Multimodal Entity Linking with Dynamic Modality Selection and Interactive Prompt Learning

Anonymous ACL submission

Abstract

Recent advances in Multimodal Entity Linking (MEL) utilize multimodal information to link target mentions to corresponding entities. However, existing methods uniformly adopt a "one-size-fits-all" approach, ignoring individual sample needs and modality-induced noise. Also, the commonly used separate large-scale visual and text pre-trained models for feature extraction do not address inter-modal heterogeneity and the high computational cost of fine-tuning. To resolve these two issues, this paper introduces a novel approach named Multimodal Entity Linking with Dynamic Modality Selection and Interactive Prompt Learning (DSMIP). First, we design three expert networks that utilize different subsets of modalities to tackle the task and train them individually. In particular, for the multimodal expert network, we extract multimodal features of entities and mentions by updating multimodal prompts and set up a coupling function to realize the interaction of prompts between modalities. Subsequently, to select the best-suited expert network for each specific sample, we devise a Modality Selection Gating Network to gain the optimal one-hot selection vector by applying a specialized reparameterization technique and a two-stage training. Experimental results on three public benchmark datasets demonstrate that our solution outperforms the majority of state-of-the-art baselines and surpasses all baselines in settings with low training resources.

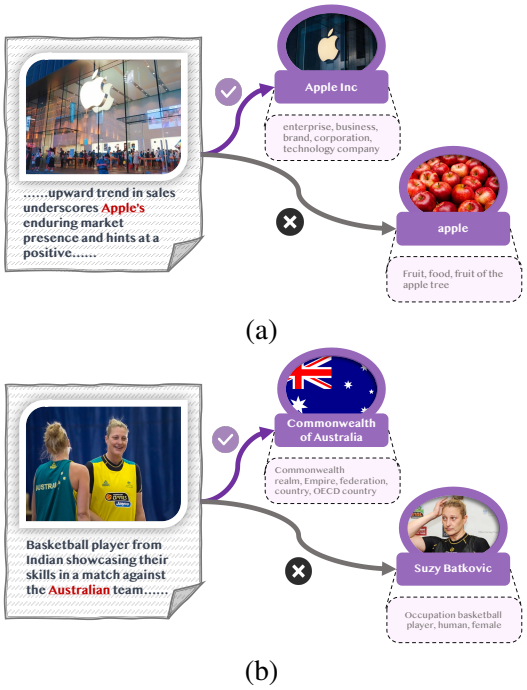

Figure 1: Two examples of Multimodal Entity Linking. On the left are the textual context and image from a corpus, with the mention word underscored, on the right are the entity name, attributes, and image from a knowledge base. In (a), all two modalities are needed to correctly link the mention "Apple" to the company "Apple Inc." In (b), the text modality alone suffices to correctly link to the country "Australia," but adding the visual modality leads to an erroneous link to the female basketball player "Suzy Batkovic."

1 Introduction

Entity Linking (EL), also known as entity disambiguation, aims to map mentions within unstructured data from sources such as social media, news, or web content to the correct entities in a structured Knowledge Graph (KG), which benefits numerous downstream tasks, including information extraction (Hoffart et al., 2011), question answering (Yih et al., 2015) and semantic search (Hasibi et al., 2016). Traditional EL approaches primarily rely on the textual context of mentions to link to the correct knowledge base entities. However, in recent years, there has been an increasing amount of oneline information being conveyed through images, on the other hand, the textual context of mentions often fails to eliminate ambiguity, posing challenges to text-based methods as is shown in Figure 1(a). Therefore, an increasing attention has drawn to the research of Multimodal Entity Linking (MEL). Despite considerable improvements (Zhu et al., 2024) have been made in MEL, these methods still exhibit

several notable limitations:

Firstly, for the real-world data, two obstacles are considered the most significant and demanding in this task: (1) *Mention Ambiguity:* Ambiguity exists in both text and image for mentions. Textual mentions and contexts are often brief and may contain abbreviations, which is common in social media and web news. Also the related images might correspond to more than one entity (e.g. different characters played by the same actor). These ambiguities make it infeasible to identify the correct entity by uni-modal for some complex samples. (2) *Noise of sample data:* For MEL task, context serves as a crucial resource for disambiguation and searching for the correct entity. However, as text and image contexts are typically sourced from the internet, not all modalities' contexts work positively for the task. Textual context might contain information irrelevant to the mention word, and low-quality visual context can easily act as noise, affecting the accuracy of linking. Figure 1 visually illustrates that disparate samples may encounter distinct obstacles: for samples like (a), a single textual modality can lead to mention ambiguity, whereas multimodal data assists the model in learning richer representations and linking to the correct entity easier. For samples like (b), using the single text modality could easily identify the ground-trurh entity due to the strong specificity of the mention word, whereas the visual modality might be counterproductive. **Employing a universal model for all samples struggles to balance modality-assisted disambiguation and modality-induced noise.** Existing methods are static in essence, processing all instances with a single framework. Therefore, a dynamic method is required to select which modalities to use under different samples, which can both filter noisy modal and utilize multimodal information for disambiguation when necessary.

