Quality-Controlled Multimodal Emotion Recognition in Conversations with Identity-Based Transfer Learning and MAMBA Fusion

Zanxu Wang[1], Homayoon Beigi[1,2]
[1]Columbia University, New York, USA
[2]Recognition Technologies, Inc., New York, USA
[1]zw2864@columbia.edu, [2]beigi@recotechnologies.com

Abstract—**This paper addresses data quality issues in multimodal emotion recognition in conversation (MERC) through systematic quality control and multi-stage transfer learning. We implement a quality control pipeline for MELD and IEMOCAP datasets that validates speaker identity, audio-text alignment, and face detection. We leverage transfer learning from speaker and face recognition, assuming that identity-discriminative embeddings capture not only stable acoustic and facial traits but also person-specific patterns of emotional expression. We employ RecoMadeEasy® engines for extracting 512-dimensional speaker and face embeddings, fine-tune MPNet-v2 for emotion-aware text representations, and adapt these features through emotion-specific MLPs trained on unimodal datasets. MAMBA-based trimodal fusion achieves 64.8% accuracy on MELD and 74.3% on IEMOCAP. These results show that combining identity-based audio and visual embeddings with emotion-tuned text representations on a quality-controlled subset of data yields consistent competitive performance for multimodal emotion recognition in conversation and provides a basis for further improvement on challenging, low-frequency emotion classes.**

Index Terms—**Multimodal Emotion Recognition, Data Quality, Transfer Learning, State Space Model, Modality Fusion**

I. Introduction

Multimodal emotion recognition in conversation (MERC) leverages complementary information from text, audio, and visual modalities to identify emotional states in dialogue contexts [1], [2]. While recent architectures demonstrate incremental improvements in MERC performance, such as MGCMA [3] and MMGAT-EMO [4], fundamental data quality issues in benchmark datasets remain unaddressed.

Large-scale conversational emotion datasets such as MELD [5] serve as primary benchmarks for MERC research. However, these datasets exhibit systematic quality issues that propagate through modeling pipelines: speaker labels conflate distinct individuals (e.g., "Waiter" assigned to multiple actors), audio-text misalignment occurs due to imprecise video segmentation, and face tracking often captures non-speaking individuals or fails entirely. Despite these issues affecting a substantial portion of samples, few studies systematically address data curation.

Additionally, beyond data quality, most MERC approaches rely on generic pre-trained models for each modality that lack emotion-specific discriminability. While numerous unimodal emotion datasets exist with prototypical emotional expressions—acted speech in RAVDESS [6], posed facial expressions in CK+ [7], emotion-labeled text in GoEmotions [8]—their potential to inform multimodal models through transfer learning remains largely unexplored.

The selection of appropriate pre-trained models for transfer learning is critical. We leverage the RecoMadeEasy® speaker and face recognition models built by Recognition Technologies, Inc. [9], based on the hypothesis that identity-preserving representations inherently capture fine-grained acoustic and facial patterns that correlate with emotional expression. Speaker embeddings [15] encode prosodic variations, voice quality, and temporal dynamics—all crucial for emotion recognition. Similarly, face recognition models must learn subtle facial muscle movements and expression dynamics to distinguish individuals, providing rich features for emotion classification.

Finally, fusion strategies predominantly employ attention-based cross-modal transformers [10], [11] with quadratic complexity, raising questions about efficiency in practical settings. Recent efficient sequence models like MAMBA [12] offer linear complexity but have not been widely used for multimodal emotion fusion.

This paper addresses these challenges through an integrated methodology with three core contributions:

1. Systematic Data Quality Control: We develop a systematic quality control pipeline for MELD and IEMOCAP that resolves speaker disambiguation, audio-text alignment verification, and face-speaker validation.

2. Emotion-Specific Multi-Stage Transfer Learning: We propose leveraging various auxiliary unimodal emotion datasets to inject emotion-discriminative information into each modality's representation before fusion. This multi-stage approach uses emotion-specific text datasets to fine-tune language models, vocal emotion corpora to adapt speaker embeddings, and facial expression databases to condition face

features.

3. Efficient Fusion with State Space Model: We demonstrate that a single MAMBA block achieves competitive performance with linear complexity, and report empirical results on quality-controlled data subsets while identifying remaining challenges for future investigation.

II. RELATED WORK

A. Transfer Learning in Speech Emotion Recognition

Transfer learning has emerged as a dominant paradigm for addressing data scarcity in emotion recognition. Recent work by Jakubec et al. [13] demonstrates that speaker embedding models (d-vector, x-vector, r-vector) pre-trained on large speaker verification [15] datasets significantly improve speech emotion recognition (SER). The advantage lies in these models' ability to capture speaker-specific acoustic patterns including prosodic variations and voice quality characteristics that correlate with emotional states [14].

Padi et al. [14] proposed leveraging ResNet-based models trained on speaker recognition [15] tasks, incorporating statistics pooling layers to handle variable-length inputs. Their approach eliminates the need for sequence truncation commonly used in SER systems, achieving competitive results on IEMOCAP. This supports our hypothesis that speaker identity models capture emotion-relevant acoustic features.

