# When Tone and Words Disagree: Towards Robust Speech Emotion Recognition under Acoustic-Semantic Conflict

**Anonymous ACL submission**

## Abstract

Speech Emotion Recognition (SER) systems often assume congruence between vocal emotion and lexical semantics. However, in real-world interactions, acoustic-semantic conflict is common yet overlooked, where the emotion conveyed by tone contradicts the literal meaning of spoken words. We show that state-of-the-art SER models, including ASR-based, self-supervised learning (SSL) approaches and Audio Language Models (ALMs), suffer performance degradation under such conflicts due to semantic bias or entangled acoustic–semantic representations. To address this, we propose the **Fusion Acoustic-Semantic (FAS)** framework, which explicitly disentangles acoustic and semantic pathways and bridges them through a lightweight, query-based attention module. To enable systematic evaluation, we introduce the **Conflict in Acoustic-Semantic Emotion (CASE)**, the first dataset dominated by clear and interpretable acoustic-semantic conflicts in varied scenarios. Extensive experiments demonstrate that FAS consistently outperforms existing methods in both in-domain and zero-shot settings. Notably, on the CASE benchmark, conventional SER models fail dramatically, while FAS sets a new SOTA with 59.38% accuracy. Our code and datasets is available at https://anonymous.4open.science/r/FAS-Anonymous.

## 1 Introduction

Speech Emotion Recognition (SER), a fundamental task in affective computing, aims to automatically identify the emotional state of a speaker from their vocal expressions. Current SER methods (Ma et al., 2024; Chen et al., 2024; Elizalde et al., 2023; Hsu et al., 2021; Radford et al., 2023; Chen et al., 2022) have demonstrated remarkable performance on standard academic benchmarks such as IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2019). However, this success is largely confined to scenarios of acoustic-semantic congruence, where the prosodic cues in speech (acoustics) align with the literal meaning of the spoken content (semantics)—for instance, expressing "What a beautiful day!" in a joyful tone.

The crux of the problem is that emotion is inherently complex and frequently manifests under conditions of acoustic-semantic conflict. Real-world communication is replete with nuanced expressions like sarcasm, schadenfreude (gloating) or cold fury, where a speaker's true emotion, conveyed through acoustic cues, is decoupled or even antithetical to the semantic content of their utterance. For example, upon learning that a colleague has been promoted, someone might say *"Congratulations on your promotion!"* in a flat or resentful tone, subtly revealing underlying envy rather than joy. In these prevalent yet challenging scenarios, the performance of current SER methods collapses.

This vulnerability is systemic across current SER paradigms. speech-text pre-trained encoders (Radford et al., 2023; Elizalde et al., 2023) exhibit a strong semantic bias, causing them to be "poisoned" by the literal meaning of words. Self-supervised learning (SSL) methods (Hsu et al., 2021; Baevski et al., 2020; Schneider et al., 2019; Chen et al., 2022) produce entangled representations where affective and semantic information are conflated, making ambiguity difficult to resolve. Furthermore, explicit multimodal approaches (Zhang et al., 2025; Cheng et al., 2024) struggle, as their current modality fusion mechanisms often lack a robust strategy to arbitrate between conflicting signals, defaulting to the misleading modality. Recent ALMs, despite their impressive capabilities, depend on LLM-aligned encoders (e.g. Whisper (Radford et al., 2023), CLAP (Elizalde et al., 2023)) that prioritize semantics over prosody, depriving the LLM of affective cues during emotional conflict. All these brittleness limit the reliable application of SER models in uncon-

strained real-world environments.

What's more, current SER evaluation is limited by datasets biases: most benchmarks feature predominantly congruent emotional expressions. While in-the-wild datasets (e.g., MELD, IEMOCAP) contain occasional acoustic-semantic conflicts, such cases are sparse and unstructured. Critically, no existing resource provides a high-density, controlled setting to systematically evaluate robustness under emotional conflict—leaving a key aspect of real-world performance unassessed.

To address these challenges, this paper introduces a dual-pronged contribution. First, we propose an innovative **Fusion Acoustic-Semantic (FAS)** framework, designed to explicitly disentangle acoustic and semantic information from speech. The FAS uniquely employs an audio tokenizer, inspired by recent advances in Text-to-Speech generation, to extract low-dimensional acoustic tokens, while concurrently utilizing a pre-trained module to capture high-dimensional semantic information. A lightweight, query-based module is then introduced to integrate disentangled features and make robust predictions.

Second, to validate our proposed framework and to provide a much-needed resource for the community, we released the **Conflict in Acoustic-Semantic Emotion (CASE)** Benchmark. Unlike conventional datasets, CASE is constructed with a high concentration of logical, interpretable, and scenario-driven conflict samples. It serves not only as a challenging testbed for evaluating model robustness but also as a valuable corpus for researchers to study the interplay between acoustics and semantics in human emotion expression. The main contributions of this paper are summarized as follows:

- We are the first to systematically investigate the problem of affective-semantic conflict in SER, showing that existing methods degrade significantly on such samples.
- We introduce the **CASE** benchmark, the first standardized dataset designed to assess the robustness of SER models against acoustic-semantic conflict.
- We propose the **FAS** framework, designed to resolve emotional ambiguity by disentangling acoustic cues and semantic content. Through extensive experiments, we demonstrate the superiority of our FAS over SOTA baselines on all benchmarks.

## 2 Related Work

### 2.1 Speech Emotion Recognition

Recent advances in Speech Emotion Recognition (SER) have been predominantly driven by large-scale pre-trained models like WavLM (Chen et al., 2022), Whisper (Radford et al., 2023) and HuBERT (Hsu et al., 2021). A significant body of work (Ma et al., 2024; Chen et al., 2024; Qi et al., 2025) focuses on transferring, distilling and adapting the representations from these powerful encoders to SER task.

However, ASR-based encoders such as Whisper (Radford et al., 2023) and CLAP (Elizalde et al., 2023), while powerful, exhibit a strong semantic bias due to their pre-training objective, making them highly susceptible to being misguided by the literal meaning of spoken words. Furthermore, a fundamental limitation of SSL-based encoders (Hsu et al., 2021; Chen et al., 2022; Schneider et al., 2019; Baevski et al., 2020) is their production of entangled representations. Within these learned features, affective prosody is inseparably mixed with phonetic content, rather than being explicitly disentangled. This conflation severely limits their robustness, particularly when faced with the challenge of affective-semantic conflict. Our work directly addresses this representation entanglement and semantic bias by proposing a novel fusion framework.

### 2.2 Neural Audio Tokenization

The field of Text-To-Speech generation and Audio Editing has spurred the development of high-fidelity neural audio tokenizers. Discrete audio tokenizers, such as EnCodec (Défossez et al., 2022), XCodec (Ye et al., 2025a,b) and VibeVoice (Peng et al., 2025), built upon the VQ-VAE framework (Van Den Oord et al., 2017), excel at discretizing waveforms for high-quality signal reconstruction. More recently, VAE-based continuous tokenizers have also been proposed to better unify semantic and acoustic information for joint understanding and generation tasks (Yan et al., 2025; Jia et al., 2025), such as MingTok-Audio (Yan et al., 2025). A common characteristic of these generation-focused approaches is their ability to distill speech into a low-dimensional, acoustically-clean representation. Unlike the high-dimensional,

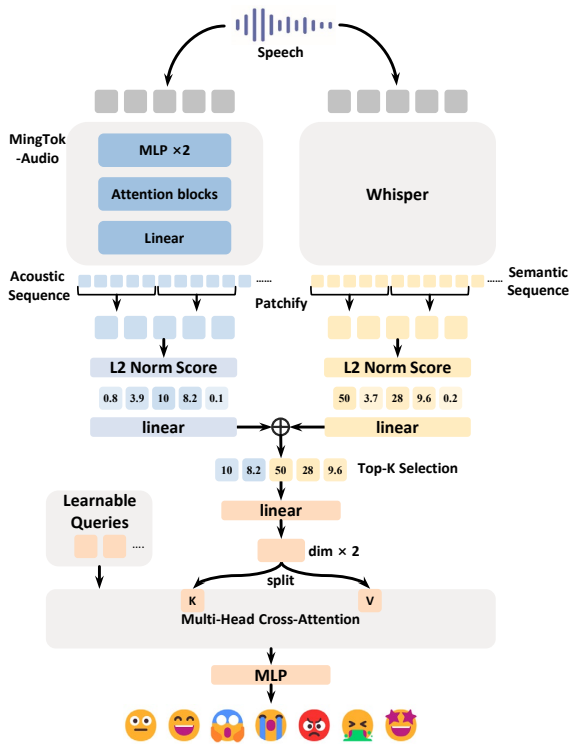



Figure 1: Overview of the proposed Fusion Acoustic-Semantic (FAS) framework. It disentangles speech into acoustic (MingTok-Audio) and semantic (Whisper) representations. The Fusion Module, guided by learnable queries and cross-attention to dynamically distill a unified feature for robust emotion recognition.

semantically-entangled features from recognition-oriented encoders, this compact representation explicitly captures fine-grained prosody and speaker identity, which are crucial for high-quality synthesis.

