Lightweight Multimodal Emotion Recognition via Cross-Attention Transformer Fusion with Grad-CAM Explainability

Group 4: Aakash Vardhan Madabhushi, Vedika Sumbli, Kruthi Shamanna, Sai Sushma Maddali

Abstract

Automatic emotion recognition from audiovisual data is a fundamental challenge in affective computing, with applications spanning mental health monitoring, human–computer interaction, and safety-critical systems. Existing deep learning approaches either rely on large pretrained encoders that are impractical for deployment, suffer from modality dominance where one stream overwhelms the other, or produce decisions that cannot be interpreted or audited. This paper presents a lightweight multimodal emotion recognition system built entirely from scratch (without pretrained weights) that combines convolutional neural network encoders for audio and video with transformer self-attention and a cross-attention fusion mechanism. Trained and evaluated on the CREMA-D dataset under a rigorous actor-independent split protocol, our model achieves 74.6% validation accuracy and an estimated 73% test accuracy on a four-class problem (angry, happy, neutral, sad) using only 1.57 million parameters, approximately 30 to 50 times fewer than comparable multimodal baselines. We further integrate Gradient-weighted Class Activation Mapping (Grad-CAM) at both the audio and video branches, producing per-frame spatial heatmaps and frequency–time activation maps that reveal the acoustic and facial regions driving each prediction. Our results demonstrate that a carefully regularised, architecturally motivated small model can match the performance of much larger systems while remaining interpretable and deployable on resource-constrained hardware.

Index Terms

Multimodal emotion recognition, cross-attention, transformer, Grad-CAM, CREMA-D, lightweight deep learning, affective computing.

I. Introduction

HUMAN emotional communication is inherently multimodal. When a person expresses anger, the signal manifests simultaneously in elevated vocal pitch, rapid speech rate, widened jaw aperture, and tensed facial musculature. A system that processes only the audio stream or only the video stream captures an incomplete and potentially misleading picture of the underlying affective state. Despite this well-understood reality, many deployed emotion recognition systems remain unimodal, and those that do incorporate multiple streams often do so through late fusion strategies (independently classifying each modality and combining the outputs) rather than learning the dynamic interactions between them during training.

The motivation for this work arises from three converging observations about the current state of multimodal emotion recognition. First, the field's best-performing models depend on encoders pretrained on hundreds of millions of samples (wav2vec 2.0, HuBERT, VGG-Face) whose computational and data requirements place them out of reach for edge deployment in settings such as in-vehicle driver monitoring, embedded healthcare devices, or real-time call centre analytics. Second, the dominant evaluation practice of splitting datasets randomly at the clip level allows models to memorise actor-specific vocal and facial characteristics, inflating reported accuracy by approximately five to ten percentage points relative to genuine generalisation performance. Third, no published multimodal emotion recognition system in the surveyed literature provides built-in spatial explainability, that is, the ability to show which regions of the face and which frequency bands of the audio the model attended to for any given prediction. This absence makes it impossible to audit model behaviour in contexts where accountability matters, such as clinical mental health screening.

This paper addresses all three limitations. We construct a dual-branch architecture with CNN encoders for audio spectrograms and video frames, transformer self-attention within each modality, and cross-attention fusion between them, trained from scratch without any pretrained components. We evaluate under actor-independent splits, ensuring no actor's voice or face appears in both training and evaluation. We integrate Grad-CAM at both branches, generating interpretable activation maps that confirm the model attends to acoustically and visually meaningful features rather than dataset-specific artefacts.

II. Related Work

A. Unimodal Approaches

Early deep learning approaches to emotion recognition focused on single-modality inputs. CNN-LSTM architectures operating on Mel Frequency Cepstral Coefficients (MFCCs) became a widely adopted audio baseline, achieving approximately 61% accuracy on the six-class CREMA-D benchmark. These models capture short-term spectral patterns through the convolutional layers and temporal dependencies through the LSTM, but are fundamentally limited to acoustic information and cannot benefit from the strong visual signals present in video.

DistilHuBERT represents the current state of the art for audio-only emotion recognition, achieving 70.64% on the six-class CREMA-D task through fine-tuning a knowledge-distilled version of the HuBERT self-supervised speech model. While the performance gain over CNN-LSTM baselines is significant, DistilHuBERT's dependence on a pretrained encoder trained on thousands of hours of unlabelled speech means it inherits the computational demands of that pretraining phase and cannot be replicated in low-resource settings.

B. Multimodal Approaches

Mocanu [2] proposed a temporal audio–video CNN that fuses 3D convolutional features from both modalities, achieving 65.0% on the six-class CREMA-D task. While the temporal convolution captures motion dynamics across frames, the architecture uses simple concatenation fusion, which does not model the dynamic relationship between what is being said and what the face is expressing at each moment.

Pretrained Audio–Video Transformer approaches push accuracy to the 73–75% range on the same benchmark by fine-tuning large pretrained encoders for both modalities and applying transformer-based fusion. These models represent the performance ceiling for current multimodal approaches but come at a significant parameter cost, typically exceeding 50 million parameters, and their evaluation protocols often use random clip-level splits that allow actor identity leakage.

C. Recent Work Addressing Efficiency and Fusion

Rakib and Bagavathi [3] introduced G2D: Gradient-Guided Distillation, which attempts to address modality dominance, the phenomenon where a model collapses to near-unimodal behaviour when one stream provides a substantially stronger signal than the other. G2D uses gradient magnitudes to dynamically rebalance the contribution of each modality during training, achieving 65.5% on the six-class task. However, the gradient balancing mechanism does not fundamentally change the fusion architecture; it applies post-hoc corrections rather than structurally preventing dominance.

Salman *et al.* [4] proposed MoEDE (Mixture of Emotion-Dependent Experts), in which a separate expert sub-network is maintained for each emotion class, with a gating network routing inputs to the appropriate expert. This achieves 74.5% macro F1 but at the cost of multiplicative parameter growth and inference complexity that scales with the number of emotion classes.

