symmetry

MDPI

Article

DRAV-GON: Structural Symmetry in Cross-Modal Interaction for Robust Audio-Visual Emotion Recognition

Ruijia Fan [1], Yang Chen [2,*], Guoquan Wang [3] and Wenshuai Wang [3]

[1] Grid Planning & Research Center, Guangdong Power Grid Corporation, Guangzhou 510000, China; fanruijia@alumni.pku.edu.cn
[2] Faculty of Computational Mathematics and Cybernetics, Shenzhen MSU-BIT University, Shenzhen 518172, China
[3] Open Laboratory on Human-Robot Interaction, Shenzhen Graduate School, Peking University, Shenzhen 518055, China; gqwang109@gmail.com (G.W.); wangws@stu.pku.edu.cn (W.W.)
* Correspondence: chenyang@smbu.edu.cn

Abstract

Audio-visual emotion recognition supports affective computing and human–computer interaction, but semantic information asymmetry between audio and visual modalities hinders the extraction of a joint audio-visual representation. We propose DRAV-GON, a framework that maps heterogeneous data into a shared representation space and leverages structurally symmetric cross-modal interaction for modality feature fusion. We propose a Heterogeneous Acoustic Representation Learner, which strengthens acoustic representations by combining handcrafted spectral descriptors with pretrained embeddings. We propose a structurally symmetrical Gumbel-Guided Attention Fusion module to fuse features from the two modalities. Furthermore, we propose Dynamic Input Perturbation to simulate degraded or missing modalities during training, improving the robustness of the model. Extensive experiments demonstrate competitive performance under clearly specified evaluation protocols. Specifically, under 5-fold cross-validation, Variant 2 of DRAV-GON reaches 93.45% $\pm$ 1.66% mean accuracy (best fold 96.11%) on the PKU-ER dataset, and Variant 1 achieves 91.08% $\pm$ 4.55% (best fold 95.00%) on the RAVDESS dataset. Robustness evaluations across six test-time corruption conditions characterize both the benefits and the remaining modality dependence of the framework.

Keywords: structural symmetry; multimodal learning; audio-visual emotion recognition; bidirectional cross-modal attention; heterogeneous representation learning; Gumbel-Softmax; modality corruption

Check for updates

Academic Editors: Haobing Liu, Yanwei Yu, Yuxi Wang, Tianzi Zang and Zhixun Su

Received: 8 October 2026
Revised: 10 November 2026
Accepted: 20 November 2026
Published: 26 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Human emotional communication naturally relies on multiple modalities interacting closely [1]. Artificial empathy research in therapy and healthcare further motivates the development of technologies for affective interpersonal interaction [2]. As key components of multimodal affective signals, acoustic patterns (e.g., vocal prosody) and facial movements constitute primary carriers of emotional states, forming an integrated perceptual system in humans. The central role of facial appearance in conveying emotion is further evidenced by recent advances in programmable soft-mask technologies [3]. However, when processing such multimodal data computationally, acoustic-visual streams frequently suffer from partial corruption or asynchrony in real-world environments due to hardware limitations or background interference, degrading machine recognition robustness. Visual cues may

become unavailable due to occlusion, forcing reliance on auditory cues alone. Conversely, degraded or absent audio signals may require exclusive dependence on visual information. Existing methods primarily adopt two strategies: (1) ensemble-based techniques [4], or (2) Transformer-based multimodal architectures [5] that theoretically support human-initiated modal switching. However, neither approach can handle unpredictable real-world scenarios involving random modality corruption or dropout during inference. Audio-visual emotion recognition has emerged as a critical technology for human–computer interaction, yet its real-world deployment remains constrained by critical technical bottlenecks. Specific problems and solutions are described below:

Problem 1: Modal Information Asymmetry. Audio and visual signals can provide unequal predictive information in multimodal representation learning, and the stronger modality depends on the dataset and architecture. In the visual-dominant settings that motivate this design, conventional spectral audio features may provide less discriminative information than pretrained visual representations [6]. Figure 1(a2) illustrates this feature-extraction setting. Certain audio-visual models adopt the visual modality as the dominant branch and use its output directly as the final prediction [7]. Such asymmetry can produce unequal representation learning and gradient contributions.

Solution 1. To bridge this gap, we design the Heterogeneous Acoustic Representation Learner (HAR-Learner). HAR-Learner combines interpretable handcrafted spectral features, including mel-frequency cepstral coefficients (MFCC) [8] and the constant-Q transform (CQT) [9], with data-driven features from pretrained models such as the audio pretraining model with acoustic tokenizers (BEATs) [10] to transfer high-level semantic information. This design is intended to enrich the acoustic representation before cross-modal interaction, rather than to impose equal predictive performance on the audio and visual branches.

Problem 2: Cross-Modal Representation-Propagation Dilemma. Many audio-visual emotion recognition systems combine pretrained or fixed feature extractors with a downstream fusion module. Such pipelines can be effective, but their fixed intermediate representations cannot be jointly adapted with the downstream fusion network. This motivates our joint optimization of the trainable acoustic projections, visual temporal model, fusion module, and classifiers.

Another point to consider is that current methods typically use fixed fusion approaches, such as concatenation or Transformer-based fusion methods, as shown in Figure 1(a3). While simple fixed fusion methods such as concatenation are easy to implement, they face difficulties in preserving the intrinsic connections between modalities and eliminating redundant information. Although Transformer-based fusion works well for tasks needing fine-grained weight distribution, it struggles with clear-cut decisions and tends to leave behind noise. These issues demonstrate the need for improvements in current fusion methods.

Solution 2. We develop an end-to-end adaptive architecture unifying feature extraction and cross-modal fusion through differentiable operations. The feature fusion uses paired interaction paths for audio and visual features. The Gumbel-regularized mechanism in Figure 1(b3) applies reciprocal cross-modal modulation and stochastic gating before concatenating the two representations. Moreover, the network is optimized through joint classifier training, differing from traditional approaches that rely solely on a single loss function, as shown in Figure 1(a4) vs. (b4).

Problem 3: Fragility to Modal Distortion and Incompleteness. Current models, despite achieving high performance under ideal multi-modal conditions, perform poorly when confronted with modality distortions, such as sensor failures, environmental noise, or occlusion events, that disrupt audio or visual inputs and lead to catastrophic accuracy

degradation. This limitation originates from the unrealistic assumption of modal completeness during training: models implicitly assume that all modalities are always complete, as shown in Figure 1(a1). As a result, feature encoders become overly dependent on fixed input patterns, making them less able to handle incomplete or distorted sensory data. While recent modality-robust training schemes attempt to mitigate this issue through modality dropout and attention-fusion variants [6], their performance remains suboptimal.

Solution 3. We propose Dynamic Robust Audio-Visual Emotion Recognition with a Gumbel-Optimized Cross-Modal Network (DRAV-GON). Its Dynamic Input Perturbation (DIP) strategy is shown in Figure 1(b1). DIP randomly masks individual modalities during training. This approach encourages the model to develop compensatory mechanisms across modalities, supporting emotion inference when inputs are missing.

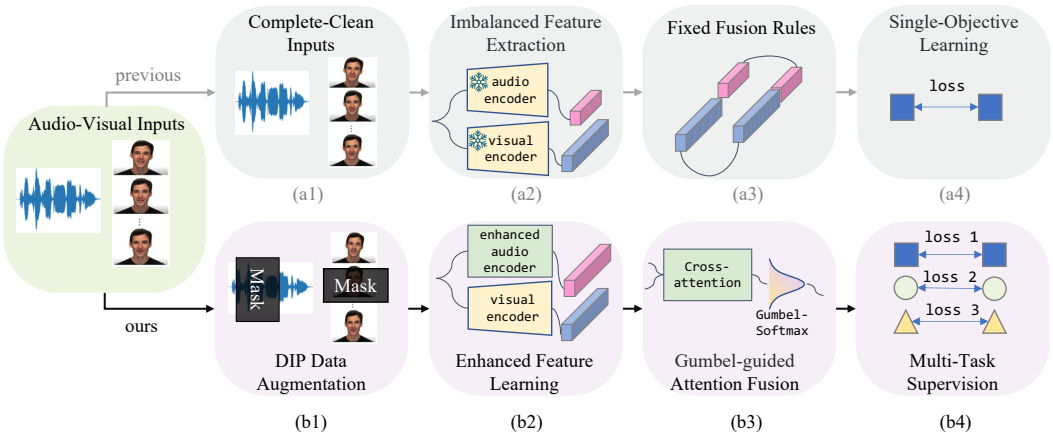


Figure 1. Comparison of the conventional methods and the proposed method. Conventional methods typically (**a1**) process clean, complete inputs, extract isolated features (resulting in (**a2**) imbalanced representations), and employ (**a3**) a fixed fusion network with (**a4**) a single-objective loss function. The snowflake symbol (**a2**) denotes that the model weights are frozen during training. Our DRAV-GON leverages (**b1**) Dynamic Input Perturbation for augmentation, (**b2**) an enhanced feature learner, and (**b3**) an adaptive fusion network to generate robust multi-modal representations. (**b4**) Multi-task supervision is conducted. The model is optimized through joint classifier training.

In summary, we have made the following contributions:

- We propose DRAV-GON, which jointly optimizes the trainable feature-learning, fusion, and classification components for robust audio-visual emotion recognition.
- Dynamic Input Perturbation (DIP) introduces modality-level replacement, attenuation, and zero-masking during training to simulate modality degradations and evaluate robustness under matched and unseen corruption operators.
- HAR-Learner enriches acoustic representations by combining handcrafted spectral features with pretrained embeddings.
- We propose Gumbel-Guided Attention Fusion (G^2AF), which combines bidirectional cross-attention, reciprocal modulation, and stochastic gating for cross-modal representation learning.
- Under five-fold cross-validation, our method achieves 93.45% $\pm$ 1.66% mean accuracy (best fold 96.11%) on the Peking University Emotion Recognition dataset (PKU-ER) and 91.08% $\pm$ 4.55% (best fold 95.00%) on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). We report these results separately from the matched single-split comparisons because the evaluation protocol, acoustic backbone, and visual sampling rate differ.

2. Related Work

Feature Extraction. Feature extraction in affective computing has evolved from manual engineering to data-driven approaches, alongside a broader progression from unimodal analysis to multimodal fusion [11]. Early studies primarily employed expert-designed features. For audio analysis, the open-source Speech and Music Interpretation by Large-space Extraction (openSMILE) toolkit [12] provides standardized low-level descriptors of (1) prosody, including fundamental frequency contours and energy dynamics, (2) voice quality, and (3) spectral characteristics, such as MFCC [8] and CQT [9]. For visual analysis, texture descriptors include local binary patterns (LBP) [13] and histograms of oriented gradients (HOG) [14], while scale-invariant feature transform (SIFT)-based facial modeling improves robustness to viewpoint changes [15]. Although computationally efficient and interpretable, these methods struggle to capture high-level affective semantics. Deep learning enables automatic representation learning. For audio analysis, convolutional neural networks (CNNs) [16] can process raw speech, while long short-term memory (LSTM) [17] networks and gated recurrent units (GRUs) [18] improve temporal modeling [19]. BEATs [10] learns representations directly from waveforms. Beyond speech, USV-DETR [20,21] demonstrates high-resolution, densely supervised detection of ultrasonic vocalizations, highlighting the broader progress of learned acoustic representations. In visual analysis, hybrid CNN-LSTM architectures are adopted for spatiotemporal expression analysis [22,23]. Attention-integrated, multi-scale feature encoding has also advanced fine-grained visual recognition tasks such as fruit and vegetable instance segmentation [24]. While improving semantic capabilities, these approaches increase model complexity and reduce interpretability. Related advances in physiological emotion recognition include domain generalization with residual networks [25] and adaptive spatial–temporal graph learning for EEG signals [26], while acoustic-inspired brain-to-sentence decoding extends learned signal representations to language reconstruction [27]. In power-system fault detection, time–frequency embedded deep learning [28], DSP-enhanced Transformers [29], and FD-Mamba with a neural observer and frequency-enhanced updates [30] provide further examples of combining signal-processing structure with learned temporal representations. Persistent challenges include cross-modal representation disparities and noise robustness, setting higher standards for the next generation of affective computing systems.

Multimodal Fusion and Structural Symmetry. Multimodal learning encompasses representation, alignment, and fusion as interconnected challenges [31]. Fusion can operate on input features, intermediate representations, or predictions. Chumachenko et al. [6] compare three audio–visual fusion variants, and Transformer-based methods use attention to model dependencies between modality features [7,32]. MultiMAE-DER [33] uses unified encoders in an early fusion framework for cross-domain emotion recognition. Zhou et al. [34] propose dynamic graph learning with pairwise cross-modal alignment for accurate and robust multimodal emotion recognition in human–robot interaction. Their method emphasizes pairwise alignment and robustness-oriented interaction, whereas DRAV-GON employs structurally symmetric bidirectional attention, reciprocal Gumbel-guided modulation, and training-time modality perturbation. The two designs address complementary aspects of robust cross-modal fusion.

Fan et al. [35] extend E-Branchformer to facial, acoustic, and posture processing, with structurally corresponding modality branches and a shared fusion module. Their experiments use CREMA-D and an extended dataset incorporating eNTERFACE'05. This work connects symmetry in multimodal architecture design to emotion recognition. DRAV-GON places the structural correspondence within its bidirectional fusion paths and retains heterogeneous feature encoders. Its evaluation focuses on audio-visual inputs under clean and simulated corruption conditions.

Qu and Wang [36] factorize physiological representation learning into cross-channel interaction and within-channel temporal-spectral modeling. Their framework uses hierarchical cross-modal fusion and jointly predicts valence, arousal, and liking from EEG and other physiological signals on DEAP. This study is relevant to the separation of modality-level representation learning from cross-modal interaction and to joint training. Its input modalities and prediction tasks differ from those evaluated here, so its reported accuracies are not direct benchmarks for DRAV-GON.

Gumbel Distribution and Gumbel-Softmax. The Gumbel distribution (Type-I Extreme Value distribution) is widely used in statistical modeling of extreme events and discrete optimization. Its key property is the Gumbel-Max trick [37], which enables reparameterization of categorical sampling from categorical distributions. This was later extended by the Gumbel-Softmax distribution [38], which introduces a continuous relaxation via softmax with temperature control. The technique has become fundamental in reinforcement learning, generative models, and discrete neural representation learning [39], bridging gradient-based optimization with discrete latent variables.

