

Beyond Boundaries: Navigating the Complexity of Multimodal Learning with Heterogeneity Tackling and Label-Guided Fusion

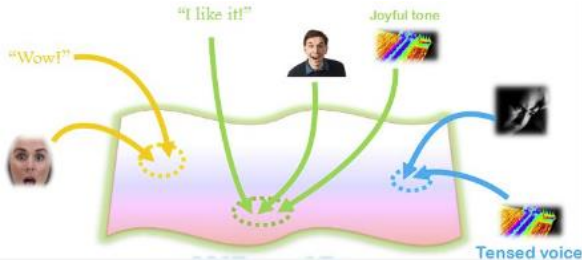
Meng CHEN

<https://chenmengdx.github.io/research/>

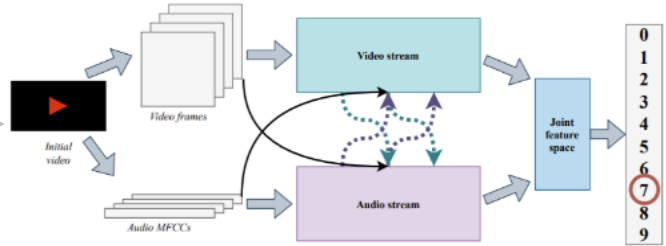
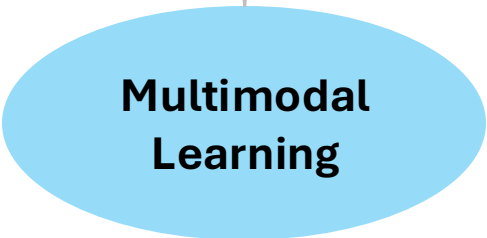
2024/04/12

Multimodal Intelligence

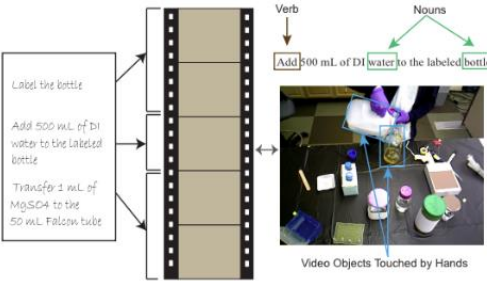
- Multimodality is a new AI paradigm where various modalities (text, speech, videos, images) are combined with multiple intelligence processing algorithms to achieve higher performance
- Multimodal applications currently include various discriminative tasks such as information retrieval, mapping and fusion



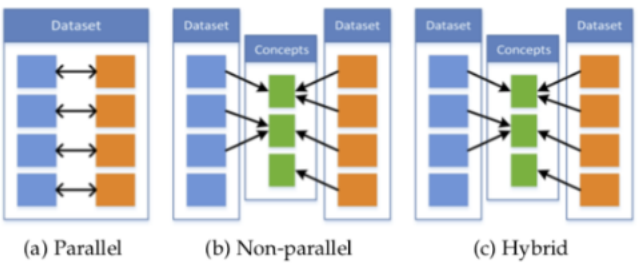
Representation



Fusion



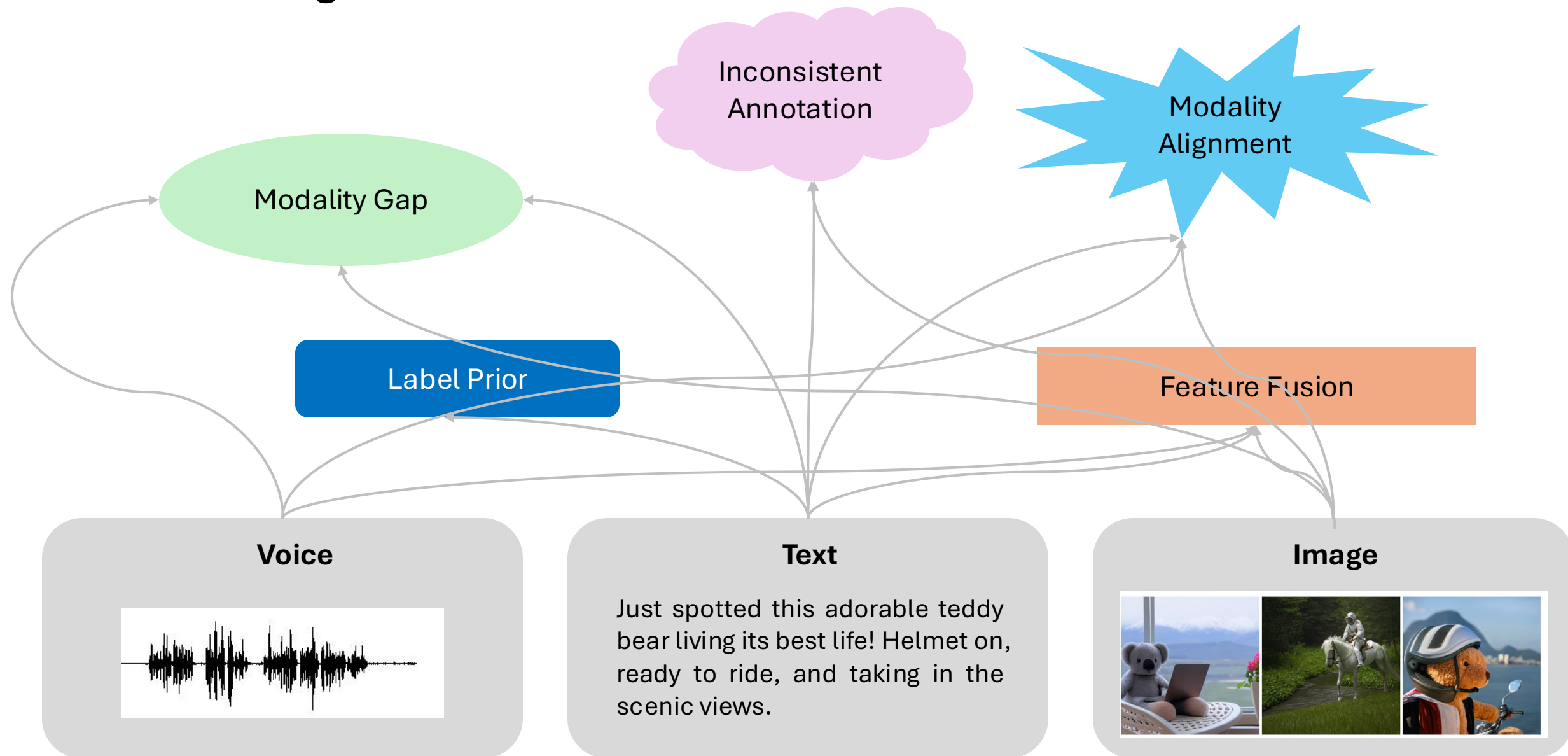
Alignment



Co-learning

[Zhang et al, 2020]

Main Challenges



Multimodal Fusion

Similarity

- Inner product: $\mathbf{u}\mathbf{v}$

Linear / sum

- Concat: $W[\mathbf{u}, \mathbf{v}]$
- Sum: $W\mathbf{u} + V\mathbf{v}$
- Max: $\max(W\mathbf{u}, V\mathbf{v})$

Multiplicative

- Multiplicative: $W\mathbf{u} \odot V\mathbf{v}$
- Gating: $\sigma(W\mathbf{u}) \odot V\mathbf{v}$
- LSTM-style: $\tanh(W\mathbf{u}) \odot V\mathbf{v}$

Attention

- Attention: $\alpha W\mathbf{u} + \beta V\mathbf{v}$
- Modulation: $[\alpha\mathbf{u}, (1-\alpha)\mathbf{v}]$

Bilinear

- Bilinear: $\mathbf{u}W\mathbf{v}$
- Bilinear gated: $\mathbf{u}W\sigma(\mathbf{v})$
- Low-rank bilinear: $\mathbf{u}U^T V\mathbf{v} = P(U\mathbf{u} \odot V\mathbf{v})$
- Compact bilinear: $\text{FFT}^{-1}(\text{FFT}(\Psi(\mathbf{x}, \mathbf{h}_1, \mathbf{s}_1)) \odot \text{FFT}(\Psi(\mathbf{x}, \mathbf{h}_2, \mathbf{s}_2)))$

Early, Middle, and Late Fusion

Suppose we have a binary classifier MLP and two input vectors.

