arXiv:2601.13659v1 [cs.CL] 20 Jan 2026

时空先解耦再交互：面向多模态情感分析的解耦表征学习

Chunlei Meng[1], *Ziyang Zhou*[2], *Lucas He*[3], *Xiaojing Du*[4], *Chun Ouyang*[1 †], *Zhongxue Gan*[1]

[1]复旦大学， [2]汕头大学， [3]伦敦大学学院， [4]南澳大利亚大学

摘要

多模态情感分析融合语言、视觉和声学信息。主流方法基于模态不变与模态特定因子分解，或采用复杂融合，但仍依赖时空混合建模。这忽视了时空异质性，导致时空信息不对称，进而限制了模型性能。因此，我们提出 TSDA（时空先解耦再交互），在任何交互发生之前，将每个模态显式解耦为时间动态和空间结构上下文。对于每种模态，时间编码器和空间编码器将信号分别投影到独立的时间表征和空间表征中。随后，因子一致的跨模态对齐仅对齐不同模态间的时间特征及其对应的时间特征，并仅对齐空间特征及其对应的空间特征。因子特定监督和去相关正则化可减少跨因子泄漏，同时保留互补性。之后，门控重耦合模块将对齐流重新耦合，以供任务使用。大量实验表明，TSDA 优于基线方法。消融分析研究证实了该设计的必要性和可解释性。

索引词— 多模态情感分析，时间-空间解耦，表征学习

1. 引言

多模态情感分析（MSA）整合语言、视觉和听觉证据以推断情感，并支持人机交互、辅助智能体和媒体理解等应用 [1, 2, 3, 4]。与单模态学习相比，多模态模型能够利用不同来源的互补线索，但由于各模态的特征分布和信号可靠性存在异质性，表征学习仍然困难 [5, 6, 7, 8]。

近年来，研究人员提出了大量多模态模型来解决这一问题。已有工作主要沿着两个方向展开。第一个方向是在共享潜在空间中学习跨模态交互。例如，解耦模态蒸馏将知识从强模态迁移到弱模态，同时解耦教师和学生的目标，以稳定不平衡情况下的训练 [9]。基于专家的框架为各模态分配专属专家，并学习如何聚合这些专家，从而平衡专业化与覆盖度 [10]。对比特征分解引入对比目标，将共同信号与独有信号分离，从而在融合过程中明确决策边界 [11]。第二个方向通过将每种模态分解为不变子空间和特有子空间，在保留特有线索的同时缩小分布差异。MISA 在共享不变子空间中对齐模态，并在私有子空间中施加正交性约束，为下游融合提供互补视角 [12]。后续基于解耦的方法通过差异正则化或对比正则化，强化不变对齐并提升私有空间中的多样性 [13, 14]。以语言为中心的变体将有偏的文本特征分解，以提高跨领域的稳健性和迁移能力 [15, 16]。

尽管取得了稳步进展，大多数流程仍将每种模态编码为单一的混合嵌入后再进行交互，迫使模型联合学习时间动态与空间或结构上下文。对于情感分析而言，这种设计并不稳健：决定性证据往往存在于稀疏的时间模式中，例如韵律突发、犹豫和微表情；而身份、音色和句子级极性等静态上下文则十分丰富，且往往主导优化过程 [17]。当这些线索纠缠在一起时，目标函数在很大程度上由高方差的静态成分解释，从而削弱决定极性的时间轨迹，导致预测不稳定 [18, 19]。我们将这一问题称为时空信息不对称，并通过有选择地破坏一个因素、同时保留另一个因素的扰动来诊断这一问题，具体包括时间打乱和静态交换。这些观察结果促使我们提出一种明确建模时间因素与空间因素，并在融合前避免二者混合的表述方式。