3. Methodology

3.1. Overview of DRAV-GON

As illustrated in Figure 2, DRAV-GON combines three components for robust audio-visual emotion recognition. At the input stage, DIP simulates sensory degradation through random modality dropout and noise injection during training. Feature extraction uses modality-specific pathways: visual streams use CNN [16] architectures, while the audio stream combines MFCC [8] and CQT [9] descriptors with deep neural embeddings. This stage produces visual features $\mathbf{F}_v$ and audio features $\mathbf{F}_a$. The G^{2}AF module adjusts cross-modal interaction weights through temperature-annealed attention masking and generates fused features $\mathbf{G}_{av}$. Joint optimization of the primary and auxiliary modality-specific classification losses mitigates gradient asymmetry caused by the dominance of a single modality while improving feature discriminability.

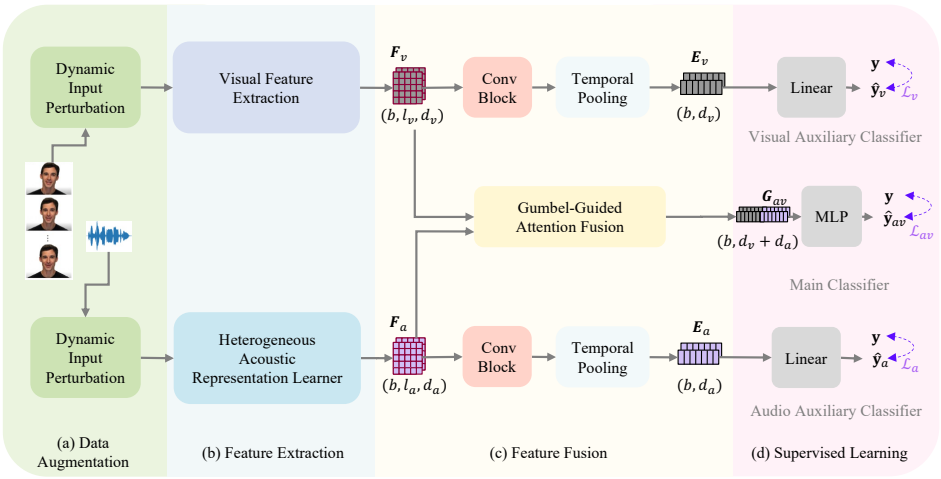


Figure 2. Framework of DRAV-GON. The proposed end-to-end framework comprises four core stages: (**a**, the green block) Dynamic Input Perturbation Augmentation (DIP): Emulates real-world modal distortions by applying random modality dropout and noise injection during training (Algorithm 1). (**b**, the blue block) Dual-Modal Feature Extraction: Processes visual inputs via a convolution module while extracting auditory features through a Heterogeneous Acoustic Representation Learner (combining traditional descriptors with deep representations). (**c**, the yellow block) Adaptive Fusion: Employs Gumbel-guided attention to enable robust cross-modal interaction. (**d**, the pink block) Joint Learning: Optimizes the network using a combined loss function from both main and auxiliary classifiers.

Algorithm 1: Dynamic Input Perturbation.

Input: Original batch: audio features $\mathbf{X}_a \in \mathbb{R}^{b\times l_a\times d_a}$, visual features $\mathbf{X}_v \in \mathbb{R}^{b\times l_v\times c\times h\times w}$, labels $\mathbf{y} \in \mathcal{Y}^b$; Perturbation mode $m \in \{\text{noise}, \text{mask}\}$

Output: Augmented batch $(\widetilde{\mathbf{X}}_a, \widetilde{\mathbf{X}}_v, \widetilde{\mathbf{y}})$

▷ Notation: b = batch size, d_a = audio feature channels ($820 = 10_{\text{MFCC}} + 42_{\text{CQT}} + 768_{\text{embedding}}$), l_a = audio frames (156 for a 3.6 s clip at 22.05 kHz), c = visual channels (3), l_v = sampled video frames, h, w = frame size (224×224); $\sigma \sim \text{Uniform}(\text{Perm}(Kb))$ is a random permutation of the Kb sample indices, drawn independently at each iteration and applied identically to the audio, visual and label tensors so that the sample-to-label pairing is preserved;

if $m = \textit{noise}$ **then**

`// Noise injection Mode`

Sample $\mathbf{N}_a, \mathbf{N}_v$ with entries i.i.d. $\sim \mathcal{N}(0,1)$, of the same shape as $\mathbf{X}_a$ and $\mathbf{X}_v$, respectively;

$\widetilde{\mathbf{X}}_a \leftarrow \text{Concat}_{\text{dim=0}}(\mathbf{X}_a,\ \mathbf{N}_a,\ \mathbf{X}_a)$ `// clean, fully corrupted, clean`

$\widetilde{\mathbf{X}}_v \leftarrow \text{Concat}_{\text{dim=0}}(\mathbf{X}_v,\ \mathbf{X}_v,\ \mathbf{N}_v)$ `// clean, clean, fully corrupted`

$K \leftarrow 3$ `// sample expansion factor`

else

`// Hybrid Masking`

Sample mixing coefficients $\mathbf{c}_a \sim \mathcal{U}(0,1)^{b\times 1\times 1}$, and set the complementary visual coefficients to $\mathbf{c}_v = \text{Reshape}(1 - \mathbf{c}_a)$;

Broadcast them to the feature shapes, $\mathbf{C}_a \in \mathbb{R}^{b\times l_a\times d_a}$ and $\mathbf{C}_v \in \mathbb{R}^{b\times l_v\times c\times h\times w}$;

$\widetilde{\mathbf{X}}_a \leftarrow \text{Concat}_{\text{dim=0}}(\mathbf{X}_a,\ \mathbf{X}_a \odot \mathbf{C}_a,\ \mathbf{0},\ \mathbf{X}_a)$ `// clean, attenuated, missing, clean`

$\widetilde{\mathbf{X}}_v \leftarrow \text{Concat}_{\text{dim=0}}(\mathbf{X}_v,\ \mathbf{X}_v \odot \mathbf{C}_v,\ \mathbf{X}_v,\ \mathbf{0})$ `// clean, complementary attenuated, clean (audio-missing), missing`

$K \leftarrow 4$ `// sample expansion factor`

Expand labels by concatenating K batch copies: $\widetilde{\mathbf{y}} \leftarrow \text{Concat}_{\text{dim=0}}(\underbrace{\mathbf{y}, \ldots, \mathbf{y}}_{K \text{ copies}})$;

Draw $\sigma \sim \text{Uniform}(\text{Perm}(Kb))$;

return $(\widetilde{\mathbf{X}}_a[\sigma],\ \widetilde{\mathbf{X}}_v[\sigma],\ \widetilde{\mathbf{y}}[\sigma])$

3.2. Dynamic Input Perturbation

Traditional multi-modal learning methods typically assume complete and reliable modality inputs, an assumption that often fails in real-world scenarios where modalities may suffer from distortions such as sensor noise, data corruption, or partial missingness. To address this limitation, we propose a DIP scheme that systematically enhances model robustness by simulating diverse distortion patterns during training. This approach introduces two complementary augmentation modes to construct a continuous perturbation space while maintaining cross-modal semantic consistency.

The perturbation strategy differs from simple data augmentation techniques. Specifically, the visual branch operates at the tensor level on multi-frame video clips rather than in pixel space. The auditory branch modifies audio features obtained through hand-crafted methods and pretrained encoders rather than simply injecting noise into raw audio signals.

3.2.1. Noise Injection Mode

We first construct perturbed samples by introducing Gaussian noise while preserving semantic consistency. Given a batch $\mathcal{B} = \{(\mathbf{X}_a^{(i)}, \mathbf{X}_v^{(i)}, y^{(i)})\}_{i=1}^{b}$, where $\mathbf{X}_a \in \mathbb{R}^{b\times l_a\times d_a}$ denotes audio features, $\mathbf{X}_v \in \mathbb{R}^{b\times l_v\times c\times h\times w}$ denotes visual features, and $\mathbf{y} \in \mathcal{Y}^b$ denotes class labels,

let $\mathbf{N}_a \sim \mathcal{N}(0, \mathbf{I})^{b \times l_a \times d_a}$ and $\mathbf{N}_v \sim \mathcal{N}(0, \mathbf{I})^{b \times l_v \times c \times h \times w}$ be noise tensors. We generate a clean, audio-corrupted, and visual-corrupted copy of each sample:

$$\begin{cases} \widetilde{\mathbf{X}}_a = [\mathbf{X}_a \| \mathbf{N}_a \| \mathbf{X}_a] \in \mathbb{R}^{3b \times l_a \times d_a}, \\ \widetilde{\mathbf{X}}_v = [\mathbf{X}_v \| \mathbf{X}_v \| \mathbf{N}_v] \in \mathbb{R}^{3b \times l_v \times c \times h \times w}, \\ \widetilde{\mathbf{y}} = [\mathbf{y} \| \mathbf{y} \| \mathbf{y}] \in \mathcal{Y}^{3b}. \end{cases} \tag{1}$$

In Equation (1), the double vertical bar $\|$ denotes tensor concatenation along the batch dimension, which joins tensors on the first axis while keeping all other dimensions unchanged. The resulting batch contains clean samples first, followed by samples with noisy audio and then samples with noisy visual input.

3.2.2. Hybrid Masking

To handle both partial and complete modality corruption, we introduce a hybrid masking mechanism that dynamically interpolates between modalities. Here, "masking" always refers to complete modality masking: an entire modality (all frames of the audio features or all frames of the visual tensor) is zeroed or attenuated as a whole, rather than partially occluded in time or space. First, audio mixing coefficients $\mathbf{c}_a \sim \mathcal{U}(0, 1)^{b \times 1 \times 1}$ are sampled, and the complementary visual coefficients are reshaped as $\mathbf{c}_v = \text{Reshape}(1 - \mathbf{c}_a) \in \mathbb{R}^{b \times 1 \times 1 \times 1 \times 1}$.

Then, augmented samples with explicitly broadcast coefficients are generated:

$$\begin{cases} \mathbf{C}_a = \text{Broadcast}_{b \times l_a \times d_a}(\mathbf{c}_a) \in \mathbb{R}^{b \times l_a \times d_a}, \\ \mathbf{C}_v = \text{Broadcast}_{b \times l_v \times c \times h \times w}(\mathbf{c}_v) \in \mathbb{R}^{b \times l_v \times c \times h \times w}. \end{cases} \tag{2}$$

The stochastic weighting mechanism prevents over-reliance on any single modality by enforcing dynamic feature recalibration. The following feature-concatenation operation is aligned with the implementation:

$$\begin{cases} \widetilde{\mathbf{X}}_a = \left[\mathbf{X}_a \| \mathbf{X}_a \odot \mathbf{C}_a \| \mathbf{0}^{b \times l_a \times d_a} \| \mathbf{X}_a\right] \in \mathbb{R}^{4b \times l_a \times d_a}, \\ \widetilde{\mathbf{X}}_v = \left[\mathbf{X}_v \| \mathbf{X}_v \odot \mathbf{C}_v \| \mathbf{X}_v \| \mathbf{0}^{b \times l_v \times c \times h \times w}\right] \in \mathbb{R}^{4b \times l_v \times c \times h \times w}, \\ \widetilde{\mathbf{y}} = [\mathbf{y} \| \mathbf{y} \| \mathbf{y} \| \mathbf{y}] \in \mathcal{Y}^{4b}. \end{cases} \tag{3}$$

The symbol $\odot$ denotes the Hadamard product (element-wise multiplication). The two concatenations are position-wise paired: the third entry deliberately combines a zeroed audio stream with intact visual input (audio-missing, visual-present), while the fourth is its visual-missing counterpart.

The implementation procedure is formally described in Algorithm 1.

3.3. Feature Extraction

3.3.1. Visual Feature Extraction

The visual branch comprises two key components: (1) frame-level visual feature extraction and (2) temporal modeling of video sequences. Departing from conventional approaches that decouple feature extraction (e.g., facial landmarks or action units) from multi-modal fusion [40,41], our end-to-end trainable model unifies these stages, enabling joint optimization of feature learning and fusion.

Our visual processing pipeline integrates frame-level feature extraction with temporal modeling in an end-to-end framework. Building on EfficientFace [42] for spatial feature extraction, we extend its architecture with temporal convolutional blocks to handle video sequences. Each input video is decomposed into 30 frames, which are independently

processed through 2D convolutions to generate frame-wise embeddings. These temporal features are then concatenated and fed through four 1D convolutional blocks (detailed in Table 1), each containing sequential operations of convolution, batch normalization, and Rectified Linear Unit (ReLU) [43] activation. This design achieves efficient temporal modeling while maintaining compatibility with pretrained 2D networks, and eliminates the computational overhead of 3D convolutions without sacrificing recognition accuracy under modality distortions.

Table 1. Architecture of the visual feature extraction branch.

Component	Operations
Feature extraction	EfficientFace Module [42]
Stage1	Reshape 2×[Conv1D(k3,d64,s1) + BN1D + ReLU]
Stage2	2×[Conv1D(k3,d128,s1) + BN1D + ReLU]
Prediction	Temporal Pooling + Linear

Conv: k3 (kernel size = 3), d64 (dimension = 64), s1 (stride = 1).

3.3.2. Heterogeneous Acoustic Representation Learner

The pursuit of optimal audio emotion recognition hinges on effective feature extraction, but faces two competing constraints: the need for high-quality representations versus the impracticality of large labeled datasets. Our framework addresses this divide through a novel synthesis of pretraining and feature compression, systematically transforming raw audio into compact yet discriminative features for downstream classification.

This solution emerges from carefully reconciling two approaches to feature engineering. Conventional techniques like MFCC [8] provide explicable, physics-inspired representations through mathematically defined spectral transformations, though their manual parameterization requires domain expertise. Modern data-driven alternatives such as BEATs [10] instead employ self-supervised pretraining to automatically discover generalizable patterns directly from waveforms, significantly reducing annotation needs. While spectrogram-based methods retain an advantage in interpretability, pretrained models offer greater data efficiency, illustrating the trade-off between manual feature engineering and neural adaptability.

