Computer Science and Application 计算机科学与应用, 2026, 16(1), 281-294
Published Online January 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.161023

Hans汉斯

CMPTA：预训练大模型在多模态情感分析任务中的应用研究

李志豪[1]，智　宇[1]，陈　昂[1,2]

[1]温州大学计算机与人工智能学院，元宇宙与人工智能研究中心，浙江　温州
[2]温州大学元宇宙与人工智能研究院，浙江　温州

收稿日期：2025年12月16日；录用日期：2026年1月15日；发布日期：2026年1月22日

摘　要

大语言模型(LLMs)在自然语言处理领域取得了显著进展，但将其有效迁移至多模态情感分析(MSA)任务仍面临巨大挑战。主要难点在于如何弥合异构模态(如视觉、音频)特征与预训练文本大模型语义空间之间的鸿沟。现有方法多依赖复杂的深度融合网络或昂贵的全量微调，难以充分利用大模型的推理与泛化能力。为此，本文提出了一种轻量级的跨模态伪Token适配器(Cross-Modal Pseudo-Token Adapter, CMPTA)。该方法并不破坏大模型的原有参数，而是通过高效的注意力机制，将非文本模态特征转化为LLM可理解的“伪Token”(Pseudo-Tokens)，并以软提示(Soft Prompts)的形式注入文本输入序列，从而实现多模态信息与文本语义的深度对齐。此外，本文还系统探究了伪Token数量对模型语义对齐效果的影响规律。实验结果表明，CMPTA能够有效激发大模型的多模态情感理解能力，其性能优于当前的先进基线方法，验证了该框架的有效性与泛化能力。

关键词

多模态情感分析，大语言模型，伪Token，参数高效微调，跨模态适配器

CMPTA: Exploring the Application of Pre-Trained Large Language Models in Multimodal Sentiment Analysis

Zhihao Li[1], Yu Zhi[1], Ang Chen[1,2]

[1]Research Center for Metaverse and Artificial Intelligence, College of Computing Science and Artificial Intelligence, Wenzhou University, Wenzhou Zhejiang
[2]Institute of Metaverse and Artificial Intelligence, Wenzhou University, Wenzhou Zhejiang

Received: December 16, 2025; accepted: January 15, 2026; published: January 22, 2026

文章引用: 李志豪, 智宇, 陈昂. CMPTA: 预训练大模型在多模态情感分析任务中的应用研究[J]. 计算机科学与应用, 2026, 16(1): 281-294. DOI: 10.12677/csa.2026.161023

Abstract

Large Language Models (LLMs) have achieved remarkable progress in Natural Language Processing, yet effectively adapting them to Multimodal Sentiment Analysis (MSA) tasks remains a significant challenge. The core difficulty lies in bridging the gap between heterogeneous modal features (e.g., visual, acoustic) and the semantic space of pre-trained text models. Existing approaches often rely on complex deep fusion networks or expensive full fine-tuning, failing to fully leverage the reasoning and generalization capabilities of LLMs. To address this, we propose a lightweight Cross-Modal Pseudo-Token Adapter (CMPTA). Instead of disrupting the original parameters of the LLM, this method employs an efficient attention mechanism to transform non-textual modal features into "Pseudo-Tokens" understandable by the LLM. These tokens are then injected into the text input sequence as Soft Prompts, achieving deep alignment between multimodal information and textual semantics. Furthermore, we systematically investigate the impact of the number of pseudo-tokens on semantic alignment. Experimental results demonstrate that CMPTA effectively stimulates the multimodal sentiment understanding capability of LLMs, outperforming state-of-the-art baselines, thereby validating the effectiveness and generalization ability of the framework.

Keywords

Multimodal Sentiment Analysis, Large Language Models, Pseudo-Token, Parameter-Efficient Fine-Tuning (PEFT), Cross-Modal Adapter

1. 引言

人工智能对人类情感的综合分析能力十分重要，近年来，多模态情感分析在自然语言处理领域越来越引人注意，面对各种各样的限制，早期的情感分析仅仅只能对文本进行综合分析，这种方式其实与人类真实的情感表达差距甚大，相同的一句话，通过不同的脸部表情与不同的语气说出往往表达的是不同的情感[1]，因此，通过多种信息对情感进行分析十分必要，这也为更优质的人机交互体验提供了一个方向。

随着深度学习技术的不断发展，研究者逐渐意识到单一模态在情感识别中的局限性，而多模态信息的融合能够更全面地刻画人类情绪状态。文本模态往往蕴含语义信息，语音模态体现语调、节奏与声音强弱等情绪线索，而视觉模态则呈现面部表情、眼神、姿态等外在情绪表现[2]。将这些来源不同、表达方式不同但互补的模态进行联合建模，不仅能够提高情感识别的准确率，也使得模型更贴近真实的人类认知过程。因此，多模态情感分析正逐渐成为人工智能研究领域的重要方向，其在智能客服、教育辅导、心理健康监测、社交机器人等多种应用场景中都具有巨大的发展潜力[3]。随着多模态大模型的兴起，这一领域也正迈向更加细致、智能、可泛化的情感理解时代[4]。

2. 相关工作

目前，针对多模态情感分析任务出现了不少杰出的工作，他们大多数是使用一些传统的方法，包括基于对比学习的方法[5]-[6]，基于监督学习的方法[7]-[9]，基于图学习的方法[10]-[12]等等。基于预训练

的大语言模型的方法通常指定设计好的大语言模型[13]-[15]，例如 Bert [16]或者 Llama [17]，这些方法通过将非文本模态的信息转化为与文本模态维度等效的信息，再将其与文本模态信息进行融合，最后达到能在大语言模型中进行情感分析任务的目的。

