中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-15

论文引用格式:Liu Jinbai, Liu Xin, Peng Shujuan. Multimodal sarcasm detection driven by local semantic conflict awareness and modality rebalancing[J/OL]. Journal of Image and Graphics, XXXX:1-15. DOI: 10.11834/jig.260500. (刘金柏, 柳欣, 彭淑娟. 局部语义冲突感知与模态重平衡驱动的多模态讽刺检测[J/OL]. 中国图象图形学报,XXXX:1-15. DOI: 10.11834/jig.260500.)[DOI:10.11834/jig.260500]

局部语义冲突感知与模态重平衡驱动的多模态讽刺检测

刘金柏,柳欣,彭淑娟*

华侨大学计算机科学与技术学院,厦门　361021

摘　要: **目的**　多模态讽刺检测需联合分析图像与文本信息,识别字面含义与真实态度相背离的讽刺表达。现有方法多基于全局表示或词元级特征进行跨模态交互,难以充分刻画短语级局部语义冲突;同时,联合训练中容易出现文本模态主导,导致视觉信息学习不足。针对上述问题,提出一种结合局部语义冲突建模与模态重平衡的多模态讽刺检测模型。**方法**　首先通过局部语义冲突感知模块将细粒度局部特征聚合为语义连贯的文本短语与视觉短语,并采用去相关正则减少不同短语表示间的语义冗余。其次,联合建模图文短语间的语义相关性与差异性,抑制无关短语对产生的差异干扰,并聚合具有判别性的局部冲突线索。最后,动态原型引导的模态重平衡策略分别维护文本与视觉类别原型,根据原型分类难度动态调节两模态的监督强度,以缓解文本主导并促进视觉分支学习。**结果**　在MMSD1.0(Multimodal Sarcasm Detection Dataset 1.0)和MMSD2.0数据集上,所提模型的准确率分别达到94.14%和86.20%,F1值分别达到92.78%和85.27%,均优于现有对比方法。**结论**　所提模型能够有效增强多模态讽刺检测中局部图文冲突信息的建模与利用,并缓解联合训练中的模态学习失衡,从而提升模型对复杂图文讽刺语义的判别能力。代码开源地址为:https://doi.org/10.57760/sciencedb.j00240.0009p。

关键词: 多模态讽刺检测;局部语义冲突;短语级表示;跨模态不一致;槽注意力;模态重平衡

Multimodal sarcasm detection driven by local semantic conflict awareness and modality rebalancing

Liu Jinbai, Liu Xin, Peng Shujuan*

School of Computer Science and Technology, Huaqiao University, Xiamen 361021, China

Abstract: **Objective** Multimodal sarcasm detection identifies discrepancies between literal content and underlying attitudes by jointly analyzing images and text, with decisive cues often concentrated in local textual expressions and visual regions. Existing methods mainly model image-text relations through global fusion or direct interaction between text tokens and image patches. Global interaction may weaken subtle local semantic differences after feature aggregation, whereas token-level interaction can produce fragmented correspondences that are difficult to organize into semantically coherent units. Joint multimodal training suffers from modality imbalance. Text often provides more direct and stable discriminative

收稿日期:2026-08-10;**修回日期**:2026-11-02

*通信作者:彭淑娟　pshujuan@hqu.edu.cn

基金项目:国家自然科学基金项目(62576143, 62476103);厦门市科技计划项目(3502Z20251019);华侨大学中央高校基本科研业务费专项资金资助项目(2026HQYJ01)

Supported by: National Natural Science Foundation of China (62576143, 62476103); Xiamen Science and Technology Project (3502Z20251019); Fundamental Research Funds for the Central Universities of Huaqiao University (2026HQYJ01)

© 中国图象图形学报版权所有

cues than images, causing optimization to rely excessively on textual information and leaving the visual branch insufficiently trained. To address these issues, a multimodal sarcasm detection model is proposed to improve phrase-level local semantic conflict modeling while dynamically coordinating the learning progress of textual and visual modalities. **Method** Visual and textual encoders first extract global representations together with fine-grained image-patch and text-token sequences. The global representations are projected and aligned using bidirectional contrastive learning to strengthen sample-level image-text semantic correspondence, while the local feature sequences are retained for phrase-level modeling. Slot attention is then applied separately to the two modalities. Multiple learnable slots compete for local features and are iteratively updated, organizing discrete tokens and image patches into a fixed number of semantically coherent textual and visual phrases. A decorrelation regularizer reduces redundancy among local slot representations and encourages complementary phrase features. A cross-modal relevance matrix measures semantic correspondence between textual and visual phrases, whereas a semantic difference matrix describes their relative deviation in the feature space. Because a large semantic difference may also result from semantically unrelated phrases, correlation-gated fusion combines relevance and difference information to suppress irrelevant differences caused by unrelated phrase pairs and retain semantically related yet conflicting local cues. The resulting textual and visual conflict representations are aggregated for sarcasm classification. To alleviate modality imbalance, class prototypes are maintained separately for textual and visual representations and updated with class-wise feature centers through an exponential moving average. Prototype-based classification losses characterize the relative learning difficulty of the two modalities. The modality with the larger prototype classification loss is treated as the slower-learning modality and receives stronger prototype supervision, while the faster-learning modality is regularized during the early training stage to reduce excessive confidence and premature dominance. The overall objective jointly optimizes multimodal classification, global image-text alignment, phrase decorrelation, and prototype-guided modality rebalancing. **Result** Experiments are conducted on the public Multimodal Sarcasm Detection Dataset 1.0 (MMSD1.0) and Multimodal Sarcasm Detection Dataset 2.0 (MMSD2.0) using the official training, validation, and test splits. MMSD2.0 reduces spurious correlations and revises noisy annotations, providing a more challenging benchmark with fewer superficial cues. The proposed model is compared with representative text-only, image-only, attention-based multimodal, and graph-based multimodal methods. Accuracy, precision, recall, and F1 score are used for evaluation. On MMSD1.0, the model achieves an accuracy of 94.14% and an F1 score of 92.78%; on MMSD2.0, it achieves an accuracy of 86.20% and an F1 score of 85.27%. In terms of F1 score, these results exceed the best-performing compared baselines by 0.85 and 1.16 percentage points on MMSD1.0 and MMSD2.0, respectively. Ablation experiments verify the contributions of the major components. Removing local semantic conflict-aware phrase aggregation reduces F1 by 2.87 and 3.79 percentage points on MMSD1.0 and MMSD2.0, respectively, producing the largest degradation. Removing correlation-gated fusion decreases F1 by 1.02 and 1.73 percentage points, respectively, confirming the importance of using semantic relevance to filter irrelevant differences. Removing dynamic modality rebalancing decreases F1 by 1.75 percentage points on MMSD1.0 and 2.65 percentage points on MMSD2.0, indicating the benefit of coordinating modality-specific learning states. Comparison among different phrase aggregation strategies on MMSD2.0 shows that the adopted slot-based aggregation reaches an F1 score of 85.27%, outperforming cross-attention aggregation, the strongest alternative, by 0.71 percentage points. Parameter analysis indicates that moderate settings for the number of phrase slots and iterative updates provide a better balance between semantic coverage and representation concentration. Training-process analysis shows that modality rebalancing improves visual discriminative learning while maintaining stable learning in the textual branch. **Conclusion** The results demonstrate that organizing fragmented token- and patch-level features into semantically coherent phrase units provides a basis for modeling sparse local image-text conflicts. Jointly considering cross-modal relevance and semantic difference helps distinguish informative conflicts from differences caused mainly by semantic irrelevance, thereby improving the utilization of local conflict cues. The results also indicate that modality imbalance is closely related to the joint optimization process; dynamically coordinating modality-specific learning progress can reduce excessive dependence on textual cues and provide more effective learning signals for visual information. Future work can explore adaptive phrase aggregation with variable numbers of semantic units and more stable estimation of modality learning states to improve adaptability to diverse multimodal data distributions. Code is available athttps://doi.org/10.57760/sciencedb.j00240.0009p.