To balance the asymmetry between visual and acoustic modalities, we develop a hybrid auditory feature extractor that combines data-driven and knowledge-based approaches. As shown in Figure 3, the architecture contains two complementary branches: the learned branch and the spectral branch.

The **learned branch** processes raw audio through a sequential pipeline. Input waveform segments are transformed into intermediate embeddings by a frozen pretrained acoustic encoder, either BEATs [10] or emotion2vec+ [44]: $\mathbf{a}_{\text{sem}} = \text{Enc}(x)$. These embeddings are then resampled to consistent dimensions and fed into a 1D convolution encoder. The two-stage pipeline produces discriminative semantic features $\mathbf{f}_{\mathbf{sem}}$ that offer the advantages of computationally efficient feature extraction, reduced dependence on large training sets, and flexible feature adaptation through subsequent tuning. Although freezing the BEATs weights makes feature extraction computationally efficient, the resulting features may exhibit domain-specific biases due to the mismatch between BEATs' general audio pretraining corpus and our target emotional audio dataset. This problem is mitigated by our complementary spectral processing path. To examine this domain bias directly, we additionally replaced the frozen BEATs encoder with emotion2vec+ [44], a self-supervised model pretrained on emotional speech, under an identical architecture and training protocol.

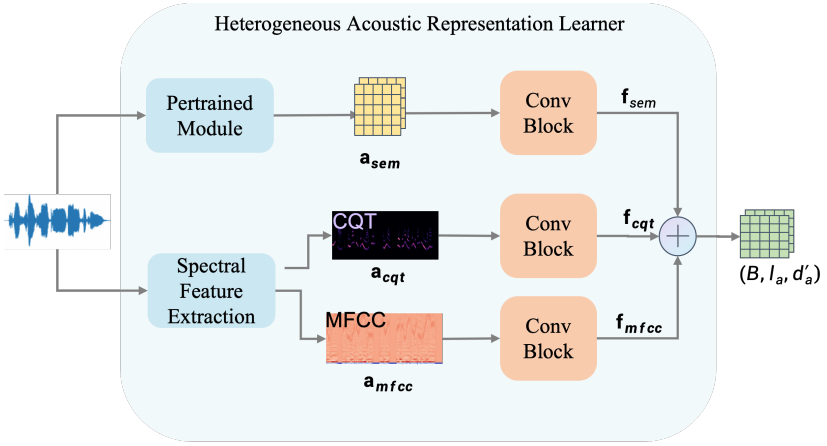


Figure 3. Heterogeneous Acoustic Representation Learner, an auditory feature extractor that combines spectral and learned representations.

The **spectral branch** processes traditional acoustic features (MFCC and CQT spectrograms) through dedicated CNN encoders, generating physically interpretable features f_{mfcc} and f_{cqt}. These engineered features compensate for potential biases in data-driven representations while maintaining computational efficiency. The implementation details are shown in Table 2.

Table 2. Architecture of the auditory feature extraction branch.

Component	Branch	Operations
Feature Extraction	BEATs [10]/emotion2vec+ [44] MFCC [8] CQT [9]	Raw waveform processing 10 coefficients 42 bins
Stage 1	All branches	Reshape 2×[Conv1D(k3,d64,s1) + BN1D + ReLU]
Fusion	Element-wise addition of all branch outputs	
Stage 2	2×[Conv1D(k3,d128,s1) + BN1D + ReLU]	
Prediction	Temporal Pooling + Linear	

Conv: k3 (kernel size = 3), d64 (dimension = 64), s1 (stride = 1).

3.3.3. Structurally Symmetrical Cross-Attention Feature Interaction

Given audio features $\mathbf{F}_a \in \mathbb{R}^{b \times l_a \times d_a}$ and visual features $\mathbf{F}_v \in \mathbb{R}^{b \times l_v \times d_v}$, we compute bidirectional cross-modal attention through two architectural variants.

Variant 1 implements a full-structured attention block that processes layer-normalized audio-visual features through a multi-component operator, which integrates single-head cross-attention with residual connections and MLP projection.

Variant 2 employs a minimal architecture where features are directly processed by a single-head attention layer followed solely by linear projection, deliberately excluding normalization and skip connections to isolate core attention effects.

3.3.4. Gumbel-Guided Attention Fusion

In this module, we aim to use Gumbel-Softmax to adaptively enhance both audio and visual features, whose architecture is illustrated in Figure 4. Unlike the previous methods [7], we leverage a symmetrical architecture: we use audio features to gate visual features and visual features to gate audio features, and then concatenate the gated audio and visual features to obtain the final fused feature.

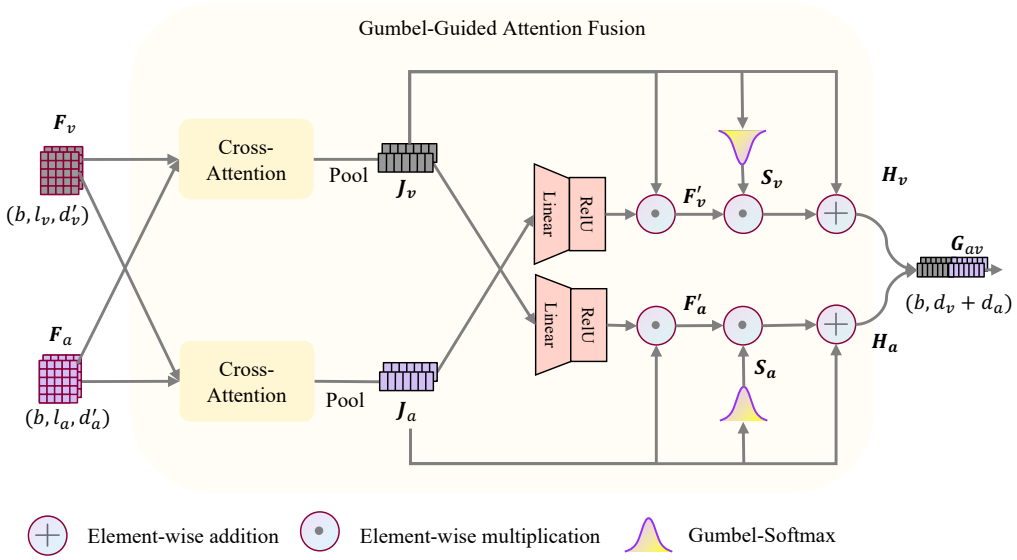


Figure 4. Architecture of the G^{2}AF module.

Specifically, the attention-enhanced features $\mathbf{h}_{av}$ and $\mathbf{h}_{va}$ are pooled temporally: $\mathbf{J}_a = \text{MeanPool}(\mathbf{h}_{av})$, $\mathbf{J}_v = \text{MeanPool}(\mathbf{h}_{va})$.

Step 1: Feature Interaction. We transform and interact features dimensionally:

$$\mathbf{F}'_a = \mathbf{J}_a \odot \text{ReLU}(\text{Linear}(\mathbf{J}_v)), \quad \mathbf{F}'_v = \mathbf{J}_v \odot \text{ReLU}(\text{Linear}(\mathbf{J}_a)). \tag{4}$$

Step 2: Gumbel-Softmax (GS) Sampling. Stochastic gates are sampled from the GS distribution:

$$\begin{aligned} \mathbf{S}_a \sim \text{GS}(\mathbf{J}_a; \tilde{\tau}), \qquad \mathbf{S}_v \sim \text{GS}(\mathbf{J}_v; \tilde{\tau}), \\ \tau(t) = \max\left(\tau_0 - \gamma\, t,\ \tau_{\min}\right), \qquad \tau_{\min} = 0.1, \\ \tilde{\tau} = \text{clip}\left(\tau(t),\, 0.1,\, 1.0\right), \end{aligned} \tag{5}$$

Here, t is the training step, $\tau_0 = 1.0$ is the initial temperature, and γ is the per-step decrement. Both streams share a single schedule, which is hand-designed and not trainable.

Step 3: Residual Fusion. The final features integrate stochastic selection and original signals. The output is the concatenation $\mathbf{G}_{av} = \text{Concat}[\mathbf{H}_v; \mathbf{H}_a]$, where $\mathbf{H}_a = \mathbf{F}'_a \odot \mathbf{S}_a + \mathbf{J}_a$, $\mathbf{H}_v = \mathbf{F}'_v \odot \mathbf{S}_v + \mathbf{J}_v$.

In addition, the network employs visual and audio auxiliary classifiers to enhance the main classifier through a unified multi-task framework. This architecture concurrently handles three interrelated tasks: audio-visual fusion (av), audio-only (a), and visual-only (v) processing. Each task $m \in \mathcal{T} = \{av, a, v\}$ operates through a dedicated branch containing shared feature extraction layers, task-specific parameters θ_m in deeper network stages, and independent classification heads that generate predictions $\hat{y}_m$. The task-specific loss L_m is the standard unweighted multiclass cross-entropy loss between the logits of task m and the shared ground-truth emotion label; no class weighting is applied.

To autonomously balance task priorities, we implement uncertainty-aware weighting through learnable parameters σ_m. These weights are embedded in the combined loss function:

$$L_{\text{comb}} = \sum_{m \in \mathcal{T}} \left(\frac{1}{2\sigma_m^2} L_m + \ln(1 + \sigma_m^2) \right), \tag{6}$$

where the $\frac{1}{2\sigma_m^2}$ term inversely scales each task's influence based on its learned uncertainty. Tasks with higher σ_m (*i.e.*, greater uncertainty) receive lower weighting during training. Simultaneously, the bounded regularizer $\ln(1 + \sigma_m^2)$ avoids the gradient-instability issues present in earlier formulations [45], creating a self-stabilizing optimization mechanism.

This design eliminates manual loss-weighting strategies while ensuring balanced multi-task learning.

The model jointly optimizes three parameter sets through gradient descent: shared backbone layers that capture cross-modal features, task-specific parameters θ_{av}, θ_a, θ_v that refine modality-discriminative patterns, and uncertainty weights σ_{av}, σ_a, σ_v that dynamically calibrate task importance. This coordinated optimization enables simultaneous improvement of all tasks, through which feature representations evolve to serve both shared and task-specific objectives. All uncertainty parameters are initialized as $\sigma_{av} = \sigma_a = \sigma_v = 1$ and optimized jointly with the network parameters using SGD without weight decay.

4. Experiments and Discussions

4.1. Dataset and Evaluation Metrics

The RAVDESS-Speech database contains 1440 audio clips recorded by 24 professional actors, annotated with eight distinct emotional states: neutral, calm, happy, sad, angry, fearful, disgusted, and surprised [46]. Unlike previous studies [47,48] that employed seven-class classification schemes, this research adopts an eight-category framework. To simulate real-world emotion recognition scenarios, we implemented a speaker-independent data partitioning strategy where training, validation, and test sets are strictly separated by speaker identities. This ensures that both the audio segments and the speakers in the test and validation sets remain unseen during training. After deduplication, the dataset comprises 1440 clips: 96 neutral and 192 for each of the other seven classes. We evaluate with an actor-wise 5-fold cross-validation in which all samples of a given actor remain within the same fold; the anonymised actor IDs assigned to each fold are {10, 11, 17, 19, 23}, {01, 03, 08, 14, 20}, {02, 04, 07, 12, 13}, {05, 06, 09, 15, 22} and {16, 18, 21, 24}, with male/female compositions of 1/4, 3/2, 3/2, 2/3 and 3/1, respectively, so that every fold contains actors of both sexes. Per-fold test splits contain 300 clips (folds 0-3) or 240 clips (fold 4). RAVDESS contains acted emotional expressions: each label denotes the emotion category a professional actor was instructed to portray using fixed lexical statements, so the ground truth is defined by the instructed category, validated by external perceptual raters [46].

The PKU Emotion Recognition Dataset (PKU-ER) comprises Chinese audio-visual recordings captured from 14 participants in human–computer interaction scenarios [49]. This balanced dataset categorizes emotions into three distinct classes (positive, negative, neutral) with uniform intensity levels and a nearly equal distribution of samples across categories. PKU-ER is a Chinese-language corpus collected in human–computer interaction scenarios; its cultural aspect refers to the language and interaction context, which complement the English, scripted RAVDESS. Since PKU-ER does not release per-clip speaker identities, a strictly subject-wise partition cannot be constructed; we therefore adopt a class-stratified 5-fold cross-validation in which each fold's class ratio closely matches the global distribution. Neither dataset captures spontaneous or self-reported emotional states. Consequently, the model learns to predict the annotation scheme of each dataset, which encodes acted expressions rather than internally experienced affect. For both datasets, we report the mean and standard deviation over the five folds together with the per-fold range [minimum, maximum], and we additionally report balanced accuracy (the mean per-class recall) to provide a reliable interpretation under any class imbalance. No class-balancing strategy (class-weighted loss, over-/under-sampling, or balanced batch construction) was applied during training on either dataset.

4.2. Experimental Settings

Each fusion model variant was trained using the two training schemes for enhancing modality robustness described in Section 3.2, with quantitative performance evaluations provided.

To avoid ambiguity and evaluate model robustness under simulated, real-world-inspired sensor degradation conditions, we explicitly separate the training-time augmentation schemes from the test-time corruption conditions. The three training schemes are: ODT (Original Data Training, no perturbation), NIT (Noise-Injected Training: feature-level Gaussian full-noise replacement of one modality per augmented copy, Section 3.2.1), and HM (Hybrid Masking: complete modality zero-masking combined with coefficient attenuation, Section 3.2.2). As shown in Table 3, the seven test-time conditions are uniformly named as follows: C0 (complete input), VN (visual full-noise replacement), VA (visual additive noise), VZ (visual zero-mask), and AN (audio full-noise replacement), AA (audio additive noise), AZ (audio zero-mask). For each test condition, one modality is corrupted by one of three operators. The three operators are: full-noise replacement, which replaces the entire modality input (visual tensor or audio feature tensor) with i.i.d. Gaussian noise $\mathcal{N}(0, \mathbf{I})$; additive noise, which adds i.i.d. Gaussian noise to the modality input; and zero-mask, which replaces the entire modality input with zeros. Each operator is applied to either the visual or the audio modality (the application domain), at one of several noise levels (the severity). In all seven conditions, every test sample of the evaluated fold is corrupted. NIT exposes the model to noise-type perturbations during training, so VN, VA, AN and AA are matched conditions for NIT whereas VZ and AZ are unseen; the reverse holds for HM, and robustness claims are stated accordingly.