近几年来，研究人员更加关注使用预训练的大语言模型来进行情感分析任务，并且取得越来越好的效果，主要贡献如下。

DialogueLLM [18]直接从对话文本中识别情感，通过学习将多模态对话特征空间与基于大规模情感标注语料预训练的情感特征空间对齐来解决该问题。

Emotion-LLaMA [19]构建了新的情感分析数据集，通过将音频、视觉与文本特征对齐到统一的情感语义空间，实现细粒度情感识别与可解释推理。

ECPEC [20]提出基于知识蒸馏与 LLM 的多模态对话情绪–原因对抽取方法，通过教师模型生成情绪摘要并优化学生模型，实现情绪与成因联合识别。

M5SER [13]提出两阶段对比学习框架，将多语种语音与情绪词对齐，实现跨语言零样本语音情绪识别。

DeepMLF [21]通过可学习融合 token 在若干层解码器中渐进式深度融合音视频与文本，仅用二十个 token 即实现多模态情感分析。

MERITS-L [22]先用大模型给语音识别文本打伪标签预训练文本编码器，再与语音嵌入做分层上下文建模和交叉注意力融合。

3. 方法

本文提出的 CMPTA 模型，整体架构如图 1 所示，编码器由预训练大语言模型的文本嵌入层(Text Embedder)、基于 LSTM [23]的时序特征对齐层以及跨模态伪 Token 适配器(CMPT)组成，解码器为预训练好的大语言模型。

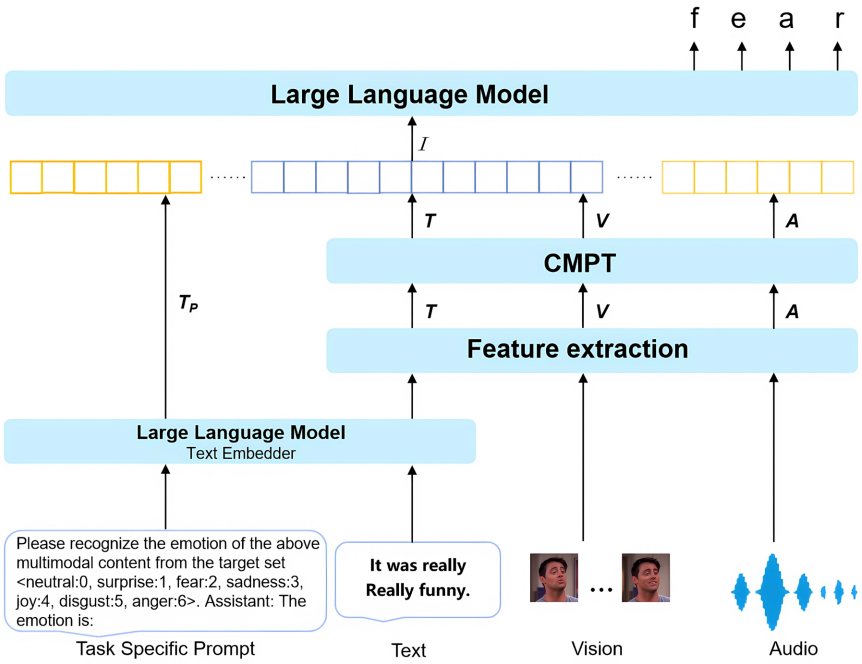


Figure 1. Cross-modal Pseudo-Token adapter framework
图 1. 跨模态伪 Token 适配器框架概览

3.1. CMPTA 的编码器

3.1.1. 文本嵌入层

文本嵌入层是预训练语言模型中将离散符号映射到连续向量空间的关键模块。对于输入的文本序列 $X=(x_1,x_2,\cdots,x_T)$ 其中每个 x_i 都是词表中的索引，嵌入层将其转换为维度为 d 的向量表示，使模型能够在连续空间中捕捉语义与句法特征，并支持基于梯度的优化。

设预训练模型的嵌入矩阵为 $E\in R^{|V|\times d}$，其中 $|V|$ 为词表大小、d 为嵌入维度，则嵌入操作可表示为：

$$h_i=E[x_i],i=1,\cdots,T \tag{1}$$

其中 $E[x_i]$ 表示嵌入矩阵中第 x_i 行的向量。对于批量输入张量 $x\in R^{B\times T}$，其对应的嵌入输出为：

$$H=Embed(X)\in R^{B\times T\times d} \tag{2}$$

3.1.2. 时序特征对齐层

虽然预训练大语言模型本身具备强大的长序列建模能力，但原始的视觉和音频特征序列往往存在采样率高、噪声大、时序冗余等问题。直接将其输入 LLM 会导致上下文窗口被无效信息占据，降低推理效率。

因此，我们在 LLM 之前引入了一个时序建模模块，并以双向 LSTM 作为主要实现形式，用于对非文本模态特征进行时序压缩与去噪。通过进一步的消融实验，对比单向 LSTM 与双向 LSTM 在该模块中的表现，以分析不同时间上下文建模方式对整体性能的影响。它将长序列的原始模态特征编码为更紧凑的上下文表示，在保留关键情感动态的同时，对齐了非文本模态与文本模态的信息密度，为后续生成高质量的伪 Token 奠定基础。输入序列表示为：

$$X=(x_1,\cdots,x_T),x_t\in R^{d_{in}} \tag{3}$$

双向 LSTM 同时从前向与后向建模序列的动态变化，其输出为：

$$h_t=\left[\vec{h}_t;\overleftarrow{h}_t\right] \tag{4}$$

其中前向与后向隐藏状态互补地编码了局部与长程依赖关系，为后续任务提供更具表达力的特征。

为进一步提升表示质量，拼接后的隐藏状态经过一个线性映射与归一化层：

$$\widetilde{h}_t=\mathrm{LayerNorm}(Wh_t+b) \tag{5}$$

该步骤在压缩维度的同时，使特征分布更加稳定，有助于加快模型收敛。此外，在训练阶段使用 Dropout，以降低过拟合风险。考虑到实际输入序列具有不同长度，在输出阶段根据每个样本的真实长度 L 将超出有效范围的时间步置零，从而保证特征序列仅由有效语义位置构成。这使得模型能够灵活地处理可变长度输入，同时保持特征对齐与结构一致性。