© 中国图象图形学报版权所有

Key words: multimodal sarcasm detection; local semantic conflict; phrase-level representation; cross-modal incongruity; slot attention; modality rebalancing

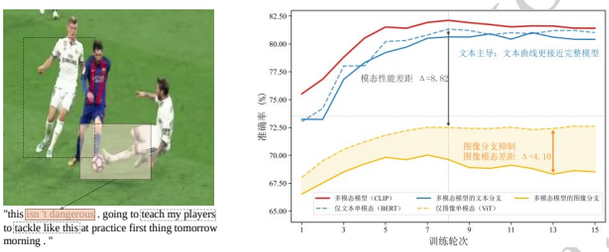
论文引用格式:DOI:10. 11834/jig. 260500

0 引　言

讽刺是常见的修辞表达,其核心在于字面含义与真实态度之间存在显著偏离(Farabi 等,2024),理解该类表达需要结合语境、语气及外部常识进行语用推断(Sarsam 等,2020)。因此,讽刺检测任务旨在从文本及相关上下文中识别隐含语义反转,是自然语言理解与情感计算领域中一项基础且具挑战性的研究问题。提升讽刺检测性能有助于缓解情感分析中的语义反转问题,还可提升复杂社交媒体场景下的情感识别与理解能力(陶建华 等,2024),并为意见挖掘、社交媒体内容分析及用户意图理解等实际应用提供可靠的语义基础(Chen 等,2024)。

早期研究主要聚焦文本讽刺检测,通常从情感极性冲突、句法与标点模式以及上下文依赖等角度挖掘讽刺线索(Sarsam 等,2020)。随着社交媒体内容形态的演进,异构模态信息的关联与联合表征逐渐成为多模态理解的重要基础(刘华峰 等,2023),用户也更倾向于使用图文组合表达真实意图(Cai 等,2019),多模态讽刺检测任务由此兴起。其核心在于联合建模图像与文本信息,刻画模态内与模态间的不一致性,据此推断讽刺意图。现有研究主要从跨模态交互融合和图结构关系建模等方向展开,并逐渐向多层级、多视角的不一致建模拓展。前者通过跨模态注意力与交互融合建模图文语义关联及潜在不一致线索(Tian 等,2023);后者将图文细粒度单元及其关系构建为图,通过结构化信息传播刻画跨模态依赖与冲突线索(Zhu 等,2024)。尽管现有方法在跨模态建模方面取得了一定进展,但在局部冲突线索的刻画以及联合训练中的有效利用方面仍存在不足。

第一,现有方法难以刻画短语级局部语义冲突。现有方法多侧重全局融合表示,或在词元与图像块等离散局部单元上建模跨模态不一致。前者依赖整体特征聚合,容易弱化局部语义差异,使细微的讽刺触发线索在全局表示中被掩盖。后者虽能捕捉细粒度关联,但其交互通常停留在离散局部单元层面,难以将局部线索组织为语义连贯的文本短语与图像局部短语,从而限制短语层面的跨模态语义冲突建模。以图 1(a)样本为例,文本短语“isn’t dangerous”表达明确正向判断,图像中运动员高速滑铲所呈现的高风险局部场景可视为对应视觉短语。二者在短语层面形成的冲突构成样本讽刺意图的重要来源。由于缺乏对短语级语义单元的显式聚合与冲突刻画,现有范式难以稳定定位并利用此类关键冲突线索,进而影响细粒度讽刺信息的挖掘。



(a) 短语级别不一致示例　　(b) 模态不平衡现象

((a) phrase-level incongruity; (b) modality imbalance)

图 1　多模态讽刺检测面临的两个关键挑战

Fig. 1　Two key challenges in multimodal sarcasm detection

第二,多模态讽刺检测的联合训练中普遍存在模态不平衡现象(Huang 等,2022),即文本模态在判别中占据主导地位,并对图像模态的优化产生抑制作用。如图 1(b)所示,在多模态模型中,文本分支的性能曲线几乎与完整模型重合,并与纯文本模型高度一致,而图像分支在多模态模型中的性能则明显低于其对应的纯图像模型。这种文本主导、视觉受抑的现象表明,联合建模过程中,文本强模态占据主导地位并抑制视觉弱模态的判别贡献,导致图像线索在多模态表示中被弱化,从而削弱模型对跨模态不一致线索的利用能力。

针对上述问题,提出局部语义冲突感知与模态重平衡驱动的多模态讽刺检测框架,从表征粒度与优化过程两个层面进行改进。首先,提出局部语义冲突感知模块,采用槽注意力(Locatello 等,2020)对词元序列与图像块序列进行竞争式聚合,形成语义连贯的文本短语表示与视觉短语表示。并结合去相关正则降低短语间语义冗余,使局部冲突线索在短

© 中国图象图形学报版权所有

语层面得到更集中的表征。在此基础上，设计相关性门控融合，对图文短语交互加权，抑制无关匹配并突出与局部冲突相关的短语对。其次，为减轻联合训练中的文本主导现象，提出动态原型引导的模态重平衡机制，通过维护文本与图像的类别原型自适应调节两模态的监督信号强度，促进弱模态获得更充分的有效更新并提升其判别贡献。

本文的主要贡献包括3个方面：

(1)提出局部语义冲突感知方法。针对现有方法难以充分刻画短语级局部语义冲突的问题，采用槽注意力对文本词元与图像块进行竞争式迭代聚合，并结合去相关正则减少不同短语表示间的冗余，形成语义连贯的文本短语与视觉短语表示。

(2)设计相关性门控融合策略。针对图文局部差异存在语义无关信息干扰的问题，联合建模图文短语间的语义相关性与差异，并利用相关性约束差异信息，突出语义相关但表达不一致的局部短语对。

(3)提出动态原型引导的模态重平衡机制。针对联合训练中文本模态容易占据主导的问题，分别维护文本与视觉类别原型，根据原型分类难度动态调整两模态监督强度，缓解模态学习不平衡并促进视觉分支学习。

1 相关工作

1.1 多模态讽刺检测

多模态讽刺检测旨在联合建模图文信息，挖掘模态内及跨模态语义不一致，从而识别讽刺意图。围绕图文不一致信息的建模，现有研究主要形成了跨模态交互融合和图结构关系建模两条技术路线，并逐渐向多层级、多视角的细粒度不一致建模拓展。

一类方法以跨模态交互为核心，在连续表示空间中学习图文关联并挖掘潜在不一致线索。Pan 等(2020)引入模态间注意力与共同注意力，同时捕获跨模态矛盾和文本内部冲突。针对固定交互方式适应性不足的问题，Tian 等(2023)通过动态路由与分层共同注意力自适应调整跨模态交互。随后，不一致建模逐渐由单一语义关系拓展至多维信息。Wen 等(2023)分别建模事实语义矛盾与情绪冲突；Qiao 等(2023)通过互增强机制促进局部与全局线索协同；Qin 等(2023)从文本、图像与跨模态交互3个视角联合捕捉讽刺信息。

另一类方法借助图结构显式刻画模态内部及跨模态关系，并通过结构化信息传播进行不一致推理。Liang 等(2021)构建模态内图与跨模态图进行交互式图卷积融合，随后将建模粒度推进至对象—词语层面(Liang 等，2022)。Zhu 等(2024)结合普通图与超图增强高阶关系表达；Wei 等(2024)利用样本间语义关系构建全局信息，辅助讽刺判断。

在上述两类技术路线基础上，近期研究进一步从动态关系、多层级和多视角等方面拓展不一致建模。Guo 等(2025)从词元—图像块、实体—对象及情感等多个视角联合学习跨模态不一致。相关研究表明，多模态讽刺检测已由全局交互逐步发展至词元、图像块、对象和实体等细粒度关系建模。

综上，现有方法已逐步细化跨模态不一致建模，但局部交互仍多基于离散单元，难以形成语义连贯的局部表示。为此，将文本词元与图像块聚合为短语级语义单元，并在短语对层面建模局部语义冲突，以增强关键图文冲突信息的表征与利用。

1.2 多模态平衡学习

在多模态联合学习中，不同模态在信息密度、任务相关性和学习难度上存在差异，容易形成不同的优化速度与判别能力。Peng 等(2022)指出，优势模态容易在联合训练中占据主导，并通过在线梯度调节协调不同模态的学习过程；Huang 等(2022)进一步从模态竞争角度指出，强模态可能抑制弱模态学习，使多模态模型难以充分利用模态间的互补信息。针对该问题，现有研究主要从优化动态调节和单模态判别能力保持两个方面展开。前者根据模态贡献或学习状态动态调整优化强度，Li 等(2023)通过自适应梯度调制改善模态间学习失衡；Fan 等(2023)借助类别原型等稳定参照维持弱模态的判别结构，减轻联合优化中的结构退化与贡献失衡。

综上，现有模态平衡方法多面向通用多模态学习，而多模态讽刺检测依赖图文语义反差，不能简单追求两模态学习强度一致。因此，需要结合任务相关的学习状态动态协调两模态监督。

