Multi-level Mixture of Experts for Multimodal Entity Linking

Zhiwei Hu
School of Computer and Information Technology
Shanxi University
Taiyuan, China
zhiweihu@whu.edu.cn

Víctor Gutiérrez-Basulto
School of Computer Science and Informatics
Cardiff University
Cardiff, UK
gutierrezbasultov@cardiff.ac.uk

Zhiliang Xiang
School of Computer Science and Informatics
Cardiff University
Cardiff, UK
xiangz6@cardiff.ac.uk

Ru Li*
School of Computer and Information Technology
Shanxi University
Taiyuan, China
liru@sxu.edu.cn

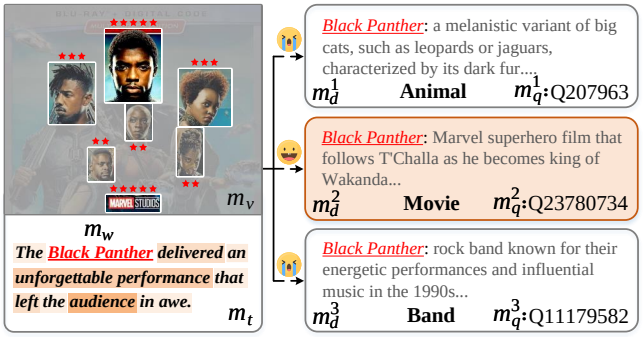
Jeff Z. Pan*
ILCC, School of Informatics
University of Edinburgh
Edinburgh, UK
http://knowledge-representation.org/j.z.pan/

Abstract

Multimodal Entity Linking (MEL) aims to link ambiguous mentions within multimodal contexts to associated entities in a multimodal knowledge base. Existing approaches to MEL introduce multimodal interaction and fusion mechanisms to bridge the modality gap and enable multi-grained semantic matching. However, they do not address two important problems: (*i*) *mention ambiguity*, *i.e.,* the lack of semantic content caused by the brevity and omission of key information in the mention's textual context; (*ii*) *dynamic selection of modal content*, *i.e.,* to dynamically distinguish the importance of different parts of modal information. To mitigate these issues, we propose a **M**ulti-level **M**ixture **o**f **E**xperts (**MMoE**) model for MEL. MMoE has four components: (*i*) the *description-aware mention enhancement* module leverages large language models to identify the WikiData descriptions that best match a mention, considering the mention's textual context; (*ii*) the *multimodal feature extraction* module adopts multimodal feature encoders to obtain textual and visual embeddings for both mentions and entities; (*iii*)-(*iv*) the *intra-level mixture of experts* and *inter-level mixture of experts* modules apply a switch mixture of experts mechanism to dynamically and adaptively select features from relevant regions of information. Extensive experiments on WikiMEL, RichpediaMEL and WikiDiverse datasets demonstrate the outstanding performance of MMoE compared to the state-of-the-art. MMoE's code is available at: https://github.com/zhiweihu1103/MEL-MMoE.

CCS Concepts

• **Information systems** → **Multimedia databases**; **Data mining**; • **Computing methodologies** → **Knowledge representation and reasoning**.



(a) Multimodal Context (b) Entity Description

Figure 1: (a) The given mention `Black Panther` within a multimodal context. For mention textual context m_t, different color depths indicate the importance for the prediction of the mention m_w. (b) Wikidata descriptions of entities called `Black Panther`. The number of ★ in m_v indicates the importance level of the corresponding region.

Keywords

Multimodal Entity Linking, Entity Linking, Multimodal Matching

ACM Reference Format:
Zhiwei Hu, Víctor Gutiérrez-Basulto, Zhiliang Xiang, Ru Li, and Jeff Z. Pan. 2025. Multi-level Mixture of Experts for Multimodal Entity Linking. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3711896.3737060

1 Introduction

Entity linking (EL) [5] is the task of recognising mentions of entities in unstructured textual content and linking them to the corresponding entities in a knowledge graph (KG) [35, 61]. EL supports various downstream information retrieval tasks, such as question answering [15, 19, 28, 55], semantic search [6, 16, 19, 34, 45], dialog systems [2, 23, 50], and so on [10, 18, 36]. With the increasing availability of multimodal content, *multimodal entity linking* (MEL), which aims to link ambiguous mentions in multimodal contexts

*Contact Authors.

KDD '25, Toronto, ON, Canada

ACM ISBN 979-8-4007-1454-2/2025/08
https://doi.org/10.1145/3711896.3737060

to entities in a multimodal KG (MMKG), has garnered increasing attention. Figure 1 shows that to successfully link the mention `Black Panther` to its corresponding entities $\{m_q^1, m_q^2, m_q^3\}$, a comprehensive understanding of the associated textual m_t and visual m_v information is required. Several approaches have been introduced to handle the MEL task, including (*i*): *vision-and-language pre-trained* [14, 20, 22, 39], (*ii*): *generative-based* [27, 33, 41], and (*iii*): *multimodal interaction-based* [1, 17, 30–32, 44, 52, 59, 60]. These frameworks have made steady progress, obtaining performance improvements in various benchmarks like WikiMEL [52], RichpediaMEL [52], and WikiDiverse [53]. However, each of these approaches have their own disadvantages: vision-and-language pretrained based methods often neglect the fine-grained interactions within modality-specific information, resulting in an inability to fully uncover the latent relationships among the internal elements of a modality. Generative-based methods face challenges in associating visual inputs with their internal entity knowledge, making it difficult to establish connections between fine-grained visual patterns and the corresponding entities. Although the prevailing approaches are based on multimodal interaction, they still suffer from two crucial issues:

▹ **Mention Ambiguity.** The context in which textual mentions occur is frequently concise and may incorporate abbreviations, idiomatic expressions, or domain-specific jargon. This scarcity and ambiguity of contextual information can render the accurate identification of the referenced entity infeasible, directly impacting the performance of a model. Figure 1 shows an intuitive example in which for the mention `Black Panther`, it is difficult to determine whether `Black Panther` is an *animal*, a *movie* or a *band* simply based on its given textual context m_t. Indeed, m_t might relate to the film's excellence or the actors' outstanding performance, the animal's striking movements, or the band's impressive live performance.

▹ **Dynamic Selection of Modal Content.** Existing methods typically encode the full textual sequence or visual image into a single fixed-length vector for modal interaction, ignoring that different sections within a modality contribute differently to the mention prediction of the corresponding entity. Figure 1 presents a typical example illustrating the variation in the significance of different regions within the image and sections of the textual sequence. For instance, for the textual context m_t, compared to the definite article `the`, the phrase `unforgettable performance` exerts a greater impact on the interpretation of the mention `Black Panther`. Similarly, for the visual modality knowledge m_v, the regions corresponding to actor `Chadwick Aaron Boseman` and `Marvel Studios` should receive more attention compared to the region of actor `Danai Jekesai Gurira`. Thus, it is paramount to dynamically distinguish the importance of different parts of modal information.

To address these two problems, we propose a **M**ulti-level **M**ixture **o**f **E**xperts (**MMoE**) model for the MEL task, composed of the following four components. The *description-aware mention enhancement* module leverages large language models to identify the WikiData descriptions that best match a mention, to compensate for the limitation of concise mentions. Then, the *multimodal feature extraction* module utilizes a pre-trained CLIP model to represent textual tokens and visual patches for both mentions and entities. Afterwards, we design a *switch mixture of experts* (SMoE) mechanism to dynamically and adaptively select features from relevant regions of information. Specifically, we introduce SMoE for both intra and inter modality interaction.

In summary, our main contributions are three-fold:

- We propose the MMoE framework for MEL which selects the optimal expert network from the intra- and inter-modality perspectives, and dynamically and adaptively selects features from relevant regions of information.
- We investigate the issue of mention ambiguity and use the relevant WikiData description of an entity (identified using a large language model) to enhance the semantic richness of the mention context.
- We conduct empirical and ablation experiments on three widely-used benchmarks, *i.e.,* WikiMEL, RichpediaMEL and WikiDiverse, showing the effectiveness and generality of MMoE over existing state-of-the-art models.

2 Related Work

▹ **Text-based Entity Linking.** Text-based Entity Linking (EL) focuses on identifying mentions in text and linking them to corresponding entities in knowledge graphs. Approaches to EL can be categorized in *retrieval-based methods* and *generative-based methods*. (*i*) *retrieval-based methods* leverage retrieval mechanisms to filter the most relevant entities [9, 21, 37, 40, 54]. (*ii*) *generative-based methods* utilize the intrinsic knowledge within language models to generate entities associated with mentions [7, 8, 12, 26]. While text-based entity linking methods have achieved significant advancements, they cannot be directly applied to the multimodal setting.

▹ **Multimodal Entity Linking.** Multimodal entity linking (MEL) is an extension of the text-based entity linking task, which utilizes both textual and visual modal knowledge to deal with the ambiguity of entities. Existing MEL methods can be classified into three categories: *vision-and-language pre-trained methods*, *generative-based methods*, and *multimodal interaction based methods*. (*i*) *vision-and-language pre-trained methods* use large-scale multimodal data to learn semantic alignment and complementary visual and language knowledge, thereby enhancing its understanding and generation capabilities for the MEL task [14, 20, 22, 39]. (*ii*) *generative-based methods* leverage the strong generative capabilities of large language models and the constrained decoding strategy to ensure the validity of the generated entities [24, 27, 33, 41]. (*iii*) *multimodal interaction-based methods* employ various mechanisms, such as concatenation [1, 58], attention [32, 52, 56, 57, 60], matching networks [17, 30, 31, 43] and optimal transport [59], to fusion different modal knowledge from entities and mentions. However, all of these methods still suffer from mention ambiguity and dynamic selection content issues.