因此，我们采用 TSDA（行动前时空解耦）来应对这一挑战。对于每种模态，时间编码器将有序令牌或帧映射为时间序列，空间编码器则生成与时间无关的结构集合。随后，通过因子一致跨模态对齐，将交互限制在因子一致的路径上。该方法采用具有块对角模式的掩码交叉注意力，使时间特征仅与跨模态的时间对应特征对齐，空间特征仅与空间对应特征对齐。通过令牌级的因子纯度监督和摘要级的去相关正则化，抑制残余的跨因子泄漏，从而在防止相互污染的同时保留互补性。随后，门控重耦合模块根据每个实例重新组合对齐后的时间摘要和空间摘要。正交正则项则在融合过程中保持两条流的良好条件。我们的主要贡献如下：

- 我们提出**TSDA**，这是一种新颖的**行动前时空解耦**框架，在交互前将动态与结构解耦，消除融合前的耦合并缓解时空不对称性。
- 我们提出了**因子一致性交叉模态对齐（FCCA）**，通过块对角掩码注意力分别对齐时间流和空间流。因子纯度监督与去相关进一步确保了各因子的分离，同时保留互补性。
- 我们提出了一种**门控重耦合模块（GR）**，利用按实例校准的权重整合时间摘要与空间摘要，并由因子分歧、置信度信号、干预式监督和正交性正则化共同引导，以减轻静态主导。

†通讯作者。

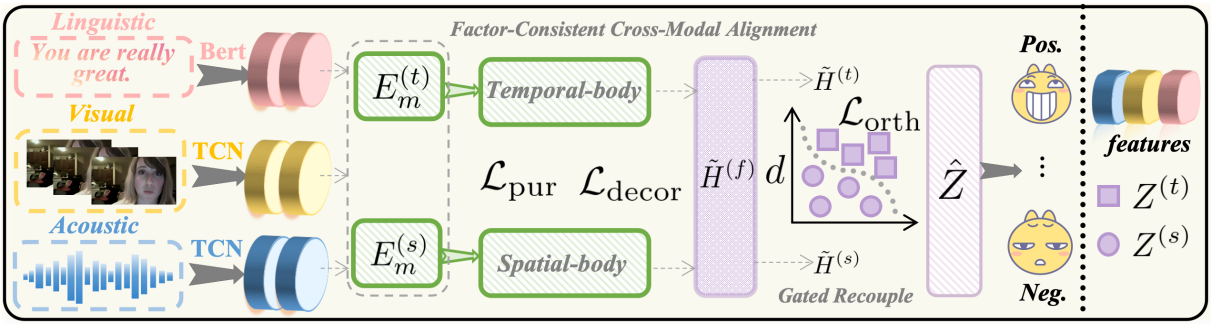

图 1。TSDA 的架构示意图：该方法将时间因素与空间因素解耦，使用 FCCA 对齐二者，并通过门控重耦合模块进行自适应融合。

2. 方法

2.1. 任务定义

令 $\mathcal{M} = \{L, V, A\}$ 表示语言、视觉和声学模态。对于每个模态 $m \in \mathcal{M}$，我们观测到一个长度为 T_m的令牌序列 $X_m = \{x_{m,1}, \ldots, x_{m,T_m}\}$。目标是预测情感目标 y，用于分类或回归。每个 X_m 都包含短暂的时间线索，以及相对稳定的空间或结构上下文。传统流程会在跨模态交互之前将这些因素联合嵌入，从而使学习偏向静态成分。如图1所示，我们则在表征阶段将时间与空间分离，再进行任何跨模态交换。

2.2. 时间与空间解耦

时间模型能够捕获长程依赖，但单一的混合嵌入会混合动态线索与静态上下文，并使学习产生偏差。我们从源头着手解决这一问题，将每个输入映射为两个针对互补方面的因子专属主体：时间演化与空间结构。对于每种模态 m，时间编码器 $E_m^{(t)}$生成一系列时间令牌，空间编码器 $E_m^{(s)}$则生成与时间无关的结构令牌：

$$F_m^{(t)} = E_m^{(t)}(X_m) \in \mathbb{R}^{T_m \times d_t}, \quad F_m^{(s)} = E_m^{(s)}(X_m) \in \mathbb{R}^{S_m \times d_s}. \tag{1}$$