Early - mix inputs:

- $\sigma(W_2\sigma(W_1[\mathbf{u},\mathbf{v}]+b_1)+b_2)$

Middle - concatenate features:

- $\sigma(W_2[\sigma(W_1[\mathbf{v}]+b_1), \sigma(W'_1[\mathbf{v}]+b'_1)] +b_2)$

Late - combine final scores:

- $1/2 (\sigma(W_2\sigma(W_1[\mathbf{u}]+b_1)+b_2) + \sigma(V_2\sigma(V_1[\mathbf{u}]+b'_1)+b'_2))$

More details

Douwe Kiela's talk at CS224n of Stanford University: <https://www.youtube.com/watch?v=5vfIT5LOkR0&t=5s>



Stanford CS224N NLP with Deep Learning | 2023 | Lecture 16 - Multimodal Deep Learning, Douwe Kiela

Our Main Works

Tacking Modality Gap

G2SAM: Graph-Based Global Semantic Awareness Method for Multimodal Sarcasm Detection

Yiwei Wei*, Shaozu Yuan*, Hengyang Zhou, Longbiao Wang, Zhiling Yan, Ruosong Yang, Meng Chen

The 38th Annual AAAI Conference on Artificial Intelligence (AAAI 2024)

Tackling Modality Heterogeneity with Multi-View Calibration Network for Multimodal Sentiment Detection

Yiwei Wei*, Shaozu Yuan*, Ruosong Yang, Lei Shen, Zhangmeizhi Li, Longbiao Wang, Meng Chen

The 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)

Label-guided fusion

MNER-QG: An End-to-End MRC framework for Multimodal Named Entity Recognition with Query Grounding

Meihuizi Jia, Lei Shen, Xin Shen, Lejian Liao, Meng Chen, Xiaodong He, Zhendong Chen, Jiaqi Li

The 37th Annual AAAI Conference on Artificial Intelligence (AAAI 2023)

Leveraging Label Information for Multimodal Emotion Recognition

Peiyong Wang, Sunlu Zeng, Junqing Chen, Lu Fan, Meng Chen, Youzheng Wu, Xiaodong He

The 24th INTERSPEECH Conference (Interspeech 2023)

Multimodal Sentiment Detection

Problem Definition

Input: <Text, Image> pair

Output: {Positive/Neutral/Negative}



Text: Mario finally has a smile on his face!

Sentiment Label: (Positive)



Text: The reason why this devoted dog is in critical condition will make you cry.

Sentiment Label: (Negative)

Figure 1: Examples of multimodal sentiment detection.

Motivation & Solution

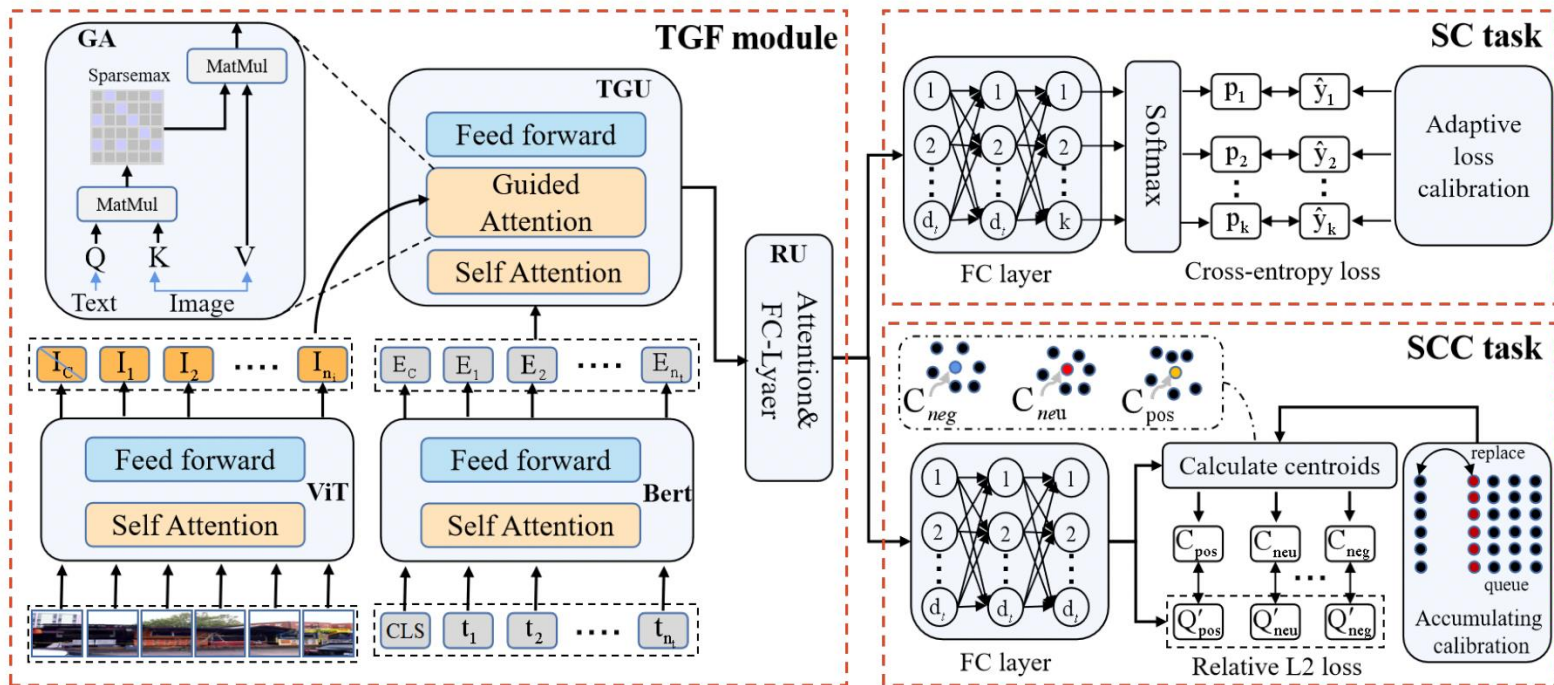
Motivation:

- Introducing redundant visual features during feature fusion
- Causing feature shift in the representation space
- Leading to inconsistent annotations for different modal data

Our Approach:

- Propose a Text-Guided Fusion (TGF) module to leverage text data to dominate the fusion process
- Propose a Sentiment-based Congruity Constraint (SCC) task to restrain representation space
- Introduce an adaptive loss calibration (ALC) strategy to calibrate the training loss

Multi-View Calibration Network



$$\hat{y}_i = (1 - \alpha) \cdot I(y_i = 1) + \left(\frac{\alpha}{K - 1}\right) \cdot I(y_i = 0) \quad (7)$$

$$L_{sc} = \sum ((1 - \alpha) * L_i \cdot I(y_i = 1) + \alpha \cdot L_i \cdot I(y_i = 0)) \quad (8)$$

$$L = \lambda_{sc} L_{sc} + \lambda_{scc} L_{scc} \quad (9)$$

$$X_t = \{E_C, E_1, E_2, \dots, E_{n_t}\} = BERT(T; \theta_t^{bert}) \quad (1)$$

$$X_i = \{I_C, I_1, I_2, \dots, I_{n_i}\} = ViT(I; \theta_i^{vit}) \quad (2)$$

$$TGU(\hat{X}_t, X_i)^N = [\hat{X}_f = \alpha(X_t, X_t, X_t), X_g = \gamma(X_f, X_i, X_i), FFN(X_g)].$$

$$C_{(0|1|2)} = \frac{\sum_{j=1}^B I(Y(j) = (0|1|2)) \cdot \hat{Q}_j}{\sum_{j=1}^B I(Y(j) = (0|1|2))} \quad (3)$$

$$D = \frac{\{\|\hat{Q}_i - C_{Y(i)}\|_2^2\}_{i=1}^B}{\sqrt{t}} \quad (4)$$

$$L_{scc} = - \sum_{i=1}^B \log\left(\frac{\exp(-d_i)}{\sum_{i=1}^B \exp(-d_i)}\right) \quad (5)$$

Experimental Results

Datasets:

- MVSA-Single (Niu et al., 2016)
- MVSAMultiple (Niu et al., 2016)
- HFM (Cai et al., 2019)

Baselines:

Unimodal Baselines.