3 Method

In this section, we introduce the MEL problem definition (§3.1) and the four components of our MMoE (§3.2-§3.5).

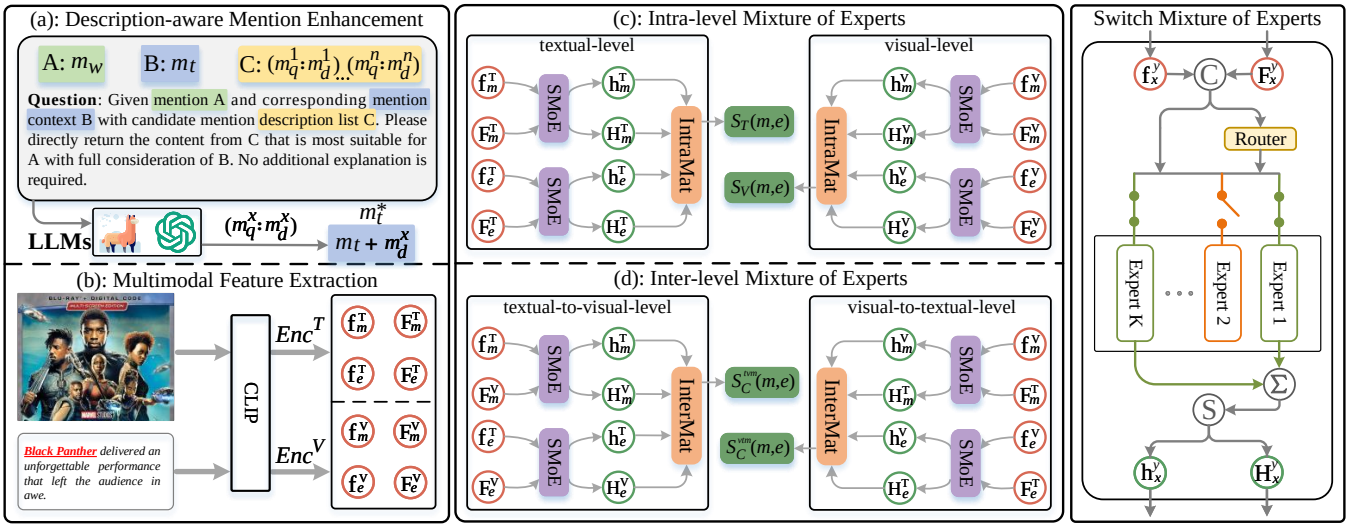


Figure 2: The structure of the MMoE model, containing four components: (*a*): Description-aware Mention Enhancement (DME), (*b*): Multimodal Feature Extraction (MFE), (*c*): Intra-level Mixture of Experts (IntraMoE) and (*d*): Inter-level Mixture of Experts (InterMoE). IntraMat and InterMat are short for Intra-level Feature Matching and Inter-level Feature Matching modules.

3.1 Problem Definition

Intuitively, given a mention m within a multimodal context, the *multimodal entity linking* (MEL) task aims to retrieve the ground truth entity e from the set of entities $\mathcal{E}$ of a knowledge base that is the most relevant to m. Each entity $e \triangleq (e_n, e_a, e_v)$ is composed of an entity name e_n, textual attribute e_a and a visual image e_v. Each mention is defined as a triple $m \triangleq (m_w, m_t, m_v)$, with m_w a mention word, m_t a textual context, and m_v a visual image. Formally, MEL can be formulated as maximizing the log-likelihood over the training set $\mathcal{D}$ while optimizing the model parameters θ as follows:

$$\theta^* = \max_{\theta} \sum_{(m,e)\in\mathcal{D}} \log p_\theta(e|m,\mathcal{E}) \tag{1}$$

where θ^* denotes the final parameters, and $p_\theta(e|m,\mathcal{E})$ is a score function used to calculate the similarity between mention m and the candidate entity $e \in \mathcal{E}$.

3.2 Description-aware Mention Enhancement

Textual mentions and associated contexts are frequently concise and may include abbreviations. For instance, in the WikiDiverse dataset mention contexts have an average length of 11 words. This brevity can give rise to the "*mention ambiguity*" phenomenon. More precisely, a limited number of words fails to provide sufficient semantic information, complicating the extraction of adequate knowledge from the context surrounding a mention. Figure 1 provides a clear illustration of this issue, where the mention word m_w is `Black Panther` and the mention textual context m_t is `The Black Panther delivered...`. Given the succinctness of m_t, it is challenging to ascertain whether `Black Panther` refers to an animal, a movie, or a band, because m_t might relate to the awe-inspiring movements or posture displayed by the animal, the brilliance of the film or the outstanding performance of actors, or the band's outstanding performance and musical impact during the show. So, such ambiguity negatively impacts the effectiveness of frameworks for the MEL task.

However, WikiData [49] provides a description for each entity, which might serve to mitigate the mention ambiguity issue. Nonetheless, a single entity name may correspond to several entities within WikiData, each associated with its own description. For instance, the mention `Black Panther` in Figure 1 can refer to the WikiData entities `Q207963`, `Q23780734`, and `Q11179582`. Selecting the most appropriate description in a given context is thus a critical challenge to overcome for the succesful use of WikiData information. Recently, large language models (LLMs), such as ChatGPT [33] and LLaMA [46], have demonstrated strong capabilities in knowledge storage and semantic evaluation. Motivated by this, we introduce the LLM-based **D**escription-aware **M**ention **E**nhancement (**DME**) module, designed to augment the knowledge associated with a mention. Specifically, starting with the mention word m_w and mention textual context m_t, to generate the description-aware enhanced mention content m_t^*, we proceed through three steps as depicted in Figure 2(a):

(1) Given the mention word m_w, we retrieve entities from WikiData that share the same name with m_w. The set of all these entities is denoted as $M_w = \{(m_q^1, m_d^1)...(m_q^i, m_d^i)...(m_q^n, m_d^n)\}$, where n is the number of retrieved entities, and m_d^i represents the description corresponding to the i-th entity with id m_q^i. For instance, as shown in Figure 1, m_q^1 is `Q207963` and the corresponding description m_d^1 is `Black Panther: a melanistic variant...`.

(2) Given M_w, we use the prompt in Figure 2(a), to employ LLMs to identify the entity-description pair that is the most pertinent for the mention m_w and its textual context m_t, we denote the most appropriate entity-description pair as (m_q^x, m_d^x). For example, as shown in Figure 1, the most relevant WikiData entity id and

description corresponding to mention `Black Panther` is m_q^2 and m_d^2.

(3) We concatenate the mention textual context m_t with the most relevant mention description m_d^x to obtain the description-aware enhanced mention textual context $m_t^* = m_t + m_d^x$.

3.3 Multimodal Feature Extraction

Given the entity e and mention m with their corresponding textual and visual knowledge, we generate their initial embeddings as done in [17, 31, 59], as shown in Figure 2(b). Specifically, we adopt the contrastive language-image pretraining architecture (CLIP) [39] as the multimodal encoder, which contains a pre-trained BERT [12] textual encoder Enc^T and a pre-trained ViT [13] visual encoder Enc^V:

$$\begin{cases} \mathbf{F}_m^T = Enc^T([\texttt{CLS}]\, m_w\, [\texttt{SEP}]\, m_t^*) \\ \mathbf{F}_e^T = Enc^T([\texttt{CLS}]\, e_n\, [\texttt{SEP}]\, m_a) \\ \mathbf{F}_m^V = Enc^V(m_v),\ \mathbf{F}_e^V = Enc^V(e_v) \end{cases} \tag{2}$$

where $\mathbf{F}_m^T \in \mathbb{R}^{L_1 \times d}, \mathbf{F}_e^T \in \mathbb{R}^{L_2 \times d}$ represent sequential textual features for the mention and entity sentence, L_1 and L_2 denote the length of the token sequence of the mention and entity, respectively, and d is the dimension of embeddings. $\mathbf{F}_m^V \in \mathbb{R}^{P_1 \times d}, \mathbf{F}_e^V \in \mathbb{R}^{P_2 \times d}$ denote visual features for the mention and entity patches, P_1 and P_2 represent the number of patches of the mention and entity, respectively. In addition to these fine-grained features, we also obtain coarse-grained features $\mathbf{f}_m^T, \mathbf{f}_e^T, \mathbf{f}_m^V, \mathbf{f}_e^V \in \mathbb{R}^d$ for a more comprehensive semantic understanding, which can be achieved by using the embedding of the special token `[CLS]` for the textual encoder Enc^T and the special token `[EOS]` for the visual encoder Enc^V.