B. Face Recognition for Emotion Transfer

The relationship between face recognition and emotion recognition has been extensively studied. Knyazev et al. [16] demonstrated that industry-level face recognition networks pre-trained on large-scale datasets (e.g., VGG-Face2) outperform emotion-specific models when fine-tuned for facial emotion recognition (FER). This finding suggests that learning to distinguish facial identities requires capturing subtle muscular movements and expression dynamics that directly transfer to emotion recognition.

Ngo and Yoon [17] showed that an SE-ResNet-50 model pre-trained on the VGG-Face2 database, combined with a novel weighted-cluster loss function, effectively transfers high-level facial features to FER tasks. The success of face recognition models in emotion tasks stems from their need to be invariant to expressions while still encoding them—creating representations that separate identity from emotion while preserving both.

C. Sentence Transformers for Emotion-Aware Text

While BERT and its variants dominate text emotion recognition, recent advances in sentence transformers offer superior semantic representations. MPNet [18] combines the strengths of BERT's bidirectional context and XLNet's autoregressive modeling, addressing the limitations of masked language modeling through permuted language modeling. The all-mpnet-base-v2 model, fine-tuned on over 1 billion sentence pairs using contrastive learning [19], produces 768-dimensional embeddings that capture nuanced semantic relationships.

The choice of MPNet-v2 over alternatives is motivated by several factors: (1) the contrastive training objective naturally aligns with emotion recognition's need to distinguish subtle semantic differences at utterance level; (2) extensive evaluation shows MPNet-based models achieve state-of-the-art performance on semantic textual similarity tasks [19], [20], which correlate with emotion understanding.

D. Quality Issues in MERC Datasets

Despite widespread use of MELD and IEMOCAP, few studies systematically address their quality issues. Busso et al. [21] acknowledged annotation challenges in the original IEMOCAP paper, noting inter-annotator agreement issues for ambiguous emotions. Recent work highlights that fear and disgust categories in MELD barely exceed chance-level recognition even with state-of-the-art models [22], suggesting the possibility of fundamental data problems rather than model limitations.

The lack of systematic quality control might lead to inflated performance claims. Models trained on corrupted data may learn spurious correlations by using audio-text misalignment as unintended features. Our work aims to narrow this gap through comprehensive validation of modality alignment and speaker identity.

III. METHODOLOGY

A. Problem Formulation

Given a conversational sequence $U = [u_1, u_2, \ldots, u_N]$ with N utterances and corresponding speakers $S = [s_1, s_2, \ldots, s_N]$, each utterance u_i comprises three modalities: text t_i, audio a_i, and visual (face) v_i. The goal is to predict emotion label $y_i \in \mathcal{Y}$ for each utterance, where $|\mathcal{Y}| = 7$ for MELD and $|\mathcal{Y}| = 4$ for IEMOCAP.

B. Multi-Stage Emotion Representation Learning

Our proposed framework (Figure 1) consists of a hierarchical multimodal pipeline across three stages.

1. Foundation Embedding Extraction

We begin by extracting unimodal embeddings from pre-trained speaker recognition and face recognition models, as well as a fine-tuned MPNet-v2 model, as foundation embeddings for each modality.

a) **Speaker Embeddings**

We leverage the RecoMadeEasy® speaker recognition engine [9] to extract embeddings from audio. Specifically, we pass both unimodal audio datasets (e.g., CREMA-D and RAVDESS) and audio from MELD and IEMOCAP to the engine, obtaining 512-dimensional speaker embeddings per utterance.

b) **Face Embeddings**

For visual features, extracted facial images from MELD, IEMOCAP, and additional unimodal facial expression datasets (e.g., CK+ and RAF-DB) are processed through the RecoMadeEasy® face recognition engine to obtain 512-dimensional face embeddings.
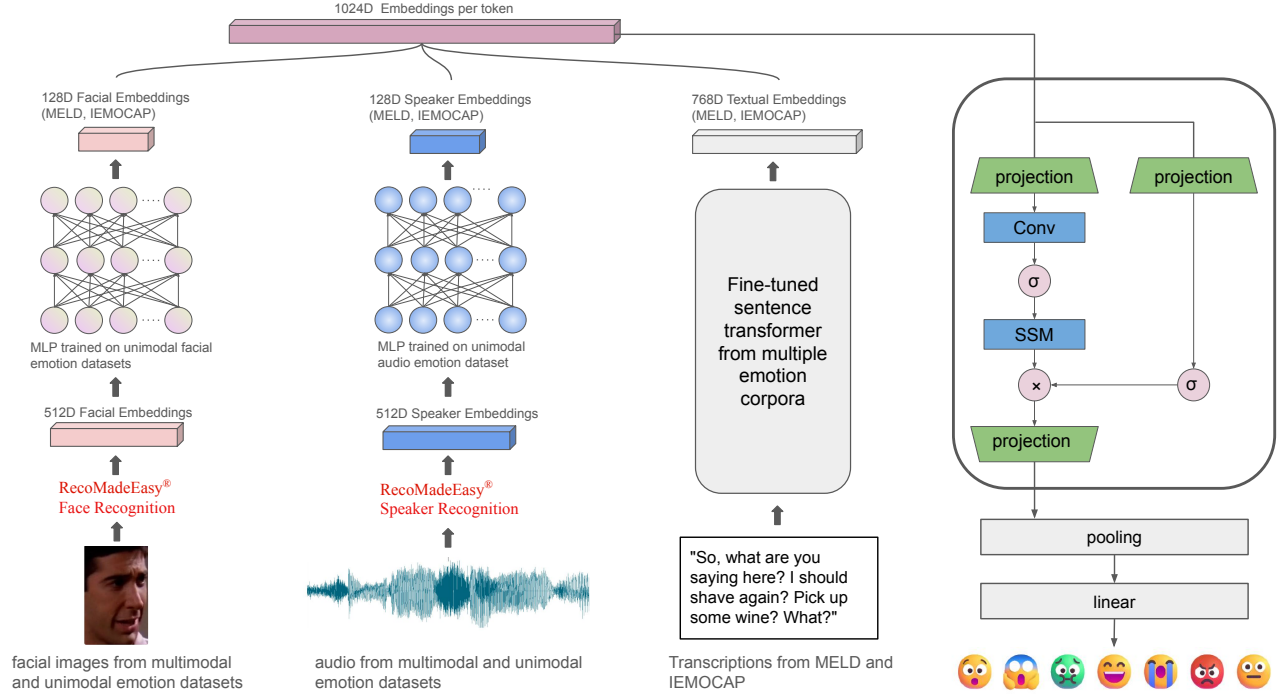
c) **Textual Embeddings**

