当语调与词语不一致：面向声学-语义冲突下的鲁棒语音情感识别

Dawei Huang[1,2], Yongjie Lv[1], RuijieXiong[1],Chunxiang Jin[1],Xiaojiang Peng[2]*, [1] Inclusion AI，蚂蚁集团，[2] 深圳大学，

arXiv:2601.04564v1 [cs.SD] 8 Jan 2026

摘要

语音情感识别（SER）系统通常假设语音情绪与词汇语义一致。然而，在现实世界的交互中，声学-语义冲突十分常见，却常被忽视；这种冲突表现为语调传达的情绪与口语词语的字面含义相矛盾。我们发现，最先进的语音情感识别模型，包括基于自动语音识别的方法、自监督学习（SSL）方法和音频语言模型（ALM），都会因语义偏差或声学与语义表征纠缠而在此类冲突下性能下降。为解决这一问题，我们提出**声学-语义融合（FAS）**框架，该框架明确分离声学与语义通路，并通过轻量级的基于查询的注意力模块将二者连接起来。为实现系统评估，我们构建了**声学-语义情感冲突（CASE）**数据集，这是首个以不同场景中清晰且可解释的声学-语义冲突为主的数据集。大量实验表明，FAS 在域内和零样本设置下均持续优于现有方法。值得注意的是，在 CASE 基准测试中，传统语音情感识别模型表现极差，而 FAS 则以 59.38% 的准确率达到新的最先进水平。我们的代码和数据集可在以下地址获取：https://github.com/24DavidHuang/FAS。

1 引言

语音情感识别（SER）是情感计算中的一项基础任务，旨在从说话者的声音表达中自动识别其情感状态。当前的语音情感识别方法（Ma 等人，2024年； Chen 等人，2024年； Elizalde 等人，2023年；Hsu 等人，2021； Radford 等人，2023年；Chen 等人，2022）在 IEMO-CAP（Busso 等人，2008）和 MELD（Poria 等人，2019年）等标准学术基准测试中展现出卓越的性能。然而，这种成功很大程度上局限于声学-语义一致性的情境，即语音中的韵律线索（声学）与所说内容的字面含义（语义）相符——例如，用喜悦的语气说出“多么美好的一天！”。

问题的关键在于，情感本质上十分复杂，并且经常在声学-语义冲突的情况下表现出来。现实世界的交流中充满了诸如讽刺、幸灾乐祸或冷酷的愤怒等细微表达：说话者通过声学线索传达的真实情感，与其话语的语义内容相互脱节，甚至截然相反。例如，得知同事获得晋升后，有人可能会用平淡或怨恨的语气说"恭喜你晋升！"，从而含蓄地流露出潜在的嫉妒，而非喜悦。在这些常见却极具挑战性的情境中，当前语音情感识别方法的性能会骤然崩溃。

这种脆弱性普遍存在于当前语音情感识别范式中。语音-文本预训练编码器（Radford 等，2023；Elizalde 等人，2023）表现出明显的语义偏差，因而容易被词语的字面含义“误导”。自监督学习（SSL）方法（Hsu 等人，2021；Baevski 等人，2020；Schneider 等人，2019；Chen 等人，2022）会生成纠缠表征，将情感信息与语义信息混为一谈，导致难以消解歧义。此外，显式多模态方法（Zhang 等人，2025年；Cheng 等人，2024年）也面临挑战，因为其现有的模态融合机制往往缺乏稳健的策略来裁决相互冲突的信号，最终默认采用具有误导性的模态。近期的音频语言模型（ALMs）虽然能力出众，却依赖与大语言模型对齐的编码器（例如 Whisper（Radford 等，2023）、CLAP（Elizalde 等人，2023））；这些编码器重语义、轻韵律，使大语言模型在情感冲突中无法获得情感线索。所有这些脆弱性都限制了语音情感识别模型在非受

*通讯作者。

限的真实环境中的可靠应用。

此外，当前语音情感识别评估受数据集偏差限制：大多数基准测试主要包含情感表达一致的样本。虽然自然场景数据集（如 MELD、IEMO-CAP）中偶尔会出现声学与语义冲突，但此类案例既稀少又缺乏系统性。更关键的是，现有资源均未提供高密度、可控的设置，无法系统评估情感冲突下的稳健性，因此现实场景表现中的一个关键方面尚未得到评估。

为应对这些挑战，本文提出了双管齐下的贡献。首先，我们提出一种创新的 **融合声学-语义（FAS）**框架，旨在明确分离语音中的声学信息和语义信息。受文本到语音生成领域近期进展的启发，FAS 独特地采用音频标记器提取低维声学令牌，同时利用预训练模块捕获高维语义信息。随后，我们引入一个轻量级的基于查询的模块，用于整合解耦后的特征并进行稳健预测。

其次，为了验证我们提出的框架，并为学界提供亟需的资源，我们发布了**声学-语义情感冲突（CASE）**基准测试。与传统数据集不同，CASE 的构建包含大量逻辑清晰、可解释且由具体情境驱动的冲突样本。它不仅是评估模型鲁棒性的高难度测试平台，也是研究者探讨人类情感表达中声学与语义相互作用的宝贵语料库。本文的主要贡献总结如下：

- 我们首次系统地研究了语音情感识别中的情感-语义冲突问题，并表明现有方法在此类样本上的性能会显著下降。
- 我们推出**CASE** 基准测试，这是首个旨在评估语音情感识别模型应对声学-语义冲突时鲁棒性的标准化数据集。
- 我们提出了**FAS** 框架，旨在通过解耦声学线索与语义内容来消除情感歧义。通过大量实验，我们证明了我们的 FAS 在所有基准测试中均优于最先进的基线方法。

2 相关工作

2.1 语音情感识别

近年来，语音情感识别（SER）的进展主要得益于 WavLM（Chen 等人，2022）、Whisper（Radford 等人，2023年）和 HuBERT（Hsu 等人，2021）等大规模预训练模型。大量研究（Ma 等人，2024年；Chen 等人，2024年；Qi 等人，2025年）着重于将这些强大编码器的表征迁移、蒸馏并适配到语音情感识别任务中。

然而，基于自动语音识别的编码器（例如 Whisper （ Radford 等人，2023 年）和 CLAP（Elizalde 等人，2023年）虽然性能强大，却因其预训练目标而表现出显著的语义偏差，因而极易受到口语字面含义的误导。此外，自监督学习编码器的一个根本局限（Hsu 等人，2021；Chen 等人，2022；Schneider 等人，2019年；Baevski 等人，2020）在于它们会生成纠缠表征。在这些习得的特征中，情感韵律与语音内容无法分离地混合在一起，而没有得到明确的解耦。这种混杂严重限制了它们的鲁棒性，尤其是在面临情感-语义冲突时。针对这一问题，我们提出了一种新颖的融合框架，直接解决这种表示纠缠和语义偏差。

2.2 神经音频标记化

文本转语音生成和音频编辑领域推动了高保真神经音频分词器的发展。基于离散音频的分词器，例如 EnCodec（Défossez 等，2022）、XCodec（Ye 等，2025a，b）和 VibeVoice（Peng 等，2025年），构建于 VQ-VAE框架（Van Den Oord 等，2017）之上，擅长将波形离散化，以实现高质量的信号重建。近年来，人们还提出了基于 VAE 的连续分词器，以更好地统一语义和声学信息，用于联合理解与生成任务（Yan 等，2025年；Jia 等，2025年），例如 MingTok-Audio（Yan 等，2025年）。这些面向生成的方法的共同特征是能够将语音提炼为低维且声学纯净的表征。不同于识别导向的编码器所产生的高维、

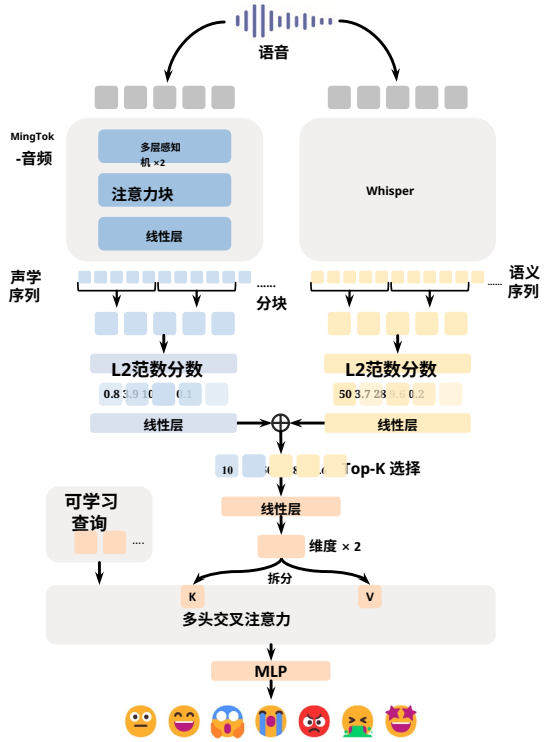


图 1：所提出的融合声学-语义（FAS）框架概览。该框架将语音解耦为声学（MingTok-Audio）和语义（Whisper）表征。融合模块在可学习查询的引导下，通过交叉注意力动态提炼出统一特征，以实现稳健的情感识别。