3.4 Intra-level Mixture of Experts

Considering that not every part (region/section) of unimodal information contributes equally to the final prediction, different tokens and patches in the textual and visual information characterize the entity or mention content to different extents. For example, as shown in Figure 1, for the mention context m_t, it is obvious that `unforgettable performance` should be given higher attention than the definite article `the`. To help dealing with such redundant noise, we believe specialized screening and learning of unimodal information through sample adaptive mechanisms are paramount. To achieve this vision, we need to address two important questions: (1) *What guides the selection of specific regions within a modality?* (2) *How to dynamically learn the importance of different regions of information in various samples?*

▹ **SMoE: Switch Mixture of Experts.** To address these challenges, we introduced the **S**witch **M**ixture **o**f **E**xperts (**SMoE**) mechanism to dynamically learn the contribution of different regions of information within a modality. To solve the first problem, we use coarse-grained and fine-grained features to guide the supervision of specific regions. To solve the second problem, we introduce a mixture of experts mechanism to adaptively and dynamically select information from different regions. Specifically, SMoE is composed of the following three steps:

(1) *Multi-granular Fusion of Features.* Given the coarse-grained and fine-grained features $\mathbf{f}_x^y$, $\mathbf{F}_x^y$, we first perform a concatenation operation and then apply multi-layer perceptron (MLP) layers to obtain the fusion feature $\mathbf{P}_x^y = \texttt{MLP}([\mathbf{F}_x^y; \mathbf{f}_x^y])$. This is simple and intuitive, and can also achieve a bidirectional effect from coarse-grained to fine-grained and vice versa in subsequent operations.

(2) *Dynamic Selection of Features.* The main components of a SMoE layer are a network $\mathbf{R}$ as a sparse router and an expert network $\mathbf{E}$. Given a token representation p_x^y from $\mathbf{P}_x^y$, we process it sparsely using k out of the K available experts to obtain the final representation q_x^y, which is formally defined as:

$$\begin{cases} \mathbf{R}(p_x^y) = \texttt{Softmax}(\mathbf{W}_\mathbf{R} \otimes p_x^y) \\ q_x^y = \sum_{i=1}^{k} \mathbf{R}_i \circledast \mathbf{E}_i,\ \mathbf{E}_i = \texttt{FFN}_i(p_x^y) \end{cases} \tag{3}$$

where $\otimes$ and $\circledast$ denote matrix and element-wise multiplication respectively. $\mathbf{W}_\mathbf{R}$ is a learnable parameter, $\texttt{FFN}_i$ represents a feed-forward layer and its parameters are independent between different experts, $\mathbf{R}_i$ denotes the activation value of the i-th expert, considering that the acquisition of the $\mathbf{R}_i$ value is adaptive, so its behavior is intrinsically dynamic. The final output q_x^y of k activated experts are the linearly weighted combination of each expert's output $\mathbf{E}_i$ on the token output by the router $\mathbf{R}_i$. By performing similar operations on each token we can obtain the representation $\mathbf{Q}_x^y$ after multiple layers of SMoE.

(3) *Multi-granular Split of Features.* Considering that the dynamic selection process in Step (2) is completed by coarse-grained and fine-grained features $\mathbf{f}_x^y$, $\mathbf{F}_x^y$, the interaction between $\mathbf{f}_x^y$ and $\mathbf{F}_x^y$ is realized to a certain extent, and the introduction of the multi-layer MoE further diversifies such interaction. We further use a split operation based on the dimension to restore $\mathbf{Q}_x^y$ to $\mathbf{h}_x^y$ and $\mathbf{H}_x^y$ to facilitate the subsequent calculation of feature matching scores. This is formally expressed as $[\mathbf{h}_x^y; \mathbf{H}_x^y] = \texttt{Split}(\mathbf{Q}_x^y)$.

▹ **IntraMat: Intra-level Feature Matching.** To fully consider the unimodal correlations between mentions and entities, we design the **Intra**-level Feature **Mat**ching (**IntraMat**) module, which is composed of the *Coarse-grained Matching (CM)* and *Fine-grained Matching (FM)* sub-modules. As shown in Figure 2(c), given the mention m with coarse-grained and fine-grained textual embeddings $\mathbf{f}_m^T$, $\mathbf{F}_m^T$, and entity e with coarse-grained and fine-grained textual embeddings $\mathbf{f}_e^T$, $\mathbf{F}_e^T$, we first leverage the SMoE mechanism to obtain the enhanced embeddings $\mathbf{h}_m^T$, $\mathbf{H}_m^T$ and $\mathbf{h}_e^T$, $\mathbf{H}_e^T$. Then we integrate two types of matching scores as follows:

- *Coarse-grained Matching (CM).* We directly perform the dot product $\odot$ between the mention and entity coarse-grained textual embeddings $\mathbf{h}_m^T$ and $\mathbf{h}_e^T$ formulated as:

$$\mathcal{S}_T^{\texttt{cm}}(m, e) = \mathbf{h}_e^T \odot \mathbf{h}_m^T \tag{4}$$

- *Fine-grained Matching (FM).* We establish the fine-grained textual unimodal correlation assignments based on an attention

mechanism [48] as follows:

$$\begin{cases} \mathbf{M} = \mathbf{H}_e^T \otimes \mathbf{W}_q, \mathbf{K} = \mathbf{H}_m^T \otimes \mathbf{W}_k, \mathbf{V} = \mathbf{H}_m^T \otimes \mathbf{W}_v \\ \mathbf{A}_{m\to e}^T = \texttt{Softmax}(\mathbf{X}),\ \mathbf{X} = \mathbf{M} \otimes \mathbf{K}/\sqrt{d} \\ \mathbf{G}_{m\to e}^T = \texttt{mean}(\mathbf{A}_{m\to e}^T \otimes \mathbf{V}) \\ \mathcal{S}_T^{\texttt{fm}}(m,e) = \mathbf{h}_e^T \odot \mathbf{G}_{m\to e}^T \\ \mathcal{S}_T(m,e) = (\mathcal{S}_T^{\texttt{cm}}(m,e) + \mathcal{S}_T^{\texttt{fm}}(m,e))/2 \end{cases} \tag{5}$$

where $\mathbf{W}_q$, $\mathbf{W}_k$, $\mathbf{W}_v$ are learnable parameters, $\mathbf{A}_{m\to e}^T$ is the assigned correlation weight from mention textual tokens embeddings $\mathbf{H}_m^T$ to entity textual tokens embeddings $\mathbf{H}_e^T$. $\mathbf{G}_{m\to e}^T$ represents the aggregated representation of $\mathbf{H}_m^T$ and $\mathbf{H}_e^T$ after attention interaction. Afterwards, the intra-level textual matching score $\mathcal{S}_T(m,e)$ is defined as the average of the coarse-grained score $\mathcal{S}_T^{\texttt{cm}}(m,e)$ and fine-grained score $\mathcal{S}_T^{\texttt{fm}}(m,e)$.
Similarly, the intra-level visual matching score $\mathcal{S}_V(m,e)$ is defined as the average of $\mathcal{S}_V^{\texttt{cm}}(m,e)$ and $\mathcal{S}_V^{\texttt{fm}}(m,e)$.

3.5 Inter-level Mixture of Experts

Relying exclusively on unimodal content often leads to ambiguous interpretations. Consider the example presented in Figure 1, when constrained only to the textual modality m_t, it remains uncertain whether the mention `Black Panther` refers to an *animal* or a *movie*. However, incorporating the visual modality m_v helps to alliviate this ambiguity, substantially increasing the likelihood that `Black Panther` is identified as a *movie*. This enhancement stems from the fact that visual information provides spatial structural knowledge absent in the textual representation, while textual information offers semantic details not captured visually. A synergistic integration of both modalities compensates for the deficiencies inherent to each of them, leading to a more robust understanding and interpretation of the content.

▹ **InterMat: Inter-level Feature Matching.** To alleviate this content gap between mentions and entities, we introduce an adaptive multimodal feature interaction mechanism: the **Inter**-level Feature **Mat**ching (**InterMat**) module contains the *textual visual matching (TVM)* and *visual textual matching (VTM)* sub-modules. Considering that they have similar operations except for the input, we focus on the implementation process of TVM. Specifically, as shown in Figure 2(d), given the input embeddings $\{\mathbf{f}_m^T, \mathbf{F}_m^V, \mathbf{f}_e^T, \mathbf{F}_e^V\}$ associated with mention m and entity e, similar to IntraMat, we leverage SMoE to obtain enhanced embeddings $\{\mathbf{h}_m^T, \mathbf{H}_m^V, \mathbf{h}_e^T, \mathbf{H}_e^V\}$. Then, we apply a gated function based on the extracted embeddings $(\mathbf{h}_m^T, \mathbf{H}_m^V)$ as follows:

$$\begin{cases} \mathbf{h}_m^{T\star} = \texttt{Tanh}(\mathbf{h}_m^T) \\ \mathbf{H}_m^{V\star} = \texttt{Softmax}(\mathbf{h}_m^T \otimes \mathbf{H}_m^V) \otimes \mathbf{H}_m^V \\ \mathbf{E}_m^{\texttt{tvm}} = \texttt{LayerNorm}(\mathbf{h}_m^{T\star} \circledast \mathbf{h}_m^T + \mathbf{H}_m^{V\star}) \end{cases} \tag{6}$$

By replacing inputs $\mathbf{h}_m^T$ and $\mathbf{H}_m^V$ with $\mathbf{h}_e^T$ and $\mathbf{H}_e^V$ in the operations in Equation 6, we can also obtain the entity-side context embedding $\mathbf{E}_e^{\texttt{tvm}}$, then the score is calculated by the dot product:

$$\mathcal{S}_C^{\texttt{tvm}}(m,e) = \mathbf{E}_e^{\texttt{tvm}} \odot \mathbf{E}_m^{\texttt{tvm}} \tag{7}$$

Using a similar approach, we can obtain the visual textual matching score $\mathcal{S}_C^{\texttt{vtm}}(m,e)$, and acerage $\mathcal{S}_C^{\texttt{tvm}}(m,e)$ and $\mathcal{S}_C^{\texttt{vtm}}(m,e)$ to get the final matching score $\mathcal{S}_C(m,e)$ of the InterMoE module.