Fig. 1. Overview of the proposed three-stage quality-controlled multimodal emotion recognition pipeline. Stage 1 extracts foundation embeddings from RecoMadeEasy® speaker and face recognition engines [9] and a fine-tuned MPNet-based sentence transformer for MELD and IEMOCAP utterances. Stage 2 adapts the 512-dimensional audio and visual features into 128-dimensional emotion embeddings using modality-specific MLPs trained on unimodal emotion datasets. Stage 3 fuses 768-dimensional text, 128-dimensional speaker, and 128-dimensional facial embeddings into 1024-dimensional token representations and applies a single MAMBA block, pooling, and a linear classification head for utterance-level emotion prediction.

Textual embeddings are derived from the `sentence-transformers/all-mpnet-base-v2` model, which we fine-tune for seven-class emotion classification (surprise, fear, disgust, happiness, sadness, anger, neutral) using a supervised learning setup with balanced sampling from multiple emotion corpora. A classification head with two linear layers is added on top of the pretrained MPNet encoder, with the final layer containing seven neurons, one for each emotion class. The input text is tokenized using the MPNet tokenizer with padding to a maximum length and truncation enabled. Training is performed for three epochs using the AdamW optimizer with a learning rate of 2×10^{-5}, a batch size of 16, and weight decay of 0.01. We save the best checkpoint based on validation loss at the end of each epoch. This fine-tuning procedure yields a training accuracy of 90.08% and a test accuracy of 83.41% on the aggregated emotion corpora. After training, we discard the classification head and use the 768-dimensional pooled output of the encoder as the textual embedding for each utterance. The transcripts of IEMOCAP and MELD dialogues are passed through this fine-tuned MPNet model to produce 768-dimensional textual embeddings that encode both semantic and emotional content.

2. Modality-Specific Adaptation

Once we obtain the initial 512-dimensional embeddings for the audio (speaker) and facial (image) modalities from the Stage 1 feature extraction process, we focus on adapting these embeddings to better reflect emotional content. Our goal is to transform the raw modality-specific features into representations that are sensitive to emotional patterns.

Dedicated unimodal emotion datasets contain clearer, more prototypical expressions of each emotion (for example, exaggerated vocal tones in RAVDESS and posed facial expressions in CK+) compared to natural conversational data, where emotions are often subtle or mixed. This transfer learning strategy teaches the model what each emotion looks and sounds like before applying it to noisy multimodal conversational context.

To achieve this, we train feedforward multi-layer perceptron (MLP) classifiers on unimodal emotion recognition datasets, each containing single-modality emotional cues. For facial expression recognition, we train an MLP on 512-dimensional face embeddings extracted from static facial images for seven-class emotion classification. The MLP architecture consists of three fully connected layers with dimensions $512 \rightarrow 256 \rightarrow 128 \rightarrow 7$, with batch normalization and ReLU activation after each hidden layer. Dropout with a rate of 0.3 is applied after each hidden layer to reduce overfitting. The model is trained only on non-MELD datasets (including CK+ and RAF-DB) using the Adam optimizer with an initial learning rate of 1×10^{-3} and weight decay of 1×10^{-5}. The learning rate is adjusted

by a ReduceLROnPlateau scheduler with a factor of 0.1 and patience of 5 epochs. Training uses cross-entropy loss with a batch size of 32 for up to 100 epochs, with early stopping triggered after 10 epochs without improvement in validation F1-score. This setup achieves about 82.5% validation accuracy on the non-MELD facial expression test set.

We apply the same MLP architecture, hyperparameters, and training procedure to 512-dimensional audio embeddings derived from emotion speech datasets (CREMA-D, RAVDESS, SAVEE, and TESS). This audio MLP reaches 69% validation accuracy on the non-MELD audio test sets.

For both audio and face, we use the 128-dimensional penultimate layer of the MLP as the transformed emotion embedding. These 128-dimensional vectors serve as the final unimodal representations used in the subsequent multimodal fusion stage, as they capture the emotion-relevant structure learned from unimodal data while remaining compact.