语义纠缠特征，这种紧凑表征能够明确捕捉细粒度的韵律和说话人身份，而这些对于高质量合成至关重要。

尽管其主要应用于音频生成任务，但作为判别式任务（如语音情感识别）的解耦声学表征来源，其潜力尚未得到探索。我们率先将音频词元重新用作专门建模声学情感特征的路径，旨在解决语音情感识别中的情感歧义，而传统方法对此无能为力。

3 方法

3.1 声学-语义融合框架概述

当前语音情感识别模型的局限性源于其依赖单一或纠缠的表征。我们的核心见解是有效融合两种异质且随时间变化的特征流：语义特征 $F_{\text{sem}} \in \mathbb{R}^{T_{\text{sem}} \times D_{\text{sem}}}$和声学特征 $F_{\text{aco}} \in \mathbb{R}^{T_{\text{aco}} \times D_{\text{aco}}}$，其中 T 和 D 分别表示序列长度和特征维度。

如图 1所示，我们提出了**声学-语义融合（FAS）**，这是一种两阶段融合框架：首先从特征流中智能蒸馏出显著令牌，然后通过交叉注意力机制将其连接起来。流程如下：

1. **分块化：** 为高效处理长序列，我们首先执行分块化步骤。每个序列 $F \in \mathbb{R}^{T \times D}$ 都按 $s = 5$ 的比例进行下采样。由此生成较短的块序列 $F' \in \mathbb{R}^{(T/s) \times D}$。随后，将这些块序列投影到统一隐藏维度 d：

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

其中 $f_{\text{aco}} \in \mathbb{R}^{T'_{\text{aco}} \times d}$和 $f_{\text{sem}} \in \mathbb{R}^{T'_{\text{语义}} \times d}$，且 $T' = T/s$。

2. **令牌蒸馏：** 鉴于情绪线索分布稀疏，我们引入一种非均匀令牌选择策略，从每个序列中识别并仅保留信息量最高的“重点”令牌。具体做法如下：

(a) **显著性评分：** 我们为序列中的每个词元 f_t 计算能量分数 s_t 。受情绪激昂事件往往与较高激活程度相关这一发现的启发，我们使用 L2范数作为其分数的近似值：

$$s_t = \|f_t\|_2 \tag{2}$$

这种快速的非参数方法能够有效捕捉声学流和文本流中的高能量时刻。

(b) **Top-K 选择：** 随后，我们选取显著性分数最高的 k个词元，其中 k_{aco} 和 k_{sem} 是序列长度，分别用于反映每条路径的信息密度差异。由此得到两个精简后的序列：

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

这一蒸馏过程在保留情感相关性最高的时间信息的同时，大幅缩短了序列长度，优于均匀压缩技术。

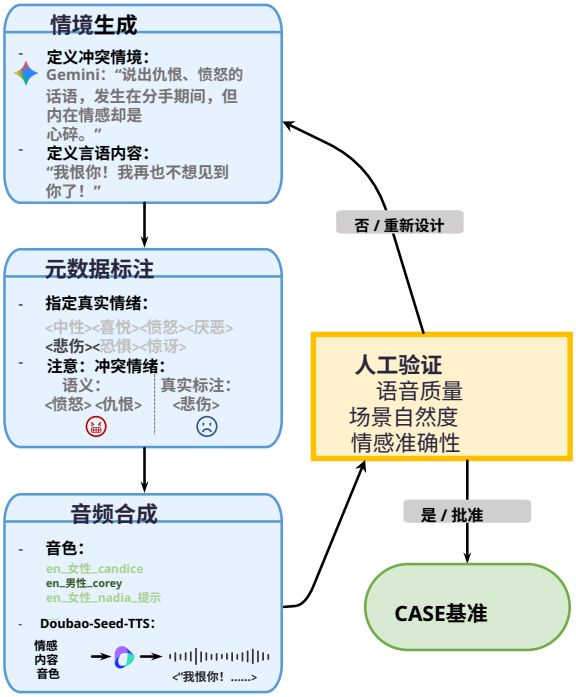


图2：CASE基准的构建流程。

3. **可学习查询：** 将蒸馏后的序列拼接成一个上下文序列 $C \in \mathbb{R}^{(k_{\text{aco}}+k_{\text{sem}})\times d}$。随后，我们采用受 Q-Former 类架构启发的融合模块（Li 等人，2023年）。一组 n 可学习查询 $Q_{\text{learn}} \in \mathbb{R}^{n\times d}$会主动查询该上下文，以生成学习得到的表示。上下文 C 经过投影后生成键（K）和值（V）矩阵，并通过交叉注意力机制计算最终融合后的词元：

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

最后，使用一个简单的多层感知机（MLP）作为融合向量之上的预测头，生成最终的情绪概率，$P(y|X)$，涵盖7种情绪类别。

3.2 声学-语义情感冲突（CASE）基准

为了严格评估各方法应对情感-语义冲突的鲁棒性，我们构建了**声学-语义情感冲突（CASE）**基准测试。这是一个专门的测试平台，旨在系统地探究语音情感识别模型在复杂情绪场景中的局限性。图2所示的构建流程遵循逻辑连贯性和高冲突密度原则，确保每个样本都基于合理的真实场景。

流程始于场景和文本生成。人类专家在 Gemini-2.5-pro 的辅助下（Team 等，2023年；Comanici 等，2025年），创作了带有鲜明情绪倾向的话语文本。识别文本内容内在传达的情感（例如，文本"I'm going to give you one last chance"表达的愤怒），并将其作为构建冲突的语义锚点。随后，在元数据标注阶段，研究人员有意设计核心冲突。对于每个样本，我们指定一种真实声学情感，并严格要求它必须与文本内在的语义情感相冲突。为提升多样性，还从包含 21 种多情绪语音的集合中随机选取一种说话者音色。随后，将这组完整的元数据——文本、目标情感和音色——输入最先进的文本转语音模型 Doubao-Seed-TTS 2.0（Anastassiou 等，2024年）进行音频生成。

最后，所有合成样本均由一个由12名人类专家组成的小组进行了严格的人工核验。评估重点在于，即使文本与音频所表达的情绪相矛盾，音频是否仍能清晰地传达预期的声学情绪。如果音频韵律给人的感受较弱、含糊不清，或被语义内容掩盖，则会弃用相应样本。这一严格的质量控制流程确保了 CASE 中的每个样本都清晰呈现出具有挑战性的情感冲突，最终形成了一个包含378个高质量样本的基准测试。

4 实验

4.1 数据集

表1 概述了我们实验中使用的数据集，这些数据集分为两组：用于训练、域内测试和零样本评估。

为构建泛化能力强的模型，我们整合了多个开源数据集，形成一个总计超过 66 小时的大规模异构训练语料库。其中包括 MER2024（Lian 等人，2024年），一个基于视频的多语言情感识别语料库；IEMOCAP（Busso 等人，2008），一个包含自然情感语音的双人对话数据集；CMU-MOSEI（Bagher Zadeh 等人，2018），规模最大的情感分析数据集，涵盖多种主题和说话者；MELD（Poria 等人，2019年），一个源自《老友记》的电视对话数据集；RAVDESS（Living- stone 和 Russo，2018），一个由专业演员演绎情绪语音和歌曲的数据库

数据集	#Emo	话语	#小时	训练集	Test	Lang
训练集与域内测试集						
IEMOCAP(Busso 等人,2008)	5	10,039	7.0	✓	-	英语
CMU-MOSEI（Bagher Zadeh 等人，2018）	7	5,239	9.3	✓	-	英语
MER2024（Lian 等人，2024年）	6	5,030	5.9	✓	-	多语言
MELD（Poria 等人，2019年）	7	13,847	12.2	11.2	1.0	英语
RAVDESS（Livingstone 和 Russo，2018）	8	2452	2.8	2.3	0.5	英语
ESD(Zhou 等人, 2022)	5	17,500	29.0	23.7	5.3	多语言
零样本测试集						
CASE（我们的方法）	7	378	0.32	-	✓	多语言
Emo-Emilia（Zhao 等人，2025年）	7	1400	3.29	-	✓	多语言
EMOVO（Costantini 等人，2014）	7	588	0.51	-	✓	意大利语
EmoDB（Burkhardt 等人，2005）	7	535	0.41	-	✓	德语