While their primary application has been in audio generative tasks, their potential as a source of disentangled acoustic representation for discriminative tasks like SER has remained unexplored. We pioneer the repurposing of audio tokens as a dedicated pathway for modeling acoustic affective features, aiming to resolve emotional ambiguity in SER where traditional methods fail.

## 3 Methods

### 3.1 Overview of Fusion Acoustic-Semantic Framework

The limitations of current SER models stem from their reliance on a single or entangled representation. Our core insight lies in effectively fusing two heterogeneous and temporally-varying feature streams: the semantic features $F_{\text{sem}} \in \mathbb{R}^{T_{\text{sem}} \times D_{\text{sem}}}$ and the acoustic features $F_{\text{aco}} \in \mathbb{R}^{T_{\text{aco}} \times D_{\text{aco}}}$, where $T$ and $D$ represent the sequence length and feature dimension, respectively.

As depicted in Figure 1, we propose the **Fusion Acoustic-Semantic (FAS)**, a two-stage fusion framework that first intelligently distills salient tokens from feature streams and then bridges them using a cross-attention mechanism. The process is as follows:

1. **Patchification:** To efficiently handle long sequence, we first apply a patchification step. Each sequence $F \in \mathbb{R}^{T \times D}$ is downsampled by a factor of $s = 5$. This creates a shorter sequence of patches, $F' \in \mathbb{R}^{(T/s) \times D}$. Subsequently, these patch sequences are projected into the unified hidden dimension, $d$:

$$f_{\text{aco}} = F'_{\text{aco}} W_{\text{aco}}; \quad f_{\text{sem}} = F'_{\text{sem}} W_{\text{sem}} \quad (1)$$

where $f_{\text{aco}} \in \mathbb{R}^{T'_{\text{aco}} \times d}$ and $f_{\text{sem}} \in \mathbb{R}^{T'_{\text{sem}} \times d}$, with $T' = T/s$.

2. **Token Distillation:** Recognizing that emotional cues are sparsely distributed, we introduce a non-uniform token selection strategy to identify and retain only the most informative "highlight" tokens from each sequence. This is achieved by:

   (a) **Saliency Scoring:** We compute a energy score $s_t$ for each token $f_t$ in a sequence. Inspired by findings that emotionally charged events often correlate with higher activation, we use the L2 Norm as a proxy for its score:

$$s_t = \|f_t\|_2 \quad (2)$$

   This fast, non-parametric method effectively captures moments of high energy in the acoustic stream and text stream.

   (b) **Top-K Selection:** We then select the $k$ tokens with the highest saliency scores, where $k_{\text{aco}}$ and $k_{\text{sem}}$ are sequence lengths chosen to reflect the different information densities of each pathway. This results in two condensed sequences:

$$\begin{aligned} f'_{\text{aco}} &\in \mathbb{R}^{k_{\text{aco}} \times d} \\ f'_{\text{sem}} &\in \mathbb{R}^{k_{\text{sem}} \times d} \end{aligned} \quad (3)$$

This distillation process drastically reduces sequence length while preserving the most emotionally relevant temporal information, which is a improvement over uniform compression techniques.

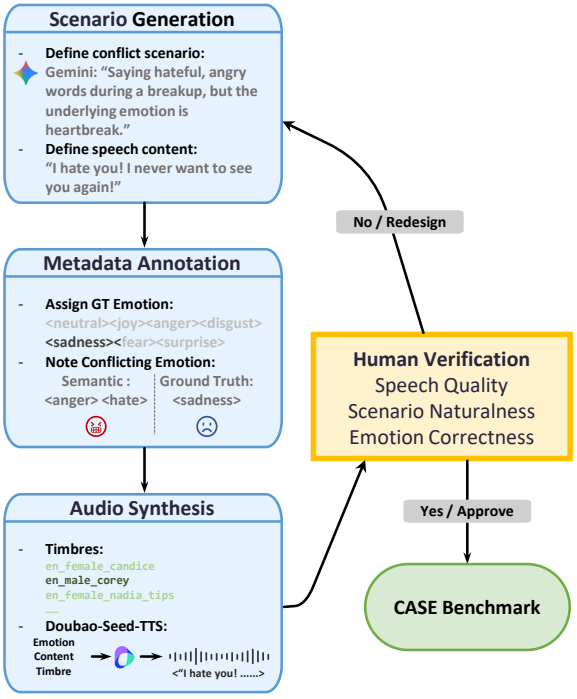



Figure 2: Pipeline of construction of CASE benchmark.

3. **Learnable Queries:** The distilled sequences are concatenated to a context sequence $C \in \mathbb{R}^{(k_{\text{aco}}+k_{\text{sem}})\times d}$. We then employ a fusion module inspired by Q-Former-like (Li et al., 2023) architectures. A set of $n$ learnable queries, $Q_{\text{learn}} \in \mathbb{R}^{n\times d}$, actively interrogate this context to produce a learned-representation. The context $C$ is projected to generate Key ($K$) and Value ($V$) matrices, and a cross-attention mechanism computes the final fused tokens:

$$F_{\text{fused}} = \text{Attn}(Q_{\text{learn}}, CW_K, CW_V) \quad (4)$$

Finally, a simple Multi-Layer Perceptron (MLP) acts as the prediction head on top of the fused vector to yield the final emotion probabilities, $P(y|X)$, across the 7 emotion categories.

### 3.2 Conflict in Acoustic-Semantic Emotion (CASE) Benchmark

To rigorously evaluate methods robustness against affective-semantic conflict, we constructed the **Conflict in Acoustic-Semantic Emotion (CASE)** Benchmark, a specialized testbed designed to systematically probe SER model limitations in complex emotional scenarios. The construction process in Figure 2 was guided by the principles of logical coherence and high conflict density, ensuring that every sample is grounded in a plausible real-world scenarios.

The process began with scenario and text generation. Human experts, assisted by Gemini-2.5-pro (Team et al., 2023; Comanici et al., 2025), crafted utterance texts that were laden with clear emotional sentiment. The emotion inherently implied by this verbal content was identified (e.g., *angry* for the text "I'm going to give you one last chance") and served as the semantic anchor for creating the conflict. Subsequently, in the metadata annotation stage, the core conflict was deliberately engineered. For each sample, we designated a ground-truth acoustic emotion with the strict constraint that it must conflict with the inherent semantic emotion of the text. To enhance diversity, a speaker timbre was also randomly selected from a pool of 21 multi-emotion voices. This complete set of metadata—text, target emotion, and timbre—was then fed into the state-of-the-art TTS model Doubao-Seed-TTS 2.0 (Anastassiou et al., 2024) for audio generation.

Finally, all synthesized samples underwent a rigorous manual verification by a panel of 12 human experts. The evaluation focused on whether the audio successfully projected the intended acoustic emotion with clarity, even when contradicted by the text. Samples were discarded if the acoustic prosody was perceived as weak, ambiguous, or overshadowed by the semantics. This stringent quality control process ensures that every sample in CASE presents a clear and challenging instance of affective conflict, culminating in a final benchmark of 378 high-quality samples.

## 4 Experiments

### 4.1 Datasets

Table 1 provides an overview of the datasets used in our experiments, they are categorized into two groups: for training, in-domain testing, and for zero-shot evaluation.

To build a generalized model, we aggregated multiple open-source datasets into a large-scale, heterogeneous training corpus totaling over 66 hours. This includes MER2024 (Lian et al., 2024), a multilingual video-based emotion recognition corpus; IEMOCAP (Busso et al., 2008), a dyadic conversational dataset of naturalistic emotional speech; CMU-MOSEI (Bagher Zadeh et al., 2018), the largest sentiment analysis dataset with diverse topics and speakers; MELD (Poria et al., 2019), a TV dialogue dataset from Friends; RAVDESS (Livingstone and Russo, 2018), a database of acted emo-

| Train & In-Domain Test Sets | | | | | | |
|---|---|---|---|---|---|---|
| **Dataset** | **#Emo** | **Utts.** | **#Hours** | **Train** | **Test** | **Lang** |
| IEMOCAP (Busso et al., 2008) | 5 | 10,039 | 7.0 | ✓ | - | English |
| CMU-MOSEI (Bagher Zadeh et al., 2018) | 7 | 5,239 | 9.3 | ✓ | - | English |
| MER2024 (Lian et al., 2024) | 6 | 5,030 | 5.9 | ✓ | - | Multilingual |
| MELD (Poria et al., 2019) | 7 | 13,847 | 12.2 | 11.2 | 1.0 | English |
| RAVDESS (Livingstone and Russo, 2018) | 8 | 2452 | 2.8 | 2.3 | 0.5 | English |
| ESD (Zhou et al., 2022) | 5 | 17,500 | 29.0 | 23.7 | 5.3 | Multilingual |
| **Zero-Shot Test Sets** | | | | | | |
| **CASE** (Ours) | 7 | 378 | 0.32 | - | ✓ | Multilingual |
| Emo-Emilia (Zhao et al., 2025) | 7 | 1400 | 3.29 | - | ✓ | Multilingual |
| EMOVO (Costantini et al., 2014) | 7 | 588 | 0.51 | - | ✓ | Italian |
| EmoDB (Burkhardt et al., 2005) | 7 | 535 | 0.41 | - | ✓ | German |

Table 1: Overview of datasets used for training and evaluation. "#Emo" denotes the number of emotion classes, "Utts." refers to the number of utterances, and "#Hours" indicates the duration. The "Train" and "Test" columns specify the usage of each dataset in hours or with a "✓" for full inclusion.

tional speech and song by 24 professional actors; and ESD (Zhou et al., 2022), a multilingual emotional speech dataset with 350 parallel utterances from 10 English and 10 Chinese native speakers.