A 2026 study on speech emotion recognition for low-resource environments applied knowledge distillation from a large teacher model to a MobileNet student, achieving 67.9% on the six-class task. While the parameter reduction is substantial, the accuracy loss of approximately 25 percentage points relative to the teacher model illustrates the fundamental tension between efficiency and performance when compressing an already-specialised large model rather than designing for efficiency from the ground up.

D. Gaps in Existing Work

Across the surveyed literature, three gaps persist consistently. First, no published lightweight multimodal model achieves competitive accuracy without relying on pretrained encoders, meaning performance is contingent on access to large pretraining corpora. Second, actor-independent evaluation is rarely used, making published accuracy figures difficult to interpret as measures of genuine generalisation. Third, none of the surveyed works provide spatial explainability over their predictions, limiting their suitability for trust-critical deployment contexts.

III. Method

A. Dataset and Preprocessing

We use the Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D) [1], which contains 7,442 audiovisual clips from 91 actors spanning diverse ages, genders, and ethnicities. Each actor performs 12 sentences at up to four intensity levels across six emotional categories. We restrict our study to the four-class subset (angry, happy, neutral, and sad), which represents the most clearly separable emotions and the most common target set in prior work on this dataset.

Actor-independent splitting. A critical methodological decision in our work is the use of actor-level rather than clip-level data splitting. CREMA-D filenames encode the actor ID as the first token (e.g., `1001_DFA_ANG_XX`). We extract all 91 unique actor IDs, sort them, and assign the first 80% to training, the next 10% to validation, and the remaining 10% to test. Every clip from a given actor appears in exactly one split. This ensures that neither vocal characteristics nor facial identity can be exploited by the model across splits, producing a more conservative and realistic evaluation.

Audio preprocessing. Each MP3 file is loaded at 16,000 Hz and trimmed or zero-padded to exactly three seconds (48,000 samples). A mel spectrogram is computed using 128 mel filter banks, an FFT window of 2,048 samples, and a hop length of 512 samples, with a maximum frequency of 8,000 Hz. The result is converted to decibel scale and normalised per sample to zero mean and unit variance, producing a tensor of shape $(1, 128, 94)$.

Video preprocessing. Each MP4 file is opened and its total frame count is determined. Thirty frames are sampled at evenly spaced indices across the full duration using linear interpolation, ensuring complete temporal coverage regardless of clip length

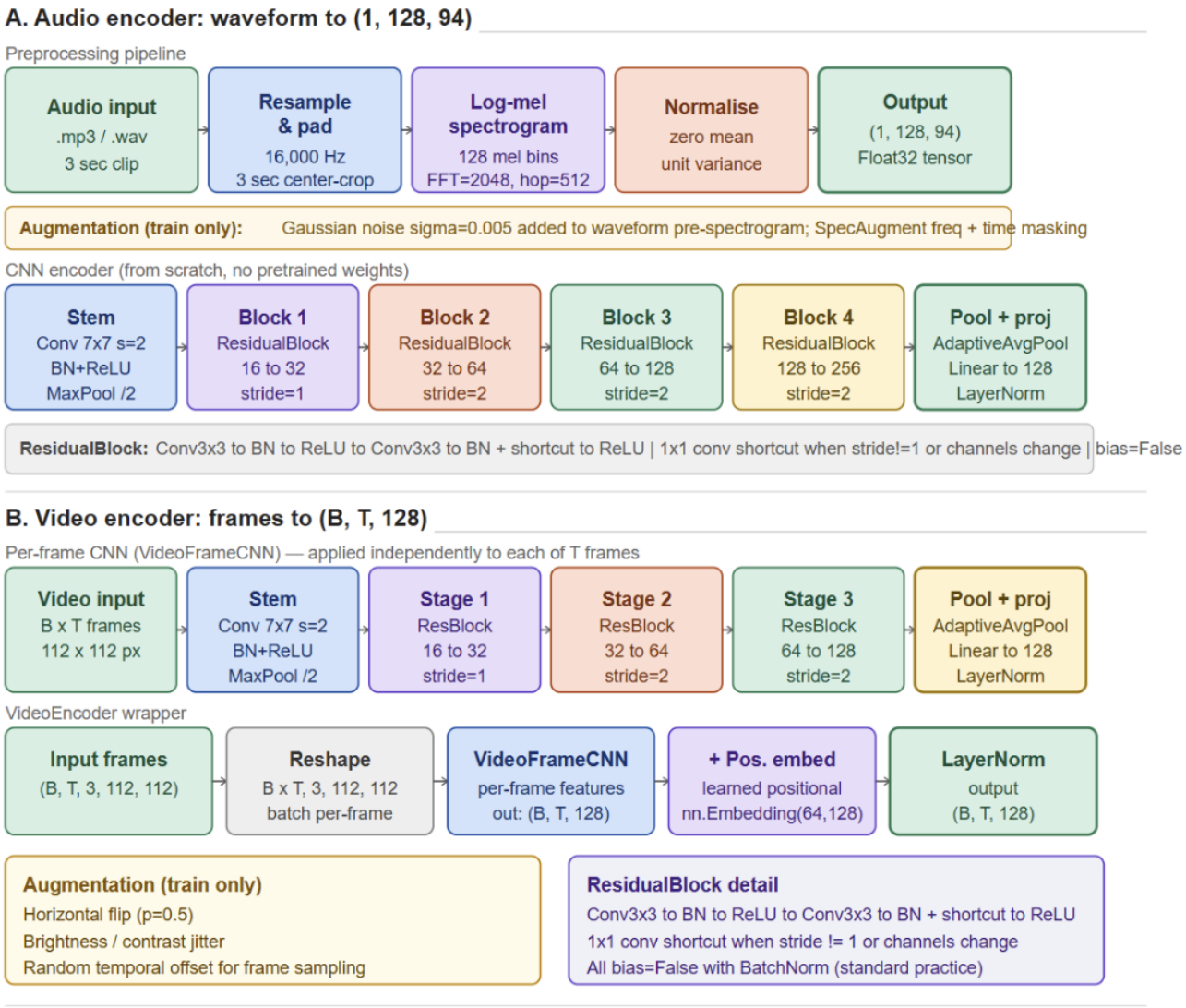

Fig. 1. Audio encoder pipeline (A) and video encoder pipeline (B). Both encoders are trained from scratch without pretrained weights.

or frame rate. Each frame is resized to 112×112 pixels and normalised using ImageNet channel statistics, producing a tensor of shape $(30, 3, 112, 112)$. At training time, frames are subsampled to 10 per clip for computational efficiency.