其中，T_m 和 S_m 分别表示时间令牌和空间令牌的数量，d_t 和 d_s 分别表示嵌入宽度。空间粒度 S_m取决于模态，例如，视频的图像块或区域、音频的片段，以及语言的句子或短语令牌。在实际应用中，$E_m^{(t)}$和 $E_m^{(s)}$是单模态骨干网络上的轻量级头部：时间头部沿序列轴保留顺序，而空间头部则将局部证据聚合为稳定的结构令牌。这种因子分解为后续的因子一致性对齐做好准备，并防止动态信息与上下文过早协同适应。

2.3. 因子一致性交叉模态对齐

解耦后，每种模态 $m \in \{L, V, A\}$ 都会产生一个时间序列 $F_m^{(t)} \in \mathbb{R}^{T_m \times d_t}$和一个空间集合 $F_{(m}{}^{s} \in \mathbb{R}^{S_m \times d_s})$。传统融合会在同一空间中混合各个因子，从而引发跨因子干扰和静态主导。FCCA 通过采用块对角消息传递结构运行两个独立的跨模态注意力，使同类因子彼此对齐。这从形式上保证了因子一致性。令 $H^{(t)} = \text{concat}(\{F_m^{(t)}\}_m)$ 且 $H^{(s)} = \text{concat}(\{F_m^{(s)}\}_m)$。对于因子 $f \in \{t, s\}$，我们计算掩码注意力和摘要

$$\tilde{H}^{(f)} = \text{softmax}\left(\frac{Q^{(f)} K^{(f)\top}}{\sqrt{d_f}} + \log M^{(f)}\right) V^{(f)} \tag{2}$$

$$Z^{(f)} = \text{Pool}(\tilde{H}^{(f)}), \tag{3}$$

其中，$Q^{(f)}$，$K^{(f)}$，$V_{(f)}$是 $H^{(f)}$的线性投影，$M^{(f)}$ 是一个二值掩码，仅允许同一因子内进行注意力计算。该掩码使算子相对于因子索引呈块对角结构，因此对齐期间不会交换时间到空间的消息。最后，我们得到表征信息 $Z^{(f)}$。

残余泄漏通过两个层面加以控制。在令牌级，判别器 $D : \mathbb{R}^{d_f} \to [0, 1]$ 预测对齐后的令牌是否源自时间流，我们通过优化纯度目标来实现这一点：

$$\mathcal{L}_{\text{pur}} = \mathbb{E}\Big[-\log D(\tilde{h}^{(t)})\Big] + \mathbb{E}\Big[-\log\big(1 - D(\tilde{h}^{(s)})\big)\Big]. \tag{4}$$

在摘要级，我们通过惩罚二阶依赖和非线性依赖来限制因子间的重叠：

$$\mathcal{L}_{\text{decorr}} = \lambda_c \cos^2\big(Z^{(t)}, Z^{(s)}\big) + \lambda_h \,\text{HSIC}\big(Z^{(t)}, Z^{(s)}\big). \tag{5}$$

综合而言，块对角注意力提供结构约束，纯度监督确保因子可辨识性，而去相关则减少摘要之间的冗余；三者共同维持互补性，同时防止跨因子污染。

2.4. 门控重耦合

我们使用反映可靠性的实例级权重，重新组合对齐的时间摘要和空间摘要。当时间线索稀疏时，普通的标量门控往往会过度侧重静态上下文，因此我们利用 FCCA 和干预式监督中的信号对门控进行校准。令 $Z^{(t)}$ 和 $Z^{(s)}$ 为对齐摘要；定义分歧度 $d = 1 - \cos\big(Z^{(t)}, Z^{(s)}\big)$ 以及因子置信度 $c_t = \text{mean}\big(D(\tilde{H}^{(t)})\big)$ 和 $c_s = \text{mean}\big(1 - D(\tilde{H}^{(s)})\big)$。门控接收$\phi = [Z^{(t)}; Z^{(s)}; d; c_t; c_s]$ 并输出 $g = \sigma(w^\top \phi + b) \in (0, 1)$。我们利用投影头 $U_t \in \mathbb{R}^{d \times d_t}$和 $U_s \in \mathbb{R}^{d \times d_s}$ 构造 $\hat{Z} = g U_t Z^{(t)} + (1 - g) U_s Z^{(s)}$，并施加正交性正则项 $\mathcal{L}_{\text{orth}} = \|U_t^\top U_s\|_F^2$，以降低共线性并提升门控可辨识性。门控校准采用两种干预：时间

