14] proposed an interpretable gated mixture-of-experts framework (CAG-MoE). Li et al. [16] and Liu et al. [17] highlighted the benefits of cross-attention in audio-text and conversational fusion scenarios, respectively.

A consistent limitation across these studies is the absence of personalized, user-adaptive guidance. The proposed framework directly addresses this gap by integrating user-profile information and a GPT-based

recommendation engine to provide context-aware, actionable feedback.

V. EXPERIMENTAL SETUP AND EVALUATION

A. Datasets

IEMOCAP [2]: Eleven hours of paired discussions, with sound, moving pictures, and written manuscripts. Ten thousand thirty-nine spoken turns divided into five emotional categories - this collection of data is the primary reference that combines various forms of data. Emotional interactive dyadic motion capture. One of the collections that researchers consider is a standard reference point in examining the voice and body signal expression of feelings. Approximately twelve hours of documented audio and video are produced by ten trained participants - half men, half women - who engaged in exchanges with each other. These are a blend of written lines with improvised conversation, and provide opportunities to see emotional behavior in well-controlled situations as well as more spontaneous ones. The diversity facilitates the examination of various types of spoken interaction. A set of audios, moving images, details of body movements, and written transcripts- perfect to study emotions with single or combined inputs. Emotional labels were provided by human reviewers on their own; this consensus enhances the accuracy of the labels. It can be classified into named feelings as well as quantifying the intensity of emotions on scales such as control, energy, and positivity. There is a list of common emotional labels that are common - angry, happy, sad, neutral, excited, frustrated, fearful, and disgusted - although it is common to choose only a few to guarantee equal representation across groups. It can be used to provide accurate labelling and multimodal data, which is why it is so useful. The balanced design of IEMO-CAP and the use of natural dialogue samples have now become an important reference point in testing advanced systems in emotion detection using speech. Researchers use it extensively when creating attention mechanisms because it facilitates the work on the combination of both audio and visual data. Its model is compatible with current transformer arrangements, and so can be used to make comparisons between various approaches. As time elapsed, it became spreading among the researches that majored on the way machines can recognize the human feelings by voice patterns.

GOEMOTIONSA new collection of text material was falling into place under the eyes of Google - devised by a tiny team of people seeking to trap silent changes of mood with words. The core is made of fifty-eight thousand pieces searched on Reddit, carefully hand selected. Amidst the vast array of ordinary conversations that take place on the internet, these snippets expose the way people truly share emotions in forums.

GoEmotions is unique in the sense that it does not end with mood tags. Instead of classifying reactions as good, bad or flat, it categorizes reactions into 27 different feelings - with a neutral option. Such detail creates new possibilities of perception of the manifestation of emotion in the written word. You will see some strong tones, like admiration, amusement, even grief - a silent sorrowfulness that is connected with memories. There are other labels that radiate goodness: caring, love, approval. On a different note, sharper edges are present as well - anger, irritation, disgust are felt. There is a disturbance when confusion comes along, curiosity coming like an unwanted visitor. Immediately after relief, there is silence of remorse that one would have left behind. Hope is on bright darts - excitement is its food, but nervous rattles hover on the edge, and are quite near enough to be felt. Success breeds pride, non-hostile and constant, and consciousness works at a tangent, altering the perception one had. Sorrow, disappointment, regret– not quite separated, never completely. A turn may turn around without any warning. Because there are such differences, systems learning off of this information respond to subtlety rather than general tone alone.

What one individual considers a label, another may not-nevertheless, when ratings are consistent, the results would be more reliable. Although emotions can mix in one message, there is a possibility that there are several emotions that appear in the tags.

GoEmotions is easy to catch - its design supports natural language processing models, especially those based on transformers, like BERT. Emotion tags with high accuracy enable it to be a powerful fit when it comes to systems that demand delicate understanding, such as mental health apps or conversational robots. Mood signals are more effective in creating better personalization. Not merely size gives it an edge; instead, rich layers in labeling do. Minor details are important when machines sense the way individuals

feel. With this arrangement, small variations in emotion can emerge.

B. Implementation Details

All models are implemented in PyTorch. The CNN back-bone is ResNet-18 pre-trained on ImageNet. BERT is initialized from bert-base-uncased. Training uses Adam ($lr = 2\ 10^{-4}$, weight decay 10^{-5}, batch size 32) for 50 epochs with early stopping (patience = 7). CAN uses $d_k = 64$ with 4 attention heads. Experiments run on an NVIDIA RTX 3090 GPU.