3. Multimodal Fusion for Classification

We employ a single MAMBA block ($d_{\text{state}} = 64$, expand_factor $= 2$, $d_{\text{conv}} = 4$) followed by mean pooling and a linear classification head for utterance-level emotion recognition. The model's input dimension (d_{model}) varies based on the modality combination: 1024-dimensional for text+speaker+face, 768-dimensional for text-only, and other configurations for different multimodal setups. Each input utterance is represented as a variable-length sequence of concatenated multimodal embeddings, with padding and attention masking applied to handle sequences of different lengths. We train on MELD and IEMOCAP training sets separately using AdamW optimizer with a learning rate of 1×10^{-5}, batch size 32, and early stopping (patience 10 epochs) based on validation weighted F1. For MELD (7 emotion classes), we address severe class imbalance using weighted cross-entropy loss with class weights of [15.0, 15.0, 6.0, 1.0, 3.0, 6.0, 4.0] for surprise, fear, disgust, happiness, sadness, anger, and neutral respectively, combined with label smoothing of 0.2. For IEMOCAP (4 emotion classes with excitement merged into happiness), we use standard cross-entropy loss with the same label smoothing of 0.2.

a) **Text-only:** We feed each token's 768-dimensional embedding sequentially into the model.

b) **Face-only:** We feed each video frame's 128-dimensional reduced feature vector into the model.

c) **Text + Face Fusion:** For a given utterance, we first determine the correspondence between text tokens and video frames. With 15 tokens and 4 frames, we match the first 4 tokens to frame 1, the next 4 to frame 2, the next 4 to frame 3, and the final 3 to frame 4. Each token's 768-dimensional embedding is concatenated with its corresponding frame's 128-dimensional face representation, yielding an 896-dimensional vector per token.

d) **Speaker Embedding Integration:** When including speaker information, we concatenate the same speaker embedding vector for the entire utterance alongside the text and/or face features. This maintains a consistent speaker representation within the fused input sequence.

C. Datasets

We employ a comprehensive collection of multimodal and unimodal emotion datasets for our experiments. The multimodal datasets (MELD and IEMOCAP) serve as our target evaluation benchmarks, while the unimodal datasets are used for fine-tuning MPNet-v2 and training the face and speaker MLPs in Stage 2 of our framework. Figure 2 illustrates the distribution of emotion categories across all auxiliary unimodal datasets, demonstrating our balanced sampling strategy across modalities.
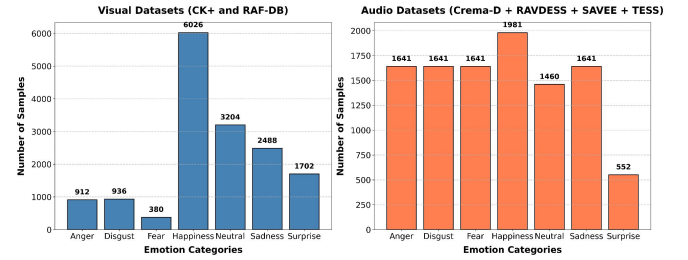
Text Datasets: Blanched sampling from 5 sources with each emotion 6.2k utterances.

Name	anger	disgust	fear	joy	neutral	sadness	surprise
Crowdflower (2016)	Yes	-	-	Yes	Yes	Yes	Yes
Emotion Dataset, Elvis et al. (2018)	Yes	-	Yes	Yes	-	Yes	Yes
GoEmotions, Demszky et al. (2020)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
ISEAR, Vikash (2018)	Yes	Yes	Yes	Yes	-	Yes	-
SemEval-2018, El-reg, Mohammad et al. (2018)	Yes	-	Yes	Yes	-	Yes	-

Fig. 2. Distribution of emotion categories across auxiliary unimodal datasets. **Top left:** Visual datasets (CK+ and RAF-DB) showing 15,648 total samples. **Top right:** Audio datasets (CREMA-D, RAVDESS, SAVEE, and TESS) with 10,557 samples. **Bottom:** Text dataset composition showing balanced sampling from five sources (CrowdFlower, CARER, GoEmotions, ISEAR, and SemEval-2018) totaling 6.2k utterances across seven emotion categories.

1. Target Conversational Datasets

MELD (Multimodal EmotionLines Dataset) [5] is a large-scale multimodal dataset derived from TV show dialogues, providing synchronized audio, video and text data for more than 13,000 utterances in 1,400 dialogues. It includes seven emotion categories: anger, disgust, fear, joy, neutral, sadness, and surprise. The dataset's focus on conversational context makes it particularly suitable for emotion recognition in multi-turn dialogues, where emotions vary dynamically across speakers and modalities.

IEMOCAP (Interactive Emotional Dyadic Motion Capture Database) [21] is a widely used multimodal dataset containing approximately 12 hours of audio, video, and text from both scripted and improvised dyadic conversations between actors. It encompasses nine emotion categories: angry, fearful, disappointed, happy, sad, surprised, excited, frustrated, and neutral. The dataset provides detailed facial expressions, vocal signals, and transcriptions, making it ideal for comprehensive multimodal emotion analysis.

2. Auxiliary Unimodal Emotion Datasets

To enhance modality-specific representations, we leverage several established unimodal emotion datasets across visual, audio, and text modalities. As shown in Figure 2, we apply balanced sampling strategies to ensure comparable representation across emotion categories within each modality.

