From Text, Speech, to Multimodal Learning

Meng CHEN
https://chenmengdx.github.io/research/

20/06/2025

My Research

2019

- [Liu et al, IJCAI 2019]*
- [Liu et al, NLPCC 2019]
- [Guo et al, NLPCC 2019]*
- [Liu et al, MIPR 2019]*

2020

- [Liu et al, ACM MM 2020]*
- [Le et al, ECAI 2020]*
- [Chen et al, LREC 2020]
- [Shen et al, IJCNN 2020]*
- [Liu, Chen et al, NLPCC 2020]

2021

- [Yuan et al, ACM MM 2021]*
- [Shen et al, ACM MM 2021]
- [Liu et al, ICASSP 2021]*
- [Yuan et al, ICCV 2021]*
- [Zhang et al, CIKM 2021]*

2022

- **[Zhang et al, NAACL 2022]***
- [Liu et al, COLING 2022]*
- **[Yang et al, ICASSP 2022]***
- [Wang et al, ICASSP 2022]*
- [Zhu et al, Interspeech 2022]*
- [Fu et al, Interspeech 2022]
- [Yuan et al, IJCAI 2022]*
- **[Jia et al, ACM MM 2022]**
- [Le et al, CIKM 2022]*
- [Yuan et al, ICME 2022]*
- [Jia et al, LREC 2022]*

2023

- **[Jia et al, AAAI 2023]***
- [Bi et al, ACL 2023]*
- **[Zhang et al, ACL 2023]***
- **[Wei et al, ACL 2023]**
- **[Wang et al, ICASSP 2023]***
- [Fu et al, ICASSP 2023]
- [Wei et al, ICASSP 2023]
- [Fu et al, Interspeech 2023]
- **[Wang et al, Interspeech 2023]***
- [Wu et al, Interspeech 2023]*
- **[Zhang et al, CIKM 2023]***
- [Wei et al, Neurocomputing 2023]

2024

- [Wei et al, AAAI 2024]
- [Jia et al, IEEE Transactions on Multimedia]*
- [Yuan et al, IEEE Transactions on Multimedia]
- [Wei et al, IEEE Transactions on Circuits and Systems for Video Technology]

2025

- [Zhang et al, ACL 2025]*
- [Chen et al, NAACL 2025]*
- [Wu et al, AAAI 2025]*

NN PLM LLM

** Work that I was corresponding author*

Language Understanding

Dialog-Post: Multi-Level Self-Supervised Objectives and Hierarchical Model for Dialogue Post-Training
Zhenyu Zhang, Lei Shen, Yuming Zhao, Meng Chen*, Xiaodong He
The 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)

POSPAN: Position-Constrained Span Masking for Language Model Pre-training
Zhenyu Zhang, Lei Shen, Yuming Zhao, Meng Chen*, Xiaodong He
The 32nd ACM International Conference on Information and Knowledge Management (CIKM 2023)

Label Anchored Contrastive Learning for Language Understanding
Zhenyu Zhang, Yuming Zhao, Meng Chen*, Xiaodong He
2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2022)

Dialogue Pre-training: Existing Drawbacks & Motivations

Characteristics of dialogues

- Hierarchical semantic structure (Serban et al., 2016; Xing et al., 2018; Zhang et al., 2019), i.e., dialogue → utterance → token
- Multi-facet attributes (See et al., 2019; Shen et al., 2021a), such as speaker-shift, content-relatedness, fact-awareness, and coherence

Motivations

- How can we improve our modeling of the hierarchical semantic relations in dialogues?
- Is it possible to design auxiliary pre-text tasks that capture the multi-faceted attributes of dialogues?
- With the classic token/span masking method, are we overlooking anything?

HSSA: Hierarchical Segment-wise Self-Attention Network

- HSSA model contains several layers, and each layer is a block consisting of inner-segment self-attention, intersegment self-attention, segment updater, and feedforward sub-layers
- HSSA can reduce the memory cost from $O(n^2)$ to $O(nB + (\frac{n}{B})^2 + n)$

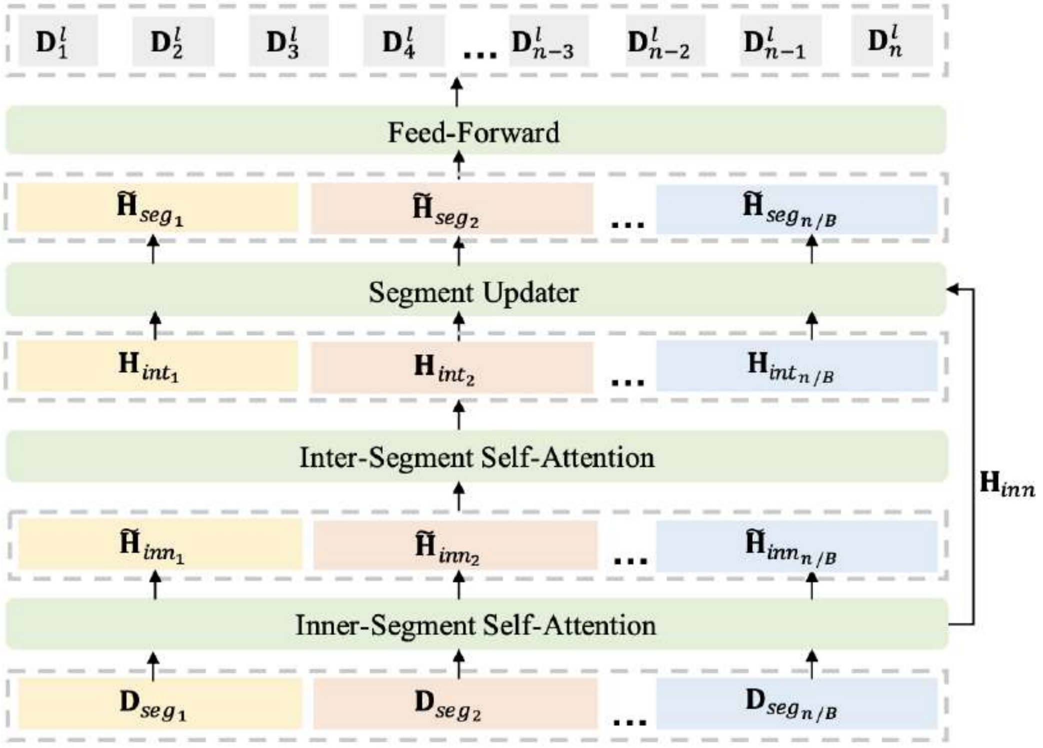


Figure 2: Overview of a HSSA layer.

$$\mathbf{H}_{inn_i} = \mathrm{SA}(\mathbf{D}_{seg_i}) \in \mathbb{R}^{B\times d}$$

$$\mathrm{Agg}(\mathbf{H}_{inn_i}) = \frac{1}{\sum e^{\mathbf{M}_j}} \sum_{j=1}^{B} \mathbf{H}_{inn_{i,j}} * e^{\mathbf{M}_j},$$

$$\alpha_{ij} = \mathrm{softmax}(\frac{\mathrm{Agg}(\mathbf{H}_{inn_i})\mathbf{H}_{inn_{i,j}}^T}{\sqrt{d}}), j \in [1, B],$$

$$\tilde{\mathbf{H}}_{inn_i} = \mathbf{W}_p(\sum_{j=1}^{B} \mathbf{H}_{inn_{i,j}} * \alpha_{ij})^T + \mathbf{b}_p,$$

$$\tilde{\mathbf{H}}_{inn} = [\tilde{\mathbf{H}}_{inn_1}, \tilde{\mathbf{H}}_{inn_2}, ..., \tilde{\mathbf{H}}_{inn_{n/B}}],$$

$$\mathbf{H}_{int} = \mathrm{SA}(\tilde{\mathbf{H}}_{inn})$$

$$\tilde{\mathbf{H}}_{seg_{i,j}} = \beta_{i,j} * \mathbf{H}_{int_i} + \mathbf{H}_{inn_{i,j}},$$

$$\beta_{i,j} = \mathrm{softmax}(\frac{\mathbf{H}_{inn_{i,j}}\mathbf{H}_{int_i}^T}{\sqrt{d}}), j \in [1, B].$$

SSOs: Multi-level self-supervised objectives

$$\mathcal{L} = \mathcal{L}_{DSM} + \mathcal{L}_{DRM} + \mathcal{L}_{DUC} + \mathcal{L}_{DUP} + \mathcal{L}_{DCL}$$

- We design five multilevel SSOs to post-train the dialogue encoder, which consist of two token-level SSOs, one utterance-level SSO, and two dialogue-level SSOs

- Apply the popular continuous multi-task learning (CMTL) framework for model training, which can pre-train models with multitask objectives efficiently and prevent knowledge forgetting of previous tasks when training with the current task objective(s)

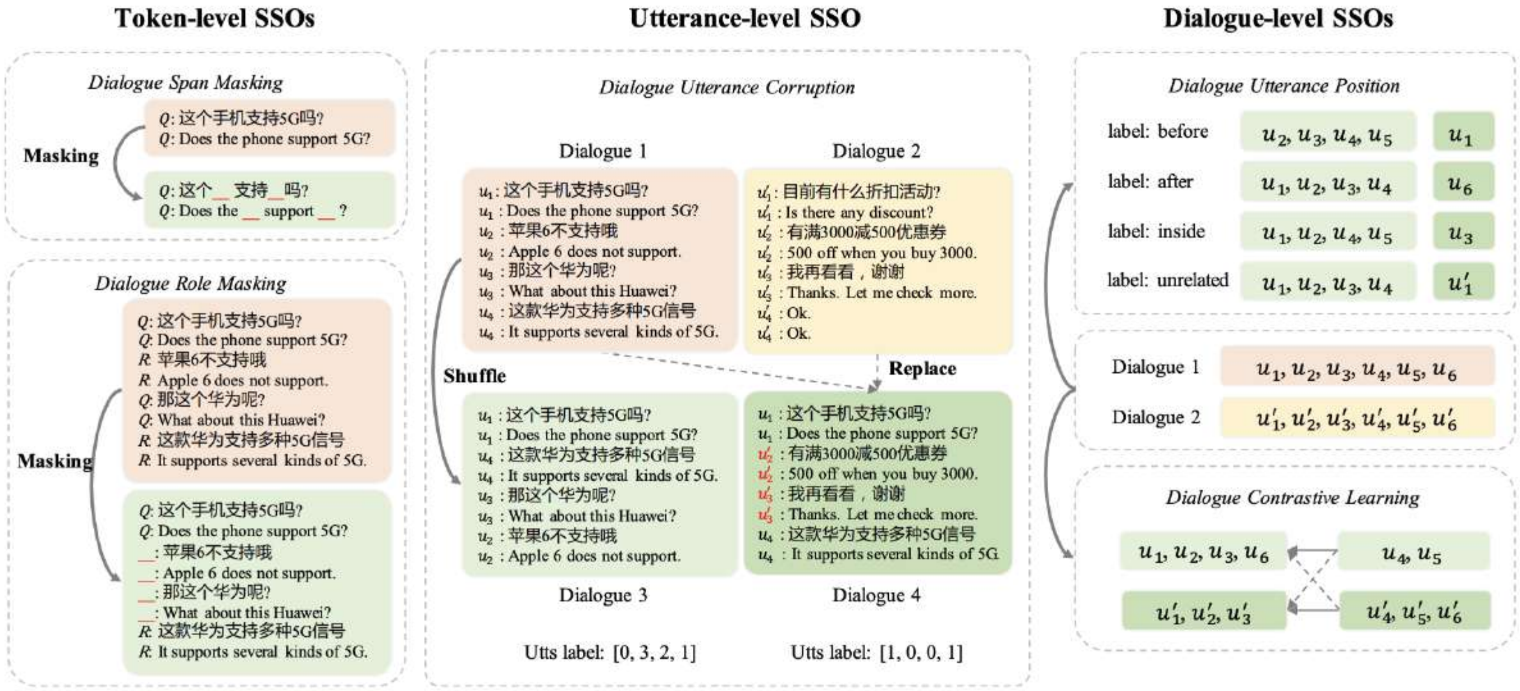


Figure 1: Illustration of multi-level SSOs in DIALOG-POST. Q and R represent speaker roles. u_i represents utterance. The utterance/dialogue in green color represents the corrupted utterance/dialogue.