表1。基准测试上的性能比较，对齐/未对齐设置下的结果分别为 a/b （a：对齐， b：未对齐）。

方法	CMU-MOSI				CMU-MOSEI			
	平均绝对误差（↓）	准确率$_7$ （%）	准确率$_2$ (%)	F1分数 (%)	平均绝对误差 (↓)	准确率$_7$ (%)	准确率$_2$ (%)	F1分数 (%)
LMF [20]	0.931 / 0.963	36.9 / 31.1	78.7 / 79.1	78.7 / 79.1	0.564 / 0.565	52.3 / 51.9	84.7 / 83.8	84.5 / 83.9
MuLT [21]	0.936 / 0.933	35.1 / 33.2	80.0 / 80.3	80.1 / 80.3	0.572 / 0.556	52.3 / 53.2	82.7 / 84.0	82.8 / 84.0
TFN [22]	0.953 / 0.995	31.9 / 35.3	78.8 / 76.5	78.9 / 76.6	0.574 / 0.573	50.9 / 50.2	80.4 / 84.2	80.7 / 84.0
MISA [12]	0.754 / 0.742	41.8 / 43.6	84.2 / 83.8	84.2 / 83.9	0.543 / 0.557	52.3 / 51.0	85.3 / 84.8	85.1 / 84.8
FDMER [13]	- / 0.725	- / 44.2	- / 84.6	- / 84.7	- / 0.536	- / 53.8	- / 84.1	- / 84.0
ConFEDE [11]	- / 0.742	- / 46.3	- / 84.2	- / 84.2	- / 0.523	- / 54.9	- / 81.8	- / 82.3
Self-MM [5]	0.738 / 0.724	45.3 / 45.7	84.9 / 83.4	84.9 / 83.6	0.540 / 0.535	53.2 / 52.9	84.5 / 85.3	84.3 / 84.8
MMIN [4]	- / 0.741	- / -	83.5 / 85.5	83.5 / 85.51	- / 0.542	- / -	83.8 / 85.9	83.9 / 85.76
DMD [9]	0.721 / 0.721	46.2 / 46.7	83.2 / 84.0	83.2 / 84.0	0.546 / 0.536	52.4 / 53.1	84.8 / 84.7	84.7 / 84.7
DEVA [6]	- / 0.730	- / 46.3	- / 84.4	- / 84.5	- / 0.541	- / 52.3	- / 83.3	- / 82.9
DLF [15]	- / 0.731	- / 47.1	- / 85.1	- / 85.1	- / 0.536	- / 53.9	- / 84.4	- / 85.3
EMOE [10]	0.710 / 0.697	47.7 / 47.8	85.4 / 85.4	85.4 / 85.3	0.536 / 0.533	54.1 / 53.9	85.3 / 85.5	85.3 / 85.5
TSDA（我们的方法）	**0.695/ 0.680**	**48.6 / 48.5**	**86.3 / 86.5**	**86.2/ 86.5**	**0.529 / 0.527**	**54.9/ 54.9**	**86.3/ 86.4**	**86.2 /86.5**

打乱（破坏顺序但保留静态结构），以及静态交换（替换空间摘要，同时保留时间轨迹）；如果门控对被扰乱的因子赋予较高权重，则会受到惩罚；软先验则鼓励在 $(c_t - c_s)$ 和 d 较大时，g 也较大。

2.5. 训练目标

令 $\mathcal{L}_{\text{task}}$ 表示分类的交叉熵或回归的均方误差。总体损失为

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \alpha\,\mathcal{L}_{\text{pur}} + \beta\,\mathcal{L}_{\text{decorr}} + \gamma\,\mathcal{L}_{\text{orth}}, \tag{6}$$