Our evaluation is twofold. For in-domain testing, we use the official test splits of MELD, RAVDESS, and ESD to assess performance on distributions similar to training. More critically, our zero-shot evaluation (Table 1) measures language generalization on out-of-domain data. CASE serves as the primary testbed, designed to probe robustness against acoustic-semantic conflicts. We also include Emo-Emilia (Zhao et al., 2025), EMOVO (Costantini et al., 2014), and EmoDB (Burkhardt et al., 2005)—covering Mandarin, Italian, and German—to evaluate cross-lingual and cross-corpus generalization, contrasting with the more spontaneous nature of our training data. This rigorous zero-shot protocol is essential for verifying whether the model has learned transferable representations of emotion, rather than overfitting to the characteristics of the training sets.

### 4.2 Implementation Details

Our experiments were conducted on $8 \times$ NVIDIA A6000 GPUs. A fixed random seed of 42 was used for all experiments to ensure reproducibility. For the Semantic Pathway, we use the encoder from the pre-trained Whisper-large (Radford et al., 2023) model to extract a 1280-dimensional feature. For the Acoustic Pathway, we employ the MingTok-Audio (Yan et al., 2025) tokenizer to extract a 64-dimensional feature. Then our proposed FAS is trained from scratch on these pre-computed features. This approach accelerates experimentation by decoupling the heavy feature extraction from the training of the lightweight fusion module. As shown in Table 3, the fusion module is configured with a unified hidden dimension of $d = 512$. The entire model is trained end-to-end using the AdamW optimizer (Loshchilov and Hutter, 2017) with an initial learning rate of $2 \times 10^{-4}$ and a weight decay of $1 \times 10^{-4}$. We use a global batch size of 2048 and train for 100 epochs, with Cross-Entropy Loss as the optimization objective.

### 4.3 Evaluation Metrics and Comparison baselines

We employ two standard metrics for the evaluation: Accuracy (ACC) and Unweighted Average F1 score. Accuracy provides a global measure of correctness, while F1 is crucial for assessing performance on imbalanced datasets.

Our comparative evaluation includes a rigorous selection of strong baselines across diverse representational paradigms. Our baselines span multiple paradigms: (1) self-supervised speech models (HuBERT (Hsu et al., 2021), WavLM (Chen et al., 2022), wav2vec 2.0 (Baevski et al., 2020)); (2) large-scale speech-text pre-trained encoders (Whisper (Radford et al., 2023), CLAP (Elizalde et al., 2023)); (3) neural audio tokenizers for TTS (EnCodec (Défossez et al., 2022), VibeVoice (Peng et al., 2025), MingTok-Audio (Yan et al., 2025)); (4) current SOTA SER methods (Emotion2Vec (Ma et al., 2024), Vesper (Chen et al., 2024)); and (5) Audio Language Models (ALMs) such as $C^2$SER 7B (Zhao et al., 2025), Qwen2-Audio-Instruct 7B (Wang et al., 2024) and Qwen2.5-Omni 7B (Xu et al., 2025), evaluated via their official pipeline

| Modality | | Methods | In-Domain | | | | | | Zero-Shot | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | MELD | | RAVDESS | | ESD | | CASE | | Emo-Emilia | | EMOVO | | EmoDB | |
| Sem. | Aco. | | ACC | F1 | ACC | F1 | ACC | F1 | ACC | F1 | ACC | F1 | ACC | F1 | ACC | F1 |
| *SSL Models* | | | | | | | | | | | | | | | | |
| ✓ | ✓ | HuBERT | 45.99 | 38.07 | 69.96 | 68.79 | 80.10 | 79.13 | 32.90 | 32.24 | 34.36 | 29.95 | 29.05 | 19.10 | 52.99 | 53.57 |
| ✓ | ✓ | WavLM | 44.64 | 34.09 | 61.90 | 60.69 | 77.43 | 76.63 | 34.20 | 33.92 | 35.64 | 29.33 | 30.74 | 21.40 | 64.18 | 56.74 |
| ✓ | ✓ | wav2vec 2.0 | 37.44 | 28.36 | 23.59 | 16.44 | 21.81 | 19.57 | 25.59 | 22.83 | 19.64 | 16.02 | 21.11 | 15.28 | 25.37 | 22.09 |
| *Semantic Models* | | | | | | | | | | | | | | | | |
| ✓ | | Whisper | 49.59 | 43.62 | 62.30 | 60.61 | 84.53 | 83.92 | 47.26 | 44.97 | 50.50 | 42.29 | **43.07** | **34.38** | 68.84 | 63.46 |
| ✓ | | CLAP | 43.74 | 34.96 | 47.38 | 44.83 | 59.97 | 59.40 | 34.46 | 31.93 | 24.93 | 18.83 | 34.63 | 26.57 | 43.28 | 41.45 |
| *Neural Audio Tokenizers* | | | | | | | | | | | | | | | | |
| | ✓ | EnCodec | 34.38 | 29.41 | 22.78 | 18.44 | 28.73 | 25.18 | 24.54 | 18.81 | 15.64 | 11.47 | 24.66 | 18.33 | 29.10 | 23.87 |
| | ✓ | Vibevoice | 39.06 | 30.43 | 37.90 | 31.48 | 37.61 | 36.40 | 30.03 | 25.90 | 19.36 | 15.35 | 22.47 | 17.95 | 28.73 | 20.92 |
| | ✓ | MingTok-Audio | 41.94 | 29.76 | 36.49 | 26.12 | 29.93 | 28.80 | 29.24 | 25.61 | 21.07 | 16.34 | 21.45 | 16.34 | 37.31 | 34.02 |
| *Other Open-Sourced SER Methods* | | | | | | | | | | | | | | | | |
| | | Emotion2Vec | 45.04 | 45.49 | 70.06 | 68.84 | 51.39 | 50.87 | 31.48 | 28.42 | 52.79 | 50.44 | 33.53 | 29.01 | 71.21 | 76.07 |
| | | Vesper † | 25.00 | 45.70 | - | - | - | - | - | - | - | - | - | - | - | - |
| *Audio Language Models* | | | | | | | | | | | | | | | | |
| ✓ | ✓ | $C^2$SER † | 51.39 | 27.45 | - | - | **93.86** | 68.19 | - | - | 68.29 | 61.28 | 37.59 | 27.33 | - | - |
| ✓ | | Qwen2-Audio | 35.29 | 29.91 | **85.74** | **86.59** | 36.99 | 23.35 | 32.53 | 27.08 | 69.64 | **68.81** | 26.87 | 20.29 | 74.21 | 70.29 |
| ✓ | | Qwen2.5-Omni | **54.06** | 36.05 | 75.35 | 74.98 | 51.60 | 35.70 | 34.66 | 30.21 | **70.64** | 68.03 | 27.89 | 20.03 | **87.85** | **85.97** |
| ✓ | ✓ | **FAS (Ours)** | 51.89 | **48.42** | 76.61 | 76.19 | 87.27 | **86.72** | **59.38** | **55.08** | 51.14 | 42.92 | 40.03 | 33.39 | 68.10 | 65.07 |

Table 2: In-domain and Zero-shot generalization performance across all datasets. The table is categorized by model paradigm. The first two columns indicate whether the baseline model primarily captures Semantic (**Sem.**) or Acoustic (**Aco.**) information. For each benchmark, we report both Accuracy (**ACC**) and F1 Score (**F1**). **AVG** represents the average ACC and F1 across these three in-domain sets. "†" denotes their results are from the official Versper (Chen et al., 2024) and $C^2$SER(Explicit CoT) (Zhao et al., 2025) paper. The best results are **bolded** and the second-best results are underlined.

| Hyperparameters | FAS | Concat&Gated |
|---|---|---|
| Hidden Dimension ($d$) | 512 | 512 |
| Query Length ($N_q$) | 2 | - |
| Dropout Rate | 0.4 | 0.4 |
| Optimizer | AdamW | |
| Learning Rate | $2 \times 10^{-4}$ | |
| LR Schedule | Cosine | |
| Weight Decay | $1 \times 10^{-4}$ | |
| Global Batch Size | 2048 | |
| Loss | Cross-Entropy | |
| Epochs | 100 | |
| Warmup Ratio | 0.05 | |
| Sample Rate | 16000 | |

Table 3: Hyperparameter settings for the experiments. The 'Concat & Gated' column specifies the shared parameters for the Concatenation and Gated Fusion baselines, which are used for the ablation studies in Section 4.5.1.

with standardized emotion prompts.

To ensure a fair and rigorous comparison across these baselines, for all non-ALM models, we freeze the pre-trained encoder, extract mean-pooled utterance embeddings, and train a lightweight two-layer classifier. For task-specific models trained on different emotion taxonomies, their outputs are projected onto a unified label space consistent with our target benchmarks for comparison.