2 本文模型

如图2所示，整体模型包括多模态特征编码与对齐、局部语义冲突感知、动态原型引导的模态重平衡和相关性门控融合4部分。首先提取图文全局及

© 中国图象图形学报版权所有

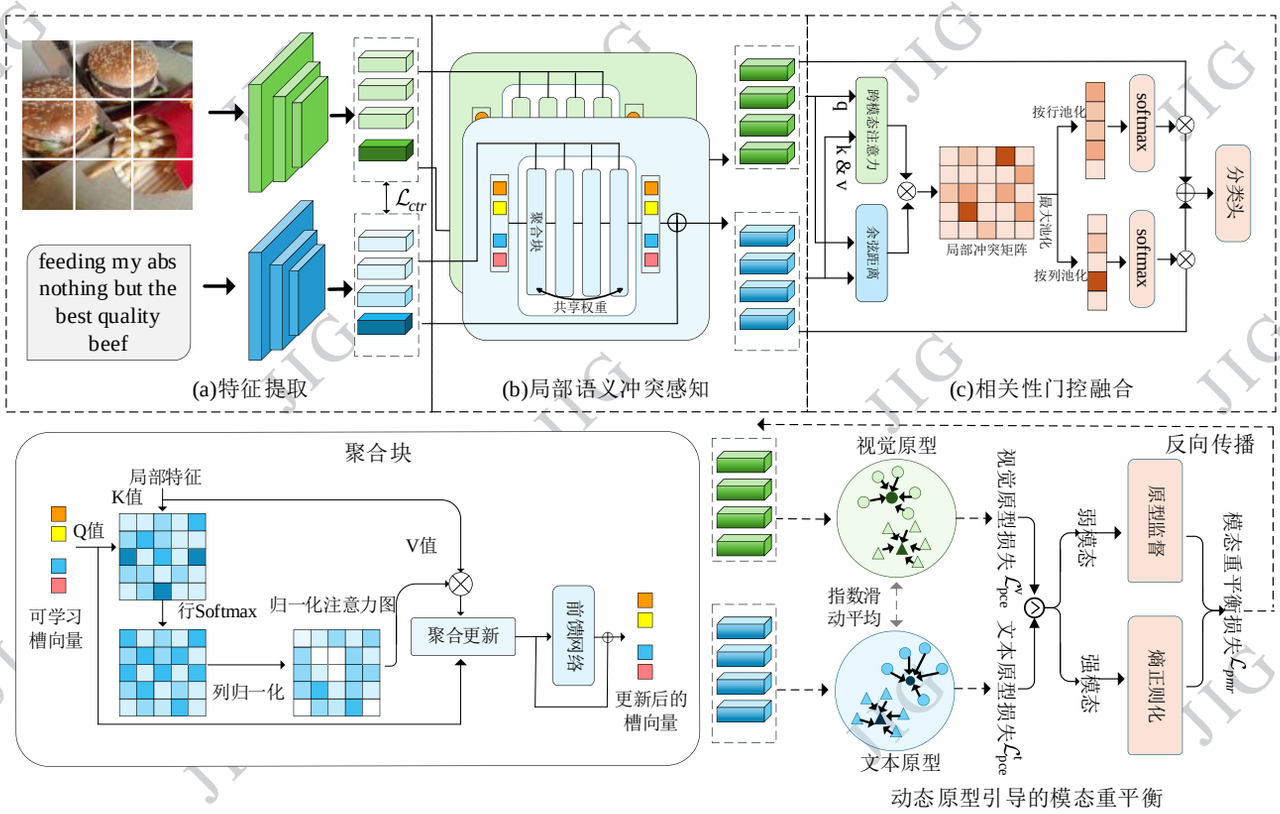
局部特征，并通过全局对比学习增强语义对应；随后，将词元与图像块聚合为短语级语义单元，根据两模态原型分类难度动态调节监督强度，并通过相关性门控融合形成用于讽刺分类的整体表示。

2.1 多模态特征编码与对齐

给定图文样本对$(\boldsymbol{I},\boldsymbol{T})$，其中，$\boldsymbol{I}$和$\boldsymbol{T}$分别表示输入图像及其对应文本。为提取图像与文本中的语义特征，分别采用ViT-B/32（vision Transformer）（Dosovitskiy 等，2021）和RoBERTa-base（robustly optimized BERT approach）（Liu 等，2019）作为视觉编码器和文本编码器。编码得到的图像全局和局部特征分别记为$\boldsymbol{g}^{\mathrm{v}}\in\mathbf{R}^{d}$和$\boldsymbol{H}^{\mathrm{v}}\in\mathbf{R}^{L_{\mathrm{v}}\times d}$，文本全局和局部特征分别记为$\boldsymbol{g}^{\mathrm{t}}\in\mathbf{R}^{d}$和$\boldsymbol{H}^{\mathrm{t}}\in\mathbf{R}^{L_{\mathrm{t}}\times d}$。其中，$d$为特征维度，$L_{\mathrm{v}}$和$L_{\mathrm{t}}$分别为图像块数和词元数。

为将两种模态的全局特征映射至可比较的共享表示空间，分别采用由两层全连接层构成的视觉投影头$\phi_{\mathrm{v}}(\cdot)$和文本投影头$\phi_{\mathrm{t}}(\cdot)$，经投影和ℓ_2归一化后得到单位长度对齐向量$\boldsymbol{z}^{\mathrm{v}}=\mathrm{norm}_2(\phi_{\mathrm{v}}(\boldsymbol{g}^{\mathrm{v}}))$和$\boldsymbol{z}^{\mathrm{t}}=\mathrm{norm}_2(\phi_{\mathrm{t}}(\boldsymbol{g}^{\mathrm{t}}))$。其中，$\mathrm{norm}_2(\cdot)$表示$\ell_2$归一化操作。对于包含$B$个图文样本对的训练批次，记第$i$个样本图像和文本对齐向量分别为$\boldsymbol{z}_i^{\mathrm{v}}$和$\boldsymbol{z}_i^{\mathrm{t}}$。采用双向InfoNCE（information noise contrastive estimation）损失在全局层面对图文表示进行对齐（Radford 等，2021），将配对的图文表示作为正样本，同一批次内不配对的图文表示作为负样本。图像到文本方向的对齐损失定义为：

$$\mathcal{L}_{\mathrm{ctr}}^{\mathrm{v}\to\mathrm{t}}=-\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\mathrm{sim}(\boldsymbol{z}_i^{\mathrm{v}},\boldsymbol{z}_i^{\mathrm{t}})/\tau_{\mathrm{ctr}})}{\sum_{j=1}^{B}\exp(\mathrm{sim}(\boldsymbol{z}_i^{\mathrm{v}},\boldsymbol{z}_j^{\mathrm{t}})/\tau_{\mathrm{ctr}})}\tag{1}$$



((a) Feature extraction; (b) Local semantic conflict awareness; (c) Correlation-gated fusion)

图2 局部语义冲突感知与模态重平衡驱动的多模态讽刺检测框架图

Fig. 2 Overall framework of the proposed model with local semantic conflict awareness and modality rebalancing

式中，j为批次内候选文本索引，$\mathrm{sim}(\cdot,\cdot)$表示余弦相似度，$\tau_{\mathrm{ctr}}$为可学习温度系数，$B$表示批次大小。

同理，得到文本到图像方向的对齐损失$\mathcal{L}_{\mathrm{ctr}}^{\mathrm{t}\to\mathrm{v}}$。最终，对齐损失取平均得到$\mathcal{L}_{\mathrm{ctr}}=(\mathcal{L}_{\mathrm{ctr}}^{\mathrm{v}\to\mathrm{t}}+\mathcal{L}_{\mathrm{ctr}}^{\mathrm{t}\to\mathrm{v}})/2$。该约束在全局层面增强配对图文表示的语义对应关系，为后续短语级冲突建模提供整体语义基础。

2.2 局部语义冲突感知

为将离散的词元与图像块组织为可用于局部冲

© 中国图象图形学报版权所有

突建模的短语级语义单元，采用槽注意力分别对文本和图像局部特征进行多轮迭代聚合。文本与图像分支结构相同但参数独立，各分支内不同轮次的聚合块权重共享（Kim 等，2023）。下文以文本分支为例进行说明。

对于编码得到的文本局部特征 $\boldsymbol{H}^{\mathrm{t}}$ 和全局特征 $\boldsymbol{g}^{\mathrm{t}}$，模块维护 K 个可学习槽向量，其初始表示记为 $\boldsymbol{S}^{(0)} \in \mathbf{R}^{K\times d}$。在第 l 轮迭代中（$l = 1, \cdots, L$），通过多头交叉注意力计算槽向量与各词元之间的注意力分配，并据此聚合词元信息以更新槽表示。设注意力头数为 h，每个注意力头的子空间维度为 $d_h = d/h$。为简化表示，以下省略注意力头索引。经层归一化和线性映射后，查询、键和值分别为：

$$\begin{aligned}\boldsymbol{q}^{(l)} &= \mathrm{LN}(\boldsymbol{S}^{(l-1)})\boldsymbol{W}_{\mathrm{q}}\\ \boldsymbol{k}^{(l)} &= \mathrm{LN}(\boldsymbol{H}^{\mathrm{t}})\boldsymbol{W}_{\mathrm{k}}, \boldsymbol{v}^{(l)} = \mathrm{LN}(\boldsymbol{H}^{\mathrm{t}})\boldsymbol{W}_{\mathrm{v}}\end{aligned} \tag{2}$$

式中，LN（layer normalization）表示层归一化；$\boldsymbol{W}_{\mathrm{q}}, \boldsymbol{W}_{\mathrm{k}}, \boldsymbol{W}_{\mathrm{v}} \in \mathbf{R}^{d\times d_h}$ 分别表示查询、键和值的可学习投影矩阵，$\boldsymbol{q}^{(l)} \in \mathbf{R}^{K\times d_h}$，$\boldsymbol{k}^{(l)}, \boldsymbol{v}^{(l)} \in \mathbf{R}^{L_{\mathrm{t}}\times d_h}$。

随后，在每个注意力头内计算槽与词元之间相似度。与常规注意力沿词元维度进行 softmax 归一化不同，该聚合过程首先沿槽维度进行归一化，使多个槽竞争解释同一词元，得到词元到槽分配权重：

$$e_{n,k}^{(l)} = \frac{\left\langle \boldsymbol{k}_n^{(l)}, \boldsymbol{q}_k^{(l)} \right\rangle}{\sqrt{d_h}}, a_{n,k}^{(l)} = \frac{\exp(e_{n,k}^{(l)})}{\sum_{k'=1}^{K}\exp(e_{n,k'}^{(l)})} \tag{3}$$

式中，n 和 k 分别表示词元索引和槽索引，分母对全部 K 个槽求和；$e_{n,k}^{(l)}$ 和 $a_{n,k}^{(l)}$ 分别表示第 l 轮中第 n 个词元与第 k 个槽之间的相似度得分和分配权重；$\langle\cdot,\cdot\rangle$ 表示向量内积。该竞争机制促使不同槽聚合互补的局部语义内容，从而形成差异化的短语级表示。