- CNN (Kim, 2014)
- Bi-LSTM (Zhou et al., 2016)
- BERT (Jacob Devlin, 2019)
- ResNet (He et al., 2016)
- ViT (Dosovitskiy et al., 2020)

Multimodal Baselines.

- MultiSentiNet (Xu and Mao, 2017)
- HSAN (Xu, 2017)
- Co-MN-Hop6 (Xu et al., 2018)
- MGNNs (Yang et al., 2021)
- CLMLF (Li et al., 2022)

Modality	Model	MVSA-Single		MVSA-Multiple		Model	HFM	
		Acc	F1	Acc	F1		Acc	F1
Text	CNN	0.6819	0.5590	0.6564	0.5766	CNN	0.8003	0.7532
	BiLSTM	0.7012	0.6506	0.6790	0.6790	BiLSTM	0.8190	0.7753
	BERT	0.7111	0.6970	0.6759	0.6624	BERT	0.8389	0.8326
Image	ResNet-50	0.6467	0.6155	0.6188	0.6098	ResNet-50	0.7277	0.7138
	ViT	0.6378	0.6226	0.6194	0.6119	ViT	0.7309	0.7152
Multimodal	MultiSentiNet	0.6984	0.6984	0.6886	0.6811	Concat(2)	0.8103	0.7799
	HSAN	0.6988	0.6690	0.6796	0.6776	Concat(3)	0.8174	0.7874
	Co-MN-Hop6	0.7051	0.7001	0.6892	0.6883	MMSD	0.8344	0.8018
	MGNNs	0.7377	0.7270	0.7249	0.6934	D&R Net	0.8402	0.8060
	CLMLF	0.7533	0.7346	0.7200	0.6983	CLMLF	0.8543	0.8487
	CLMLF ¹	0.7378	0.7291	0.7112	0.6863	CLMLF ¹	0.8489	0.8446
	MVCN	0.7606	0.7455	0.7207	0.7001	MVCN	0.8568	0.8523

Table 1: Experimental results of different models on MVSA-Single, MVSA-Multiple and HFM datasets.

Ablation Study & Discussion

Model	MVSA-Single		MVSA-Multiple		HFM	
	Acc	F1	Acc	F1	Acc	F1
BERT	0.7111	0.6970	0.6759	0.6624	0.8389	0.8326
ViT	0.6378	0.6226	0.6194	0.6119	0.7309	0.7152
MFS	0.7217	0.7205	0.7063	0.6851	0.8434	0.8375
TGF	0.7396	0.7355	0.7095	0.6887	0.8493	0.8451
TGF, SCC	0.7515	0.7403	0.7158	0.6929	0.8522	0.8499
TGF, SCC+AC	0.7563	0.7422	0.7188	0.6971	0.8568	0.8523
TGF, SCC+AC, ALC	0.7606	0.7455	0.7207	0.7001	-	-

Table 3: Ablation results of our MVCN. Here, MFS denotes equally fusing the image and text features as Li et al. (2022). And TGF, SCC, AC, ALC are respectively our four designs: text-guided fusion module, sentiment-based congruity constraint task, accumulating calibration strategy, and adaptive loss calibration strategy. Note that the experiment for the ALC model on the HFM dataset is missing due to the absence of the unimodal labels.

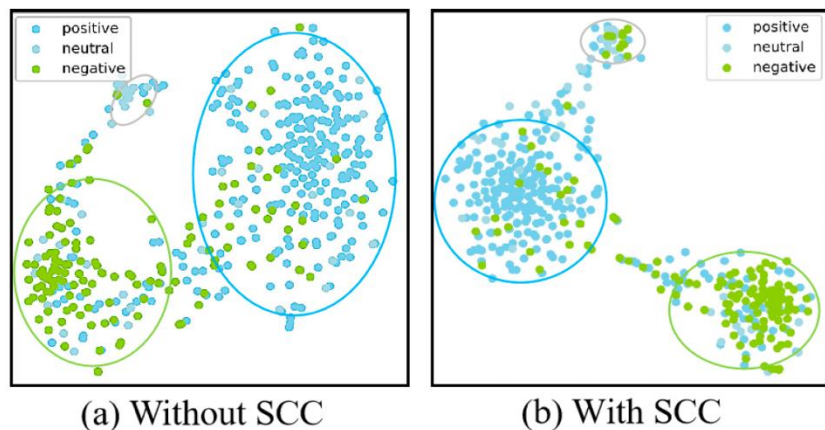


Figure 5: Visualization of representation. Different colored dots represent samples with different categories.

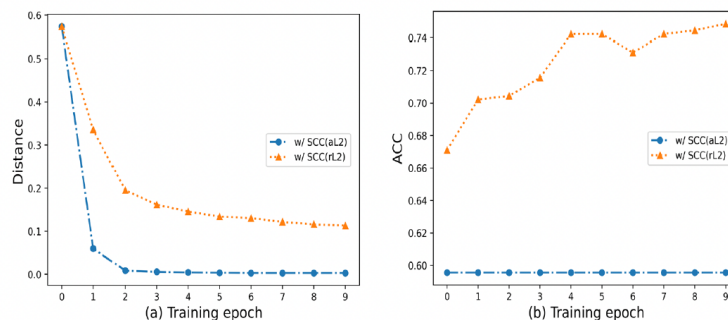


Figure 3: The training curves for the average semantic distance between the samples and corresponding centroids in (a), and the accuracy in (b) with different constraint methods. Here, aL2 and rL2 represent optimizing SCC with absolute L2 and relative L2.

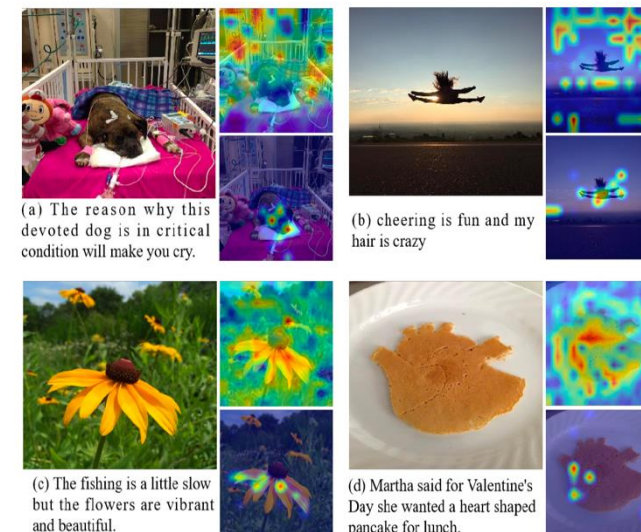


Figure 4: Attention visualization for sampling cases with Self-Attention (the upper one) and Sparse-Attention in TGF (the lower one).

