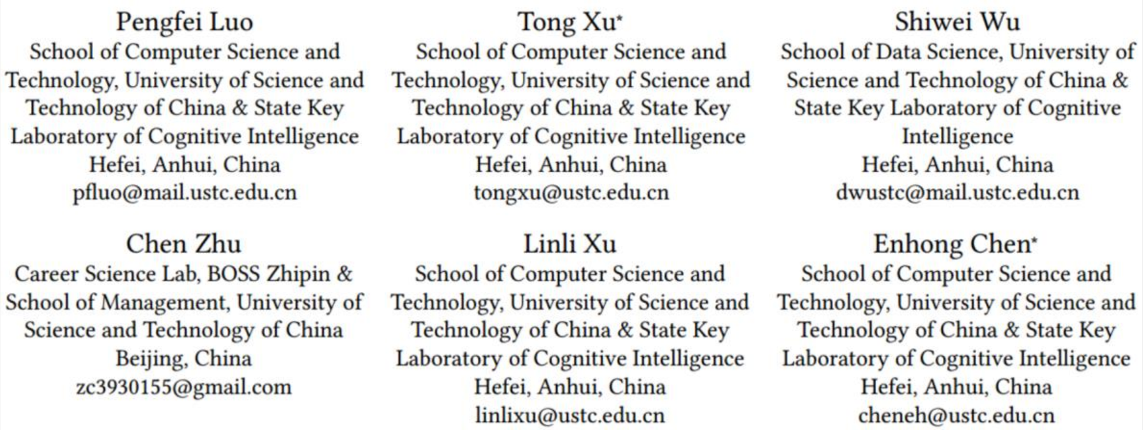
Multi-Grained Multimodal Interaction Network for Entity Linking



Pengfei Luo
School of Computer Science and Technology, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, Anhui, China
pfluo@mail.ustc.edu.cn

Tong Xu*
School of Computer Science and Technology, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, Anhui, China
tongxu@ustc.edu.cn

Shiwei Wu
School of Data Science, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, Anhui, China
dwustc@mail.ustc.edu.cn

Chen Zhu
Career Science Lab, BOSS Zhipin & School of Management, University of Science and Technology of China
Beijing, China
zc3930155@gmail.com

Linli Xu
School of Computer Science and Technology, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, Anhui, China
linlixu@ustc.edu.cn

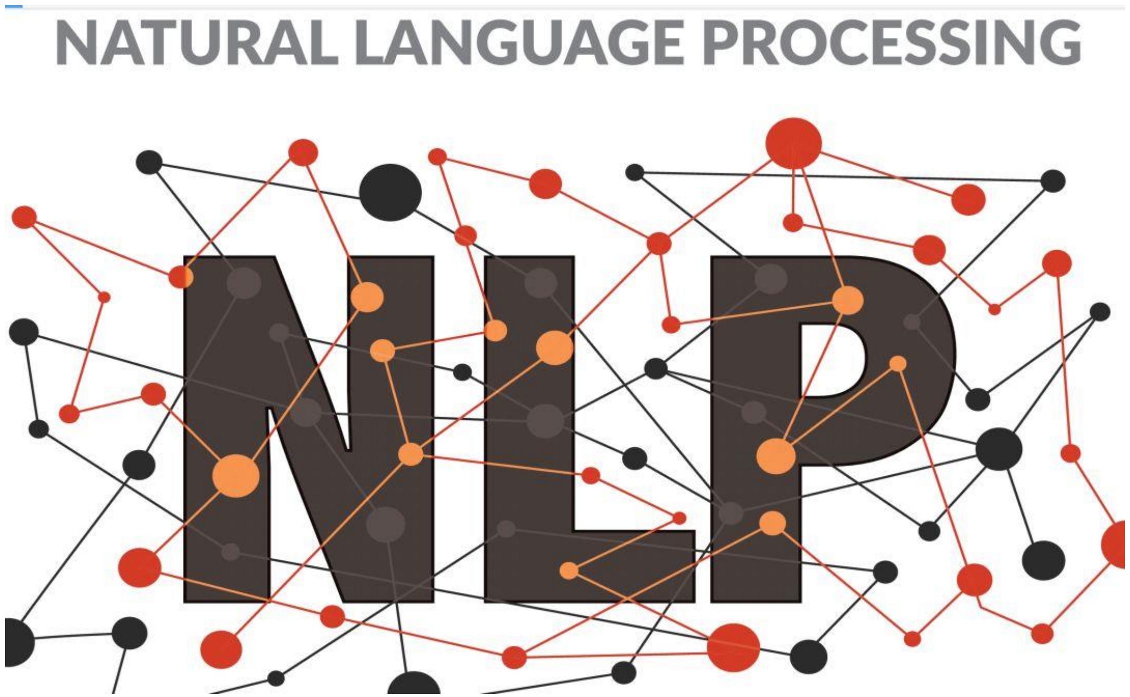
Enhong Chen*
School of Computer Science and Technology, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, Anhui, China
cheneh@ustc.edu.cn

