arXiv:2602.04904v1 [cs.LG] 3 Feb 2026

DCER：通过双阶段压缩与基于能量的重建实现鲁棒多模态融合

Yiwen Wang Jiahao Qin*

*通讯作者：Jiahao.Qin19@gmail.com

摘要

多模态融合面临两项鲁棒性挑战：噪声输入会降低表征质量，而模态缺失则会导致预测失败。我们提出**DCER**，这是一个通过双阶段压缩和基于能量的重建来应对这两项挑战的统一框架。压缩阶段分两个层面进行：模态内频率变换（对音频采用小波变换，对视频采用离散余弦变换）在去除噪声的同时保留任务相关模式；跨模态瓶颈令牌则促使模型真正进行整合，而非走模态特定捷径。对于缺失模态，该框架通过对学习得到的能量函数执行梯度下降来恢复表征，并利用最终能量实现内在不确定性量化（与预测误差 $\rho > 0.72$ 相关）。在 CMU-MOSI、CMU-MOSEI 和 CH-SIMS 上进行的实验表明，该方法在所有基准测试中均达到最先进的性能，并呈现 U 形鲁棒性模式：在模态完整和模态高度缺失的情况下，多模态融合都更具优势。代码将在 GitHub 上发布。

关键词：多模态学习、多模态融合、缺失模态、情感分析

1 引言

多模态学习已成为理解复杂人类行为的一种基础范式。在情感分析中，面部表情揭示情绪状态，语音韵律传达强度，而口语则提供语义语境。有效整合这些异质信号，比单独使用任何一种模态都能获得更准确、更鲁棒的预测 [1, 2]。

然而，多模态融合面临两项根本挑战（图1a）。首先，标准融合方法——拼接、交叉注意力或张量积——使模型能够为每种模态学习独立的处理路径。尽管所有模态均可用时这类方法表现出色，但它们会产生脆弱的表征，在受到噪声干扰或出现缺失数据时便会失效。其次，原始多模态输入包含大量冗余：音频信号在不同时间尺度上编码情绪，但大多数频率分量都是噪声；视频帧在空间上信息丰富，但情绪线索集中在特定面部区域。若缺乏有原则的压缩，模型便难以识别任务相关模式。

这些挑战有一个共同的解决方案：压缩。通过迫使信息经过容量受限的通道，我们可以同时消除冗余并防止捷径学习。信息瓶颈原理 [3] 为此提供了理论依据，不过我们采用的是架构层面的方法，而非显式的互信息优化。

我们认为，有效的多模态融合需要在不同粒度上进行双阶段压缩。第一阶段在各模态内部进行：频域变换（音频采用小波，视频采用离散余弦变换）将任务相关模式集中到紧致子空间中，同时丢弃噪声。第二阶段在模态之间进行：可学习查询令牌同时关注所有模态，迫使信息经过固定容量瓶颈，从而防止形成彼此独立的处理路径。

另一个相关但独立的问题是评估方法：用于缺失模态评估的标准零值掩码协议可能会高估模型的鲁棒性，因为用零替换缺失特征会让模型将零信号当作捷径来利用。我们的初步分析显示，与噪声掩蔽相比，这种做法会使相关性指标虚高 15–51%，因此有必要在两种协议下进行验证。

我们将这些思路具体实现为 **DCER**(**D**双阶段压缩**C**与**E**基于能量的**R**重建），其概述见图1b，详细内容见图 2。我们的贡献包括：（1）提出一种双阶段压缩框架，将频域预处理与基于注意力的瓶颈相结合，以实现鲁棒的表征学习；（2）提出一种用于缺失模态的基于能量的重建算法，并对内在不确定性进行量化；（3）在 CMU-MOSI、CMU-MOSEI 和 CH-SIMS 上取得最先进的结果，在存在缺失模态的情况下仍表现出很强的鲁棒性。