In addition, we introduce a waveform-level acoustic-noise robustness protocol: additive white Gaussian noise is injected into the raw waveform after reading the audio file and before MFCC/CQT/emotion2vec+ extraction. For the additive-noise operator we use seven severity levels (SNR = clean, 20, 15, 10, 5, 0, -5 dB), and the noise is drawn with a fixed seed per file and per SNR level to guarantee reproducibility.

At the start of each training fold, seed 1 is reset for PyTorch 2.1.2 (CPU/CUDA) and NumPy 1.26.4 to control weight initialization, data shuffling, DIP augmentation, and Gumbel sampling; DataLoader workers use derived seeds for spatial augmentation. Experiments were run on a single NVIDIA GeForce RTX 4090 GPU (24 GB, CUDA 11.8) manufactured by NVIDIA Corporation (Santa Clara, CA, USA). Under the full NIT configuration, training and evaluation took on average 51.4 min per fold, approximately 4 h 17 min for five folds.

Table 3. Mapping between the training-time augmentation schemes and the seven test-time corruption conditions. For every condition the corruption operator, the application domain, the severity and the matched training scheme are specified. In all conditions except C0, every test sample of the evaluated fold is corrupted.

Condition	Corruption Operator	Application Domain	Severity	Matched Scheme
C0	none (complete input)	–	–	–
VN	full-noise replacement	visual tensor	i.i.d. $\mathcal{N}(0, \mathbf{I})$, whole tensor	NIT
VA	additive Gaussian noise	visual tensor	i.i.d. $\mathcal{N}(0, \mathbf{I})$ added	NIT
VZ	zero-mask	visual tensor	entire tensor set to zero	HM
AN	full-noise replacement	audio feature tensor	i.i.d. $\mathcal{N}(0, \mathbf{I})$, whole tensor	NIT
AA	additive Gaussian noise	audio feature tensor	i.i.d. $\mathcal{N}(0, \mathbf{I})$ added	NIT
AZ	zero-mask	audio feature tensor	entire tensor set to zero	HM

NIT exposes the model to noise-type perturbations during training, so VN, VA, AN and AA are matched conditions for NIT whereas VZ and AZ are unseen; the reverse holds for HM. The audio conditions act on the extracted acoustic feature tensor, whereas the waveform-level SNR protocol of Section 4.2 corrupts the raw waveform before feature extraction.

4.3. Performance Under Clean Settings

This subsection reports clean-condition results under two evaluation protocols, which are deliberately kept apart because they differ in the acoustic backbone, the visual sampling rate and the evaluation protocol; their numbers are therefore not directly comparable. Table 4 uses the traditional setting: frozen BEATs acoustic backbone, one held-out test split per dataset. This setup follows the experimental protocol adopted in [6,49,50]. Table 5 adopts a different setting: the emotion2vec+ acoustic backbone, trained with NIT under 5-fold cross-validation. For each fold it reports the accuracy, and additionally summarizes the results as mean $\pm$ std, the min–max range across the five folds, and balanced accuracy.

Effect of the training scheme. In Table 4, NIT improves the fused (AV) accuracy over ODT on both datasets and for both variants: by 6.24% and 2.08% for Variant 1 and Variant 2 on RAVDESS, and by 3.52% and 13.07% on PKU-ER. Injecting noise during training therefore acts as a regularizer that also pays off on clean test data. HM behaves differently across datasets: on RAVDESS it reduces the AV accuracy of Variant 1 by 12.92% relative to ODT, whereas on PKU-ER it raises it to 90.45%, the best Variant 1 result on that dataset.

Effect of the architecture. In Table 4, under NIT, Variant 2 clearly outperforms Variant 1 on PKU-ER (94.47% versus 89.44%), while Variant 1 is the stronger choice on RAVDESS (86.66% versus 81.66%). The advantage is specific to the fused head: under ODT the visual-only head of Variant 1 reaches 90.45% on PKU-ER, well above the 78.89% of Variant 2. The two variants therefore prioritise different feature-learning patterns, and each dataset is best served by a different main configuration, which is why the main configuration of RAVDESS is Variant 1 and that of PKU-ER is Variant 2 in the remainder of this paper. The corresponding ablation results are reported in Appendix A.

Modality dominance and dispersion. Table 5 quantifies the modality distribution under the 5-fold protocol. On RAVDESS the audio-only head (85.98%) clearly exceeds the visual-only head (61.80%), so the dataset is audio-dominant; on PKU-ER the ordering is reversed (visual 92.01% versus audio 69.14%). This difference explains the divergent behavior of HM reported above: on an audio-dominant dataset, masking a whole modality removes information that the remaining stream cannot replace, whereas on a visual-dominant dataset the same operation acts as a regularizer that forces cross-modal compensation. Appendix B confirms this with the seven-condition corruption study, where damaging audio costs the fused model 30% to 50% on RAVDESS while damaging video costs only 2% to 7%. Balanced accuracy is close to overall accuracy on both datasets (90.69% versus 91.08% on RAVDESS; 93.43% versus 93.45% on PKU-ER), showing that the reported accuracy is not carried by a single dominant class. The fold-to-fold spread is larger on RAVDESS ([83.75%, 95.00%]) than on PKU-ER ([92.18%, 96.11%]), which reflects the actor-wise partition: every RAVDESS fold is evaluated on unseen speakers, and fold 4 contains only four actors.

For Table 5, Variant 1 with NIT is evaluated on RAVDESS and Variant 2 with NIT on PKU-ER. Both use one Gumbel-sampling fusion head and auxiliary A and V heads; visual input comprises 30 MTCNN-cropped 224 $\times$ 224 frames, and audio input comprises MFCC (n = 10), CQT (n_{bins} = 42), and a frozen emotion2vec+ embedding. RAVDESS uses actor-wise 5-fold validation with 24 actors grouped as whole persons (300 test clips in folds 0–3; 240 in fold 4). PKU-ER uses class-stratified 5-fold validation because per-clip speaker identities are unavailable. Values are held-out-fold accuracies; BA denotes fold-level mean per-class recall, subsequently averaged over folds. Appendix C provides the corresponding per-class recalls and confusion matrices.

Five-fold ablation. We conducted a new batch of 80 runs (16 configurations $\times$ five actor-wise RAVDESS folds) under the 30-frame, emotion2vec+ protocol. The fusion and acoustic-representation choices were varied independently, and the one-way interaction controls used both the original depth and deeper, capacity-compensated counterparts.

All comparisons use identical fold assignments. Table 6 reports the compact summary; Appendix A defines every configuration and provides all 80 fold-level records, including balanced accuracy. The full reference was independently rerun within this batch, so its result is not mixed with Table 5.

Table 4. Accuracy of the proposed method: a comparison across its two variants and three training schemes on the RAVDESS and PKU-ER datasets in clean settings.

	Dataset	RAVDESS			PKU-ER		
Variant	**Training Scheme**	**AV**	**A**	**V**	**AV**	**A**	**V**
Var 1	ODT (Baseline)	80.42	68.33	**70.83**	85.92	**78.39**	90.45
	NIT (Noise-Injected)	**86.66**	**70.83**	70.41	89.44	75.88	90.95
	HM (Hybrid Masking)	67.50	55.83	62.92	90.45	77.38	**93.96**
Var 2	ODT (Baseline)	79.58	67.91	68.75	81.40	74.37	78.89
	NIT (Noise-Injected)	81.66	65.83	67.08	**94.47**	77.89	92.96
	HM (Hybrid Masking)	82.08	70.82	70.00	87.93	66.33	84.42

Configuration. Datasets: RAVDESS, 8 classes; PKU-ER, 3 classes. Architecture: Variant 1 (full-structured cross-attention block with layer normalisation, residual connections and MLP projection) and Variant 2 (single-head attention followed by linear projection), both with Gumbel-sampling fusion and one attention head; each variant emits three prediction heads AV, A and V. Training: SGD (momentum 0.9, weight decay 10^{-3}), initial lr 0.04, ReduceLROnPlateau (patience 10), batch size 4, 100 epochs; Gumbel temperature annealed from $\tau_0 = 1.0$ to 0.1 during training, $\tau = 0.1$ at evaluation; every cell is an independent training run under the indicated scheme (ODT/NIT/HM) and the best-validation checkpoint is evaluated on the held-out test split. Visual input: 15 frames sampled per clip, MTCNN face crop, 224 × 224, EfficientFace initialized from AffectNet7. Audio input: 3.6 s waveforms; MFCC ($n = 10$) + CQT ($n_{\text{bins}} = 42$) + BEATs embedding (768) = 820 channels × 156 frames; the BEATs encoder is frozen. **Protocol and metrics.** One held-out split per dataset: RAVDESS speaker-independent (train, validation and test actors disjoint); PKU-ER split by clip, since per-clip speaker identities are not released. This is a single-split protocol, hence no dispersion across splits is reported. All values are test accuracies (%); AV = fused head, A = audio-only head, V = visual-only head; bold marks the best value in each column.

Table 5. Five-fold cross-validation results of the proposed method with 30-frame visual input and the emotion2vec+ acoustic backbone: RAVDESS (actor-wise 5-fold, Variant 1 + NIT) and PKU-ER (stratified 5-fold, Variant 2 + NIT). BA = balanced accuracy. No class-balancing strategy was used during training.

Dataset	**Fold**	**AV**	**A**	**V**	**AV BA**
RAVDESS	0	95.00	87.33	60.67	95.31
	1	93.67	87.67	69.33	94.06
	2	89.67	82.67	61.00	89.38
	3	93.33	91.00	59.67	93.44
	4	83.75	81.25	58.33	81.25
	mean ± std [range]	91.08 ± 4.55 [83.75, 95.00]	85.98 ± 3.98	61.80 ± 4.34	90.69 ± 5.73
PKU-ER	0	93.96	73.08	91.21	93.82
	1	92.82	67.40	93.37	92.78
	2	96.11	68.33	93.33	96.08
	3	92.18	69.27	88.27	92.17
	4	92.18	67.60	93.85	92.30
	mean ± std [range]	93.45 ± 1.66 [92.18, 96.11]	69.14 ± 2.32	92.01 ± 2.33	93.43 ± 1.62

Configuration. Datasets: RAVDESS, 8 classes; PKU-ER, 3 classes. Architecture: the main configuration of each dataset, i.e., Variant 1 on RAVDESS and Variant 2 on PKU-ER, both with Gumbel-sampling fusion and one attention head; each emits three prediction heads AV, A and V, where A and V are the auxiliary single-modality heads branching off before the fusion module, not separately trained unimodal models. Training scheme: NIT (Noise-Injected Training); SGD (momentum 0.9, weight decay 10^{-3}), initial lr 0.04, ReduceLROnPlateau (patience 10), batch size 4, 100 epochs; Gumbel temperature annealed from $\tau_0 = 1.0$ to 0.1 during training, $\tau = 0.1$ at evaluation; the best-validation checkpoint of each fold is evaluated on that fold's held-out test split; no class-balancing strategy was used. Visual input: 30 frames sampled per clip, MTCNN face crop, 224 × 224, EfficientFace initialized from AffectNet7. Audio input: 3.6 s waveforms; MFCC ($n = 10$) + CQT ($n_{\text{bins}} = 42$) + emotion2vec+ embedding (768) = 820 channels × 156 frames; the emotion2vec+ encoder is frozen. **Protocol and metrics.** All folds use identical hyper-parameters and temperature schedules. RAVDESS is partitioned by actor and PKU-ER by class-stratified clips, because per-clip speaker identities are unavailable for PKU-ER. For fold k, F_k is used for testing, the adjacent group $F_{(k+1) \bmod 5}$ for validation, and the remaining groups for training; validation therefore uses actor groups for RAVDESS and clip-level groups for PKU-ER. The training seed is fixed at 1. Reported values are held-out-fold test accuracies (%), summarized across folds as mean ± sample standard deviation with the accuracy range. Balanced accuracy (BA) is calculated within each fold as mean per-class recall and then averaged across folds. No training sample is used for evaluation.

Table 6. Five-fold RAVDESS ablation and one-way interaction controls. AV, A, and V denote the accuracies of the audio–visual fusion, audio-only, and visual-only classification heads, respectively. Values are reported as mean ± sample standard deviation over the five actor-wise folds. ΔAV is the fused-head accuracy difference from the independently rerun full reference (P1-1), in percentage points. IDs correspond to the complete configuration key and fold-level records in Appendix A.

ID	Controlled Configuration	AV	A	V	ΔAV
P1-1	Full reference: G^2AF, full HAR, NIT, two-way, depth 1	91.32 ± 3.53	87.65 ± 3.55	59.23 ± 3.11	–
P1-3	Attention-only fusion (without Gumbel gating)	89.30 ± 4.10	87.33 ± 4.06	60.80 ± 7.22	−2.02
P1-5	Concatenation fusion (without G^2AF)	89.88 ± 6.02	87.73 ± 5.05	62.00 ± 6.26	−1.43
P1-2	Learned-only HAR (without handcrafted branch)	90.47 ± 4.73	85.13 ± 2.23	60.33 ± 5.91	−0.85
P4-1	Handcrafted-only HAR (without learned branch)	74.60 ± 6.22	42.63 ± 4.17	68.55 ± 6.43	−16.72
P2-1	ODT (without DIP)	89.53 ± 2.98	85.48 ± 3.70	60.82 ± 9.64	−1.78
P3-1	AV-only interaction, depth 1	91.07 ± 2.22	86.08 ± 4.57	63.40 ± 3.55	−0.25
P3-2	AV-only interaction, depth 2 (capacity-compensated)	91.08 ± 4.02	87.33 ± 2.79	56.90 ± 7.14	−0.23
P3-3	VA-only interaction, depth 1	89.13 ± 5.14	87.37 ± 3.67	62.00 ± 7.44	−2.18
P3-4	VA-only interaction, depth 4 (capacity-compensated)	91.05 ± 3.33	86.62 ± 3.41	56.73 ± 11.13	−0.27

Configuration. All rows belong to the same new 80-run batch and use 30-frame visual input and identical actor-wise fold assignments. The full reference is therefore reported independently of the earlier five-fold result in Table 5. AV-only and VA-only follow the interaction-path notation of Section 3. The deeper one-way variants provide depth-based capacity compensation; exact parameter equality is not assumed. **Datasets:** RAVDESS, 8 classes. **Training:** SGD (momentum 0.9, weight decay 10^{-3}), initial lr 0.04, decayed by a factor of 10 at epochs 40, 55, 65 and 70; batch size 4, 100 epochs; Gumbel temperature annealed from $\tau_0 = 1.0$ to 0.1 during training, $\tau = 0.1$ at evaluation; every cell is an independent training run under the indicated scheme (fusion/acoustic composition/interaction direction/perturbation scheme), and the best-validation checkpoint is evaluated on the held-out test split. **Visual input:** 30 frames sampled per clip, MTCNN face crop, 224 × 224, EfficientFace initialized from AffectNet7. **Audio input:** 3.6 s waveforms; MFCC ($n = 10$) + CQT ($n_{\text{bins}} = 42$) + emotion2vec+ embedding (768) = 820 channels × 156 frames; the emotion2vec+ encoder is frozen. **Protocol and metrics:** Actor-independent 5-fold cross-validation: for fold k, testing = F_k, validation = $F_{(k+1) \bmod K}$, and the remaining actors are used for training; only the clean (both-modality) test condition is evaluated. All values are test accuracies (%); AV = fused head, A = audio-only head, V = visual-only head; per-fold accuracy and balanced accuracy are reported and summarized as mean ± sample standard deviation across the five folds; the training seed is fixed (seed = 1) and shared by all cells.