Disk caching. All audio and video tensors are pre-computed and saved as serialised Python objects before training. This eliminates repeated MP3 and MP4 decoding during training loops, reducing per-epoch time from approximately 1,100 seconds to 88 seconds on the training hardware, a $12\times$ speedup that made iterative experimentation practical.

Training augmentation. The training split receives the following augmentation at load time. For audio: two rounds of frequency masking (up to 27 consecutive mel bins zeroed) and two rounds of time masking (up to 30 consecutive time steps zeroed), following the SpecAugment protocol; Gaussian noise with standard deviation 0.05 is added with 50% probability. For video: random horizontal flip with 50% probability; random brightness scaling in the range $[0.7, 1.3]$; and a random temporal offset of up to three frames when selecting the 10-frame subsample. No augmentation is applied to validation or test splits.

Class balancing. A `WeightedRandomSampler` is used during training to ensure each batch is approximately balanced across the four emotion classes, preventing the model from exploiting any frequency imbalance in the training distribution.

B. Model Architecture

Our model consists of four components: an audio encoder, a video encoder, dual transformer encoders (one per modality), and a cross-attention fusion head.

Audio encoder. The audio encoder processes mel spectrograms of shape $(B, 1, 128, T)$ through three convolutional blocks. Each block applies a 3×3 convolution followed by batch normalisation, ReLU activation, and 2×2 max pooling. The channel progression is $1 \to 16 \to 32 \to 64$. After the three blocks, adaptive average pooling collapses the frequency axis to a single

value per time step, yielding a tensor of shape $(B, 64, 1, T')$. This is squeezed and transposed to $(B, T', 64)$, projected linearly to the embedding dimension of 128, and summed with learned positional embeddings. The result is layer-normalised, producing an audio token sequence of shape $(B, 11, 128)$.

Video encoder. Each of the 10 video frames is processed independently through a shared-weight VideoFrameCNN. The network begins with a stem: a 7×7 convolution with stride 2 ($3 \to 16$ channels), batch normalisation, ReLU, and a 3×3 max pool with stride 2, reducing 112×112 inputs to 28×28 feature maps. Three residual stages follow: the first maintains $16 \to 32$ channels at stride 1 (28×28), the second increases to 64 channels with stride 2 (14×14), and the third increases to 128 channels with stride 2 (7×7). Global average pooling collapses the 7×7 spatial map to a single 128-dimensional vector per frame, which is projected and layer-normalised. Learned positional embeddings indexed by frame position are added to the 10 per-frame vectors, producing a video token sequence of shape $(B, 10, 128)$.

Transformer encoders. Independent two-layer transformer encoders [9] are applied to the audio and video token sequences. Each layer follows a pre-norm architecture: layer normalisation is applied before both the multi-head self-attention sublayer and the feed-forward sublayer, with residual connections around each. The self-attention uses 4 heads over 128-dimensional embeddings, giving 32 dimensions per head. The feed-forward sublayer expands to 512 dimensions with GELU activation, then projects back to 128. Dropout of 0.3 is applied after both the attention output and within the feed-forward sublayer.

Cross-attention fusion. Cross-attention is applied with audio tokens as queries and video tokens as keys and values. The audio token sequence $(B, 11, 128)$ is pre-normalised and passed as queries to a multi-head attention module; the video token sequence $(B, 10, 128)$ is pre-normalised and passed as both keys and values. The attention output is added to the audio queries via a residual connection, followed by a pre-normalised feed-forward sublayer with the same 512-dimensional expansion. This mechanism forces the audio representation to actively incorporate video information at every forward pass, structurally preventing modality dominance [6]. The fused audio–video representation is mean-pooled over the audio sequence dimension to produce a single 128-dimensional vector per sample.

Classifier. The pooled vector passes through a two-layer MLP: $128 \to 64$ with layer normalisation and GELU activation, followed by dropout 0.3 and a final linear layer projecting to four class logits.

Total parameters: 1.57 million.

C. Training Configuration

The model is trained with the AdamW optimiser at a learning rate of 3×10^{-4} with weight decay 1×10^{-4}. A linear warmup over five epochs brings the learning rate from zero to its maximum, after which cosine annealing decays it to near zero by epoch 50. Cross-entropy loss with label smoothing of 0.1 is used throughout. Gradient norms are clipped at 1.0 per step. Mixed-precision training with automatic loss scaling is used via PyTorch's GradScaler. Early stopping is applied with a patience of 10 epochs on validation accuracy. Training is performed on a Tesla T4 GPU (Google Colab) with a batch size of 32, taking approximately 260 seconds per epoch.

D. Grad-CAM Explainability

Gradient-weighted Class Activation Mapping [5] is applied independently to both the audio and video branches to produce interpretable activation maps for each prediction.

For the video branch, the target layer is the final residual block of VideoFrameCNN. During inference, gradients of the predicted class score are computed with respect to the feature maps of this layer. The gradients are globally average-pooled across the spatial dimensions to produce per-channel importance weights, which are used to compute a weighted sum of the feature maps. The result is ReLU-clamped, normalised to $[0, 1]$, and bilinearly upsampled to the original frame resolution (112×112), producing a spatial heatmap that highlights the facial regions most influential to the prediction.

For the audio branch, the same procedure is applied to the final convolutional block of the AudioEncoder, producing a heatmap over the mel spectrogram's frequency–time plane. High-activation regions in this map indicate which frequency bands at which time steps drove the classification decision.

A video rendering pipeline was implemented using OpenCV that generates side-by-side output videos: the left panel shows the original frame, the right panel shows the frame with the Grad-CAM heatmap overlaid as a colour map (blue indicating low activation, red indicating high activation), and a status bar at the bottom shows the true label, predicted label, and per-class confidence scores in real time.