Image	Text	MVCN	CLMLF
	Why buy ordinary art when you can have your art on sailcloth. vibrant colors!!	Positive	Negative
	My job is not to simply defend a position when having a heated discussion.	Negative	Neutral
	William Bruce Ellis Ranken's ebullient portrait of the young Ernest Thesiger. Born this day in 1879.	Positive	Neutral

Figure 6: Case Study of MVCN and CLMLF.

Multimodal Sarcasm Detection

Problem definition:

Input: <Text, Image> pair

Output: {Sarcasm, Non-sarcasm}

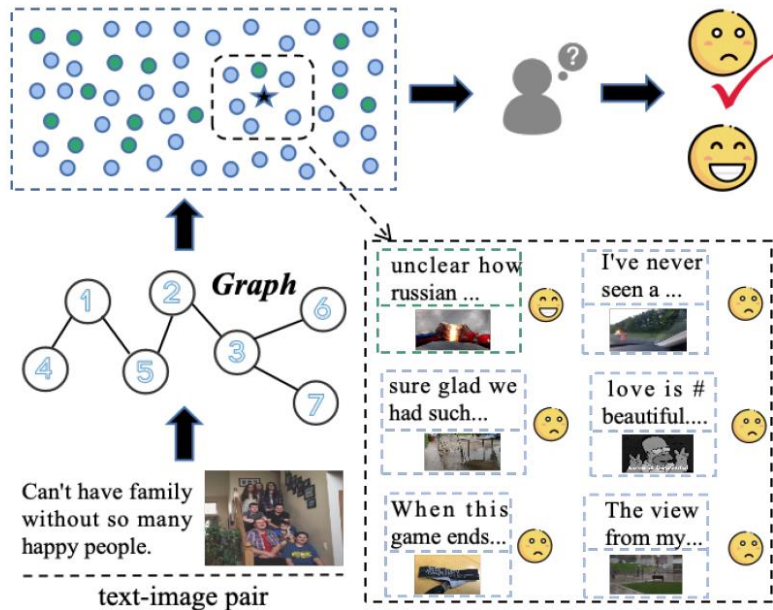


Figure 1: Examples of applying global semantic congruity for multimodal sarcasm prediction, where blue dots indicate sarcasm features while the green dots denote non-sarcasm.

Motivation:

- Multimodal posts with the same labels typically exhibit more analogous graph representations than other posts
- Sarcasm is often a subtle emotion, sometimes not so intense

Our approach:

- Introduce a novel inference paradigm to multimodal sarcasm detection by applying graph-based global semantic awareness
- Propose a Fine-grained Graph-aligned (FGM) model, a simple yet effective framework to align and fuse fine-grained unimodal graphs into a graph-based global space to capture contradictory sentiment cues
- Introduce Label-aware Graph Contrastive Learning (LGCL), which constrains graph-based representations with the same labels to be more similar in the semantic space

Graph-Based Global Semantic Awareness Model

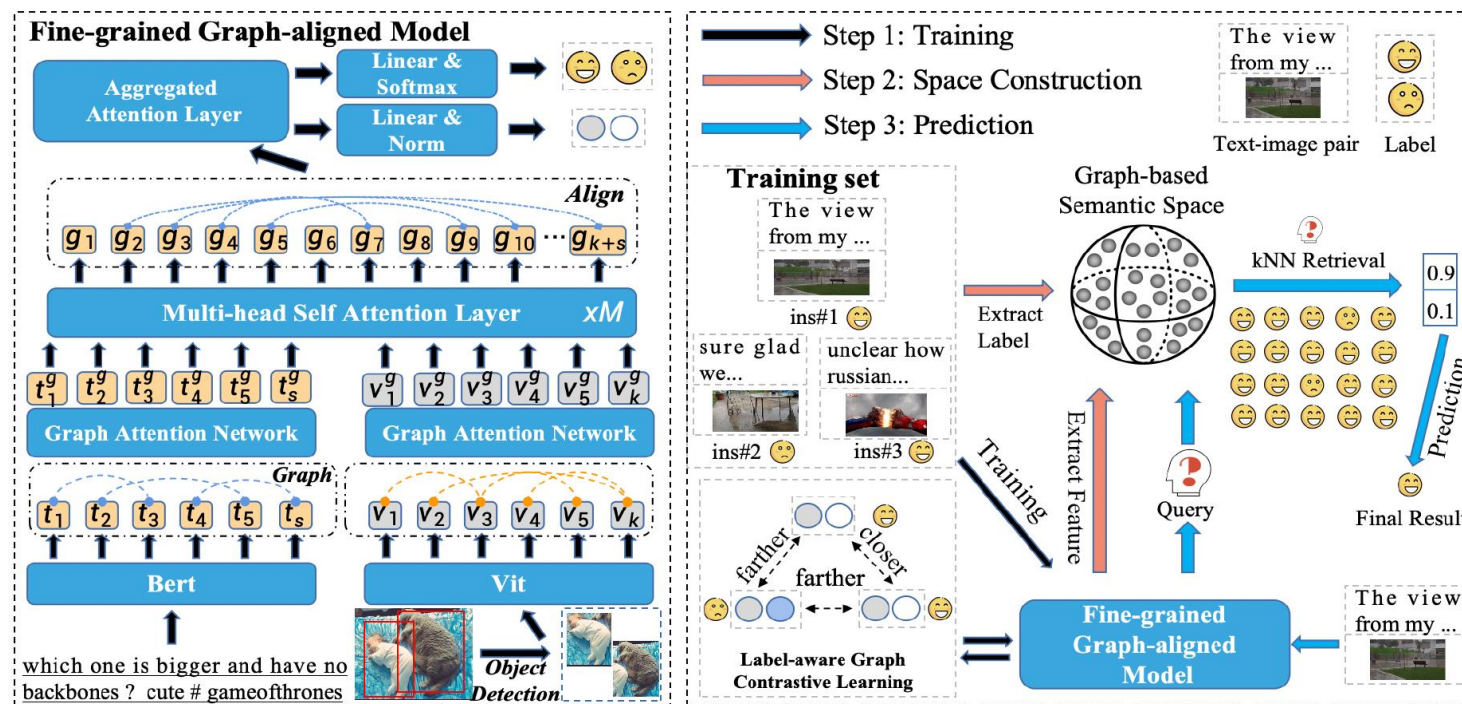


Figure 2: The overall architecture of our model. The left figure presents fine-grained graph-aligned model to generate multi-modal graph-based representation. And the right figure illustrates how to generate the result with graph-based global semantic congruity, where this inference process can be aware via label-aware graph contrastive learning during the training stage.

$$G = \text{softmax}\left(\frac{(G^{[v,t]}W_q)^T(G^{[v,t]}W_k)}{\sqrt{d}}\right)(G^{[v,t]}W_v) \quad (1)$$

$$\tilde{r}_i = \text{GELU}(g_iW_1 + b_1)W_2 + b_2 \quad (2)$$

$$q = \text{GELU}\left(\left(\sum_{i=1}^{k+s} \exp\left(\frac{\tilde{r}_i}{\sum_{j=1}^{k+s} \tilde{r}_j}\right)(g_i)\right)W_3 + b_3\right) \quad (3)$$

$$\hat{y}_{kNN} = \sum_{i=1}^k \alpha_i y_i, \quad \alpha_i = \frac{e^{-\|q_i - q\|_2^2 / \tau}}{\sum_j e^{-\|q_j - q\|_2^2 / \tau}} \quad (4)$$

$$\begin{cases} \hat{y} = 1, & \text{if } \hat{y}_{kNN} \geq 0.5 \\ \hat{y} = 0, & \text{if } \hat{y}_{kNN} < 0.5 \end{cases} \quad (5)$$

$$L_{ce} = \text{CrossEntropy}(GELU(qW_{ce} + b_{ce})) \quad (6)$$

$$L = L_{ce} + \gamma L_{LGCL} \quad (7)$$

Experimental Results

Datasets:

