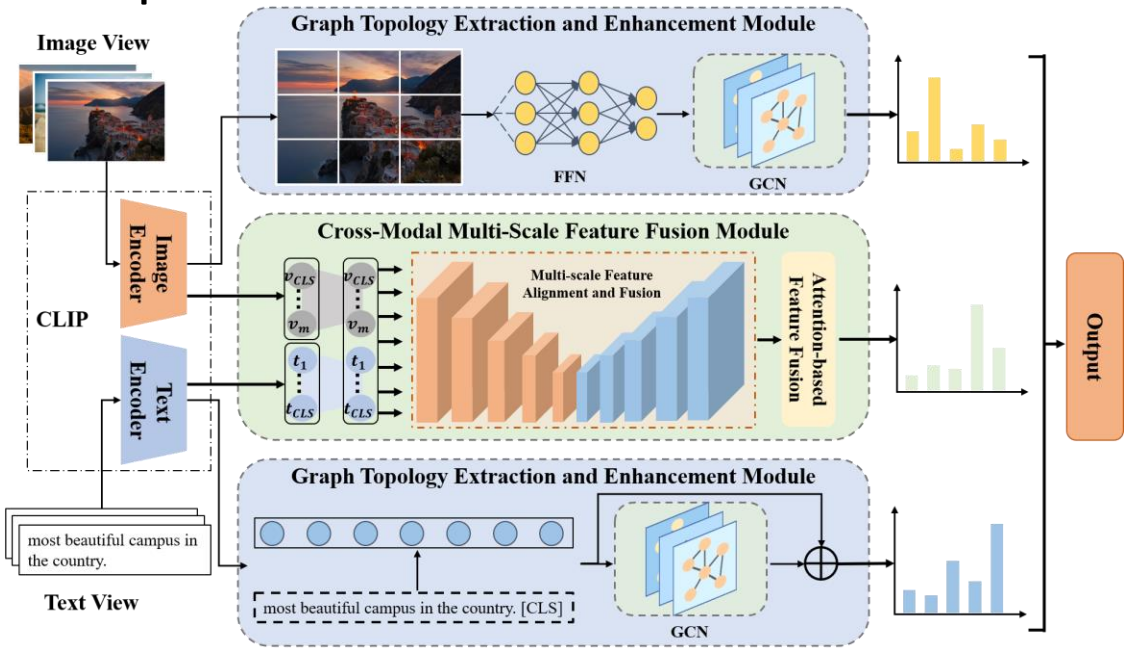
# DCPNet: A Comprehensive Framework for Multimodal Sarcasm Detection via Graph Topology Extraction and Multi-Scale Feature Fusion

**Youjiang FANG, Chuanbin LIU, Liang ZHANG, Shihao WANG, Wenyuan ZHANG, Yuxin WANG, Xiaopeng WEI, Xin YANG**

Frontiers of Computer Science, DOI: 10.1007/s11704-025-50196-4

# Problems & Ideas

- Problems of conventional MSD approaches:
  - Struggle to capture fine-grained intra-modal semantic structures;
  - Fail to model cross-modal semantic incongruities.
- Ideas: A unified model that integrates structural modeling and cross-modal alignment via graph topology and multi-scale fusion for improved sarcasm detection.




The architecture of the proposed DCPNet. CLIP is used to extract visual and textual representations, the GTEE module is utilized to extract deep features and graph topology features within each modality, the CMFF module aligns and merges features from multiple scales of text and images. Additionally, the attention mechanism assigns appropriate weights to both text and visual features, ensuring that comprehensive contextual information is captured.

# Main Contributions

- Contributions:
  - A novel multimodal sarcasm detection framework, DCPNet, that integrates unimodal features with cross-modal fused features, significantly enhancing sarcasm detection accuracy;
  - A GTEE module that extracts detailed cues from unimodal features;
  - A CMFF module that effectively aligns cross-modal information, improving semantic understanding in both text and image modalities.

| Modality | Method | MMSD | | | | MMSD 2.0 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Acc(%) | F1(%) | P(%) | R(%) | Acc(%) | F1(%) | P(%) | R(%) |
| Text | TextCNN† [43] | 80.03 | 75.32 | 74.29 | 76.39 | 71.61 | 69.52 | 64.62 | 75.22 |
| | Bi-LSTM† [44] | 81.90 | 77.53 | 76.66 | 78.42 | 72.48 | 68.05 | 68.02 | 68.08 |
| | SMSD† [45] | 80.90 | 75.82 | 76.46 | 75.18 | 73.56 | 69.97 | 68.45 | 71.55 |
| | RoBERTa† [46] | **93.97** | **92.45** | **90.39** | **94.59** | 79.66 | 76.21 | 76.74 | 75.70 |
| Image | ResNet* [47] | 71.53 | 66.53 | 64.41 | 68.80 | 66.44 | 61.65 | 62.06 | 61.25 |
| | VIT† [48] | 67.83 | 63.40 | 57.93 | 70.07 | 72.02 | 69.72 | 65.26 | 74.83 |
| Text-Image | HFM† [28] | 83.44 | 80.18 | 76.57 | 84.15 | 70.57 | 66.88 | 64.84 | 69.05 |
| | Att-BERT† [49] | 86.05 | 82.92 | 80.87 | 85.08 | 80.03 | 77.04 | 76.28 | 77.82 |
| | CMGCN* [29] | 86.43 | 86.12 | 85.21 | 87.04 | 79.83 | 76.90 | 75.82 | 78.01 |
| | HKE† [50] | 87.36 | 81.84 | 86.48 | 84.09 | 76.50 | 72.25 | 73.48 | 71.07 |
| | Multi-view CLIP† [19] | 88.33 | 85.55 | 82.66 | 88.65 | 85.64 | 84.10 | 80.33 | 88.24 |
| | DCPNet | **89.48** | **88.10** | **87.46** | **88.73** | **86.71** | **87.02** | **85.57** | **88.51** |

Experimental results on the MMSD and MMSD2.0 datasets.