表1：训练和评估所用数据集概览。“#Emo”表示情绪类别数量，“Utts.”表示话语数量，“#Hours”表示时长。“Train”和“Test”列表示各数据集的使用时长，或以“✓”表示完整纳入。

；以及 ESD（Zhou 等人，2022），一个多语言情感语音数据集，包含 10 位英语母语者和 10 位中文母语者的 350 条平行话语。

我们的评估分为两部分。对于域内测试，我们使用 MELD、RAVDESS 和 ESD 的官方测试集划分，以评估模型在与训练数据分布相似的数据上的表现。更重要的是，我们的零样本评估（表1）衡量模型在域外数据上的语言泛化能力。CASE 是主要测试平台，旨在探究模型应对声学-语义冲突的鲁棒性。我们还纳入了 Emo-Emilia（Zhao 等人，2025年）、EMOVO（Costantini 等人，2014）和 EmoDB（Burkhardt 等人，2005），涵盖普通话、意大利语和德语，用于评估跨语言和跨语料库泛化；与之相比，我们的训练数据具有更强的自发性。严格的零样本评估方案对于验证模型是否学到了可迁移的情绪表征，而非过拟合训练集的特征，至关重要。

4.2 实现细节

我们的实验在 8 ×NVIDIA A6000 GPU 上进行。所有实验均使用固定随机种子42 以确保可复现性。对于语义路径，我们使用预训练 Whisper-large（Radford 等人，2023年）模型的编码器提取 1280 维特征。对于声学路径，我们采用 MingTok-Audio（Yan 等人，2025年）分词器提取 64 维特征。随后，我们使用这些预计算特征从头开始训练所提出的 FAS。通过将繁重的特征提取与轻量级融合模块的训练解耦，这种方法加快了实验进度。 如表 3所示，融合模块的统一隐藏维度配置为 $d = 512$。整个模型采用 AdamW 优化器（LoshchilovandHutter，2017），以端到端方式进行训练，初始学习率为 2×10^{-4}，权重衰减为 1×10^{-4}。全局批次大小设为2048，训练100 轮，并以交叉熵损失作为优化目标。

4.3 评估指标与比较基线

我们采用两个标准指标进行评估：准确率（准确率）和非加权平均 F1 分数。准确率提供了对整体正确性的衡量，而 F1 分数对于评估不平衡数据集上的性能至关重要。

我们的比较评估严谨地选取了涵盖多种表征范式的强基线方法。基线方法涵盖多个范式：(1) 自监督语音模型（HuBERT（Hsu 等人，2021）、WavLM（Chen 等人，2022）、wav2vec 2.0（Baevski 等人，2020））；(2) 大规模语音-文本预训练编码器（Whisper（Radford 等，2023）、CLAP（Elizalde 等人，2023））；(3) 用于文本转语音（TTS）的神经音频分词器（EnCodec（Défossez 等人，2022）、VibeVoice（Peng 等人，2025）、MingTok-Audio（Yan 等，2025））；(4) 当前最先进水平的语音情感识别方法（Emotion2Vec（Ma 等人，2024）、Vesper（Chen 等，2024））；以及 (5) 音频语言模型，例如 C^{2}SER 7B（Zhao 等，2025）、Qwen2-Audio-Instruct 7B（Wang 等人，2024）和 Qwen2.5-Omni 7B（Xu 等人，2025），并通过各自的官方流程进行评估

模态		方法	域内						零样本							
			MELD		RAVDESS		ESD		CASE		Emo-Emilia		EMOVO		EmoDB	
语义。	声学。		准确率	F1分数	准确率	F1分数	准确率	F1分数	准确率	F1分数	准确率	F1分数	准确率	F1分数	准确率	F1分数
自监督学习模型																
✓	✓	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
语义模型																
✓		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
神经音频分词器																
	✓	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
其他开源语音情感识别方法																
		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	-	-	-	-	-	-	-	-	-	-	-	-
音频语言模型																
✓	✓	C²语音情感识别 †	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（我们的方法）**	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

表 2：所有数据集上的域内和零样本泛化性能。该表按模型范式分类。前两列表示基线模型主要捕获语义（**语义**）还是声学（**声学**）信息。对于每个基准测试，我们均报告准确率（**准确率**）和 F1 分数（**F1 分数**）。**平均值**表示这三个域内数据集上的准确率和 F1 分数的平均值。“†”表示其结果来自 Versper（Chen 等人，2024年）和 C²语音情感识别（显式思维链）（Zhao et al.，2025年）论文。最佳结果以**粗体**标出，次佳结果以下划线标出。

超参数	FAS	拼接与门控
隐藏维度 (d)	512	512
查询长度 (N_q)	2	-
丢弃率	0.4	0.4
优化器	AdamW	
学习率	2×10^{-4}	
学习率调度	余弦	
权重衰减	1×10^{-4}	
全局批次大小	2048	
Loss	交叉熵	
训练轮数	100	
预热比例	0.05	
采样率	16000	

表3：实验的超参数设置。“拼接与门控”列列出了拼接和门控融合基线所共用的参数，这些参数用于第4.5.1节的消融研究。

并使用标准化的情绪提示。

为确保对这些基线方法进行公平、严谨的比较，对于所有非音频语言模型，我们冻结预训练编码器，提取经均值池化的语句嵌入，并训练一个轻量级两层分类器。对于采用不同情绪分类体系训练的任务专用模型，其输出将映射到与我们的目标基准测试一致的统一标签空间，以便进行比较。

4.4 主要结果

我们在域内（MELD、RAVDESS、ESD）和零样本（CASE、Emo- Emilia、EMOVO、EmoDB）设置下评估 FAS。如表 2所示，FAS 在各项评测中均取得最先进的结果：其域内平均准确率达到 **71.92 %**，优于自监督学习模型、语义编码器、音频分词器，甚至优于 Qwen2-Audio 和 Qwen2.5-Omni 等大型音频语言模型（ALM）。值得注意的是，尽管 Qwen2.5-Omni 在高资源数据集上表现出色（例如，在 EmoDB 上的准确率为 87.85%），但在 CASE（34.66%）和 EMOVO（27.89%）等具有挑战性的零样本基准测试中表现不佳。相比之下，FAS 的表现稳健——在 CASE 上达到 **59.38% 的最先进准确率**，并且**在所有零样本任务上的平均准确率达到 54.66%** ——展现出其在分布偏移情况下的泛化能力。这种性能差距揭示了当前音频语言模型的一项根本权衡：其架构针对与大语言模型骨干网络的对齐进行了优化，因而强调文本语义，却忽视了声学韵律所承载的情感细微差异。

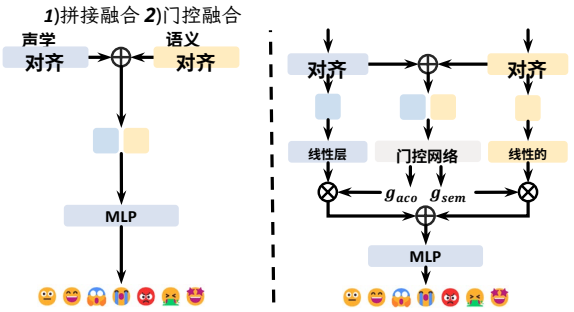


图 3：不同融合策略示意图，包括拼接融合和门控融合。

通过显式建模韵律与语义之间的交互，FAS 弥合了这一差距，使其能够在熟悉和未见过的场景中实现可靠的情感识别。更多实验结果，包括损失曲线、混淆矩阵和特征图可视化，见附录第A.2节。

4.5 消融研究

为验证 FAS 框架的优越性，我们开展了一系列全面的消融实验，以回答以下三个核心问题：

（1）所提出的 FAS 与其他策略相比表现如何？（2）FAS 框架是一种通用的“即插即用”解决方案，还是其成功依赖于特定的模型组合？（3）FAS 的内部配置如何影响其性能？

4.5.1 FAS 的有效性

方法	参数	CASE		MELD		RAVDESS	
		ACC	F1分数	准确率	F1分数	准确率	F1
拼接	1.22M	53.65	50.77	52.88	48.50	75.40	75.22
门控	1.65M	53.12	50.84	**52.70**	47.92	73.79	73.59
不使用 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**

表4：在 CASE（零样本）、MELD 和 RAVDESS 上对融合方法进行消融研究，所有方法均基于 Whisper 和 MingTok-Audio 编码器构建。“无 Top-K”表示移除令牌选择（改用随机令牌）；“无 Q_{learn}”表示移除可学习查询。最佳结果以**粗体**标示。