With full HAR and NIT fixed, replacing G^2AF with attention-only or concatenation fusion reduces mean AV accuracy by 2.02 and 1.43 points, respectively; the paired fold-level differences are not statistically significant (two-sided paired t-test: $p = 0.113$ and $p = 0.583$). With G^2AF and NIT fixed, learned-only HAR trails full HAR by 0.85 points ($p = 0.553$), whereas handcrafted-only HAR trails it by 16.72 points, with lower AV accuracy in every fold. The latter comparison gives $p = 0.013$ with a paired t-test and $p = 0.0625$ with a Wilcoxon signed-rank test. It also reduces audio-only accuracy from 87.65% to 42.63%, identifying the pretrained acoustic branch as the largest numerical contributor in this comparison.

At depth 1, the AV-only and VA-only variants obtain 91.07% and 89.13% AV accuracy, respectively, compared with 91.32% for the two-way reference. With depth-based capacity compensation, the AV-only (depth 2) and VA-only (depth 4) variants reach 91.08% and 91.05%. The corresponding numerical two-way differences are only 0.23 and 0.27 points ($p = 0.854$ and $p = 0.737$). Furthermore, the cross-attention modules contain 157.5 k parameters in the two-way reference, 115.8 k and 231.6 k in AV-only depths 1 and 2, and 41.7 k and 166.8 k in VA-only depths 1 and 4. These counts concern only the cross-attention modules, not total system computation or training efficiency. Overall, these results support the two-way structure as an effective fusion strategy with a compact cross-attention design, suggesting that its performance is not simply attributable to increased model capacity.

4.4. Performance Under Modality Distortion Settings

Tables 7 and 8 show how the three training schemes respond to visual and audio test-time corruption, respectively.

Table 7. Full-head classification accuracy under visual corruption on the RAVDESS test set. VN, VA, and VZ denote visual full-noise replacement, additive noise, and zero-masking, respectively; audio input remains intact.

Variant	Training Scheme	VN			VA			VZ		
		AV	A	V	AV	A	V	AV	A	V
Var 1	ODT (Baseline)	32.08	68.33	13.33	42.91	68.33	20.00	45.41	68.33	13.33
	NIT (Noise-Injected)	74.58	70.83	36.66	65.41	70.83	36.66	39.16	70.83	13.33
	HM (Hybrid Masking)	24.58	55.83	13.33	60.00	55.83	42.50	48.75	55.83	13.33
Var 2	ODT (Baseline)	35.83	67.91	13.33	66.67	67.91	20.00	50.00	67.91	13.33
	NIT (Noise-Injected)	71.25	65.83	13.33	76.25	65.83	48.75	32.50	65.83	13.33
	HM (Hybrid Masking)	39.58	70.83	13.75	73.33	70.83	39.58	74.16	70.83	13.33

Table 8. Full-head classification accuracy under audio corruption on the RAVDESS test set. AN, AA, and AZ denote audio full-noise replacement, additive noise, and zero-masking, respectively; visual input remains intact.

Variant	Training Scheme	AN			AA			AZ		
		AV	A	V	AV	A	V	AV	A	V
Var 1	ODT (Baseline)	45.41	11.25	70.83	48.33	13.75	70.83	32.08	13.33	70.83
	NIT (Noise-Injected)	73.75	16.66	70.41	55.00	15.83	70.41	63.75	13.33	70.41
	HM (Hybrid Masking)	23.75	13.33	62.92	24.16	13.33	62.92	65.41	13.33	62.91
Var 2	ODT (Baseline)	22.08	12.91	68.75	17.08	12.91	68.75	13.75	13.33	68.75
	NIT (Noise-Injected)	69.16	12.50	67.08	39.16	16.66	67.08	50.41	13.33	67.08
	HM (Hybrid Masking)	13.33	13.33	70.00	13.33	13.33	70.00	70.83	13.33	70.00

Both tables use the traditional RAVDESS protocol: a frozen BEATs acoustic backbone, 15-frame visual input, and one speaker-independent held-out test split. ODT, NIT, and HM denote Original Data Training, Noise-Injected Training, and Hybrid Masking. VN and AN replace the affected input with i.i.d. Gaussian noise $\mathcal{N}(0, \mathbf{I})$; VA and AA add the same noise; VZ and AZ replace it with zeros. AV, A, and V are the fusion, audio-only, and visual-only heads, respectively.

No single scheme dominates in Tables 7 and 8, and the ranking follows the type of corruption. NIT is strongest when one modality is entirely replaced by noise, because this is the perturbation it saw during training: Variant 1 rises from 32.08% (ODT) to 74.58% under visual replacement and from 45.41% to 73.75% under audio replacement. HM is strongest when one modality is removed altogether, because it forces cross-modal compensation: Variant 2 reaches 74.16% under visual zero-masking and 70.83% under audio zero-masking, against 32.50% and 50.41% for NIT. Under additive noise, where the damage is partial, the two schemes are comparable and their ordering depends on which modality is attacked. The training scheme should therefore be selected for the distortion it is expected to meet, rather than treated as a general-purpose regularizer.

4.5. Comparison with State-of-the-Art Methods

Tables 9 and 10 summarize recognition accuracy and reported trainable parameter counts on RAVDESS and PKU-ER. We distinguish the single-split comparisons marked with † from our five-fold evaluations. Results obtained with different data partitions, acoustic backbones, or visual sampling settings provide context but do not establish a ranking under a common protocol. The reported human recognition accuracy is included as a reference, not as a matched human–model evaluation.

Table 9. Performance of DRAV-GON and state-of-the-art methods on the RAVDESS dataset.

Model	Year	Venue	Accuracy	Params (Trainable)
MulT [40] †	2019	ACL	78.50%	-
MMTM [51] †	2020	CVPR	73.12%	31.97 M
MSAF [52] †	2020	-	74.86 %	25.94 M
TAV [7] †	2020	ICASSP	77.75%	-
TVA [7] †	2020	ICASSP	76.00%	-
ERANNs [53] †	2022	CVPR	74.80%	23.60 M
John et al. [54] †	2022	ICPR	51.01%	-
Franceschini et al. [55] †	2022	ICPR	78.54%	-
Chumachenko et al. [6] †	2022	ICPR	81.58%	-
Guo et al. [49] †	2022	CAAI	78.45%	18.40 M
AttA-Net [50] †	2024	ICASSP	82.62%	2.33 M
AttA-Net (non-end-to-end) [50] †	2024	ICASSP	88.00%	0.06 M
Dabbabi et al. [56] (subject-dependent)	2024	J.Syst.Sci.Syst.Eng	77.87%	-
Human performance [46]	2018	PLOS ONE	80.00%	-
Our method † (8 classes, subject-independent)	2026	-	86.67%	2.25 M
Our method (8 classes, subject-independent, actor-wise 5-fold)	2026	-	91.08% ± 4.55%	2.25 M

† configuration using the same and specified training, validation, and test sets, and the sets are subject-independent. The DRAV-GON row marked with † reports single-split accuracy. The actor-wise five-fold row reports mean ± std (per-fold range [83.75, 95.00]) under the separate emotion2vec+ and 30-frame protocol. These rows should not be used interchangeably when comparing methods. Human recognition accuracy is reported under the source study's evaluation conditions.

The listed RAVDESS results include 78.50% for MulT [40], 73.12% for MMTM [51], and 77.75% and 76.00% for TAV and TVA [7], respectively. Chumachenko et al. [6] report 81.58%, while the Conformer-based network [57] of Guo et al. [49] reaches 78.45% with 18.40 M trainable parameters. Franceschini et al. [55] report 78.54%. Middya et al. [48] report 86%, but their evaluation uses a 10% test split without ensuring speaker independence between training and test sets. That result is not directly comparable with our speaker-independent evaluations.

Recent advances have demonstrated an increasing emphasis on balancing model efficiency and accuracy. AttA-Net [50] introduces adaptive attention mechanisms to achieve 82.62% accuracy with only 2.33 M parameters. Meanwhile, Dabbabi et al. [56] leverage large-model distillation to reach 77.87% accuracy.

In the single-split comparison marked with †, DRAV-GON attains 86.67% on RAVDESS, surpassing the end-to-end AttA-Net (82.62%). The non-end-to-end AttA-Net achieves a marginally higher 88.00%, which we attribute to its more sophisticated feature-extraction pipeline. Our separate actor-wise five-fold evaluation yields 91.08% ± 4.55% (per-fold range $[83.75, 95.00]$), measured over held-out actor groups. The reported human recognition accuracy of 80.00% is included only as contextual information, because the human and model evaluations are not matched and do not support a claim of model superiority over human performance. Moreover, DRAV-GON contains only 2.25 M trainable parameters, in contrast to 31.97 M for MMTM and 2.33 M for the end-to-end AttA-Net. These figures characterize the trainable model size and our method is advantageous in this regard as well.

On PKU-ER, the single-split results marked with † in Table 10 are 94.47% for DRAV-GON, 92.60% for AttA-Net, and 90.53% for Guo et al [49]. These values support a numerical improvement within the stated single-split comparison; no statistical significance claim is made from this split alone. Under the separate stratified five-fold protocol, DRAV-GON achieves 93.45% ± 1.66% (per-fold range [92.18, 96.11]; best fold 96.11%). This result summarizes variability across clip-level partitions and should not be used to claim superiority over the listed single-split baselines. The model has 2.25 M trainable pa-

rameters; its accuracy and parameter count are reported separately from claims about runtime efficiency.

Table 10. Performance of DRAV-GON and state-of-the-art methods on the PKU-ER dataset.

Model	Year	Venue	Accuracy	Params (Trainable)
MMTM [51] †	2020	CVPR	83.58%	31.97 M
MSAF [52] †	2020	-	83.12%	25.94 M
ERANNs [53] †	2022	CVPR	86.94%	23.60 M
Guo et al. [49] †	2022	CAAI	90.53%	18.40 M
AttA-Net [50] †	2024	ICASSP	92.60%	2.33 M
Our method † (3 classes, subject-dependent)	2026	-	94.47%	2.25 M
Our method (3 classes, subject-dependent, stratified sampling)	2026	-	93.45% ± 1.66%	2.25 M

† configuration using the same and specified training, validation, and test sets. The DRAV-GON row marked with † reports single-split accuracy. The stratified five-fold row reports mean ± std (per-fold range [92.18, 96.11]; best fold 96.11%) under the separate emotion2vec+ and 30-frame protocol and is not a matched comparison with the single-split results. PKU-ER does not release per-clip speaker identities; hence, it is subject-dependent and the stratified protocol is used (Section 4.1).

4.6. Visualization Results

To assess the effectiveness of the G^2AF module, we present two representative cases comparing DRAV-GON's performance with its auxiliary unimodal classifiers, as illustrated in Table 11. While actors in the dataset co-created emotional responses through interaction, their actual emotional manifestations often varied across modalities. The results demonstrate that when unimodal classifiers produce inconsistent predictions, DRAV-GON successfully synthesizes cross-modal information to achieve accurate emotion recognition. These examples are qualitative illustrations only; the aggregate behavior is quantified over the complete test set below.

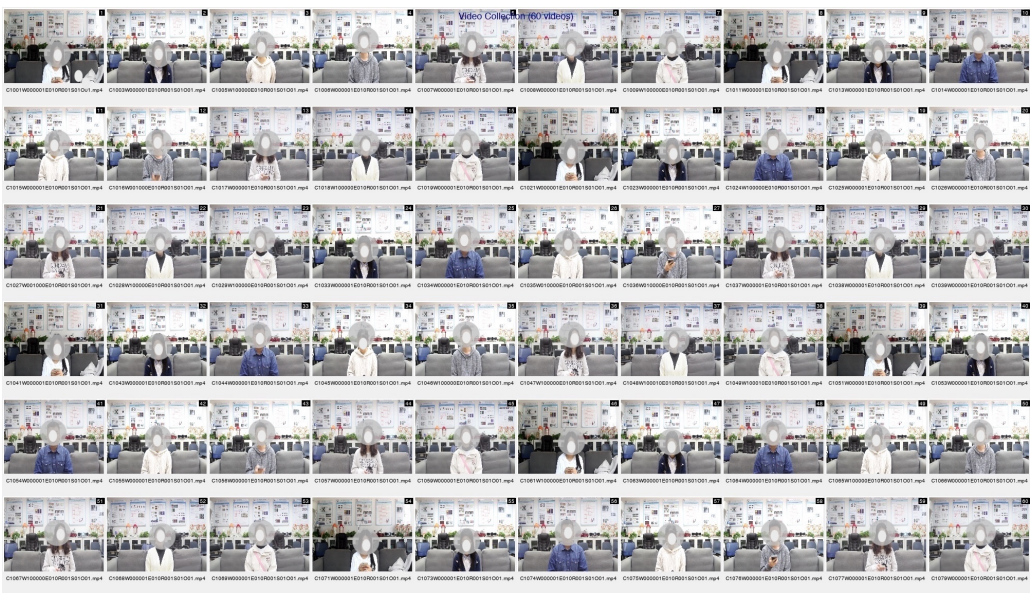
To characterize how the fused classifier resolves disagreement between the auxiliary unimodal classifiers, we analyzed the per-sample predictions of all three heads over the complete test sets of both datasets (RAVDESS: Variant 1 + NIT, actor-wise 5-fold, all clean test predictions; PKU-ER: Variant 2 + NIT, stratified 5-fold, all clean test predictions). As reported in Table 12, on RAVDESS the audio and visual heads disagree on 46.39% of the samples; when they disagree, the fused head is still correct on 85.63% of them (rescue rate), and when they agree the fused accuracy reaches 96.89%. On PKU-ER the disagreement rate is 32.85% with a rescue rate of 86.49%. The fusion head flips a correct unimodal prediction into an incorrect one on only 5.63% (RAVDESS) and 3.44% (PKU-ER) of all samples. These results quantify the contribution of multimodal fusion beyond anecdotal cases, and the rescue behavior is consistent across both datasets despite their opposite modality dominance.