Secondly, it's essential to deeply mine the multimodal information of both mentions and entities for MEL. To this end, the model is required to not only leverage the semantic information of each but also recognize the interrelations between modalities. The fact that image features and word token embeddings reside in their respective spaces poses a challenge to construct a unified representation, so it is necessary to model the interaction of the modalities and enhance the inter-model effect. Quite a few recent MEL research, in the feature extraction stage, use large-scale pre-trained models to extract text and image features, then merge uni-modal features and fine-tune the encoder. However, existing methods extract uni-modal features independently, which can easily overlook the interactive clues hidden between modalities. Additionally, fine-tuning pre-trained models also leads to extensive computational costs. **Using separate large-scale pre-trained models for visual and textual feature extraction fail to tackle inter-modal heterogeneity and the high computational demands of fine-tuning.** Therefore, a method is needed to fully interact and align different modalities' information during the feature extraction stage in a cost-effective manner.

To tackle the above issues, this paper proposes a novel multi-modal entity linking method **Multimodal Entity Linking with Dynamic Modality Selection and Interactive Prompt Learning (DSMIP)**. First, we train three expert networks with different modal subsets for the entity linking task, each of which includes a feature encoding module and a matching score calculation module. To extract a unified multi-modal representation and eliminate heterogeneity between modalities, we introduce low-overhead multi-modal prompt learning for feature encoding in multimodal network. A coupling function is used to establish the interaction between textual and visual prompts during the feature extraction stage. Then, to eliminate the ambiguity of mentions, as well as dynamically select the required modalities when calculating the scores for each mention-entity sample pair, we design a modality selection gating network and update it with a reparameterization technique during back-propagation to generate discrete one-hot decisions. Moreover, during the training process, to avoid the dependency of a specific modality and enhance robustness, we adopted a two-stage training strategy and employed contrastive training loss to compute the final matching scores for entity-mention similarity. In summary, the contribution of this paper can be summarized as follows:

- We propose a dynamic modality selection gating network to solve MEL task, which select the optimal expert network for each individual sample, flexibly leveraging multimodal data for disambiguation and filtering out noisy modalities.
- Within the multimodal expert network, we adopt a method based on multimodal prompt learning to diminish computational overhead and implement a coupling function for the interaction of prompts across modalities.

- Experimental result on three public MEL datasets demonstrate that our method surpasses the current baseline models in performance.

2 Related Work

An increasing number of MEL methods have been proposed in the past few years, which incorporate additional multimodal information to help resolve the ambiguity of entities. Moon et al. (2018) first proposed a multimodal entity linking task. Adjali et al. (2020) introduced a Twitter dataset for social media and learned a dual-branch neural network to minimize triplet loss. Gan et al. (2021) modeled the text-visual mention alignment as a bipartite graph-matching problem and addressed it using an optimal transport-based linking method. Wang et al. (2022b) presented a novel WIKIDiverse dataset and investigated intra-modal and inter-modal attention to better align the two modalities. Wang et al. (2022a) leveraged transformers for fine-grained cross-modal relation mining in MEL tasks, employing gated fusion and contrastive training for meaningful multimodal representations. Luo et al. (2023) investigated entity-mention feature interactions across modalities via three interaction units of different granularities. Xing et al. (2023) explicitly modeled four different types of alignment of mention-entity by constructing graph convolutional networks (GCN). However, the aforementioned MEL methods commonly applied a uniform framework for all instances, thereby neglecting noise present in the modality of certain samples. Recently, research on Dynamic Neural Networks (Masoudnia and Ebrahimpour, 2014; Han et al., 2022) and its multimodal applications (Panda et al., 2021; Xue and Marculescu, 2023) has emerged, offering valuable insights for sample-targeted disambiguation of mentions and entities.

In recent years, with the advent of large-scale pre-trained models such as BERT, ViT, and CLIP, which have leveraged the abundant data available on the internet to enable the model to learn a wealth of knowledge. An increasing number of MEL methods have begun to adopt these model for feature extraction. Recent works (Luo et al., 2023; Yang et al., 2023) commonly used pre-trained BERT to extract textual features and pre-trained ViT or CLIP visual encoders for visual features, fine-tuning the models for multimodal entity linking tasks. Due to the high computational resource consumption and catastrophic forgetting issues during fine-tuning, recent studies have proposed language prompt learning (Zhou et al., 2022; Ju et al., 2022; Khattak et al., 2023), which involved constructing prompt tokens and automatically updating prompts to adjust the Vision-and-Language Pre-training (VLP) model while keeping the original weights frozen during fine-tuning. From the perspective of MEL, the idea of prompt learning can help reduce the fine-tuning overhead when using pre-trained models to extract multimodal features and also effectively address the heterogeneity problem between text and vision in MEL tasks through the interaction of prompts.