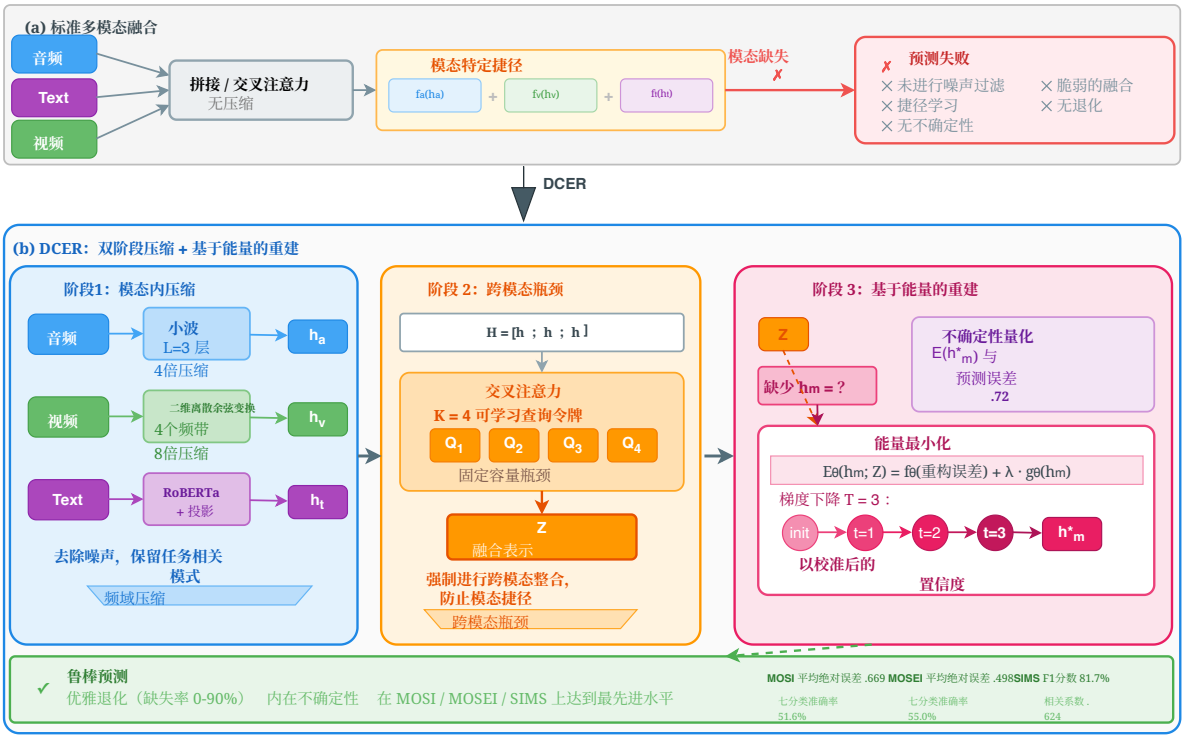

图1：**动机与概述。（a）**标准多模态融合缺乏噪声过滤能力，并会学习模态特定捷径，导致模态缺失时预测失败。**（b）**DCER 通过三个阶段解决这些问题：模态内频率压缩去除噪声，跨模态瓶颈令牌促成真正的整合，基于能量的重建则在校准不确定性的支持下恢复缺失模态（$\rho > 0.72$）。

2 相关工作

多模态情感分析（MSA）。早期方法采用拼接 [4] 或张量分解[1] 进行融合。多模态变换器（MulT） [2] 引入了有方向性的跨模态注意力，使每种模态都能关注其他模态。MAG-BERT [5] 将多模态适配门与预训练语言模型相结合。Self-MM [6] 利用自监督辅助任务学习模态特定表示。近期研究探索了提示学习（MPLMM [7]）和双层特征重建（EMT [8]）。受脑计算启发的方法 [9] 提出了预定义的多模态联合表示空间，利用神经可塑性原理增强跨模态学习。尽管这些方法在完整数据上的表现出色，但缺乏有原则的机制来防止模态特定捷径。

缺失模态处理。相关方法分为三类：（1）生成式方法使用变分自编码器（VAEs）或生成对抗网络（GANs）重建缺失特征 [10], ，但难以处理高维生成。（2）模态不变方法学习共享表示 [11]„ 但会牺牲模态特定信息。（3）鲁棒融合方法设计能够优雅退化的架构 [12]。DCER 将重建（通过基于能量的推断）与鲁棒融合（通过双重瓶颈）相结合，从而兼具强大的完整数据性能和优雅退化能力。

瓶颈架构。 信息瓶颈原理 [3]通过互信息优化，对压缩与预测之间的权衡进行形式化。深度信息瓶颈方法 [13] 会显式优化变分界，而架构方法则通过容量约束实现类似效果。Perceiver [14] 采用带可学习查询的跨注意力，将其作为通用感知的潜在瓶颈。DCER 沿用了这种架构瓶颈理念，但有两点不同：（1）我们在跨模态融合前加入模态特定频率压缩，以利用已知的信号结构；（2）我们加入基于能量的重建，以处理缺失的

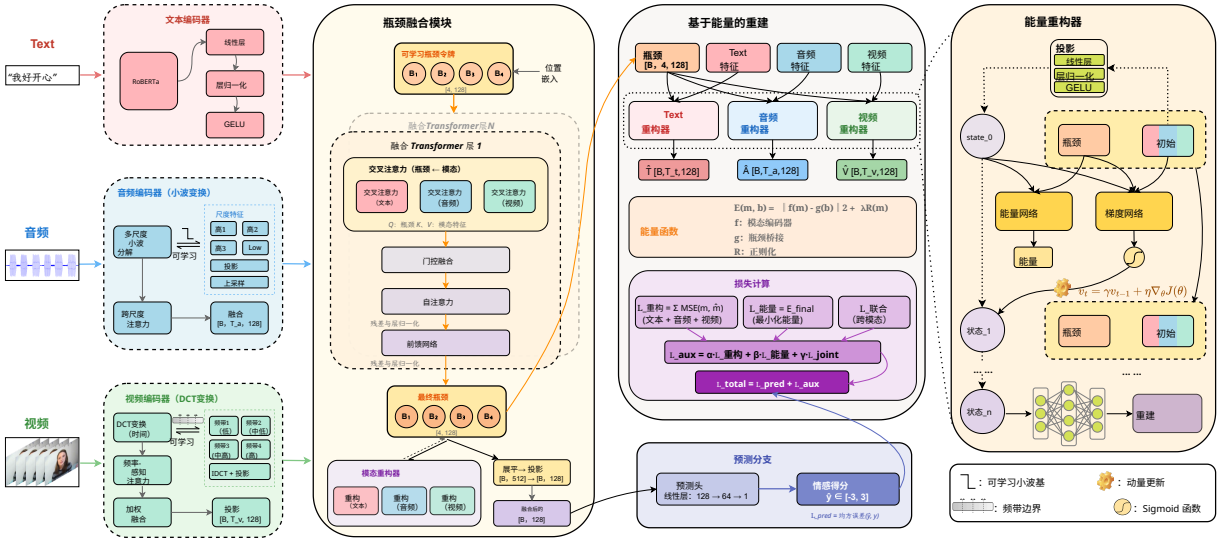

图 2：**DCER 架构**。**左侧**：模态特定编码器应用频率变换（音频采用小波变换，视频采用离散余弦变换）进行模态内压缩。**中间**：可学习瓶颈令牌通过交叉注意力关注所有模态，实现跨模态瓶颈。**右侧**：基于能量的重建通过不确定性量化实现缺失模态处理。