To further evaluate the effectiveness of the proposed HAR-Learner method, this study conducted t-SNE visualizations [58] of the features from both the visual branch and the audio branch of our method trained on the RAVDESS dataset with the traditional protocol and the ODT training scheme. Different colors and numbers represent different emotion categories, with eight distinct colors and Arabic numerals from 0 to 7 indicating the corresponding emotion classes. The t-SNE visualization reveals the clustering results for different emotions, where overlapping points between clusters suggest potential ambiguity in emotion discrimination.

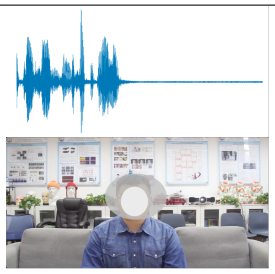
As shown in Figure 5a,b, when only traditional MFCC [8] is used as the audio feature extraction method without incorporating our proposed spectral learning-based approach, the visual features demonstrate more pronounced inter-class clustering compared to the audio features. With our proposed spectral learning-based auditory feature extraction method, the audio features exhibit significantly improved clustering, as illus-

trated in Figure 5c,d, reducing the gap in clustering performance between audio and visual features.

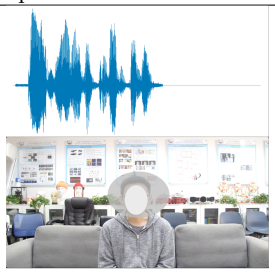
Table 11. Case study for DRAV-GON on the PKU-ER dataset.



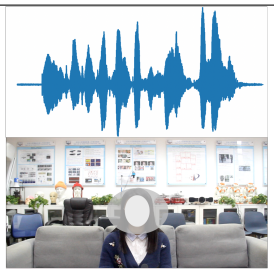
Samples from the PKU-ER Dataset.



Label:
Positive
DRAV-GON:
Positive
Audio classifier:
Neutral
Visual classifier:
Positive



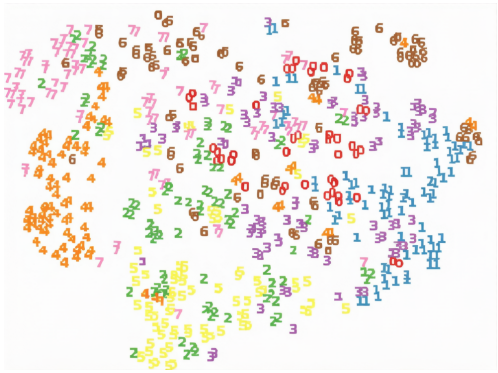
Label:
Neutral
DRAV-GON:
Neutral
Audio classifier:
Neutral
Visual classifier:
Negative



Label:
Negative
DRAV-GON:
Negative
Audio classifier:
Neutral
Visual classifier:
Negative

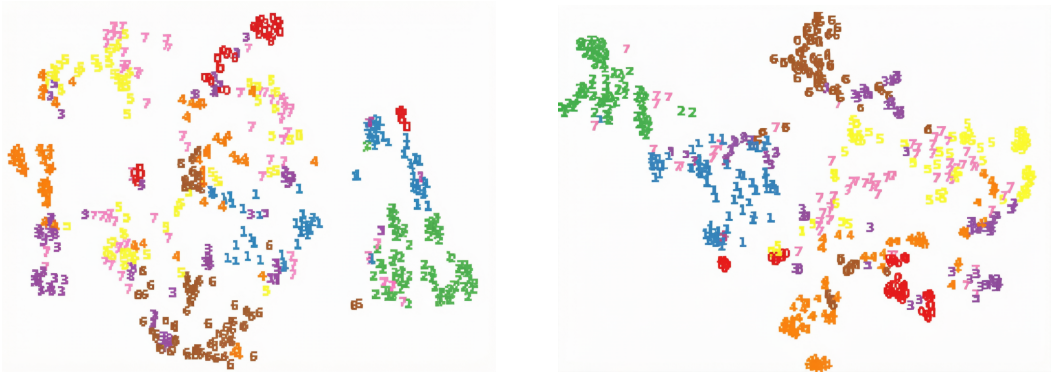


(**a**)



(**b**)

Figure 5. *Cont.*



(**c**) (**d**)

Figure 5. t-SNE feature visualization comparison between the baseline and proposed methods. (**a**,**b**) Baseline method: audio features show weaker inter-class clustering than visual features. (**c**,**d**) Proposed method: audio feature clustering is significantly improved, narrowing the gap with visual features. (**a**) t-SNE visualization of visual features under the baseline method. (**b**) t-SNE visualization of audio MFCC features under the baseline method. (**c**) t-SNE visualization of visual features under the proposed method. (**d**) t-SNE visualization of audio features extracted by the proposed HAR-Learner. In all panels, each color corresponds to one of the eight emotion categories: red for neutral (0), blue for calm (1), green for happy (2), purple for sad (3), orange for angry (4), yellow for fear (5), brown for disgust (6), and pink for surprise (7).

Table 12. Quantitative fusion-behavior analysis over the complete test sets of RAVDESS (Variant 1 + NIT, actor-wise 5-fold) and PKU-ER (Variant 2 + NIT, stratified 5-fold), clean condition.

Metric	RAVDESS	PKU-ER
A/V disagreement rate	46.39%	32.85%
Rescue rate (AV correct despite disagreement)	85.63%	86.49%
AV accuracy under agreement	96.89%	97.02%
AV accuracy under disagreement	85.63%	86.49%
Complementarity gain	43.33%	30.72%
Fusion flips correct $\rightarrow$ wrong	5.63%	3.44%

5. Conclusions

We presented DRAV-GON for audio-visual emotion recognition with heterogeneous representations and structurally symmetric cross-modal interaction. HAR-Learner combines spectral descriptors and pretrained embeddings to enrich acoustic representations. G^2AF leverages a symmetrical architecture, which consists of bidirectional attention, reciprocal modulation, and gated residual processing, to dynamically fuse audio and visual features. DIP exposes the trainable pipeline to degraded inputs.

Our model achieves mean accuracies of 93.45%$\pm$1.66% on PKU-ER and 91.08%$\pm$4.55% on RAVDESS under 5-fold cross-validation. In the independent ablation batch, removing the pretrained acoustic branch produces the largest observed performance reduction.

Limitations and future work: The reported experiments cover controlled, acted-emotion datasets and simulated input corruption. They do not establish performance under the range of lighting, pose, occlusion, accents, recording devices, or multi-person scenes encountered in deployment. The labels describe each dataset's annotated expression categories, and transfer to spontaneous emotion remains untested. Some negative categories have lower recall than others, motivating closer study of class-specific errors and finer-grained affective labels. PKU-ER lacks per-clip speaker identities, so its stratified clip-level evaluation cannot establish subject-independent generalization.

Robustness also depends on the corruption and training scheme. Noise-injected training performs well on noise replacement, while hybrid masking is better suited to complete zero-masking in the reported single-split experiments. In the waveform-noise tests, fused prediction can fall below the intact visual-only head when audio is severely degraded. The stochastic gates therefore do not guarantee selection of the most reliable modality in every condition.

Structural symmetry is defined at the level of paired interaction paths. In the controlled comparison with depth-based capacity compensation, the bidirectional model achieved the highest mean accuracy, but its small numerical gains over the unidirectional controls were not statistically significant on RAVDESS. It also required fewer cross-attention parameters than either capacity-compensated unidirectional counterpart, suggesting that its performance was not simply a result of increased model capacity. These findings thus point to structural symmetry chiefly as a compact fusion design that preserves accuracy comparable to the matched one-way controls, rather than as a source of measurable accuracy gains. The parameter comparison is limited to the cross-attention modules and should not be interpreted as evidence of lower total system computation or improved training efficiency. Additional seeds, datasets, and corruption conditions are needed to assess the generality of these findings.

Supplementary Materials: The following supporting information can be downloaded at: https://github.com/NariFan/DRAVGON, accessed on 19 November 2026.

Author Contributions: Conceptualization, R.F. and Y.C.; methodology, R.F., Y.C., G.W. and W.W.; software, R.F.; validation, R.F., Y.C., G.W. and W.W.; formal analysis, R.F.; investigation, R.F.; resources, Y.C.; data curation, R.F.; writing—original draft preparation, R.F.; writing—review and editing, R.F., Y.C., G.W. and W.W.; visualization, R.F.; supervision, Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China (No. 2025YFE0113400) and the National Natural Science Foundation of China (Grant No. T2600728).

Data Availability Statement: This study used the RAVDESS and PKU-ER datasets. RAVDESS is available at https://doi.org/10.5281/zenodo.1188976. PKU-ER was originally introduced in [49] and is not deposited in a public repository; it is available from the corresponding author upon reasonable request, subject to the dataset provider's terms of use. The source code is available in the Supplementary Materials.

Acknowledgments: Kimi 3.1 Pro (Moonshot AI), an artificial intelligence (AI) assistant, was used during manuscript writing to assist with writing the abstract, improving grammar and wording, and preparing and checking the LaTeX layout. Citations to the tool identify the affected sections; they acknowledge writing assistance and do not provide evidence for scientific claims. The authors remain responsible for the scientific content, the accuracy of the reported results, and the final manuscript.

Conflicts of Interest: Author Ruijia Fan was employed by Grid Planning & Research Center Guangdong Power Grid Corporation Guangzhou, China. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A. Five-Fold Ablation Study and Complete Records

The ablation uses the actor-wise five-fold RAVDESS protocol. Table A1 defines the 16 configurations, and Table A2 reports all 80 fold-level results. The P1 configurations independently vary fusion and acoustic representation, P2 varies the training scheme, P3 compares one-way and two-way interaction, and P4 provides additional acoustic, fusion, training, and directionality controls. All results are measured on clean held-out-fold inputs. The full reference (P1-1) is an independent rerun within this 80-run batch.

For the directionality comparison, P3-1 and P3-3 retain the reference depth, whereas P3-2 and P3-4 deepen the respective one-way paths for capacity compensation. The remaining architecture, training scheme, and fold assignments are controlled within each comparison. This is depth-based capacity control, not a claim of exact trainable-parameter equality. Section 4.3 discusses the resulting effect sizes and their exploratory paired tests.

Table A1. Configuration key for the five-fold RAVDESS ablation batch. All configurations use 30-frame visual input and identical actor-wise fold assignments; attention-based configurations use one head, and configurations retaining the learned acoustic branch use the frozen emotion2vec+ encoder. Full HAR combines learned and handcrafted acoustic branches; learned-only and handcrafted-only retain the respective branch. AV-only and VA-only follow the interaction-path notation in Section 3.

ID	Fusion	Acoustic Representation	Training	Interaction	Depth
P1-1	G^{2}AF	Full HAR	NIT	Two-way	1
P1-2	G^{2}AF	Learned only	NIT	Two-way	1
P1-3	Attention	Full HAR	NIT	Two-way	1
P1-4	Attention	Learned only	NIT	Two-way	1
P1-5	Concat	Full HAR	NIT	Two-way	1
P1-6	Concat	Learned only	NIT	Two-way	1
P2-1	G^{2}AF	Full HAR	ODT	Two-way	1
P2-3	G^{2}AF	Full HAR	HM	Two-way	1
P3-1	G^{2}AF	Full HAR	NIT	AV only	1
P3-2	G^{2}AF	Full HAR	NIT	AV only	2
P3-3	G^{2}AF	Full HAR	NIT	VA only	1
P3-4	G^{2}AF	Full HAR	NIT	VA only	4
P4-1	G^{2}AF	Handcrafted only	NIT	Two-way	1
P4-2	Attention	Handcrafted only	NIT	Two-way	1
P4-3	G^{2}AF	Full HAR	ODT	AV only	1
P4-4	G^{2}AF	Full HAR	ODT	VA only	1

Table A2. Complete fold-level records for the 80-run RAVDESS ablation batch. AV, A, and V are accuracies (%) for the fused, audio-only, and visual-only heads; BA denotes balanced accuracy (%). IDs are defined in Table A1.