3.6 Overall Score and Loss Function

The symmetrical design of the IntraMoE and InterMoE modules introduces comparable features as well as similarity measures from different perspectives. We further collect all matching scores and obtain the overall matching score $\mathcal{S}_O(m,e)$. Inspired by [31, 59], we design a contrastive training objective on the different matching scores as follows:

$$\begin{cases} \mathcal{S}_O(m,e) = \sum\limits_{X\in\{T,V,C\}} \mathcal{S}_X(m,e) \\ \mathcal{L}_X = -\log \dfrac{\exp(\mathcal{S}_X(m,e))}{\sum\limits_{e'\in\mathcal{E}} \exp(\mathcal{S}_X(m,e'))} \\ \mathcal{L} = \sum\limits_{X\in\{O,T,V,C\}} \mathcal{L}_X \end{cases} \tag{8}$$

where $\mathcal{E}$ denotes all candidate entities, e' serve as the in-batch negative samples for entity e. $\mathcal{L}$ is the final loss of our model, which considers the overall loss $\mathcal{L}_O$, unimodal losses $\mathcal{L}_T$, $\mathcal{L}_V$, and multimodal loss $\mathcal{L}_C$.

4 Experiments

To evaluate the effectiveness of MMoE, we aim to explore the following four research questions:

- **RQ1 (Effectiveness):** How MMoE performs compared to the state-of-the-art under different conditions?
- **RQ2 (Ablation studies):** How different components contribute to MMoE's performance?
- **RQ3 (Parameter sensitivity):** How hyper-parameters influence MMoE's performance?
- **RQ4 (Complexity analysis and Case Study):** What is the amount of computation and parameters used by MMoE? We provide additional results and a case study in **Appendix B** and **Appendix C**.

4.1 Experimental Setup

▹ **Datasets.** We conduct experiments on three well-known datasets, **WikiMEL** [52], **RichpediaMEL** [52], and **WikiDiverse** [53]. The **WikiMEL** dataset [52] contains more than 22K samples, where entities are collected from WikiData [49] and the corresponding textual and visual knowledge from WikiPedia; **RichpediaMEL** [52] contains more than 17K samples, where entities are gathered from Richpedia [51] and the corresponding multimodal content is obtained from WikiPedia; **WikiDiverse** [53] contains over 15K samples, where multimodal information is collected from WikiNews and covers various topics. We follow the training, validation, and test set splitting by [17, 24, 31, 44, 59]. The statistics of three datasets are summarized in Table 1.

▹ **Evaluation Metrics.** Given the similarity scores between a mention and candidate entities, we sort these scores in descending order to calculate **MRR** and **Hits@*n*** (abbreviated sometimes as **H@*n***,

Statistics	WikiMEL	RichpediaMEL	WikiDiverse
# Num. of sentences	22,070	17,724	7,405
# Num. of entities	109,976	160,935	132,460
# entities with Img.	67,195	86,769	67,309
# Num. of mentions	25,846	17,805	15,093
# Img. of mentions	22,136	15,853	6,697
# mentions in train	18,092	12,463	11,351
# mentions in valid	2,585	1,780	1,664
# mentions in test	5,169	3,562	2,078

Table 1: Statistics of three different datasets. "Num." and "Img." denote number and image, respectively.

$n \in \{1, 3, 5\}$). MRR stands for the inverse of the rank for the first correct entity, while Hits@n represents the rank for the first correct entity ranked among the top-n answers. For all metrics, the larger the values, the better the performance.

▹ **Implementation Details.** We conduct experiments on four 24G GeForce RTX 4090 GPUs, adopt AdamW [29] as optimiser, and use the grid search method to select the optimal hyperparameter setting for the MMoE model on different datasets. Consistent with [31, 59], we apply the pre-trained CLIP-ViT-Base-Pathch32 model as the embedding encoder for textual and visual information. For the textual modality, we set up parameter experiments with different max text length. For the visual modality, we first scale the images into 224×224 resolution to conform with the input of the ViT model, then we set the number of patches to 32. We adopt "`gpt-3.5-turbo-0613`" in the DME module as the LLM to order the description list. We focus on the hyperparameters that have a greater impact on experimental performance, including: *the number of Experts K*, *the number of top-experts k*, *the embedding dimension d*, and *the max text length*, the corresponding search range of candidate values for different hyperparameters are shown in Table 2.

Hyperparameters	Search Space
# Number of Experts K	{2, 4, 6, 8, 10}
# Number of Top-experts k	{1, 2, 3, 4}
# Embedding Dimension d	{48, 64, 80, 96, 112}
# Max Text Length	{20, 30, 40, 50, 60}

Table 2: The main hyperparameters search space of MMoE model.

▹ **Baselines and Evaluations.** We compare our method with three types of baselines. (*i*) *vision-and-language pre-trained methods*: including **CLIP** [39], **ViLT** [20], **ALBEF** [22], and **METER** [14]; (*ii*) *generative-based methods*: including **GPT-3.5** [33], **GEMEL** [41], and **GELR** [27]; (*iii*) *multimodal interaction based methods*: including **DZMNED** [32], **JMEL** [1], **VELML** [60], **GHMFC** [52], **MIMIC** [31], **FissFuss** [30], **M^{3}EL** [17], **MELOV** [44] and **OTMEL** [59]. We note that although DWE [43] and UniMEL [24] also tackle the MEL task, they use different dataset sizes than that from the general benchmark [17, 24, 31, 44, 59] and different evaluation indicators. Thus, we do not consider them as baselines. Consistent with all baselines, we use the MRR and Hits@n (abbreviated as H@n, $n \in \{1, 3, 5\}$) as the evaluation metrics. Details about the description of baselines can be found in **Appendix A**.

4.2 Main Results

To address **RQ1**, we conduct experiments on the three datasets, the corresponding results are shown in Table 3 and Table 4.

▹ **Overall Performance.** Table 3 reports the performance of MMoE and baselines on three datasets. We can observe the following two findings: on the one hand, regardless of whether the DME module is incorporated, MMoE consistently achieves the best performance on all datasets, in particular, we observe that MMoE respectively achieves a 1.16% and 1.70% improvement on MRR and Hits@1 on WikiMEL compared to the best performing baseline. On the other hand, incorporating the DME module into MIMIC, M^{3}EL and MMoE results in notable performance improvements, indicating the necessity of enriching semantic information within mention context. Especially for the WikiDiverse dataset, the model incorporating the DME module demonstrates a significant improvement. We believe that the main reason is that in WikiDiverse the textual mention context not only has a shorter average sequence length (with an average of 11 words) but also contains a high number of abbreviations and acronyms. This greatly limits the availability of semantically relevant textual knowledge, leading to increased ambiguity when predicting the correct entities. DME module effectively incorporates WikiData entity descriptions into the semantic representation process of mentions, which complements the semantic knowledge of the mention context. Moreover, even without the DME module, MMoE still achieves the best performance under the same experimental conditions, and it attains optimal results when the DME is integrated with the MoE mechanism, underscoring the critical role of the interplay between DME and the other MMoE components.

▹ **Low Resource Settings.** Following [17, 31], we conduct additional experiments on the RichpediaMEL and WikiDiverse datasets in low-resource scenarios, where only 10% and 20% of the training data is used as the new training set while keeping the validation and test sets unchanged. The corresponding experimental results are shown in Table 4. We have the following two main findings: on the one hand, MMoE consistently achieves the best performance, however, no baseline consistently ranks second, indicating distinct model focuses and overly tight model-dataset coupling. On the other hand, for the RichpediaMEL dataset, the performance gains obtained by increasing the data from 10% to 20% are limited. This might be due to the more complete modal knowledge available in RichpediaMEL. For instance, in RichpediaMEL, 89.04% of the mentions come with visual images, whereas in WikiDiverse only 44.37%.

4.3 Ablation Studies

To address **RQ2**, we conduct ablation experiments to verify the contribution of each component of MMoE, including: a) impact of different losses; b) impact of different modules. The corresponding results for three datasets are shown in Table 5.