code: https://github.com/pengfei-luo/MIMIC

2023. 9. 07 • ChongQing

2023_KDD

Reported by Junhao Cao



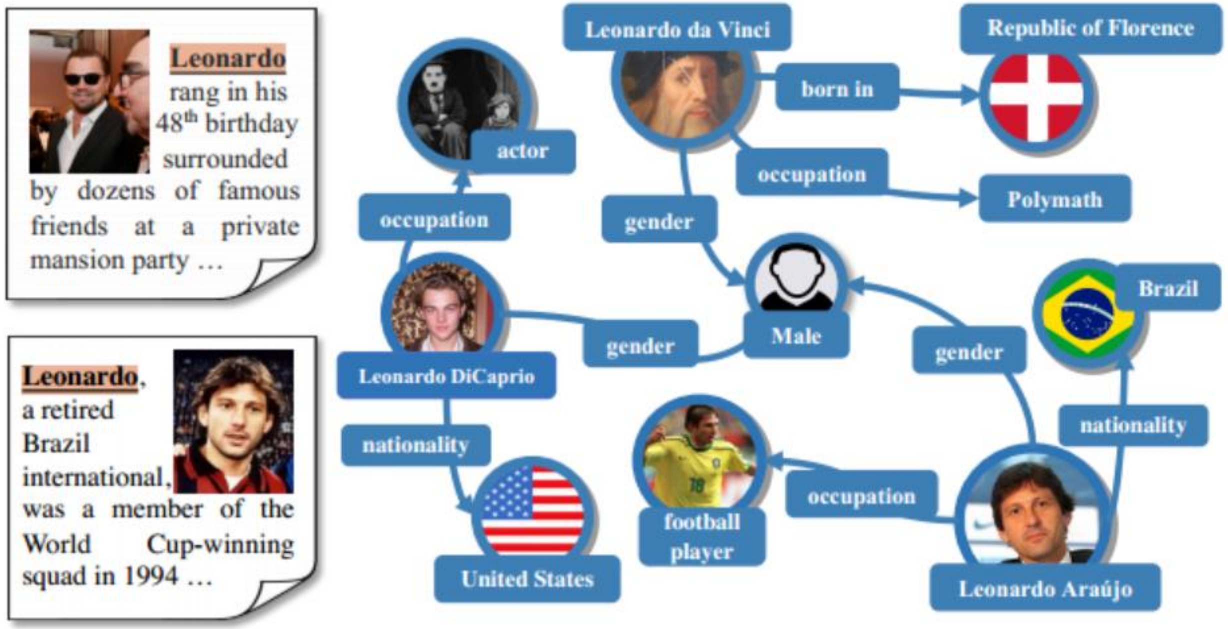
1.Introduction

2.Method

3.Experiments

Introduction

Problem Statement: given a mention $\mathbf{M}_j$, the task of multimodal entity linking targets to retrieve the ground truth entity $\mathbf{E}_i$ from the entity set $\mathcal{E}$ of knowledge base.



Leonardo rang in his 48th birthday surrounded by dozens of famous friends at a private mansion party …

Leonardo, a retired Brazil international, was a member of the World Cup-winning squad in 1994 …

Figure 1: Examples of multimodal entity linking. Left: two multimodal mentions. Right: multimodal knowledge graph.

Method

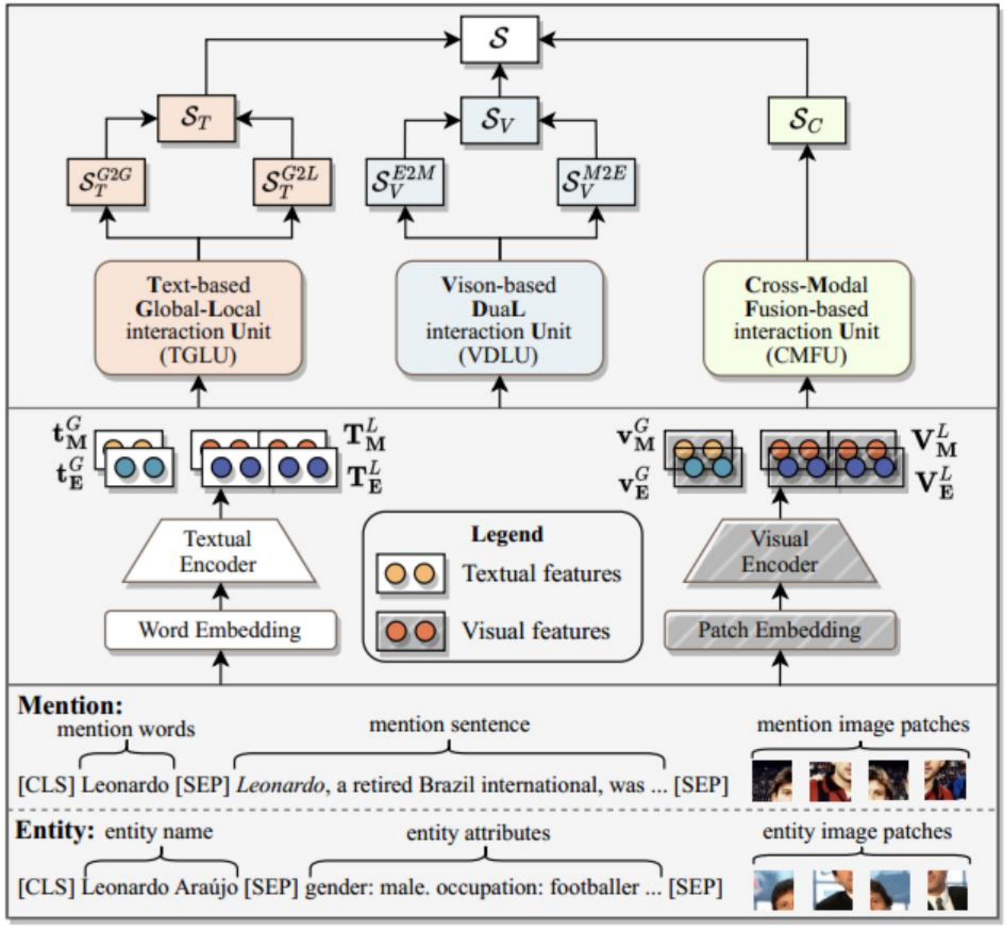

$\mathcal{S}$
$\mathcal{S}_T$
$\mathcal{S}_V$
$\mathcal{S}_C$
$\mathcal{S}_T^{G2G}$
$\mathcal{S}_T^{G2L}$
$\mathcal{S}_V^{E2M}$
$\mathcal{S}_V^{M2E}$
Text-based Global-Local interaction Unit (TGLU)
Vison-based DuaL interaction Unit (VDLU)
Cross-Modal Fusion-based interaction Unit (CMFU)
$\mathbf{t}_{\mathrm{M}}^{G}$
$\mathbf{t}_{\mathrm{E}}^{G}$
$\mathbf{T}_{\mathrm{M}}^{L}$
$\mathbf{T}_{\mathrm{E}}^{L}$
$\mathbf{v}_{\mathrm{M}}^{G}$
$\mathbf{v}_{\mathrm{E}}^{G}$
$\mathbf{V}_{\mathrm{M}}^{L}$
$\mathbf{V}_{\mathrm{E}}^{L}$
Textual Encoder
Word Embedding
Legend
Textual features
Visual features
Visual Encoder
Patch Embedding
Mention:
mention words
mention sentence
mention image patches
[CLS] Leonardo [SEP] *Leonardo*, a retired Brazil international, was ... [SEP]
Entity: entity name
entity attributes
entity image patches
[CLS] Leonardo Araújo [SEP] gender: male. occupation: footballer ... [SEP]

Figure 2: An overview of MIMIC. The bottom part is the input layer. The middle part is the encoding layer. The upper part is the multi-grained multimodal interaction layer.