3.1.3. 跨模态伪 Token 适配器

为了弥合异构模态(视觉、音频)与预训练大语言模型语义空间之间的鸿沟，我们设计了一个跨模态伪 Token 适配器(CMPT Adapter)。不同于传统方法中设计复杂的深层融合网络，本模块的核心目标是充当“桥梁”，将非文本模态特征高效映射为 LLM 可理解的软提示(Soft Prompts)。

对于任意两种模态 X 与 Y 令其特征序列分别为：

$$X=\{x_1,\cdots,x_T\},Y=\{y_1,\cdots,y_T\} \tag{6}$$

如图 2 所示，该适配器采用参数高效的注意力交互机制。它并不试图替代 LLM 的推理能力，而是通

过多头注意力机制(MHA)进行计算，对于任意两个模态 X 与 Y，跨模态注意力计算定义为：

$$Attn\left(Q_Y,K_X,V_X\right)=\text{softmax}\left(\frac{Q_Y K_X}{\sqrt{d}}\right)V_X \tag{7}$$

其中，$X,Y\in\left(T,V,A\right)$，$X\neq Y$，d 为隐藏层维度的大小，例如，当 Y 为文本模态时，X 作为视觉(V)和音频(A)模态分别经过(7)式计算得到 t_1 与 t_2，Y 为视觉模态和音频模态时以此类推，该过程生成的六类特征向量被定义为伪 Token：

$$\left\{t_1,t_2,v_1,v_2,a_1,a_2\right\} \tag{8}$$

可进一步输入到后续融合层、多模态 Transformer 或分类器中，用于增强下游任务性能。

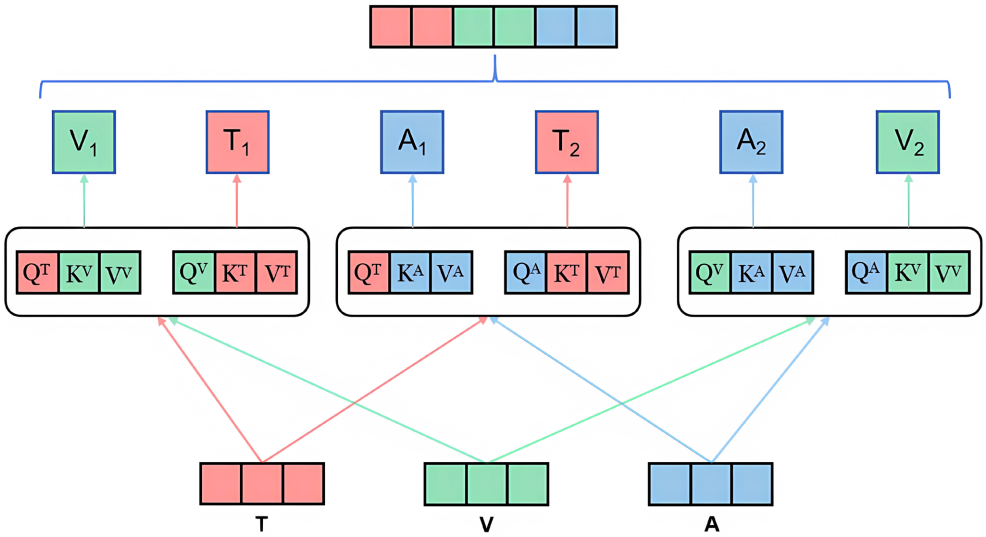


Figure 2. Cross-modal Pseudo-Token adapter
图 2. 跨模态伪 token 适配器

3.2. CMPTA 的解码器

在 CMPTA 框架中，解码器承担着将多模态融合后的表示映射为自然语言输出的核心职责。本研究采用 Qwen-1.8 作为解码器主干。Qwen 系列是基于 Transformer 解码器结构的大规模自回归语言模型，通过海量中英文语料预训练，具备强大的语义建模能力、知识泛化能力以及对齐上下文的文本生成能力。选择 Qwen-1.8 作为解码器主要基于以下两点考虑：其一，模型规模适中、参数量与计算需求平衡，便于与多模态特征融合层进行端到端训练；其二，Qwen 在开放域对话、知识问答与指令跟随任务中表现优异，可有效提升本研究任务中的文本推理与回答的自然性和一致性。

3.3. 损失函数

模型在前向过程中直接接受预融合后的模态嵌入作为输入，并利用与之对齐的真实标签进行自回归训练。由于给定了真实标签，大语言模型会自动计算跨序列的交叉熵损失。具体而言，模型对每个时间步输出原始线性分数，并通过标准的下一个 token 预测任务最小化损失，公式如下：

$$L=-\sum_t \log p_\theta\left(y_t \mid x\le t\right) \tag{9}$$

其中被 mask 的标签不参与损失计算。

4. 实验

4.1. 数据集

本研究在两个经典的多模态情感分析数据集上验证模型性能，分别为 SIMS-V2 与 MELD。如表 1 所示，SIMS-V2 是中文多模态情绪数据集，样本来自短视频片段，包含文本、视觉与音频三模态信息，并采用连续情感强度标注，能够细粒度刻画情绪变化。MELD 则源自电视剧《Friends》的多角色对话场景，同样提供文本、视觉和语音模态，但使用 7 类离散情感标签，并包含跨轮次对话上下文。两个数据集在语言、情感标注体系、场景、数据来源等方面具有互补性，为全面评估模型的跨场景泛化能力提供了可靠基础。

Table 1. Summary of the SIMS-V2 and MELD datasets
表 1. SIMS-V2 和 MELD 数据集的统计情况

数据集(Dataset)	训练集(Train)	验证集(Valid)	测试集(Test)	总计(Total)	语言(Language)
SIMS-V2	2722	647	1034	4403	Chinese
MELD	9989	1109	2610	13708	English

4.2. 评价指标

本研究根据任务特性分别采用平均绝对误差(MAE)与加权 F1 分数(WF1)作为主要评价指标。对于采用连续情感强度标注的数据集，使用 MAE 衡量预测值与真实值之间的平均绝对偏差，刻画模型在回归情感强度方面的误差表现。其定义为：

$$\mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}\left|y_i-\hat{y}_i\right| \tag{10}$$