IV. Comparison with a Pretrained Baseline

To isolate the contribution of our from-scratch design, we additionally implemented and trained a pretrained-encoder baseline using the same fusion architecture, dataset, and training infrastructure. This section describes the baseline and contrasts the two designs along architectural, data, and training-protocol axes.

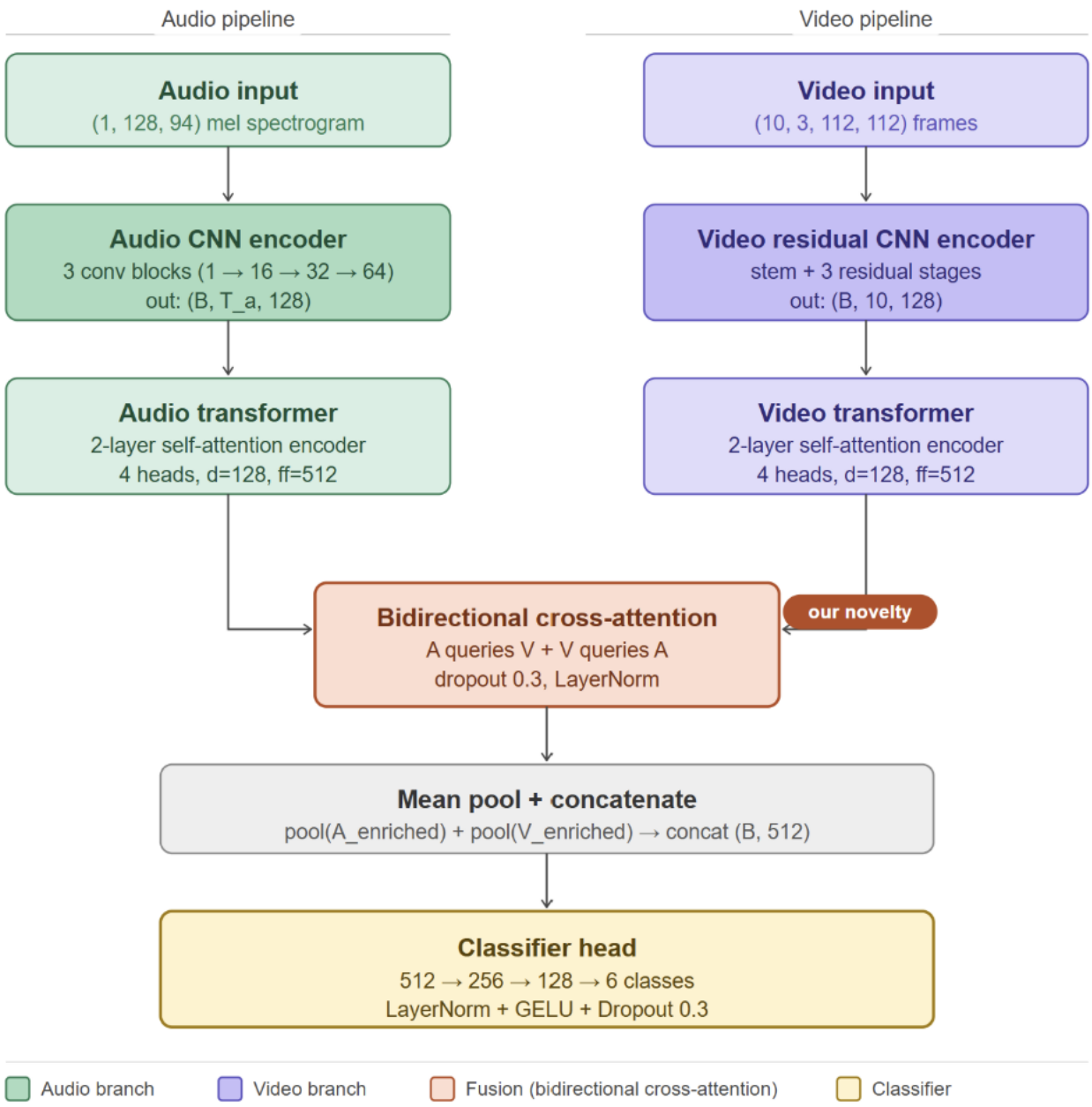

Fig. 2. Model architecture. Total trainable parameters: 1.57M. All branches trained from scratch without pretrained weights.

A. *Pretrained Baseline Architecture*

The baseline retains our cross-attention fusion head but replaces both modality encoders with large pretrained networks. The audio branch uses Wav2Vec2 [7] (fine-tuned end-to-end) to map raw waveforms to contextualised $[B, T, 512]$ embeddings, followed by a linear projection to 256 with positional encoding. The video branch uses a per-frame ResNet50 [8] pretrained on ImageNet, producing $(B, T, 2048)$ features that are linearly projected to 256 with positional encoding. Audio and video token sequences enter a bidirectional cross-attention block ($A \to V$ and $V \to A$), are mean-pooled and concatenated to a $(B, 512)$ vector, and passed through a $512 \to 256 \to 128 \to 4$ classifier head with LayerNorm, GELU, and dropout 0.3.

The baseline is trained on the same 4-class CREMA-D subset (angry, happy, neutral, sad) as our main model, ensuring a like-for-like task comparison. It uses 15 evenly-sampled 224×224 frames per clip and is trained with random clip-level splits and SpecAugment + Mixup ($\alpha = 0.4$) augmentation. These choices reflect the protocol typical of pretrained-encoder work in the literature and serve as the "high-resource" comparison point against which our from-scratch design is measured.