A. Speaker Facing Away (Face Undetectable)

dia207_utt9 (MELD train)   dia572_utt6 (MELD train)

B. Off-screen Speaker (Phone Call Scenes)

dia184_utt1 (MELD train)   dia9_utt2 (MELD train)   dia8_utt0 (MELD train)

C. Successful Speaker Identification

Multiple subjects present; correct speaker identified via Facenet-512 and yolo-v8 (frames at 4.5s, 5.5s, 6.5s in dia443_utt7 from MELD training set). Green boxes indicate matched speaker.

D. Audio-Text Misalignment Detected

Whisper Text: Yeah, it really has been great, too.
Original Text: Yeah!
Cosine Similarity: 0.3011
Levenshtein Similarity: 0.1111

dia948_utt7 (MELD train)

Fig. 3. Examples of quality control challenges and solutions in MELD dataset processing. **(A) Speaker Facing Away:** Utterances where the speaker's face is not visible (red box indicates detection failure) are removed to ensure reliable facial expression analysis. **(B) Off-screen Speaker:** Phone call scenes where the speaking character is not visible on camera are filtered out. **(C) Successful Speaker Identification:** Multi-person scenes where YOLOv8 detects multiple faces and Facenet-512 correctly identifies the target speaker (green boxes) based on embedding similarity across temporal frames (4.5s, 5.5s, 6.5s). **(D) Audio-Text Misalignment:** Example of detected misalignment between original transcript ("Yeah!") and Whisper transcription ("Yeah, it really has been great, too."), identified through low cosine similarity (0.30) and Levenshtein similarity (0.11) scores.

a) **Facial Expression Datasets**

CK+ (Extended Cohn-Kanade Dataset) [7] contains 593 video sequences from 123 subjects displaying posed facial expressions across seven emotions. Each sequence begins with a neutral expression and progressively intensifies to a peak expression, providing valuable data for expression dynamics.

RAF-DB (Real-World Affective Faces Database) [23] is a large-scale dataset comprising approximately 30,000 crowd-labeled facial images covering both basic emotions (anger, disgust, fear, happiness, sadness, surprise, neutral) and compound emotions (e.g., happily surprised, sadly fearful). Its real-world collection conditions enhance model generalization.

b) **Vocal Expression Datasets**

CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) [24] includes over 7,000 audio-visual recordings from 91 actors expressing six emotions: anger, disgust, fear, happiness, neutral, and sadness. The dataset's diversity in actor demographics provides robust vocal expression patterns.

RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) [6] contains recordings from 24 professional actors producing both speech and song across eight emotions, including calm, happy, sad, angry, fearful, surprised, and disgusted, with two intensity levels (normal and strong).

TESS (Toronto Emotional Speech Set) [25] comprises 2,800 audio clips from two female speakers aged 26 and 64, each expressing seven emotions. The age diversity makes it particularly valuable for studying vocal emotion cue variations across different age groups.

SAVEE (Surrey Audio-Visual Expressed Emotion Database) [26] includes recordings of four male actors expressing seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. The dataset provides high-quality audio recordings suitable for acoustic feature extraction.

c) **Textual Emotion Datasets**

CrowdFlower (Emotion in Text Dataset, 2016) is a corpus of 40,000 tweets labeled with 13 emotion categories (e.g., anger, joy, sadness, fear, enthusiasm, etc.) via crowd-sourcing. We map these labels into our seven-class taxonomy and sample 6.2k utterances to balance with other sources.

CARER (Contextual Affect Dataset) [27] covers five basic emotions (anger, fear, joy, sadness, and surprise) with sentences collected from diverse domains, providing varied linguistic expressions of emotion.

GoEmotions [8] is a large-scale dataset of 58,000 Reddit comments labeled with 27 fine-grained emotion categories. We map its labels to our seven-emotion taxonomy, leveraging its rich emotional diversity and conversational context.

ISEAR (International Survey on Emotion Antecedents and Reactions) [28] contains self-reported emotional experiences across multiple emotions (anger, disgust, fear, joy, sadness) from participants in 37 countries. The dataset includes contextual descriptions of emotion triggers, providing valuable semantic information.

SemEval-2018 Task 1 (Affect in Tweets) [29] provides emotion intensity ratings for tweets in four emotions (anger, fear, joy, sadness) and multi-label classification for 11 emotions. We discretize the intensity ratings into categorical labels and select tweets covering our seven-emotion set.

We standardize all datasets to a unified seven-emotion taxonomy (anger, disgust, fear, joy/happiness, neutral, sadness, surprise) where applicable. As illustrated in Figure 2, our text modality employs balanced sampling from five sources, resulting in approximately 6.2k utterances per emotion. The visual datasets (CK+ and RAF-DB) provide 15,648 samples, while the audio datasets (CREMA-D, RAVDESS, TESS, and

SAVEE) contribute 10,557 samples.

D. Data Quality Control Pipeline

Our quality control pipeline ensures the integrity, quality, and modality alignment of data from the MELD and IEMO-CAP datasets. Figure 3 illustrates key challenges and solutions in our data cleaning process, including speaker identification in multi-person scenes, handling off-screen speakers, and detecting audio-text misalignments. The objective is to generate a clean, speaker-specific, and temporally aligned corpus of audiovisual utterances suitable for emotion recognition.