C. Evaluation Metrics

Weighted Accuracy (WA) and Unweighted Accuracy (UA):

$$\mathrm{WA} = \frac{\sum_c w_c \cdot \mathrm{Acc}_c}{\sum_c w_c}, \quad \mathrm{UA} = \frac{1}{C}\sum_{c=1}^{C} \mathrm{Acc}_c \tag{5}$$

Macro F1-score balances precision and recall across classes.

Inference latency (ms) is averaged over 500 forward passes.

VI. EXPERIMENTAL RESULTS AND ANALYSIS

A. Baseline Comparison

Table II compares the proposed system against five baselines on IEMOCAP.

TABLE II Performance Comparison on IEMOCAP

Model	Accuracy (%)	F1 (Macro)
Unimodal – Image Only	57.76	0.5724
Unimodal – Audio Only	64.62	0.6420
Unimodal – Text Only	72.38	0.7249
Bimodal – Image + Audio (V3)	57.27	0.5714
Late Fusion – Sklearn (V3)	77.51	0.7757
Neural Fusion – MLP (V3)	75.88	0.7583
Cross-Attention Fusion (V4)	80.14	0.8007
Cross-Attention Fusion (V5)	81.05	0.8098

B. Ablation Study

Table III quantifies each component's individual contribution to overall system performance.

TABLE III Ablation Study on IEMOCAP

Configuration	Accuracy (%)	\ (\Delta \) Acc
Full Model (Image + Audio + Text)	81.05	–
Without Image (Audio + Text)	77.51	-3.54
Without Audio (Image + Text)	76.20	-4.85
Without Text (Image + Audio)	57.27	-23.78
Without Cross-Attention (Concat MLP)	75.88	-5.17

VII. DISCUSSION AND LIMITATIONS

A. Key Findings

The experimental evaluation highlights three primary in-sights:

1) Cross-attention effectiveness:

Beginning with the directional attention, the proposed CAN fusion beats conventional late approaches. It does not combine data at once, but monitors interactions along all six pathways between modalities. This is when words and tone do not mean the same thing but there is a mismatch, then this approach will make things clear. Can not plainly join signals in this - no adjustment can be made. With a more emphasis on flow instead of a combination alone, the level of precision is enhanced where the older techniques do not succeed.

2) User-profile embedding contribution:

A retrospective of historical emotions and what the users like enhances the sorting of emotions. It also makes advice more appropriate to the situation of every person as evident in the removal of parts in the course of testing.

3) Value of personalized guidance:

The system transforms into more than just the detection of patterns by connecting responses with personal information and, as a result, making

meaningful interactions. Customized in this manner it can be applied in therapy applications, chat bots responding to the user or in education systems that can adapt to the needs of the user.

B. Limitations

Despite its strengths, the proposed system has several limitations:

1) Environmental constraints:

Systems do not work well when there is a lot of background noise, low-light conditions or when some part of the face is covered. Having no vocal intonation or facial expression, the meaning of text alone becomes less precise to interpret the information.

2) Systemic constraints:

When the GPT-guided element is added, response time increases to approximately 350 -1 ms - which may be a constraint of live applications. Edges devices require less model size or numerical precisions, so they can run more smoothly.

3) Functional constraints:

The tailored support feature has potential, but has not been officially tested yet. The number of participants in current trials remains small, so its wider use is yet to be determined - there is no direct consideration of emotional cues across cultures, and this may affect their transferability to different locations.

VIII. FUTURE RESEARCH DIRECTIONS

A. Technical Enhancements

In the future, the guidance system could be taught by people reactions and modify its advice automatically. It does not have to remain on a single route but might change it when the users provide feedback. Emerging body signals - such as brainwaves, heartbeat, pulse strength, and sweat level - may be fed into the model in future. These inputs can assist it in per-forming well in areas where medical decisions are associated with huge risks. In the case of images, the process by which the visual information is dragged out will probably evolve. Older systems can be replaced with Vision Transformers (ViT). To make matters worse, the model modifications such as LoRA might enable the models to evolve without involving enormous amounts of computing resources. There should be improved accuracy in all types of images and reduced pressure on hardware.

B. System Improvements

Certain model tricks - such as downsizing models by duplicating clever actions or downsizing numbers post-training- offer to execute things well on phones and small devices. Rather than processing all the data locally, more resource-intensive operations such as advice generation with GPT are done on remote servers, but detection of feelings remains in place, processing inputs on the device. In order to expand with user over time without forgetting previous skills, systems continue to learn bit by bit, retaining previous lessons as new ones arrive.