POSPAN: Position-Constrained Span Masking

- Existing span masking only considers span length with some discrete distributions, while the dependencies among spans are ignored
- We present POSPAN, a general framework to allow diverse position-constrained span masking strategies via the combination of span length distribution and position constraint distribution

- **Case 1**: There are barely any dependency or semantic relationship between S_i and S_j, i.e., we can predict S_i and S_j independently without knowing each other.
- **Case 2**: $S_i \rightarrow S_j$, i.e., S_i is the premise of S_j. When S_i appears, S_j will appear most of the time.
- **Case 3**: $S_j \rightarrow S_i$, i.e., S_j is the premise of S_i.

$$
\begin{aligned}
P(S_i, S_j|R_{ij}) &= \frac{P(R_{ij}|S_i, S_j) * P(S_i, S_j)}{P(R_{ij})}, \\
\log P(S_i, S_j|R_{ij}) &\propto \underbrace{\log P(R_{ij}|S_i, S_j)}_{①} + \underbrace{\log P(S_i, S_j)}_{②},
\end{aligned} \tag{1}
$$

$$
\begin{aligned}
&\sum_{i,j} \log P(S_i, S_j) = \frac{(M-1)\log P(S_1, S_2, ..., S_M)}{2} \\
&\propto \sum_{i=1}^{M} \log P(S_i). \\
&\mathcal{L}_S :\rightarrow max(\mathbb{E}[\log P(S_i|len_i)]), \ \ where \\
&\mathbb{E}[\log P(S_i|len_i)] = \mathbb{E}_{len_i \sim F_M}(\sum_{l=0}^{len_i-1} \log P(x_{i+l})).
\end{aligned} \tag{2}
$$

$$
\begin{aligned}
&P(R_{ij}|d) = P(R_{ij}|x_{pos_i-1}, ..., x_{pos_j+len_j}) \\
&= \frac{P(x_{pos_i-1}, ..., x_{pos_j+len_j}|R_{ij}) * P(R_{ij})}{P(x_{pos_i-1}, ..., x_{pos_j+len_j})} \\
&= P(R_{ij}) * \frac{P(S_i, S_j|R_{ij}) * \prod_{k=pos_j-d}^{pos_j-1} P(x_k|R_{ij})}{P(S_i, S_j) * \prod_{k=pos_j-d}^{pos_j-1} P(x_k)} \\
&= P(R_{ij}|S_i, S_j) * \frac{\prod_{k=pos_j-d}^{pos_j-1} P(x_k|R_{ij})}{\prod_{k=pos_j-d}^{pos_j-1} P(x_k)} \\
&where\ d \sim F_D.
\end{aligned} \tag{3}
$$

$$
P(R_{ij}|S_i, S_j) \propto P(R_{ij}|d). \tag{4}
$$

Finally, the pre-training with masked language modeling can be decomposed into two losses:

$$
\begin{aligned}
&\mathcal{L} = \mathcal{L}_R + \mathcal{L}_S, \\
&\mathcal{L}_R :\rightarrow max(\mathbb{E}[\log P(R_{ij}|F_D)]),
\end{aligned} \tag{5}
$$

Experiments

- **Datasets**
  - Pre-training: JDDC (Chen et al., 2020) and ECD (Zhang et al., 2018)
  - POSPAN: 9 public NLU tasks
- **Evaluation**
  - Dialogue Representation Evaluation: SR & STS
  - Dialogue Understanding Evaluation: IC, Senti, CtxQ, CtxR

Method	JDDC			ECD		
	Corr.	MAP	MRR	Corr.	MAP	MRR
BERT (Devlin et al., 2019)	72.60	53.03	66.99	74.26	59.32	76.89
ELECTRA (Clark et al., 2020)	71.05	52.21	66.30	73.07	56.07	76.14
ERNIE (Sun et al., 2019, 2020)	72.73	52.96	66.79	74.29	59.11	76.87
UMS (Whang et al., 2021)	74.69	56.39	70.33	75.23	60.99	78.06
TOD-BERT (Wu et al., 2020)	78.43	60.15	74.32	80.17	65.78	80.22
PLATO (Bao et al., 2020b, 2021)	73.48	53.86	68.00	74.65	60.52	77.16
DialBERT (Zhang et al., 2021)	76.55	58.83	72.09	78.65	62.23	78.64
DomainAP (Wu et al., 2021)	76.54	59.27	72.36	78.99	62.85	79.08
DialCSE (Liu et al., 2021)	81.22	68.02	79.52	83.94	69.32	81.20
DIALOG-POST-BERT	82.78	69.91	79.83	**83.96**	**71.78**	**81.78**
DIALOG-POST	**82.90**	**69.95**	**79.87**	83.91	71.65	81.72

Table 2: Evaluation results on semantic retrieval (SR) and dialogue-based semantic textual similarity (D-STS) tasks.

Task	Class	Metric	Train	Test
J/D-STS	-	Corr.	-	2,000
J/SR	-	MAP/MRR	-	6,970
E/D-STS	-	Corr.	-	1,000
E/SR	-	MAP/MRR	-	4,243
IC	30	F1	4.7K	988
Senti	7	ACC	2.7K	342
CtxQ	2	AUC	4.1K	620
CtxR	2	AUC	4K	593

Table 3: Details of evaluation tasks. "J" and "E" represent JDDC and ECD.

Method	IC	Senti	CtxQ	CtxR	Average
BERT (Devlin et al., 2019)	86.0 ± 0.3	71.9 ± 1.8	87.9 ± 1.1	80.0 ± 0.9	81.5
ELECTRA (Clark et al., 2020)	87.4 ± 0.5	72.5 ± 0.6	88.9 ± 0.5	81.7 ± 1.5	82.6
ERNIE (Sun et al., 2019, 2020)	87.2 ± 0.3	73.4 ± 1.0	89.2 ± 1.2	82.9 ± 0.4	83.2
UMS (Whang et al., 2021)	86.8 ± 0.3	71.2 ± 1.0	88.8 ± 0.8	84.0 ± 0.1	82.7
TOD-BERT (Wu et al., 2020)	87.4 ± 0.9	74.8 ± 1.2	87.8 ± 0.7	82.8 ± 0.5	83.2
PLATO (Bao et al., 2020b, 2021)	86.5 ± 0.4	73.1 ± 0.1	88.9 ± 0.4	82.2 ± 0.4	82.7
DialBERT (Zhang et al., 2021)	88.5 ± 0.4	73.5 ± 0.5	87.5 ± 0.4	81.9 ± 0.5	82.8
DomainAP (Wu et al., 2021)	87.9 ± 0.4	73.8 ± 0.5	89.1 ± 0.4	83.7 ± 0.2	83.6
DialCSE (Liu et al., 2021)	86.8 ± 0.3	73.6 ± 0.5	90.7 ± 0.8	85.6 ± 0.2	84.2
DIALOG-POST-BERT	91.3 ± 0.7	$\mathbf{78.3}\pm0.9$	92.0 ± 0.6	87.3 ± 0.8	87.2
DIALOG-POST	$\mathbf{91.8}\pm0.5$	78.1 ± 0.5	$\mathbf{92.4}\pm0.7$	$\mathbf{87.9}\pm0.5$	**87.5**

Table 4: Evaluation results on dialogue understanding tasks (all with significance value $p < 0.05$).

Ablation Study

Ablation of HSSA

- We stack 10 layers of HSSA blocks and 2 layers of Transformer blocks, the last 2 Transformer layers are devised to capture the full dialogue semantics based on the global self-attention (SA) mechanism. Here, we first replace the last 2 Transformer layers with 2 HSSA layers (denoted as “w/o trs”)
- The performance of Senti becomes slightly better with all HSSA blocks. Since the input of Senti task is an utterance without context, it is possible that the 12-layer HSSA focusing on the local attention has some advantages

Model	JDDC			ECD		
	Corr.	MAP	MRR	Corr	MAP	MRR
HSSA	**82.90**	**69.95**	**79.87**	**83.91**	**71.65**	**81.72**
w/o trs	78.92	65.40	76.31	79.84	68.25	78.86
w/o updater	74.20	65.61	74.35	75.67	67.33	77.85
w/o $\mathbf{H}_{int}$	58.75	49.83	65.74	56.92	59.86	74.99
w/o $\mathbf{H}_{inn}$	45.97	48.64	63.22	29.65	49.57	69.02

Table 9: Experimental results of HSSA Ablation Study on all dialogue representation tasks.

Model	IC	Senti	CtxQ	CtxR	Average
HSSA	**91.8**	78.1	**92.4**	**87.9**	**87.5**
w/o trs	91.0	**78.5**	91.2	87.2	87.0
w/o updater	88.6	77.6	90.5	86.5	85.8
w/o $\mathbf{H}_{int}$	86.8	75.2	87.9	82.7	83.2
w/o $\mathbf{H}_{inn}$	76.6	68.9	82.4	73.0	75.2

Table 10: Experimental results of HSSA Ablation Study on all dialogue understanding tasks.

Ablation of SSOs

- We remove one training objective each time while keeping the remaining four, each training objective contributes to the overall performance to some extent, indicating the multi-level SSOs are complementary
- DCL brings the most benefits, which implies the effectiveness of DCL on capturing the content-relatedness of context-context pairs

Method	JDDC			ECD		
	Corr.	MAP	MRR	Corr.	MAP	MRR
DIALOG-POST	**82.90**	**69.95**	**79.87**	83.91	71.65	**81.72**
w/o DRM	82.84	69.93	79.90	**83.95**	71.64	81.72
w/o DSM	82.76	69.16	78.65	83.62	**71.69**	81.24
w/o DUC	81.96	69.25	79.69	83.91	71.64	81.72
w/o DUP	81.75	68.99	79.13	83.58	71.18	81.71
w/o DCL	77.98	61.21	75.33	80.16	67.35	79.06

Table 11: Experimental results of SSOs Ablation Study on all dialogue representation tasks.

Method	IC	Senti	CtxQ	CtxR	Average
DIALOG-POST	**91.8**	**78.1**	**92.4**	**87.9**	**87.5**
w/o DRM	91.2	77.9	91.8	87.0	87.0
w/o DSM	91.0	77.4	90.9	86.9	86.6
w/o DUC	89.7	77.4	90.3	85.1	85.6
w/o DUP	91.0	77.8	91.2	86.7	86.7
w/o DCL	89.0	77.0	89.6	86.5	85.5

Table 12: Experimental results of SSOs Ablation Study on all dialogue understanding tasks.

Experimental Results of POSPAN

- All post-training models improve upon the strong baseline DeBERTaV3, highlighting the effectiveness of post-training.
- Span-level masking methods outperform single-token masking, showing their advantage in capturing critical language semantics.
- POSPAN achieves the best performance across tasks, demonstrating the importance of position constraints in span masking.

Notation	Distribution	F_M	F_D
Pois	Poisson	$\lambda = 4$	$\lambda = 5$
Norm	Normal	σ=1,μ=4	σ=1,μ=5
Geo	Geometric	p=0.2	p=0.1
Rand	Uniform	a=1,b=5	a=4,b=6

Table 1: Hyper-parameters of different distributions. We tune hyper-parameters of the distributions via grid search and find the best settings.