### 4.4 Main Results

We evaluate FAS on both in-domain (MELD, RAVDESS, ESD) and zero-shot (CASE, Emo-Emilia, EMOVO, EmoDB) settings. As shown in Table 2, FAS achieves state-of-the-art results across the board: it obtains an average in-domain ACC of **71.92%**, outperforming SSL models, semantic encoders, audio tokenizers, and even large audio-language models (ALMs) like Qwen2-Audio and Qwen2.5-Omni. Notably, while Qwen2.5-Omni excels on high-resource datasets (e.g., 87.85% ACC on EmoDB), it underperforms on challenging zero-shot benchmarks such as CASE (34.66%) and EMOVO (27.89%). In contrast, FAS delivers robust performance—reaching **59.38% SOTA ACC on CASE** and **54.66% average ACC** across all zero-shot tasks—demonstrating its ability to generalize under distribution shift. This performance gap reveals a fundamental trade-off in current ALMs: their architecture is optimized for alignment with the LLM backbone, which emphasizes textual semantics while drop the affective nuances carried

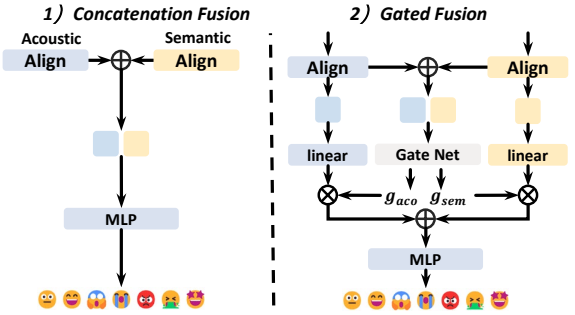



Figure 3: Illustration of different fusion strategies, including Concatenation Fusion and Gated Fusion.

by acoustic prosody. By explicitly modeling interactions between prosody and semantics, FAS bridges this gap, enabling reliable emotion recognition in both familiar and unseen scenarios. More experimental results including loss curves, confusion matrices, and visualizations of features map are provided in the *Appendix Section A.2*.

## 4.5 Ablation Study

To validate the superiority within our FAS framework, we conduct a series of comprehensive ablation studies to answer three central questions:

(1) How is the proposed FAS compared to other strategies? (2) Is the FAS framework a generalizable "plug-and-play" solution, or is its success tied to a specific model pair? (3) How does the internal configuration of the FAS influence its ability?

### 4.5.1 Efficacy of FAS framework

| Methods | Param | CASE | | MELD | | RAVDESS | |
|---|---|---|---|---|---|---|---|
| | | ACC | F1 | ACC | F1 | ACC | F1 |
| Concat | 1.22M | 53.65 | 50.77 | 52.88 | 48.50 | 75.40 | 75.22 |
| Gated | 1.65M | 53.12 | 50.84 | **52.70** | 47.92 | 73.79 | 73.59 |
| w/o Top-K | 3.45M | 55.47 | 51.57 | 52.70 | **48.60** | 73.79 | 73.32 |
| w/o $Q_{learn}$ | 0.82M | 55.99 | 52.48 | 48.65 | 44.71 | 67.34 | 66.95 |
| FAS | 3.45M | **59.38** | **55.08** | 51.89 | 48.42 | **76.61** | **76.19** |

Table 4: Ablation study of fusion methods on CASE (zero-shot), MELD, and RAVDESS, all built upon Whisper and MingTok-Audio encoders. "w/o Top-K" removes token selection (uses random token); "w/o $Q_{\text{learn}}$" removes learnable queries. Best results are **bolded**.

To validate the core design of the FAS framework, we compare against both classical fusion baselines and key architectural ablations. As shown in Table 4, naive strategies like concatenation or gating offer limited gains, confirming that passive combination is insufficient for cross-type feature integration. Critically, removing either component of FAS leads to significant performance drops: (1) *w/o Top-K*—which random select tokens without energy scores—underperforms FAS by up to 3.91% ACC on CASE; (2) *w/o* $Q_{learn}$—which replaces learnable queries with original inputs, suffers a severe drop on RAVDESS (-9.27% ACC). FAS achieves consistent gains over strong baselines with only a negligible increase in parameters, demonstrating that its superiority stems from the synergistic design of token distillation and learnable queries.

### 4.5.2 Framework Generalizability

To demonstrate the generalization of our FAS framework, we evaluate its "plug-and-play" capability with diverse acoustic and semantic backbones. For the **acoustic pathway**, we substitute MingTok-Audio with VibeVoice and XCodec2. For the **semantic pathway**, we compare Whisper and CLAP—two contrastively trained models with distinct training objectives. As shown in Table 6, FAS consistently enables strong cross-modal fusion across all combinations. FAS outperforms single-pathway baselines (marked with "–"), confirming that gains stem from fusion rather than individual encoders. FAS w/ (Whisper+XCodec2) achieves the best MELD performance (52.34% ACC), while FAS w/ (Whisper+VibeVoice) yields the highest RAVDESS score (80.04% ACC)—both surpassing our default MingTok+Whisper pairing on their respective datasets. On the zero-shot CASE benchmark, MingTok+Whisper remains optimal (59.38% ACC), suggesting that tokenized continuous acoustic tokens better support cross-lingual transfer when paired with a strong semantic encoder. These results confirm that FAS effectively bridges heterogeneous features, and its performance scales with encoder quality rather than being tied to a fixed encoder pair.

### 4.5.3 Ablation on Hyper-parameters of FAS

Table 5 presents a systematic ablation on the number of retained acoustic ($k_{\text{aco}}$) and semantic ($k_{\text{sem}}$) tokens. Specifically, increasing $k_{\text{sem}}$ (e.g., from 8 to 16) consistently improves performance on zero-shot benchmarks—most notably on CASE (+0.79% ACC) and EmoDB (+1.35% ACC)—suggesting that richer semantic context enhances cross-lingual and cross-corpus transfer. In contrast, enlarging $k_{\text{aco}}$ provides marginal or even negative gains on in-domain datasets such as MELD and RAVDESS, indicating diminishing returns from redundant acous-

| $k_{aco}$ | $k_{sem}$ | MELD | | RAVDESS | | ESD | | CASE | | Emo-Emilia | | EMOVO | | EmoDB | | AVG | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ACC | F1 | ACC | F1 | ACC | F1 | ACC | F1 | ACC | F1 | ACC | F1 | ACC | F1 | ACC | F1 |
| 8 | 8 | 51.62 | 47.99 | 77.82 | 77.51 | 86.84 | 86.25 | 58.59 | 54.58 | 50.64 | 42.66 | 38.68 | 31.57 | 66.23 | 62.47 | 61.49 | 57.58 |
| 16 | 16 | 51.89 | 48.35 | 77.02 | 76.59 | 87.20 | 86.72 | 56.77 | 52.70 | 51.29 | 42.84 | 39.36 | 32.11 | 66.79 | 64.27 | 61.47 | 57.65 |
| 16 | 8 | 51.35 | 47.73 | 78.23 | 77.75 | 87.16 | 86.66 | 56.25 | 52.53 | 51.21 | 43.05 | 38.85 | 31.14 | 67.16 | 64.22 | 61.46 | 57.58 |
| 8 | 16 | 51.89 | 48.42 | 76.61 | 76.19 | 87.27 | 86.72 | **59.38** | **55.08** | 51.14 | 42.92 | 40.03 | 33.39 | **68.10** | **65.07** | **62.06** | **58.26** |
| 32 | 16 | 50.90 | 47.44 | 78.83 | 78.47 | 86.70 | 86.40 | 55.21 | 51.47 | **52.36** | **43.80** | **41.89** | **35.30** | 64.74 | 61.88 | 61.52 | 57.82 |
| 16 | 32 | **52.70** | **49.11** | **79.03** | **78.77** | **87.66** | **87.20** | 56.25 | 52.37 | 52.29 | 43.76 | 39.70 | 32.04 | 64.37 | 61.98 | 61.71 | 57.89 |

Table 5: Ablation on FAS hyper-parameters: acoustic sequence length ($k_{\text{aco}}$), semantic sequence length ($k_{\text{sem}}$). Hidden dimension ($d = 512$) and the length of learnable queries ($N_q = 2$) are fixed across all benchmarks. Best results per dataset are marked **bolded**.

| Acoustic | Semantic | CASE (ACC / F1) | MELD (ACC / F1) | RAVDESS (ACC / F1) |
|---|---|---|---|---|
| - | Whisper | 47.26/44.97 | 49.59/43.62 | 62.30/60.61 |
| Vibevoice | Whisper | 58.07/53.35 | 51.53/48.06 | **80.04/79.71** |
| XCodec2 | Whisper | 58.33/54.46 | **52.34/48.87** | 79.03/78.76 |
| - | CLAP | 34.46/31.93 | 43.74/34.96 | 47.38/44.83 |
| MingTok | CLAP | 33.85/31.03 | 40.83/36.06 | 62.50/62.04 |
| Vibevoice | CLAP | 36.72/34.19 | 35.43/34.07 | 63.31/62.72 |
| XCodec2 | CLAP | 32.55/30.63 | 43.26/38.14 | 59.07/58.81 |
| **MingTok** | **Whisper** | **59.38 /55.08** | 51.89/48.42 | 76.61/76.19 |

Table 6: Generalization of the FAS framework across diverse acoustic and semantic encoders on CASE (zero-shot), MELD, and RAVDESS. Results demonstrate that FAS consistently enables effective cross-type fusion regardless of encoder architecture.