- HFM (Cai, Cai, and Wan 2019)

Unimodal Baselines:

- Bi-LSTM (Graves and Schmidhuber 2005)
- BERT (Devlin et al. 2018)
- Resnet
- ViT (Dosovitskiy et al. 2020)

Multimodal Baselines:

- HFM (Cai, Cai, and Wan 2019)
- Res-BERT
- Att-BERT (Pan et al. 2020)
- InCrossMGs (Liang et al. 2021)
- CMGCN (Liang et al. 2022)
- HKEmodel (Liu, Wang, and Li 2022)
- MILNet (Qiao et al. 2023)
- DIP (Wen, Jia, and Yang 2023)

MODALITY	METHOD	Acc(%)	Pre(%)	Rec(%)	F1(%)	Macro-average		
						Pre(%)	Rec(%)	F1(%)
image	Resnet	64.76	54.41	70.80	61.53	60.12	73.08	65.97
	ViT	67.83	57.93	70.07	63.43	65.68	71.35	68.40
text	Bi-LSTM	81.90	76.66	78.42	77.53	80.97	80.13	80.55
	BERT	83.85	78.72	82.27	80.22	81.31	80.87	81.09
image+text	HFM	83.44	76.57	84.15	80.18	79.40	82.45	80.90
	Res-BERT	84.80	77.80	84.15	80.85	78.87	84.46	81.57
	Att-BERT	86.05	78.63	83.31	80.90	80.87	85.08	82.92
	InCrossMGs*	86.10	81.38	84.36	82.84	85.39	85.80	85.60
	CMGCN*	86.54	-	-	82.73	-	-	-
	HKEmodel*	87.36	81.84	86.48	84.09	-	-	-
	MILNet* [†]	88.72	84.97	87.79	86.37	87.75	88.29	88.04
	DIP	89.59	87.76	86.58	87.17	88.46	89.13	89.01
	Ours*	90.48	87.95	89.02	88.48	89.44	89.79	89.65

Table 2: Experimental results for sarcasm detection. We use * to indicate the graph-based models. [†] indicates the reproduced results by unifying the textual backbone with the previous works.

Model	ACC(%)	Pre(%)	Rec(%)	F1(%)
MILNet	88.72	84.97	87.79	86.37
FGM	89.01	85.73	87.15	86.43
FGM+kNN	89.86	86.82	87.89	87.35
FGM+LGCL	89.33	86.29	87.56	86.92
FGM+kNN+LGCL	90.48	87.95	89.02	88.48

Table 3: The ablation results of our model. To show the superiority of FGM, we also provide the result of the previous SOTA graph-based model MILNet for comparison.

Model	ACC(%)	Pre(%)	Rec(%)	F1(%)
ViT	67.25	57.48	69.93	63.10
G ² SAM(ViT)	68.75	58.20	70.59	63.79
BERT	84.74	79.27	83.52	81.34
G ² SAM(BERT)	85.65	80.09	84.31	82.15
DIP	89.59	87.76	86.58	87.17
G ² SAM(DIP)	90.21	88.01	87.45	87.73

Table 4: Performance of three different types of models equipped with our G²SAM for sarcasm detection.

Discussion & Analysis




Text-Image Pairs	kNN Prediction				GT
	Instance	label	Instance	label	
 Can't have family without so many happy people.	ins#1	1	ins#6	1	Sarcasm 1
	ins#2	1	ins#7	1	
	ins#3	1	ins#8	1	
	ins#4	1	ins#9	0	
	ins#5	1	ins#10	1	
 Happy fourth birthday to one of the cutest and happiest kids.	ins#1	0	ins#6	1	Non-sarcasm 0
	ins#2	0	ins#7	0	
	ins#3	0	ins#8	0	
	ins#4	0	ins#9	0	
	ins#5	0	ins#10	0	
 This is a man who obviously knows how to make good decisions.	ins#1	1	ins#6	1	Sarcasm 1
	ins#2	1	ins#7	0	
	ins#3	0	ins#8	1	
	ins#4	1	ins#9	1	
	ins#5	0	ins#10	0	

Figure 4: User study for sampled instances. Here, we provide the top 10 retrieved kNN instances for analysis.

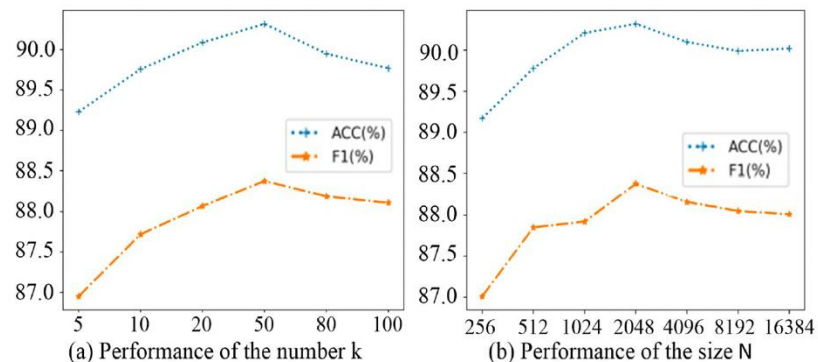


Figure 3: The curve of performance for multimodal sarcasm detection with different settings.

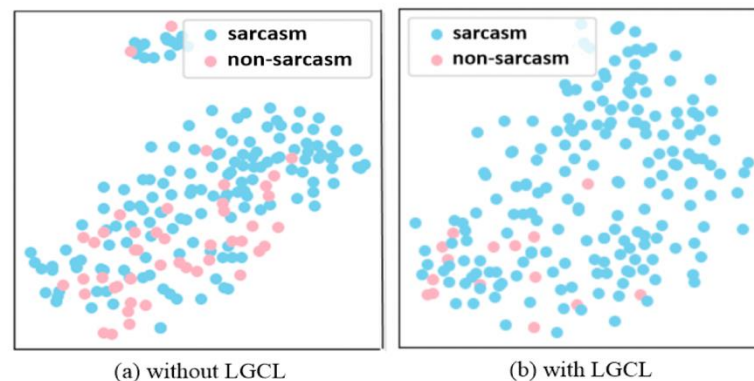


Figure 5: Distribution of the retrieved top 200 nearest neighbors instances for a sarcasm case. Here, the LGCL is removed in Figure (a).

Multimodal Named Entity Recognition with Query Grounding

Problem definition:

Input: <Image, Text> pair

Output: Entity type and value in the text

Motivation:

- Implicitly alignments inside a sentence-image pair is hard to interpret and evaluate
- Two-stage manner (detecting regions w/ tools then recognizing entity) brings error propagation

Our Approach:

- Propose a novel end-to-end MRC framework for Multimodal Named Entity Recognition
- Utilize label to provide prior knowledge of entity types and visual regions, and further enhance representations of both text and image



Figure 1: Two examples of MNER-QG with entity type “ORG”, “PER”, and “OTHER”.