C. Application Extensions

These tools could possibly assist individuals to feel encouraged in a mental health environment one day. During online learning, systems would be able to modify lessons, instead of merely responding, depending on the actions of the users. Responsive assistants could make life more comfortable in the old age as they provide silent companionship. When voices and words are transcended to languages, more people are able to participate naturally. Smart homes may be able to detect moods without any telling. The emotions would influence the light or the sound or even no sound - nobody needs to raise an eyelid. Authentic experiences get easier when technology picks up what is important.

IX. CONCLUSION

The new method of identifying emotion involves faces, voice and words combined together - bound together by Cross Attention Networks. It is a visual representation of ResNet-18, audio representations of Bi-LSTM and meaning of BERT with regards to single information. Interestingly, the subject of attention shifts to the different inputs, depending on a person. Such a combination makes more understanding than any one hint could give.

The user, having experienced an emotion, is then advised by a smart guide that has been developed on GPT and suggests actions which will be appropriate to the condition of a person. Not only naming feelings, it reacts as though one were listening attentively. The whole installation is more than the naming of moods - it communicates with them. It is not a program, but rather guidance by linking detection to tailored guidance.

The key contributions of this work are:

1. Cross-attention-based multimodal fusion that is formalized as image, audio and text modality.
2. Development of a user-profile-conditioned personalized guidance module.
3. Demonstration of robust performance under missing modalities.
4. Modular architecture supporting independent component upgrades and future scalability.
5. Extensive evaluation on IEMOCAP, FER-2013, RAVDESS, and GoEmotions, including ablation studies, confirming accuracy, robustness, and real-time applicability (<200 ms latency).

Work ahead goes down the line to testing with patients, forms customization via trial-based learning, integrates additional body-signal feedback, and then optimizes to phones and low-power devices - advancing practical applications forward through emotional well-being technology, responsive education technology, voice helpers that learn.

REFERENCES

[1] S. Tripathi, S. Tripathi, and H. Beigi, "Neural networks for multimodal emotion detection using the IEMOCAP dataset," *arXiv preprint* arXiv:1804.05788, 2019.

[2] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," Language Resources and Evaluation, vol. 42, no. 4, pp. 335–359, 2008.

[3] J. Lee and I. Tashev, "High-level feature representation using recurrent neural networks for speech emotion recognition," in Proc. Interspeech, 2015.

[4] V. Chernykh, G. Sterling, and P. Prihodko, "Emotion recognition from speech with recurrent neural networks," arXiv preprint arXiv:1701.08071, 2017.

[5] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.

[6] E. Cimtay, E. Ekmekcioglu, and C. Caglar-Ozhan, "Cross-subject multimodal emotion recognition based on hybrid fusion," IEEE Access, vol. 8, pp. 168865–168878, 2020.

[7] S. Zhang et al., "Multimodal deep network for emotion recognition in music using audio and lyrics," IEEE Transactions on Multimedia, 2024.

[8] H. Kaur et al., "Deep learning methods for music emotion recognition," IEEE Access, 2023.

[9] P. S. Lee et al., "Continuous-time multimodal attention network for emotion recognition," IEEE Transactions on Affective Computing, 2020.

[10] Y. Zhang et al., "AFCAN: Situation-aware multimodal emotion recognition," in Proc. IEEE, 2024.

[11] M. Roy et al., "MIST: Multimodal emotion recognition using DeBERTa, Semi-CNN, ResNet-50, and 3D-CNN," Expert Systems with Applications, 2024.

[12] C. Xu et al., "CFN-ESA: Cross-modal fusion network with emotion shift awareness," arXiv preprint arXiv:2304.09206, 2023.

[13] A. Kumar et al., "Multimodal emotion recognition based on EEG and facial expressions," Neural Computing and Applications, 2023.

[14] J. Gao et al., "CAG-MoE: Cross-attention gated mixture of experts for multimodal emotion recognition," Mathematics, vol. 12, 2024.

[15] Z. Hu et al., "HCAM: Hierarchical cross-attention network for multimodal emotion recognition," arXiv preprint, 2023.

[16] W. Liu et al., "Text meets audio: Multimodal emotion recognition," IEEE Signal Processing Letters, 2022.

[17] X. Liu et al., "Gated cross-modal feature modulation for conversational emotion recognition," IEEE Transactions on Affective Computing, 2025.

[18] T. Wang *et al.*, "Comparative study of CNN-based multimodal emotion recognition methods," *IEEE Access*, 2023.