▹ **Impact of Different Losses.** The overall loss $\mathcal{L}$ in Equation 8 comprises four components: $\mathcal{L}_T$, $\mathcal{L}_V$, $\mathcal{L}_C$, and $\mathcal{L}_O$. We validate the necessity of each component by individually removing them. The experimental results are presented in the upper half of Table 5. It is evident that, despite the presence of some outliers, the removal of each loss component generally leads to a performance degradation. An example of an outlier is observed in the WikiDiverse dataset,

Methods	WikiMEL				RichpediaMEL				WikiDiverse			
Metrics	MRR	H@1	H@3	H@5	MRR	H@1	H@3	H@5	MRR	H@1	H@3	H@5
CLIP [39]♢	88.23	83.23	92.10	94.51	77.57	67.78	85.22	90.04	71.69	61.21	79.63	85.18
ViLT [20]♢	79.46	72.64	84.51	87.86	56.63	45.85	62.96	69.80	45.22	34.39	51.07	57.83
ALBEF [22]♢	84.56	78.64	88.93	91.75	75.29	65.17	82.84	88.28	69.93	60.59	75.59	81.30
METER [14]♢	79.49	72.46	84.41	88.17	74.15	63.96	82.24	87.08	63.71	53.14	70.93	77.59
GPT-3.5-Turbo [41]♦	–	73.80	–	–	–	–	–	–	×	×	×	×
GEMEL [41]♦	–	82.60	–	–	–	–	–	–	×	×	×	×
GELR [27]♦	–	84.80	–	–	–	–	–	–	×	×	×	×
DZMNED [32]♢	84.97	78.82	90.02	92.62	76.63	68.16	82.94	87.33	67.59	56.90	75.34	81.41
JMEL [1]♢	73.39	64.65	79.99	84.34	60.06	48.82	66.77	73.99	48.19	37.38	54.23	61.00
VELML [60]♢	83.42	76.62	88.75	91.96	77.19	67.71	84.57	89.17	66.13	54.56	74.43	81.15
GHMFC [52]♢	83.36	76.55	88.40	92.01	80.76	72.92	86.85	90.60	70.99	60.27	79.40	84.74
FissFuse [30]♦	92.02	87.89	95.36	–	–	–	–	–	×	×	×	×
MELOV [44]♦	92.32	88.91	95.61	96.58	88.80	84.14	92.81	94.89	76.57	67.32	83.69	87.54
OT-MEL [59]♦	92.59	88.97	95.63	96.96	88.27	83.30	92.39	94.83	75.43	66.07	82.82	87.39
MIMIC [31]♢	91.82	87.98	95.07	96.37	86.95	81.02	91.77	94.38	73.44	63.51	81.04	86.43
MIMIC+DME	92.37	88.66	95.67	97.10	87.49	81.64	92.22	94.69	82.38	75.70	87.30	90.57
Δ	+0.55	+0.68	+0.60	+0.73	+0.54	+0.62	+0.45	+0.31	+8.94	+12.19	+6.26	+4.14
M^{3}EL [17]♦	92.30	88.84	95.20	96.71	88.26	82.82	92.73	**95.34**	81.29	74.06	86.57	90.04
M^{3}EL+DME	92.51	88.91	95.49	96.98	88.84	83.89	92.84	94.95	83.53	76.71	88.79	92.25
Δ	+0.21	+0.07	+0.29	+0.27	+0.58	+1.07	+0.11	-0.39	+2.24	+2.65	+2.22	+2.21
MMoE	92.53	89.07	95.36	97.02	89.04	84.36	92.93	94.64	81.72	74.59	87.25	90.33
MMoE+DME	**93.75**	**90.77**	**96.27**	**97.41**	**89.86**	**85.51**	**93.43**	95.31	**84.23**	**77.57**	**89.12**	**92.49**
Δ	+1.22	+1.70	+0.91	+0.39	+0.82	+1.15	+0.50	+0.67	+2.51	+2.98	+1.87	+2.16
improvement (%)	+1.16	+1.70	+0.60	+0.31	+0.82	+1.15	+0.50	-0.03	+0.70	+0.86	+0.33	+0.24

Table 3: Evaluation of different models on three MEL datasets, ♢ results are from [31], ♦ results are from the corresponding papers. Best scores are highlighted in bold. Δ represents the performance improvement before and after adding the DME module. The *improvement* (%) represents the performance improvement over the state-of-the-art modes. – means that the corresponding models does not provide the corresponding dataset results, and × means that the dataset size used by the corresponding models is different from other baselines and cannot be directly compared. FissFuse [30], MELOV [44], and OT-MEL [59] do not have open source code, so we could not perform +DME experiments.

where the exclusion of the visual modality-related loss results in a slight improvement. This phenomenon can be attributed to two potential reasons: on the one hand, the availability of visual information in WikiData is sparse, with only 44.37% of the mentions containing visual content, which is considerably lower than the 85.65% and 89.04% in WikiMEL and RichpediaMEL, respectively. On the other hand, the visual information may not be closely related to the mentions, introducing noise that could adversely affect performance. Despite these data-induced outliers, the overall trend indicates that removing any loss component results in performance degradation.

▹ **Impact of Different Modules.** To better understand the contribution to the performance of each module in the MMoE framework, we individually remove the `IntraMoE` for the textual and visual modalities, and the `InterMoE` for textual-visual modality interaction. The experimental results are presented in Table 5. The results show that removing module generally leads to performance degradation, with the most significant drop occurring when `IntraMoE-T` is removed. This is mainly because the textual information is more relevant for the MEL task than the visual one, thus having a greater impact on accuracy. Further analysis reveals that removing a module of a given modality has a greater impact than only removing the corresponding modality's loss. The primary reason lies in the fact that removing a modality implies that the corresponding modal knowledge is not fed into the model, whereas removing the modality's loss allows the knowledge to be passed into the model without explicit supervision on the modal knowledge. However, the modal knowledge still participates in the interactions among various modules of the model, implicitly contributing to the training process through these interactions.

4.4 Parameter Sensitivity Analysis

To address **RQ3**, we carry out parameter sensitivity experiments on the WikiMEL, RichpediaMEL and WikiDiverse datasets, mainly includes the following four aspects: a) effect of the number of experts and top-experts; b) effect of learning rates; c) effect of embedding dimension; and d) effect of max text length. The corresponding results for three datasets are shown in Figure 3.

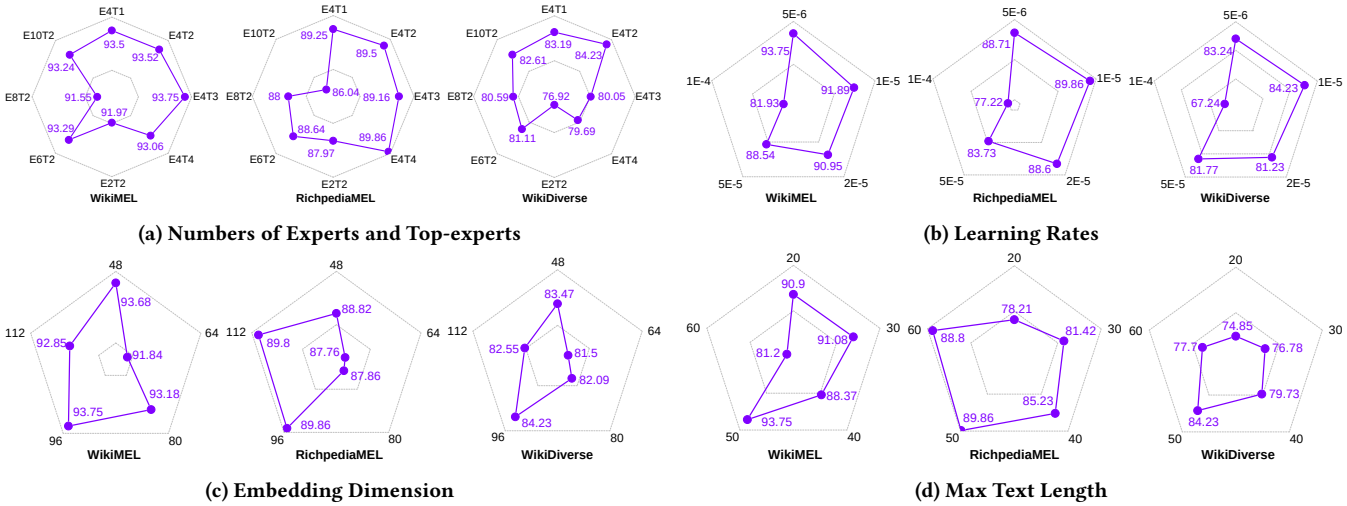
▹ **Effect of the Numbers of Experts and Top-experts.** In the SMoE mechanism, two critical hyperparameters, *i.e.,* the number of experts and the number of top-experts jointly determine the

Methods	RichpediaMEL(10%)				RichpediaMEL(20%)				WikiDiverse(10%)				WikiDiverse(20%)			
Metrics	MRR	H@1	H@3	H@5	MRR	H@1	H@3	H@5	MRR	H@1	H@3	H@5	MRR	H@1	H@3	H@5
DZMNED	31.79	22.57	34.95	41.33	47.01	36.38	52.25	58.28	19.99	11.45	22.52	29.50	40.97	28.73	47.35	56.69
JMEL	25.01	16.70	27.68	33.63	39.38	28.92	43.35	50.59	28.26	19.97	32.19	37.58	39.05	29.26	44.23	49.90
VELML	35.52	27.15	38.60	43.99	59.24	48.85	64.91	71.76	40.70	30.51	46.20	52.36	54.76	43.65	61.36	67.66
GHMFC	76.69	68.00	83.38	87.73	80.42	72.57	86.69	90.15	59.56	48.08	66.31	74.25	63.46	51.73	71.85	78.54
MIMIC	74.62	64.49	82.03	87.59	82.73	75.60	88.63	91.72	69.70	60.54	76.18	81.33	63.46	61.01	77.67	83.35
M^3EL	78.68	69.54	85.77	89.70	82.04	74.20	88.32	91.89	74.61	66.36	80.70	84.31	77.90	70.36	83.16	87.63
MMoE	**86.08**	**80.63**	**90.54**	**93.12**	**86.51**	**81.19**	**90.82**	**93.18**	**78.04**	**70.93**	**82.96**	**86.72**	**81.25**	**74.25**	**86.62**	**89.89**

Table 4: Evaluation of different models in low resource settings on four MEL datasets, all the baselines results are from [17]. FissFuse [30], MELOV [44], and OT-MEL [59] do not provide access to their code, so low resource experimental results could not be obtained.