对于使用离散情感类别标注的数据集，采用 WF1 分数反映模型在类不平衡条件下的整体分类性能。WF1 分数对各类别值按照样本数量加权：

$$\mathrm{WF1}=\sum_{c=1}^{c} w_c\cdot \mathrm{Fl}_c, w_c=\frac{n_c}{\sum_{j=1}^{C} n_j} \tag{11}$$

MAE 关注预测偏差规模，而 WF1 能有效处理类别分布不均带来的偏差，两者结合能够全面评估模型在连续与离散情感任务上的表现。

4.3. 实验设置

本文的 CMPTA 模型接受文本、视觉和音频特征作为输入，对于整个数据集，首先将数据集分为训练集、验证集和测试集，训练集用于模型的训练阶段，验证集用于训练过程中检验模型的阶段性效果，实验结果是模型在测试集上的表现。所有的模型训练和测试皆是在一个装有 windows 系统上的设备完成的，该设备配备 GEFORCE RTX 4090 显卡。模型训练的超参数见表 2 所示：

Table 2. Experimental hyperparameters and prompt settings for the SIMS-V2 and MELD datasets
表 2. SIMS-V2 和 MELD 数据集的实验超参数和提示词设置概览

超参数(Hyperparameter)	数值(Value)
特征维度	1024
优化器	AdamW

续表

学习率	$5e^{-4}$
迭代次数	150 Epochs
批次大小	16
提示词(Prompt)	
数据集(Dataset)	**提示词(Prompt)**
SIMS-V2	'请对上述多模态内容的情感强度进行预测，范围在[−1.0, +1.0]之间。响应：情感为'
MELD	'Please recognize the emotion of the above multimodal content from the target set <neutral:0, surprise:1, fear:2, sadness:3, joy:4, disgust:5, anger:6>. Assistant: The emotion is'

4.4. 实验结果

我们在 SIMS-V2 和 MELD 这两个数据集上分别进行了实验，所有的实验均保证了训练集、验证集和测试集拥有相同方式的划分，对于 SIMS-V2 这个数据集，我们做的是情感回归任务，评价指标为平均绝对误差，结果如表 3 所示，我们的方法 MAE 值为 0.308，显著优于对比方法，说明在情感回归任务中，对多模态信息的互补性建模更加充分，能更好地捕捉情感强度的细微变化。对于 MELD 数据集，我们做的是情感分类任务，评价指标为加权 F1 分数，结果如表 4 所示，我们的方法 WF1 分数为 59.49，超过现有方法，表明模型在复杂对话场景下能够更有效融合多模态信息，并缓解说话人变化和情境噪声带来的影响。

总体而言，在回归型(SIMS-V2)和对话型分类(MELD)两类特性差异明显的数据集上均取得稳定提升，说明我们的方法具有较好的跨数据集泛化能力和融合有效性。

Table 3. The results on the SIMS-V2 dataset
表 3. 在 SIMS-V2 上的实验结果

方法	MAE↓	Corr
LMF [24]	0.343	0.638
Self-MM [25]	0.335	0.640
MAG-BERT [26]	0.334	0.691
MSE-LLaMA2-7B [27]	0.382	0.553
CMPTA (Ours)	0.308	0.686

Table 4. The results on the MELD dataset
表 4. 在 MELD 上的实验结果

方法	WF1↑
TFN [28]	57.74
MMGCN [29]	58.31
GA2MIF [30]	58.94
CMPTA (Ours)	59.49

4.5. 消融实验

我们在 SIMS-V2 和 MELD 数据集上，分别对时序特征对齐层、跨模态伪 Token 适配器和伪 token 生成数量进行了消融研究，结果如表 5 和表 6 所示。

Table 5. Ablation results on SIMS-V2
表 5. 在 SIMS-V2 上的消融实验结果

时序特征对齐层的消融研究	
是否启用	MAE
是	0.308
否	0.412
LSTM 类型	MAE
双向	0.308
单向	0.328
跨模态伪 Token 适配器的消融研究	
是否启用	MAE
是	0.308
否	0.380
伪 token 生成数量的消融研究	
数量	MAE
5	0.313
6	**0.308**
7	0.309
8	0.311

Table 6. Ablation results on MELD
表 6. 在 MELD 上的消融实验结果

时序特征对齐层的消融研究	
是否启用	MAE
是	59.49
否	41.19
LSTM 类型	MAE
双向	59.49
单向	53.89
跨模态伪 Token 适配器的消融研究	
是否启用	MAE
是	59.49
否	51.31
伪 token 生成数量的消融研究	
数量	WF1
5	54.20
6	**59.49**

续表

7	55.67
8	55.23

4.5.1. 时序特征对齐层的消融研究

为了提高模型的健壮性，我们在特征对齐层没有使用基于 Transformer 的编码器，尽管基于 Transformer 的编码器在很多任务上表现出色，但在本任务的模态对齐阶段，特征序列相对较短且样本量有限。相比之下，双向 LSTM 具有更强的归纳偏置，在小样本下更易收敛且不易过拟合[23]。为了提高模型的健壮性和训练效率，本工作选择了双向 LSTM 作为对齐层，如表 5 和表 6 所示，在使用双向 LSTM 时，模型在 SMIS-V2 和 MELD 数据集上性能都有不同程度的提高。

在 SMIS-V2 数据集上，从表 5 可知，当启用时序特征对齐层时，MAE 为 0.308，而移除后性能显著下降至 0.412。这说明高质量的时序特征是情感回归的基础，直接使用原始或弱处理特征难以有效表达情感相关信息，特征提取模块在降低回归误差方面起到了关键作用。

在 MELD 数据集上，从表 6 可知，启用时序特征对齐层时，WF1 分数达到 59.49，而移除后骤降至 41.19，性能下降极为明显。这表明在 MELD 这种多说话人、情绪变化复杂的对话场景中，高质量时序特征对情感判别至关重要，缺乏有效时序特征会严重削弱模型的判别能力。