为验证 FAS 框架的核心设计，我们将其与经典融合基线方法及关键架构消融实验进行比较。如表 4所示，拼接或门控等朴素策略带来的收益有限，这证实了被动组合不足以实现跨类型特征整合。关键的是，移除 FAS 的任一组成部分都会导致性能显著下降：（1）*w/o Top-K*——随机选择令牌而不使用能量分数——在 CASE 上的准确率最多比 FAS 低 3.91%；（2）*w/o* Q_{learn}——以原始输入替代可学习查询——在 RAVDESS 上性能大幅下降（准确率降低 9.27%）。FAS 的表现始终优于强基线方法，而参数量仅略有增加，这表明其优势源于令牌蒸馏与可学习查询的协同设计。

4.5.2 框架泛化能力

为展示 FAS 框架的泛化能力，我们评估了其与多种声学和语义骨干网络的“即插即用”能力。对于**声学路径**，我们用 VibeVoice 和 XCodec2 替换 MingTok-Audio。对于**语义路径**，我们比较了 Whisper 和 CLAP——这两种对比训练模型具有不同的训练目标。如表6所示，FAS 在所有组合中均能实现强大的跨模态融合。FAS 的表现优于单路径基线（以“–”标记），这证实了性能提升源于融合，而非单个编码器。FAS 搭配（Whisper+XCodec2）在 MELD 上取得最佳表现（准确率为 52.34%），而 FAS 搭配（Whisper+VibeVoice）在 RAVDESS 上取得最高分（准确率为 80.04%）——两者在各自的数据集上均优于默认的 MingTok+Whisper 配对。在零样本 CASE 基准上，MingTok+Whisper 仍为最佳组合（准确率为 59.38%），这表明与强大的语义编码器配对时，标记化连续声学令牌更有利于跨语言迁移。这些结果证实，FAS 能有效融合异构特征，其性能随编码器质量提升，而非受限于固定的编码器组合。

4.5.3 FAS 超参数消融实验

表5对保留的声学（k_{aco}）和语义（k_{sem}）词元数量进行了系统的消融实验。具体而言，增加 k_{sem}（例如，从 8 增加到 16）能够持续提升零样本基准测试的性能——其中 CASE（+0.79% 准确率）和 EmoDB（+1.35% 准确率）的提升最为显著——这表明更丰富的语义上下文有助于跨语言和跨语料库迁移。相比之下，扩大k_{aco} 在 MELD 和 RAVDESS 等域内数据集上的收益很小，甚至会带来负面影响，说明冗余声

k_{aco}	k_{sem}	MELD		RAVDESS		ESD		CASE		Emo-Emilia		EMOVO		EmoDB		AVG	
		准确率	F1分数	准确率	F1分数	准确率	F1分数	准确率	F1分数	准确率	F1分数	准确率	F1分数	准确率	F1分数	准确率	F1分数
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**	**49.1179.03**	**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

表5：FAS 超参数消融实验：声学序列长度（k_{aco}）、语义序列长度（k_{sem}）。所有基准测试中的隐藏维度（$d = 512$）和可学习查询的长度（$N_q = 2$）均固定。每个数据集上的最佳结果以**粗体**标出。

声学	语义	CASE （准确率 / F1 分数）	MELD （准确率 / F1 分数）	RAVDESS （准确率 / F1 分数）
-	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

表6：FAS 框架在 CASE（零样本）、MELD 和 RAVDESS 上不同声学与语义编码器之间的泛化能力。结果表明，无论编码器架构如何，FAS 都能持续实现有效的跨类型融合。

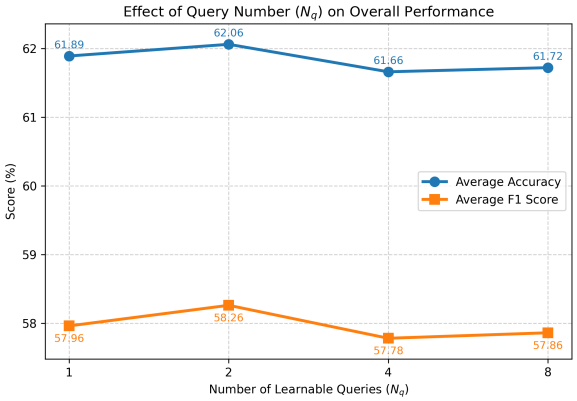


图 4：可学习查询数量的消融实验（N_q）。准确率和 F1 分数为所有数据集上的平均值。

声学帧。 $k_{\textbf{aco}} = 8, k_{\textbf{sem}} =$16取得了最佳的总体平均表现（62.06%准确率），这揭示了一种非对称设计原则：预先保留更多语义令牌，比保留额外的声学令牌更有益，无论是在域内还是零样本设置下。

为进一步探究可学习查询（N_q）的作用，我们进行了消融研究，将 N_q 从1逐步增加到8。如图4所示，我们发现性能在 $N_q = 2$时趋于饱和，此时平均准确率（62.06%）和F1分数（58.26%）均达到最佳。将 N_q 增加到4或8并未带来提升，反而常常导致性能略有下降；而 $N_q = 1$ 已取得有竞争力的结果（准确率为61.89%）。这证实了语音情感识别作为一种低复杂度的话语级分类任务，只需极少的查询容量；增加查询会引入冗余，却无法提升泛化能力。

5 结论

在这项工作中，我们探讨了语音情感识别中的一个长期被忽视的挑战——声学-语义冲突：声学韵律所传达的情感与字面含义相矛盾。我们证明，当前的语音情感识别方法——从基于自动语音识别的方法到自监督学习模型——在这类场景中都十分脆弱，原因在于语义偏差或纠缠表征。为应对这一问题，我们提出了**FAS** 框架，通过基于查询的融合模块，显式地解耦并连接声学路径与语义路径。与此同时，我们还推出**CASE**，这是首个专为评估模型在情感冲突下的鲁棒性而设计的基准测试。大量实验表明，FAS 框架在域内和零样本设置下均优于最先进的基线方法。尽管 FAS 作为一种轻量级语音情感识别方法表现出色，但其作为端到端音频语言模型中的集成组件所具备的潜力仍有待探索，可作为未来研究方向。

局限性

尽管我们的工作为声学-语义冲突下的鲁棒语音情感识别提出了全新视角，但仍存在若干局限，界定了当前研究的范围，也为未来研究指明了有前景的方向。首先，虽然 CASE 基准涵盖多种语言，包括英语、普通话和具有代表性的汉语方言，但其语言覆盖范围仍然有限。在更广泛的语系、声调系统或文化语境中，声学-语义冲突现象可能呈现不同形式。其次，CASE 主要被设计为诊断性评估套件，而非大规模训练资源。它包含不到 400 条经过人工核验的高质量冲突样本，可作为受控测试平台，用于探究模型鲁棒性，但规模不足以单独充当训练数据。

最后，我们当前的表述聚焦于声学信号与语义信号之间的二元紧张关系，这涵盖了一类普遍且影响深远的冲突（例如讽刺、礼貌性掩饰）。然而，现实世界中的情感表达还可能涉及其他情境线索，例如说话人身份或对话历史，而这些因素在 FAS 中并未得到明确建模。未来的系统若能纳入这些更丰富的信号，便有望实现更加细致的冲突消解。

伦理考量

我们采取了多项措施，以确保本研究符合伦理规范。所有用于模型训练的源数据集（IEMOCAP、MELD、CMU-MOSEI 等）均为公开可用的学术语料，且已在情感计算领域得到广泛采用。我们新引入的 CASE基准完全由合成数据构成，基于脚本化文本提示，使用商用级文本转语音引擎生成。因此，其中不包含任何个人身份信息（PII）或真实个人的录音，从而降低了收集敏感情感数据所引发的隐私问题。

CASE 的人工验证流程涉及 12 位专家标注员，他们从本机构的语言学同事中招募而来，并获得了明确的标注指南。报酬按标准学术费率支付，对其人口特征及任务复杂度而言公平且合理。他们的任务仅限于评估合成数据的感知质量和情感清晰度，不涉及披露任何个人信息。

我们承认，强健的语音情感识别技术可能被滥用。能够不受言语内容影响、准确推断真实情绪的系统，可能在未经当事人同意的情况下，被用于监控、操纵性广告或高风险审讯。为降低这些风险，我们强调，本研究仅用于研究目的，旨在加深对情绪表达的基础认识。我们倡导制定并落实严格的伦理准则和监管框架，以规范此类技术在现实应用中的部署，并确保用户同意、透明度以及退出权。