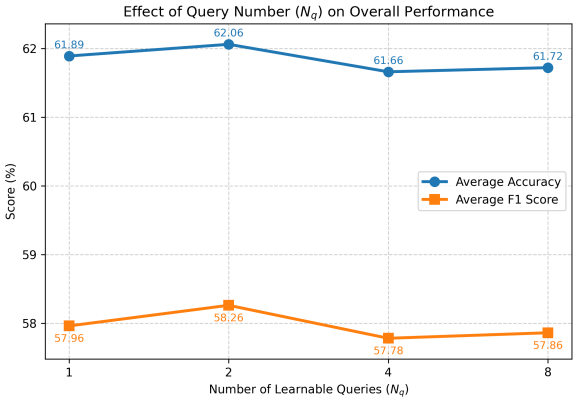



Figure 4: Ablation on the number of learnable queries ($N_q$). Average Accuracy and F1 scores are computed across all datasets.

tic frames. The best overall average performance (62.06% ACC) is achieved with $k_{\textbf{aco}} = 8$, $k_{\textbf{sem}} = 16$, revealing an asymmetric design principle: *preserving more semantic tokens is more beneficial than retaining additional acoustic ones* for both in-domain and zero-shot settings.

To further investigate the role of learnable queries ($N_q$), we conducted an ablation study varying $N_q$ from 1 to 8. As shown in Figure 4, we find that performance saturates at $N_q = 2$, with the best average ACC (62.06%) and F1 (58.26%). Increasing $N_q$ to 4 or 8 yields no gain—often slight degradation—while $N_q = 1$ already achieves competitive results (61.89% ACC). This confirms that SER, as a low-complexity utterance-level classification task, requires only minimal query capacity; additional queries introduce redundancy without improving generalization.

## 5 Conclusion

In this work, we address an overlooked challenge in SER-Acoustic-Semantic Conflict, where acoustic prosody conveys an emotion that contradicts the literal meaning. We demonstrate that current SER methods—ranging from ASR-based to SSL models—are brittle in such scenarios due to semantic bias or entangled representations. To tackle this, we propose the **FAS** framework, which explicitly disentangles and bridges acoustic and semantic pathways using a query–based fusion module. Along with the **CASE**, the first benchmark specifically designed to evaluate model robustness under emotional conflicts. Extensive experiments show FAS framework outperforms SOTA baselines across both in-domain and zero-shot settings. While FAS demonstrates strong performance as a lightweight SER method, its potential as an integrated component within end-to-end ALMs remains unexplored, which could be investigated in the future work.

## Limitations

While our work introduces a novel perspective on robust speech emotion recognition under acoustic-semantic conflict, several limitations delineate the current scope of our investigation and suggest promising avenues for future research. First, although the CASE benchmark incorporates multiple languages—including English, Mandarin, and representative Chinese dialects—its linguistic coverage remains limited. The phenomena of acoustic-semantic conflict may manifest differently across a broader range of language families, tonal systems, or cultural contexts. Second, CASE is designed primarily as a diagnostic evaluation suite, not a large-scale training resource. With fewer than 400 high-quality, human-verified conflict samples, it provides a controlled testbed for probing model robustness but is insufficient in scale to serve as standalone training data.

Finally, our current formulation centers on the binary tension between acoustic and semantic signals, which captures a prevalent and impactful class of conflicts (e.g., sarcasm, polite masking). However, real-world emotional expression can involve additional contextual cues—such as speaker identity or conversational history—that are not explicitly modeled in FAS. Incorporating these richer signals could enable even more nuanced conflict resolution in future systems.

## Ethical Considerations

We have taken several measures to ensure the ethical integrity of this research. All source datasets used for model training (IEMOCAP, MELD, CMU-MOSEI, etc.) are publicly available academic corpora that have been widely adopted in the affective computing community. Our newly introduced CASE benchmark is entirely synthetic, generated from scripted text prompts using a commercial-grade TTS engine. Consequently, it contains no personally identifiable information (PII) or recordings of real individuals, thereby mitigating privacy concerns associated with collecting sensitive emotional data.

The human verification process for CASE involved 12 expert annotators—recruited from our institution's pool of linguistics co-workers—who were provided with clear annotation guidelines and compensated at a standard academic rate, which is fair and adequate for their demographic and task complexity. Their task was limited to evaluating the perceptual quality and emotional clarity of the synthetic audio, not to disclose any personal information.

We acknowledge the potential for misuse of robust SER technology. A system capable of accurately inferring true emotions despite verbal content could be deployed in surveillance, manipulative advertising, or high-stakes interrogation settings without the subject's consent. To mitigate these risks, we emphasize that our work is intended solely for research purposes to improve the fundamental understanding of emotion expression. We advocate for the development and enforcement of strict ethical guidelines and regulatory frameworks governing the deployment of such technologies in real-world applications, ensuring user consent, transparency, and the right to opt-out.

## References


Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, and 1 others. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, Benjamin Weiss, and 1 others. 2005. A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Weidong Chen, Xiaofen Xing, Peihao Chen, and Xiangmin Xu. 2024. Vesper: A compact and effective pretrained model for speech emotion recognition. *IEEE Transactions on Affective Computing*, 15(3):1711–1724.

Zebang Cheng, Shuyuan Tu, Dawei Huang, Minghan Li, Xiaojiang Peng, Zhi-Qi Cheng, and Alexander G. Hauptmann. 2024. Sztu-cmu at mer2024: Improving emotion-llama with conv-attention for multimodal emotion recognition. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, MRAC '24, page 78–87, New York, NY, USA. Association for Computing Machinery.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, Massimiliano Todisco, and 1 others. 2014. Emovo corpus: an italian emotional speech database. In *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)*, pages 3501–3504. European Language Resources Association (ELRA).

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *Preprint*, arXiv:2210.13438.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Dongya Jia, Zhuo Chen, Jiawei Chen, Chenpeng Du, Jian Wu, Jian Cong, Xiaobin Zhuang, Chumin Li, Zhen Wei, Yuping Wang, and Yuxuan Wang. 2025. DiTAR: Diffusion transformer autoregressive modeling for speech generation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 27255–27270. PMLR.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Zheng Lian, Haiyang Sun, Licai Sun, Zhuofan Wen, Siyuan Zhang, Shun Chen, Hao Gu, Jinming Zhao, Ziyang Ma, Xie Chen, Jiangyan Yi, Rui Liu, Kele Xu, Bin Liu, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. 2024. Mer 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. *Preprint*, arXiv:2404.17113.

Steven R Livingstone and Frank A Russo. 2018. 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*, 13(5):e0196391.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15747–15760.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Zhiliang Peng, Jianwei Yu, Wenhui Wang, Yaoyao Chang, Yutao Sun, Li Dong, Yi Zhu, Weijiang Xu, Hangbo Bao, Zehua Wang, Shaohan Huang, Yan Xia, and Furu Wei. 2025. Vibevoice technical report. *Preprint*, arXiv:2508.19205.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 527–536.

Xin Qi, Yujun Wen, Pengzhou Zhang, and Heyan Huang. 2025. Mfgcn: Multimodal fusion graph convolutional network for speech emotion recognition. *Neurocomputing*, 611:128646.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Aaron Van Den Oord, Oriol Vinyals, and 1 others. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Canxiang Yan, Chunxiang Jin, Dawei Huang, Haibing Yu, Han Peng, Hui Zhan, Jie Gao, Jing Peng, Jingdong Chen, Jun Zhou, and 1 others. 2025. Minguniaudio: Speech llm for joint understanding, generation and editing with unified representation. *arXiv preprint arXiv:2511.05516*.

Zhen Ye, Peiwen Sun, Jiahe Lei, Hongzhan Lin, Xu Tan, Zheqi Dai, Qiuqiang Kong, Jianyi Chen, Jiahao Pan, Qifeng Liu, and 1 others. 2025a. Codec does matter: Exploring the semantic shortcoming of codec for audio language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25697–25705.

Zhen Ye, Xinfa Zhu, Chi-Min Chan, Xinsheng Wang, Xu Tan, Jiahe Lei, Yi Peng, Haohe Liu, Yizhu Jin, Zheqi Dai, and 1 others. 2025b. Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis. *arXiv preprint arXiv:2502.04128*.

Zongmeng Zhang, Wengang Zhou, Jie Zhao, and Houqiang Li. 2025. Robust multimodal large language models against modality conflict. In *Forty-second International Conference on Machine Learning*.

Zhixian Zhao, Xinfa Zhu, Xinsheng Wang, Shuiyuan Wang, Xuelong Geng, Wenjie Tian, and Lei Xie. 2025. Steering language model to stable speech emotion recognition via contextual perception and chain of thought. *arXiv preprint arXiv:2502.18186*.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 2022. Emotional voice conversion: Theory, databases and esd. *Speech Communication*, 137:1–18.

# A Appendix

## A.1 Case Study

To illustrate the challenge of acoustic-semantic conflict in emotion recognition, consider the example shown in Figure 5: "Well done—now we're both trapped." While the lexical content ("well done") typically signals positive sentiment, the speaker's angry tone reveals a negative emotional state. This discrepancy poses a significant challenge for existing models.