Methods	WikiMEL			RichpediaMEL			WikiDiverse		
Metrics	MRR	Hits@1	Hits@3	MRR	Hits@1	Hits@3	MRR	Hits@1	Hits@3
w/o $\mathcal{L}_T$	92.29 ↓1.46	88.95 ↓1.82	95.03 ↓1.24	87.38 ↓2.48	82.03 ↓3.48	91.77 ↓1.66	82.66 ↓1.57	75.75 ↓1.82	87.73 ↓1.39
w/o $\mathcal{L}_V$	92.17 ↓1.58	88.82 ↓1.95	94.74 ↓1.53	88.45 ↓1.41	83.60 ↓1.91	92.36 ↓1.07	84.88 ↑0.65	78.78 ↑0.21	89.89 ↑0.77
w/o $\mathcal{L}_C$	91.55 ↓2.20	88.16 ↓2.61	94.27 ↓2.00	85.30 ↓4.56	80.71 ↓4.80	88.66 ↓4.77	83.38 ↓0.85	76.85 ↓0.72	88.59 ↓0.53
w/o $\mathcal{L}_O$	92.53 ↓1.22	89.07 ↓1.70	95.36 ↓0.91	89.04 ↓0.82	84.36 ↓1.15	92.93 ↓0.50	83.66 ↓0.57	77.09 ↓0.48	88.88 ↓0.24
w/o IntraMoE-T	88.05 ↓5.70	84.50 ↓6.27	90.54 ↓5.73	78.64 ↓11.22	73.67 ↓11.84	81.47 ↓11.96	66.08 ↓18.15	56.69 ↓20.88	71.75 ↓17.37
w/o IntraMoE-V	90.36 ↓3.39	86.32 ↓4.45	93.31 ↓2.96	87.64 ↓2.22	82.82 ↓2.69	91.16 ↓2.27	84.29 ↑0.06	78.25 ↑0.68	88.88 ↓0.24
w/o InterMoE	91.41 ↓2.34	87.95 ↓2.82	94.00 ↓2.27	88.30 ↓1.56	83.04 ↓2.47	92.76 ↓0.67	82.66 ↓1.57	75.99 ↓1.58	87.58 ↓1.54
MMoE	**93.75**	**90.77**	**96.27**	**89.86**	**85.51**	**93.43**	**84.23**	**77.57**	**89.12**

Table 5: Evaluation of ablation studies on three MEL datasets.



(a) Numbers of Experts and Top-experts

(b) Learning Rates

(c) Embedding Dimension

(d) Max Text Length

Figure 3: Parameter sensitivity experiments under different conditions on WikiMEL, RichpediaMEL and WikiDiverse datasets.

model's effectiveness and efficiency. The number of experts primarily affects the model's capacity and diversity, while the number of top-experts primarily influences the computational efficiency and the balance between performance and resource consumption. Selecting an appropriate number of experts and top-experts is crucial for balancing the model's performance and efficiency. We set the number of experts to {2, 4, 6, 8, 10} and the number of top-experts to {1, 2, 3, 4}, creating 8 experimental conditions to verify the impact of these parameters. The corresponding results on three datasets are shown in Figure 3(a). We can draw the following three conclusions: (*i*): Increasing the number of experts can enhance the model's expressive capacity to capture diverse features from different perspectives, but more experts are not always better. For instance, on the WikiMEL dataset, E6T2 outperforms E8T2; (*ii*): Top experts

dynamically select suitable experts to process different samples by activating certain experts. The appropriate choice of top value is a trade-off between performance and resource efficiency. Generally, when memory capacity is large, we prefer to set a higher number of experts, but the number of top-experts is usually kept relatively small; (*iii*): Under the optimal conditions across the three datasets, the values for the number of experts and top experts differ. For WikiMEL, E4T3 achieves the best results, while for RichpediaMEL and WikiDiverse, E4T4 and E4T2 are optimal, respectively.

▹ **Effect of Learning Rates.** A larger learning rate can expedite the convergence of the model but may lead to the omission of the optimal solution during the optimization process. Conversely, a lower learning rate facilitates a more stable convergence to the optimal solution, albeit at a slower rate. Therefore, selecting the learning rate involves a trade-off between convergence speed and optimal performance. We conducted experiments with five different learning rates across three datasets, with the corresponding results illustrated in Figure 3(b). We observed that the trends across the three datasets were generally consistent with respect to the learning rates: a learning rate of 1E-4 resulted in the poorest performance, while a learning rate of 1E-5 achieved the best performance. We attribute this to the stability of the chosen optimization algorithm.

▹ **Effect of Embedding Dimension.** High-dimensional embeddings can capture more complex semantic information, enhancing the representational capacity of the model, but they also require more computational resources. In contrast, low-dimensional embeddings can effectively reduce the complexity of the model, which helps to prevent overfitting. We conduct a progressive comparison experiment on three datasets, selecting 5 different dimensionality values. The corresponding results are shown in Figure 3(c). We can observe that when the embedding dimension is set to 96, the model consistently achieves the best performance across all three datasets. Additionally, increasing the embedding dimension does not lead to an improvement in performance, which is consistent with the findings of [25]. In practice, it is crucial to select the smallest embedding dimension that yields optimal performance.

▹ **Effect of Max Text Length.** The max length of text sequences in a mention context is a significant factor influencing the representational capacity of the textual modality. Longer texts contain a greater number of words, but they may also introduce more noisy-words that are less relevant to the mentions. Selecting an appropriate maximum text length requires a comprehensive consideration of the characteristics of the dataset itself, specifically analyzing the sequence lengths of the majority of data points in the dataset. We conducted comparative experiments with 5 different maximum sequence lengths across three datasets, as shown in Figure 3(d). The results indicate that the performance improves with the increase of the length for all three datasets, peaking when the sequence length is set to 50. Further increases in sequence length beyond this point, however, lead to a decline in performance. This is reasonable, as excessively short text sequences provide limited semantic coverage and insufficient support for predicting the corresponding entities, while excessively long sequences may disrupt the representation of mentions, leading to adverse effects.

4.5 Complexity Analysis

We conduct a comparative analysis of the computational complexity of each module within the MMoE model, with the results detailed in Table 6. The computational complexity is assessed from two dimensions, the number of floating-point operations (#FLOPs) and the number of parameters (#Params). We observe that the SMoE module bears the majority of the computational burden, as evidenced by the significant reduction in complexity values upon its removal. This observation is reasonable, given that the SMoE mechanism involves a substantial number of feed-forward neural networks (FFNs), which are the primary contributors to both computational and temporal overhead. Furthermore, upon removing the IntraMoE and InterMoE modules, the computational costs were ranked as follows: `IntraMoE-T` > `InterMoE` > `IntraMoE-V`, with `IntraMoE-T` incurring in significantly higher costs than the other two. This is mainly due to the higher embedding dimensions in the intermediate layers of the SMoE module within `IntraMoE-T`.

Additionally, we compared the training convergence times between the MMoE and MIMIC. On the WikiMEL, RichpediaMEL, and WikiDiverse datasets, the convergence times for MMoE are 1.88 hours, 1.53 hours, and 1.10 hours, respectively, while for the MIMIC, are 1.63 hours, 1.15 hours, and 0.95 hours, respectively. Although the MMoE requires relatively more time to train, this computational cost remains within an acceptable range.

#FLOPs	WikiMEL	RichpediaMEL	WikiDiverse
w/o `IntraMoE-T`	3.135G	3.135G	3.135G
w/o `IntraMoE-V`	18.435G	18.435G	18.435G
w/o `InterMoE`	17.325G	17.325G	17.325G
w/o `SMoE`	481.333M	481.333M	481.333M
`MMoE`	19.443G	19.443G	19.443G
#Params	**WikiMEL**	**RichpediaMEL**	**WikiDiverse**
w/o `IntraMoE-T`	683.849K	683.849K	683.849K
w/o `IntraMoE-V`	5.378M	5.378M	5.378M
w/o `InterMoE`	5.347M	5.347M	5.347M
w/o `SMoE`	295.905K	295.905K	295.905K
`MMoE`	5.703M	5.703M	5.703M

Table 6: Amount of parameters and calculation for MMoE model without various modules.

5 Conclusions

We introduced `MMoE` which simultaneously considers mention ambiguity and dynamic selection of different regions of information. We leverage LLMs to rank WikiData entity descriptions related to mentions to alleviate the issue of insufficient semantically relevant knowledge in mention contexts. In addition, we introduce a switch mixture-of-experts mechanism to dynamically specify the importance of different textual tokens and visual patches. Extensive empirical experiments and ablation studies confirm MMoE's outstanding performance. In the future, we plan to explore tasks in specific domains, such as the biomedical and drug discovery domains. Moreover, the visual knowledge of three datasets is very sparse, thus how to complete it is an interesting research direction.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.62376144), by the Science and Technology Cooperation and Exchange Special Project of Shanxi Province (No.202204041101016), by the Key Research and Development Project of Shanxi Province (No.202102020101008), by the Shanxi Natural Language Processing Innovation Team (Shanxi Talents) Project.