参考文献

Philip Anastassiou、Jiawei Chen、Jitong Chen、Yuanzhe Chen、Zhuo Chen、Ziyi Chen、Jian Cong、Lelai Deng、Chuang Ding、Lu Gao等。2024年。Seed-TTS：一系列高质量、多用途的语音生成模型。*arXiv*预印本 *arXiv:2406.02430*。

Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli。2020。wav2vec 2.0：一种用于语音表征自监督学习的框架。神经信息处理系统进展，33:12449–12460。

AmirAli Bagher Zadeh、Paul Pu Liang、Soujanya Poria、Erik Cambria 和 Louis-Philippe Morency。2018。真实场景中的多模态语言分析：CMU-MOSEI 数据集与可解释的动态融合图。载于第*56*届计算语言学协会年会论文集（第*1*卷：长篇论文），第2236–2246页，澳大利亚墨尔本。计算语言学协会。

Felix Burkhardt、Astrid Paeschke、Miriam Rolfes、Walter FSendlmeier、Benjamin Weiss 和1others。2005。德语情感语音数据库。载于*Interspeech*，第5卷，第1517–1520页。

Carlos Busso、Murtaza Bulut、Chi-Chun Lee、Abe Kazemzadeh、Emily Mower、Samuel Kim、Jean- nette N Chang、Sungbok Lee 和 Shrikanth S Narayanan。2008。Iemocap：交互式情感双人动作捕捉数据库。《语言资源与评估》，42(4):335–359。

Sanyuan Chen、Chengyi Wang、Zhengyang Chen、Yu Wu、Shujie Liu、Zhuo Chen、Jinyu Li、Naoyuki Kanda、Takuya Yoshioka、Xiong Xiao 和其他 1 位作者。2022。Wavlm：面向全栈语音处理的大规模自监督预训练。*IEEE*《信号处理精选主题期刊》，16(6):1505–1518。

Weidong Chen、Xiaofen Xing、Peihao Chen 和 Xiang-min Xu。2024年。Vesper：一种用于语音情感识别的紧凑高效预训练模型。*IEEETransactionson Affective Computing*，15(3):1711–1724。

Zebang Cheng、Shuyuan Tu、Dawei Huang、Minghan Li、Xiaojiang Peng、Zhi-Qi Cheng 和 Alexander G. Hauptmann。2024年。Sztu-cmu 在 MER2024 中：利用卷积注意力改进 E motion-LLaMA，以实现多模态情感识别。载于第二届多模态与负责任情感计算国际研讨会论文集，MRAC ’24，第 78–87 页，纽约，美国纽约州。计算机协会。

Gheorghe Comanici、Eric Bieber、Mike Schaekermann、Ice Pasupat、Noveen Sachdeva、Inderjit Dhillon、Marcel Blistein、Ori Ram、Dan Zhang、Evan Rosen 和其他 1 位作者。2025年。Gemini 2.5：借助高级推理、多模态性、长上下文和下一代智能体能力，拓展前沿。*arXiv* 预印本 *arXiv:2507.06261*。

Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, Massimiliano Todisco 和另外 1 位作者。2014年。Emovo语料库：意大利语情感语音数据库。载于第九届国际语言资源与评估会议（*LREC’14*）论文集，第 3501–3504 页。欧洲语言资源协会（ELRA）。

Alexandre Défossez、Jade Copet、GabrielSynnaeve 和 Yossi Adi。2022 年。高保真神经音频压缩。预印本，arXiv:2210.13438。

Benjamin Elizalde、Soham Deshmukh、Mahmoud Al Ismail 和 Huaming Wang。2023年。通过自然语言监督学习音频概念的 CLAP。载于 *ICASSP 2023——2023*年 *IEEE* 声学、语音与信号处理国际会议（*ICASSP*），第 1–5 页。

Wei-Ning Hsu、Benjamin Bolte、Yao-Hung Hubert Tsai、Kushal Lakhotia、Ruslan Salakhutdinov 和 Abdel-rahman Mohamed。2021年。HuBERT：通过隐藏单元掩码预测进行自监督语音表示学习。*IEEE/ACM* 音频、语音与语言处理汇刊，29:3451–3460。

Dongya Jia、Zhuo Chen、Jiawei Chen、Chenpeng Du、Jian Wu、Jian Cong、Xiaobin Zhuang、Chu min Li、Zhen Wei、Yuping Wang 和 Yuxuan Wang。2025年。DiTAR：用于语音生成的扩散变换器自回归建模。载于第 *42* 届国际机器学习会议论文集，第 267 卷，载于机器学习研究论文集，第 27255–27270 页。PMLR。

Junnan Li、Dongxu Li、Silvio Savarese 和 Steven Hoi。2023年。Blip-2：利用冻结图像编码器和大型语言模型引导语言-图像预训练。在国际机器学习会议，第 19730–19742 页。PMLR。

Zheng Lian、Haiyang Sun、Licai Sun、Zhuofan Wen、Siyuan Zhang、Shun Chen、Hao Gu、Jin ming Zhao、Ziyang Ma、Xie Chen、Jiangyan Yi、Rui Liu、Kele Xu、Bin Liu、Erik Cambria、Guoying Zhao、Björn W. Schuller 和 Jianhua Tao。2024年。Mer 2024：半监督学习、噪声鲁棒性和开放词汇多模态情感识别。预印本，arXiv:2404.17113。

Steven R Livingstone 和 Frank A Russo。2018年。瑞尔森情感语音与歌曲视听数据库（RAVDESS）：北美英语中面部和声音表达的动态多模态集合。*PLOS ONE*，13(5)：e0196391。

Ilya Loshchilov 和 Frank Hutter。2017年。解耦权重衰减正则化。*arXiv* 预印本 *arXiv:1711.05101*。

Ziyang Ma、Zhisheng Zheng、Jiaxin Ye、Jinchao Li、Zhifu Gao、Shiliang Zhang 和 Xie Chen。2024年。emotion2vec：用于语音情感表征的自监督预训练。载于计算语言学协会论文集：*ACL 2024*，第 15747–15760 页。

Leland McInnes、John Healy 和 James Melville。2018。UMAP：用于降维的均匀流形近似与投影。*arXiv* 预印本 *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 和 Furu Wei。2025 年。Vibevoice 技术报告。预印本，arXiv:2508.19205。

Soujanya Poria、Devamanyu Hazarika、Navonil Ma- jumder、Gautam Naik、Erik Cambria 和 Rada Mi- halcea。2019年。Meld：用于对话情感识别的多模态多方数据集。载于计算语言学协会第*57*届年会论文集，第527–536页。

Xin Qi、Yujun Wen、Pengzhou Zhang 和 Heyan Huang。2025年。 Mfgcn：用于语音情感识别的多模态融合图卷积网络。 神经计算，611：128646。

Alec Radford、Jong Wook Kim、Tao Xu、Greg Brock- man、Christine McLeavey 和 Ilya Sutskever。2023年。通过大规模弱监督实现稳健的语音识别。载于国际机器学习会议，第28492–28518页。PMLR。

Steffen Schneider、Alexei Baevski、Ronan Collobert 和 Michael Auli。2019年。wav2vec：用于语音识别的无监督预训练。*arXiv* 预印本 *arXiv:1904.05862*。

Gemini 团队、Rohan Anil、Sebastian Borgeaud、Jean- Baptiste Alayrac、Jiahui Yu、Radu Soricut、Johan Schalkwyk、Andrew M Dai、Anja Hauth、Katie Mil- lican 等人。2023年。Gemini：一系列能力卓越的多模态模型。*arXiv* 预印本 *arXiv:2312.11805*。

Aaron Van Den Oord、Oriol Vinyals 等人。2017。神经离散表示学习。神经信息处理系统进展，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 Z hou, Jingren Zhou, and Junyang Lin. 2024年。Qwen2-vl：增强视觉语言模型在任意分辨率下对世界的感知能力。预印本，arXiv:2409.12191。

Jin Xu、Zhifang Guo、Jinzheng He、Hangrui Hu、Ting He、Shuai Bai、Keqin Chen、Jialin Wang、Yang Fan、Kai Dang及其他 1 位作者。2025年。Qwen2.5-omni 技术报告。*arXiv* 预印本 *arXiv:2503.20215*。

Canxiang Yan、Chunxiang Jin、Dawei Huang、Haib- i ng Yu、Han Peng、Hui Zhan、Jie Gao、Jing Peng、Jingdong Chen、Jun Zhou及其他 1 位作者。2025年。Ming- UniAudio：采用统一表示的语音大语言模型，用于联合理解、生成与编辑。*arXiv* 预印本 *arXiv:2511.05516*。

Zhen Ye、Peiwen Sun、Jiahe Lei、Hongzhan Lin、Xu Tan、Zheqi Dai、Qiuqiang Kong、Jianyi Chen、Jiahao Pan、Qifeng Liu及其他1位作者。2025a。编解码器至关重要：探索音频语言模型中编解码器的语义缺陷。载于《*AAAI*人工智能会议论文集》，第39卷，第25697–25705页。