Our Proposed Model

MNER-QG is a multi-task framework: Query Grounding, Existence Detection, and Entity Span Prediction

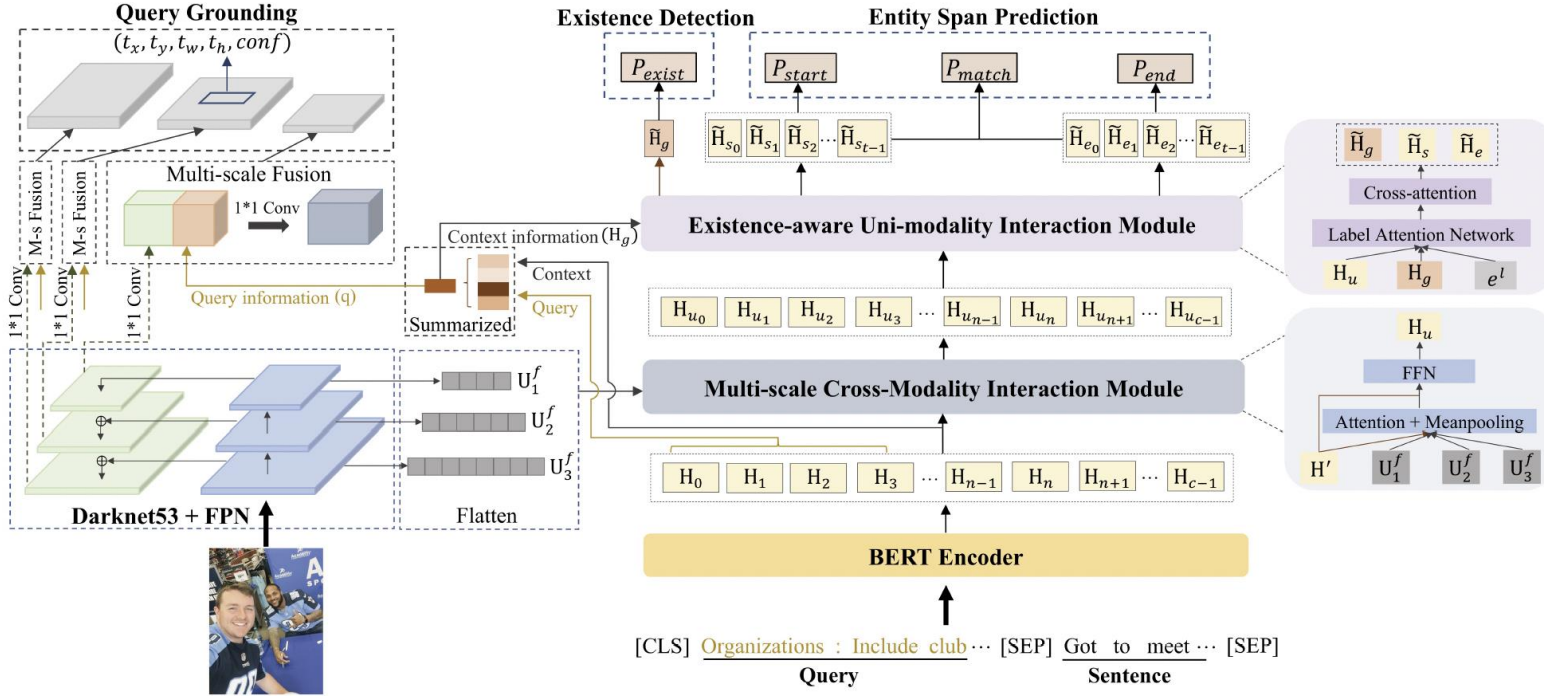


Figure 2: Overview of our MNER-QG framework (M-s Fusion denotes Multi-scale Fusion).

$$\alpha = \text{softmax}(\text{MLP}(\mathbf{Q})), \quad \mathbf{q} = \sum_{k=0}^{m-1} \alpha_k \mathbf{Q}[k, :] \quad (1)$$

$$\mathbf{Z}_s = \text{softmax}\left(\frac{\mathbf{Q}_s \mathbf{K}_g^T}{\sqrt{d_k}}\right) \mathbf{V}_g, \tilde{\mathbf{H}}_s = \text{LN}(\mathbf{H}_s + \mathbf{Z}_s), \quad (2)$$

$$P_{exist} = \text{softmax}(\tilde{\mathbf{H}}_g \mathbf{W}_{exist}) \quad (3)$$

$$P_{start} = \text{softmax}_{\text{eachrow}}(\tilde{\mathbf{H}}_s \mathbf{W}_s) \quad (4)$$

$$P_{match} = \text{sigmoid}(\mathbf{W}_m [\tilde{\mathbf{H}}_s; \tilde{\mathbf{H}}_e]) \quad (5)$$

$$\mathcal{L} = \omega_f \mathcal{L}_{QG} + \lambda_1 \mathcal{L}_{ED} + \lambda_2 \mathcal{L}_{ESP} \quad (6)$$

Experimental Results

Dataset

- Twitter2015 (Zhang et al. 2018)
- Twitter2017 (Lu et al. 2018)

For query grounding, both weak supervisions and manual annotations were applied to acquire explicit text-image alignment data

Methods	Twitter2015							Twitter2017						
	Single Type (<i>F1</i>)				Overall			Single Type (<i>F1</i>)				Overall		
	PER	LOC	ORG	OTH.	<i>Pre.</i>	<i>Rec.</i>	<i>F1</i>	PER	LOC	ORG	OTH.	<i>Pre.</i>	<i>Rec.</i>	<i>F1</i>
BiLSTM-CRF	76.77	72.56	41.33	26.80	68.14	61.09	64.42	85.12	72.68	72.50	52.56	79.42	73.43	76.31
CNN-BiLSTM-CRF	80.86	75.39	47.77	32.61	66.24	68.09	67.15	87.99	77.44	74.02	60.82	80.00	78.76	79.37
HBiLSTM-CRF	82.34	76.83	51.59	32.52	70.32	68.05	69.17	87.91	78.57	76.67	59.32	82.69	78.16	80.37
BERT	84.72	79.91	58.26	38.81	68.30	74.61	71.32	90.88	84.00	79.25	61.63	82.19	83.72	82.95
BERT-CRF	84.74	80.51	60.27	37.29	69.22	74.59	71.81	90.25	83.05	81.13	62.21	83.32	83.57	83.44
T-NER	83.64	76.18	59.26	34.56	69.54	68.65	69.09	-	-	-	-	-	-	-
MNER-QG-Text (Ours)	84.72	81.13	60.07	39.23	76.35	69.46	72.74	91.33	85.23	81.75	68.41	87.12	84.03	85.55
GVATT-HBiLSTM-CRF	82.66	77.21	55.06	35.25	73.96	67.90	70.80	89.34	78.53	79.12	62.21	83.41	80.38	81.87
AdaCAN-CNN-BiLSTM-CRF	81.98	78.95	53.07	34.02	72.75	68.74	70.69	89.63	77.46	79.24	62.77	84.16	80.24	82.15
GVATT-BERT-CRF	84.43	80.87	59.02	38.14	69.15	74.46	71.70	90.94	83.52	81.91	62.75	83.64	84.38	84.01
AdaCAN-BERT-CRF	85.28	80.64	59.39	38.88	69.87	74.59	72.15	90.20	82.97	82.67	64.83	85.13	83.20	84.10
MT-BERT-CRF	85.30	81.21	61.10	37.97	70.84	74.80	72.58	91.47	82.05	81.84	65.80	84.60	84.16	84.42
UMT-BERT-CRF	85.24	81.58	63.03	39.45	71.67	75.23	73.41	91.56	84.73	82.24	70.10	85.28	85.34	85.31
ATTR-MMKG-MNER	84.28	79.43	58.97	41.47	74.78	71.82	73.27	-	-	-	-	-	-	-
UMGF	84.26	83.17	62.45	42.42	74.49	75.21	74.85	91.92	85.22	83.13	69.83	86.54	84.50	85.51
MAF	84.67	81.18	63.35	41.82	71.86	75.10	73.42	91.51	85.80	85.10	68.79	86.13	86.38	86.25
MNER-QG (Ours)	85.31	81.65	63.41	41.32	77.43	72.15	74.70	92.92	86.19	84.52	71.67	88.26	85.65	86.94
MNER-QG (Oracle) (Ours)	85.68	81.42	63.62	41.53	77.76	72.31	74.94	93.17	86.02	84.64	71.83	88.57	85.96	87.25

Table 2: Results on two MNER datasets. We refer to the results of UMGF from Zhang et al. (2021) and other results from Xu et al. (2022). Our model achieves a statistically significant improvement with p-value<0.05 under a paired two-sided t-test.