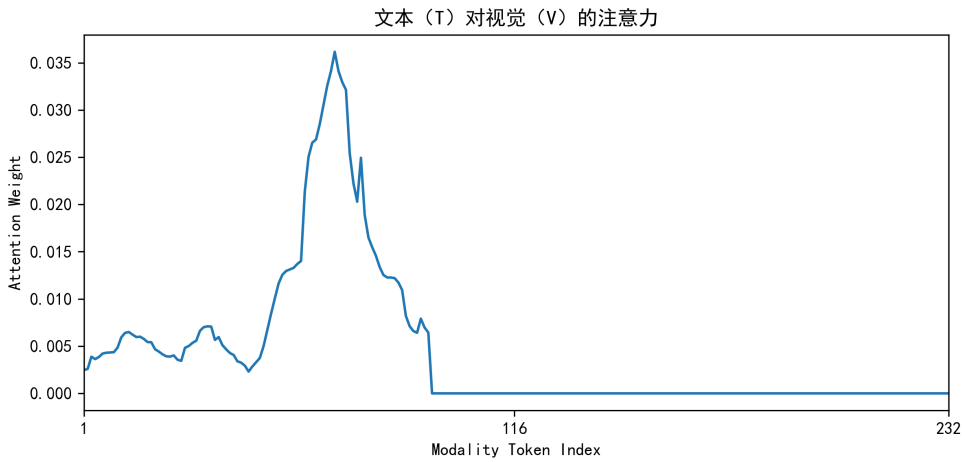
4.5.2. 跨模态伪 Token 适配器的消融研究

在 SMIS-V2 数据集上，从表 5 可知，启用跨模态伪 Token 适配器可将 MAE 从 0.380 降至 0.308，性能提升明显，表明该适配器在建模跨模态交互和全局语义对齐方面具有重要贡献。缺少 CMPT 时，模型对不同模态间互补信息的利用不足，导致情感强度预测精度下降。

在 MELD 数据集上，从表 6 可知，移除跨模态伪 Token 适配器后，WF1 从 59.49 降至 51.31，说明 CMPT 在 MELD 上同样发挥了重要作用。该模块有助于建模多模态间的深层交互关系，从而更好地应对对话中模态信息不一致和上下文复杂的问题。

4.5.3. 跨模态伪 Token 适配器的定性研究

我们在 SMIS-V2 和 MELD 数据集上对跨模态伪 Token 适配器进行了定性分析。在模型训练过程中，我们随机抽取了一些样本，以观察文本模态(T)在训练过程中是如何利用视觉模态(V)和音频模态(A)的信息进行特征增强和对齐的。通过可视化 T 对 V 和 T 对 A 的注意力权重，我们能够直观地理解模型在跨模态交互中的信息流动，从而验证伪 Token 适配器在多模态特征融合中的有效性。



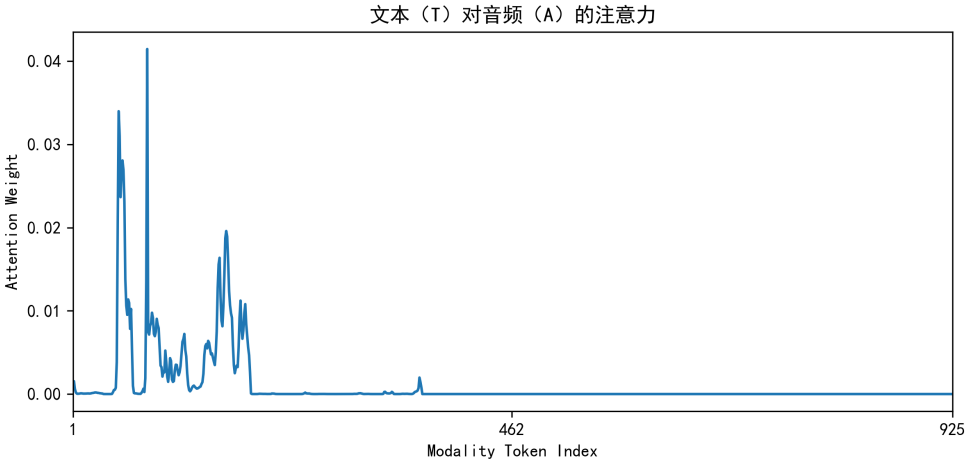


Figure 3. Visualization of visual and audio attention weights on SMIS-V2
图 3. SMIS-V2 上对视觉和音频注意力权重的可视化