Zhen Ye、Xinfa Zhu、Chi-Min Chan、Xinsheng Wang、Xu Tan、Jiahe Lei、Yi Peng、Haohe Liu、Yizhu Jin、Zheqi Dai及其他1位作者。2025b。Llasa：扩展基于Llama的语音合成的训练时与推理时计算。*arXiv*预印本 *arXiv:2502.04128*。

Zongmeng Zhang、Wengang Zhou、Jie Zhao 和Houqiang Li。2025。抵御模态冲突的鲁棒多模态大型语言模型。载于第四十二届国际机器学习会议。

Zhixian Zhao、Xinfa Zhu、Xinsheng Wang、Shuiyuan Wang、Xuelong Geng、Wenjie Tian 和 Lei Xie。2025年。通过上下文感知和思维链引导语言模型，实现稳定的语音情绪识别。*arXiv* 预印本 *arXiv:2502.18186*。

Kun Zhou、Berrak Sisman、Rui Liu 和 Haizhou Li。2022。情感语音转换：理论、数据库与 ESD。*Speech Communication*，137：1–18。

A 附录

A.1 案例研究

为说明情感识别中声学与语义冲突所带来的挑战，请看图5所示示例：“干得好——现在我们俩都被困住了。”虽然词汇内容（“干得好”）通常传达积极情感，但说话者的愤怒语调揭示了负面的情感状态。这种不一致给现有模型带来了重大挑战。

以往的方法（例如 Whisper）主要依赖从文本转录中获取的语义理解，因此容易受到强烈语言先验的影响而产生错误分类。它们会将“做得好”解读为本质上积极的表达，从而错误地预测为快乐（见图 5）。类似地，自监督学习模型（如 HuBERT 和 WavLM）虽然能够提取细粒度的声学特征，但其处理的是纠缠表征，其中语音和韵律信息并未清晰分离。因此，即使模型捕捉到了愤怒韵律，也缺乏明确的机制来解决冲突线索，最终只能给出模棱两可的预测。

相比之下，我们提出的**FAS 框架**通过显式建模两条并行路径来解决这一问题。借助 Top-K 选择和基于查询的融合，FAS 能够检测差异并应用置信度感知决策规则。

A.2 补充分析

A.2.1 损失曲线

为了更好地了解我们提出的**FusionAcoustic-Semantic (FAS)** 框架与其他策略相比的学习行为，我们绘制了训练损失曲线。

图6 展示了多种方法的训练损失轨迹，包括我们提出的**FAS**、其消融变体（采用拼接融合和门控融合）以及若干强基线方法。结果表明，尽管 FAS 初始收敛较慢，但最终达到的损失平台显著低于其他方 法。相比之下，拼接融合和门控融合在早期阶段收敛更快，但稳定在较高的损失值上，这表明它们的泛化能力可能较差。这种延迟收敛可能是因为通过 Q_{learn} 融合模块学习对齐关系较为复杂，因此需要更多训练轮数才能稳定。不过，优化完成后，学习到的表征展现出更强的判别能力。

A.2.2 空间可视化

为了定性评估我们的 FAS 框架在学习判别性情感表征方面的效果，我们使用统一流形近似与投影（UMAP）（McInnes 等人，2018）。

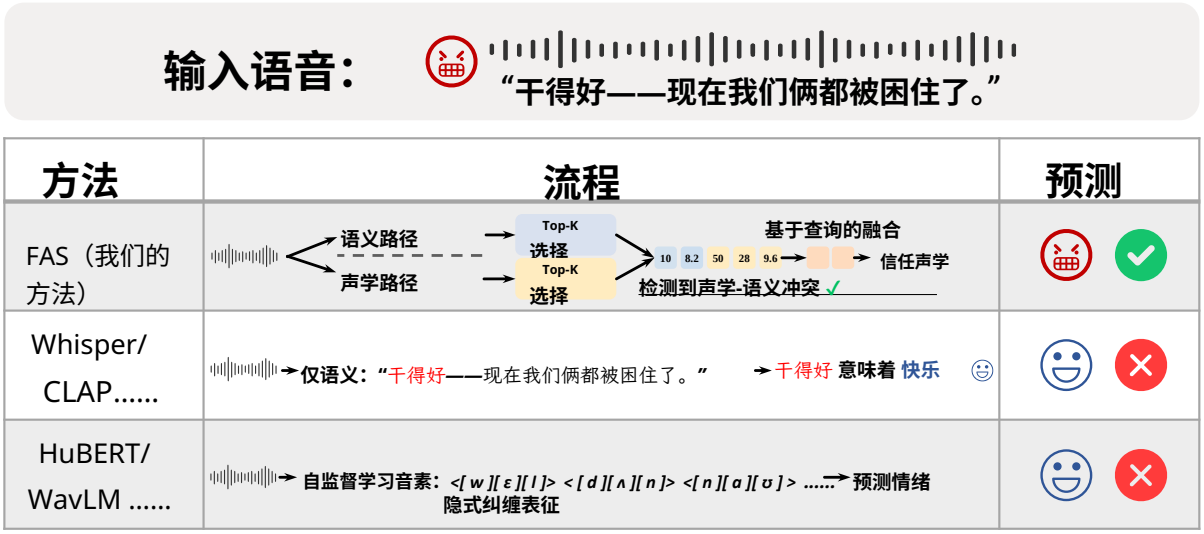


图 5：在一条存在冲突的话语上对语音情感识别方法进行比较。FAS（本文方法）显式地解耦声学路径和语义路径，通过基于查询的融合检测二者之间的冲突，并优先利用声学信号正确预测情绪。

从每个模型的最后一个隐藏层提取特征，并投影到二维空间，以便在 CASE基准上进行比较。

如图 7 所示，HuBERT（Hsu 等人，2021）和 Wav2Vec 2.0（Baevski 等人，2020）等基线模型呈现出高度混杂且难以区分的簇，表明其从其他因素中解耦情感内容的能力有限。相比之下，我们的 FAS 框架生成了清晰分离且结构良好的簇，与真实声学情绪标签呈现出很强的相关性。

我们进一步在 EmoDB（Burkhardt 等人，2005）上评估了我们的框架。EmoDB 是一个数据规模有限且录音质量较低的德语情感语音语料库。如图 8 所示，尽管受这些挑战影响，整体聚类结构不够明显，但我们的 FAS 框架仍保持了相对连贯的情绪簇，优于大多数基线方法。这表明我们的方法在不同语言和数据条件下均具有稳健性，进一步凸显了其实用性。

A.2.3 混淆矩阵分析

为进一步评估 FAS 框架的性能，我们将其在 CASE基准和 RAVDESS 数据集上的混淆矩阵与强基线方法进行比较。如图 9所示，FAS 在主要类别上取得了最高准确率，尤其是在 愤怒、悲伤和中性类别上；与其他方法相比，高唤醒情绪（例如愤怒 与惊讶）之间的混淆显著更少。值得注意的是，没有任何方法能够正确预测恐惧 或厌恶样本，这可能是由于 CASE基准的高难度，以及训练数据稀缺，难以有效解耦声学和语义线索。

在结构良好的 RAVDESS 数据集上的进一步验证（图10）证实了 FAS 的稳健性：其非对角线错误极少，表现优于 WavLM（10e）、VibeVoice（9c）和 Wav2Vec 2.0（10f）等基线方法，尤其擅长区分中性和悲伤等细微情绪。

相比之下，音频语言模型（ALM）的表现则喜忧参半。如图11所示，Qwen2-Audio-Instruct 和 Qwen2.5-Omni 等模型在情绪表达模式一致的数据集（如 Emo-Emilia）上表现尚可，因为大规模预训练使其能够“记住”常见的声学—语义映射。然而，在 CASE 这类高冲突或零样本条件下，它们会严重混淆语义相近的情绪（例如愤怒与厌恶），这揭示了它们从根本上依赖词汇内容而非韵律情感。这种脆弱性凸显了与大语言模型对齐的架构中隐式融合的局限性。