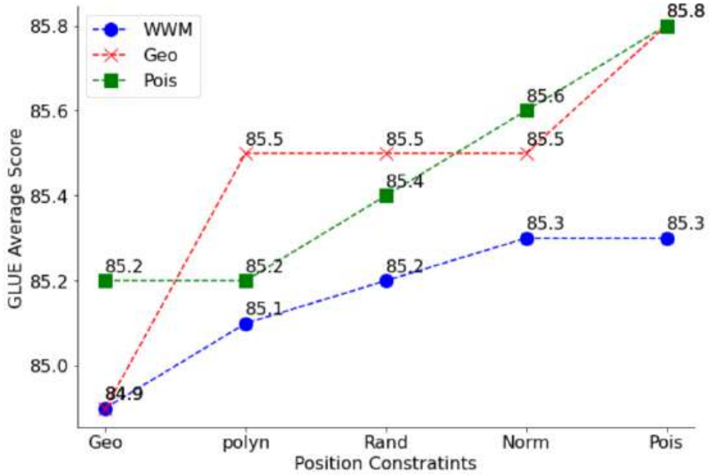


Figure 1: The model performance of POSPAN with different position constraints (x-axis).

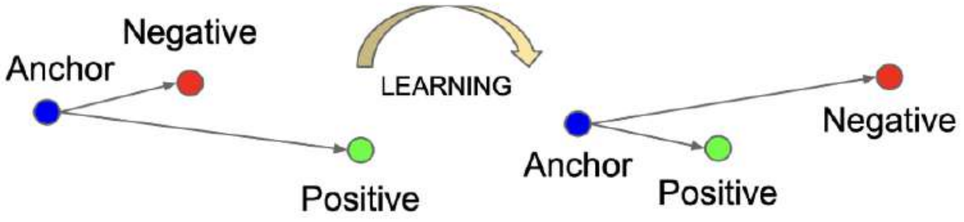
Method	CoNLL	MNLI(m/mm)	MRPC	QNLI	BoolQ	COPA	ReCoRD	SQuAD	RACE
DeBERTaV3 (He et al., 2021a)	94.9	88.1/88.3	87.0	92.4	80.1	70.3	56.5/44.6	84.8/82.0	52.0
MLM (Devlin et al., 2019)	95.3	88.2/88.5	88.4	92.5	80.5	70.9	56.3/44.9	84.8/82.1	52.1
Fixed	95.3	88.2/88.6	88.2	92.8	80.6	72.9	56.5/44.9	84.7/82.2	52.2
N-gram (Cui et al., 2020)	95.3	88.2/88.5	88.6	93.0	81.2	73.5	56.7/45.2	84.9/82.2	52.4
WWM (Cui et al., 2021)	95.2	88.2/88.5	88.0	92.7	80.8	71.8	56.4/44.7	84.8/82.2	52.3
Geo (Joshi et al., 2020)	95.7	88.5/88.7	88.9	93.1	81.3	73.2	56.8/45.1	85.0/82.5	52.5
Pois (Lewis et al., 2020)	95.6	88.4/88.7	87.5	93.0	81.0	73.9	56.7/45.1	85.1/82.5	52.3
POSPAN(WWM-*Norm*)	95.5	88.3/88.5	88.5	93.1	80.9	73.3	56.9/45.0	84.8/82.3	52.5
POSPAN(*Geo-Pois*)	**95.9**	88.8/89.0	**89.2**	**93.4**	81.6	**75.7**	**57.3/45.6**	85.4/82.5	52.8
POSPAN(*Pois-Pois*)	95.8	**88.9/89.3**	88.2	93.2	**81.9**	75.6	57.1/45.3	**85.6/82.7**	**53.1**

Table 2: Experimental results of POSPAN. POSPAN(*Geo-Pois*) denotes $F_M \sim Geo$ and $F_D \sim Pois$. CoNLL and SQuAD represent ConNLL 2003 and SQuAD v2.0. MNLI (m/mm) represents the two versions of MNLI, MNLI-matched and MNLI-mismatched. The complete evaluation results are reported in Appendix A.4.

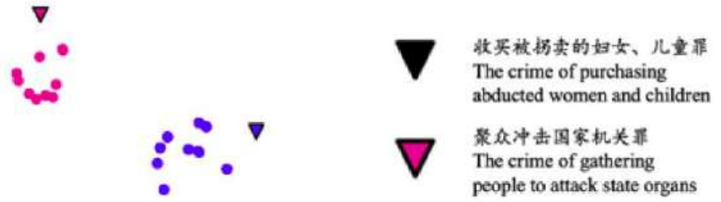
Method	MNLI (m/mm)	QNLI	QQP	MRPC	RTE	CoLA	SST-2	STS-B	Avg.
BERT-base (Devlin et al., 2019)	74.4/75.5	85.3	81.6	78.3	63.1	58.1	91.4	88.7	77.3
MLM (Devlin et al., 2019)	74.8/75.8	86.3	83.1	77.2	64.1	57.9	91.6	88.3	77.7
Fixed	74.6/75.6	86.4	83.2	80.6	62.8	59.5	92.1	89.9	78.3
N-gram (Cui et al., 2020)	74.5/75.2	86.4	83.2	80.4	64.0	59.3	91.8	90.3	78.4
WWM (Cui et al., 2021)	74.5/75.7	85.9	82.6	77.6	63.4	**61.6**	92.0	90.2	78.2
Geo (Joshi et al., 2020)	74.9/75.8	86.1	82.5	81.0	64.4	60.2	91.5	90.4	78.5
Pois (Lewis et al., 2020)	75.2/75.5	86.9	82.9	81.2	63.9	60.8	92.1	90.0	78.7
POSPAN(WWM-*Norm*)	76.0/**76.9**	**87.4**	83.5	78.5	65.9	60.8	**93.1**	90.5	79.2
POSPAN(*Geo-Pois*)	75.9/76.2	87.2	83.9	82.4	64.3	59.9	92.1	91.2	79.2
POSPAN(*Pois-Pois*)	**76.2**/76.7	87.3	**84.1**	**82.4**	**66.1**	59.4	92.9	**91.4**	**79.6**

Table 7: Experimental results of POSPAN in GLUE with BERT as base model. POSPAN(*Geo-Pois*) denotes $F_M \sim Geo$ and $F_D \sim Pois$. MNLI (m/mm) represents the two versions of MNLI, MNLI-matched and MNLI-mismatched.

Label Anchored Contrastive Learning



(Image source: Schroff et al. 2015)



Preliminary

- Contrastive Learning (CL) involves bringing an anchor and a 'positive' sample closer in the embedding space while distancing the anchor from 'negative' samples. In self-supervised CL, positive pairs are created through data augmentations, and negative pairs are formed with random samples from the mini-batch.
- Supervised Contrastive Learning (SCL) uses label information to form positive pairs, pulling examples from the same class closer and separating those from different classes, utilizing label semantics instead of just data augmentation.
- Label Embedding (LE) focuses on learning label representations in classification tasks, capturing class information to enhance task understanding.

Motivation

- CL under supervised learning is not fully explored because the label information can be better utilized.
- On the one hand, labels are usually not merely categorical indices in the label vocabulary, but also contain specific semantic meanings, especially in the language understanding tasks. Thus labels can be used as positive/negative samples or anchors when calculating contrastive loss.
- On the other hand, label embedding enjoys a built-in ability to leverage alternative sources of information related to labels, such as class hierarchies or textual descriptions.

Our Proposed Approach (LaCon)

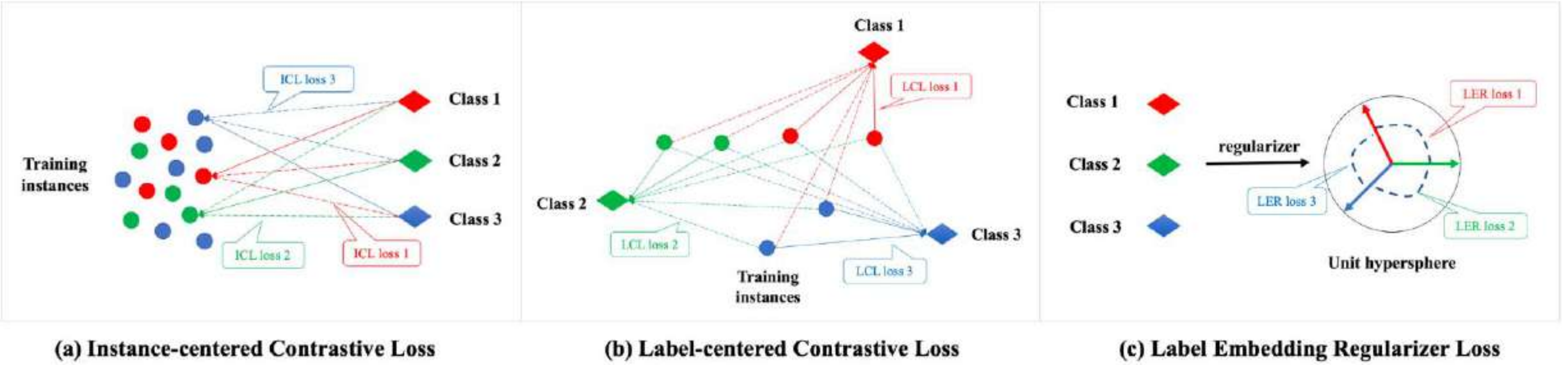


Figure 1: Overview of LaCon. The full line is the similarity between a instance and corresponding label, and the dash line is the similarity between the mismatched instance and label. The lines with the same color denote the per-instance or per-label loss.

$$\mathcal{L}'_{ICL} = -\frac{1}{N}\sum_{k=1}^{m}\sum_{x_i,y_i} log\frac{exp(sim(h^k_{x_i}, l^k_{y_i})/\tau)}{\sum_{1\leq p\leq C} exp(sim(h^k_{x_i}, l^k_p)/\tau)}$$

$$\mathcal{L}_{LCL} = -\frac{1}{|P|}\sum_{p\in P}\sum_{a\in A(p)} log\frac{exp(sim(L_p, H_a)/\tau)}{\sum_{b\in B(p)} exp(sim(L_p, H_b)/\tau)}$$

$$\mathcal{L}_{LER} = avg(\sum_{i\neq j}(exp(1.0 + sim(L_i, L_j)) - 1.0)$$

$$pred = \arg_{max}(j)\{score_j | sim(H_x, L_j)\}$$

$$\mathcal{L} = \mathcal{L}'_{ICL} + \mathcal{L}_{LCL} + \lambda * \mathcal{L}_{LER}$$

$$m = conv_{max}(G), \text{ where } G_{ij} = \frac{< L_i, E_j >}{||L_i|| \cdot ||E_j||}$$

$$z = \sum_i \beta_i E_i, \text{ where } \beta = softmax(m)$$

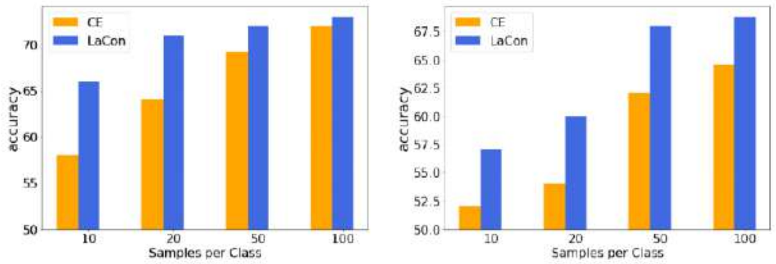
Experiments

Methods	YelpRev	DBPedia	Tnews	QNLI	RTE	QQP	MRPC	CoLA
CE	82.0±0.5	98.7±0.3	54.5±0.3	87.1±0.2	67.3±1.9	82.2±0.5	85.6±1.6	60.9±0.8
LEAM	82.1±0.6	98.7±0.5	54.1±0.3	87.2±0.7	67.3±1.3	81.9±0.5	85.6±1.3	60.9±1.0
LSAN	82.2±0.6	98.7±0.7	54.9±0.8	87.1±0.3	69.7±1.0	81.2±0.5	86.1±0.7	61.6±0.9
CE+CL	82.2±0.6	98.5±0.5	53.9±0.5	87.3±0.3	67.8±1.5	82.4±0.3	83.1±0.7	61.1±0.7
CE+SCL	81.4±0.8	98.5±0.6	54.6±0.2	87.7±0.1	69.1±2.2	82.5±0.6	88.1±0.9	62.3±0.6
LaCon-vanilla	82.3±0.5	98.9±0.5	**56.8±0.6**	88.1±0.2	71.4±0.7	82.8±0.5	87.5±0.9	62.4±1.0
LaCon-fusion	**83.1±0.8**	**99.5±0.2**	56.7±0.3	**88.4±0.3**	**72.2±0.9**	**83.7±0.5**	**88.6±0.7**	**62.8±0.5**

Table 2: The experimental results for the Language Understanding Tasks. Best scores for each dataset are highlight in **bold** (all with significance value $p < 0.05$).