ID	Fold	AV	A	V	BA$_{AV}$	BA$_A$	BA$_V$
P1-1	0	93.33	92.00	60.00	91.56	86.25	60.94
P1-1	1	90.33	85.67	57.00	90.94	83.75	59.69
P1-1	2	91.00	85.00	64.00	91.56	79.69	62.81
P1-1	3	95.67	91.00	56.00	95.94	91.56	53.44
P1-1	4	86.25	84.58	59.17	85.55	83.98	60.16
P1-2	0	96.33	85.00	65.00	96.56	84.38	66.25
P1-2	1	91.67	86.33	55.33	91.88	80.94	54.38
P1-2	2	89.33	85.00	65.33	87.81	79.69	64.06
P1-2	3	91.67	87.67	52.67	91.25	82.19	49.38
P1-2	4	83.33	81.67	63.33	81.25	76.56	62.89
P1-3	0	91.67	94.00	63.33	91.88	91.88	64.69
P1-3	1	92.00	85.33	59.00	92.50	83.44	61.56
P1-3	2	88.33	84.00	71.33	88.12	81.88	70.62
P1-3	3	92.00	88.33	58.67	92.50	84.06	60.00
P1-3	4	82.50	85.00	51.67	82.03	84.38	52.34
P1-4	0	90.00	92.33	56.00	88.75	86.56	58.12
P1-4	1	94.33	87.33	64.33	94.69	81.88	64.69
P1-4	2	90.33	85.00	66.00	90.62	79.69	63.12
P1-4	3	91.00	82.33	64.00	89.38	80.00	64.69
P1-4	4	79.17	81.67	44.17	78.12	76.56	44.92

Table A2. *Cont.*

ID	Fold	AV	A	V	BA_{AV}	BA_A	BA_V
P1-5	0	89.00	92.67	67.67	89.69	89.38	69.38
P1-5	1	96.00	88.00	64.67	96.25	86.88	65.00
P1-5	2	94.00	83.67	67.00	94.38	78.75	66.25
P1-5	3	90.00	92.67	54.00	90.62	90.94	55.62
P1-5	4	80.42	81.67	56.67	78.12	76.56	55.86
P1-6	0	90.67	92.00	67.67	90.31	86.25	68.44
P1-6	1	89.33	81.33	58.67	88.75	82.50	58.75
P1-6	2	85.33	81.00	60.67	85.00	81.56	58.75
P1-6	3	94.00	88.33	62.67	93.75	84.06	62.50
P1-6	4	87.50	75.83	62.08	84.77	75.78	62.50
P2-1	0	89.33	91.00	65.67	90.00	86.88	67.19
P2-1	1	93.33	87.33	67.33	92.19	84.06	66.56
P2-1	2	90.00	82.67	68.33	90.31	82.19	69.69
P2-1	3	90.00	84.33	57.33	90.62	85.31	58.75
P2-1	4	85.00	82.08	45.42	82.81	76.95	42.58
P2-3	0	91.67	87.33	70.33	90.31	83.12	71.25
P2-3	1	91.67	89.00	60.67	92.19	88.12	62.81
P2-3	2	85.67	84.67	71.33	86.25	82.81	71.25
P2-3	3	90.67	82.67	58.33	91.25	83.75	59.06
P2-3	4	82.08	82.08	56.25	81.25	77.34	58.59
P3-1	0	87.67	93.00	65.33	85.31	89.69	66.25
P3-1	1	92.33	84.67	61.67	92.50	80.31	62.81
P3-1	2	93.00	85.33	68.67	92.50	81.56	69.06
P3-1	3	92.33	87.00	61.33	92.81	81.56	62.50
P3-1	4	90.00	80.42	60.00	89.45	78.91	57.81
P3-2	0	96.00	90.67	52.33	94.06	85.00	49.06
P3-2	1	92.67	87.00	65.67	92.81	83.12	64.38
P3-2	2	89.00	84.33	61.00	86.56	79.06	58.12
P3-2	3	92.33	89.67	58.00	92.81	86.56	57.50
P3-2	4	85.42	85.00	47.50	85.55	84.38	47.66
P3-3	0	92.67	91.33	57.00	91.56	85.94	56.56
P3-3	1	90.00	87.67	63.33	89.38	82.19	62.50
P3-3	2	86.67	85.00	74.00	87.50	83.12	73.12
P3-3	3	94.67	90.33	60.67	93.75	86.88	60.62
P3-3	4	81.67	82.50	55.00	79.30	77.34	53.12
P3-4	0	93.00	91.67	69.00	93.12	88.75	68.75
P3-4	1	92.67	86.67	66.33	93.12	81.25	67.19
P3-4	2	89.00	86.00	57.33	88.12	83.75	55.94
P3-4	3	94.33	86.67	46.00	92.81	81.25	43.75
P3-4	4	86.25	82.08	45.00	83.98	76.95	46.88
P4-1	0	69.33	36.00	63.67	70.00	33.75	64.69
P4-1	1	81.00	45.67	75.00	81.25	43.44	75.62
P4-1	2	80.33	46.67	76.00	80.31	44.38	75.31
P4-1	3	67.33	42.33	62.67	67.81	40.94	62.81
P4-1	4	75.00	42.50	65.42	76.56	39.84	67.58
P4-2	0	74.67	40.00	71.00	75.00	37.81	71.56
P4-2	1	81.00	47.00	73.67	80.62	45.31	73.44
P4-2	2	82.67	50.00	70.33	81.25	47.81	69.69
P4-2	3	73.67	43.33	63.33	74.38	42.19	64.38
P4-2	4	75.42	44.17	68.33	76.95	41.41	70.31

Table A2. *Cont.*

ID	Fold	AV	A	V	BA_{AV}	BA_A	BA_V
P4-3	0	89.00	90.00	65.33	89.69	86.25	67.19
P4-3	1	89.00	85.33	58.00	88.44	85.62	58.44
P4-3	2	87.33	85.67	66.67	86.88	85.31	67.19
P4-3	3	96.67	87.67	58.33	96.88	82.50	59.69
P4-3	4	89.17	83.75	59.58	88.67	82.81	59.77
P4-4	0	91.33	88.33	64.67	91.88	84.38	66.25
P4-4	1	92.67	87.00	65.33	92.19	81.56	61.56
P4-4	2	89.00	85.67	71.67	89.69	85.00	71.56
P4-4	3	96.00	89.00	69.33	96.25	86.25	70.62
P4-4	4	83.33	81.67	45.83	82.03	76.56	48.05

Configuration. All rows belong to the same new 80-run batch and use 30-frame visual input and identical actor-wise fold assignments. The full reference is therefore reported independently of the earlier five-fold result in Table 5. AV-only and VA-only follow the interaction-path notation of Section 3. The deeper one-way variants provide depth-based capacity compensation; exact parameter equality is not assumed. **Datasets:** RAVDESS, 8 classes. **Training:** SGD (momentum 0.9, weight decay 10^{-3}), initial lr 0.04, decayed by a factor of 10 at epochs 40, 55, 65 and 70; batch size 4, 100 epochs; Gumbel temperature annealed from $\tau_0 = 1.0$ to 0.1 during training, $\tau = 0.1$ at evaluation; every cell is an independent training run under the indicated scheme (fusion/acoustic composition/interaction direction/perturbation scheme), and the best-validation checkpoint is evaluated on the held-out test split. **Visual input:** 30 frames sampled per clip, MTCNN face crop, 224 × 224, EfficientFace initialized from AffectNet7. **Audio input:** 3.6 s waveforms; MFCC ($n = 10$) + CQT ($n_{\text{bins}} = 42$) + emotion2vec+ embedding (768) = 820 channels × 156 frames; the emotion2vec+ encoder is frozen. **Protocol and metrics:** Actor-independent 5-fold cross-validation: for fold k, testing = F_k, validation = $F_{(k+1) \bmod K}$, and the remaining actors are used for training; only the clean (both-modality) test condition is evaluated. All values are test accuracies (%); AV = fused head, A = audio-only head, V = visual-only head; per-fold accuracy and balanced accuracy are reported and summarized as mean ± std across the five folds; the random seed is fixed (seed = 1) and shared by all cells.

Appendix B. Quantified Robustness under Modality Distortion Settings

Tables A3–A6 quantify the robustness of the models and report the sensitivity of the Gumbel schedule.

Table A3 shows that on RAVDESS damaging audio costs the fused model 30% to 50% while damaging video costs only 2–7 points, which follows directly from the audio dominance shown in Section 4.3. Tables A4 and A5 report the waveform-level protocol, in which noise is added before feature extraction and the visual stream is untouched. On RAVDESS the audio-only head collapses from 85.98% to 15.68% between clean and −5 dB, whereas the fused head degrades from 91.37% to 52.93%, i.e., about 45% less than the audio-only head. On PKU-ER the same protocol shows the roles reversed: the audio-only head drops from 69.14% to 33.29%, the visual-only head is unaffected at 92.01%, and the fused head falls only from 93.56% to 80.45%. Table A6 shows that varying the initial Gumbel temperature τ_0 and the decay rate γ moves the AV accuracy within a 0.83-point band, so the default $\tau_0 = 1.0$ is a representative operating point rather than a tuned optimum.

Table A3. Robustness to test-time modality corruption on RAVDESS. Corruption is applied at test time only; drops are relative to the complete input measured in the same evaluation run.

Test-Time Condition	AV Accuracy	Drop vs. Complete
Complete input (C0)	91.03	–
Audio full-noise replacement (AN)	60.72	−30.31
Audio additive noise (AA)	41.42	−49.61
Audio zero-mask (AZ)	49.92	−41.11
Visual full-noise replacement (VN)	87.17	−3.86
Visual additive noise (VA)	88.65	−2.38
Visual zero-mask (VZ)	84.03	−7.00

Table A4. Waveform-level acoustic noise on RAVDESS. Additive white Gaussian noise is injected into the raw waveform before feature extraction; every test sample is corrupted, and the visual input is untouched.

SNR (dB)	Clean	20	15	10	5	0	−5
AV	91.37	78.98	73.03	67.05	51.28	50.13	52.93
A	85.98	74.40	67.12	55.48	23.85	15.73	15.68
V	61.80	61.80	61.80	61.80	61.80	61.80	61.80

Table A5. Waveform-level acoustic noise on PKU-ER under stratified 5-fold evaluation. Additive white Gaussian noise is injected into the raw waveform before feature extraction; every test sample is corrupted, and the visual input is untouched.

SNR (dB)	Clean	20	15	10	5	0	−5
AV	93.56	74.12	75.46	81.02	87.59	87.79	80.45
A	69.14	34.85	38.28	40.61	38.29	30.97	33.29
V	92.01	92.01	92.01	92.01	92.01	92.01	92.01

Table A6. Sensitivity to the Gumbel gate schedule on RAVDESS (actor-wise fold 0). Each column is a complete training run differing only in the initial temperature τ_0 and the per-step decrement γ; the evaluation temperature is fixed at $\tau = 0.1$ for all three columns.

τ_0 (γ)	0.5 (0.0005)	1.0 (0.002)	1.5 (0.0005)
AV accuracy	95.42	95.00	95.83

Configuration for Tables A3–A6. Datasets: RAVDESS, 8 classes; PKU-ER, 3 classes. Architecture: the main configuration of each dataset, i.e., Variant 1 on RAVDESS and Variant 2 on PKU-ER, both with Gumbel-sampling fusion and one attention head; the three prediction heads AV, A and V are supervised jointly, A and V being the auxiliary single-modality heads branching off before the fusion module. Training scheme: NIT; the best-validation checkpoint of each fold is evaluated on that fold's test split; no class-balancing strategy was used. Visual input: 30 frames per clip, MTCNN face crop, 224 × 224, EfficientFace initialized from AffectNet7. Audio input: 3.6 s waveforms; MFCC (n = 10) + CQT (n_{bins} = 42) + emotion2vec+ embedding (768) = 820 channels × 156 frames; the emotion2vec+ encoder is frozen. Metrics. All values are test accuracies (%); no training sample is used for evaluation. Folding. Tables A3–A5 report means over the five folds (RAVDESS: actor-wise; PKU-ER: stratified by clip). Table A6 is a single-fold probe: each of its cells is an independent 100-epoch training run, so the schedule sweep was restricted to RAVDESS actor-wise fold 0; its three cells are to be compared with one another, and not with the five-fold means in Tables A3–A5. Conditions. In Table A3 the corruption operator, its domain and its severity follow the definitions of Section 4.2, and every test sample of the evaluated fold is corrupted; in Tables A4 and A5 the noise level is given as the signal-to-noise ratio of the corrupted waveform and only the audio channel is corrupted. Gumbel schedule in Table A6. The temperature follows $\tau(t) = \max(\tau_0 - \gamma t, \tau_{\min})$ with $\tau_{\min} = 0.1$, and the gate clips the effective value to $[0.1, 1.0]$; consequently $\tau_0 = 1.5$ is not a hotter gate than $\tau_0 = 1.0$ but a schedule that dwells at the clipping ceiling for longer. τ_0 and γ vary jointly; the evaluation temperature τ is clipped to 0.1 inside the gate.

Appendix C. Per-Class Recall Analysis

An analysis of the confusion matrices in Figure A1 reveals per-class behavior. These recalls are computed after pooling the clean per-sample predictions from the waveform-noise evaluation across all five folds. On RAVDESS, the pooled recalls range from 85.4% to 96.4% (neutral 88.5%, calm 85.4%, happy 93.8%, sad 87.5%, angry 96.4%, fearful 92.2%, disgust 95.3%, surprised 92.7%). On PKU-ER, the pooled recalls are 95.5% for positive, 91.1% for negative, and 94.1% for neutral, giving a pooled macro recall of 93.57%. This pooled value differs in aggregation from Table 5, where BA of 93.43% and overall accuracy of 93.45% are averages of fold-level values.

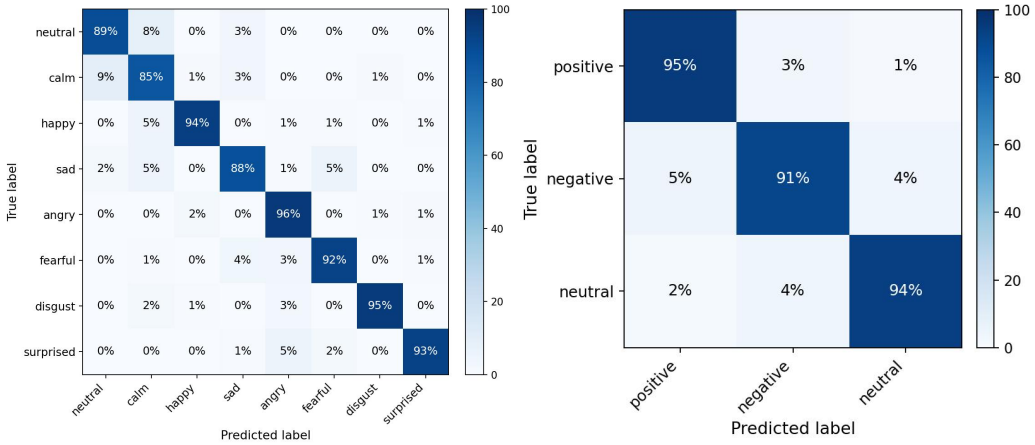


Figure A1. Matrices are recomputed over the merged 5-fold clean test predictions with the emotion2vec+ pipeline and the Variant 1 + NIT scheme (**left**). Matrices are recomputed over the merged 5-fold clean test predictions with the emotion2vec+ pipeline and the Variant 2 + NIT scheme (**right**). Confusion matrices generated from the RAVDESS and PKU-ER datasets are shown on the left and right, respectively.