Previous approaches such as Whisper, which primarily rely on semantic understanding derived from text transcripts, are prone to misclassification due to their strong language priors. They interpret "well done" as inherently positive, leading to an incorrect prediction of happiness (see Figure 5). Similarly, SSL models like HuBERT and WavLM, though capable of extracting fine-grained acoustic features, operate on entangled representations where phonetic and prosodic information are not cleanly separated. As a result, even when they capture the angry prosody, the model lacks explicit mechanisms to resolve the conflict cues, defaulting to ambiguous predictions.

In contrast, our proposed **FAS framework** addresses this issue by explicitly modeling two parallel pathways. Through Top-K selection and query-based fusion, FAS detects discrepancies and applies a confidence-aware decision rule.

## A.2 Additional Analyses

### A.2.1 Loss Curve

To better understand the learning behavior of our proposed **Fusion Acoustic-Semantic (FAS)** framework compared to alternative strategies, we plot the training loss curves.

Figure 6 shows the training loss trajectories of multiple methods, including our proposed **FAS**, its ablation variants (w/ Concatenation Fusion and Gated Fusion), and several strong baselines. The results reveal that while FAS exhibits slower initial convergence, it eventually achieves a significantly lower final loss plateau compared to other methods. In contrast, both Concatenation and Gated Fusion converge more rapidly in the early stages but stabilize at higher loss values, indicating potentially poorer generalization. This delayed convergence may be attributed to the complexity of learning alignment through the $Q_{learn}$ fusion module, which requires more epochs to stabilize. However, once optimized, the learned representations demonstrate superior discriminative power.

### A.2.2 Space Visualization

To qualitatively assess the effectiveness of our FAS framework in learning discriminative emotion representations, we visualize the utterance-level embeddings using Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018).

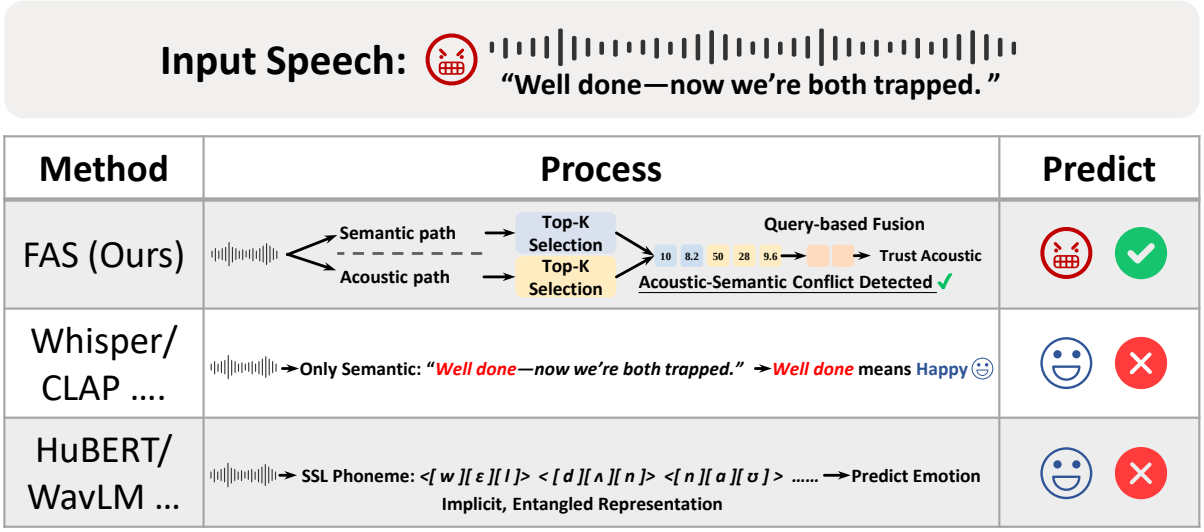



Figure 5: A comparison of SER methods on a conflicting utterance. FAS (Ours) explicitly disentangles acoustic and semantic pathways, detects their conflict via query-based fusion, and prioritizes the acoustic signal to correctly predict emotion.

Features are extracted from the final hidden layer of each model and projected onto a 2D space for comparison on the CASE benchmark.

As shown in Figure 7, baseline models such as HuBERT (Hsu et al., 2021) and Wav2Vec 2.0 (Baevski et al., 2020) exhibit highly mixed and indistinct clusters, indicating limited ability to disentangle emotional content from other factors. In contrast, our FAS framework yields clearly separated and well-structured clusters that strongly correlate with the ground-truth acoustic emotion labels.

We further evaluate our framework on EmoDB (Burkhardt et al., 2005), a German emotional speech corpus with limited data size and lower recording quality. As shown in Figure 8, while the overall clustering structure is less pronounced due to these challenges, our FAS framework still maintains relatively coherent emotion clusters, outperforming most baselines. This demonstrates the robustness of our approach across different languages and data conditions, reinforcing its practical applicability.

#### A.2.3 Confusion Matrix Analysis

To further evaluate the performance of our proposed FAS framework, we compare its confusion matrix against strong baselines on the CASE and RAVDESS datasets. As shown in Figure 9, FAS achieves the highest accuracy on major classes—particularly *anger*, *sadness*, and *neutral*—with significantly less confusion between high-arousal emotions (e.g. *anger* vs. *surprise*) than other models. Notably, no method correctly predicts any samples of *fear* or *disgust*, likely due to the high difficulty of the CASE benchmark and the scarcity of training data that effectively decouples acoustic and semantic cues.

Further validation on the well-structured RAVDESS dataset (Figure 10) confirms FAS's robustness: it exhibits minimal off-diagonal errors and outperforms baselines such as WavLM (10e), VibeVoice (9c), and Wav2Vec 2.0 (10f), especially in distinguishing subtle emotions like *neutral* and *sadness*.

In contrast, Audio Language Models (ALMs) show a mixed profile. As illustrated in Figure 11, models like Qwen2-Audio-Instruct and Qwen2.5-Omni achieve reasonable performance on datasets with consistent emotion-expression patterns—such as Emo-Emilia—where their massive pre-training allows them to "memorize" common acoustic-semantic mappings. However, under high-conflict or zero-shot conditions like CASE, they exhibit severe confusion between semantically proximate emotions (e.g. *angry* vs. *disgust*), revealing a fundamental reliance on lexical content over prosodic affect. This brittleness highlights the limitation of implicit fusion in LLM-aligned architectures.

FAS's consistent accuracy across both structured and ambiguous settings underscores its superior ability to capture fine-grained affective cues by explicitly modeling the interaction—and potential conflict—between acoustic and semantic modali-

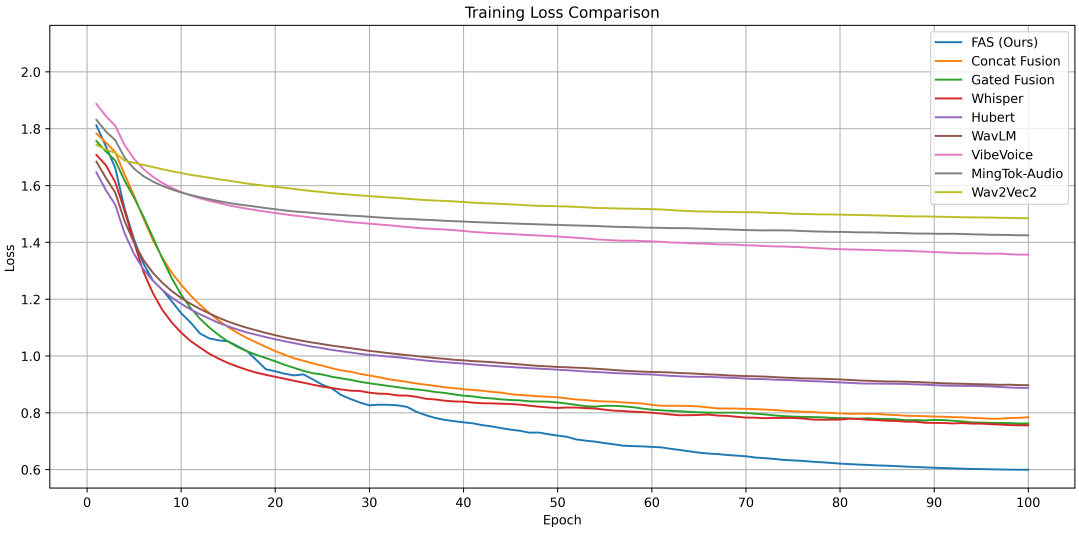



Figure 6: Training loss curves for different SER strategies. FAS demonstrates slower convergence but lower final loss.