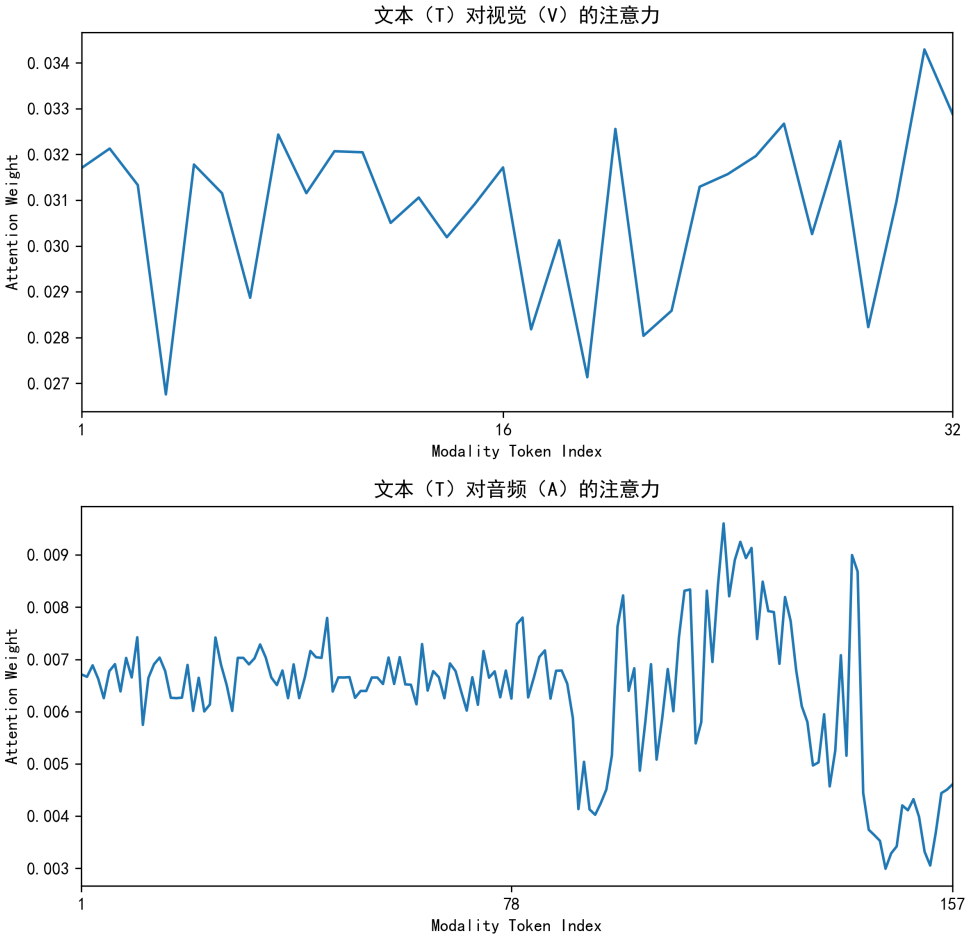


Figure 4. Visualization of visual and audio attention weights on MELD
图 4. MELD 上对视觉和音频注意力权重的可视化

在 SMIS-V2 数据集上，注意力图的横坐标表示 Token 序列索引，其中偏后的索引位置对应的是占位 Token (用于序列长度统一)，这些 Token 不参与实际信息交互，因此没有有效注意力权重。不同模态横坐

标范围不一致，是由于不同模态数据本身的序列长度不一致，最终导致了生成的伪 Token 的实际长度也不同。从图 3 可知，在训练过程中模型对视觉和音频模态的伪 Token 有着不同程度的关注度，在有效伪 Token 的范围内尤其明显，最高点往往意味着模型捕捉到了与文本模态信息最为密切相关的视觉和音频模态信息，充分说明跨模态伪 Token 适配器对 LLM 进行多模态情感分析具有十分重要的作用。

在 MELD 数据集上，注意力图的横坐标表示 Token 序列索引，不同模态横坐标范围不一致，是由于不同模态数据本身的序列长度不一致，最终导致了生成的伪 Token 的实际长度也不同。通过对图 4 的分析，我们能够直观地看到文本模态在训练过程中是如何利用视觉和音频模态进行特征增强与跨模态对齐的：高注意力值的区域对应文本模态与视觉模态之间的强相关信息，而音频模态的注意力则在部分关键时间步上体现作用。整体来看，注意力图验证了伪 Token 适配器在多模态特征融合中的有效性，同时也反映了各模态在不同数据集上的信息贡献差异。

4.5.4. 伪 Token 生成数量的消融研究

我们在两个数据集上就伪 token 的数量进行了消融实验，分为四种情况，如表 5 和表 6 所示，不同伪 token 数量下性能差异较小，但 6 个伪 token 时取得最优结果，这与跨模态伪 Token 生成数量一致，充分说明了我们提出的 CMPTA 的有效性。结果展示数量过少限制了跨模态信息表达能力，而数量过多可能引入冗余信息，反而影响回归效果，说明模型在表达能力与噪声控制之间存在一个合理的平衡点。

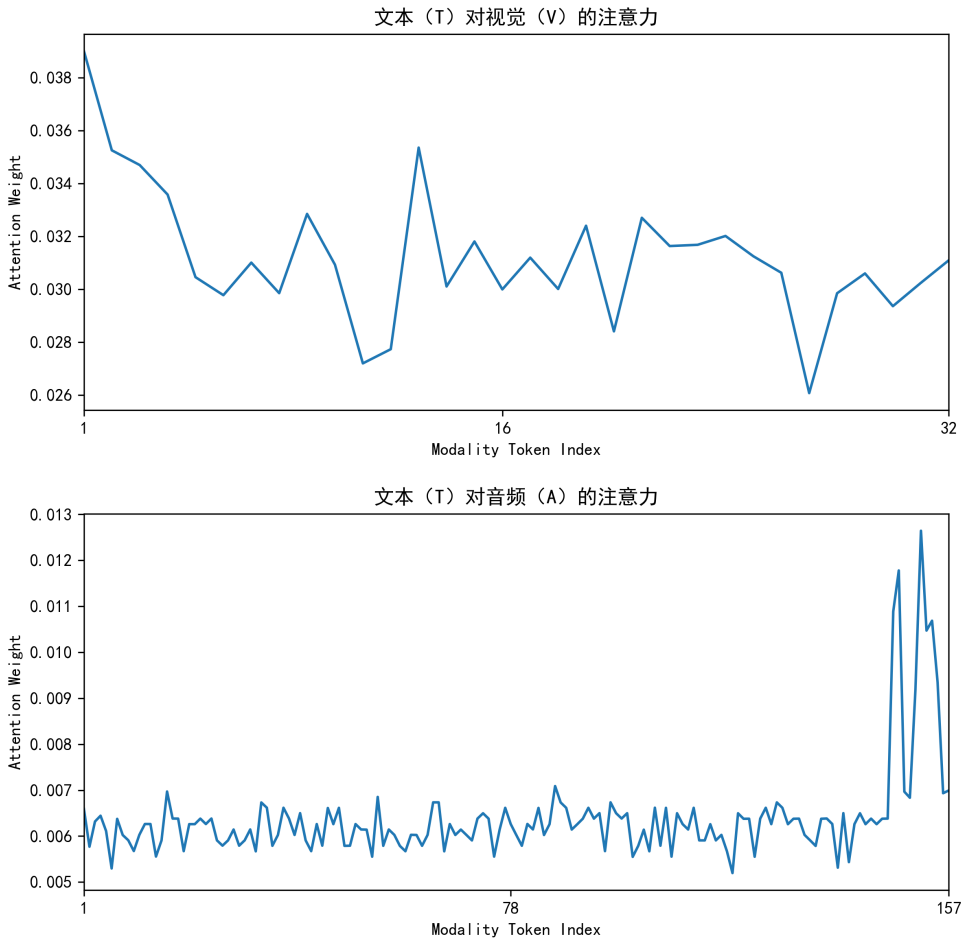


Figure 5. Visualization of visual and audio attention weights for a sarcastic sample in the MELD dataset
图 5. MELD 中一个反讽样本对视觉和音频注意力权重的可视化