随后，沿词元维度对分配权重再次归一化，并利用归一化权重对值向量进行加权平均：

$$D_k^{(l)} = \sum_{n=1}^{L_{\mathrm{t}}} a_{n,k}^{(l)}, \boldsymbol{U}_k^{(l)} = \sum_{n=1}^{L_{\mathrm{t}}} \frac{a_{n,k}^{(l)}}{D_k^{(l)} + \varepsilon}\boldsymbol{v}_n^{(l)} \tag{4}$$

式中，$D_k^{(l)}$ 表示第 k 个槽在词元维度上的权重归一化因子，$\boldsymbol{U}_k^{(l)}$ 表示第 k 个槽聚合得到的特征向量，ε 为防止分母为零的微小常数。将各注意力头的聚合结果拼接后，得到聚合矩阵 $\boldsymbol{U}^{(l)} \in \mathbf{R}^{K\times d}$。

随后通过输出投影与残差连接更新槽表示，并用前馈网络增强非线性表达能力：

$$\begin{aligned}\boldsymbol{S}^{(l)} = \boldsymbol{S}^{(l-1)} &+ \boldsymbol{U}^{(l)}\boldsymbol{W}_{\mathrm{o}}\\ &+\mathrm{FFN}[\mathrm{LN}(\boldsymbol{S}^{(l-1)} + \boldsymbol{U}^{(l)}\boldsymbol{W}_{\mathrm{o}})]\end{aligned} \tag{5}$$

式中，$\boldsymbol{W}_{\mathrm{o}} \in \mathbf{R}^{d\times d}$ 为输出投影矩阵，FFN（feed-forward network）表示两层前馈网络。

经过 L 轮迭代后，得到文本局部槽表示 $\boldsymbol{S}^{(L)}$，其中每一行对应一个文本短语级语义单元。为补充短语表示中的全局语境，将文本全局特征 $\boldsymbol{g}^{t}$ 在槽维度复制并加入各槽表示：

$$\boldsymbol{S}^{t} = \boldsymbol{S}^{(L)} + 1_K(\boldsymbol{g}^{t})^{\top} \tag{6}$$

式中，$1_K \in \mathbf{R}^{K\times 1}$ 为全 1 向量，$\boldsymbol{S}^{\mathrm{t}} \in \mathbf{R}^{K\times d}$ 为最终文本短语表示。图像分支采用结构相同但参数独立的聚合过程，得到视觉短语表示 $\boldsymbol{S}^{\mathrm{v}} \in \mathbf{R}^{K\times d}$。

尽管槽间竞争能够促使不同槽关注不同的局部单元，多轮迭代后仍可能出现不同槽聚合相近语义的现象。为减少槽表示之间的冗余，对加入全局特征前的局部槽表示 $\boldsymbol{S}^{(L)}$ 施加去相关约束。首先对每个槽向量进行层归一化，并沿特征维分别进行 ℓ_2 归一化，得到 $\bar{\boldsymbol{S}} = \mathrm{norm}_2(\mathrm{LN}(\boldsymbol{S}^{(L)}))$，文本分支去相关正则定义为：

$$\mathcal{L}_{\mathrm{dec}}^{\mathrm{t}} = \frac{1}{K^2}\left\| \bar{\boldsymbol{S}}\bar{\boldsymbol{S}}^{\top} - \boldsymbol{I}_K \right\|_{\mathrm{F}}^{2} \tag{7}$$

式中，$\boldsymbol{I}_K \in \mathbf{R}^{K\times K}$ 为 K 阶单位矩阵，$\|\cdot\|_{\mathrm{F}}^{2}$ 表示 Frobenius 范数的平方，矩阵 $\bar{\boldsymbol{S}}\bar{\boldsymbol{S}}^{\top}$ 表示槽向量之间的余弦相似度矩阵，式（7）对其中的非对角元素进行平方惩罚。文本和图像分支分别计算去相关正则，二者之和作为最终去相关损失，即 $\mathcal{L}_{\mathrm{dec}} = \mathcal{L}_{\mathrm{dec}}^{\mathrm{t}} + \mathcal{L}_{\mathrm{dec}}^{\mathrm{v}}$。最终得到的文本短语表示 $\boldsymbol{S}^{\mathrm{t}}$ 与视觉短语表示 $\boldsymbol{S}^{\mathrm{v}}$，作为后续短语对相关性与语义差异建模的输入。

2.3 动态原型引导的模态重平衡

在联合训练中，文本分支通常更快形成判别性表征，使模型倾向依赖文本线索并抑制视觉分支学习。为缓解这一现象，在文本与视觉表征空间内分别维护类别原型，以两模态的原型分类难度刻画相对学习进度，并据此动态调节监督强度。该模块仅作为训练辅助分支，用于协调两模态的优化过程。

具体而言，记 $m \in \{\mathrm{t}, \mathrm{v}\}$ 分别表示文本和图像模态，对于经过局部语义冲突感知模块后的短语集合 $\boldsymbol{S}^{m}$，进一步通过注意力池化得到模态全局表征 $\boldsymbol{f}^{m} \in \mathbf{R}^{d}$，作为后续原型建模与重平衡的输入，从而使重平衡机制建立在短语级语义建模上。

随后，在每个模态空间内维护一组类别原型 $\{\boldsymbol{c}_y^{m}\}_{y\in\{0,1\}}$。对于当前批次内属于类别 y 的样本集合，对其对应的模态级表征进行批内平均，得到该类别在当前批次下的中心表示 $\bar{\boldsymbol{f}}_y^{m}$，并采用指数滑动平

© 中国图象图形学报版权所有

均将该中心融入历史原型，同时对更新后的原型进行单位化以保持尺度一致性：

$$\boldsymbol{c}_y^m \leftarrow \mathrm{norm}_2(\mu \boldsymbol{c}_y^m + (1-\mu)\bar{\boldsymbol{f}}_y^m) \tag{8}$$

式中，μ为动量系数。基于原型集合，对每个样本的模态全局表征计算原型分类概率。记第i个样本的真实类别为$y_i \in \{0,1\}$，采用余弦相似度引入缩放系数γ定义为：

$$p_{i,y}^m = \frac{\exp(\gamma \cdot \cos(\boldsymbol{f}_i^m, \boldsymbol{c}_y^m))}{\sum_{y' \in \{0,1\}} \exp(\gamma \cdot \cos(\boldsymbol{f}_i^m, \boldsymbol{c}_{y'}^m))} \tag{9}$$

式中，$p_{i,y}^m$表示第i个样本在模态m下被判为类别y的原型分类概率，分母对全部候选类别求和。在此基础上，以原型分类交叉熵作为原型判别难度度量，定义模态m原型交叉熵损失为：

$$\mathcal{L}_{\mathrm{pce}}^m = -\frac{1}{B}\sum_{i=1}^{B} \log p_{i,y_i}^m \tag{10}$$

当某一模态尚未形成清晰的类内结构时，其样本与正确类别原型相似度较低、分类分布更分散，从而$\mathcal{L}_{\mathrm{pce}}^m$更大，可反映该模态较慢的学习进度。在每个批次内比较两模态的原型交叉熵损失$\mathcal{L}_{\mathrm{pce}}^{\mathrm{t}}$和$\mathcal{L}_{\mathrm{pce}}^{\mathrm{v}}$，不设置显式判定阈值，将损失较大的模态视为当前阶段的弱模态m_{w}，另一模态视为强模态m_{s}，并据此动态调整两模态的监督强度。类别原型采用指数滑动平均更新，以减弱单批次波动对判定的影响。基于该判定，对弱模态施加原型分类监督，同时在训练早期采用强模态原型分类分布的熵进行适度抑制，避免其过快收敛。具体定义为：

$$\mathcal{H}(\boldsymbol{p}^m) = -\frac{1}{B}\sum_{i=1}^{B}\sum_{y\in\{0,1\}} p_{i,y}^m \log p_{i,y}^m \tag{11}$$

在最小化目标中减去熵项以降低其过度自信。综合上述，模态重平衡损失定义为：

$$\mathcal{L}_{\mathrm{pmr}} = \mathcal{L}_{\mathrm{pce}}^{m_{\mathrm{w}}} - \lambda_{\mathrm{ent}}\mathcal{H}(\boldsymbol{p}^{m_{\mathrm{s}}}) \tag{12}$$

式中，λ_{ent}为强模态熵正则权重。第一项增强弱模态的原型判别学习，第二项在训练早期约束强模态的过度自信，从而动态协调两模态的训练进度。

2.4 相关性门控融合

获得短语级文本表示$\boldsymbol{S}^{\mathrm{t}}$与视觉短语表示$\boldsymbol{S}^{\mathrm{v}}$后，在短语对层面建模局部冲突。考虑到语义无关短语对可能产生较大差异，采用跨模态相关性对语义差异加权，突出语义相关但表达不一致的短语对。