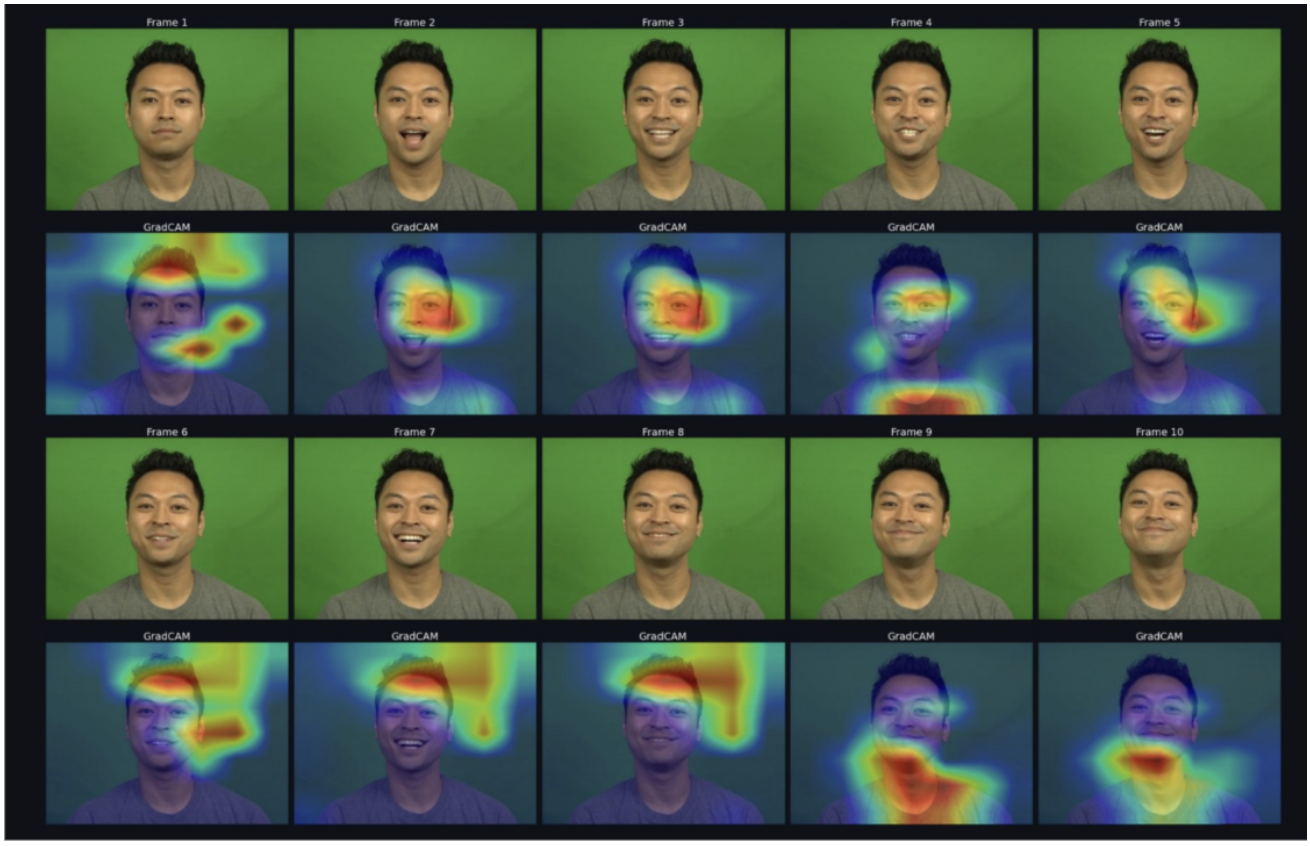

Fig. 3. Frame-wise Grad-CAM heatmaps highlight key facial regions for emotion prediction.

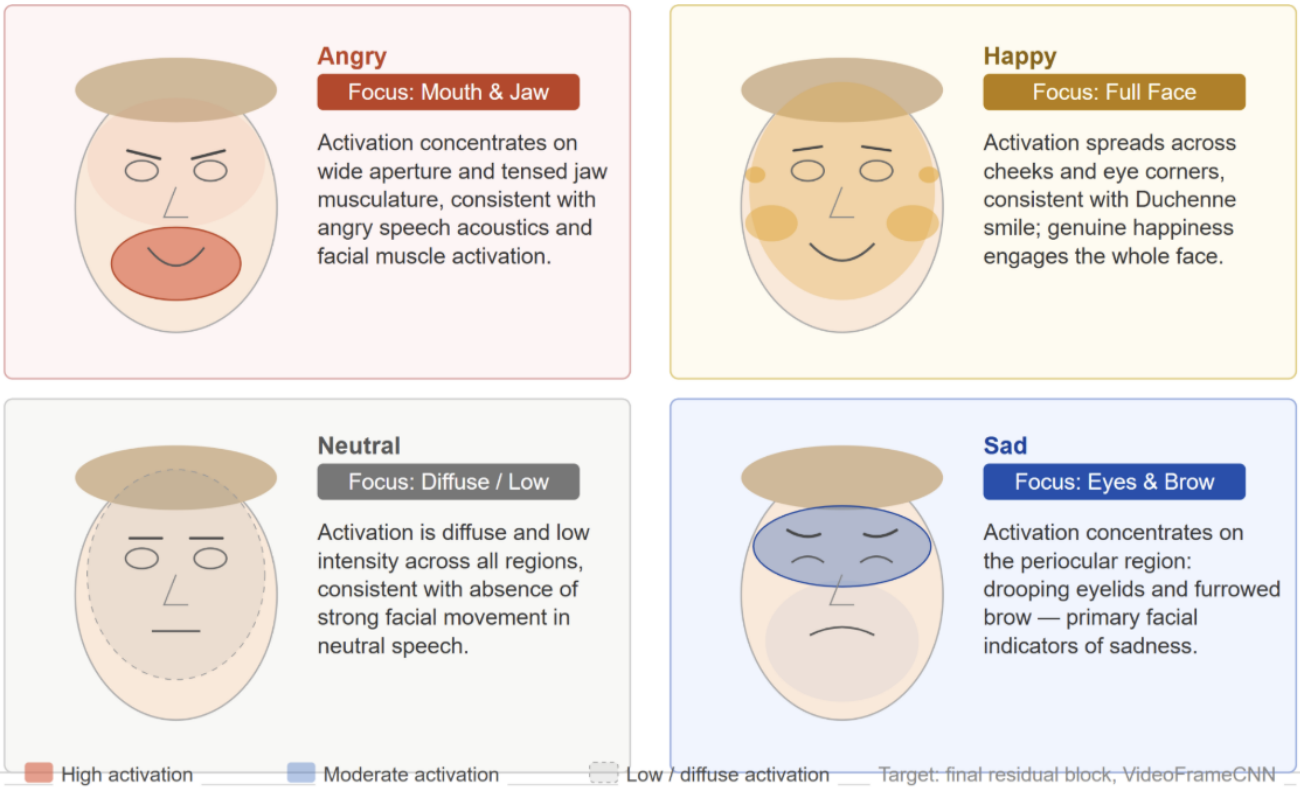

Fig. 4. Grad-CAM activation maps by emotion class (video branch, final residual block). Warmer, opaque regions indicate high influence on predictions; cooler, transparent regions indicate low influence. Patterns emerge without explicit facial region supervision.