References

[1] Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020. Multimodal Entity Linking for Tweets. In *ECIR*. 463–478.
[2] Ali Ahmadvand, Harshita Sahijwani, Jason Ingyu Choi, and Eugene Agichtein. 2019. ConCET: Entity-Aware Topic Classification for Open-Domain Conversational Agents. In *CIKM*. 1371–1380.
[3] Meta AI. 2024. Introducing meta llama 3: The most capable openly available LLM to date.
[4] Ujjwala Anantheswaran, Himanshu Gupta, Kevin Scaria, Shreyas Verma, Chitta Baral, and Swaroop Mishra. 2024. Investigating the Robustness of LLMs on Math Word Problems. *arXiv* (2024).
[5] Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. ReFinED: An Efficient Zero-shot-capable Approach to End-to-End Entity Linking. In *NAACL*. 209–220.
[6] Ilaria Bordino, Yelena Mejova, and Mounia Lalmas. 2013. Penguins in sweaters, or serendipitous entity search on user-generated content. In *CIKM*. 109–118.
[7] Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive Entity Retrieval. In *ICLR*. 1–20.
[8] Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2022. Multilingual Autoregressive Entity Linking. In *TACL*. 274–290.
[9] Yixin Cao, Lei Hou, Juanzi Li, and Zhiyuan Liu. 2018. Neural Collective Entity Linking. In *COLING*. 675–686.
[10] Jiaoyan Chen, Freddy Lecue, Jeff Z. Pan, Ian Horrocks, and Huajun Chen. 2018. Knowledge-based Transfer Learning Explanation. In *KR*. 349–358.
[11] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *EMNLP*. 1724–1734.
[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*. 4171–4186.
[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*. 1–22.
[14] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. 2022. An Empirical Study of Training End-to-End Vision-and-Language Transformers. In *CVPR*. 18145–18155.
[15] Biralatei Fawei, Jeff Z. Pan, Martin J. Kollingbaum, and Adam Z. Wyner. 2020. A Semi-automated Ontology Construction for Legal Question Answering. *New Generation Computing* (2020), 453–478.
[16] Bowei He, Xu He, Yingxue Zhang, Ruiming Tang, and Chen Ma. 2023. Dynamically Expandable Graph Convolution for Streaming Recommendation. In *WWW*. 1457–1467.
[17] Zhiwei Hu, Víctor Gutiérrez-Basulto, Ru Li, and Jeff Z. Pan. 2024. Multi-level Matching Network for Multimodal Entity Linking. *arXiv* (2024).
[18] Zhiwei Hu, Víctor Gutiérrez-Basulto, Zhiliang Xiang, Ru Li, and Jeff Z. Pan. 2024. Leveraging Intra-modal and Inter-modal Interaction for Multi-Modal Entity Alignment. *arXiv* (2024).
[19] Wenyu Huang, Guancheng Zhou, Mirella Lapata, Pavlos Vougiouklis, Sebastien Montella, and Jeff Z Pan. 2025. Prompting Large Language Models with Knowledge Graphs for Question Answering involving Long-tail Facts. *Knowledge-Based Systems* (2025).
[20] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *ICML*. 5583–5594.
[21] Phong Le and Ivan Titov. 2018. Improving Entity Linking by Modeling Latent Relations between Mentions. In *ACL*. 1595–1604.
[22] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *NeurIPS*. 9694–9705.
[23] Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2018. Knowledge-aware Multimodal Dialogue Systems. In *MM*. 801–809.
[24] Qi Liu, Yongyi He, Tong Xu, Defu Lian, Che Liu, Zhi Zheng, and Enhong Chen. 2024. UniMEL: A Unified Framework for Multimodal Entity Linking with Large Language Models. In *CIKM*. 1909–1919.
[25] Xiao Liu, Shiyu Zhao, Kai Su, Yukuo Cen, Jiezhong Qiu, Mengdi Zhang, Wei Wu, Yuxiao Dong, and Jie Tang. 2022. Mask and Reason: Pre-Training Knowledge Graph Transformers for Complex Logical Queries. In *KDD*. 1120–1130.
[26] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* (2019).
[27] Xinwei Long, Jiali Zeng, Fandong Meng, Jie Zhou, and Bowen Zhou. 2024. Trust in Internal or External Knowledge? Generative Multi-Modal Entity Linking with Knowledge Retriever. In *ACL*. 7559–7569.
[28] Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-Based Knowledge Conflicts in Question Answering. In *EMNLP*. 7052–7063.
[29] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *ICLR*. 1–11.
[30] Pengfei Luo, Tong Xu, Che Liu, Suojuan Zhang, Linli Xu, Minglei Li, and Enhong Chen. 2024. Bridging Gaps in Content and Knowledge for Multimodal Entity Linking. In *MM*. 9311–9320.
[31] Pengfei Luo, Tong Xu, Shiwei Wu, Chen Zhu, Linli Xu, and Enhong Chen. 2023. Multi-Grained Multimodal Interaction Network for Entity Linking. In *KDD*. 1583–1594.
[32] Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal Named Entity Disambiguation for Noisy Social Media Posts. In *ACL*. 2000–2008.
[33] OpenAI. 2023. Chatgpt.
[34] Jeff Pan, Stuart Taylor, and Edward Thomas. 2009. Reducing Ambiguity in Tagging Systems with Folksonomy Search Expansion. In *ESWC*. 669–683.
[35] Jeff Z. Pan, Guido Vetere, Jose Manuel Gomez-Perez, and Honghan Wu (Eds.). 2017. *Exploiting Linked Data and Knowledge Graphs for Large Organisations*. Springer.
[36] Jeff Z. Pan, Mei Zhang, Kuldeep Singh, Frank Van Harmelen, Jinguang Gu, and Zhi Zhang. 2019. Entity Enabled Relation Linking. In *ISWC*. 523–538.
[37] Matthew E. Peters, Mark Neumann, Robert L. Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge Enhanced Contextual Word Representations. In *EMNLP*. 43–54.
[38] Flavio Petruzzellis, Alberto Testolin, and Alessandro Sperduti. 2024. Assessing the Emergent Symbolic Reasoning Abilities of Llama Large Language Models. *arXiv* (2024).
[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*. 8748–8763.
[40] Chenwei Ran, Wei Shen, and Jianyong Wang. 2018. An Attention Factor Graph Model for Tweet Entity Linking. In *WWW*. 1135–1144.
[41] Senbao Shi, Zhenran Xu, Baotian Hu, and Min Zhang. 2024. Generative Multimodal Entity Linking. In *COLING*. 7654–7665.
[42] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*. 1–14.
[43] Shezheng Song, Shan Zhao, Chengyu Wang, Tianwei Yan, Shasha Li, Xiaoguang Mao, and Meng Wang. 2024. A Dual-Way Enhanced Framework from Text Matching Point of View for Multimodal Entity Linking. In *AAAI*. 19008–19016.
[44] Xuhui Sui, Ying Zhang, Yu Zhao, Kehui Song, Baohang Zhou, and Xiaojie Yuan. 2024. MELOV: Multimodal Entity Linking with Optimized Visual Features in Latent Space. In *ACL*. 816–826.
[45] Edward Thomas, Jeff Z. Pan, and Derek H. Sleeman. 2007. ONTOSEARCH2: Searching Ontologies Semantically. In *OWLED*.
[46] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv* (2023).
[47] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv* (2023).
[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NeurIPS*. 5998–6008.

[49] Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* (2014), 78–85.
[50] Hongru Wang, Wenyu Huang, Yufei Wang, Yuanhao Xi, Jianqiao Lu, Huan Zhang, Nan Hu, Zeming Liu, Jeff Z. Pan, and Kam-Fai Wong. 2025. Rethinking Stateful Tool Use in Multi-Turn Dialogues: Benchmarks and Challenges. In *ACL*.
[51] Meng Wang, Haofen Wang, Guilin Qi, and Qiushuo Zheng. 2020. Richpedia: A Large-Scale, Comprehensive Multi-Modal Knowledge Graph. *Big Data Res.* (2020), 1–11.
[52] Peng Wang, Jiangheng Wu, and Xiaohang Chen. 2022. Multimodal Entity Linking with Gated Hierarchical Fusion and Contrastive Training. In *SIGIR*. 938–948.
[53] Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. 2022. WikiDiverse: A Multimodal Entity Linking Dataset with Diversified Contextual Topics and Entity Types. In *ACL*. 4785–4797.
[54] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In *EMNLP*. 6397–6407.
[55] Wenhan Xiong, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. Improving Question Answering over Incomplete KBs with Knowledge-Aware Reader. In *ACL*. 4258–4264.
[56] Dongjie Zhang and Longtao Huang. 2022. Multimodal Knowledge Learning for Named Entity Disambiguation. In *EMNLP*. 3160–3169.
[57] Gongrui Zhang, Chenghuan Jiang, Zhongheng Guan, and Peng Wang. 2023. Multimodal Entity Linking with Mixed Fusion Mechanism. In *DASFAA*. 607–622.
[58] Li Zhang, Zhixu Li, and Qiang Yang. 2021. Attention-Based Multimodal Entity Linking with High-Quality Images. In *DASFAA*. 533–548.
[59] Zefeng Zhang, Jiawei Sheng, Chuang Zhang, Liangyunzhi Liangyunzhi, Wenyuan Zhang, Siqi Wang, and Tingwen Liu. 2024. Optimal Transport Guided Correlation Assignment for Multimodal Entity Linking. In *ACL*. 4103–4117.
[60] Qiushuo Zheng, Hao Wen, Meng Wang, and Guilin Qi. 2022. Visual Entity Linking via Multi-modal Learning. *Data Intell.* (2022), 1–19.
[61] Jeff Z.Pan, Diego Calvanese, Thomas Eiter, Ian Horrocks, Michael Kifer, Fangzhen Lin, and Yuting Zhao (Eds.). 2017. *Reasoning Web: Logical Foundation of Knowledge Graph Construction and Querying Answering*. Springer.