具体而言，以文本短语表示为查询，以视觉短语表示为键和值，通过多头交叉注意力建模图文短语间的相关性(Li 等，2022)。将各注意力头的权重取平均，得到语义相关性矩阵$\boldsymbol{A}^{\mathrm{tv}} \in \mathbf{R}^{K\times K}$。其中，$A_{ij}^{\mathrm{tv}}$表示第$i$个文本短语与第$j$个视觉短语之间的关联强度。记$\boldsymbol{s}_i^{\mathrm{t}}, \boldsymbol{s}_j^{\mathrm{v}} \in \mathbf{R}^d$分别为相应的文本短语表示和视觉短语表示，并采用余弦距离构造语义差异矩阵$\boldsymbol{C} \in \mathbf{R}^{K\times K}$，其中$C_{ij} = 1 - \cos(\boldsymbol{s}_i^{\mathrm{t}}, \boldsymbol{s}_j^{\mathrm{v}})$。最后，将相关性与语义差异逐元素结合，得到局部冲突矩阵：

$$\boldsymbol{M}^{\mathrm{inc}} = \boldsymbol{A}^{\mathrm{tv}} \odot \boldsymbol{C} \in \mathbf{R}^{K\times K} \tag{13}$$

式中，$\odot$表示逐元素乘积。全局InfoNCE约束负责样本级对齐，式(13)利用短语间相关性对语义差异加权，二者分别作用于全局对齐和局部冲突建模。

为提取判别相关的冲突信息，沿视觉短语维对局部冲突矩阵$\boldsymbol{M}^{\mathrm{inc}}$进行最大池化，为每个文本短语保留最大冲突得分。随后，通过softmax将各文本短语的冲突得分归一化为权重，并对文本短语表示进行加权融合，得到文本侧冲突表示：

$$\boldsymbol{r}^{\mathrm{t}} = \mathrm{softmax}[\mathcal{P}_v(\boldsymbol{M}^{\mathrm{inc}})]^{\top}\boldsymbol{S}^{\mathrm{t}} \tag{14}$$

式中，$\mathcal{P}_v$表示沿视觉短语维进行最大池化。同理，沿文本短语维对$\boldsymbol{M}^{\mathrm{inc}}$进行最大池化，并利用所得权重对视觉短语表示进行加权融合，得到视觉侧冲突表示$\boldsymbol{r}^{\mathrm{v}}$。随后将二者拼接并进行融合映射：

$$\boldsymbol{Z}^{\mathrm{inc}} = \psi_{\mathrm{f}}(\boldsymbol{r}^{\mathrm{t}} \oplus \boldsymbol{r}^{\mathrm{v}}) \tag{15}$$

式中，$\oplus$表示特征拼接，ψ_{f}表示由线性层、层归一化和GELU(Gaussian error linear unit)激活组成的融合映射。$\boldsymbol{Z}^{\mathrm{inc}}$为用于讽刺分类的整体冲突表示。

随后，将$\boldsymbol{Z}^{\mathrm{inc}}$输入分类头，经softmax得到两类预测概率$\hat{\boldsymbol{p}}_i = \mathrm{softmax}(\mathrm{Classifier}(\boldsymbol{Z}_i^{\mathrm{inc}}))$，模型根据真实标签与预测概率计算交叉熵损失：

$$\mathcal{L}_{\mathrm{cls}} = -\frac{1}{B}\sum_{i=1}^{B}\log \hat{p}_{i,y_i} \tag{16}$$

式中，$y_i \in \{0,1\}$为真实标签，$\hat{p}_{i,y_i}$为第i个样本在真实类别上的预测概率。

最终，模型以多模态讽刺分类为主要优化目标，并联合全局图文对齐、短语去相关和模态重平衡约束进行端到端训练。总体损失定义为：

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{ctr}}\mathcal{L}_{\mathrm{ctr}} + \lambda_{\mathrm{dec}}\mathcal{L}_{\mathrm{dec}} + \lambda_{\mathrm{pmr}}\mathcal{L}_{\mathrm{pmr}} \tag{17}$$

式中，λ_{ctr}、λ_{dec}、λ_{pmr}分别为全局对齐、去相关和模态重平衡损失的权重。训练过程中，$\mathcal{L}_{\mathrm{cls}}$用于优化讽刺分类，其他损失作为辅助约束共同参与训练。

© 中国图象图形学报版权所有

3 实验及结果分析

3.1 数据集

为评估所提模型性能，选取MMSD1.0（Cai等，2019）与MMSD2.0（Qin等，2023）两个基准数据集进行实验。每个样本由图像文本对构成，并标注为讽刺或非讽刺。MMSD2.0进一步移除误导性图文信号并修正部分噪声标注，数据分布更具挑战性。具体统计信息见表1。

3.2 实验设置

采用Accuracy、Precision、Recall和F1评价模型性能，其中讽刺样本为正类，并以Accuracy和F1作为主要评价指标。本文实验结果均为5次独立实验的平均值。图像与文本特征维度均设为$d = 768$。局部语义冲突感知模块中，文本与图像模态的短语槽位数量均设为$K = 4$，聚合迭代轮次设定为$L = 4$。模态重平衡模块中，类别原型更新动量系数μ设为0.99，缩放因子γ为20.0，模态熵正则权重$\lambda_{ent} = 0.1$，熵正则在前5个训练轮次启用。联合损失中，$\lambda_{pmr} = 0.5$、$\lambda_{ctr} = 0.2$、$\lambda_{dec} = 0.05$。训练采用AdamW（adam with decoupled weight decay）优化器，视觉与文本编码器均参与端到端微调，编码器学习率为2×10^{-5}，其余模块为1×10^{-4}，权重衰减系数为2×10^{-2}，批大小为32，训练15轮。实验基于PyTorch 1.12.0，在NVIDIA GeForce RTX 3090 GPU上完成。

表1 MMSD1.0与MMSD2.0数据集统计信息

Table 1 Statistics of the MMSD1.0 and MMSD2.0 datasets

数据集		样本总数	讽刺样本	非讽刺样本
MMSD1.0	训练集	19816	8642	11174
	验证集	2410	959	1451
	测试集	2409	959	1450
MMSD2.0	训练集	19816	9576	10240
	验证集	2410	1042	1368
	测试集	2409	1037	1372

3.3 对比模型

为验证所提模型在多模态讽刺检测任务中的有效性，在MMSD1.0与MMSD2.0上与单模态及多模态方法进行比较。单模态方面，选取ResNet（residual network）（He等，2016）与ViT（Dosovitskiy等，2021）作为视觉基线。TextCNN（text convolutional neural network）（Kim，2014）、Bi-LSTM（bidirectional long short term memory）（Graves和Schmidhuber，2005）与SMSD（self-matching sarcasm detection model）（Xiong等，2019）作为文本基线。

多模态方面，选取不同跨模态建模策略的代表性方法进行比较，包括HFM（hierarchical fusion model）（Cai等，2019）、Att-BERT（attention BERT）（Pan等，2020）、InCrossMGs（in-modal and cross-modal graphs）（Liang等，2021）、CMGCN（cross-modal graph convolutional network）（Liang等，2022）、DIP（dual incongruity perceiving）（Wen等，2023）、DynRT-Net（dynamic routing Transformer network）（Tian等，2023）、MILNet（mutual-enhanced incongruity learning network）（Qiao等，2023）、MV-CLIP（multi-view CLIP）（Qin等，2023）、DGLF（dual graph-based learning framework）（Zhu等，2024）、G^{2}SAM（graph-based global semantic awareness method）（Wei等，2024）、MoBA（mixture of bi-directional adapter）（Xie等，2024）、HIAN（hybrid interactive attention network）（Bao等，2025）和MICL（multimodal incongruities via contrastive learning）（Guo等，2025）。这些方法涵盖跨模态交互融合、图结构关系建模以及多层级和多视角不一致学习等代表性方法。

3.4 实验结果分析

表2列出了所提方法与已有方法在MMSD1.0和MMSD2.0上的性能比较结果，其中部分基线结果引自Qin等（2023），其余结果来自对应文献。观察得到以下结论：

（1）本文方法在两个数据集上的Accuracy和F1均取得最优结果。在MMSD1.0上，F1值达到92.78%，较DynRT-Net和MICL分别提升0.85和2.45个百分点；在去除伪相关线索、数据分布更复杂的MMSD2.0上，F1值达到85.27%，较最佳基线MoBA提升1.16个百分点。结果表明，所提模型在两个基准数据集上均保持领先性能。

（2）跨模态信息能够提升讽刺检测性能，但两模态间存在明显差异。由表2可见，多数多模态基线优于单模态模型，说明异构模态信息能够补充单一模态的语义表达，丰富模型对复杂讽刺语境的理解。