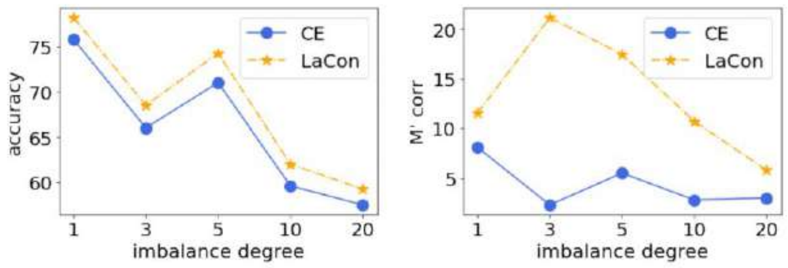
Methods	MRPC	RTE	CoLA
LaCon-vanilla	**87.5±0.8**	**71.4±0.7**	**62.4±1.1**
LaCon w/ $\mathcal{L}_{ICL}$	87.0±1.2	69.2±1.4	61.5±0.9
$-\mathcal{L}'_{ICL}$	86.6±1.3	68.1±0.9	61.3±0.7
$-\mathcal{L}_{LCL}$	87.1±0.6	70.5±0.8	61.2±1.1
$-\mathcal{L}_{LER}$	87.3±1.1	70.2±1.3	62.2±1.6
$-g$	86.8±0.6	69.6±0.6	62.2±0.9
BERT w/ g	84.9±1.7	66.5±2.1	61.0±1.2

Table 3: Ablation study. Best scores for each dataset are highlight in **bold** (all with significance test $p < 0.05$).



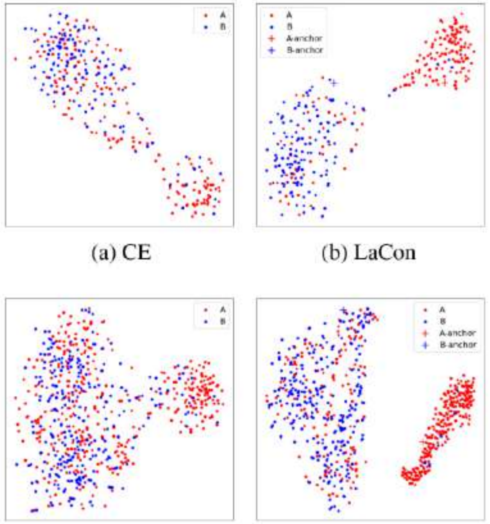
(a) QQP (b) RTE

Figure 2: Few-shot learning with different number of training samples.



(a) QNLI (b) CoLA

Figure 3: LaCon with Different Imbalance Degree (ρ).



(a) CE (b) LaCon

(c) CE (d) LaCon

Figure 4: Visualization of label and instance representations for MRPC (a&b) and CoLA (c&d) using T-SNE.

Speech Processing

Improving Disfluency Detection with Multi-scale Self Attention and Contrastive Learning
Peiying Wang, Chaoqun Duan, Meng Chen*, Xiaodong He
2023 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2023)

Leveraging Label Information for Multimodal Emotion Recognition
Peiying Wang, Sunlu Zeng, Junqing Chen, Lu Fan, Meng Chen*, Youzheng Wu, Xiaodong He
The 24th INTERSPEECH Conference (Interspeech 2023)

Gated Multimodal Fusion with Contrastive Learning for Turn-taking Prediction in Human-Robot Dialogue
Jiudong Yang*, Peiying Wang*, Yi Zhu, Mingchao Feng, Meng Chen*, Xiaodong He
2022 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2022)

Disfluency Detection

- Disfluency detection aims to remove the non-fluent word sequence from a sentence. Disfluency consists of three distinct parts: **interregnum**, **reparandum**, and **repair**.
- Specifically, the interregnum refers to filled pauses and discourse cue words, such as “uh”, “I mean”, etc.
- The reparandum means what the speaker wants to replace, and the repair is the content that the speaker intends to adopt to replace the reparandum.

Motivation:

- Previous works either design hand-crafted features or adopt CNN models to acquire repeating patterns based on the **word-to-word match patterns**, which may cause under-tagging and over-tagging problems

- Under-tagging: missing out some disfluencies
- Over-tagging: recognizing some correct phrases as disfluencies

Utt_1: I was we were so glad to meet her		
Out_1: we were so glad to meet her	(Correct)
Utt_2: I camp every month camp at least one weekend	
Out_2: I ~~every~~ camp at least one weekend	(Under-tagging)
Utt_3: I ’m sure within those people ’s minds it ’s justified	
Out_3: I ’m sure within those it ’s justified	(Over-tagging)

Table 1: Examples of Switchboard. Phrases with red color is the reparandum and the one with blue color is the repair.

- The word-to-word relations is not enough -> under-tagging
- Lacking constraints to keep the output fluent version consistent with the input in semantics -> over-tagging

Our Proposed Approach: MSAT + CL

- To tackle under-tagging issue, we propose a novel multi-scale self-attention (MSAT) module to acquire relations among different phrases, which can effectively capture the "rough copy" in the input.
- To tackle over-tagging issue, we devise an auxiliary CL loss [19] to constrain the training, which takes the fluent version of the input as a positive sample and delete some words from it to build a negative sample.

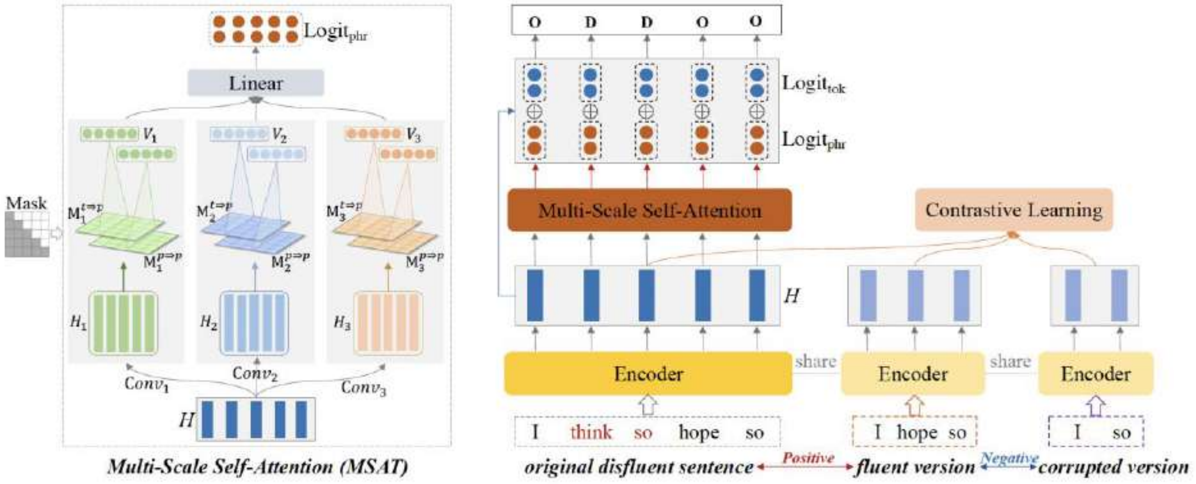


Fig. 1: The architecture of the model. The left part is multi-scale self-attention module and the right part is contrastive learning.

$$\mathbf{M}_f^{t\Rightarrow p} = \mathrm{softmax}\left(\frac{(\mathbf{H}\mathbf{W}_q)(\mathbf{H}_f\mathbf{W}_k)^T}{\sqrt{d}}\right),$$
$$\mathbf{M}_f^{p\Rightarrow p} = \mathrm{softmax}\left(\frac{(\mathbf{H}_f\mathbf{W}_q)(\mathbf{H}_f\mathbf{W}_k)^T}{\sqrt{d}}\right),$$

$$\mathbf{V}_f^{t\Rightarrow p} = \mathrm{Max\_Pooling}(\mathbf{M}_f^{t\Rightarrow p}),$$
$$\mathbf{V}_f^{p\Rightarrow p} = \mathrm{Max\_Pooling}(\mathbf{M}_f^{p\Rightarrow p}),$$

$$\mathcal{L}_{cl} = -\sum_{S\in\mathbb{S}} \log \frac{e^{\mathrm{sim}(\mathbf{H},\mathbf{H}^+)/\tau}}{e^{\mathrm{sim}(\mathbf{H},\mathbf{H}^+)/\tau} + e^{\mathrm{sim}(\mathbf{H},\mathbf{H}^-)/\tau}}$$

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda\mathcal{L}_{cl}$$

Experiments

- We conduct extensive experiments on both public dataset (Switchboard, 70k, English) and in-house dataset (Waihu, 24k, Chinese)
- Results show that our method outperforms the baselines and achieves better performance especially on long disfluency patterns

Models	SWBD			Waihu		
	P	R	F1	P	R	F1
HFLSTM* [2]	91.80	80.60	85.90	-	-	-
ACNN [4]	89.50	80.00	84.50	71.21	71.20	71.61
Trans-based [24]	91.10	84.10	87.50	-	-	-
MTL [5]	93.40	87.30	90.20	76.56	73.12	74.80
LSTM	90.56	74.80	81.93	71.46	58.28	64.20
w/ MSAT	91.06	77.34	83.62	72.29	62.95	67.30
CNN	90.94	74.85	82.11	77.53	60.93	68.23
w/ MSAT	91.01	79.96	85.52	77.01	68.41	72.28
BERT	93.92	87.44	90.56	79.12	74.79	76.90
w/ MSAT	**94.83**	88.42	91.51	81.73	75.17	78.32
BERT+MSAT+CL	94.03	**89.30**	**91.61**	**82.91**	**75.39**	**78.97**

Table 3: Performance comparison on SWBD and Waihu (* denotes performance is evaluated on combination of interregnum and reparandum).

Models	SWBD			Waihu		
	P	R	F1	P	R	F1
BERT	93.92	87.44	90.56	79.12	74.79	76.90
w/ 1-gram	93.72	**89.20**	91.40	81.70	74.48	77.92
w/ 2-gram	93.62	88.78	91.13	81.09	73.56	77.14
w/ 3-gram	93.93	88.74	91.26	81.21	73.44	77.13
w/ 4-gram	93.14	88.92	90.98	81.46	74.23	77.67
w/ MSAT	**94.83**	88.42	**91.51**	**81.73**	**75.17**	**78.32**

Table 4: Effects of different scales of phrases.

Neg. Sample	SWBD			Waihu		
	P	R	F1	P	R	F1
BERT	93.92	87.44	90.56	79.12	74.79	76.90
CL w/ RD	94.19	88.46	91.24	81.38	**75.49**	78.32
CL w/ ID	**94.52**	**88.67**	**91.48**	**82.81**	74.54	**78.46**

Table 5: Comparison of CL with different strategies of generating negative samples.

BERT: I camp every month camp at least one weekend
Ours: I camp every month camp at least one weekend
BERT: I work in I 'm a on the professional administrative
Ours: I work in I 'm a on the professional administrative

Table 6: Comparison of the BERT and our method on two examples of the test set of SWBD (Highlighted words mean the ground truth and underlined text denotes the prediction).