1. Face Extraction from MELD.

a) **Database Construction for Unique Speakers:** MELD contains over 300 distinct speakers, but in many cases different individuals share identical speaker labels. To address this ambiguity, we created a speaker database with corrected annotations:

- We re-labeled speakers who shared the same name by appending a dialogue-specific ID, resulting in 352 uniquely identified speakers.
- Entries with multiple speakers (e.g., "All," "Guys," or "Phoebe and Rachel") were removed to ensure that each utterance corresponded to a single speaker.

b) **Sampling and Face Detection:** For each of the 352 speakers, we sampled up to 15 video clips. We used the DeepFace library with a YOLOv8 backend to detect faces as follows:

- Extract one frame per second, starting from 0.5 seconds into the utterance.
- Save every detected face in each frame to the corresponding speaker's folder.

All cropped face images were inspected to remove incorrect detections and misaligned utterances (Figure 3A-B):

- Utterances where the speaker never appeared (e.g., speaker off-screen or on the phone) were discarded.
- Utterances where the video and the transcription were not aligned were removed.

From these steps, we eliminated 182 utterances.

c) **Ensuring Audio-Text Alignment in MELD:** MELD's original video cropping sometimes resulted in misalignments between audio and text, especially for short utterances (1–2 seconds long). To address this, we re-transcribed the audio using the Whisper API and compared the newly generated transcriptions to the original MELD annotations. Ninety-one utterances were removed due to extremely short or inaudible audio, where Whisper yielded empty transcriptions. To detect more subtle misalignments (Figure 3D), we employed two metrics:

- **Cosine Similarity (Textual Embeddings):** Identified utterances with semantically different transcriptions.
- **Levenshtein Similarity:** Captured utterances that differed significantly in length, indicating misalignment.

We passed each utterance's Whisper transcription and original transcription to the MPNet-v2 model, extracted the sentence embeddings, and calculated cosine similarity scores. As shown in Figure 4, we set a threshold of 0.25 for cosine similarity and 0.3 for normalized Levenshtein similarity. Only utterances passing both thresholds were retained. The distribution reveals that most utterances exhibit high alignment (cosine similarity ¿ 0.9), while the filtering process effectively removes outliers with significant mismatches.
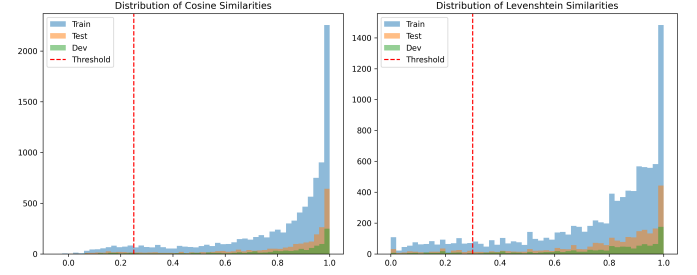

Fig. 4. Distribution of audio-text alignment metrics across MELD splits. **Left:** Cosine similarity between original and Whisper-generated transcriptions using MPNet-v2 embeddings. **Right:** Levenshtein similarity measuring character-level differences. Red dashed lines indicate filtering thresholds (0.25 for cosine similarity, 0.3 for Levenshtein). Most utterances show high alignment, while removed samples cluster near low similarity scores.

d) **Face Extraction Based on Speaker Identity:** We used the Facenet-512 model within DeepFace to generate embeddings for each speaker's sample images. By averaging these embeddings, we obtained a unique 512-dimensional representation per speaker. For each utterance, we re-detected faces using YOLOv8 and compared them to the speaker's stored embedding (Figure 3C):

- Embeddings were generated from each newly detected face (starting from 0.5 seconds into the utterance, one frame per second).
- We computed cosine similarity scores with a threshold of 0.3 to match each face to the speaker's reference embedding.
- If multiple faces were detected, we selected the one with the highest similarity. If no face passed the threshold, we applied an offset search (looking slightly before or after the initial frame) to find a valid match.

This step further filtered out utterances where no suitable face was found. The resulting subset of MELD contains utterances where all three modalities (text, audio, and video) were successfully aligned and the correct speaker's face was detected.

2. Face Extraction for IEMOCAP.

For the IEMOCAP dataset, face extraction was simpler due to fewer speakers (10 actors across 5 sessions). We applied a similar methodology using DeepFace, YOLOv8, and Facenet-512 with a cosine similarity threshold of 0.3, starting from 0.5 seconds into each utterance. Any utterances without a detected face were discarded. After this process, we retained 7,309 utterances from IEMOCAP that included all three modalities.

3. Audio Channel Selection for MELD.

MELD's original audio was recorded using a microphone array, resulting in multiple audio channels per utterance. For consistency and quality:

- From the 6-channel recordings, we selected the third channel.
- For the 2-channel recordings, we chose the channel with the highest energy level.

This ensured a standardized and relatively clean audio input across all utterances. In the end, our experiments relied on subsets of MELD and IEMOCAP where all three modalities were successfully captured and passed our quality control processes. Table I summarizes the dataset composition before and after quality control.