© 中国图象图形学报版权所有

表2　不同算法在两个数据集上的分类表现对比

Table 2　Classification performance of different methods on the two datasets

/%

模态	模型	MMSD1.0				MMSD2.0			
		准确率	精确率	召回率	F1值	准确率	精确率	召回率	F1值
文本	TextCNN	80.03	74.29	76.39	75.32	71.61	64.62	75.22	69.52
	Bi-LSTM	81.90	76.66	78.42	77.53	72.48	68.02	68.08	68.05
	SMSD	80.90	76.46	75.18	75.82	73.56	68.45	71.55	69.97
图像	ResNet	64.76	54.41	70.80	61.53	65.50	61.17	54.39	57.58
	ViT	67.83	57.93	70.07	63.40	72.02	65.26	74.83	69.72
多模态	HFM	83.44	76.57	84.15	80.18	70.57	64.84	69.05	66.88
	Att-BERT	86.05	78.63	83.31	80.90	80.03	76.28	77.82	77.04
	InCrossMGs	86.10	81.38	84.36	82.84	–	–	–	–
	CMGCN	87.55	83.63	84.69	84.16	79.83	75.82	78.01	76.90
	DIP	89.59	87.76	86.58	87.17	80.59	75.52	81.14	78.23
	DynRT-Net	93.59	90.30	**93.62**	91.93	70.37	63.02	75.15	68.55
	MILNet	89.50	85.16	89.16	87.11	–	–	–	–
	MV-CLIP	88.33	82.66	88.65	85.55	85.64	80.33	88.24	84.10
	G^2SAM	90.48	87.95	89.02	88.48	79.43	72.04	85.20	78.07
	HIAN	92.22	88.20	92.40	90.20	–	–	–	–
	MICL	92.08	90.05	90.61	90.33	–	–	–	–
	DGLF	89.01	84.96	89.10	86.98	81.52	77.98	79.23	78.60
	MoBA	88.40	82.04	88.31	84.85	85.22	79.82	88.29	84.11
	本文	**94.14**	**93.41**	92.16	**92.78**	**86.20**	**81.75**	**89.10**	**85.27**

注：加粗字体为每列最优值。

与此同时，文本基线明显强于图像基线，表明讽刺表达在很大程度上依赖文本语义。这种模态强弱差异容易使模型过度依赖文本线索，进而削弱对图像信息的建模与利用。

（3）本文方法在两数据集上表现稳定。与MMSD1.0相比，MMSD2.0减少了伪相关线索并降低表层偏置，任务难度更高。部分方法性能出现不同程度下降。所提方法在该数据集上仍取得最优结果，说明其在更严格的数据条件下仍保持较好的检测性能。

3.5　消融实验

3.5.1　消融设置

为验证所提方法中每个组件的有效性，分别去除各组件进行消融分析，模型的变体如下：

（1）去除相关性门控融合：移除相关性感知的门控融合结构，将图文短语表示直接拼接后输入多层感知机进行分类。

（2）去除动态原型引导的模态重平衡：不使用重平衡机制调节图文模态不平衡。

（3）去除局部语义冲突感知：不使用局部语义冲突感知构建固定规模的短语级语义单元。

（4）去除全局对比学习：不使用编码阶段的全局对比学习对图文全局特征进行显式语义对齐。

（5）移除去相关正则：保留短语聚合结构，但不使用去相关。

3.5.2　结果分析

消融实验结果如表3所示，得到以下结论：

（1）移除任一组件均引起性能下降，说明各模块

© 中国图象图形学报版权所有

对整体性能均有贡献。这表明模型性能提升并非依赖单一模块,而是由多个模块协同支撑。不同组件对模型性能的贡献存在差异。由表3可见,去除局部语义冲突感知后两数据集性能下降幅度最大,说明短语级局部冲突建模是性能提升的关键;模态重平衡与相关性门控融合次之,表明动态协调模态学习进度以及筛选有效图文冲突同样具有重要作用;去相关正则与全局对比学习带来的增益相对较小,主要发挥辅助作用。

表3 消融实验结果

Table 3 Ablation study results

/%

数据集	模型	准确率/%	精确率/%	召回率/%	F1值/%
MMSD1.0	去除局部语义冲突感知	91.43	89.67	90.15	89.91
	去除模态重平衡	92.38	90.94	91.12	91.03
	去除相关性门控融合	92.95	91.68	91.85	91.76
	去除去相关正则	93.18	92.05	91.93	91.99
	去除全局对比学习	93.54	92.57	**92.24**	92.40
	本文	**94.14**	**93.41**	92.16	**92.78**
MMSD2.0	去除局部语义冲突感知	82.87	77.06	86.43	81.48
	去除模态重平衡	83.84	78.53	87.16	82.62
	去除相关性门控融合	84.72	79.65	87.84	83.54
	去除去相关正则	85.13	80.12	88.46	84.08
	去除全局对比学习	85.46	80.73	88.62	84.49
	本文	**86.20**	**81.75**	**89.10**	**85.27**

(2)局部语义冲突感知、相关性门控融合与模态重平衡对性能提升起主要作用。去除局部语义冲突感知后模型性能明显下降,表明将词元与图像块组织为短语单元有助于充分利用局部线索。去除相关性门控融合后,模型将图文短语表示直接拼接并分类,性能下降表明在短语层面联合建模图文对应关系与语义差异能够提供有效判别信息。去除模态重平衡同样导致性能下降,说明联合训练中对两模态学习强度进行动态调节,可以缓解训练中对单一模态过度依赖,使另一模态获得稳定优化信号,进而提升整体效果。

3.5.3 不同短语聚合方式比较

为进一步验证短语聚合策略的合理性,在保持短语数量、去相关约束、后续融合及训练设置一致的条件下,比较固定步长池化、跨度注意力、软k-means、跨注意力和槽注意力5种聚合方式。固定步长池化将局部特征按位置等分为4组后进行平均池化;跨度注意力在预设连续区域内学习权重并加权融合;软k-means依据局部特征与4个聚类中心的相似度进行软分配和聚合;跨注意力采用4个可学习查询,通过多头注意力聚合文本词元和图像块特征。除聚合方式外,其余模型结构保持一致,各方法均输出4个短语级语义单元,并在MMSD2.0上进行实验,结果见表4。

从表4可以看出,固定步长池化仅依据位置划分局部特征,缺乏自适应信息选择能力,因此性能相对较低。跨度注意力能够在预设局部范围内突出重要特征,但语义单元边界仍受固定位置限制。软k-means可根据特征相似性自适应分组,但所得分组不一定与讽刺判别目标一致。跨注意力利用多个可学习查询从完整局部特征序列中提取信息,取得较好性能,但不同查询之间缺少显式竞争约束,可能重复关注相近内容。

表4 不同短语聚合方式的性能比较

Table 4 Performance comparison of different phrase aggregation methods

/%

聚合方式	准确率/%	精确率/%	召回率/%	F1值/%
固定步长池化	84.10	78.40	88.86	83.28
跨度注意力聚合	84.62	79.95	87.95	83.76
软k-means聚合	84.86	79.35	**89.42**	84.08
跨注意力聚合	85.48	80.71	88.81	84.56
本文	**86.20**	**81.75**	89.10	**85.27**

相比之下,槽注意力通过槽间竞争分配局部特征,并利用多轮迭代逐步细化槽表示。在采用相同去相关约束的条件下,槽注意力取得最高的准确率、精确率和F1值。相较于跨注意力,准确率和F1值分别提高0.72和0.71个百分点,表明槽间竞争与迭代更新有助于不同槽聚合相对互补的局部语义信息,更适合构建跨模态局部冲突建模的短语级表示。

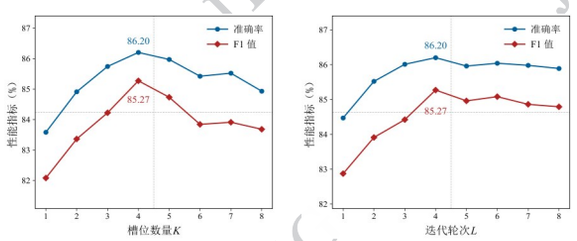
© 中国图象图形学报版权所有

3.6 实验分析

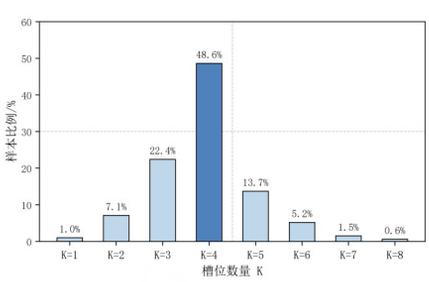
3.6.1 参数敏感性分析

为分析局部语义冲突感知模块中短语表示规模与迭代聚合过程对模型性能的影响，分别对槽位数量K与迭代轮次L进行参数敏感性实验。实验在MMSD2.0数据集上进行，在其余参数保持不变的条件下分别调整K和L，并记录不同设置下的Accuracy和F1，结果如图3所示。

关于槽位数量K，如图3(a)所示，模型性能随K增大呈先升后降趋势。当K较小时，Accuracy和F1均相对较低；随着K增加到4，两项指标达到最优。该结果表明，较少的槽位难以充分组织分散在文本词元和图像块中的局部语义信息，限制了短语级冲突线索的表达。继续增加K时，性能出现回落。说明过多槽位可能使局部语义信息被进一步分散，降低单个短语表示对关键冲突线索的聚合能力。因此，主实验设置$K=4$。



(a) 槽位数量的性能分析　(b) 迭代轮次的性能分析



(c) 样本最优槽位分布

((a) Effect of the number of slots; (b) Effect of the number of iterations; (c) Distribution of optimal slot numbers)

图3 MMSD2.0数据集上的参数分析实验

Fig. 3 Parameter analysis on MMSD2.0

关于迭代轮次L，如图3(b)所示，随着L从1增加至4，Accuracy和F1明显提升，并在$L=4$时达到最优；继续增加迭代轮次后，性能趋于平稳并出现轻微回落。结合前文槽表示的迭代更新过程可以看出，适度增加迭代轮次能够使槽向量在多轮竞争与聚合中逐步细化，从而形成更加稳定的短语级表示；当迭代次数进一步增加时，槽表示逐渐趋于稳定，新增迭代带来的有效信息有限，因此性能不再持续提升。基于上述结果，主实验设置$L=4$。