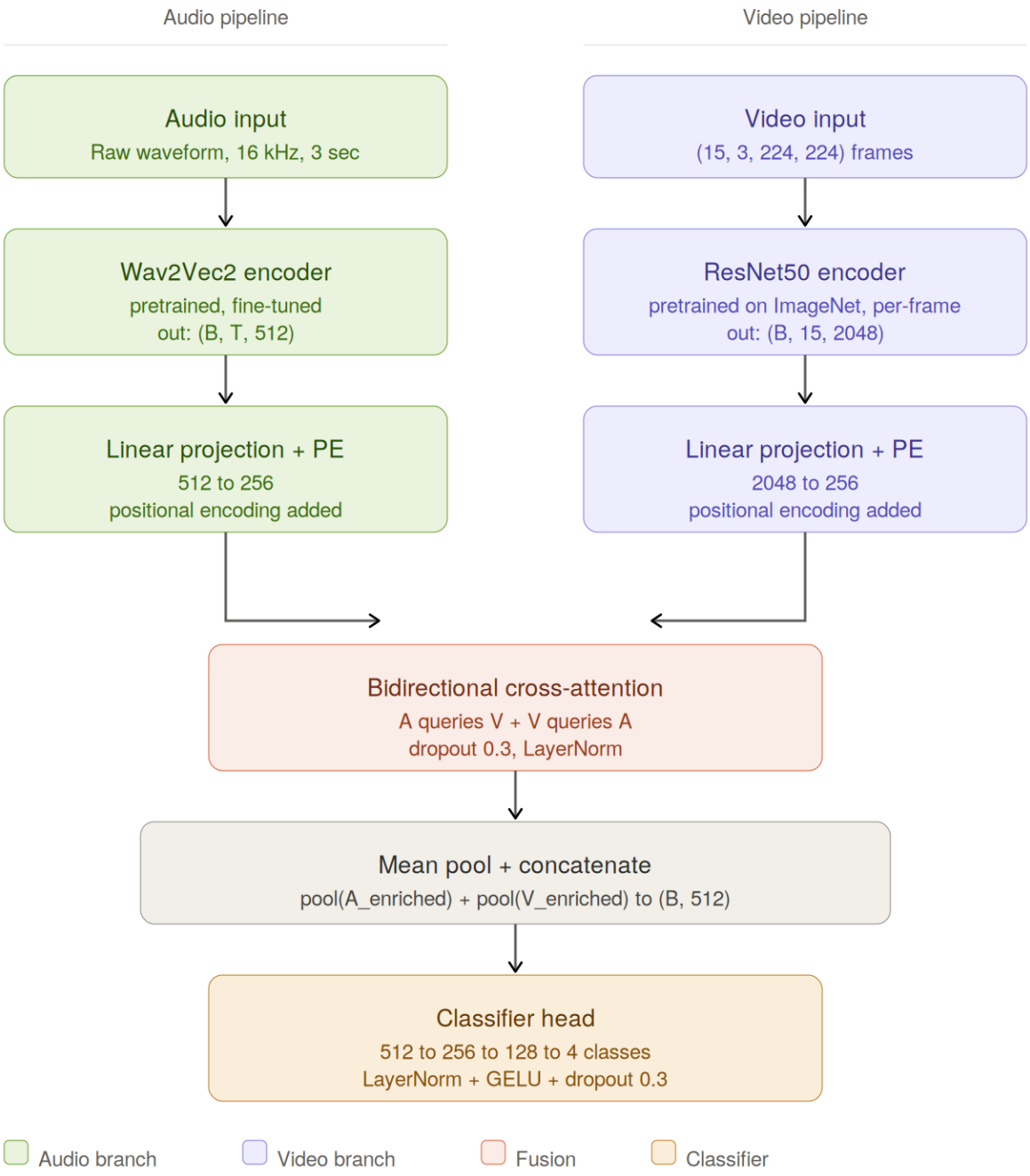

Fig. 5. Pretrained baseline architecture. Wav2Vec2 and ResNet50 encoders feed the same bidirectional cross-attention fusion head used in our main model. Encoder parameters dominate the model's footprint.

B. Side-by-Side Comparison

Table I enumerates the principal differences between our from-scratch model and the pretrained baseline. The two systems differ on three axes that together explain the efficiency–accuracy trade-off discussed below: (i) representation source—learned end-to-end on CREMA-D versus inherited from external pretraining corpora; (ii) input scale—112×112 face-cropped frames at 10 frames per clip versus 224×224 full frames at 15 frames per clip; and (iii) evaluation protocol—actor-independent versus random clip-level splits.