Method

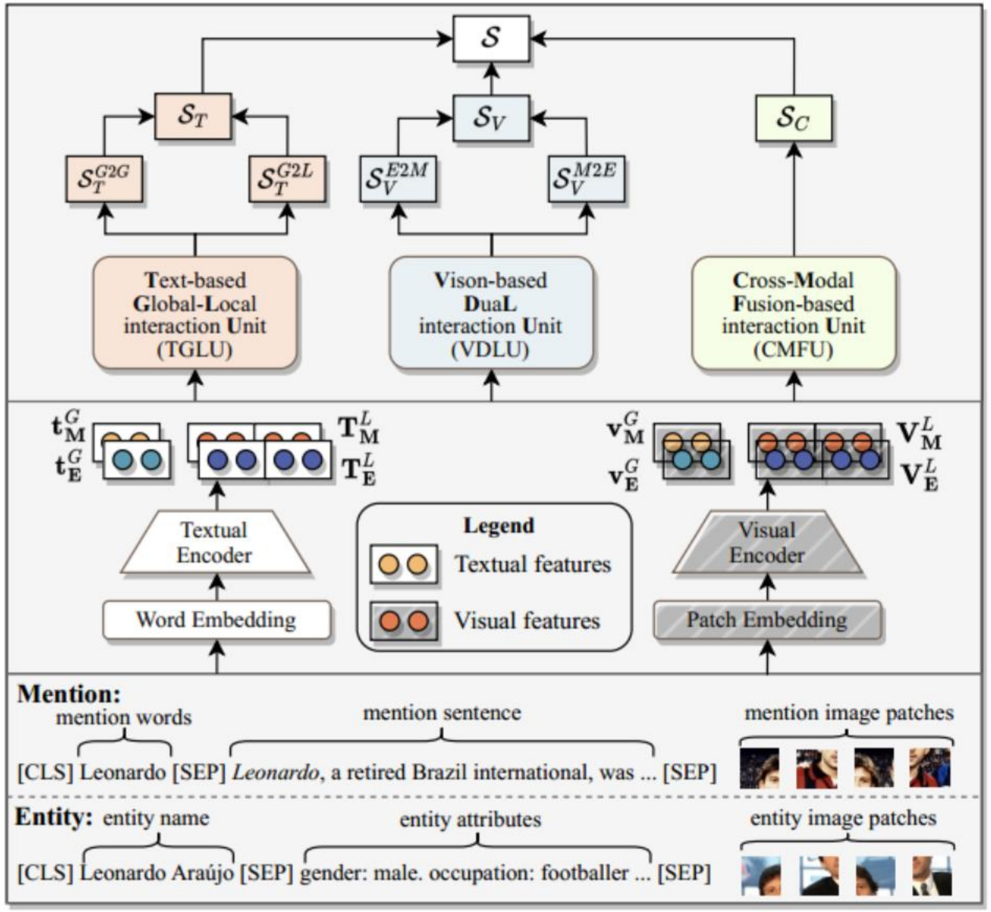

$\mathcal{S}$
$\mathcal{S}_T$
$\mathcal{S}_V$
$\mathcal{S}_C$
$\mathcal{S}_T^{G2G}$
$\mathcal{S}_T^{G2L}$
$\mathcal{S}_V^{E2M}$
$\mathcal{S}_V^{M2E}$
Text-based Global-Local interaction Unit (TGLU)
Vison-based DuaL interaction Unit (VDLU)
Cross-Modal Fusion-based interaction Unit (CMFU)
$\mathbf{t}_{\mathrm{M}}^{G}$
$\mathbf{t}_{\mathrm{E}}^{G}$
$\mathbf{T}_{\mathrm{M}}^{L}$
$\mathbf{T}_{\mathrm{E}}^{L}$
$\mathbf{v}_{\mathrm{M}}^{G}$
$\mathbf{v}_{\mathrm{E}}^{G}$
$\mathbf{V}_{\mathrm{M}}^{L}$
$\mathbf{V}_{\mathrm{E}}^{L}$
Textual Encoder
Legend
Textual features
Visual features
Visual Encoder
Word Embedding
Patch Embedding
Mention:
mention words
mention sentence
mention image patches
[CLS] Leonardo [SEP] *Leonardo*, a retired Brazil international, was ... [SEP]
Entity: entity name
entity attributes
entity image patches
[CLS] Leonardo Araújo [SEP] gender: male. occupation: footballer ... [SEP]

Figure 2: An overview of MIMIC. The bottom part is the input layer. The middle part is the encoding layer. The upper part is the multi-grained multimodal interaction layer.

$$\mathbf{E}_i = (\mathbf{e}_{n_i}, \mathbf{e}_{v_i}, \mathbf{e}_{d_i}, \mathbf{e}_{a_i}),$$

$\mathbf{E}_i$ represent entity name, entity images, entity description, and entity attributes,

$$\mathbf{M}_j = (\mathbf{m}_{w_j}, \mathbf{m}_{s_j}, \mathbf{m}_{v_j})$$

$\mathbf{m}_{w_j}, \mathbf{m}_{s_j}$ and $\mathbf{m}_{v_j}$ indicate the words of mention, the sentence in which the mention is located, and the corresponding image.