进一步考虑不同样本中局部语义线索数量和图文交互复杂程度的差异，对固定槽位数量的样本适应性进行分析。在保持其他实验设置一致的条件下，对于每个样本，比较不同槽位数量设置下模型对真实类别的预测概率，并将其中最高概率对应的槽位数量作为该样本的最优槽位配置，以刻画不同槽位规模对样本的相对适配程度。如图3(c)所示，不同样本的最优槽位数量存在一定差异，但整体主要分布在$K=3\sim5$的中等槽位范围。其中，$K=4$对应样本比例最高，为48.6%，$K=3$和$K=5$分别为22.4%和13.7%；当槽位数量进一步减小或增大时，对应样本比例明显降低。该结果表明，不同样本对局部语义表示粒度具有一定差异，但适宜的槽位规模整体较为集中，当前采用的$K=4$位于主要分布区间，并且是占比最高的单一配置。因此，固定$K=4$能够作为当前模型较为合理的统一槽位设置。对于不同样本间仍存在的槽位需求差异，后续可进一步考虑根据样本语义复杂度动态调整槽位数量。

3.6.2 模态重平衡分析

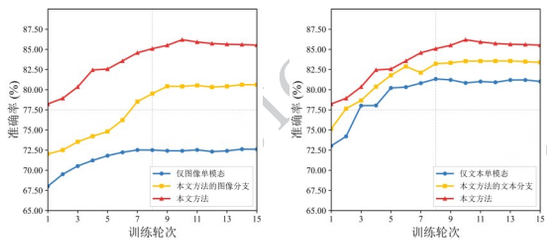
为分析动态原型引导的模态重平衡模块对联合训练中模态不平衡现象的缓解作用，在MMSD2.0数据集上跟踪训练过程中图像分支与文本分支的性能变化。具体地，分别屏蔽另一模态局部特征，得到对应的图像侧与文本侧评估分支，并记录其准确率，以对应单模态基线及完整多模态模型作为参照。

从图4(a)可以看出，模态重平衡后的图像分支整体优于图像单模态基线，且收敛更快，说明引入模态重平衡后，图像分支能够获得更充分的学习信号，缓解文本主导对视觉模态学习的抑制，使视觉线索更充分参与判别。图4(b)显示，文本单模态基线本身已具备较强的判别能力，重平衡后的文本分支仍保持较高性能，并较单模态基线有所提升。结果表明，该机制能够在保持文本分支判别能力的同时增强图像分支的有效更新，使两模态学习过程更加协调，进而提升跨模态信息的利用效果。

3.6.3 计算效率分析

为进一步分析所提模型的计算效率，在相同硬

© 中国图象图形学报版权所有



（a） 图像模态重平衡分析　　（b） 文本模态重平衡分析

((a) Visual modality rebalancing; (b) Textual modality rebalancing)

图4 模态重平衡机制分析

Fig. 4 Analysis of the modality rebalancing mechanism

件环境和实验设置下，统计模型参数量、训练时间和推理延迟，并结合分类性能进行比较，结果如表5所示。其中，单轮槽聚合将槽注意力迭代轮次设置为1，用于分析多轮迭代带来的计算开销；直接融合保留与本文方法相同的短语聚合过程，仅采用图文短语表示的直接拼接替代相关性门控融合。

由表5可见，本文方法虽较MV-CLIP具有更大的参数规模，但训练时间和推理延迟仍处于相近量级，同时取得了更高的检测性能。与单轮槽聚合相比，将迭代轮次由1增加至4后，模型参数量保持不变，表明不同迭代轮次间的参数共享避免了模型规模随迭代次数增加；训练时间和推理延迟有所增加，而准确率和F1值分别提高1.73和2.40个百分点，说明多轮竞争聚合能够以有限的额外计算开销进一步细化短语级语义表示。进一步比较直接融合与本文方法，引入相关性门控融合后，参数量由233 M增加至236 M，单轮训练时间由428 s增加至440 s，单样本推理延迟由20.05 ms增加至20.76 ms，而F1值由83.54%提高至85.27%。这是由于相关性门控主要作用于聚合后的少量短语级语义单元，因此额外计算开销相对有限。总体而言，所提模型引入的多轮槽聚合和短语级跨模态交互带来了有限的额外计算开销，同时获得了较为稳定的性能提升，在检测性能与计算效率之间取得较好平衡。

3.6.4 预训练编码器分析

为分析预训练编码器的训练方式及模型规模对检测性能的影响，在MMSD2.0数据集上比较不同编码器设置。Base采用本文默认的ViT-B/32和RoBERTa-base作为视觉与文本编码器，Large则分别采用规模更大的ViT-L/16和RoBERTa-large。对于两种编码器配置，分别设置参数冻结和端到端微调两种训练方式，Large配置下将编码器输出映射至与Base一致的特征维度。其余模型结构与训练设置保持一致，实验结果如图5所示。

表5 计算效率与性能比较

Table 5 Comparison of computational efficiency /and performance

方法	参数量/M	训练/s	推理/ms	准确率	F1值
MV-CLIP	165	416	19.42	85.64	84.10
单轮槽聚合	236	423	19.65	84.47	82.87
直接融合	233	428	20.05	84.72	83.54
本文	236	440	20.76	86.20	85.27

注：训练时间为单个训练轮次，推理延迟为单样本。

由图5可见，在Base和Large两种配置下，端到端微调均优于编码器冻结，说明任务相关的特征适配有利于提升模型性能。另一方面，采用更大规模编码器配置后模型性能仅有小幅变化，表明更强的预训练表示能够带来一定收益，但本文方法对编码器规模并不敏感。结合表3中局部语义冲突感知模块的消融结果可以看出，本文性能提升并非主要依赖更大规模的预训练编码器，而与局部冲突建模等任务相关结构设计密切相关。

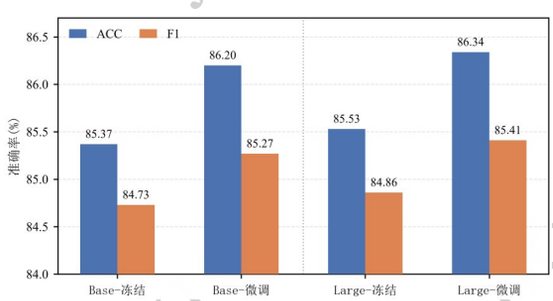


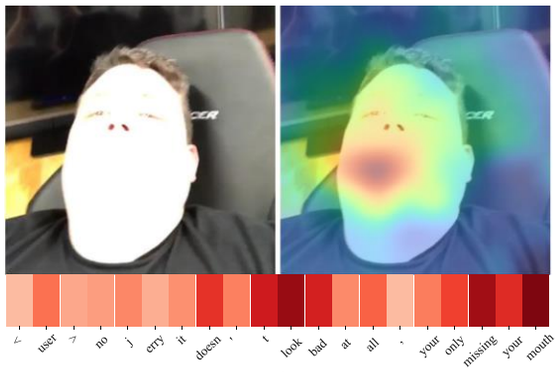
图5 不同编码器配置与训练策略下的性能比较

Fig. 5 Performance comparison under different encoder configurations and training strategies

3.6.5 局部特征聚合可视化分析

为分析槽聚合过程对局部语义信息的选择情况，选取测试集中两条代表性样本，对文本与视觉局部特征的聚合权重进行可视化。如图6所示，文本颜色深浅和图像热力区域反映局部特征在短语表示构建过程中的相对聚合权重。

© 中国图象图形学报版权所有



（a） 样本1局部特征聚合可视化



（b） 样本2局部特征聚合可视化

((a) Sample 1; (b) Sample 2)

图6 局部特征聚合可视化案例

Fig. 6 Visualization examples of local feature aggregation

在图6(a)中，文本侧的高权重主要集中在“doesn’t look bad”以及“only missing your mouth”等局部表达，其中“doesn’t look bad”提供对图像内容的评价信息，“missing your mouth”则进一步指向嘴部相关语义。对应的图像侧高权重区域主要集中在人脸，尤其是嘴部附近。从样本语义可以看出，文本采用较为轻描淡写的评价方式，而图像中的异常视觉线索与其形成明显反差。文本与图像两侧的聚合结果均覆盖了与该语义关系相关的局部内容。

在图6(b)中，文本侧较高权重主要出现在“get up”“on time”和“deadline”等词元附近，图像侧高权重区域则主要集中在人物及其情绪表达区域。文本描述较为日常的行为，而图像提供了更强的情绪与压力语境，两种信息共同构成该样本的讽刺语境。可以看到，槽聚合并未将权重均匀分散于所有局部特征，而是更集中于与样本主要语义相关的文本词元和图像区域。

总体来看，两组样本中的高权重区域与人工观察到的关键文本和视觉线索具有较好的对应关系，说明槽注意力能够从离散词元和图像块中集中保留与样本语义相关的局部信息，为后续短语级相关性计算与局部冲突建模提供较为连贯的语义表示。