模态。我们的消融实验（表 3）验证了频率压缩能显著提升性能。

频域学习。 傅里叶变换和小波变换已应用于时间序列预测 [15], 、图像分类 [16], 和音频处理 [17]。FNet [18] 以傅里叶变换取代注意力，以提高效率。近期研究将时间序列成像与 Transformer 架构相结合，用于稳健的信号分析，并证明了其在癫痫发作预测 [19] 以及通过脑电图-功能磁共振成像数据中的跨感知方法进行内言语识别 [20]方面的有效性。我们提出了一种新的动机：频率变换通过将任务相关模式集中到特定频带中实现<左右压缩 ，因而可以选择性地保留这些模式，同时舍弃噪声主导成分。

基于能量的模型。 基于能量的模型 [21] 通过能量函数隐式定义分布。近期研究将基于能量的模型应用于结构化预测、 [22], 生成 [23], 和多模态学习 [24]。我们专门将基于能量的推断用于缺失模态重建，从而提供以往多模态基于能量的模型工作中无法实现的不确定性量化。

3 方法

3.1 问题形式化

给定多模态输入 $\mathbf{X} = \{\mathbf{x}_a, \mathbf{x}_v, \mathbf{x}_t\}$，其中包括音频 $\mathbf{x}_a \in \mathbb{R}^{T_a \times D_a}$、视频 $\mathbf{x}_v \in \mathbb{R}^{T_v \times D_v}$和文本 $\mathbf{x}_t \in \mathbb{R}^{T_t \times D_t}$，我们的目标是预测情感 $y \in [-3, 3]$。在推断过程中，由于传感器故障、网络问题或隐私约束，某些模态可能缺失。

3.2 概述

DCER 处理 mu 多模态输入经由双阶段瓶颈，生成情感预测 $\hat{y} \in \mathbb{R}$:

$$\hat{y} = f_{\text{pred}}\Big(\underbrace{\text{CrossAttn}\big(\mathbf{Q}, \underbrace{[\mathbf{h}_a; \mathbf{h}_v; \mathbf{h}_t]}_{\text{Stage 1: Within-Modality}}\big)}_{\text{Stage 2: Cross-Modality}}\Big) \tag{1}$$

其中， f_{pred} ： $\mathbb{R}^{K\times D} \to \mathbb{R}$ 是预测头，以多层感知机 (MLP) 实现； $\mathbf{h}_m = \mathcal{F}_m(\mathbf{x}_m) \in \mathbb{R}^{T_m\times D}$ 是模态特定编码，而 $\mathbf{Q} \in \mathbb{R}^{K\times D}$ 是隐藏维度为 D 的可学习查询令牌。

3.3 阶段1：模态内压缩

有效的压缩应在保留任务相关信息的同时去除噪声。情感信号具有典型的频率模式：语音韵律呈现于多个时间尺度，而面部表情则集中在特定空间频带中。传感器噪声通常均匀分布于各个频率。这启发我们采用频域压缩，有选择地保留与情感相关的频带。

音频。 音频情感涉及多个时间尺度——语音细节（20-100ms）、音节级韵律（100-500ms）和话语节奏（500ms-2s）。我们采用 $L = 3$ 层离散小波变换（DWT）： $\mathcal{W}(\mathbf{x}_a) = \{c_L, d_L, \ldots, d_1\}$，其中 c_L 捕获粗略近似信息， d_ℓ 捕获尺度 ℓ 上的细节。可学习小波基（由 Daubechies-4 初始化）能够针对具体任务进行适配： $\mathbf{h}_a = \text{Proj}(\text{CrossScaleAttn}(\mathcal{W}(\mathbf{x}_a)))$.

视频。 视觉情感依赖于面部构型——其本质上属于低频空间模式。我们通过二维离散余弦变换（DCT）将视频特征分解为四个频带，并采用具有可学习频带边界的频率感知注意力： $\mathbf{h}_v = \text{Proj}(\text{FreqAttn}(\mathcal{D}(\mathbf{x}_v)))$.

文本。 文本本身已经是一种压缩符号表示。我们使用经过预训练的 RoBERTa，并采用任务特定的投影： $\mathbf{h}_t = \text{Proj}(\text{RoBERTa}(\mathbf{x}_t))$。

3.4 阶段2：跨模态瓶颈

标准融合允许模型学习独立的路径 $\hat{y} = f_a(\mathbf{h}_a) + f_v(\mathbf{h}_v) + f_t(\mathbf{h}_t)$，在完整数据上取得出色性能，但在存在缺失模态时表现不佳。为避免这一问题，我们引入 K 可学习查询令牌 $\mathbf{Q} \in \mathbb{R}^{K\times D}$，将其作为固定容量瓶颈。令 $\mathbf{H} = [\text{<style id='}\mathbf{h}_a; \mathbf{h}_v; \mathbf{h}_t]$。瓶颈的计算如下：

$$\mathbf{Z} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{H}^\top}{\sqrt{D}}\right)\mathbf{H} \in \mathbb{R}^{K\times D} \tag{2}$$

由于 $K \ll T_a + T_v + T_t$，信息流受到限制，且每个查询都必须关注所有模态——预测仅依赖于 $\mathbf{Z}$，从而防止出现模态特定捷径。我们堆叠 $L_f = 6$ 个融合Transformer层，每层依次对 $\mathbf{H}$ 应用交叉注意力、自注意力以及带残差连接的前馈网络（FFN）。

表 1：**CMU-MOSI 和 CMU-MOSEI 上的主要结果（完整模态）**。两个数据集均采用 7 类情感标签（$[-3, 3]$），并使用相同的评估指标。最佳结果以**粗体**标示，次佳结果加下划线。 †使用官方开源代码和原始超参数复现的结果。