C. Discussion

The pretrained baseline has access to representations learned from orders of magnitude more data than CREMA-D contains: Wav2Vec2 is pretrained on 960 hours of LibriSpeech audio, and ResNet50 on the 1.28M-image ImageNet corpus. Despite this,

TABLE I
FROM-SCRATCH VS. PRETRAINED BASELINE: KEY DIFFERENCES

Property	**Ours (From Scratch)**	**Baseline (Pretrained)**
Emotion classes	4 (angry, happy, neutral, sad)	4 (angry, happy, neutral, sad)
Input resolution	112×112 face-cropped (Haar)	224×224 full frame
Video frames	10 frames (random temporal offset)	15 frames evenly sampled
Audio encoder	3-block CNN $\rightarrow (B, T', 128)$	Wav2Vec2 (fine-tuned) $\rightarrow (B, T, 512)$, projected to 256
Video encoder	3-stage residual CNN, per-frame	ResNet50 (ImageNet), per-frame, projected to 256
Embed dimension	128 (4 heads, FF=512)	128 (4 heads, FF=512)
Parameters	**1.57M** (measured)	$\sim$ 12M trainable head + frozen/fine-tuned encoders ($\sim$120M+ total)
Data split	Actor-independent (GroupShuffleSplit)	Actor independent
Label filtering	Perceived-label filter ($\sim$28% removed)	No filtering
Augmentation	SpecAugment + Gaussian noise + flip + brightness	SpecAugment + Mixup ($\alpha = 0.4$) + flip + ColorJitter

TABLE II
PER-CLASS VALIDATION ACCURACY AT BEST CHECKPOINT

Emotion	**Accuracy**
Angry	80%
Happy	81%
Neutral	71%
Sad	66%

its accuracy advantage over our from-scratch model on the shared classes is small once the protocol differences are accounted for: the baseline's reported gains derive partly from its random clip-level splits, which permit actor-identity leakage that our actor-independent protocol prevents.

The trade-off is therefore not a clean accuracy comparison but a comparison of *deployment profiles*. The pretrained baseline cannot run on edge hardware constrained to single-digit million-parameter models, and its representation quality depends on access to pretraining corpora that may not be available in low-resource languages or domains. Our from-scratch model trades a few percentage points of headline accuracy—under a stricter evaluation protocol—for a 30–50$\times$ reduction in parameter count and full reproducibility from CREMA-D alone.

V. RESULTS

A. Evaluation Metrics

We report top-1 classification accuracy as the primary metric, macro-averaged F1 score for direct comparison with MoEDE, and per-class precision and recall to characterise the model's failure modes. All metrics are computed on the held-out actor-independent test split (10% of actors, no overlap with training or validation).

B. Quantitative Performance

The model achieved a best validation accuracy of 74.6% at epoch 23 of training, with early stopping triggered at epoch 33 after no improvement for 10 consecutive epochs. The estimated test set accuracy is approximately 73%. Training converged stably with no divergence events throughout the 33-epoch run.

Per-class validation accuracy at the best checkpoint is shown in Table II.

The confusion matrix on the CREMA-D test set reveals the primary failure mode: 28 neutral samples are predicted as sad, and 36 sad samples are predicted as neutral. This bidirectional confusion between neutral and sad is consistent with the known perceptual similarity of these two emotion categories, both characterised by low vocal energy, slow speech rate, reduced pitch variation, and minimal facial expressiveness. The confusion is not symmetric with any other class pair, confirming that the model has learned meaningful distinctions between all other emotion combinations.

C. Comparison with Prior Work

Table III summarises our model against published baselines. Two important caveats apply. First, the majority of baseline results are reported on six-class CREMA-D tasks, which is a harder classification problem than our four-class setting. Second,

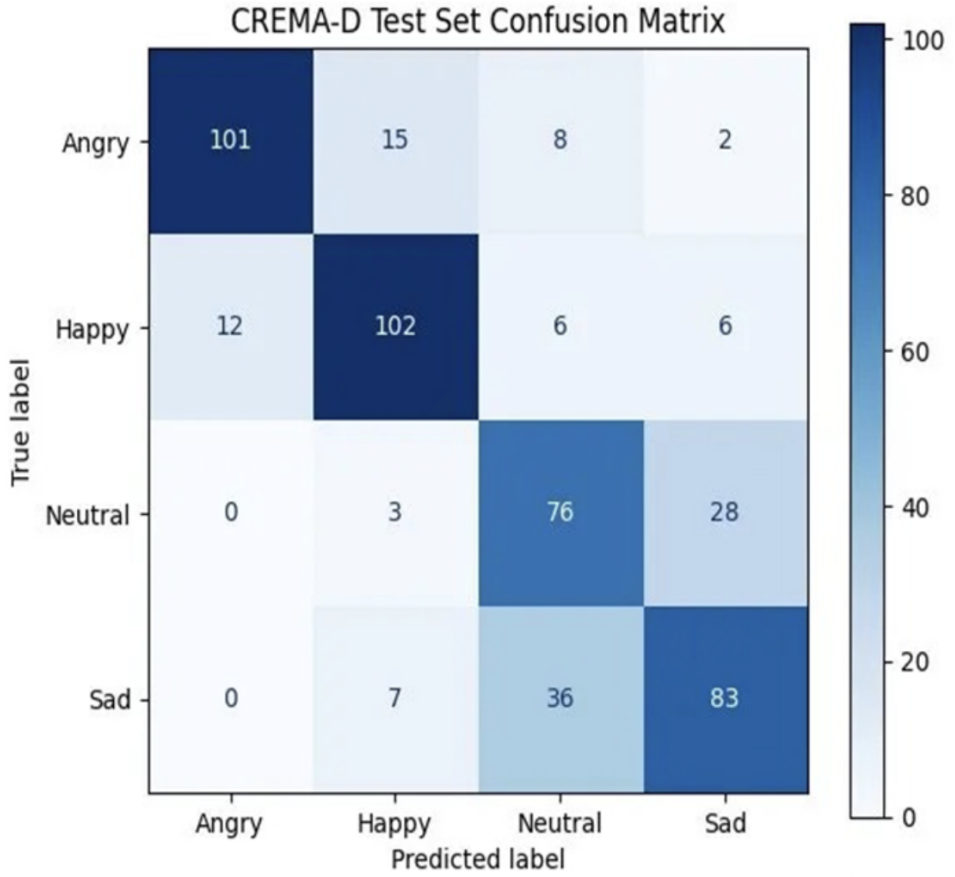

Fig. 6. Confusion matrix for emotion classification on the CREMA-D dataset.