$$\theta^* = \max_{\theta} \sum_{(\mathbf{M}_j, \mathbf{E}_i) \in \mathcal{D}} \log p_\theta \left(\mathbf{E}_i | \mathbf{M}_j, \mathcal{E}\right), \quad (1)$$

$$\mathbf{I}_{\mathbf{E}_i} = [\mathrm{CLS}]\mathbf{e}_{n_i}[\mathrm{SEP}]\mathbf{e}_{a_i}[\mathrm{SEP}], \quad (2)$$

$$\mathbf{I}_{\mathbf{M}_j} = [\mathrm{CLS}]\mathbf{m}_{w_j}[\mathrm{SEP}]\mathbf{m}_{s_j}[\mathrm{SEP}]. \quad (3)$$

$$\mathcal{S}_T = \mathcal{U}_T(\mathbf{M}, \mathbf{E}) = (\mathcal{S}_T^{G2G} + \mathcal{S}_T^{G2L})/2, \quad (4)$$

$$\mathcal{S}_V = \mathcal{U}_V(\mathbf{M}, \mathbf{E}) = (\mathcal{S}_V^{E2M} + \mathcal{S}_V^{M2E})/2, \quad (5)$$

$$\mathcal{S}_C = \mathcal{U}_C(\mathbf{M}, \mathbf{E}), \quad (6)$$

$$\mathcal{S} = \mathcal{U}(\mathbf{M}, \mathbf{E}) = (\mathcal{S}_V + \mathcal{S}_T + \mathcal{S}_C)/3, \quad (7)$$

Method

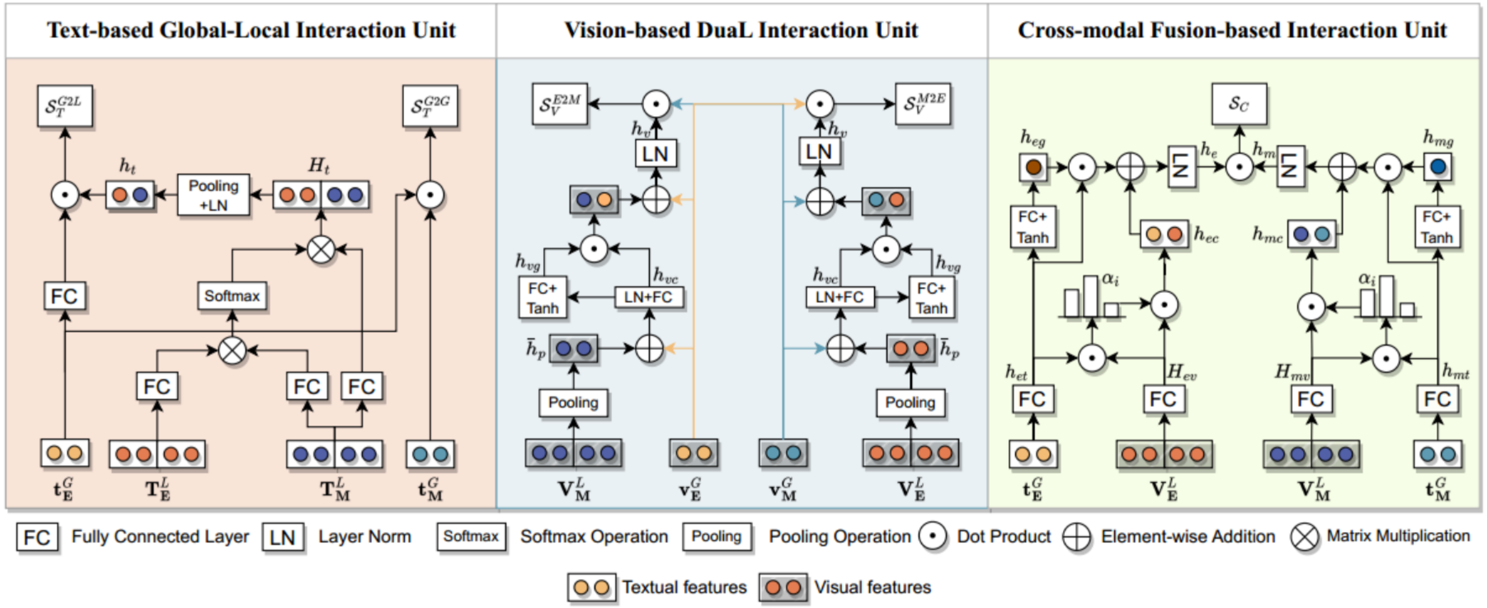


Figure 3: The designed multi-grained multimodal interaction layer, which contains three interaction units.

$$\mathcal{S}_T^{G2G} = \mathbf{t}_\mathrm{E}^G \cdot \mathbf{t}_\mathrm{M}^G. \tag{8}$$

$$\begin{aligned} Q, K, V &= \mathbf{T}_\mathrm{E}^L \boldsymbol{W}_{tq}, \mathbf{T}_\mathrm{M}^L \boldsymbol{W}_{tk}, \mathbf{T}_\mathrm{M}^L \boldsymbol{W}_{tv}, \\ H_t &= \mathrm{softmax}(\frac{QK^T}{\sqrt{d_T}})V, \end{aligned} \tag{9}$$

$$\begin{aligned} h_t &= \mathrm{LayerNorm}\left(\mathrm{MeanPooling}\left(H_t\right)\right), \\ \mathcal{S}_T^{G2L} &= \mathrm{FC}\left(\mathbf{t}_\mathrm{E}^G\right) \cdot h_t, \end{aligned} \tag{10}$$

$$\begin{aligned} \mathcal{S}_V^{E2M} &= \mathrm{DUAL}_{E2M}\left(\mathbf{v}_\mathrm{E}^G, \mathbf{v}_\mathrm{M}^G, \mathbf{V}_\mathrm{M}^L\right), \\ \mathcal{S}_V^{M2E} &= \mathrm{DUAL}_{M2E}\left(\mathbf{v}_\mathrm{M}^G, \mathbf{v}_\mathrm{E}^G, \mathbf{V}_\mathrm{E}^L\right), \end{aligned} \tag{11}$$

$$\begin{aligned} \bar{h}_p &= \mathrm{MeanPooling}\left(\mathbf{V}_\mathrm{B}^L\right), \\ h_{vc} &= \mathrm{FC}\left(\mathrm{LayerNorm}\left(\bar{h}_p + \mathbf{v}_\mathrm{A}^G\right)\right), \end{aligned} \tag{12}$$