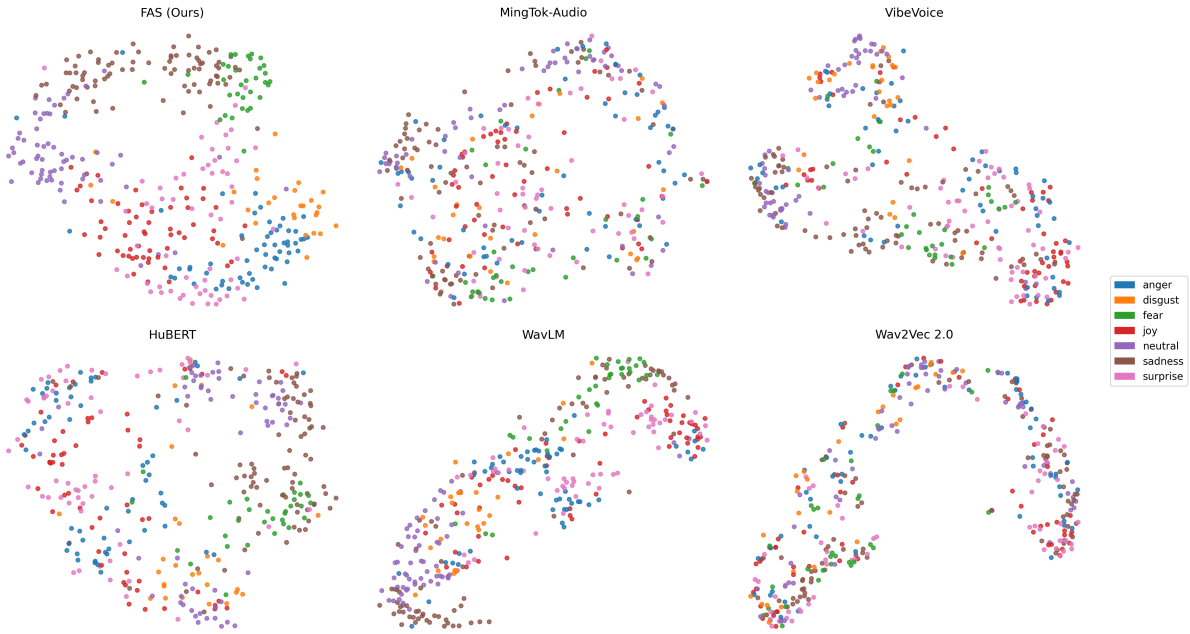



Figure 7: UMAP visualization of deep representations on CASE. Colors denote ground-truth acoustic emotion labels. FAS (top-left) achieves the most distinct emotion clusters, whereas HuBERT (bottom-left) and Wav2Vec 2.0 (bottom-right) show significant overlap.

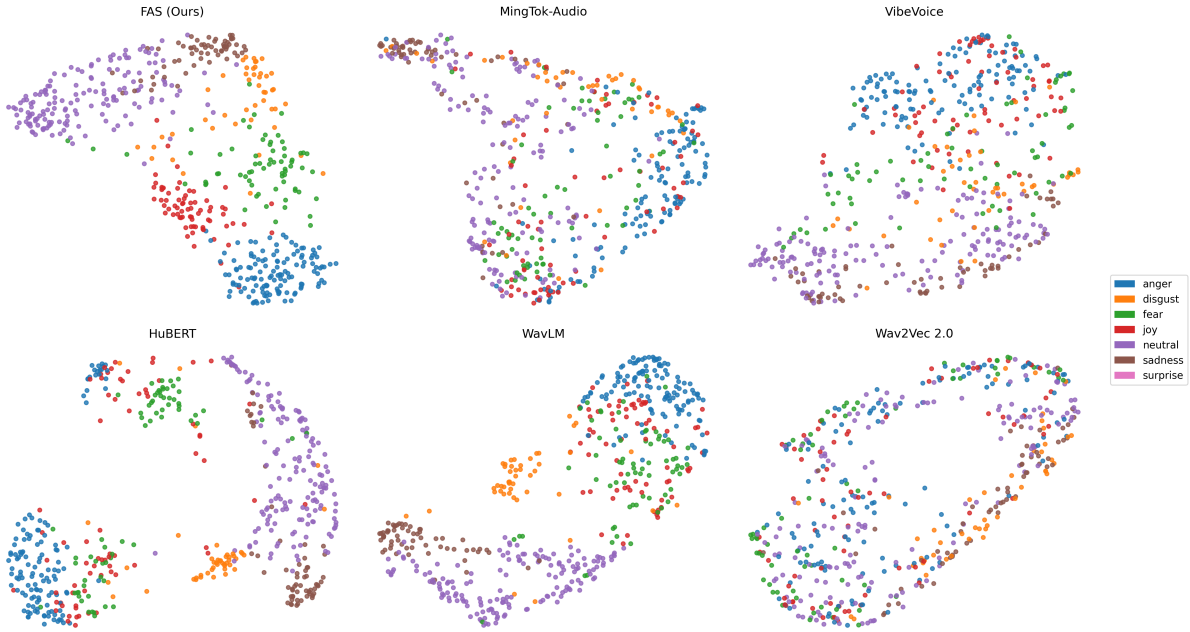



Figure 8: UMAP visualization of deep representations on EmoDB, a German emotional speech dataset. Colors denote ground-truth acoustic emotion labels. Despite the challenges of cross-lingual generalization and limited data size, our FAS framework still produces relatively structured clusters compared to baseline models, demonstrating its robustness in diverse conditions.

ties.

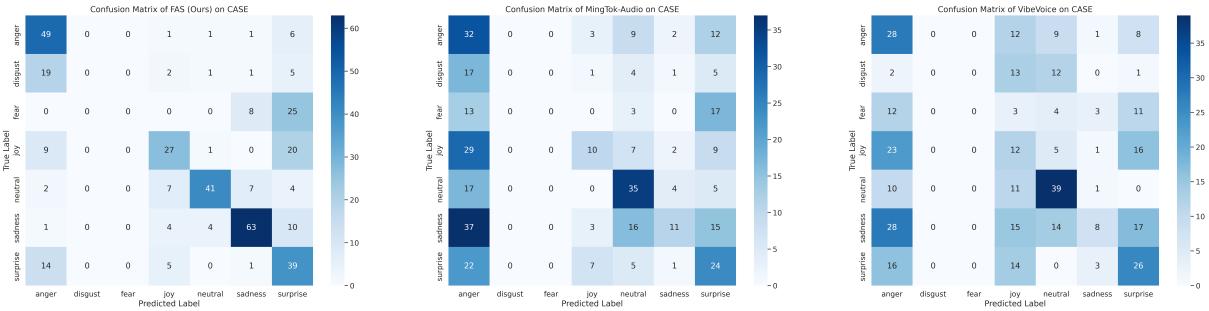
### A.3 Illustrative Samples of CASE Benchmark

To provide concrete examples of the acoustic-semantic conflict embodied in our benchmark, we list a selection of representative samples in Table 7. Each entry includes the original Chinese utterance, its ground-truth acoustic emotion (what is heard), the inherent semantic emotion (what the words imply), and a short narrative context that justifies the conflict. These scenarios are designed to be psychologically plausible, ensuring that the emotional dissonance arises from realistic human experiences rather than artificial manipulation.

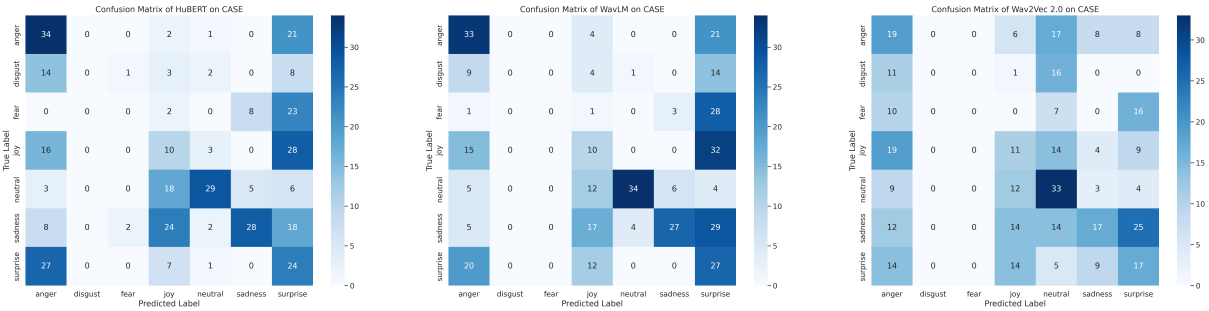
As shown, CASE encompasses a wide variety of psychologically plausible affective conflicts—ranging from fear-laden final words of love spoken by a soldier facing death (Sample 007) to cold contempt disguised as neutral inquiry after an irrational decision (Sample 011). These nuanced scenarios challenge models to disentangle acoustic prosody from semantic content, making CASE a rigorous testbed for evaluating emotional robustness in the presence of cross-modal conflict.