TABLE I
MELD AND IEMOCAP COMPOSITION BEFORE AND AFTER QUALITY CONTROL

Dataset	Split	Original	Verified
MELD	Train	9,989	7,443
	Dev	1,109	835
	Test	2,610	1,966
	Total	13,708	10,244
IEMOCAP	Total	10,039	7,309

IV. RESULTS

A. Overall Performance

Table II presents emotion recognition performance across modality configurations. On MELD, the trimodal system (T+V+A) achieves 64.8% accuracy and 64.3% weighted F1-score. Text emerges as the strongest unimodal baseline at 58.1%, substantially outperforming vision at 42.3%. The poor performance of visual features indicates IEMOCAP shows similar patterns with trimodal fusion reaching 74.3% accuracy. This suggest there might be temporal misalignment between the emotional peaks for the facial expressions and the extracted frames in our pipelines.

TABLE II
EMOTION RECOGNITION PERFORMANCE WITH FINAL FUSION BY MODALITY CONFIGURATION

Dataset	Modalities	Accuracy	W-F1
MELD	T	58.1%	57.9%
	V	42.3%	41.8%
	T+A	62.5%	62.1%
	T+V	61.2%	60.9%
	V+A	48.7%	48.2%
	T+V+A	**64.8%**	**64.3%**
IEMOCAP	T	66.2%	66.0%
	V	52.1%	51.7%
	T+A	72.9%	72.6%
	T+V	71.6%	71.3%
	V+A	58.3%	57.9%
	T+V+A	**74.3%**	**74.1%**

B. Class-wise Analysis

Figure 5 displays confusion matrices for trimodal fusion. On MELD, neutral achieves 81.9% accuracy while fear (16.7%) and disgust (29.0%) remain problematic. Fear shows substantial confusion with disgust (18.8%) and anger (17.7%), suggesting these emotions share overlapping multimodal patterns

in conversational contexts. IEMOCAP's four-class problem yields more balanced performance: neutral (74.5%), sadness (80.0%), anger (72.3%), and happy+excited (70.8%).
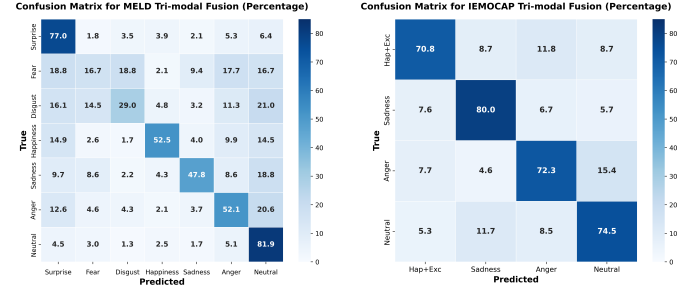

Fig. 5. Confusion matrix for trimodal fusion on MELD test set (7 classes) and IEMOCAP test set (4 classes).

V. DISCUSSION

Our choice to leverage speaker and face recognition models for emotion transfer departs from conventional approaches, grounded in the hypothesis that identity-preserving representations inherently capture emotion-relevant patterns. Speaker recognition models encode prosodic variations and voice quality while face recognition systems learn subtle facial muscle configurations—both crucial for emotion expression. Similarly, we select MPNet-v2 as a sentence encoder that combines contrastive learning, masked and permuted language modeling, and can be fine-tuned on textual emotion corpora and then used to extract emotion-aware embeddings for conversational datasets such as MELD and IEMOCAP.

The filtered utterances from benchmark data like MELD reveal systematic quality issues, with audio-text misalignment representing most of removed MELD samples. These misalignments could create spurious training signals where models learn incorrect cross-modal associations. The dominance of text (58.1%) versus poor visual performance (42.3%) in MELD challenges multimodal fusion assumptions, stemming from compressed TV footage quality, subtle conversational expressions, and temporal misalignment between emotional peaks and extracted frames. The limited fusion improvement (6.7% over text-only) questions whether current datasets truly require multimodal processing.

This work faces several limitations. We implement quality control without extensive ablation or statistical significance testing, potentially missing systematic biases in our filtering process. The distribution of removed samples across speakers, emotions, and dialogue positions remains unanalyzed. Most critically, our use of fixed speaker embeddings fails to capture temporal dynamics inherent in emotional speech—speaker embeddings provide static representations while emotions evolve over time within utterances.

Future work should address these limitations through several directions. We plan to incorporate pre-trained speech recognition models such as wav2vec 2.0 [30] to enhance the audio component, as these models provide frame-level

features that capture temporal emotional evolution unlike static speaker embeddings. This temporal richness would enable better utilization of MAMBA's state-space modeling capabilities, which currently process fixed-dimensional embeddings rather than exploiting sequential speech dynamics. Additionally, developing quality-aware training methods that utilize imperfect samples rather than discarding them could recover the filtered data. We also aim to explore self-supervised approaches learning from unlabeled conversational data and investigate continuous emotion dimensions as alternatives to problematic categorical schemes. Creating new datasets with rigorous quality control during acquisition, rather than post-hoc filtering, remains essential for advancing the field.