FAS 在结构化场景和模糊场景中均保持稳定的准确率，这凸显了其通过显式建模声学与语义模态之间的交互及潜在冲突，捕捉细粒度情感线索的卓越能力。

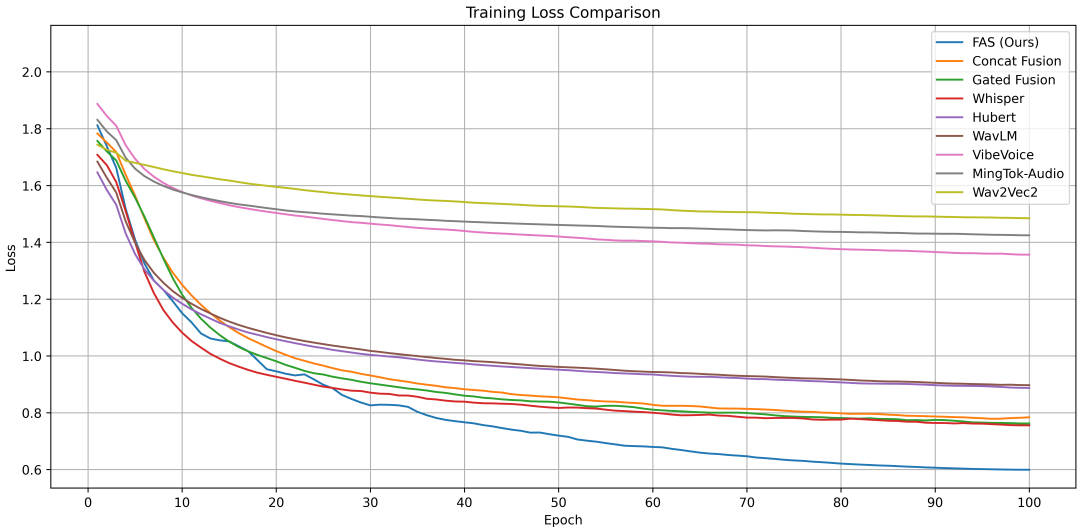


图 6：不同语音情感识别策略的训练损失曲线。FAS 收敛较慢，但最终损失较低。



图 7：CASE 上深层表示的 UMAP 可视化。颜色表示真实声学情绪标签。FAS（左上）形成的情绪簇区分度最高，而 HuBERT（左下）和 Wav2Vec 2.0（右下）则存在明显重叠。

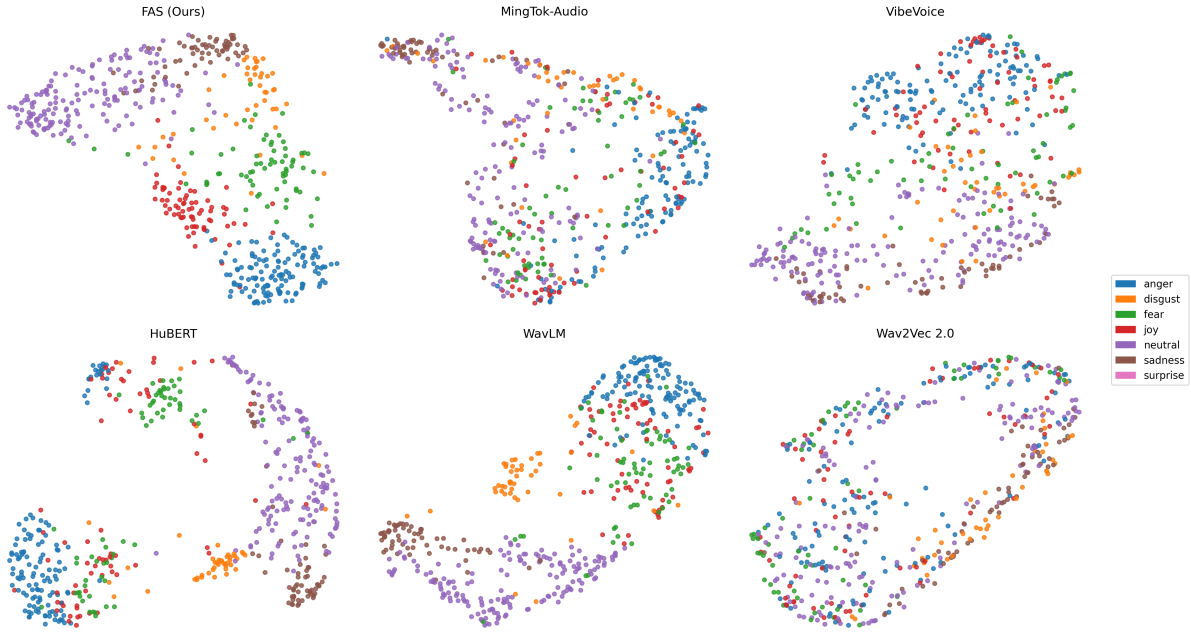


图 8：EmoDB（德语情感语音数据集）上深层表示的 UMAP 可视化。颜色表示真实声学情绪标签。尽管面临跨语言泛化和数据量有限的挑战，我们的 FAS 框架仍能生成相对结构化的聚类，相较于基线模型表现更为稳健，展现了其在多种条件下的鲁棒性。

性。

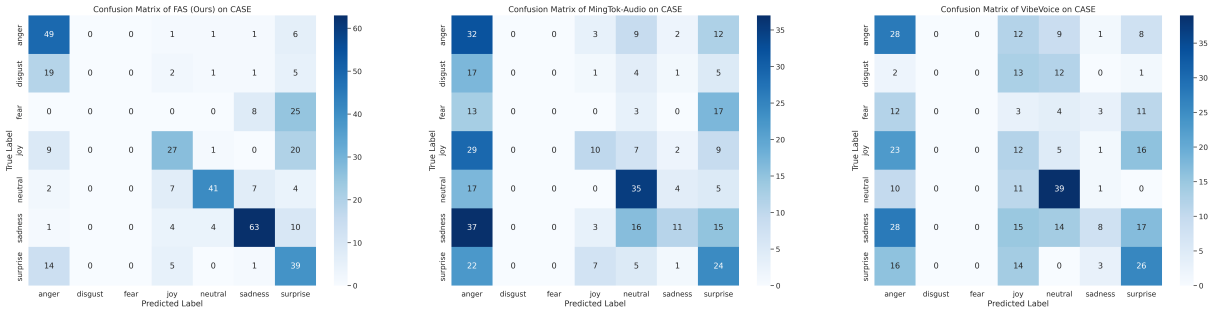
A.3 CASE基准示例

为具体展示我们基准测试中体现的声学—语义冲突，我们在表 7 中列出了一组选取的代表性样本。每个条目都包含原始中文话语、真实声学情感（听到的情绪）、内在的语义情感（言语所暗示的情绪），以及用于说明冲突的简短叙事情境。这些场景在设计上符合心理现实，确保情感失调源于真实的人类经历，而非人为操纵。

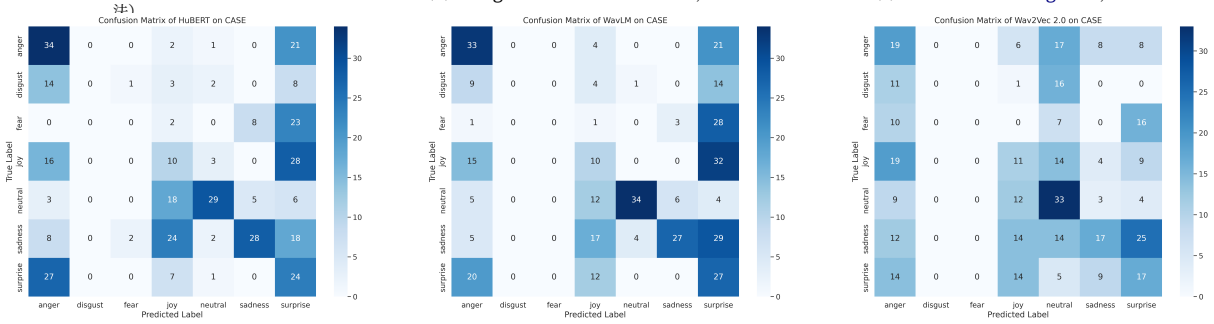
如表所示，CASE涵盖了各种心理上合理的情感冲突——从濒临死亡的士兵带着恐惧说出最后的爱意（样本 007），到做出非理性决定后以中性询问掩饰冷漠的蔑视（样本 011）。这些细腻的场景要求模型将声学韵律与语义内容区分开来，使 CASE 成为评估模型在跨模态冲突下情感鲁棒性的严格测试平台。