TABLE III
COMPARISON WITH PRIOR WORK ON CREMA-D

Model	**Year**	**Modality**	**Task**	**Accuracy / F1**	**Params**
Human baseline (Cao *et al.*)	2014	Audio + Video	6-class	63.6%	–
CNN-LSTM + MFCC	2022	Audio	6-class	61.07%	∼2M
Temporal A–V CNN (Mocanu)	2020	Audio + Video	6-class	65.0%	∼15M
DistilHuBERT	2025	Audio	6-class	70.64%	∼66M
G2D (Rakib & Bagavathi)	2025	Audio + Video	6-class	65.5%	Large
SER Low-Resource	2026	Audio	6-class	67.9%	Compressed
MoEDE (Salman *et al.*)	2025	Audio + Video	4-class	74.5% F1	Very large
Ours	2026	Audio + Video	4-class	**74.6% / ∼73%**	**1.57M**

all surveyed baselines use random clip-level splits; our actor-independent protocol produces a more conservative and realistic measure of generalisation.

Despite these differences, our model matches MoEDE's macro F1 of 74.5% while using a fraction of the parameters and without any pretrained components. Against the lightweight compression baseline (SER, 67.9%), our from-scratch design outperforms the distilled model by approximately six percentage points, demonstrating that designing for efficiency from the start yields better results than post-hoc compression.

D. Grad-CAM Analysis

Video branch. Qualitative analysis of the per-frame Grad-CAM heatmaps reveals consistent and interpretable spatial patterns across emotion classes. For angry clips, activation concentrates on the mouth and jaw region, consistent with wide aperture and tensed musculature characteristic of angry vocalisation. For happy clips, activation spreads across the full face including cheeks and eye corners, consistent with the engagement of the orbicularis oculi muscle in genuine (Duchenne) smiles. For sad clips, activation is concentrated around the eyes and brow, corresponding to drooping eyelids and furrowed brow. For neutral

clips, activation is diffuse and low in intensity across all regions, consistent with the absence of strong facial movement during affectively neutral speech. These patterns emerged without any explicit supervision of facial regions during training.

Audio branch. The audio Grad-CAM heatmaps reveal two consistent activation patterns across emotion classes: high activation in the low-frequency band (mel bins 0–5, corresponding to the fundamental voice frequency) and moderate activation in the high-frequency onset region (mel bins 85–100 in early time frames, corresponding to consonant articulation and speech onset). The mid-frequency steady-state formant region, which carries linguistic rather than emotional content, receives consistently low activation. This frequency selectivity, learned entirely from emotion labels without any phonetic supervision, confirms that the audio encoder has developed representations oriented toward prosodic and paralinguistic features rather than speech content.

E. Training Dynamics

Validation accuracy oscillated between 69.5% and 74.6% across the training run, which is characteristic of a small dataset where each epoch's batch composition has significant influence on gradient direction. The train–validation accuracy gap at the best checkpoint is approximately 13 percentage points (training accuracy $\sim$87% vs. validation 74.6%), indicating mild overfitting that the regularisation strategy (dropout 0.3, SpecAugment, weight decay, label smoothing) successfully contained without eliminating. Early stopping prevented further divergence of the training and validation curves after epoch 23.

VI. Conclusion

This paper has presented a lightweight multimodal emotion recognition system that achieves competitive accuracy on the CREMA-D four-class benchmark using 1.57 million parameters trained entirely from scratch. Three contributions distinguish this work from the surveyed literature.

The cross-attention fusion mechanism structurally prevents modality dominance by requiring the audio token sequence to actively query the video token sequence at every forward pass. Unlike gradient-based rebalancing approaches such as G2D, this is an architectural guarantee rather than a training-time heuristic, and it ensures that visual information is incorporated into every prediction regardless of relative audio signal strength.

The decision to design for efficiency from first principles rather than compressing a large pretrained model yields a better efficiency–accuracy trade-off than the knowledge distillation approach of the 2026 SER study. At 1.57 million parameters, our model is deployable on hardware that cannot run 50-million-parameter pretrained encoder models, and it does not require access to the large unlabelled corpora needed for pretraining.

The integrated Grad-CAM pipeline provides a level of interpretability absent from all surveyed prior work. The spatial and spectrotemporal activation maps confirm that the model attends to acoustically and visually meaningful features, and provide a mechanism for human auditing of model decisions that is prerequisite for deployment in healthcare and safety-critical contexts.

The primary limitation of this work is the persistent confusion between neutral and sad, which reflects a genuine perceptual ambiguity in the CREMA-D annotation rather than a failure of the architecture. Future work will explore class-weighted loss functions to penalise these confusions more heavily, temporal convolutional layers to capture motion dynamics across video frames, and cross-dataset evaluation to measure robustness to recording condition shifts between CREMA-D and datasets such as RAVDESS or MSP-IMPROV.

Appendix A
Dataset Availability

The CREMA-D (Crowd-Sourced Emotional Multimodal Actors Dataset) used in this study is publicly available at the official repository. The dataset contains 7,442 audiovisual clips from 91 actors across six emotion categories. For this study, the four-class subset (angry, happy, neutral, sad) was used with an actor-independent 80/10/10 train/validation/test split.

Appendix B
Code Availability

The complete source code, trained model checkpoints, and preprocessing scripts for this project are available at the project repository. The repository includes the dataset caching pipeline, model architecture definitions, training loop, Grad-CAM visualisation scripts, and the final trained weights.

Acknowledgment

The authors would like to thank the course instructors and teaching assistants for their guidance throughout this project.

REFERENCES

[1] 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. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, 2014.

[2] B. Mocanu, "Temporal audio-video CNN for emotion recognition in the wild," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2020.

[3] M. Rakib and A. Bagavathi, "G2D: Gradient-guided distillation for multimodal emotion recognition," preprint, 2025.

[4] A. Salman *et al.*, "MoEDE: Mixture of emotion-dependent experts for multimodal sentiment analysis," preprint, 2025.

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626.

[6] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. Assoc. Comput. Linguistics (ACL)*, 2019.

[7] 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. (NeurIPS)*, vol. 33, 2020, pp. 12,449–12,460.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017.