4.5.5. 伪 Token 生成数量的定性案例分析研究

为了直观展示 CMPTA 的有效性，我们选取了 MELD 数据集中的一个反讽样本进行注意力权重可视化。文本内容为“这也太好了吧”，但在视觉上人物眉头紧锁，音频语调阴阳怪气的。传统文本模型将其误判为“积极”，如图 5，而 CMPTA 通过引入 6 个视觉伪 Token 和音频伪 Token，关注到了额外的视觉和音频模态信息，成功修正了 LLM 的判断，将其正确识别为“厌恶”情绪。这证明了伪 Token 数量为 6 个时成功捕获了非文本模态中的关键互补信息。

实验表明，当 Token 数量少于 6 时，伪 Token 承载的信息均值被过度压缩，导致非文本模态的关键细节丢失，而当 Token 数量过多时，引入了过多的冗余信息甚至噪声，干扰了 LLM 对文本主干语义的理解。因此，6 个 Token 在保留模态互补信息与维持语义空间纯净度之间达到了最佳平衡。

5. 结论

本文提出了 CMPTA 多模态情感建模框架，旨在解决多模态情感分析中跨模态交互不足和情感表达不充分的问题。CMPTA 通过引入时序特征对齐层和跨模态伪 Token 适配器，在统一语义空间内实现了更充分、稳定的多模态信息交互。在 SIMS-V2 情感回归任务和 MELD 对话情感分类任务上的实验结果表明，CMPTA 在 MAE 和 WF1 指标上均优于多种主流方法，验证了其在不同任务形式和数据集特性下的有效性与泛化能力。消融实验进一步证明了时序特征对齐层和跨模态伪 Token 适配器在性能提升中的关键作用，同时分析了伪 token 数量对模型性能的影响，说明合理的结构设计能够在信息表达能力与噪声抑制之间取得平衡。总体而言，CMPTA 为多模态情感分析提供了一种有效且具有良好扩展性的解决思路，可为后续多模态表示学习与跨模态建模研究提供参考。

致　谢

我们感谢 MISA、Self-MM、MSE-Adapter 的作者提供他们的代码和数据集。

基金项目

本课题受到“温州大学元宇宙与人工智能研究院”的“重大课题及项目产业化专项资金”(编号：2023103)的资助。

参考文献

[1] Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E. and Hussain, A. (2023) Multimodal Sentiment Analysis: A Systematic Review of History, Datasets, Multimodal Fusion Methods, Applications, Challenges and Future Directions. *Information Fusion*, **91**, 424-444. https://doi.org/10.1016/j.inffus.2022.09.025

[2] Li, J., Wang, X., Liu, Y. and Zeng, Z. (2024) CFN-ESA: A Cross-Modal Fusion Network with Emotion-Shift Awareness for Dialogue Emotion Recognition. *IEEE Transactions on Affective Computing*, **15**, 1919-1933. https://doi.org/10.1109/taffc.2024.3389453

[3] Pan, B., Hirota, K., Jia, Z. and Dai, Y. (2023) A Review of Multimodal Emotion Recognition from Datasets, Preprocessing, Features, and Fusion Methods. *Neurocomputing*, **561**, Article 126866. https://doi.org/10.1016/j.neucom.2023.126866

[4] Yuan, Y., Li, Z. and Zhao, B. (2025) A Survey of Multimodal Learning: Methods, Applications, and Future. *ACM Computing Surveys*, **57**, 1-34. https://doi.org/10.1145/3713070

[5] Yang, J., Yu, Y., Niu, D., Guo, W. and Xu, Y. (2023) ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis. *Proceedings of the* 61*st Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), Toronto, 9-14 July 2023, 7617-7630. https://doi.org/10.18653/v1/2023.acl-long.421

[6] Hu, Z., Wang, L., Lan, Y., Xu, W., Lim, E., Bing, L., *et al*. (2023) LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*, Singapore, 6-10 December 2023, 5254-5276. https://doi.org/10.18653/v1/2023.emnlp-main.319

[7] Hyeon, J., Oh, Y., Lee, Y. and Choi, H. (2025) Enhancing Speech Emotion Recognition through Segmental Average Pooling of Self-Supervised Learning Features. 2025 *IEEE International Conference on Big Data and Smart Computing* (*BigComp*), Kota Kinabalu, 9-12 February 2025, 191-198. https://doi.org/10.1109/bigcomp64353.2025.00047

[8] Lian, Z., Sun, H., Sun, L., Wen, Z., Zhang, S., Chen, S., *et al*. (2024) MER 2024: Semi-Supervised Learning, Noise Robustness, and Open-Vocabulary Multimodal Emotion Recognition. *Proceedings of the* 2*nd International Workshop on Multimodal and Responsible Affective Computing*, Melbourne, 28 October 2024-1 November 2024, 41-48. https://doi.org/10.1145/3689092.3689959

[9] Ma, H., Wang, J., Lin, H., Zhang, B., Zhang, Y. and Xu, B. (2024) A Transformer-Based Model with Self-Distillation for Multimodal Emotion Recognition in Conversations. *IEEE Transactions on Multimedia*, **26**, 776-788. https://doi.org/10.1109/tmm.2023.3271019