其中权重 α, β, γ均为非负数。该目标确保因子一致的跨模态交互、令牌级的因子纯度，以及互补但去相关的时间和空间求和摘要。

3. 实验

3.1. 实验设置

数据集。 我们在标准基准测试上评估 TSDA。CMU-MOSI [23]: 包含 2,199 个对齐了文本、音频和视觉的片段；数据集划分为 1,284/229/686。CMU-MOSEI [24]: 包含来自 1,000 位说话者的 23,453 段视频；数据集划分为 16,326/1,871/4,659。情感标签范围从 -3 到 $+3$。**评估指标**。 我们按照 EMOE [10]报告 ACC-2、ACC-7、宏平均 F1 和平均绝对误差。**实现**。 模型使用 PyTorch 实现，并在 NVIDIA A100 GPU 上训练。我们使用 Adam，权重衰减为 1×10^{-5}，批次大小为 16，最多训练 50 轮，并采用早停。我们采用五折交叉验证。

3.2. 与最先进方法的比较

表1展示了在对齐和未对齐设置下，模型在 CMU-MOSI 和 CMU-MOSEI 上的结果。TSDA 在两个数据集的所有指标上均取得最佳表现。在 MOSI 上，TSDA 超越了最强基线方法 EMOE，将平均绝对误差从 0.710/0.697 降至 0.695/0.680，并使 ACC$_7$, ACC$_2$, 和 F1分数平均提升约 1%。在 MOSEI 上，TSDA 的平均绝对误差最低，为 0.529/0.527，同时准确率和 F1分数最高，持续优于以往的解耦方法和基于融合的方法。

表 2。TSDA 在基准测试上的消融实验。

模型	CMU-MOSI 平均绝对误差(↓)	CMU-MOSI ACC$_7$ (%)	CMU-MOSEI 平均绝对误差(↓)	CMU-MOSEI ACC$_7$ (%)
TSDA（我们的方法）	**0.680**	**48.5**	**0.527**	**54.9**
模态的重要性				
不使用 L。	1.015	35.2	0.835	39.6
不使用 A。	0.818	44.8	0.624	51.0
不使用 V。	0.842	44.4	0.647	50.4
表征的重要性				
不使用时间	0.726	46.0	0.552	52.5
无空间	0.716	46.8	0.546	53.0
无时空解耦	0.731	45.7	0.555	52.2
无 FCCA	0.728	45.5	0.552	51.9
不同融合机制				
Sum	0.711	47.8	0.537	54.0
拼接	0.714	47.5	0.538	53.9
GR（我们的方法）	0.703	47.9	0.536	54.1
正则化的重要性				
无 $\mathcal{L}_{\text{pur}}$	0.722	46.5	0.548	52.9
无 $\mathcal{L}_{\text{decorr}}$	0.713	46.9	0.541	53.3
无 $\mathcal{L}_{\text{orth}}$	0.714	47.1	0.542	53.4
交叉熵损失	0.736	45.7	0.556	52.0

这些改进得益于时间与空间解耦，它能避免静态主导偏差；因子一致性对齐可确保时间与空间交互清晰；门控重耦合则能自适应地平衡各因子的可靠性。这些机制共同带来了更忠实的情感表征，并使模型在对齐和未对齐条件下均获得稳定提升，凸显了 TSDA 相较于现有方法的稳健性和有效性。

3.3 消融研究

我们通过消融实验评估 TSDA 中各个组件的必要性。MOSI 和 MOSEI 上的结果见表 2，涵盖模态贡献、表示解耦、融合策略和正则化设计。

模态的重要性。排除任何一种模态都会导致性能大幅

下降。移除语言会导致最大的性能退化，这证实了词汇极性是情感决策的依据。排除音频或视觉会造成较小但仍显著的性能下降，这表明这些信息流为稳健的多模态推理提供了互补线索。

表征的重要性。移除时间流导致的性能下降比移除空间流更明显，这与以下假设一致：情感通常由短暂的韵律和视觉动态承载。进一步取消时间—空间解耦会降低准确率，表明因子分离本身至关重要。即使保留了解耦，移除 FCCA 仍会损害性能，这说明显式的因子一致性对齐对于防止跨因子干扰和静态主导至关重要。