Method

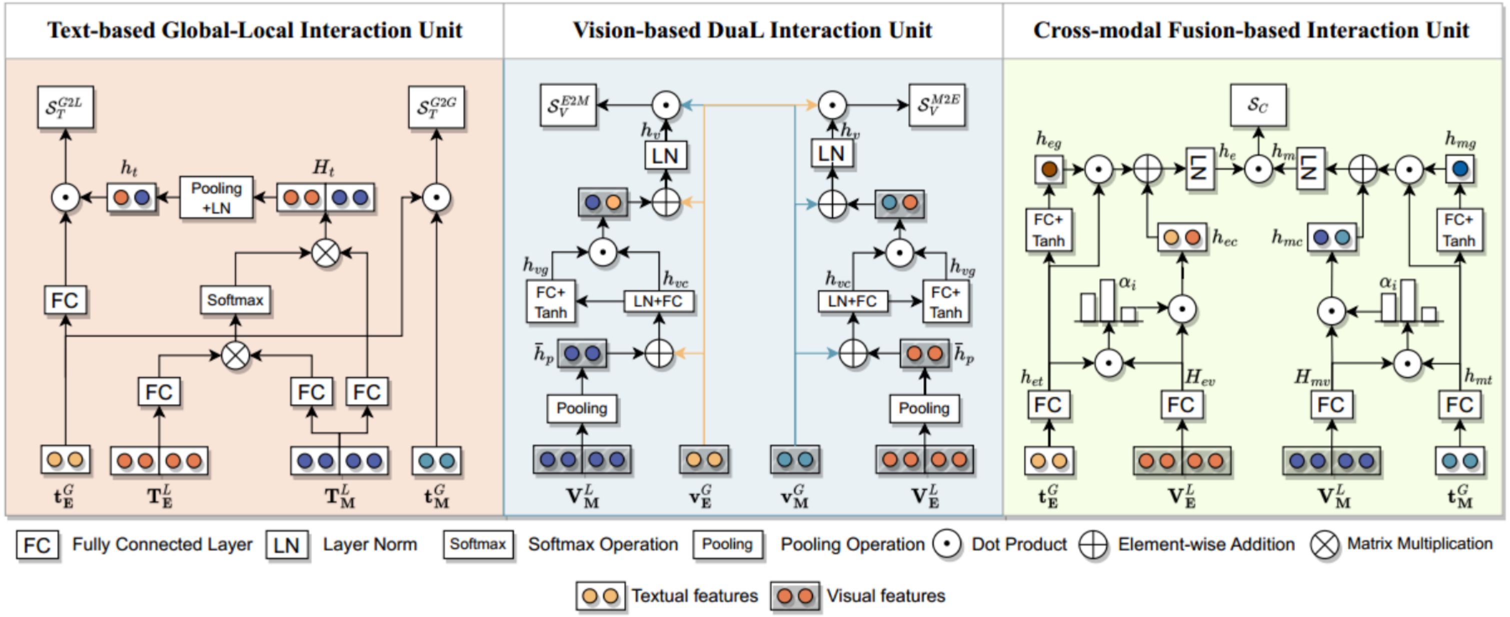


Figure 3: The designed multi-grained multimodal interaction layer, which contains three interaction units.

$$h_{vg} = \mathrm{Tanh}\left(\mathrm{FC}\left(h_{vc}\right)\right),$$
$$h_v = \mathrm{LayerNorm}\left(h_{vg} * h_{vc} + \mathbf{v}_{\mathrm{B}}^{G}\right), \tag{13}$$

$$\mathcal{S}_{\mathrm{V}}^{A2B} = h_v \cdot \mathbf{v}_{\mathrm{A}}^{G}. \tag{14}$$

$$het, hmt = \mathrm{FC}_{c1}\left(\mathbf{t}_{\mathrm{E}}^{G}\right), \mathrm{FC}_{c1}\left(\mathbf{t}_{\mathrm{M}}^{G}\right),$$
$$H_{ev}, H_{mv} = \mathrm{FC}_{c2}\left(\mathbf{V}_{\mathrm{E}}^{L}\right), \mathrm{FC}_{c2}\left(\mathbf{V}_{\mathrm{M}}^{L}\right), \tag{15}$$

$$\alpha_i = \frac{\exp\left(het \cdot H_{ev}^{i}\right)}{\sum_i^{n+1} \exp\left(het \cdot H_{ev}^{i}\right)},$$
$$hec = \sum_i^{n+1} \alpha_i * H_{ev}^{i}, i \in [1, 2, \ldots, (n+1)]. \tag{16}$$