ID	话语	真实情绪	语义情绪	情境描述
001	那辆卡车失控了，正朝我们冲过来！	惊讶	fear	一个追求肾上腺素刺激的极端主义者对危险表现出病态兴奋。
002	你们又赢了，恭喜啊。	愤怒	快乐	一个失败者向获胜者表示祝贺，语气几乎掩饰不住的怨恨。
003	他走了，再也不会回来了。	快乐	sad	一个多年来饱受压迫的人，内心感到隐秘的喜悦，因为折磨者离开了。
004	你给我站住！你到底想怎么样！	sad	愤怒	争吵得精疲力竭；愤怒已经转化为心碎与绝望。
005	任务完成，目标已清除。	sad	中性	一名杀手报告称已完成一项任务，但目标是位老朋友。
006	干得好——现在我们俩都被困住了。	愤怒	快乐	说话者讽刺地责怪自己的同伴，正是对方鲁莽的行为导致了共同的困境，掩饰带着讽刺性赞扬的挫败感。
007	妈妈，爸爸.....我爱你们。	fear	快乐	一名士兵在执行自杀式任务前，录下最后留言给他的父母；他的尽管说着充满爱意的话语。
008	不可能——他明明已经死了！	fear	惊讶	在恐怖情境中，主人公目睹一个本该被杀死的反派起身再次流露出内心深处本能的恐惧最初因震惊而倒吸一口气。
009	紧急出口被堵住了。	fear	中性	火灾期间，有人宣布唯一的逃生路线已被封锁——他的语气措辞平静，却明显透着恐慌与惧怕。
010	什么？你居然端上了这个?	hate	惊讶	一位挑剔的美食爱好者对一道令人作呕的“美食”菜肴立刻厌恶，他们的震惊很快被强烈的憎恶所取代。
011	那么，这就是你最后的决定?	hate	中性	听到一个极其不合理的选择后，说话者冷冷地说道，冷淡的确认，传达出无声的蔑视和无奈。
012	顺便说一句，大楼着火了。我们最好赶紧离开。	中性	紧张	一个带着冷面、英式幽默的角色，以极度坚忍的态度传达紧急且危及生命的消息，语气轻松、像在闲聊一样。
013	哦，是为我准备的惊喜派对？你们真是太客气了。	愤怒	中性	一个讨厌惊讶的内向者正努力保持礼貌，但他们的声音却充满了恼怒和愤怒。
014	我恨你！我再也不想见到你了！	sad	愤怒	分手时说着仇恨的话，但内心深处的情感是心碎与悲伤。
015	我喜欢。又一份电子表格。	sad	快乐	一名员工讽刺地评论自己被分派了更多繁琐工作，声音中满是忧郁。
016	你输了这场游戏。结束了。	快乐	中性	一位游戏节目主持人俏皮而欢快地向一名参赛者宣布坏消息：参赛者。
017	获胜者是.....不是你。	快乐	中性	一位游戏节目主持人俏皮地、欢快地向一位参赛者宣布坏消息。参赛者。
018	碗碟不用担心，我来洗它们。又来了。	愤怒	中性	典型的被动攻击型室友情境。话听起来似乎很热心，但语气中却满是愤怒和怨恨。
019	我听说你晋升了。我真是，真为你感到高兴。	sad	兴奋	正在祝贺一位如愿获得晋升的同事。他们试图表示支持，但声音中充满了他们自己的失望。
020	你为这么点小事气成这样游戏，真是可爱极了。	快乐	愤怒	一位朋友开玩笑地戏弄、嘲笑另一位正变得越来越烦躁的朋友玩电子游戏时。

表 7：我们提出的 CASE基准中的代表性声学-语义冲突样本。如需查看完整音频示例和更多元数据，请参阅开源代码库中公开发布的数据集文件。

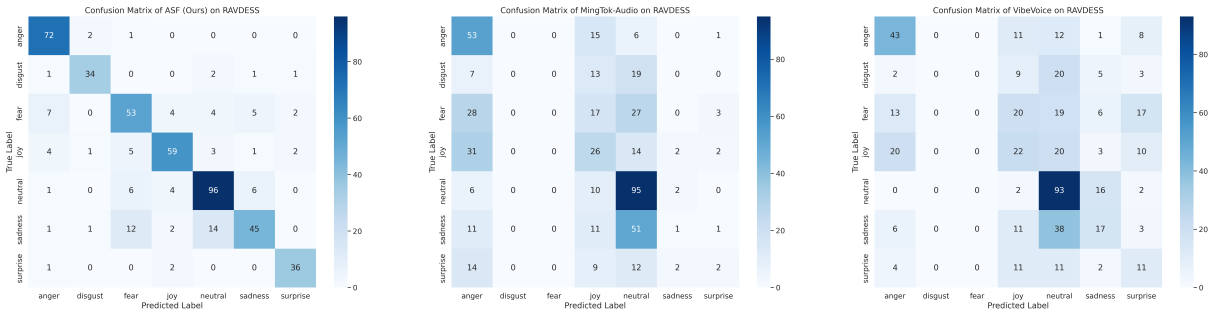


(a) FAS（我们的方法）　(b) MingTok-Audio（Yan 等人，2025年）　(c) VibeVoice（Peng 等人，2025年）

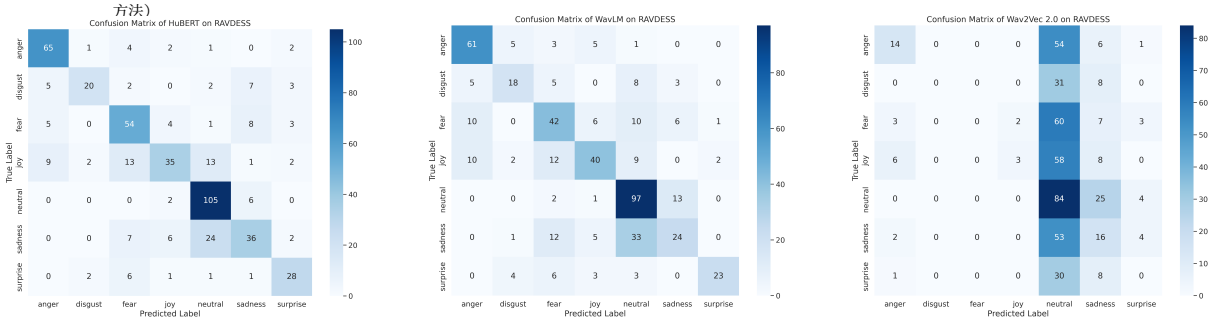


(d) HuBERT（Hsu 等人，2021）　(e) WavLM（Chen 等人，2022）　(f) Wav2Vec 2.0（Baevski 等人，2020）

图9：CASE数据集上的混淆矩阵。我们的 FAS 框架（9a）在主要类别（愤怒、悲伤、中性）上取得了最高准确率，并且高唤醒情绪之间的混淆极少。

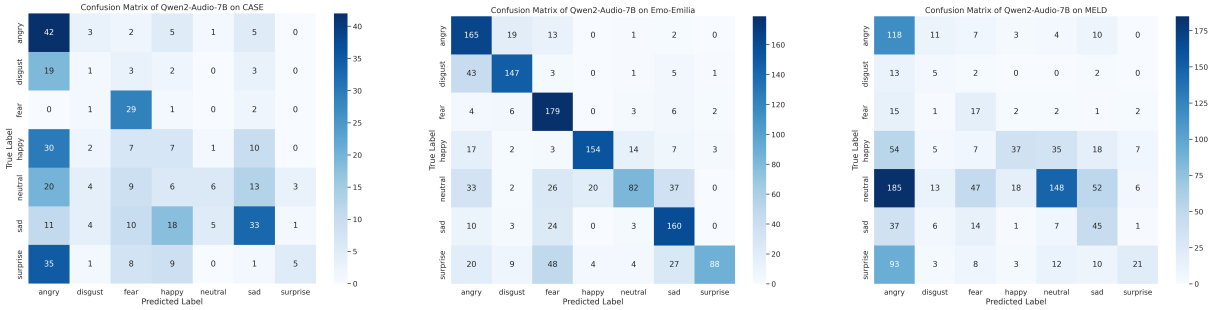


（a）FAS（我们的方法）　（b）MingTok-Audio　（c）VibeVoice



（d）HuBERT　（e）WavLM　（f）Wav2Vec 2.0

图 10：在 RAVDESS 数据集上评估的混淆矩阵。FAS 框架（10a）在所有情绪类别上均表现出色，非对角线错误极少。相比之下，其他模型（包括 MingTok-Audio（10b）、VibeVoice（10c）、HuBERT（10d）、WavLM（10e）和 Wav2Vec 2.0（10f）在相似情绪之间表现出不同程度的混淆。



(a) Qwen2-Audio 在 CASE 上 (b) Qwen2-Audio 在 Emo-Emilia 上 (c) Qwen2-Audio 在 MELD 上

(d) Qwen2.5-Omni 在 CASE 上 (e) Qwen2.5-Omni 在 Emo-Emilia 上 (f) Qwen2.5-Omni 在 MELD 上

图11：混　淆矩阵：Qwen2-Audio-Instrcut 和 Qwen2.5-Omni 在 CASE、Emo-Emilia 和 MEL 上的表现

D.