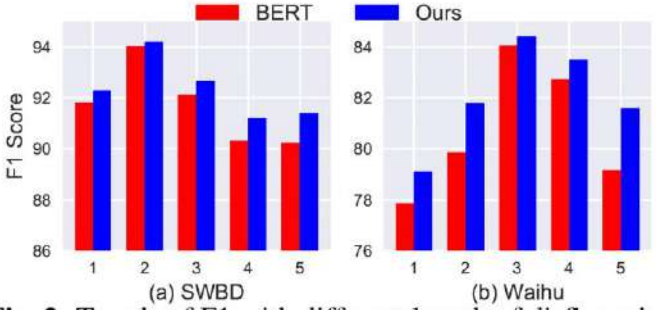


Fig. 2: Trends of F1 with different length of disfluencies.

Multimodal Emotion Recognition

Problem definition:

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

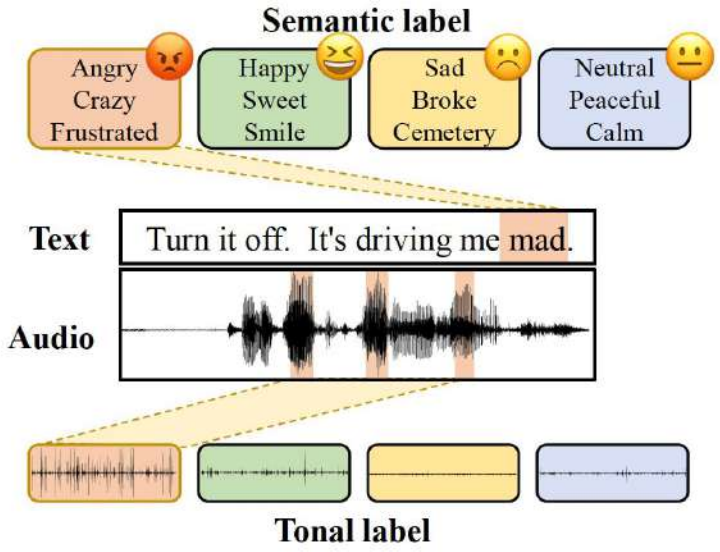


Figure 1: *Visualization of labels. The* ***semantic label*** *presents the emotion relevant words for each class, and the* ***tonal label*** *displays the waveforms generated by concatenating the key-frames under each class.*

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)

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

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

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

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

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

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

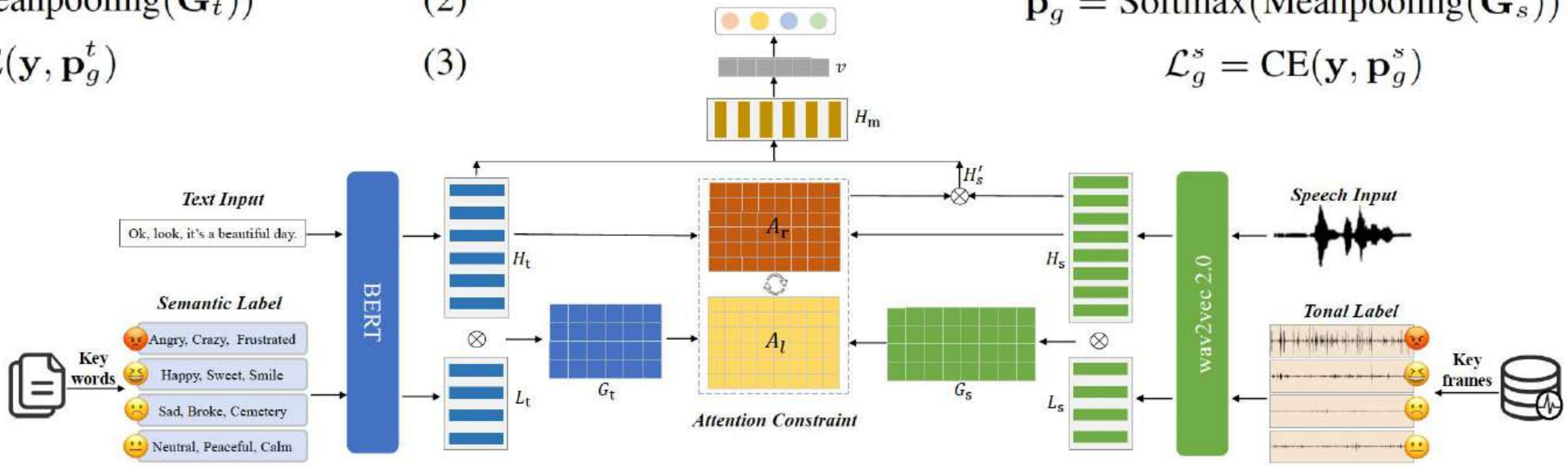


Figure 2: *The architecture of our proposed model LE-MER.*

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

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

$$\mathbf{A}_l = \mathbf{G}_t \cdot \mathbf{G}_s^T \tag{9}$$

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

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

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

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 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**

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

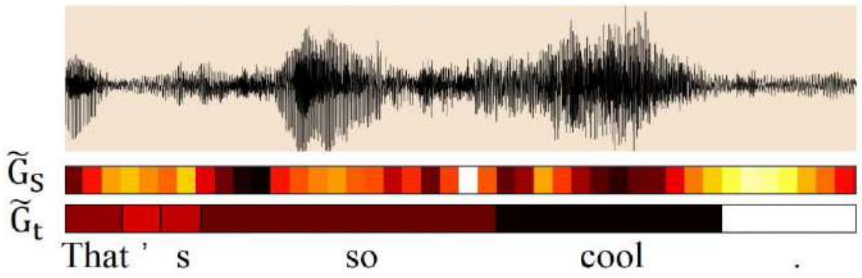


Figure 4: *Visualization of $\widetilde{\mathbf{G}}_s$ and $\widetilde{\mathbf{G}}_t$.*

Turn-taking Prediction

Background:

- **Turn-taking**, aiming to decide when the next speaker can start talking, is an essential component in building human-robot spoken dialogue systems.
- Given an utterance in a conversation, a **hold** means that the next utterance will be continued by the same speaker while a **switch** indicates that the next utterance will be uttered by the other speaker.

- **Endpointing**: end-of-turn detection, which assumes switch occurs when a speaker has stopped speaking and a period of silence comes out.
- **Barge-in**: handling user interruptions, where switch occurs when a speaker starts uttering before the other speaker finishes speaking.

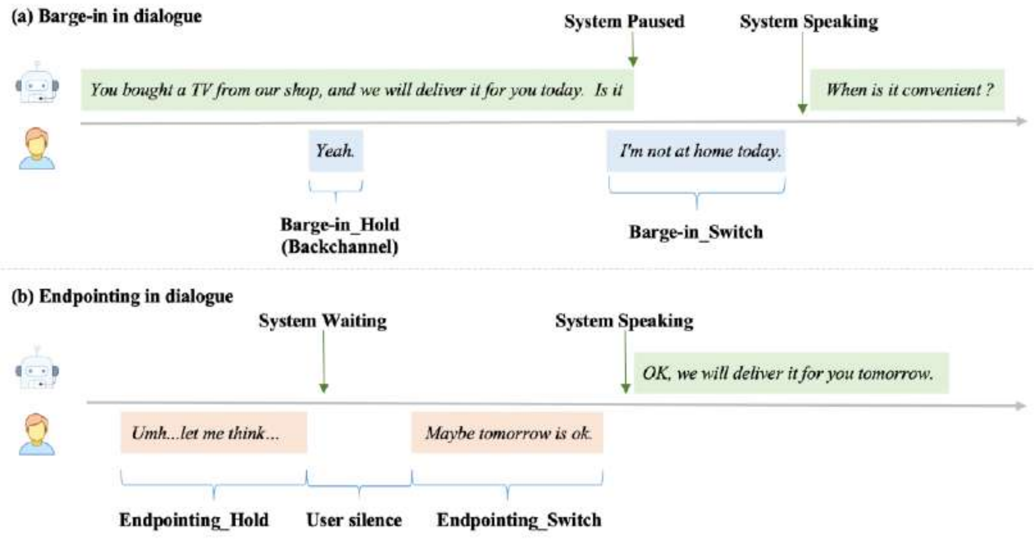


Fig. 1. Example of turn-taking in a conversation.

Gated Multimodal Fusion Model

- We collected a large-scale human-robot dialogue corpus from an online IVR system, featuring over 5,000 dialogues, including endpointing and barge-in situations
- We introduce a Gated Multimodal Fusion model (GMF) for turn-taking prediction in spoken dialogue systems
- To address class imbalance, we use data augmentation with self-supervised methods and contrastive learning to create samples for the minority class

Model	Endpointing		Barge-in	
	Acc	Macro-F1	Acc	Macro-F1
Random	0.490	0.467	0.512	0.465
MajVot$_{cls}$	0.744	0.425	0.765	0.432
LSTM$_{ens}$ [14]	0.752	0.646	0.789	0.642
MoE [16]	0.778	0.643	0.835	0.734
GMF	0.819	0.736	0.869	0.814
GMF w/ CL	**0.829**	**0.761**	**0.873**	**0.826**
w/o semantic	0.767	0.658	0.838	0.740
w/o context	0.783	0.699	0.852	0.786
w/o acoustic	0.788	0.708	0.820	0.732
w/o timing	0.791	0.707	0.853	0.791

Table 2. Turn-taking performance of different models.

Method	Endpointing		Barge-in	
	Acc	Macro-F1	Acc	Macro-F1
Concatenation	0.812	0.723	0.867	0.801
Summation	0.809	0.724	0.864	0.799
Multiplication	0.806	0.724	0.865	0.801
MFB [26]	0.813	0.728	0.861	0.798
GMF	**0.819**	**0.736**	**0.869**	**0.814**

Table 3. Turn-taking performance of different fusion methods.

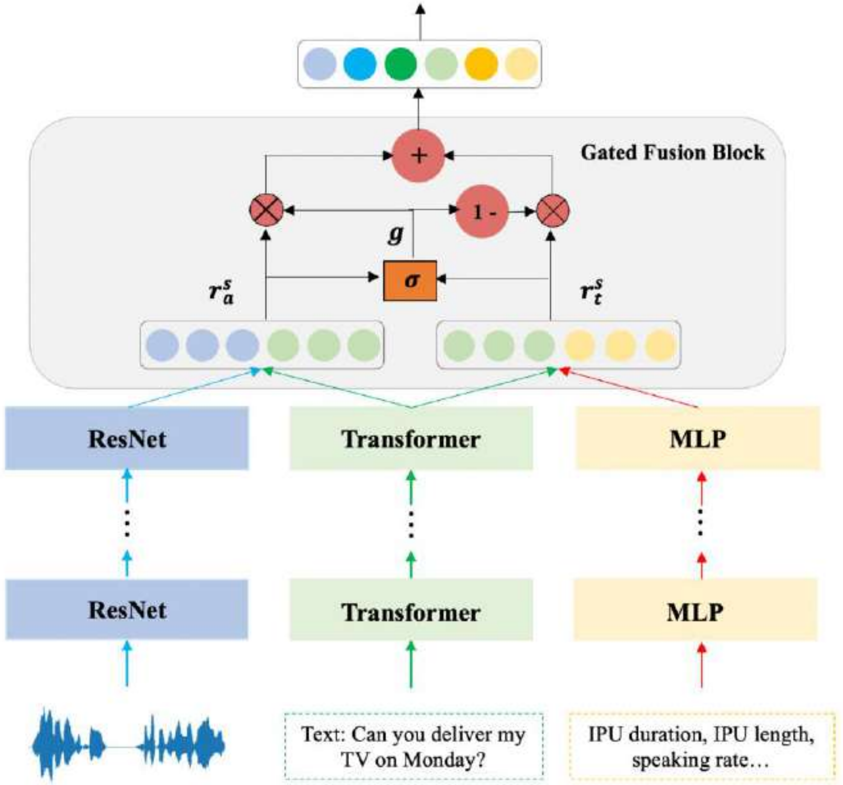
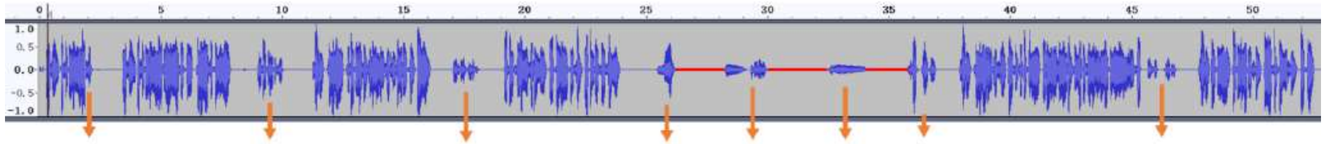