VI. CONCLUSION

We implement a systematic quality control pipeline for MERC datasets that addresses multiple problematic scenarios—speaker mislabeling, audio-text desynchronization, and face detection failures—ensuring proper alignment across all three modalities. Our approach leveraging identity-based transfer learning achieves 64.8% accuracy on MELD and 74.3% on IEMOCAP. However, significant challenges remain: sparse emotions achieve near-random accuracy, and fusion provides limited improvements over text-only baselines. These results emphasize that progress in conversational emotion recognition requires addressing fundamental data quality issues alongside architectural innovations. The field must move beyond incremental model improvements to confront the systematic data problems that limit current approaches, prioritizing the creation of properly aligned, high-quality multimodal datasets and developing methods robust to the inherent ambiguity of emotional expression in natural conversation.

REFERENCES

[1] Amith Ananthram, Kailash Karthik Saravanakumar, Jessica Huynh, and Homayoon Beigi, "Multi-Modal Emotion Detection with Transfer Learning," *Natural Language, Dialog and Speech Symposium (NDS2020)*, New York City, Nov. 13, 2020.

[2] M. Pantic and L. J. M. Rothkrantz, "Affective multimodal human-computer interaction," in *Proc. ACM Int. Conf. Multimedia*, 2005, pp. 669–676.

[3] X. Wang, S. Zhao, H. Sun, H. Wang, J. Zhou, and Y. Qin, "Enhancing Multimodal Emotion Recognition through Multi-Granularity Cross-Modal Alignment," in *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, 2025, pp. 1–5, doi: 10.1109/ICASSP49660.2025.10889156.

[4] C. Zhang, Y. Liu, and B. Cheng, "A MoE Multimodal Graph Attention Network Framework for Multimodal Emotion Recognition," in *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, 2025, pp. 1–5, doi: 10.1109/ICASSP49660.2025.10890390.

[5] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. ACL*, 2019, pp. 527–536.

[6] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," *PLoS One*, vol. 13, no. 5, 2018.

[7] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. CVPR Workshops*, 2010.

[8] D. Demszky, D. Movshovitz-Attias, J. Ko *et al.*, "GoEmotions: A dataset of fine-grained emotions," in *Proc. ACL*, 2020.

[9] Recognition Technologies, Inc., RecoMadeEasy® Speaker and Face Recognition, https://www.recotechnologies.com, Accessed 2025-11-18.

[10] Y. H. M. S. Siriwardhana, P. D. H. D. De Silva, S. M. M. Kaluarachchi, and S. A. Dharmasena, "Multimodal emotion recognition with transformer-based self-supervised feature fusion," *IEEE Access*, vol. 8, pp. 176274–176285, 2020.

[11] W. Rahman, M. K. Hasan, S. Lee *et al.*, "Integrating multimodal information in large pretrained transformers," in *Proc. ACL*, 2020, pp. 2359–2369.

[12] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv:2312.00752*, 2023.

[13] M. Jakubec, E. Lieskovska, R. Jarina, M. Spisiak, and P. Kasak, "Speech emotion recognition using transfer learning: Integration of advanced speaker embeddings and image recognition models," *Applied Sciences*, vol. 14, no. 21, p. 9981, 2024.

[14] S. Padi, S. O. Sadjadi, D. Manocha, and R. D. Sriram, "Multimodal emotion recognition using transfer learning from speaker recognition and BERT-based models," *arXiv:2202.08974*, 2022.

[15] H. Beigi, *Fundamentals of Speaker Recognition*, Springer-Verlag, New York, 2011, ISBN: 978-0-387-77591-3, http://www.fundamentalsofspeakerrecognition.org.

[16] B. Knyazev, R. Shvetsov, N. Efremova, and A. Kuharenko, "Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video," *arXiv:1711.04598*, 2017.

[17] Q. T. Ngo and S. Yoon, "Facial expression recognition based on weighted-cluster loss and deep transfer learning using a highly imbalanced dataset," *Sensors*, vol. 20, no. 9, Art. 2639, 2020.

[18] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "MPNet: Masked and permuted pre-training for language understanding," in *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020.

[19] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. EMNLP-IJCNLP*, 2019, pp. 3982–3992.

[20] N. Reimers and I. Gurevych, "Sentence Transformers for semantic textual similarity tasks," *arXiv:2104.08821*, 2021.

[21] C. Busso, M. Bulut, C.-C. Lee *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[22] T. Meng, Y. Shou, W. Ai, N. Yin, and K. Li, "Deep Imbalanced Learning for Multimodal Emotion Recognition in Conversations," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 12, pp. 6472–6487, Dec. 2024, doi: 10.1109/TAI.2024.3445325.

[23] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. CVPR*, 2017, pp. 2852–2861.

[24] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[25] K. Dupuis and M. K. Pichora-Fuller, "Toronto emotional speech set (TESS)," University of Toronto, Psychology Department, 2010.

[26] S. Haq and P. J. B. Jackson, "Speaker-dependent audio-visual emotion recognition," in *Proc. Int. Conf. Auditory-Visual Speech Process.*, 2008, pp. 53–58.

[27] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen, "CARER: Contextualized affect representations for emotion recognition," in *Proc. EMNLP*, 2018, pp. 3687–3697.

[28] K. R. Scherer and H. G. Wallbott, "Evidence for universality and cultural variation of differential emotion response patterning," *J. Personality and Social Psychology*, vol. 66, no. 2, pp. 310–328, 1994.

[29] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "SemEval-2018 Task 1: Affect in Tweets," in *Proc. SemEval*, 2018, pp. 1–17.

[30] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020.