不同的融合机制。简单求和或拼接融合的效果较差，因为这些操作无法自适应地平衡各流。所提出的门控重耦合根据可靠性对因子加权，从而提升了结果，但性能仍不及完整模型。这表明，若没有 FCCA 的结构约束，仅靠自适应重新加权并不足够；这两种机制缺一不可。

正则化的重要性。移除因子纯度监督 $\mathcal{L}_{\text{pur}}$ 会导致 MOSI 和 MOSEI 上的性能大幅下降，说明它对于保持时间因子的纯净性起着核心作用。移除去相关 $\mathcal{L}_{\text{decorr}}$ 或正交性 $\mathcal{L}_{\text{orth}}$ 也会削弱性能，不过影响较轻，且在两个数据集上相近。仅使用交叉熵进行训练的总体结果最差，这凸显了纯度、去相关和正交性相辅相成，能够共同确保因子分离与稳定性。

这些消融实验验证了 TSDA 的设计动机：解耦提供了相互区分的时间和空间主体，FCCA 确保了一致的对齐，而门控重耦合则对二者进行自适应整合。三者共同解决了时空不对称性，并在各项基准测试中带来了稳定的改进。

3.4. 可视化分析

定性分析。为证明 TSDA 的有效性，我们使用 t-SNE 对 MOSI 测试集嵌入表示进行可视化，以考察特征分布（图2）。在不使用时间-空间解耦的情况下，数据点较为分散，极性相互重叠，且强度缺乏有序性。解耦后移除 FCCA 虽然带来了部分提升，但时间流和空间流仍然相互混杂，静态线索占据主导地位。加入 GR 后，极性连续谱更加清晰，并通过降低不可靠因子的权重抑制了孤立簇，但仍存在轻微的跨因子泄漏。完整的 TSDA 产生了最紧凑、单调的梯度，并与情感强度保持一致，这正是回归任务所需的几何结构。该演进过程表明，解耦是必要的，FCCA 可防止跨因子干扰，而 GR 则能够校准逐样本可靠性。

正则化趋势。在 MOSI 和 MOSEI 训练期间，约束 $\mathcal{L}_{\text{pur}}$、$\mathcal{L}_{\text{decorr}}$ 和 $mathcal L_{\text{orth}}$ 几乎单调下降，并平稳收敛。如图3所示，用于词元识别的因子纯度分类器稳步提升，随后趋于稳定，这表明 FCCA 学习到了因子区分性表示。同时，摘要之间的余弦依赖性 $\cos^2(Z^{(t)}, Z^{(s)})$ 随着 $\mathcal{L}_{\text{decorr}}$ 同步下降，损失 $\mathcal{L}_{\text{orth}}$ 也持续下降且没有振荡。这些趋势表明，TSDA 能够实现因子分离、限制跨因子泄漏，并在面向最终任务的过程中保持稳定优化。