| ID | Utterance | GT Emo. | Semantic Emo. | Contextual Description |
|---|---|---|---|---|
| 001 | 那辆卡车失控了，正朝我们冲过来！ | surprised | fear | An adrenaline-seeking extremist expresses morbid excitement at danger. |
| 002 | 你们又赢了，恭喜啊。 | angry | happy | A loser congratulates the winner with barely concealed resentment. |
| 003 | 他走了，再也不会回来了。 | happy | sad | Someone oppressed for years feels secret joy at their tormentor's departure. |
| 004 | 你给我站住！你到底想怎么样！ | sad | angry | Exhausted from arguing; anger has turned into heartbreak and despair. |
| 005 | 任务完成，目标已清除。 | sad | neutral | A hitman reports completing a mission, but the target was an old friend. |
| 006 | Well done—now we're both trapped. | angry | happy | The speaker sarcastically blames their companion whose reckless actions led to a shared predicament, masking frustration with ironic praise. |
| 007 | Mom, Dad... I love you. | fear | happy | A soldier records a final message to his parents before a suicide mission; his voice trembles with terror despite the loving words. |
| 008 | No way—he was already dead! | fear | surprised | In a horror scenario, the protagonist witnesses a supposedly slain villain rise again, reacting with visceral fear beneath an initial gasp of shock. |
| 009 | The emergency exit is blocked. | fear | neutral | During a fire, someone announces the only escape route is sealed—their tone calm in wording but laced with palpable panic and dread. |
| 010 | What? You actually served this? | hate | surprised | A fastidious food lover reacts to a revolting "gourmet" dish with immediate disgust, their shock quickly overtaken by intense loathing. |
| 011 | So this is your final decision, then? | hate | neutral | After hearing an utterly unreasonable choice, the speaker delivers a cold, detached confirmation that conveys silent contempt and resignation. |
| 012 | By the way, the building is on fire. We should probably leave. | neutral | tension | A character with a dry, British sense of humor and extreme stoicism delivering urgent, life-threatening news in a casual, conversational tone. |
| 013 | Oh, a surprise party for me? You shouldn't have. | angry | neutral | An introvert who hates surprises is trying to be polite, but their voice is filled with irritation and anger. |
| 014 | I hate you! I never want to see you again! | sad | angry | Saying hateful words during a breakup, but the underlying emotion is one of heartbreak and sadness. |
| 015 | I love it. Another spreadsheet. | sad | happy | An employee sarcastically commenting on being assigned more tedious work, their voice full of gloom. |
| 016 | You lost the game. It's over. | happy | neutral | A game show host playfully and cheerfully announcing bad news to a contestant. |
| 017 | And the winner is... not you. | happy | neutral | A game show host playfully and cheerfully announcing bad news to a contestant. |
| 018 | Don't worry about the dishes, I'll just do them. Again. | angry | neutral | A classic passive-aggressive roommate situation. The words are seemingly helpful, but the tone is dripping with anger and resentment. |
| 019 | I heard you got the promotion. I am so, so thrilled for you. | sad | excited | Congratulating a coworker who got the promotion they wanted. They are trying to be supportive, but their voice is filled with their own disappointment. |
| 020 | You're getting so mad over this little game, it's actually adorable. | happy | angry | A friend playfully teasing and taunting another friend who is getting frustrated while playing a video game. |

Table 7: Representative acoustic-semantic conflict samples from our proposed CASE benchmark. For full audio demonstrations and additional metadata, please refer to the publicly released dataset files in the open-sourced repository.

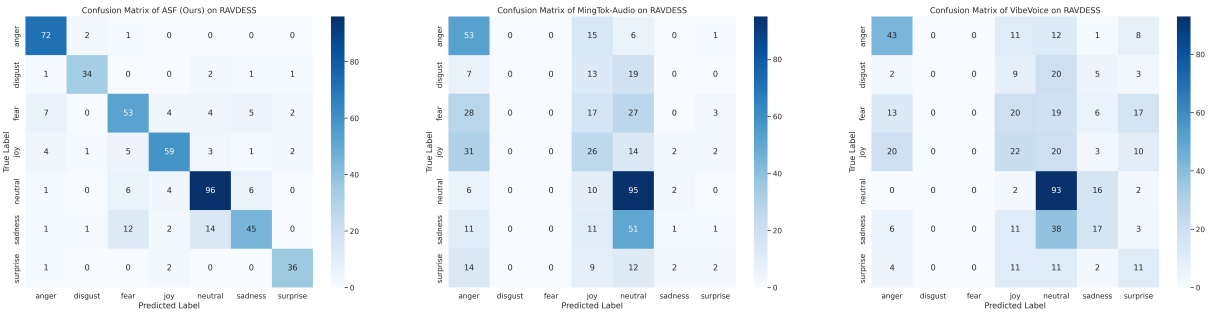



(a) FAS (Ours) (b) MingTok-Audio (Yan et al., 2025) (c) VibeVoice (Peng et al., 2025)

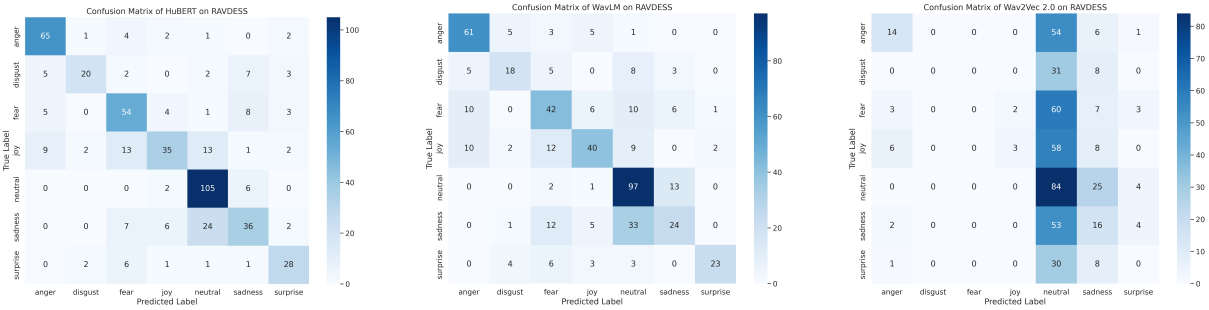



(d) HuBERT (Hsu et al., 2021) (e) WavLM (Chen et al., 2022) (f) Wav2Vec 2.0 (Baevski et al., 2020)

Figure 9: Confusion Matrices on the CASE dataset. Our FAS framework (9a) achieves the highest accuracy on major classes (*anger*, *sadness*, *neutral*) and shows minimal confusion between high-arousal emotions.

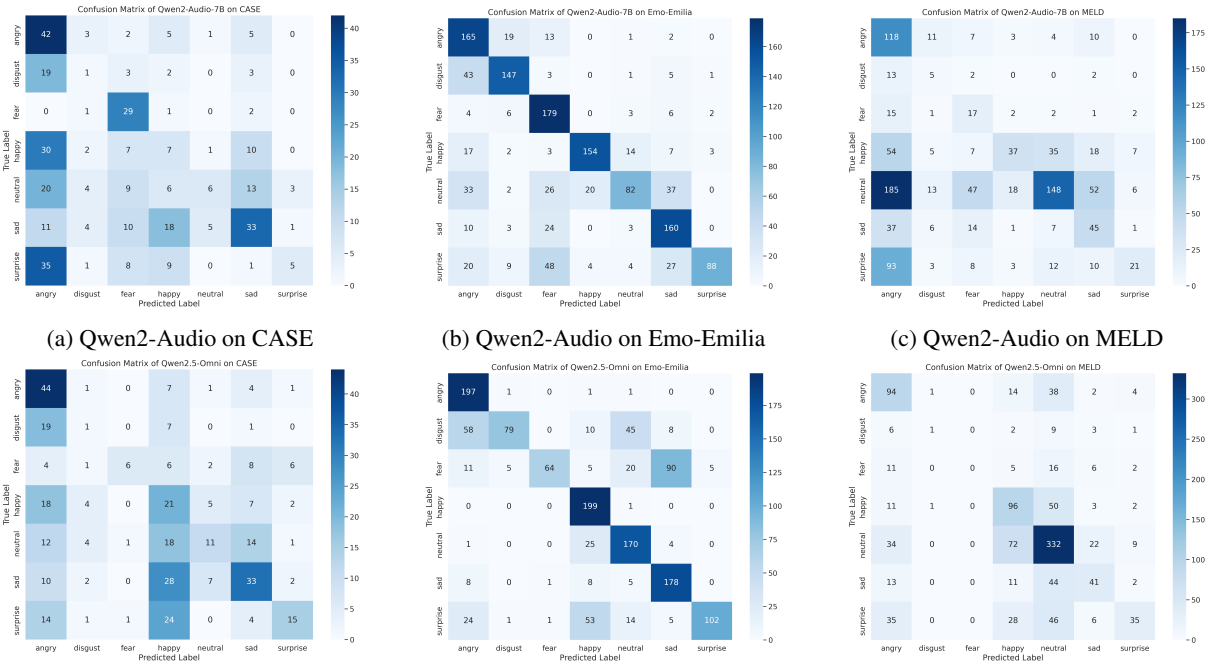



(a) FAS (Ours) (b) MingTok-Audio (c) VibeVoice




(d) HuBERT (e) WavLM (f) Wav2Vec 2.0

Figure 10: Confusion Matrices evaluated on the RAVDESS dataset. The FAS framework (10a) demonstrates superior performance across all emotion categories with minimal off-diagonal errors. In comparison, other models including MingTok-Audio (10b), VibeVoice (10c), HuBERT (10d), WavLM (10e), and Wav2Vec 2.0 (10f) show varying degrees of confusion between similar emotions.




(a) Qwen2-Audio on CASE

(b) Qwen2-Audio on Emo-Emilia

(c) Qwen2-Audio on MELD

(d) Qwen2.5-Omni on CASE

(e) Qwen2.5-Omni on Emo-Emilia

(f) Qwen2.5-Omni on MELD

Figure 11: Confusion Matrices of Qwen2-Audio-Instrcut and Qwen2.5-Omni on CASE, Emo-Emilia and MELD.