模型	CMU-MOSI						CMU-MOSEI				
	平均绝对误差↓	相关系数↑	七分类准确率	三分类准确率	二分类准确率	F1分数	平均绝对误差↓	相关系数↑	七分类准确率	二分类准确率	F1分数
MulT†19	0.904	0.680	37.7	57.9	79.3	79.2	0.549	0.748	53.1	80.4	80.1
Self-MM†21	0.788	0.742	43.3	54.5	80.8	80.1	0.588	0.738	52.3	82.5	82.2
MPLMM†24	0.880	0.677	39.5	62.8	79.8	79.9	0.547	0.749	53.4	80.1	79.8
LNLN†24	0.759	0.760	43.4	57.2	80.6	80.2	0.537	0.758	52.8	80.7	80.4
EMT†23	0.717	0.796	47.8	55.4	82.5	82.4	0.532	0.764	53.6	83.7	83.9
MMA†25	0.710	0.769	45.5	63.3	83.0	83.2	0.546	0.735	53.3	81.8	81.9
MSAmba†25	0.724	0.782	48.1	63.4	84.4	**84.8**	0.514	0.771	52.5	84.0	83.9
MTFN†25	0.736	0.767	45.1	**64.9**	83.8	84.0	0.542	0.736	52.7	83.2	82.9
DCER（我们的方法）	**0.669**	**0.823**	**51.6**	64.7	**85.4**	83.9	**0.498**	**0.806**	**55.0**	**85.7**	**84.9**

算法 1 基于能量的重建

输入： $\mathbf{h}_{\text{obs}}$、$\mathbf{Z}$、迭代次数 T、步长 η、动量 ρ
$\mathbf{h}_m^{(0)} \leftarrow \mu_\theta(\mathbf{Z}, \mathbf{h}_{\text{obs}}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2$
I)对于 $t = 1$ 到 T **执行** $\mathbf{h}_m^{(t)} \leftarrow \mathbf{h}$
$_m^{(t-1)} - (\rho \mathbf{v}^{(t-1)} + \eta \nabla E_\theta)$**结束循环 返回：**
$\mathbf{h}_m^{(T)}$, $E_\theta(\mathbf{h}_m^{(T)})$

3.5 基于能量的重建

当模态 m 缺失时，我们根据已观测模态和瓶颈$\mathbf{Z}$重建 $\mathbf{h}_m$ 。我们定义一个参数化能量函数：

$$E_\theta(\mathbf{h}_m; \mathbf{Z}) = f_\theta(\mathbf{h}_m - \text{CrossAttn}(\mathbf{h}_m, \mathbf{Z})) + \lambda_E \cdot g_\theta(\mathbf{h}_m) \tag{3}$$

其中 f_θ, g_θ 是学习得到的多层感知机，输出标量能量，且 $\lambda_E = 0.1$。这种神经参数化能够捕捉超出欧几里得距离的复杂重建模式。

重建通过带动量的梯度下降进行（算法 1）。一个学习得到的网络 μ_θ从瓶颈提供初始化。在模态完整时，初始估计准确，且$T = 0$ 次迭代即已足够。在缺失模态下，迭代优化利用跨模态一致性——$T = 3$ 显著提升性能（表 6）。最终能量提供不确定性量化：低能量表示置信的重建，高能量表明预测不可靠。我们观察到能量与预测误差之间存在强相关（$\rho > 0.72$）。

3.6 训练目标

总损失由四项组成：

$$\mathcal{L}_{\text{pred}} = \frac{1}{B} \sum_{i=1}^{B} (y_i - \hat{y}_i)^2 \tag{4}$$

$$\mathcal{L}_{\text{recon}} = \mathbb{E}_m \left[\|\mathbf{h}_m - \hat{\mathbf{h}}_m\|^2 \right] \tag{5}$$

$$\mathcal{L}_{\text{energy}} = \mathbb{E} \left[E_\theta(\mathbf{h}_m^{(T)}) \right] \tag{6}$$

$$\mathcal{L}_{\text{joint}} = \|\mathbf{Z}_{\text{full}} - \mathbf{Z}_{\text{recon}}\|^2 \tag{7}$$

ˆ其中 $\mathbf{h}_m$ 是重建表示，$\mathbf{Z}_{\text{recon}}$ 是根据重建模态计算得到的瓶颈。总损失为：

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \alpha \mathcal{L}_{\text{recon}} + \beta \mathcal{L}_{\text{energy}} + \gamma \mathcal{L}_{\text{joint}} \tag{8}$$

其中 $\alpha = 0.1$、 $\beta = 0.01$、 $\gamma = 0.05$。

4 实验

4.1 实验设置

数据集。 我们在三个多模态情感分析基准上进行评估：**CMU-MOSI** [4] 包含来自 93 条 YouTube 电影评论的 2,199 个视频片段（英语，情感范围为 $[-3, 3]$），并采用 74 维 COVAREP 声学特征和 35 维 FACET 视觉特征。**CMU-MOSEI** [25] 包含来自 1,000 位 YouTube 说话人的 23,453 个片段（英语，情感范围相同），采用与 CMU-MOSI 相同的声学和视觉特征。

CH-SIMS [26] 包含来自中国电视节目的 2,281 个片段（中文，情感标签为 $[-1, 1]$），并采用 33 维 LibROSA 声学特征和 709 维 OpenFace 2.0 视觉特征。所有数据集的文本均使用预训练 RoBERTa 嵌入（768 维）。

评估指标。 我们报告五项评估指标：MAE↓（平均绝对误差）、Corr↑ （皮尔逊相关系数）、Acc-2↑（二分类准确率，正向/负向）、Acc-7↑、Acc-3↑ 或 Acc-5↑ （多分类准确率），以及 F1↑ （二分类F1分数）。CMU-MOSI 使用七分类和三分类准确率（Acc-7、Acc-3），CMU-MOSEI 使用 Acc-7，而 CH-SIMS 使用五分类和三分类准确率（Acc-5、Acc-3）。

基线方法。 我们与八种最先进的方法进行比较：MulT [2] （跨模态Transformer）、Self-MM [6] （自监督多模态学习）、MPLMM [7] （面向缺失模态的多模态提示学习）、LNLN [11]（语言主导的抗噪学习网络）、EMT [8] （采用双层特征恢复的高效多模态Transformer）、MMA [27]（多模态适配器混合）、MSAmba [28]（用于多模态情感分析的状态空间模型）以及 MTFN [29] （采用多层特征融合的多任务融合网络）。