敏感性敏感性分析。我们通过扫描 l oss

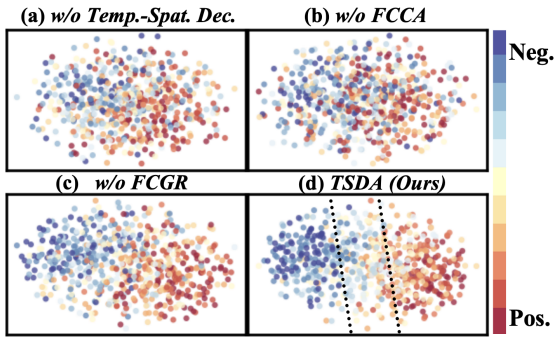

图2。MOSI 上的 t-SNE 可视化。红色表示更强的正向情感。TSDA 得到最佳结构。

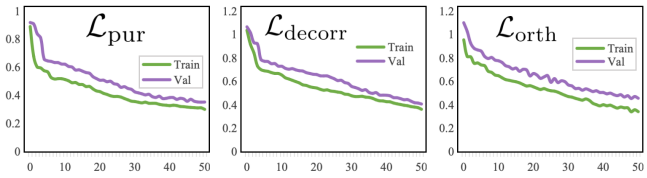

图3。MOSI 训练期间的正则化曲线。MOSEI 上也观察到了类似趋势。

公式（6）中的权重。对于每个超参数 α、β和 γ，我们在固定其他参数为主要实验所用默认值的情况下，在一定范围内改变其取值，并在对齐设置和未对齐设置下分别于 MOSI 和 MOSEI 上进行评估。图4报告了随着超参数变化而变化的平均绝对误差和 ACC_7。在所有配置中，曲线仅表现出轻微波动，表明 TSDA 对 α、β和 γ的具体取值不敏感，同时也说明观察到的性能提升源于架构，而非精细调参。

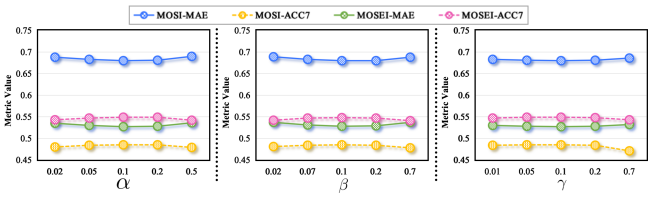

图4。TSDA 对基准测试中 α、β 和 γ 的敏感性。不同取值下，性能保持稳定。

4. 结论

本文提出了TSDA，即一种新颖的行动前时空解耦框架，通过在交互之前解耦时间动态与空间上下文，解决多模态情感分析中的时空异质性。因子一致性交叉模态对齐确保了不同因子之间的可靠对齐，而门控重耦合则自适应地平衡它们的贡献。大量实验表明，该设计能够有效缓解时空不对称性，并在基准数据集上带来稳定的性能提升。未来，我们计划将TSDA拓展至人机交互领域。

5. 参考文献

[1]Chunlei Meng、Wei Lin、Bowen Liu、Hongda Zhang、Zhongxue Gan和Chun Ouyang，“Rts-vit：用于图像分类的实时共享视觉变换器”，*IEEE* 生物医学与健康信息学期刊，2025年。[2]Paul Waligora、Muhammad Haseeb Aslam、Muham- mad Osama Zeeshan、Soufiane Belharbi、Alessandro Lameiras Koerich、Marco Pedersoli、Simon Bacon和Eric Granger，“自然场景中的联合多模态变换器情感识别”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2024年，第4625–4635页。[3] Yifan Chen、Kuntao Li、Weixing Mai、Qiaofeng Wu、Yun Xue和Fenghuan Li，“D2r：用于多模态情感检测的双分支动态路由网络”，载于*2024*年自然语言处理实证方法会议论文集，2024年。[4] Lingyong Fang、Gongshen Liu和Ru Zhang，“用于情感分析的多粒度多模态交互网络”，载于*ICASSP 2024-2024 IEEE*国际声学、语音与信号处理会议（*ICASSP*），2024年。[5] Wenmeng Yu、Hua Xu、Ziqi Yuan和Jiele Wu，“利用自监督多任务学习学习模态特定表示以进行多模态情感分析”，载于*AAAI*人工智能会议论文集，2021年，第10790–10797页。[6] Sheng Wu、Dongxiao He、Xiaobao Wang、Longbiao Wang和Jianwu Dang，“通过视觉-音频内容的文本情感描述丰富多模态情感分析”，载于*AAAI*人工智能会议论文集，2025年，第39卷。[7] Chunlei Meng、Jiacheng Yang、Wei Lin、Bowen Liu、Hongda Zhang、Zhongxue Gan等，“Cta-net：用于改进多尺度特征提取的CNN-变换器聚合网络”，*arXiv*预印本 *arXiv:2410.11428*，2024年。[8] Chunlei Meng、Jiacheng Yang、Wei Lin、Linqiang Hu、Bowen Liu、Zhuo Zou、LiDa Xu、Zhongxue Gan和Chun Ouyang，“用于表面缺陷分类的多粒度师生联合表示学习”，*Journal of Industrial Infor- mationIntegration*，第48卷，第100958页，2025年。[9] Yong Li、Yuanzhi Wang和Zhen Cui，“用于情感识别的解耦多模态蒸馏”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2023年，第6631–6640页。[10] Yiyang Fang、Wenke Huang、Guancheng Wan、Kehua Su和Mang Ye，“Emoe：模态特定增强动态情感专家”，载于计算机视觉与模式识别会议论文集，2025年。[11] Jiuding Yang、Yakun Yu、Di Niu、Weidong Guo和Yu Xu，“Confede：面向多模态情感分析的对比特征分解”，载于计算语言学协会第*61*届年会论文集，2023年，第7617–7630页。[12] Devamanyu Hazarika、Roger Zimmermann和Soujanya Po- ria，“Misa：面向多模态情感分析的模态不变及模态特定表示”，载于第*28*届*ACM*多媒体国际会议论文集，2020年，第1122–1131页。