4 结 论

针对局部图文冲突难以有效刻画以及联合训练中模态学习失衡的问题，提出局部语义冲突感知与模态重平衡相结合的多模态讽刺检测框架。该框架将离散的词元与图像块组织为短语级语义单元，并结合跨模态相关性与语义差异筛选关键局部冲突；同时，根据两模态的原型判别难度动态调节监督强度。实验结果表明，所提方法在MMSD1.0和MMSD2.0上均取得领先性能。消融实验、训练过程分析与可视化结果进一步表明，短语级建模有助于形成更有效的局部冲突表示，相关性门控有助于抑制无关差异，模态重平衡能够改善图像分支的学习状态，同时保持文本分支的判别能力。上述结果说明，从表征粒度与优化过程两个层面进行协同改进，能够提升关键图文不一致线索的利用效果。

所提方法仍采用固定数量的槽位表示不同样本中的局部语义单元，对有效短语数量差异较大的样本适应性有限；同时，模态学习状态依据批次级原型损失进行判断，可能受到批次类别分布与样本波动的影响。后续将研究动态短语聚合与更稳定的模态状态估计方法，并结合视觉语言预训练模型增强复杂场景下的跨模态语义推理能力。

参考文献(References)

Bao Y, Zhao X, Zhang P, Qi Y and Li H. 2025. HIAN: A hybrid interactive attention network for multimodal sarcasm detection. Pattern Recognition, 164: 111535 [DOI: 10.1016/j.patcog.2025.111535]

Cai Y, Cai H and Wan X. 2019. Multi-modal sarcasm detection in Twitter with hierarchical fusion model//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: ACL Press: 2506 - 2515 [DOI: 10.18653/v1/P19-1239]

Chen W, Lin F, Li G and Liu B. 2024. A survey of automatic sarcasm detection: Fundamental theories, formulation, datasets, detection methods, and opportunities. Neurocomputing, 578: 127428 [DOI: 10.1016/j.neucom.2024.127428]

© 中国图象图形学报版权所有

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations.

Fan Y, Xu W, Wang H, Wang J and Guo S. 2023. PMR: Prototypical modal rebalance for multimodal learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, BC, Canada: IEEE Computer Society Press: 20029 - 20038 [DOI: 10.1109/CVPR52729.2023.01918]

Farabi S, Ranasinghe T, Kanojia D, Kong Y and Zampieri M. 2024. A survey of multimodal sarcasm detection//Proceedings of the 33rd International Joint Conference on Artificial Intelligence. Jeju, Korea: IJCAI Press: 8020 - 8028 [DOI: 10.24963/ijcai.2024/887]

Graves A and Schmidhuber J. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18(5-6): 602 - 610 [DOI: 10.1016/j.neunet.2005.06.042]

Guo D, Cao C, Yuan F, Liu Y, Zeng G, Yu X, et al. 2025. Multi-view incongruity learning for multimodal sarcasm detection//Proceedings of the 31st International Conference on Computational Linguistics. Abu Dhabi, UAE: ACL Press: 1754 - 1766

He K, Zhang X, Ren S and Sun J. 2016. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE Computer Society Press: 770 - 778 [DOI: 10.1109/CVPR.2016.90]

Huang Y, Lin J, Zhou C, Yang H and Huang L. 2022. Modality competition: What makes joint training of multi-modal network fail in deep learning? (Provably)//Proceedings of the 39th International Conference on Machine Learning. Baltimore, Maryland, USA: PMLR: 9226 - 9259

Kim D, Kim N and Kwak S. 2023. Improving cross-modal retrieval with set of diverse embeddings//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, BC, Canada: IEEE Computer Society Press: 23422 - 23431 [DOI: 10.1109/CVPR52729.2023.02243]

Kim Y. 2014. Convolutional neural networks for sentence classification//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: ACL Press: 1746 - 1751 [DOI: 10.3115/v1/D14-1181]

Li H, Li X, Hu P, Lei Y, Li C and Zhou Y. 2023. Boosting multi-modal model performance with adaptive gradient modulation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE Computer Society Press: 22214 - 22224 [DOI: 10.1109/ICCV51070.2023.02030]

Li J, Li D, Xiong C and Hoi S. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation//Proceedings of the 39th International Conference on Machine Learning. Baltimore, Maryland, USA: PMLR: 12888 - 12900

Liang B, Lou C, Li X, Gui L, Yang M and Xu R. 2021. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs//Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event, China: ACM Press: 4707 - 4715 [DOI: 10.1145/3474085.3475190]

Liang B, Lou C, Li X, Yang M, Gui L, He Y, et al. 2022. Multi-modal sarcasm detection via cross-modal graph convolutional network//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: ACL Press: 1767 - 1777 [DOI: 10.18653/v1/2022.acl-long.124]

Liu H F, Chen J J, Li L, Bao B K, Li Z C, Liu J Y, et al. 2023. Cross-modal representation learning and generation. Journal of Image and Graphics, 28(6): 1608 - 1629 (刘华峰, 陈静静, 李亮, 鲍秉坤, 李泽超, 刘家瑛, 等. 2023. 跨模态表征与生成技术. 中国图象图形学报, 28(6): 1608-1629) [DOI: 10.11834/jig.230035]

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. 2019. RoBERTa: A robustly optimized BERT pretraining approach [EB/OL]. [2026-01-31]. https://arxiv.org/abs/1907.11692

Locatello F, Weissenborn D, Unterthiner T, Mahendran A, Heigold G, Uszkoreit J, et al. 2020. Object-centric learning with slot attention//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates, Inc.: 11525 - 11538

Pan H, Lin Z, Fu P, Qi Y and Wang W. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection//Findings of the Association for Computational Linguistics: EMNLP 2020. Online: ACL Press: 1383 - 1392 [DOI: 10.18653/v1/2020.findings-emnlp.124]

Peng X, Wei Y, Deng A, Wang D and Hu D. 2022. Balanced multi-modal learning via on-the-fly gradient modulation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, LA, USA: IEEE Computer Society Press: 8238 - 8247 [DOI: 10.1109/CVPR52688.2022.00806]

Qiao Y, Jing L, Song X, Chen X, Zhu L and Nie L. 2023. Mutual-enhanced incongruity learning network for multi-modal sarcasm detection//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, DC, USA: AAAI Press: 9507 - 9515 [DOI: 10.1609/aaai.v37i8.26138]

Qin L, Huang S, Chen Q, Cai C, Zhang Y, Liang B, et al. 2023. MMSD2.0: Towards a reliable multi-modal sarcasm detection system//Findings of the Association for Computational Linguistics: ACL 2023. Toronto, Canada: ACL Press: 10834 - 10845 [DOI: 10.18653/v1/2023.findings-acl.689]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on

© 中国图象图形学报版权所有

Machine Learning. PMLR：8748－8763

Sarsam S M，Al-Samarraie H，Alzahrani A I and Wright B. 2020. Sarcasm detection using machine learning algorithms in Twitter：A systematic review. International Journal of Market Research，62(5)：578－598［DOI：10.1177/1470785320921779］

Tao J H，Fan C H，Lian Z，Lyu Z，Shen Y and Liang S. 2024. Development of multimodal sentiment recognition and understanding. Journal of Image and Graphics，29(6)：1607－1627（陶建华，范存航，连政，吕钊，沈莹，梁山. 2024. 多模态情感识别与理解发展现状及趋势. 中国图象图形学报，29(6)：1607-1627）［DOI：10.11834/jig.240017］

Tian Y，Xu N，Zhang R and Mao W. 2023. Dynamic routing transformer network for multimodal sarcasm detection//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto，Canada：ACL Press：2468－2480［DOI：10.18653/v1/2023.acl-long.139］

Wei Y，Yuan S，Zhou H，Wang L，Yan Z，Yang R，et al. 2024. G^{2}SAM：Graph-based global semantic awareness method for multimodal sarcasm detection//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver，Canada：AAAI Press：9151－9159［DOI：10.1609/aaai.v38i8.28766］

Wen C，Jia G and Yang J. 2023. DIP：Dual incongruity perceiving network for sarcasm detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver，BC，Canada：IEEE Computer Society Press：2540－2550［DOI：10.1109/CVPR52729.2023.00250］

Xie Y，Zhu Z，Chen X，Chen Z and Huang Z. 2024. MoBA：Mixture of bi-directional adapter for multi-modal sarcasm detection//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne VIC，Australia：ACM Press：4264－4272［DOI：10.1145/3664647.3680914］

Xiong T，Zhang P，Zhu H and Yang Y. 2019. Sarcasm detection with self-matching networks and low-rank bilinear pooling//Proceedings of the World Wide Web Conference. San Francisco，CA，USA：ACM Press：2115－2124［DOI：10.1145/3308558.3313735］

Zhu Z，Shen K，Chen Z，Zhang Y，Chen Y，Jiao X，et al. 2024. DGLF：A dual graph-based learning framework for multi-modal sarcasm detection//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Miami，Florida，USA：ACL Press：2900－2912［DOI：10.18653/v1/2024.emnlp-main.170］

作者简介

刘金柏，男，硕士研究生，主要研究方向为多模态讽刺检测。E-mail：24014083037@stu.hqu.edu.cn

柳欣，男，教授，主要研究方向为信息检索、模式识别、深度学习。E-mail：xliu@hqu.edu.cn

彭淑娟，通信作者，女，教授，主要研究方向为计算机视觉、模式识别、机器学习。E-mail：pshujuan@hqu.edu.cn

© 中国图象图形学报版权所有