实现细节。 6层融合Transformer，隐藏维度为128，含4个注意力头。 $K = 4$ 个瓶颈令牌。采用AdamW优化器，学习率为$=10^{-5}$，批量大小为32，训练40个周期。完整数据使用 $T = 0$ 次能量迭代（在缺失模态场景下增加至 $T = 3$ 次）。所有实验均取5个随机种子的平均值。

4.2 主要结果

表 1 和 2 表明，DCER 在三个基准数据集上均达到了最先进水平：

英语数据集（表 1）。在 CMU-MOSI 上，DCER 取得了最佳的平均绝对误差（0.669，相比 MMA 低 5.8%）、相关系数（0.823，+3.4% 相比 EMT）、七分类准确率（51.6%， +7.3% 相比 MSAmba）以及二分类准确率（85.4%， +1.2% 相比 MSAmba）。在 F1分数方面，MSAmba 达到 84.8%，而 DCER 为 83.9%，这是 DCER 具备更优回归性能所作出的权衡。在 CMU-MOSEI 上，DCER 在所有评估指标上均处于领先地位：平均绝对误差为 0.498（相比 MSAmba 低 3.1%）、相关系数为 0.806（+4.5%）、七分类准确率为 55.0%（+2.6%）、二分类准确率为 85.7%（+2.0%）以及 F1分数为 84.9%（+1.2%）。

中文数据集（表2）。在 CH-SIMS 上，DCER 的相关系数（0.624）、Acc-5（55.5%，+22.8% 对比 MTFN）、二分类准确率（81.2%）和 F1分数（81.7%）均为最佳，同时平均绝对误差（0.405，排名第 2）和三分类准确率（64.7%，排名第 2）也具有竞争力。Acc-5 和三分类准确率的大幅提升，展现了 DCER 在细粒度情感分类方面的优势。

在不同数据集、语言和评估指标上持续取得的提升，验证了双阶段压缩对多模态融合的有效性。模态鲁棒性分析（表5）表明，在模态缺失时，多模态融合仍能保持性能，其中 T+A+V 在高缺失率下取得了最佳七分类准确率。

表2：**CH-SIMS 上的主要结果（模态完整）**。最佳结果以**粗体**标出，次佳结果以下划线标出。 _____ †复现结果。

模型	平均绝对误差↓	相关系数↑	Acc-5	三分类准确率	二分类准确率	F1分数
MulT† 19	0.434	0.599	36.3	57.9	72.6	72.2
Self-MM† 21	0.572	0.542	35.7	54.5	70.3	69.9
MPLMM† 24	0.425	0.558	37.0	62.8	71.7	71.5
LNLN† 24	0.448	0.524	36.6	57.2	71.0	71.8
EMT† 23	**0.402**	0.620	42.6	55.4	79.2	79.3
MMA† 25	0.417	<u>0.623</u>	44.1	63.3	<u>79.6</u>	<u>79.7</u>
MSAmba† 25	0.434	0.621	44.0	63.4	78.9	78.5
MTFN† 25	0.425	0.570	<u>45.2</u>	**64.9**	79.1	78.0
DCER（我们的方法）	<u>0.405</u>	**0.624**	**55.5**	<u>64.7</u>	**81.2**	**81.7**

4.3 消融研究

表3揭示了各个组件的若干见解：

阶段1（频率压缩）： 较简单的配置（S=1 或 B=2）与复杂配置表现相当，这表明情感特征集中在较粗粒度的时空尺度上。表4 将其与无频率基线（仅使用线性投影）进行了比较。频率压缩提升了完整数据（+3.7% 平均绝对误差）和极端缺失（+12.3% 平均绝对误差）下的表现，而阶段2（瓶颈）在所有条件下均带来稳定收益。

阶段2（跨模态瓶颈）： $K = 4$ 个标记在压缩与表达能力之间实现了最佳平衡。标记数量较少（$K = 2$）会导致信息损失；标记数量较多（$K = 8$）则会降低压缩收益，导致过拟合。

表3：**在 CMU-MOSI（mr=0.0）上的消融研究**。我们系统地改变各个组件，同时将其余组件固定为默认值。粗体表示每组中的最佳结果。

配置	**MAE↓**	**Acc-7↑**	**Acc-2↑**	**F1↑**
完整 DCER（默认）	0.669	51.6	85.4	83.9
Stage 1: Within-Modality Frequency Compression				
无频率项（Perceiver 风格）	0.695	50.4	83.7	82.4
音频: $S = 1$ 小波尺度	0.653	**51.9**	84.3	81.8
音频: $S = 3$ 尺度（默认）	**0.669**	51.6	**85.4**	**83.9**
视频: $B = 2$ DCT频带	0.661	51.8	84.8	82.5
视频: $B = 4$ 频带（默认）	**0.669**	51.6	85.4	**83.9**
Stage 2: Cross-Modality Bottleneck				
$K = 2$ 瓶颈标记	0.674	49.7	83.1	80.8
$K = 4$ 标记（默认）	**0.669**	**51.6**	**85.4**	**83.9**
$K = 8$ 瓶颈令牌	0.684	49.9	84.0	81.7

表4：**频率压缩消融实验（CMU-MOSI）**。“无频率压缩”使用线性投影而非小波/离散余弦变换。完整 DCER 使用 $T = 0$。最佳结果以**粗体**显示。