Entity Type	Natural Language Query
PER (Person)	Person: People’s name and fictional character.
LOC (Location)	Location: Country, city, town continent by geographical location.
ORG (Organization)	Organization: Include club, company, government party, school government, and news organization.

Table 1: Examples of transforming entity types to queries.

Methods	Twitter2015			Twitter2017		
	<i>Pre.</i>	<i>Rec.</i>	<i>F1</i>	<i>Pre.</i>	<i>Rec.</i>	<i>F1</i>
MNER-QG	77.43	72.15	74.70	88.26	85.65	86.94
- w/o QG loss	77.50	70.79	73.99	88.01	84.69	86.32
- w/o ED loss	77.53	71.20	74.23	87.81	85.28	86.53
- w/o QG+ED loss	77.17	70.29	73.57	87.63	84.47	86.02

Table 3: Ablation study of MNER-QG on test set.

Methods	Twitter2015		Twitter2017		Flickr30K
	(W.S)	(M.A)	(W.S)	(M.A)	
	A@0.5	A@0.5	A@0.5	A@0.5	A@0.5
FA-VG	50.83	63.94	56.03	71.02	68.69
MNER-QG (Ours)	54.86	67.41	58.49	73.53	-

Table 5: Results on different bounding box labels on test set (W.S and M.A denote weak supervisions and manual annotations, respectively. A@0.5 is Accu@0.5. The result of FA-VG on Flickr30K derives from Yang et al. (2019).)

Ablation Study & Discussion

Methods	Twitter2015						Twitter2017					
	MNER			QG			MNER			QG		
	Pre.	Rec.	F1	Accu@0.5	Accu@0.75	Miou	Pre.	Rec.	F1	Accu@0.5	Accu@0.75	Miou
MNER-QG-Text	76.35	69.46	72.74	-	-	-	87.12	84.03	85.55	-	-	-
MNER-VG	77.03	71.08	73.94	-	-	-	87.91	84.22	86.03	-	-	-
FA-VG	-	-	-	50.83	32.69	45.49	-	-	-	56.03	38.92	51.14
MNER-QG (Ours)	77.43	72.15	74.70	53.93 (M:54.86)	40.22 (M:41.13)	49.50 (M:50.41)	88.26	85.65	86.94	57.50 (M:58.49)	43.03 (M:43.67)	54.09 (M:55.3)

Table 4: Performance comparison of joint-training and single-training models on test set. Note that two results were provided for the QG task, one is the QG results when MNER reaches the optimum, and the other is the optimal results in the QG task. (M denotes Max).

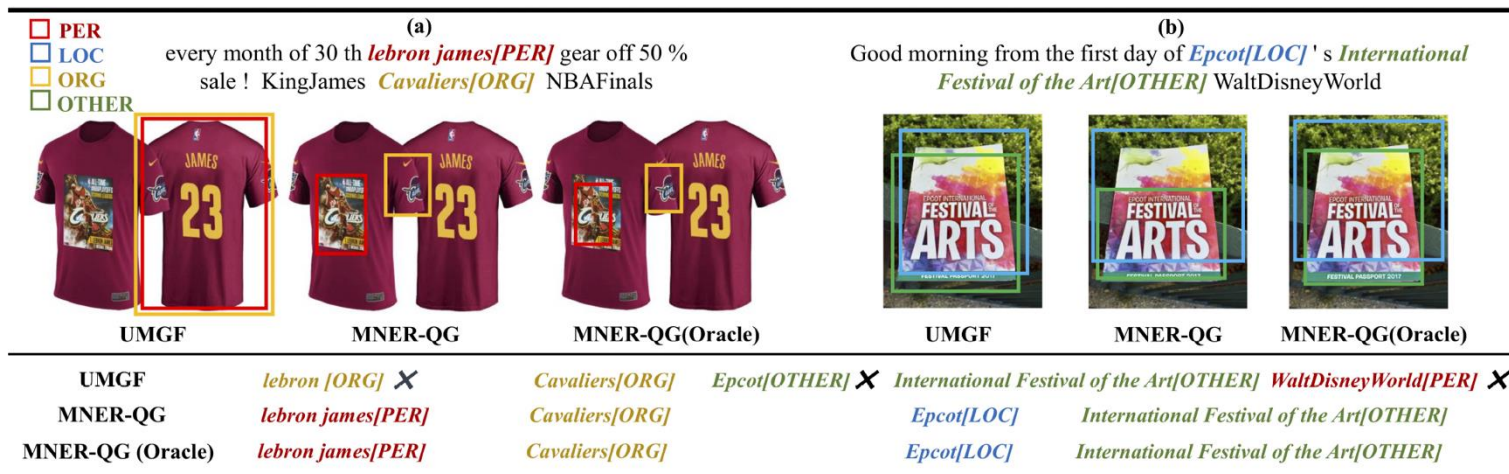


Figure 3: Example comparison among MNER-QG, MNER-QG (Oracle), and UMGF.

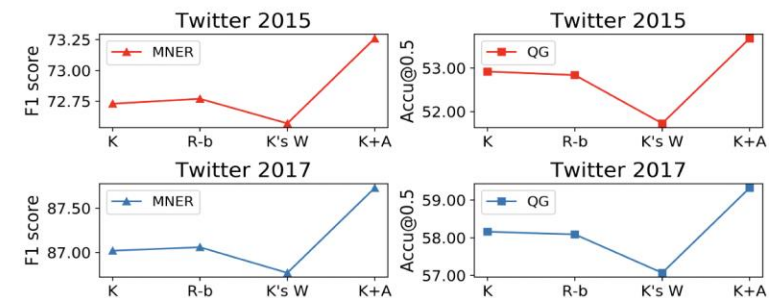


Figure 4: Results with different query transformations in MNER and QG on validation set (K, R-b, K's W, and K+A correspond to methods 1-4 of query transformations).

Multimodal Emotion Recognition

Problem definition:

- Input: <speech, text> pair
- Output: {Angry, Happy, Sad, Neutral}

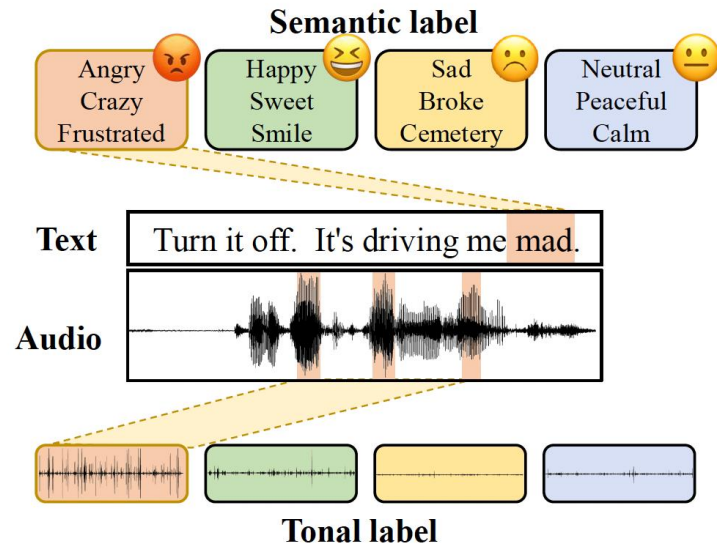


Figure 1: Visualization of labels. The *semantic label* presents the emotion relevant words for each class, and the *tonal label* displays the waveforms generated by concatenating the key-frames under each class.