References

1. Xiang, G.; Yao, S.; Wu, X.; Deng, H.; Wang, G.; Liu, Y.; Li, F.; Peng, Y. Driver multi-task emotion recognition network based on multi-modal facial video analysis. *Pattern Recognit.* **2025**, *161*, 111241. https://doi.org/10.1016/j.patcog.2024.111241.
2. Jia, T.; Pan, F.; Yang, X.; Ji, L.; Farina, D.; Li, C. Artificial empathy in therapy and healthcare: Advancements in interpersonal interaction technologies. *Cyborg Bionic Syst.* **2025**, *6*, 0473.
3. Zhong, Y.; Tang, W.; Gui, H.; Sheng, Q.; Xu, H.; Qin, K.; Guo, X.; Yang, H.; Zou, J. Human camouflage and expression via soft mask from reprogrammable chemical fluid skin. *Sci. Adv.* **2025**, *11*, eadq6141.
4. Hao, M.; Cao, W.H.; Liu, Z.T.; Wu, M.; Xiao, P. Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features. *Neurocomputing* **2020**, *391*, 42–51. https://doi.org/10.1016/j.neucom.2020.01.048.
5. Goncalves, L.; Leem, S.G.; Lin, W.C.; Sisman, B.; Busso, C. Versatile audio-visual learning for emotion recognition. *IEEE Trans. Affect. Comput.* **2025**, *16*, 306–318. https://doi.org/10.1109/TAFFC.2024.3433386.
6. Chumachenko, K.; Iosifidis, A.; Gabbouj, M. Self-attention fusion for audioVisual emotion recognition with incomplete data. In Proceedings of the International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; pp. 2822–2828. https://doi.org/10.1109/ICPR56361.2022.9956592.
7. Huang, J.; Tao, J.; Liu, B.; Lian, Z.; Niu, M. Multimodal transformer fusion for continuous emotion recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 4–8 May 2020; pp. 3507–3511. https://doi.org/10.1109/ICASSP40776.2020.9053762.
8. Sahidullah, M.; Saha, G. Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. *Speech Commun.* **2012**, *54*, 543–565. https://doi.org/10.1016/j.specom.2011.11.004.
9. Brown, J.C. Calculation of a constant Q spectral transform. *J. Acoust. Soc. Am.* **1991**, *89*, 425–434. https://doi.org/10.1121/1.400476.
10. Chen, S.; Wu, Y.; Wang, C.; Liu, S.; Tompkins, D.; Chen, Z.; Che, W.; Yu, X.; Wei, F. BEATs: Audio pre-training with acoustic tokenizers. In Proceedings of the International Conference on Machine Learning (ICML), Honolulu, HI, USA, 23–29 July 2023; pp. 5178–5193.
11. Poria, S.; Cambria, E.; Bajpai, R.; Hussain, A. A review of affective computing: From unimodal analysis to multimodal fusion. *Inf. Fusion* **2017**, *37*, 98–125.
12. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: The munich versatile and fast open-source audio feature extractor. In Proceedings of the ACM International Conference on Multimedia (ACM MM), Florence, Italy, 25–29 October 2010; pp. 1459–1462. https://doi.org/10.1145/1873951.1874246.
13. Chen, J.; Liu, X.; Tu, P.; Aragones, A. Learning person-specific models for facial expression and action unit recognition. *Pattern Recognit. Lett. PRL* **2013**, *34*, 1964–1970. https://doi.org/10.1016/j.patrec.2013.02.002.
14. Baltrušaitis, T.; Mahmoud, M.; Robinson, P. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 4–8 May 2015; pp. 1–6. https://doi.org/10.1109/FG.2015.7284869.

15. Zhang, T.; Zheng, W.; Cui, Z.; Zong, Y.; Yan, J.; Yan, K. A deep neural network-driven feature learning method for multi-view facial expression recognition. *IEEE Trans. Multimed. TMM* **2016**, *18*, 2528–2536. https://doi.org/10.1109/TMM.2016.2598092.
16. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. https://doi.org/10.1109/5.726791.
17. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.
18. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078. https://doi.org/10.3115/v1/D14-1179.
19. Ahmed, M.R.; Islam, S.; Islam, A.M.; Shatabda, S. An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition. *Expert Syst. Appl. ESWA* **2023**, *218*, 119633–119654. https://doi.org/10.1016/j.eswa.2023.119633.
20. Wei, Y.; Long, K.; Granston, A.; Rodriguez-Contreras, A. USV-DETR: High-Resolution and Densely Supervised Detection of Ultrasonic Vocalizations. In Proceedings of the Interspeech 2026, Sydney, Australia, 27 September–1 October 2026; pp. 2604–2608. https://doi.org/10.21437/Interspeech.2026-1492.
21. Wei, Y.; Long, K.; Granston, A.; Rodriguez-Contreras, A. USVexplorer: Robust Detection of Ultrasonic Vocalizations with Cross Species Generalization. In Proceedings of the ICASSP 2026—2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2026; pp. 15232–15236. https://doi.org/10.1109/ICASSP55912.2026.11463476.
22. Singh, S.K.; Kumar, M.; Khan, I.M.; Jayanthiladevi, A.; Agarwal, C. An attention-based model for recognition of facial expressions using CNN-BiLSTM. *Polytech. J.* **2025**, *15*, 4–18. https://doi.org/10.59341/2707-7799.1849.
23. He, X.; An, W.; Wang, Y.; Guo, G.; Luo, Y.; Ling, Q.; Zhou, S. Dim moving target detection with hyperspectral image sequences. *ISPRS J. Photogramm. Remote Sens.* **2026**, *239*, 524–554. https://doi.org/10.1016/j.isprsjprs.2026.06.015.
24. Wei, Y. Enhanced fruit and vegetable instance segmentation via attention integration and multi-scale feature encoding. In *Proceedings of the Applications of Machine Learning 2026*; International Society for Optics and Photonics SPIE: Bellingham, WA, USA, 2026; Volume 14218, p. 142180Z. https://doi.org/10.1117/12.3113881.
25. Li, J.; Li, J.; Wang, X.; Zhan, X.; Zeng, Z. A domain generalization and residual network-based emotion recognition from physiological signals. *Cyborg Bionic Syst.* **2024**, *5*, 0074.
26. Ye, W.; Wang, J.; Chen, L.; Dai, L.; Sun, Z.; Liang, Z. Adaptive spatial–temporal aware graph learning for EEG-based emotion recognition. *Cyborg Bionic Syst.* **2024**, *5*, 0088.
27. Feng, C.; Cao, L.; Wu, D.; Zhang, E.; Wang, T.; Jiang, X.; Chen, J.; Wu, H.; Lin, S.; Hou, Q.; et al. Acoustic inspired brain-to-sentence decoder for logosyllabic language. *Cyborg Bionic Syst.* **2025**, *6*, 0257.
28. Li, Q.; Luo, H.; Cheng, H.; Deng, Y.; Sun, W.; Li, W.; Liu, Z. Incipient Fault Detection in Power Distribution System: A Time–Frequency Embedded Deep-Learning-Based Approach. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 2507914.
29. Luo, H.; Li, Q.; Zhou, H.; Sun, W.; Li, W.; Liu, Z.; Ji, Y.; Ding, L. Weak Single Phase-to-Ground Fault Time Detection with Learnable-Parameter-Driven DSP-Enhanced Transformer. *IEEE Trans. Smart Grid* **2026**, *17*, 1693–1708.
30. Luo, H.; Feng, Q.; Sun, W.; Li, W.; Li, Q.; Zhao, W.; Liu, Z. FD-Mamba with Neural Observer and Frequency-Enhanced Update for Incipient Feeder Fault Detection. *IEEE Internet Things J.* **2026**, *13*, 20761–20772.
31. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. TPAMI* **2018**, *41*, 423–443.
32. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
33. Xiang, P.; Lin, C.; Wu, K.; Bai, O. MultiMAE-DER: Multimodal masked autoencoder for dynamic emotion recognition. In Proceedings of the International Conference on Pattern Recognition Systems (ICPRS), London, UK, 15–18 July 2024; pp. 1–7. https://doi.org/10.1109/ICPRS62101.2024.10677820.
34. Zhou, X.; Wu, J.; Xu, H.; Mei, J.; Chen, Z.; Zhang, J.; Chen, Y.; Jin, Y.; Liu, C.; Yuan, C. Accurate and robust multimodal emotion recognition for human–robot interaction via dynamic graph learning with pairwise cross-modal alignment. *Big Data Cogn. Comput.* **2026**, *10*, 276.
35. Fan, S.; Jing, J.; Wang, C. Audio-Visual Learning for Multimodal Emotion Recognition. *Symmetry* **2025**, *17*, 418. https://doi.org/10.3390/sym17030418.
36. Qu, W.; Wang, M.J.S. Structured and Factorized Multi-Modal Representation Learning for Physiological Affective State and Music Preference Inference. *Symmetry* **2026**, *18*, 488. https://doi.org/10.3390/sym18030488.
37. Huijben, I.A.; Kool, W.; Paulus, M.B.; Van Sloun, R.J. A review of the gumbel-max trick and its extensions for discrete stochasticity in machine learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 1353–1371. https://doi.org/10.1109/TPAMI.2022.3157042.
38. Jang, E.; Gu, S.; Poole, B. Categorical Reparameterization with Gumbel-Softmax. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017; pp. 1–12.

39. Balog, M.; Lakshminarayanan, B.; Ghahramani, Z.; Roy, D.M.; Tenenbaum, J.B. Tractable Learning for Complex Probability Queries. *Adv. Neural Inf. Process. Syst. NIPS* **2017**, *30*, 1–9.
40. Tsai, Y.H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, 28 July–2 August 2019; pp. 6558–6569. https://doi.org/10.18653/v1/P19-1656.
41. Li, Y.; Liu, H.; Tang, H. Multi-modal perception attention network with self-supervised learning for audio-visual speaker tracking. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Vancouver, BC, Canada, 22 February–1 March 2022; pp. 1456–1463. https://doi.org/10.1609/aaai.v36i2.20035.
42. Zhao, Z.; Liu, Q.; Zhou, F. Robust lightweight facial expression recognition network with label distribution training. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Virtual, 2–9 February 2021; pp. 3510–3519. https://doi.org/10.1609/aaai.v35i4.16465.
43. Nair, V.; Hinton, G.E. Rectified linear units improve restricted Boltzmann machines. In Proceedings of the ICML, Haifa, Israel, 21–24 June 2010; pp. 807–814.
44. Ma, Z.; Zheng, Z.; Ye, J.; Li, J.; Gao, Z.; Zhang, S.; Chen, X. emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation. In *Proceedings of the ACL 2024 Findings*; Association for Computational Linguistics: Kerrville, TX, USA, 2024. https://doi.org/10.18653/v1/2024.findings-acl.931.
45. Kendall, A.; Gal, Y.; Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7482–7491. https://doi.org/10.1109/CVPR.2018.00781.
46. Livingstone, S.R.; Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLoS ONE* **2018**, *13*, e0196391. https://doi.org/10.1371/journal.pone.0196391.
47. Kamada, S.; Ichimura, T. Multimodal adaptive structural deep belief network for emotion recognition on RAVDESS dataset. In Proceedings of the International Congress on Advanced Applied Informatics (IIAI-AAI), Takamatsu, Japan, 6–12 July 2024; pp. 292–298. https://doi.org/10.1109/IIAI-AAI63651.2024.00062.
48. Middya, A.I.; Nag, B.; Roy, S. Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. *Knowl. Based Syst. KBS* **2022**, *244*, 108580–108594. https://doi.org/10.1016/j.knosys.2022.108580.
49. Guo, P.; Chen, Z.; Li, Y.; Liu, H. Audio-visual fusion network based on conformer for multimodal emotion recognition. In Proceedings of the CAAI International Conference on Artificial Intelligence (CICAI), Beijing, China, 27–28 August 2022; pp. 315–326. https://doi.org/10.1007/978-3-031-20500-2_26.
50. Fan, R.; Liu, H.; Li, Y.; Guo, P.; Wang, G.; Wang, T. AttA-NET: Attention aggregation network for audio-visual emotion recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 8030–8034. https://doi.org/10.1109/ICASSP48485.2024.10447640.
51. Joze, H.R.V.; Shaban, A.; Iuzzolino, M.L.; Koishida, K. MMTM: Multimodal transfer module for CNN fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 14–19 June 2020; pp. 13289–13299. https://doi.org/10.1109/CVPR42600.2020.01330.
52. Su, L.; Hu, C.; Li, G.; Cao, D. MSAF: Multimodal split attention fusion. *arXiv* **2020**, arXiv:2012.07175.
53. Verbitskiy, S.; Berikov, V.; Vyshegorodtsev, V. ERANNs: Efficient residual audio neural networks for audio pattern recognition. *Pattern Recognit. Lett. PRL* **2022**, *161*, 38–44. https://doi.org/10.1016/j.patrec.2022.07.012.
54. John, V.; Kawanishi, Y. Audio and video-based emotion recognition using multimodal transformers. In Proceedings of the International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; pp. 2582–2588. https://doi.org/10.1109/ICPR56361.2022.9956730.
55. Franceschini, R.; Fini, E.; Beyan, C.; Conti, A.; Arrigoni, F.; Ricci, E. Multimodal emotion recognition with modality-pairwise unsupervised contrastive loss. In Proceedings of the International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; pp. 2589–2596. https://doi.org/10.1109/ICPR56361.2022.9956589.
56. Dabbabi, K.; Mars, A. Self-supervised learning for speech emotion recognition task using audio-visual features and distil Hubert model on BAVED and RAVDESS databases. *J. Syst. Sci. Syst. Eng.* **2024**, *33*, 576–606. https://doi.org/10.1007/s11518-024-5607-y.
57. Gulati, A.; Qin, J.; Chiu, C.C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv* **2020**, arXiv:2005.08100. https://doi.org/10.21437/Interspeech.2020-3015.
58. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res. JMLR* **2008**, *9*, 2579–2605.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.