指标 配置。		**mr=0**	**mr=0.1**	**mr=0.3**	**mr=0.5**	**mr=0.7**	**mr=0.9**
MAE↓	无频率压缩	0.695	0.722	0.906	**1.031**	**1.271**	1.606
	完整 DCER	**0.669**	**0.718**	**0.905**	1.115	1.361	**1.408**
准确率-7↑	无频率信息	50.4	44.8	40.2	**37.6**	**27.7**	17.6
	完整 DCER	**51.6**	**45.9**	**42.9**	34.7	26.8	**21.9**
准确率-2↑	无频率信息	83.7	83.4	77.0	**73.5**	63.0	52.0
	完整 DCER	**85.4**	**83.6**	**77.1**	72.7	**63.1**	**54.7**
F1↑	无频率	82.4	81.6	74.4	**71.2**	62.3	51.3
	完整 DCER	**83.9**	**81.7**	**75.1**	68.9	**62.4**	**57.7**

能量重构（表6）： 在完整数据（缺失率=0）上，$T = 0$表现最佳，因为瓶颈已捕获了足够的跨模态交互。更多迭代会降低性能，可能是因为表示被过度平滑。然而，在模态缺失（缺失率>0）的情况下，能量迭代则能发挥作用：$T = 3$ 在中等缺失率（缺失率=0.3–0.7）下，通过利用重构引导恢复缺失信息，降低了平均绝对误差。

模态贡献（表5）： 在完整数据（缺失率=0）上，仅文本的七分类准确率达到 51.0%，但平均绝对误差（0.693）表现欠佳。加入音频/视频可改善回归性能：T+A 将平均绝对误差降低 8.2%（0.636），T+V 降低 4.5%（0.662）。全融合（T+A+V）取得了最佳七分类准确率（51.6%）和二分类准确率（85.4%）。在模态缺失时，多模态优势呈 U 形：TAV 在缺失率=0 和缺失率≥0.7 时表现出色，而仅文本在中等缺失率下更稳定——这可能是因为存在部分信息时，重建噪声会干扰文本主导的预测。

表5：**模态鲁棒性（CMU-MOSI）**。多模态融合呈现 U 形优势：TAV 在完整数据（51.6%）和高缺失率（mr=0.9：49.3%，而 T 为 46.8%）下表现出色，而仅文本在中等缺失率下更稳定。

模态	缺失率	平均绝对误差↓	相关系数↑	七分类准确率	二分类准确率	F1分数
T	0	0.693	0.819	**51.0**	84.3	84.2
	0.1	0.671	0.827	48.1	83.1	83.1
	0.3	**0.644**	**0.850**	49.4	**85.7**	**85.7**
	0.5	0.645	0.842	49.7	84.7	84.6
	0.7	0.680	0.829	48.1	83.4	83.3
	0.9	0.667	0.833	46.8	83.8	83.8
T+A	0	**0.636**	**0.845**	49.6	84.3	84.0
	0.1	0.722	0.781	47.5	82.8	82.8
	0.3	0.651	0.837	48.0	85.6	85.5
	0.5	0.645	0.835	49.3	83.8	83.8
	0.7	0.658	0.827	48.0	83.1	83.0
	0.9	0.669	0.830	46.2	84.4	84.4
T+V	0	0.662	0.825	**51.0**	83.8	83.8
	0.1	0.669	0.828	49.7	82.9	82.9
	0.3	0.660	0.840	47.4	84.0	83.9
	0.5	**0.651**	0.839	48.5	83.7	83.6
	0.7	0.689	0.826	46.5	82.5	82.5
	0.9	0.668	0.838	46.8	**85.3**	**85.3**
TAV	0	0.669	0.823	**51.6**	**85.4**	83.9
	0.1	0.665	0.832	47.4	84.0	83.9
	0.3	0.681	0.823	46.7	83.1	83.0
	0.5	0.668	0.833	48.1	83.4	83.3
	0.7	0.679	0.831	48.8	83.5	83.5
	0.9	0.673	0.831	**49.3**	83.8	83.8

缺失模态下的能量迭代（表6）： 在完整数据（mr=0）下，$T = 0$ 取得最佳结果，增加迭代次数并无益处。在中等缺失率（mr=0.3–0.7）下，$T = 3$取得最佳平均绝对误差，表明迭代优化有助于恢复缺失信息。在极高缺失率（mr=0.9）下，$T = 0$ 表现最佳，因为信息不足，无法进行有效优化。

如表7所示，DCER 在两种协议下均保持出色性能，展现出真正的鲁棒性，而非针对特定协议进行优化。噪声掩蔽下性能仅有轻微退化（平均绝对误差增加 6.6%），这证实 DCER 的性能源于有原则的压缩与重建，而非利用零信号捷径。

4.4 不确定性量化

基于能量的重建的一项独特优势是能够量化内在不确定性。表 8显示，能量与预测误差高度相关（$\rho = 0.72$–0.81），无需额外校准即可可靠地估计不确定性。高能量样本的准确率较低（29–38%），而低能量样本的准确率可达 62–71%。剔除能量最高的 20% 预测结果，可将准确率提高 5–10%。在安全关键场景中，可将高能量预测结果标记出来以供人工审核。

表 6：**不同缺失率下能量迭代的影响（CMU-MOSI）**。 $T = 0$ 在完整数据下最优； $T = 3$ 有助于应对缺失模态。

迭代次数	指标	mr=0	mr=0.1	mr=0.3	mr=0.5	mr=0.7	mr=0.9
$T = 0$	MAE↓	**0.669**	0.738	0.905	1.115	1.361	**1.408**
	Acc-7↑	**51.6**	45.9	**42.9**	34.7	26.8	**21.9**
	Acc-2↑	**85.4**	**82.4**	77.1	**72.7**	**63.1**	**54.7**
	F1↑	**83.9**	80.7	**75.1**	68.9	**62.4**	**57.7**
$T = 3$	MAE↓	0.671	0.761	**0.897**	**1.068**	**1.271**	1.675
	Acc-7↑	50.3	45.8	39.5	**35.8**	**27.1**	19.1
	准确率-2↑	84.8	82.2	76.7	71.0	62.8	53.4
	F1↑	83.0	80.5	74.7	68.5	61.5	52.0
$T = 5$	MAE↓	0.682	**0.706**	0.924	1.094	1.322	1.445
	准确率-7↑	49.8	**46.9**	36.4	33.4	25.4	20.4
	准确率-2↑	84.2	82.1	**77.6**	68.2	59.5	54.8
	F1↑	81.3	**80.8**	75.5	65.8	58.0	53.5