Method

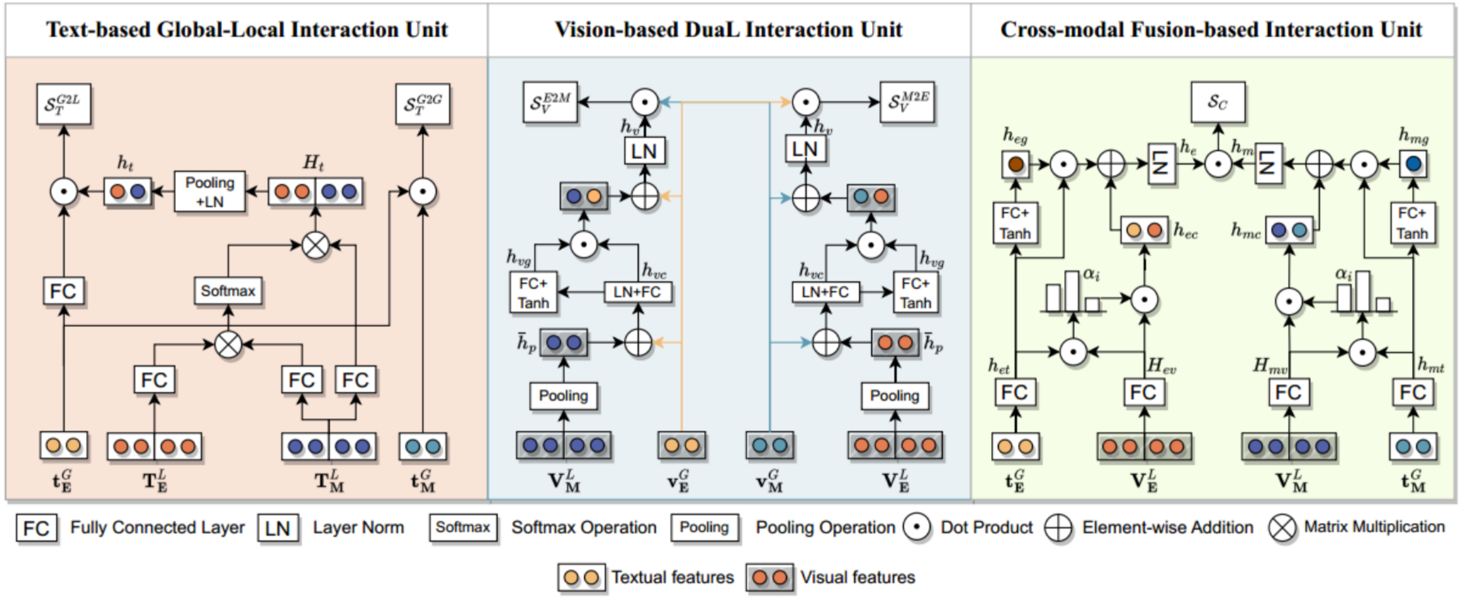


Figure 3: The designed multi-grained multimodal interaction layer, which contains three interaction units.

$$h_{eg} = \text{Tanh}\left(\text{FC}_{c3}\left(h_{et}\right)\right), \qquad (17)$$

$$h_e = \text{LayerNorm}\left(h_{eg} * h_{et} + h_{ec}\right). \qquad (18)$$

$$\mathcal{S}_C = h_e \cdot h_m. \qquad (19)$$

$$\mathcal{L}_O = -\log \frac{\exp\left(\mathcal{U}(\mathbf{M},\mathbf{E})\right)}{\sum_i \exp\left(\mathcal{U}(\mathbf{M},\mathbf{E}_i')\right)}, \qquad (20)$$

$$\mathcal{L}_X = -\log \frac{\exp\left(\mathcal{U}_X(\mathbf{M},\mathbf{E})\right)}{\sum_i \exp\left(\mathcal{U}_X(\mathbf{M},\mathbf{E}_i')\right)}, X \in \{T,V,C\}, \qquad (21)$$

$$\mathcal{L} = \mathcal{L}_O + \underbrace{\mathcal{L}_T + \mathcal{L}_V + \mathcal{L}_C}_{\text{unit-consistent loss function}}. \qquad (22)$$

Experiment

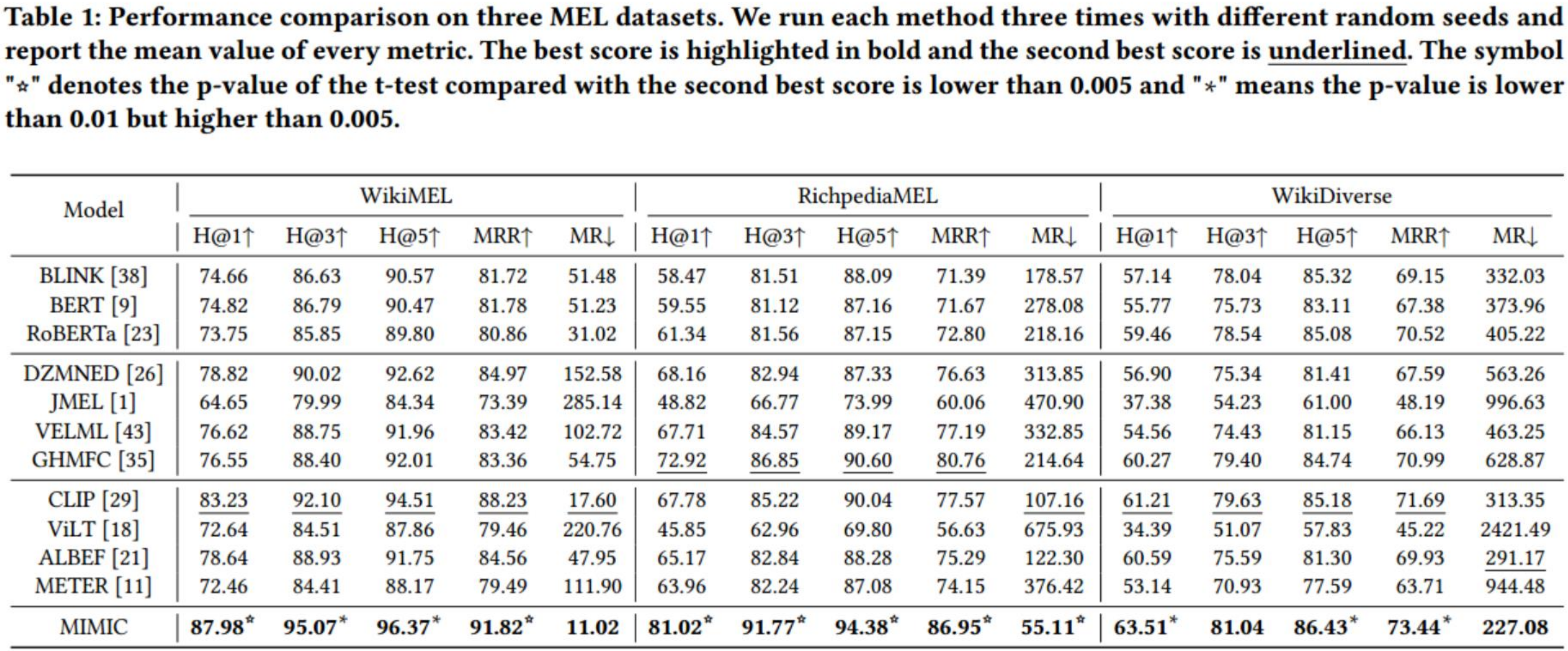


Table 1: Performance comparison on three MEL datasets. We run each method three times with different random seeds and report the mean value of every metric. The best score is highlighted in bold and the second best score is underlined. The symbol "⋆" denotes the p-value of the t-test compared with the second best score is lower than 0.005 and "∗" means the p-value is lower than 0.01 but higher than 0.005.