Fig. 2. Architecture of our proposed model GMF.



$$\mathbf{r}^s = Transformer_{Encoder}(\mathbf{e})$$
$$\mathbf{r}^a = ResNet_{Encoder}(\mathbf{f})$$
$$\mathbf{r}^t = MLP_{Encoder}(\mathbf{t})$$

$$g = \sigma(\mathbf{W}_g \cdot [\mathbf{r}_a^s, \mathbf{r}_t^s])$$
$$\mathbf{r} = g \cdot \mathbf{r}_a^s + (1-g) \cdot \mathbf{r}_t^s$$
$$\hat{\mathbf{y}} = \sigma(\mathbf{W}_f \mathbf{r} + b)$$

Multimodal Learning

Query Prior Matters: A MRC Framework for Multimodal Named Entity Recognition
Meihuizi Jia, Xin Shen, Lei Shen, Jinhui Pang, Lejian Liao, Yang Song, Meng Chen*, Xiaodong He
The 30th ACM International Conference on Multimedia (ACM Multimedia 2022)

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)

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)

Multimodal NER

Background:

- Given a sentence-image pair, MNER is required to recognize named entities of different types (mainly persons, locations, and organizations labeled as PER, LOC, and ORG respectively) in the sentence with extra image assistance

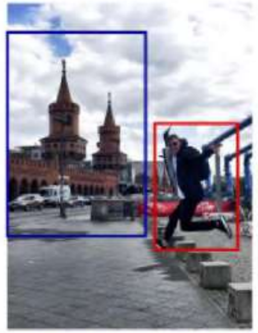
Motivations:

- Previous works use attention mechanisms to align and fuse sentence-image pairs, but these implicit alignments of entity types and image regions are difficult to interpret and evaluate.
- Extra visual grounding toolkit grounds phrases or sentences to image regions, but explicit relations between type-region pairs are not utilized. Additionally, tasks like MNER and visual grounding suffer from biased data, leading to inaccurate region detection.
- The MRC framework, known for its strong language understanding, is increasingly used in natural language tasks. Unlike sequence labeling, it better utilizes label prior information for MNER.

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

Our Approach: Two-stage MRC-based Multimodal NER

- Stage 1: fine-tune the pre-trained FA-VG model (Acc 79.96%) to locate top-k region candidates with their confidence scores for each entity
- Stage 2: a joint-training framework with Region Weights Estimation, Entity Span Prediction and Existence Detection



Query 1: "Location: Country, city, …"
Query 2: "Person: People's name …"

Formulating MNER as a QA Task ☺

Table 1: Examples for transforming entity types to queries.

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 company, government party, school government, and news organization.

Prompt!!!

Predict whether there exists the specific entity type in the utterance

Use two binary classifiers to tag the entity span from a sentence

Utilize the VG model to identify the top-k visual regions with confidence scores and encourage the model to assess the weights of these regions

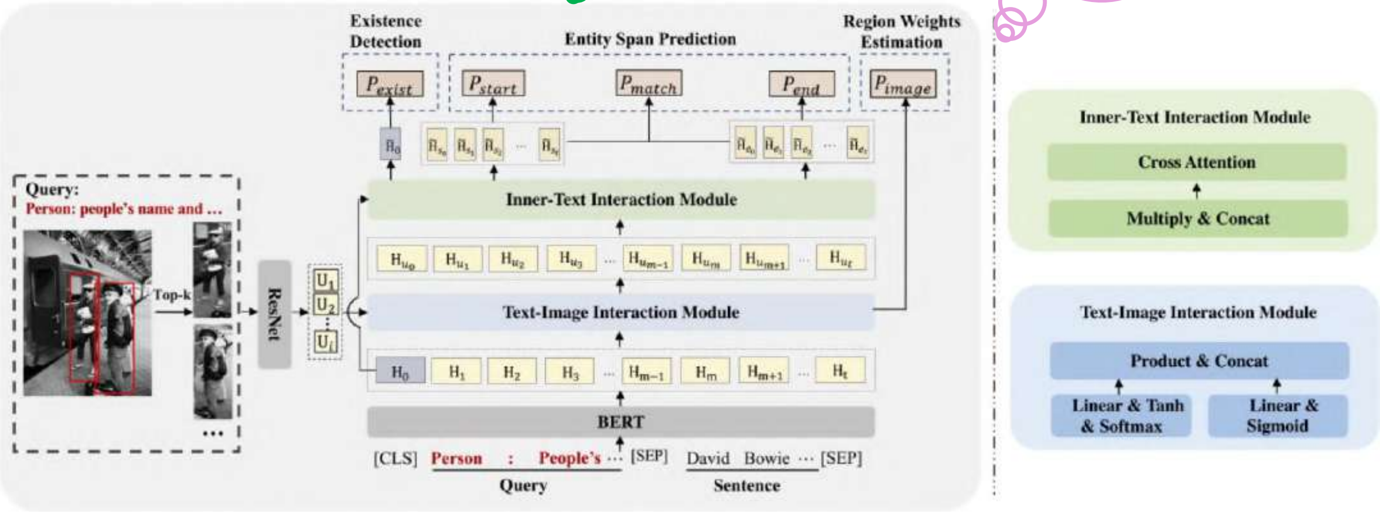


Figure 2: Overview of our MRC-MNER framework. The details of Multi-Level Modal Interaction are illustrated on the right.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{start} + \lambda_2 \mathcal{L}_{end} + \lambda_3 \mathcal{L}_{match} + \lambda_4 \mathcal{L}_{exist} + \lambda_5 \mathcal{L}_{image}$$

Experimental Results of Two-Stage Model

Table 2: Performance comparison on two MNER datasets. We refer to the results of UMGF from [40] and other results from [33].

Methods	Twitter2015							Twitter2017						
	Single Type (*F1*)				Overall			Single Type (*F1*)				Overall		
	PER	**LOC**	**ORG**	**OTH.**	*Pre.*	*Rec.*	*F1*	**PER**	**LOC**	**ORG**	**OTH.**	*Pre.*	*Rec.*	*F1*
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	-	-	-	-	-	-	-
MRC-MNER-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
MRC-MNER-VG (Ours)	84.88	81.43	61.06	39.93	78.08	70.75	74.22	91.83	85.84	83.09	72.11	88.59	84.16	86.32
MRC-MNER (Ours)	**85.71**	81.97	61.12	40.20	**78.10**	71.45	74.63	**92.64**	**86.47**	83.16	**72.66**	**88.78**	85.00	**86.85**

Table 3: Ablation study of MRC-MNER.

Methods	Twitter2015			Twitter2017		
	Pre.	*Rec.*	*F1*	*Pre.*	*Rec.*	*F1*
MRC-MNER	78.10	71.45	74.63	88.78	85.00	86.85
w/o RWE	77.24	70.86	73.91	88.11	84.26	86.14
w/o ED	77.63	70.95	74.14	88.29	84.36	86.29
w/o RWE+ED	76.82	70.24	73.38	87.72	83.99	85.81

Table 4: Results with different query transformations.

Query transformation	Twitter2015	Twitter2017	Flickr30K
	F1	*F1*	*Accuracy*
Keyword	74.03	86.11	73.37
Rule-based template filling	74.01	86.15	70.84
Keyword's Wikipedia	73.94	86.07	69.68
Keyword + Annotation	74.63	86.85	79.96

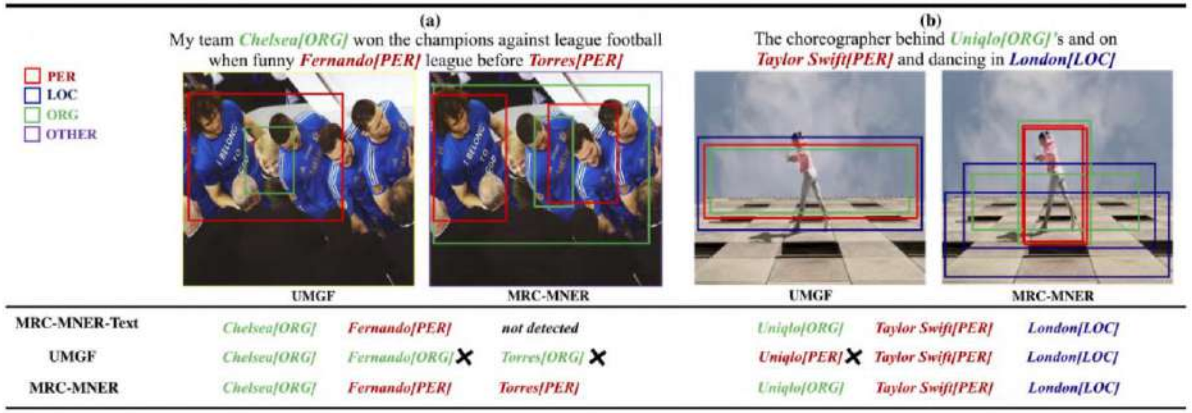


Figure 4: Example comparison among MRC-MNER, MRC-MNER-Text, and UMGF.

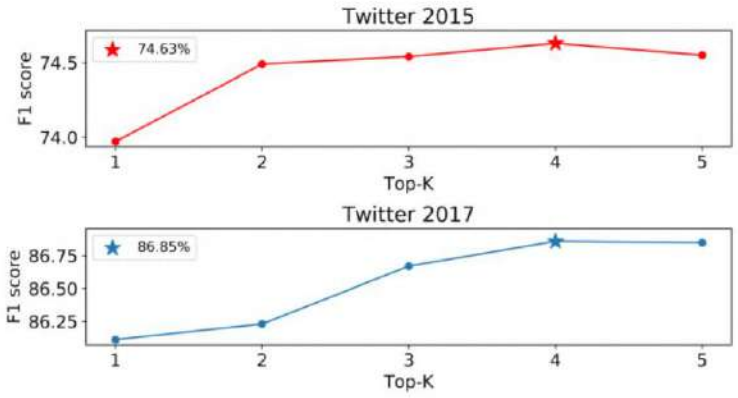


Figure 5: Results with different numbers of region candidates.

Why Not 1-Stage End-to-End Model?

Motivation:

- The two-stage manner is not fancy, which may incorporate inaccurate visual regions from the first stage will hurt the final results (error propagation)
- We propose an end-to-end MRC framework for Multimodal Named Entity Recognition with Query Grounding (MNER-QG). This joint-training approach forces the model to explicitly align entity spans with the corresponding visual regions, and further improves the performance of both named entity recognition and query grounding

Dataset Construction

- Weak supervision: utilize transfer learning to fine-tune FA-VG model to annotate visual regions for Twitter2015/2017
- Manual annotation: hire crowd-sourced workers to annotate 1.3k high-quality alignment data

Total data volume	26,311
F.30K data (unmodified)	12,504
F.30K data + modified query data	12,504
Tw.15/17 data + query + b.-box	1,303
LOC query data	2,983 (F.30K) + 700 (Tw.15/17)
ORG query data	4,191 (F.30K) + 350 (Tw.15/17)
PER query data	4,362 (F.30K) + 253 (Tw.15/17)
OTHER query data	968 (F.30K)

Table 7: Statistics of our constructed VG corpus (F.30k and Tw.15/17 denote Flickr30k and Twitter2015/2017, respectively and b.-box denotes bounding box).