表7：**不同评估规程下的性能（CMU-MOSI，mr=0.5）**。我们比较零掩蔽（将缺失特征替换为零）与噪声掩蔽（替换为高斯噪声），以验证 DCER 的鲁棒性确实存在，而非仅适用于特定协议。

协议	MAE↓	相关系数↑	七分类准确率	二分类准确率
零掩蔽	0.668	0.833	48.1	83.4
噪声掩蔽	0.712	0.801	45.3	81.7

表8：**能量-误差相关性与选择性预测**。高相关性有助于实现可靠的不确定性估计。

数据集	ρ(E, Err)	高E	低E	拒绝率@20%
MOSI	0.78	31.2%	62.4%	+8.3%
MOSEI	0.72	38.5%	64.8%	+5.1%
SIMS	0.81	29.7%	71.3%	+9.7%

5 讨论

为什么双阶段压缩有效? 两个阶段分别解决互补的问题。阶段1通过频率变换去除模态内冗余，集中保留与情绪相关的模式。阶段2通过容量受限的查询令牌，避免跨模态捷径，促使真正的融合。单独使用任何一个阶段都不够：没有跨模态瓶颈的频率压缩仍会留下走捷径的空间；没有频率压缩的瓶颈则需要处理含噪且冗余的输入。

频率压缩的理论动机。信息论视角解释了频域鲁棒性。考虑以速率 r进行随机掩蔽：

时域：如果一个判别性时间步 t^* 被掩蔽，信息损失将是灾难性的——$I(x_{\text{masked}}; y) \ll I(x; y)$。

频域：每个系数 $X_k = \sum_t x_t e^{-2\pi ikt/T}$ 都编码了全局时序信息。掩蔽会将损失分散到所有系数中，而不是集中在关键时间步上：

$$\text{Var}[I(X_{\text{masked}}; y)] < \text{Var}[I(x_{\text{masked}}; y)] \tag{9}$$

这解释了为什么小波/离散余弦变换预处理对缺失数据具有内在鲁棒性——任务相关信息分散在各个频带中，而非集中在特定时间步上。

U形多模态优势。 表5揭示了非单调模式：在完整数据下，TAV优于T-only（51.6% 对 51.0%）；在高缺失率下（mr=0.9：49.3% 对 46.8%），TAV同样表现更好；但在中等缺失率下（mr=0.3–0.5），T-only更加稳定。我们推测，这反映了重建噪声的影响：在中等缺失率下，基于能量的重建会引入扰动，从而干扰文本主导的预测；而在高缺失率下，重建信号对于维持任何跨模态信息都变得至关重要。这表明，自适应策略——例如在文本置信度较高时跳过重建——是值得未来研究的方向。

计算方面的考量。 DCER会增加适度的额外开销：小波/离散余弦变换的复杂度为$O(T \log T)$，而能量迭代会在推断过程中增加 3× 次前向传播。总推理时间约为15毫秒，而MulT为12毫秒（额外开销为25%），对于大多数应用而言是可以接受的。

局限性。 在极高缺失率（$> 70\%$）下，性能会显著下降。频率变换假设存在时间/空间结构，因此可能无法推广到所有模态类型。此外，评估仅限于情感分析；还需进一步研究更广泛的多模态任务。

6 结论

我们提出了 DCER，这一框架将频域压缩、跨模态瓶颈融合与基于能量的重建统一起来，用于多模态情感分析。DCER 在 CMU-MOSI、CMU-MOSEI 和 CH-SIMS 上达到了最先进的水平，并具备内在不确定性量化能力，无需额外校准即可实现可靠的置信度估计。除了情感分析之外，在跨模态整合之前利用模态特定信号结构（音频的时间结构、视频的空间结构）这一原则，也可能有益于其他已知存在此类结构的多模态任务，例如视听语音识别或视频问答。

影响声明

本文提出了鲁棒多模态融合方面的基础性方法进展，未引入新的数据集，也未在人类受试者身上部署系统。下文将讨论其潜在的广泛影响。

潜在的积极影响。 我们的框架提升了对缺失模态的鲁棒性，并提供内在不确定性量化能力，可提高心理健康监测和临床决策支持等安全关键应用的可靠性。在高风险场景中，标记不确定预测以供人工审核的能力可降低自动化错误的风险。

潜在的负面影响。多模态情感分析系统会处理音频、视觉和文本信号，其中可能包含敏感个人信息。鲁棒性的提升可能降低

在监控或操纵场景中部署系统的门槛。此外，情感分析模型可能会继承训练数据中存在的人口统计和文化偏差，从而导致其在不同人群中的表现存在差异。我们鼓励从业者在部署此类系统前进行全面的公平性审计并取得知情同意。我们的评估偏差发现（零掩蔽与噪声掩蔽）进一步凸显了严格评估规程的重要性，以避免高估部署就绪度。