[13] Dingkang Yang、Shuai Huang、Haopeng Kuang、Yangtao Du 和 Lihua Zhang，《用于多模态情感识别的解耦表示学习》，载于第*30*届*ACM*多媒体国际会议论文集，2022年，第1642–1651页。[14] Haoqin Sun、Shiwan Zhao、Xuechen Wang、Wenjia Zeng、Yong Chen 和 Yong Qin，《用于多模态情感识别的细粒度解耦表示学习》，载于*ICASSP 2024-2024 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2024年，第11051–11055页。[15] Pan Wang、Qiang Zhou、Yawen Wu、Tianlong Chen 和 Jingtong Hu，《DLF：解耦语言聚焦的多模态情感分析》，载于*AAAI*人工智能会议论文集，2025年，第39卷，第21180–21188页。[16] Dingkang Yang、Mingcheng Li、Dongling Xiao、Yang Liu、Kun Yang、Zhaoyu Chen、Yuzheng Wang、Peng Zhai、Ke Li 和 Lihua Zhang，《通过偏差净化实现多模态情感分析去偏》，载于欧洲计算机视觉会议。施普林格，2024年，第464–481页。[17] Zhengze Xu、Dongyue Wu、Changqian Yu、Xiangxiang Chu、Nong Sang 和 Changxin Gao，《SCTNet：结合Transformer语义信息的单分支卷积神经网络，用于实时分割》，载于*AAAI*人工智能会议论文集，2024年。[18] Jinhao Lin、Yifei Wang、Yanwu Xu 和 Qi Liu，《Semi-IIN：用于多模态情感分析的半监督模态内外交互学习网络》，载于*AAAI*人工智能会议论文集，2025年，第39卷，第1411–1419页。[19]Zirun Guo、Tao Jin、Jingyuan Chen 和 Zhou Zhao，《用于增强多模态学习的分类器引导的梯度调制》，神经信息处理系统进展，第37卷，第133328–133344页，2024年。[20] Zhun Liu、Ying Shen、Varun Bharadhwaj Lakshminarasimhan、Paul Pu Liang、Amir Zadeh 和 Louis-Philippe Morency，《利用模态特定因子实现高效低秩多模态融合》，*arXiv*预印本 *arXiv:1806.00064*，2018年。[21] Yao-Hung Hubert Tsai、Shaojie Bai、Paul Pu Liang、J Zico Kolter、Louis-Philippe Morency 和 Ruslan Salakhutdinov，《用于未对齐的多模态语言序列的多模态Transformer》，载于计算语言学协会年会论文集，2019年，第6558页。[22] 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*年自然语言处理经验方法会议论文集，2017年，第1103–1114页。[23] Amir Zadeh、Rowan Zellers、Eli Pincus 和 Louis-Philippe Morency，《视频中的多模态情感强度分析：面部表情动作与言语信息》，*IEEE*智能系统，第82–88页，2016年。[24] AmirAli Bagher Zadeh、Paul Pu Liang、Soujanya Poria、Erik Cambria 和 Louis-Philippe Morency，《自然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，2018年，第2236–2246页。