Motivation:

- Speech sequence is lengthy, how to pay attention to the key information and effectively ignore the interference of redundant information?
- Label information should be capable of helping the model locate the salient tokens/frames relevant to the specific emotion

Solution:

- Conduct label-text/speech interactions by introducing a *label-token* attention mechanism for the text and a *label-frame* one for the speech which encourages the model to pay more attention to the emotion-related tokens/frames
- Propose a novel label-guided cross-attention mechanism to fuse different modalities, capable of learning the alignment between speech and text **from the perspective of emotional space**

Our Proposed Model (LE-MER)

$$\mathbf{G}_t = \frac{\mathbf{H}_t \cdot \mathbf{L}_t^T}{\|\mathbf{H}_t\|_2 \|\mathbf{L}_t\|_2} \quad (1)$$

$$\mathbf{p}_g^t = \text{Softmax}(\text{Meanpooling}(\mathbf{G}_t)) \quad (2)$$

$$\mathcal{L}_g^t = \text{CE}(\mathbf{y}, \mathbf{p}_g^t) \quad (3)$$

$$\mathbf{G}_s = \frac{\mathbf{H}_s \cdot \mathbf{L}_s^T}{\|\mathbf{H}_s\|_2 \|\mathbf{L}_s\|_2} \quad (4)$$

$$\mathbf{p}_g^s = \text{Softmax}(\text{Meanpooling}(\mathbf{G}_s)) \quad (5)$$

$$\mathcal{L}_g^s = \text{CE}(\mathbf{y}, \mathbf{p}_g^s) \quad (6)$$

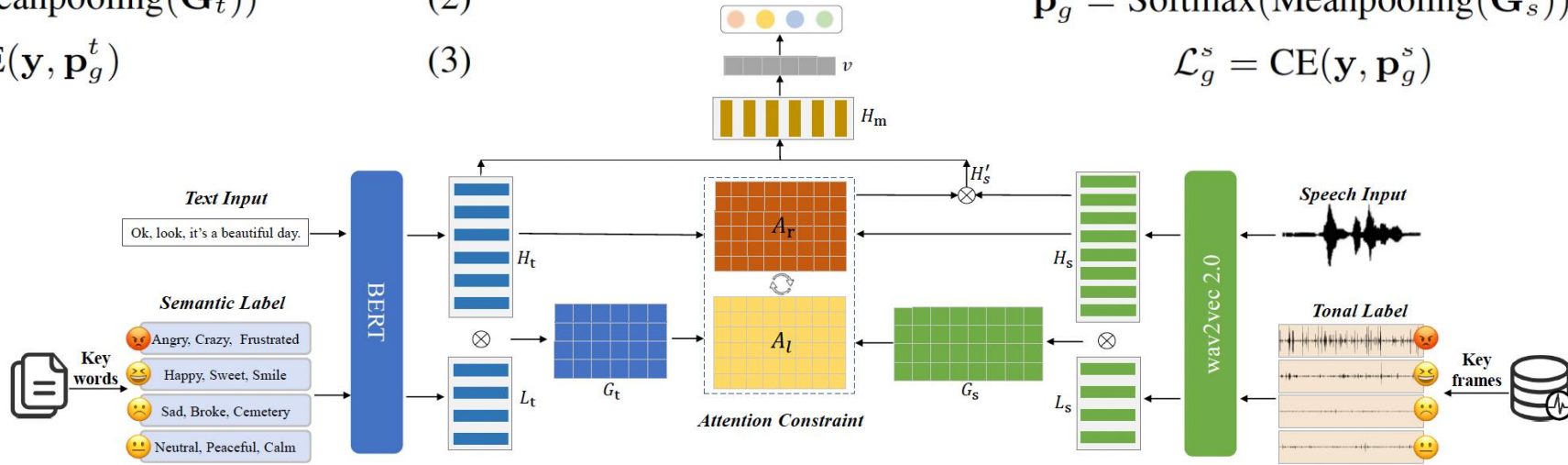


Figure 2: The architecture of our proposed model LE-MER.

$$\mathbf{A}_r = \text{Softmax}(\mathbf{H}_t \mathbf{H}_s^T \mathbf{W}) \quad (7)$$

$$\mathbf{H}'_s = \mathbf{A}_r \mathbf{H}_s^T, \mathbf{H}_m = [\mathbf{H}_t, \mathbf{H}'_s] \quad (8)$$

$$\mathbf{A}_l = \mathbf{G}_t \cdot \mathbf{G}_s^T \quad (9)$$

$$\mathcal{L}_c = \|\mathbf{A}_l - \mathbf{A}_r\|^2 \quad (10)$$

$$\mathcal{L}_m = \text{CE}(\mathbf{y}, \text{Softmax}(\text{Linear}(\mathbf{v}))) \quad (11)$$

$$\mathcal{L} = \mu_1 \mathcal{L}_m + \mu_2 \mathcal{L}_c + \mu_3 \mathcal{L}_g^t + \mu_4 \mathcal{L}_g^s \quad (12)$$

Experimental Results

Table 1: Comparison of our unimodal results on IEMOCAP dataset where “LE” denotes label embedding.

Sys.	Model	WA(%)	UA(%)
A1	BERT	67.34	67.66
A2	+ historical utterances	77.46	78.38
A3	+ historical utterances + LE (random init)	77.51	78.52
A4	+ historical utterances + LE (label words init)	78.03	78.88
A5	+ historical utterances + LE (TF-IDF init)	78.11	78.92
B1	wav2vec2.0	73.92	74.48
B2	+ 2nd stage	75.73	76.44
B3	+ 2nd stage + LE (random init)	76.20	76.80
B4	+ 2nd stage + LE (BERT embedding init)	76.48	77.14
B5	+ 2nd stage + LE (codebook init)	76.74	77.74

Table 3: Results of comparison between different fusion methods utilizing label-guided attention \mathbf{A}_l and vanilla attention \mathbf{A}_r .

Sys.	Model	WA(%)	UA(%)
C1	Attention Constraint	82.40	83.11
C2	$\mathbf{A}_r + \mathbf{A}_l$	82.39	82.75
C3	only \mathbf{A}_l	81.29	81.37
C4	only \mathbf{A}_r	81.08	81.68

Table 2: Comparison of our multimodal results with previous works on IEMOCAP dataset.

Model	WA(%)	UA(%)
Chen et al. [7]	74.30	75.30
Chen et al. [28]	74.92	76.64
Hou et al. [29]	75.60	77.60
Wu et al. [26]	77.57	78.41
Santoso et al. [30]	78.40	78.60
Li et al. [6]	80.36	81.70
Our Score Fusion	81.32	82.18
Ours	82.40	83.11

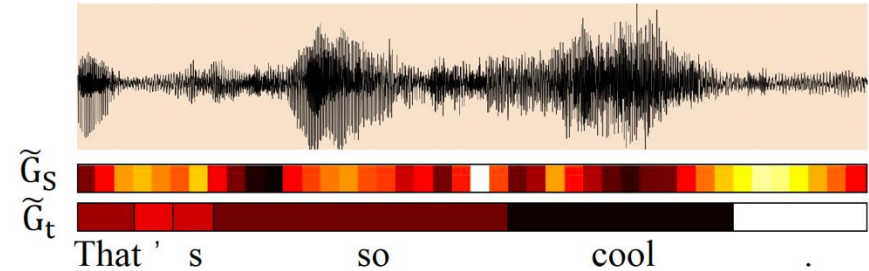


Figure 4: Visualization of $\tilde{\mathbf{G}}_s$ and $\tilde{\mathbf{G}}_t$.

Take-aways

- Given the inherent disparities among various modalities, it is crucial to bridge the modality gap for effective multimodal representation learning. Techniques such as fine-grained, graph-aligned models, text-guided fusion, semantic congruity constraints, and supervised graph contrastive learning are instrumental in mitigating this issue.
- Labels serve as valuable prior information, aiding in the feature fusion across different modalities, including text, speech, and image. The multi-task paradigm enhances knowledge transfer across diverse modalities.
- Self-supervised learning-based multimodal foundation models (GPT-4V, Gemini pro vision, LLaVA) represent the future, potentially unifying and simplifying both discriminative and generative tasks (not be discussed in this talk, maybe next time 😊).

Thanks!

Email: chenmengdx@gmail.com