Appendix

A Baselines

We compare MMoE with three types of baselines, including *vision-and-language pre-trained methods*, *generative-based methods* and *multimodal interaction based-methods*.

The **vision-and-language pre-trained methods** include:

- **CLIP** [39] adopts BERT and ViT as the textual and visual embedding encoders, and designs a contrastive learning mechanism to bring paired text-image together, and pushes unpaired texts and images away from each other.
- **ViLT** [20] focuses on correlations between multiple modalities through a series of Transformer [48] layers.
- **ALBEF** [22] uses a text-image contrastive loss to align textual and visual features and further combines them together using a multimodal Transformer encoder.
- **METER** [14] investigates how to train a performant vision-and-language Transformer, and attributes the factors affecting model performance to a vision encoder, text encoder, multimodal fusion module, and architectural design.

The **generative-based methods** include:

- **GPT-3.5** [33] is a powerful large language model developed by OpenAI.
- **GEMEL** [41] keeps the vision and language model frozen and trains a feature mapper to enable cross-modality interactions, further applies a constrained decoding strategy to efficiently search the valid entity space.
- **GELR** [27] incorporates a knowledge retriever to enhance visual entity information by leveraging external sources, and devises a prioritization scheme to manage conflicts arising from the integration of external and internal knowledge.

The **multimodal interaction based methods** include:

- **DZMNED** [32] incorporates an attention mechanism to integrate textual-level, visual-level, and character-level features.
- **JMEL** [1] utilizes both unigram and bigram embeddings as textual features, and further concatenate these features to pass through a fully connected layer.
- **VELML** [60] extracts visual features via a VGG16 [42] network and replaces a GRU [11] textual encoder with pre-trained BERT, and further designs a deep modal-attention mechanism to aggregate all modalities features.
- **GHMFC** [52] obtains the hierarchical features of textual and visual co-attention through the multi-modal co-attention mechanism, and proposes a contrastive loss to learn more generic multimodal features and reduce noise.
- **FissFuss** [30] integrates the Fission and Fusion branches, establishing dynamic features for each mention-entity pair and adaptively learning multimodal interactions to alleviate content discrepancy.
- **MELOV** [44] incorporates inter-modality generation and intra-modality aggregation, effectively leveraging the shared information from heterogeneous textual features and relevant visual details of semantic similar neighbors.
- **OT-MEL** [59] formulates the correlation assignment problem as an optimal transport problem, and exploits the multimodal correlation between mentions and entities to enhance the fine-grained matching.
- **MIMIC** [31] devises three different interaction matching modules to explore both intra-modal and inter-modal interactions among the features of entities and mentions.
- **M^{3}EL** [17] introduces an intra-modal contrastive learning mechanism to obtain better discriminative embedding representations, and devises intra-modal and cross-modal matching networks to explore different multimodal interactions.

B Case Study

To further quantitatively analyze the MMoE model, we select two test cases from the WikiDiverse dataset, with the top-3 predicted entities for each mention shown in Figure 4. Notably, both test cases require simultaneously predicting the entities corresponding to multiple mentions, yet they share identical visual images. This poses a challenge for existing MEL models to link mentions to the appropriate entities, as it requires filtering the most relevant visual knowledge based on the mentions to be predicted. However, MMoE effectively addresses the issue of fine-grained modal knowledge selection for mentions by dynamically selecting different patch regions of the visual modality according to the specific mentions being predicted. Furthermore, we observe that MMoE does not perform well for the prediction of mention *Tsinghua Science Park*. The main reason is that the image in this case depicts a building, which lacks any content relevant to mention *Tsinghua Science Park*. Consequently, MMoE struggles to capture visual knowledge associated with *Tsinghua Science Park*, making mispredictions inevitable.

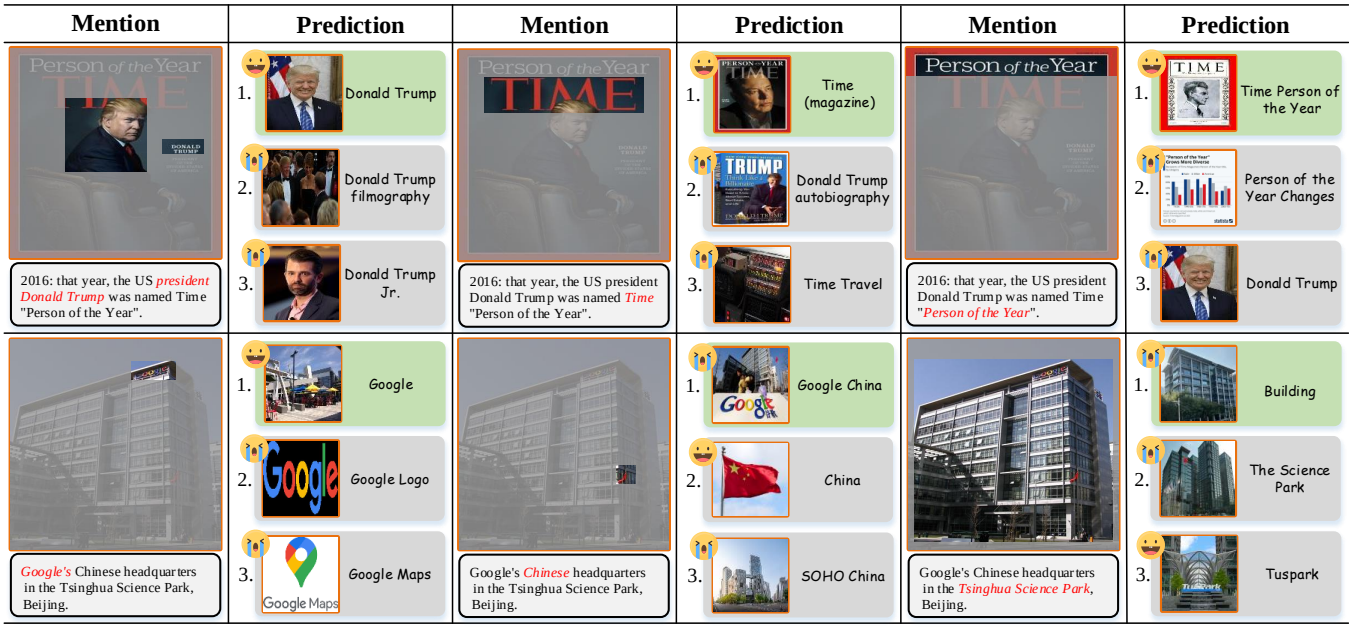


Figure 4: Case study of MMoE. These are two cases from the test set of WikiDiverse, each of them needing to predict three mentions with multimodal context. We display the top 3 ranked entities by MMoE. To better showcase the result, we omit the descriptions of mentions.

Methods	WikiMEL				RichpediaMEL				WikiDiverse			
Metrics	MRR	Hits@1	Hits@3	Hits@5	MRR	Hits@1	Hits@3	Hits@5	MRR	Hits@1	Hits@3	Hits@5
LLaMA2-7B	90.72	86.71	93.85	95.65	83.48	76.22	89.61	92.39	74.34	64.87	81.18	85.85
LLaMA3.1-8B	90.84	86.67	94.23	96.09	88.05	83.30	91.83	93.94	82.73	75.79	88.07	91.39
MMoE	**93.75**	**90.77**	**96.27**	**97.41**	**89.86**	**85.51**	**93.43**	**95.31**	**84.23**	**77.57**	**89.12**	**92.49**

Table 7: Evaluation of different LLMs of DME module on three MEL datasets.

C Additional Results

▹ **Different LLMs of DME module.** In the DME module, we use by default GPT-3.5 as the ranking model for entity descriptions to identify the most contextually appropriate descriptions. To further investigate the ability of other open-source large language models to understand the relationship between entity description knowledge and mention context, we replace GPT-3.5 with LLaMA2-7B [47] and LLaMA3.1-8B [3]. The corresponding experimental results are presented in Table 7. We can observe that the choice of ranking model has a significant impact on performance. Compared to GPT-3.5, both LLaMA2 and LLaMA3.1 perform slightly worse, with LLaMA3.1 outperforming LLaMA2. This outcome aligns with expectations, as GPT-3.5 consistently demonstrates superior performance across many task scenarios [4, 38].