Model	WikiMEL					RichpediaMEL					WikiDiverse				
	H@1↑	H@3↑	H@5↑	MRR↑	MR↓	H@1↑	H@3↑	H@5↑	MRR↑	MR↓	H@1↑	H@3↑	H@5↑	MRR↑	MR↓
BLINK [38]	74.66	86.63	90.57	81.72	51.48	58.47	81.51	88.09	71.39	178.57	57.14	78.04	85.32	69.15	332.03
BERT [9]	74.82	86.79	90.47	81.78	51.23	59.55	81.12	87.16	71.67	278.08	55.77	75.73	83.11	67.38	373.96
RoBERTa [23]	73.75	85.85	89.80	80.86	31.02	61.34	81.56	87.15	72.80	218.16	59.46	78.54	85.08	70.52	405.22
DZMNED [26]	78.82	90.02	92.62	84.97	152.58	68.16	82.94	87.33	76.63	313.85	56.90	75.34	81.41	67.59	563.26
JMEL [1]	64.65	79.99	84.34	73.39	285.14	48.82	66.77	73.99	60.06	470.90	37.38	54.23	61.00	48.19	996.63
VELML [43]	76.62	88.75	91.96	83.42	102.72	67.71	84.57	89.17	77.19	332.85	54.56	74.43	81.15	66.13	463.25
GHMFC [35]	76.55	88.40	92.01	83.36	54.75	72.92	86.85	90.60	80.76	214.64	60.27	79.40	84.74	70.99	628.87
CLIP [29]	83.23	92.10	94.51	88.23	17.60	67.78	85.22	90.04	77.57	107.16	61.21	79.63	85.18	71.69	313.35
ViLT [18]	72.64	84.51	87.86	79.46	220.76	45.85	62.96	69.80	56.63	675.93	34.39	51.07	57.83	45.22	2421.49
ALBEF [21]	78.64	88.93	91.75	84.56	47.95	65.17	82.84	88.28	75.29	122.30	60.59	75.59	81.30	69.93	291.17
METER [11]	72.46	84.41	88.17	79.49	111.90	63.96	82.24	87.08	74.15	376.42	53.14	70.93	77.59	63.71	944.48
MIMIC	**87.98***	**95.07***	**96.37***	**91.82***	**11.02**	**81.02***	**91.77***	**94.38***	**86.95***	**55.11***	**63.51***	**81.04**	**86.43***	**73.44***	**227.08**

Experiment

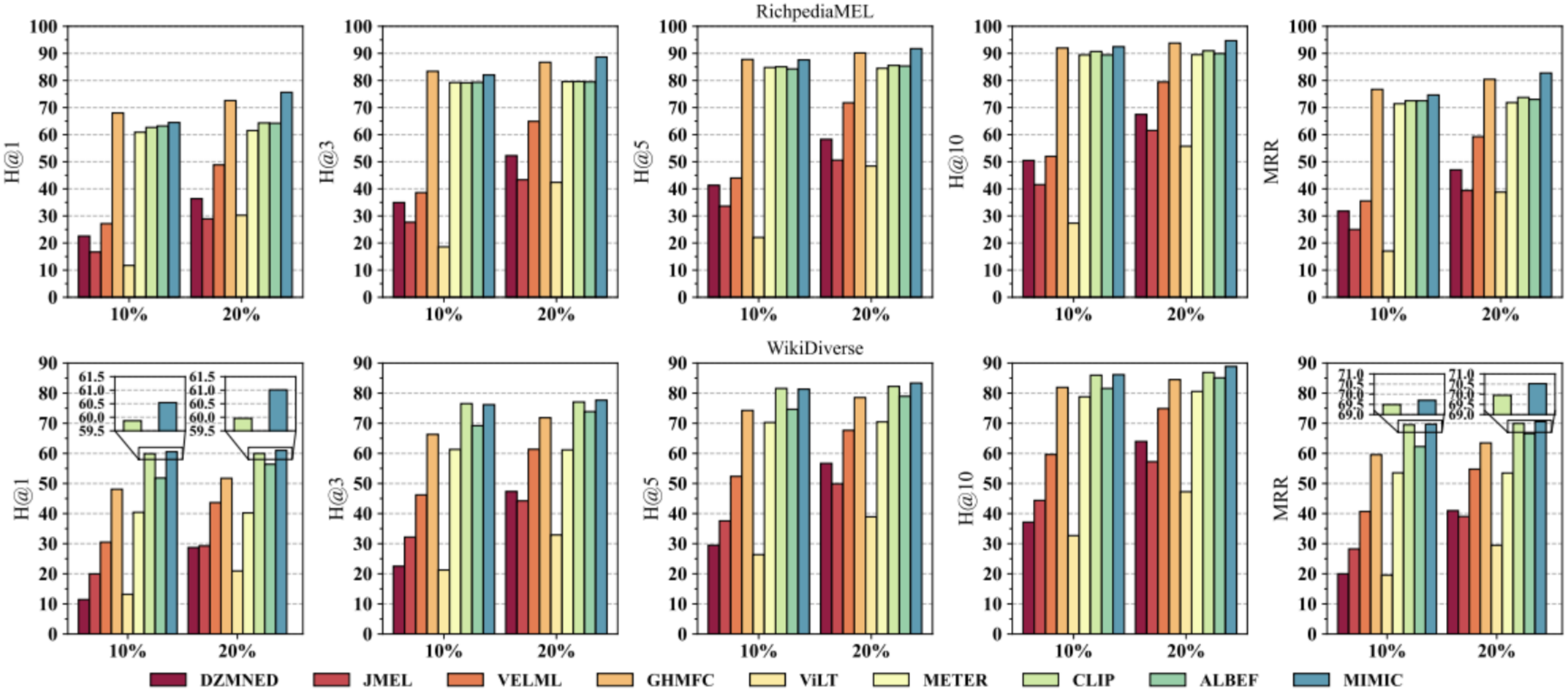
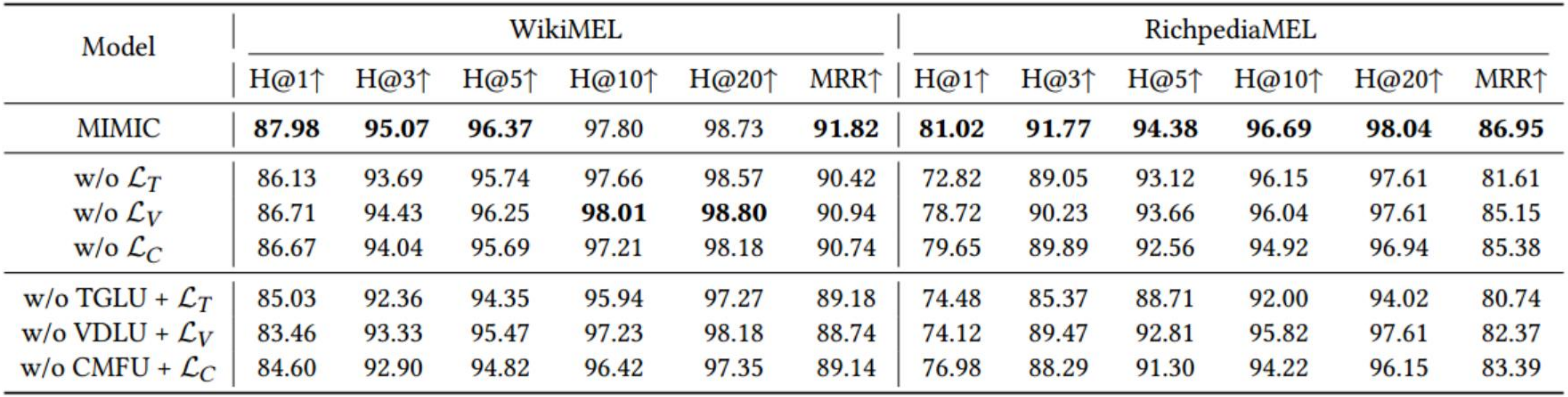