参考文献

[1] Amir Zadeh、Minghai Chen、Soujanya Poria、Erik Cambria和Louis-Philippe Morency。用于多模态情感分析的张量融合网络。载于*2017*年自然语言处理实证方法会议论文集，第1103–1114页，2017年。[2] Yao-Hung Hubert Tsai、Shaojie Bai、Paul Pu Liang、J. Zico Kolter、Louis-Philippe Morency和Ruslan Salakhutdinov。面向未对齐多模态语言序列的多模态变换器。载于计算语言学协会第*57*届年会论文集，第6558–6569页，2019年。[3] Naftali Tishby、Fernando C Pereira和William Bialek。信息瓶颈方法。*arXiv*预印本 *physics/0004057*，2000年。[4] Amir Zadeh、Rowan Zellers、Eli Pincus和Louis-Philippe Morency。视频中的多模态情感强度分析：面部手势与言语信息。载于*IEEE*智能系统，第31卷，第82–88页，2016年。[5] Wasifur Rahman、Md Kamrul Hasan、Sangwu Lee、AmirAli Bagher Zadeh、Chengfeng Mao、Louis- Philippe Morency和Ehsan Hoque。在大型预训练Transformer中整合多模态信息。载于计算语言学协会第*58*届年会论文集，第2359–2369页，2020年。[6] Wenmeng Yu、Hua Xu、Ziqi Yuan和Jiele Wu。利用自监督多任务学习为多模态情感分析学习模态特定表示。载于*AAAI*人工智能会议论文集，第35卷，第10790–10797页，2021年。[7] Zirun Guo、Tao Jin和Zhou Zhao。面向情感分析与情绪识别的缺失模态多模态提示学习。载于计算语言学协会第*62*届年会论文集（第*1*卷：长文），第1726–1736页，2024年。[8] Licai Sun、Zheng Lian、Bin Liu和Jianhua Tao。面向稳健多模态情感分析的具有双层特征恢复的高效多模态Transformer。*IEEE*情感计算汇刊，14(2):1137–1148，2023年。[9] Jiahao Qin、Feng Liu和Lu Zong。BC-PMJRS：一种受脑计算启发的预定义多模态联合表示空间，用于增强跨模态学习。神经网络，188:107449，2025年。[10] Luan Tran、Xiaoming Liu、Jiayu Zhou和Rong Jin。通过级联残差自编码器进行缺失模态插补。载于*IEEE*计算机视觉与模式识别会议论文集，第1405–1414页，2017年。[11] Haoyu Zhang、Wenbin Wang和Tianshu Yu。面向不完整数据的稳健多模态情感分析。载于神经信息处理系统进展，2024年。[12] Peng Xu、Xiatian Zhu和David A. Clifton。基于Transformer的多模态学习：综述。*IEEE*模式分析与机器智能汇刊，45(10):12113–12132，2023年。[13] Alexander A Alemi、Ian Fischer、Joshua V Dillon和Kevin Murphy。深度变分信息瓶颈。载于国际学习表征会议，2017年。[14] Andrew Jaegle、Felix Gimeno、Andrew Brock、Andrew Zisserman、Oriol Vinyals和Joao Carreira。Perceiver：基于迭代注意力的通用感知。载于国际机器学习会议，第4651–4664页，2021年。[15] Tian Zhou、Ziqing Ma、Qingsong Wen、Xue Wang、Liang Sun和Rong Jin。FEDformer：用于长期序列预测的频率增强分解Transformer。载于国际机器学习会议，第27268–27286页，2022年。

[16] Kai Xu、Minghai Qin、Fei Sun、Yuhao Wang、Yen-Kuang Chen和Fengbo Ren。在频域中学习。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第1740–1749页，2020年。[17] Neil Zeghidour、Olivier Teboul、F´elix de Chaumont Quitry和Marco Tagliasacchi。Leaf：用于音频分类的可学习前端。载于国际学习表征会议，2021年。[18] James Lee-Thorp、Joshua Ainslie、Ilya Eckstein和Santiago Ontanon。FNet：利用傅里叶变换混合标记。载于计算语言学协会北美分会*2022*年会议论文集，第4296–4313页，2022年。[19] Jiahao Qin、Yiwen Wang和Feng Liu。结合时间序列成像的双模态Transformer，用于稳健的癫痫发作预测。《应用科学》，15(3):1538，2025年。[20] Jiahao Qin、Liyuan Zong和Feng Liu。通过脑电图与功能性磁共振成像中的跨感知方法探索内言语识别。《应用科学》，14(17):7720，2024年。[21] Yann LeCun、Sumit Chopra、Raia Hadsell、M Ranzato和F Huang。基于能量的学习教程。《结构化数据预测》，1(0)，2006年。[22] David Belanger和Andrew McCallum。结构化预测能量网络。载于国际机器学习会议，第983–992页，2016年。[23] Yilun Du和Igor Mordatch。基于能量的模型的隐式生成与建模。载于神经信息处理系统进展，第32卷，2019年。[24] Hangbo Bao、Wenhui Wang、Li Dong、Qiang Liu、Owais Khan Mohammed、Kriti Aggarwal、Subhojit Som和Furu Wei。VLMo：基于模态专家混合的统一视觉语言预训练。载于神经信息处理系统进展，第35卷，2022年。[25] AmirAli Bagher Zadeh、Paul Pu Liang、Soujanya Poria、Erik Cambria和Louis-Philippe Morency。真实场景中的多模态语言分析：CMU-MOSEI数据集与可解释动态融合图。载于计算语言学协会第*56*届年会论文集，第2236–2246页，2018年。[26] Wenmeng Yu、Hua Xu、Fanyang Meng、Yilin Zhu、Yixiao Ma、Jiele Wu、Jiyun Zou和Kaicheng Yang。CH-SIMS：一个具有模态细粒度标注的中文多模态情感分析数据集。载于计算语言学协会第*58*届年会论文集，第3718–3727页，2020年。[27] Kezhou Chen、Shuo Wang、Huixia Ben、Shengeng Tang和Yanbin Hao。用于情感分析的多模态适配器混合。载于计算语言学协会美洲国家分会*2025*年会议论文集，第1822–1833页，2025年。[28] Xuan He、Hao Liang、Bo Peng、Weiquan Xie、Muhammad Haris Khan、Siyang Song和Zhiyong Yu。MSAmba：利用状态空间模型探索多模态情感分析。载于*AAAI*人工智能会议论文集，第39卷，第1309–1317页，2025年。[29] Yujian Cai、Xingguang Li、Yingyu Zhang、Jinsong Li、Fazheng Zhu和Lin Rao。基于多层特征融合与多任务学习的多模态情感分析。《科学报告》，15(1):2126，2025年。