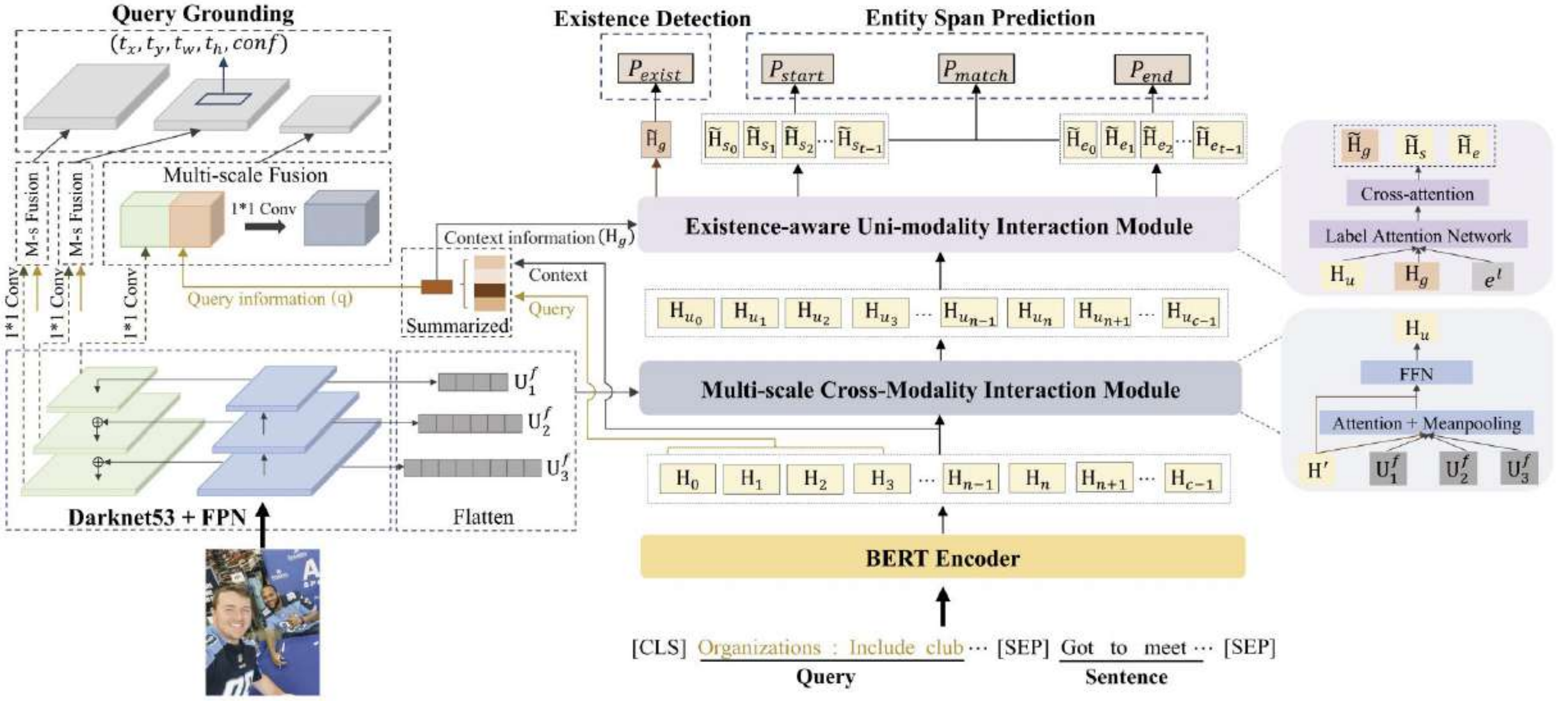


Figure 2: Overview of our MNER-QG framework (M-s Fusion denotes Multi-scale Fusion).

Experimental Results of End-to-End Model

Methods	Twitter2015							Twitter2017						
	Single Type (*F1*)				Overall			Single Type (*F1*)				Overall		
	PER	**LOC**	**ORG**	**OTH.**	*Pre.*	*Rec.*	*F1*	**PER**	**LOC**	**ORG**	**OTH.**	*Pre.*	*Rec.*	*F1*
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.

Methods	Twitter2015			Twitter2017		
	Pre.	*Rec.*	*F1*	*Pre.*	*Rec.*	*F1*
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 (W.S)	Twitter2015 (M.A)	Twitter2017 (W.S)	Twitter2017 (M.A)	Flickr30K
	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).)

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).

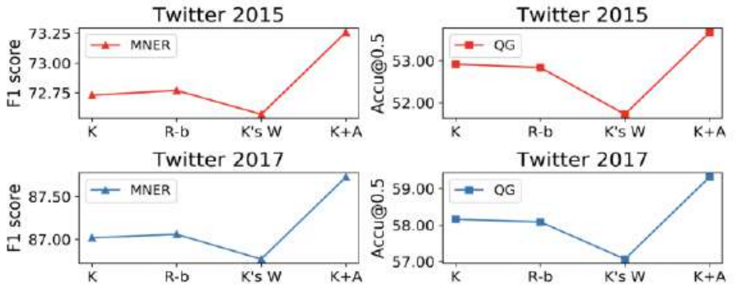


Figure 4: Results with different query transformations in MNER and QG on validation set (K, R-b, K's W, and K+A correspond to methods 1-4 of query transformations).

Multimodal Sentiment Detection

Modality Heterogeneity

- We focus on detecting the sentiment of multimodal posts in social media, i.e. given a <Text, Image> pair, predict the sentiment {Positive/Neutral/Negative}

Text: Mario finally has a smile on his face!	**Text:** The reason why this devoted dog is in critical condition will make you cry.
Sentiment Label: (Positive)	**Sentiment Label**: (Negative)

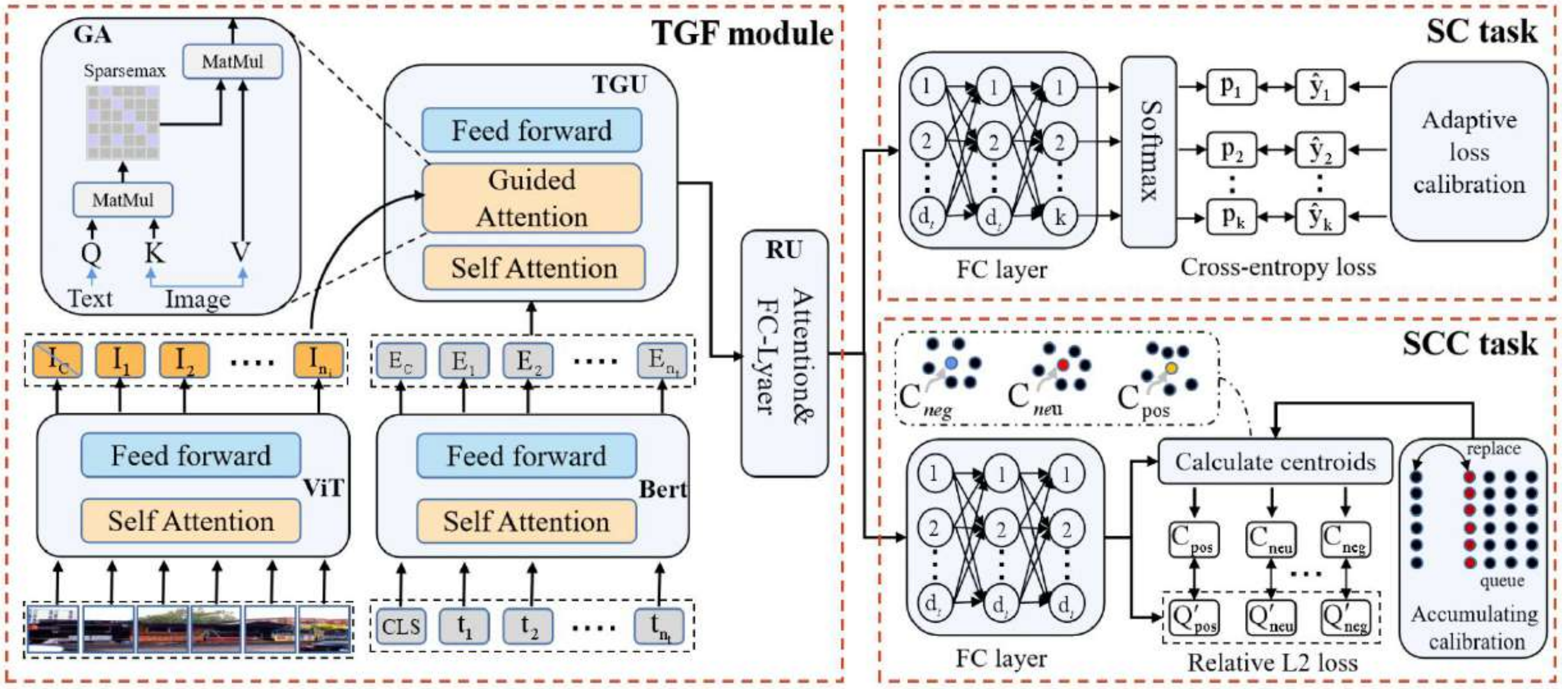
Figure 1: Examples of multimodal sentiment detection.

Problems		Solutions
Introducing redundant visual features during feature fusion	→	Propose Text-Guided Fusion module to leverage text data to dominate the fusion process
Causing feature shift in the representation space	→	Propose a Sentiment-based Congruity Constraint (SCC) task to restrain representation space
Inconsistent annotations for different modal data	→	Introduce an adaptive loss calibration (ALC) strategy to calibrate the training loss

Our Approach: Multi-View Calibration Network

To prevent sparse and redundant visual features, we introduce a **Text-Guided Fusion (TGF)** module that uses text to guide fusion. We apply Sparse-Attention with **sparsemax** to eliminate redundant features and highlight key sentiment-related image parts

We introduce an **adaptive loss calibration (ALC) strategy** to calibrate the training loss in the sentiment detection task, where the detection model is forced to be less confident for uncertain annotated labels.



To address feature shift, we introduce a **Sentiment-based Congruity Constraint (SCC)** task to refine representation space. The SCC task uses relative distance to cluster multimodal features around sentiment centroids based on sample labels. Additionally, we implement an **Accumulating Calibration (AC) strategy** to gather sampling data and compute sentiment centroids globally, overcoming minibatch limitations.

Experimental Results

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.

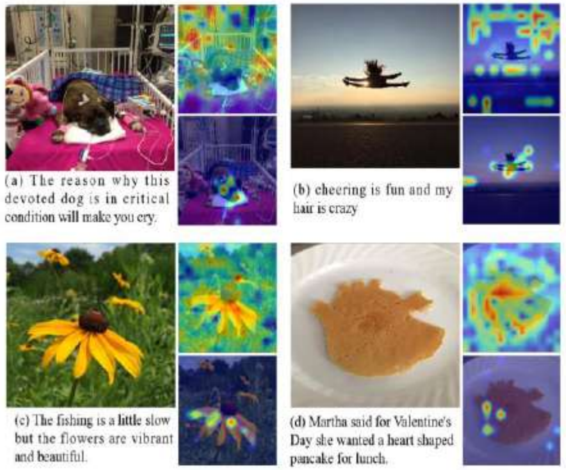
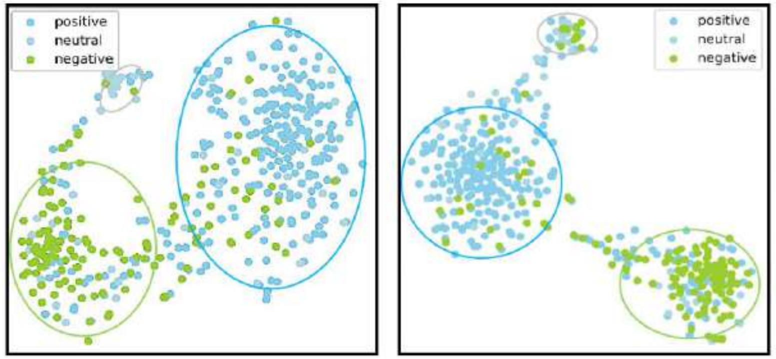


Figure 4: Attention visualization for sampling cases with Self-Attention (the upper one) and Sparse-Attention in TGF (the lower one).