Figure 4: Performance comparison of low resource settings on RichpediaMEL and WikiDiverse. Details are zoomed in for better visualization.

Experiment

Table 2: Experimental results of ablation studies. The best scores are highlighted in bold.



Model	WikiMEL						RichpediaMEL					
	H@1↑	H@3↑	H@5↑	H@10↑	H@20↑	MRR↑	H@1↑	H@3↑	H@5↑	H@10↑	H@20↑	MRR↑
MIMIC	**87.98**	**95.07**	**96.37**	97.80	98.73	**91.82**	**81.02**	**91.77**	**94.38**	**96.69**	**98.04**	**86.95**
w/o $\mathcal{L}_T$	86.13	93.69	95.74	97.66	98.57	90.42	72.82	89.05	93.12	96.15	97.61	81.61
w/o $\mathcal{L}_V$	86.71	94.43	96.25	**98.01**	**98.80**	90.94	78.72	90.23	93.66	96.04	97.61	85.15
w/o $\mathcal{L}_C$	86.67	94.04	95.69	97.21	98.18	90.74	79.65	89.89	92.56	94.92	96.94	85.38
w/o TGLU + $\mathcal{L}_T$	85.03	92.36	94.35	95.94	97.27	89.18	74.48	85.37	88.71	92.00	94.02	80.74
w/o VDLU + $\mathcal{L}_V$	83.46	93.33	95.47	97.23	98.18	88.74	74.12	89.47	92.81	95.82	97.61	82.37
w/o CMFU + $\mathcal{L}_C$	84.60	92.90	94.82	96.42	97.35	89.14	76.98	88.29	91.30	94.22	96.15	83.39

Experiment

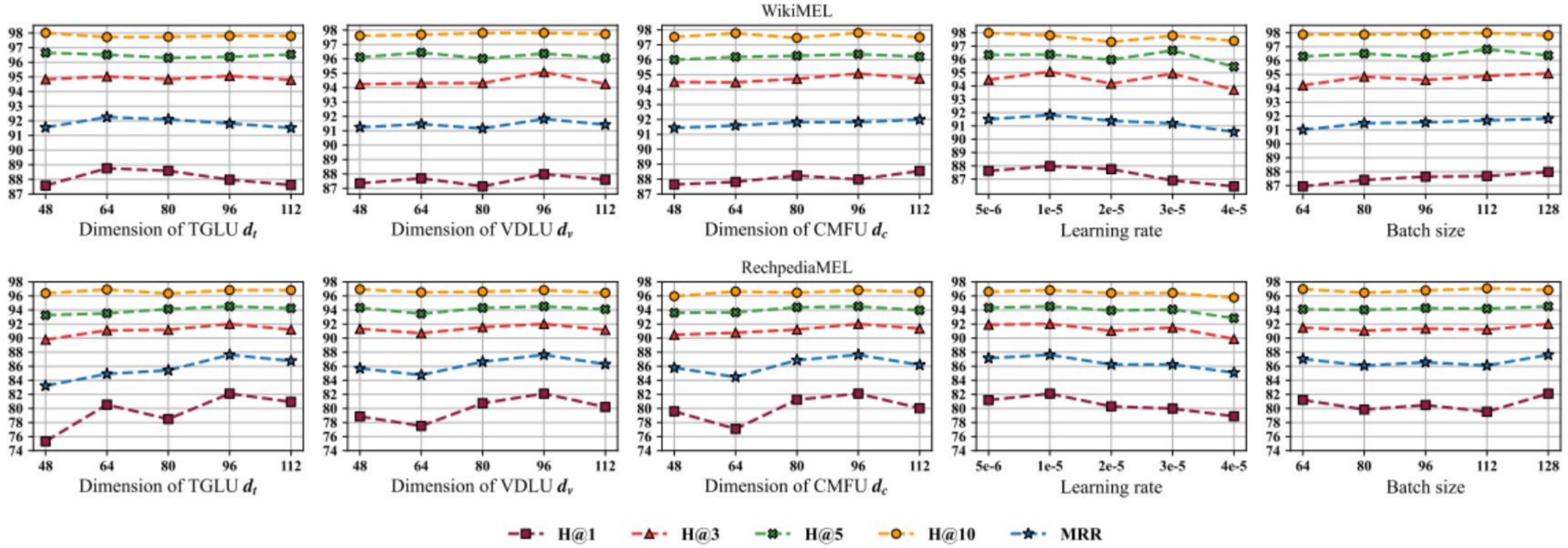


Figure 5: Parameter sensitivity analysis on WikiMEL and RichpediaMEL regarding different values.

Thank you!

CHONGQING UNIVERSITY OF TECHNOLOGY 1940

CHINESE ACADEMY OF SCIENCES

gesis
Leibniz-Institut
für Sozialwissenschaften

Forschungszentrum · Research Center
L3S