[10] Zhao, H., Ju, Y. and Gao, Y. (2024) Bilevel Relational Graph Representation Learning-Based Multimodal Emotion Recognition in Conversation. 2024 *IEEE International Conference on Multimedia and Expo* (*ICME*), Niagara Falls, 15-19 July 2024, 1-6. https://doi.org/10.1109/icme57554.2024.10688301

[11] Lian, Z., Chen, L., Sun, L., Liu, B. and Tao, J. (2023) GCNet: Graph Completion Network for Incomplete Multimodal Learning in Conversation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 8419-8432. https://doi.org/10.1109/tpami.2023.3234553

[12] Zhang, D., Chen, F. and Chen, X. (2023) DualGATs: Dual Graph Attention Networks for Emotion Recognition in Conversations. *Proceedings of the* 61*st Annual Meeting of the Association for Computational Linguistics*, Toronto, 9-14 July 2023, 7395-7408. https://doi.org/10.18653/v1/2023.acl-long.408

[13] Zou, H., Lv, F., Zheng, D., Chng, E.S. and Rajan, D. (2025) Large Language Models Meet Contrastive Learning: Zero-Shot Emotion Recognition across Languages. 2025 *IEEE International Conference on Multimedia and Expo* (ICME), Nantes, 30 June 2025-4 July 2025, 1-6. https://doi.org/10.1109/icme59968.2025.11209040

[14] Wang, L., Yang, J., Wang, Y., Qi, Y., Wang, S. and Li, J. (2024) Integrating Large Language Models (LLMs) and Deep Representations of Emotional Features for the Recognition and Evaluation of Emotions in Spoken English. *Applied Sciences*, **14**, Article 3543. https://doi.org/10.3390/app14093543

[15] Kadiyala, R.M.R. (2024) Cross-Lingual Emotion Detection through Large Language Models. *Proceedings of the* 14*th Workshop on Computational Approaches to Subjectivity*, *Sentiment*, *& Social Media Analysis*, Bangkok, 15 August 2024, 464-469. https://doi.org/10.18653/v1/2024.wassa-1.44

[16] Devlin, J., Chang, M.W., Lee, K. *et al*. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Minneapolis, 2 June-7 June 2019, 4171-4186.

[17] Touvron, H., Lavril, T., Izacard, G., *et al*. (2023) Llama: Open and Efficient Foundation Language Models. arXiv:2302.13971.

[18] Zhang, Y., Wang, M., Tiwari, P., Li, Q., Wang, B. and Qin, J. (2023) DialogueLLM: Context and Emotion Knowledge-Tuned LLaMA Models for Emotion Recognition in Conversations. arXiv:2310.11374.

[19] Cheng, Z., Cheng, Z., Hauptmann, A., He, J., Lian, Z., Lin, Y., *et al*. (2024) Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning. *Proceedings of the* 38*th International Conference on Neural Information Processing Systems*, Vancouver, 10-15 December 2024, 110805-110853. https://doi.org/10.52202/079017-3518

[20] Aruna Gladys, A., Vetriselvi, V. and Rajasekar, S.K. (2024) Multimodal Emotion Cause Pair Extraction in Conversations Using Knowledge Distillation and Large Language Models. 2024 *International Conference on Computational Intelligence and Network Systems* (*CINS*), Dubai, 28-29 November 2024, 1-8. https://doi.org/10.1109/cins63881.2024.10864457

[21] Georgiou, E., Katsouros, V., Avrithis, Y. and Potamianos, A. (2025) DeepMLF: Multimodal Language Model with Learnable Tokens for Deep Fusion in Sentiment Analysis. arXiv, arXiv:2504.11082.

[22] Dutta, S. and Ganapathy, S. (2025) LLM Supervised Pre-Training for Multimodal Emotion Recognition in Conversations. *ICASSP* 2025-2025 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Hyderabad, 6-11 April 2025, 1-5. https://doi.org/10.1109/icassp49660.2025.10889998

[23] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[24] Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Bagher Zadeh, A. and Morency, L. (2018) Efficient Low-Rank Multimodal Fusion with Modality-Specific Factors. *Proceedings of the* 56*th Annual Meeting of the Association for Computational Linguistics*, Melbourne, 15-20 July 2018, 2247-2256. https://doi.org/10.18653/v1/p18-1209

[25] Yu, W., Xu, H., Yuan, Z. and Wu, J. (2021) Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**, 10790-10797. https://doi.org/10.1609/aaai.v35i12.17289

[26] Rahman, W., Hasan, M.K., Lee, S., Bagher Zadeh, A., Mao, C., Morency, L., *et al*. (2020) Integrating Multimodal

Information in Large Pretrained Transformers. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, Online, 5-10 July 2020, 2359-2371. https://doi.org/10.18653/v1/2020.acl-main.214

[27] Yang, Y., Dong, X. and Qiang, Y. (2025) MSE-Adapter: A Lightweight Plugin Endowing LLMs with the Capability to Perform Multimodal Sentiment Analysis and Emotion Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, **39**, 25642-25650. https://doi.org/10.1609/aaai.v39i24.34755

[28] Zadeh, A., Chen, M., Poria, S., Cambria, E. and Morency, L. (2017) Tensor Fusion Network for Multimodal Sentiment Analysis. *Proceedings of the* 2017 *Conference on Empirical Methods in Natural Language Processing*, Copenhagen, 7-11 September 2017, 1103-1114. https://doi.org/10.18653/v1/d17-1115

[29] Hu, J., Liu, Y., Zhao, J. and Jin, Q. (2021) MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation. *Proceedings of the* 59*th Annual Meeting of the Association for Computational Linguistics and the* 11*th International Joint Conference on Natural Language Processing*, Online, 1-6 August 2021, 5666-5675. https://doi.org/10.18653/v1/2021.acl-long.440

[30] Li, J., Wang, X., Lv, G. and Zeng, Z. (2024) GA2MIF: Graph and Attention Based Two-Stage Multi-Source Information Fusion for Conversational Emotion Detection. *IEEE Transactions on Affective Computing*, **15**, 130-143. https://doi.org/10.1109/taffc.2023.3261279