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.



(a) Without SCC (b) With SCC

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

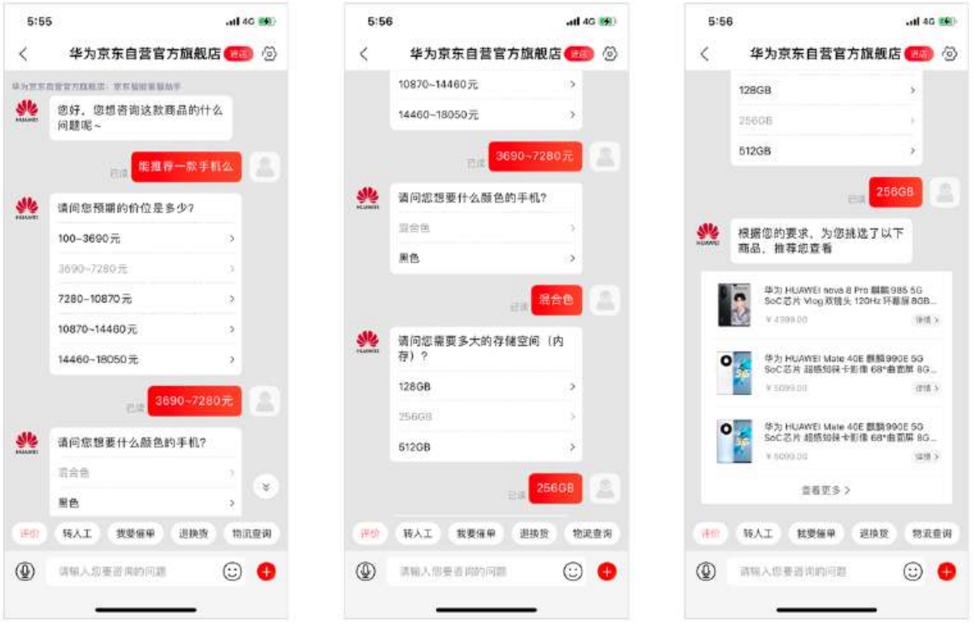
AI Applications

Text-based Chatbot for JD Smart Customer Service

- AlphaSales is a shopping assistant that not only answers questions about products, promotions, and shipping policies but also proactively recommends products to customers
- It serves 220,000 online stores on the JD.COM platform, catering to 500 million users

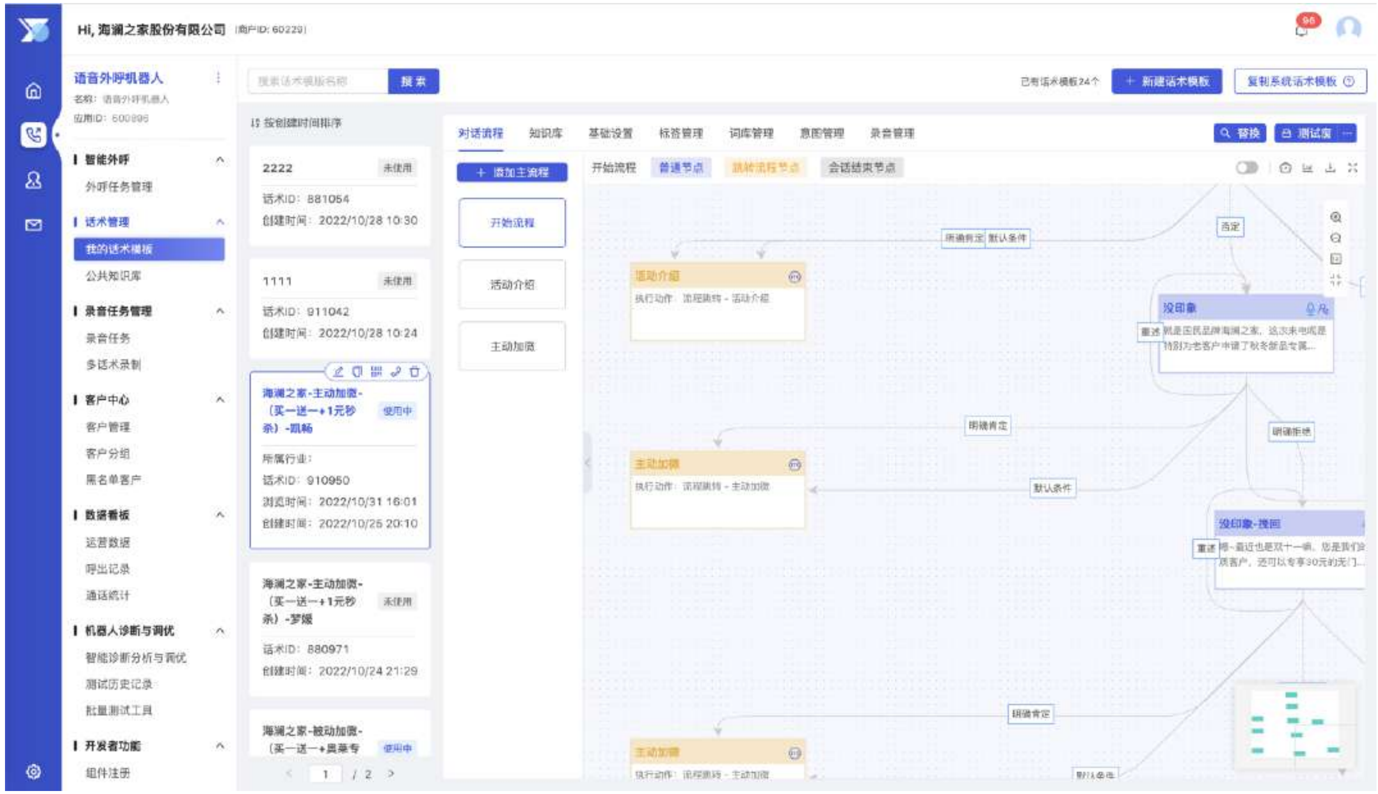
Conversational Recommendation

交互式导购



Speech-based AI Call System (Similar to Google Duplex)

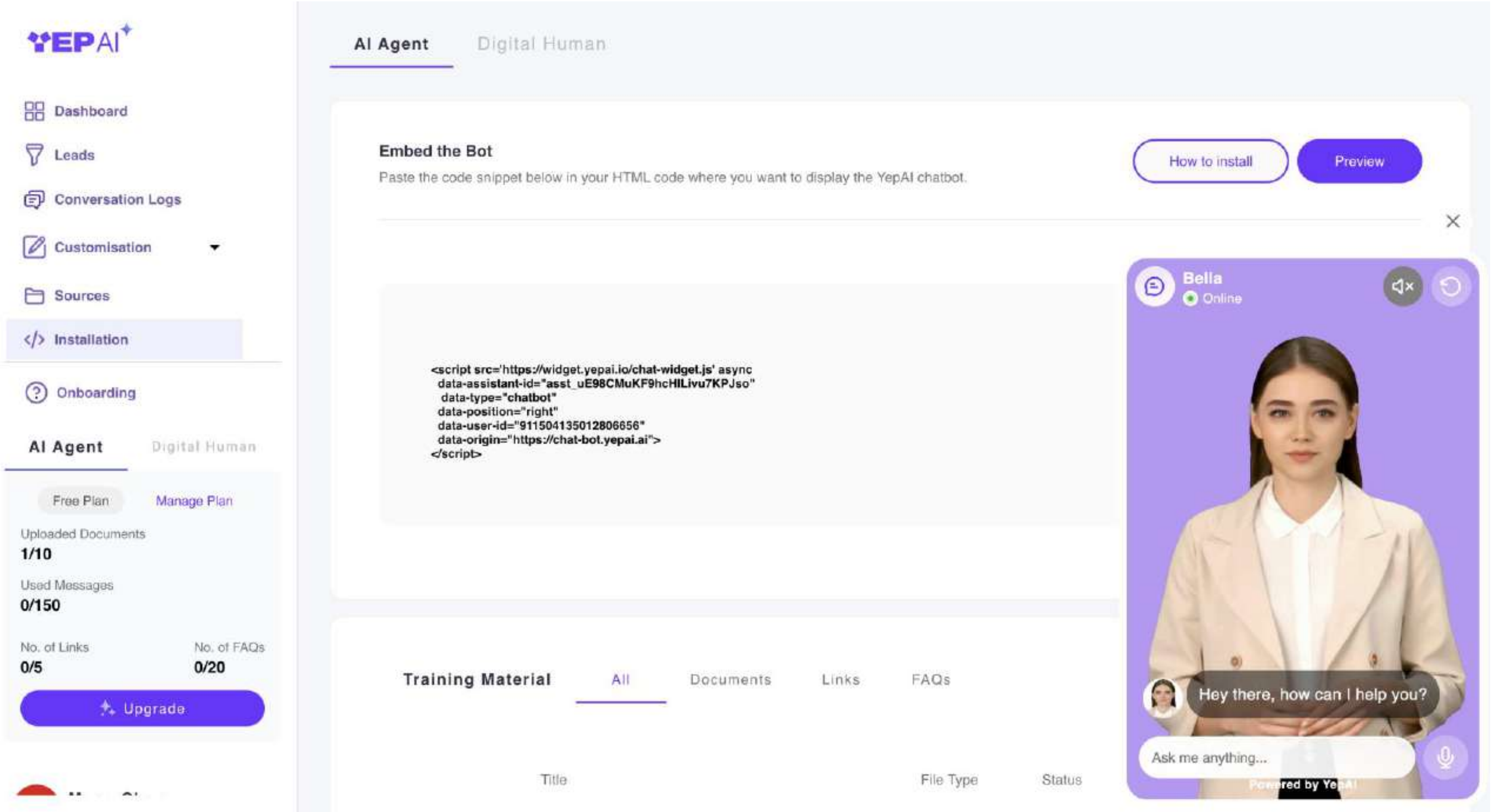
- The JD AI call system automates outbound calls to deliver essential information, including promotions and surveys, for online shops
- It serves over 200 brands, with a peak volume of more than 20 million calls in a single day

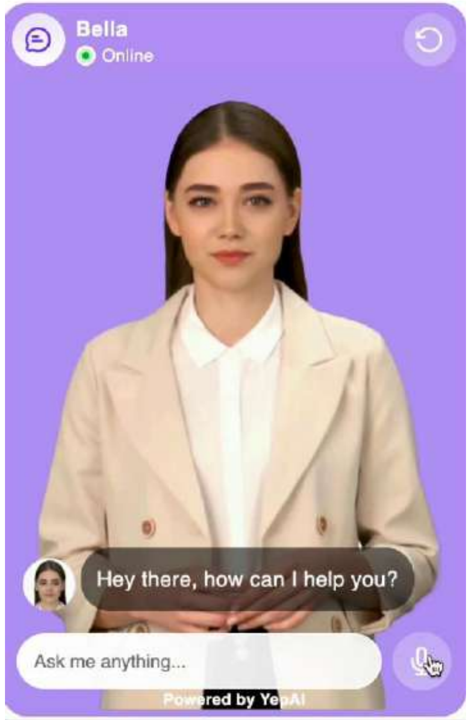


Demo audio

Avatar-based Real-time Virtual Assistant

- The Yep AI agent platform enables customers to create their own avatar-based virtual assistant in just 5 minutes!
- It is entirely powered by generative AI technology, with a latency of less than 3 seconds.





Have a try!
https://www.yepai.io

Take-aways

- The AI revolution in natural language processing (NLP) has had a profound impact on the broader AI landscape and is rapidly extending into fields like computer vision and speech.

- Bridging the modality gap is a critical challenge for achieving human-like understanding in a multimodal world.

- Multimodal and multilingual large models offer promising new opportunities to unify perception tasks, making this a valuable area for future exploration.

Q&A

Thanks!