3 Methodology

3.1 Problem Formulation

The task of Multi-modal Entity Linking (MEL) is to link the mention in the corpus dataset to the corresponding entity in the knowledge graph, we use M to denote the n input multi-modal mention samples, where each mention sample can be defined as $m = (m_w, m_t, m_v)$, containing three parts: word of mention, text context of mention, and visual context of mention. We use E to represent the knowledge base, which contains millions of entities, where each entity can be denoted as $e = (e_n, e_a, e_v)$, with each element respectively represents the name, attributes and images of entity. Our task can be expressed as follows,

$$e^*(m) = \arg\max_{e_i \in E} Score(m; e_i) \quad (1)$$

where $Score(\cdot)$ is a score function used to calculate the similarity between mentions and entities, and $e^*(m)$ denotes the ground-truth entity with the highest score that is finally selected. In this section, for the same process of mention and entity, we only exhibit the formulas for mention.

3.2 Construction Of Three Expert Networks

In this subsection, we design three pre-trained expert networks to implement entity linking task, each utilizing a different subset of modalities (only textual modality, only visual modality, and both textual and visual modalities) to calculate the similarity scores of mention-entity pairs. The three expert networks are computed independently and are chosen by the selection gating network in Section 3.3. We first developed three modules to encode features of mentions and entities as illustrated in Figure 2, then the feature representations obtained are then processed through different scoring units

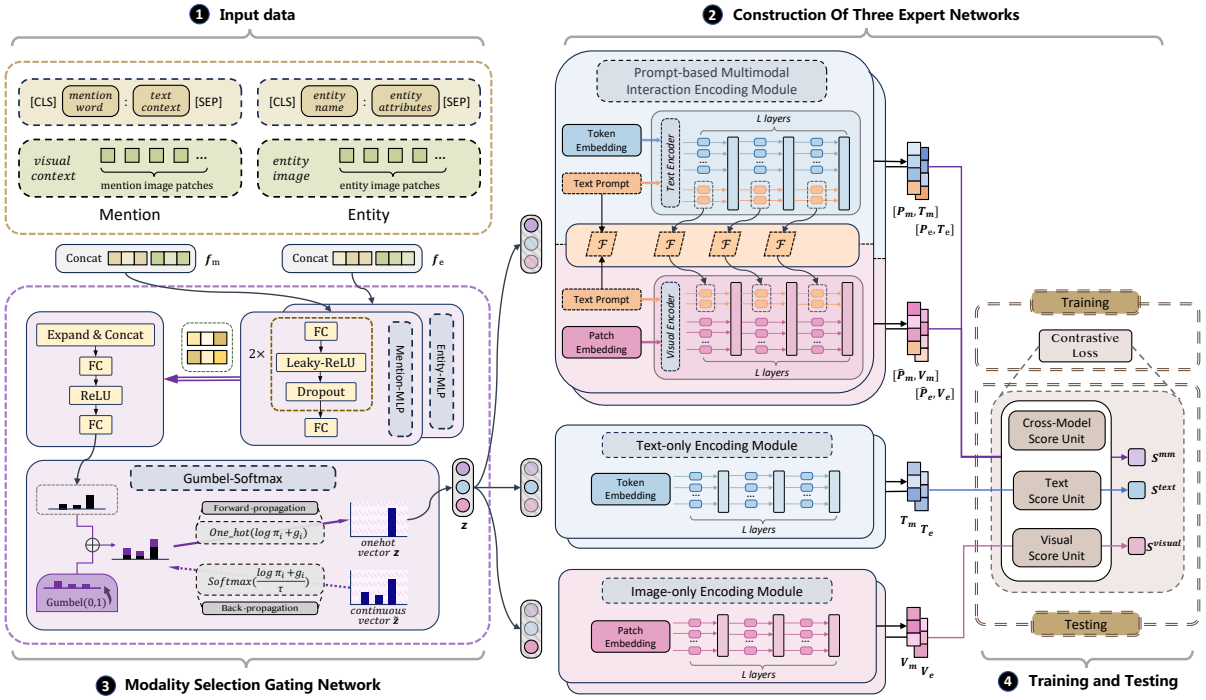

Figure 2: Overview of our proposed DSMIP framework for Multimodal Entity Linking.

to compute the similarity matching scores between mentions and entities. Three encoding modules and the scoring computation units of the different expert networks will be described in detail below.

3.2.1 Text-only Encoding Module

In this module, we first extract the textual features of mention and entity using the text encoder of CLIP, and make a template "$[CLS]$ $mention\ word$: $text\ context$ $[SEP]$" to integrate the word and text context of mention (the sentence where the mention is located) as mention's textual input, and make a template "$[CLS]$ $entity\ name$: $entity\ attributes$ $[SEP]$" to integrate the entity name and attributes as the entity's textual input. The CLIP text encoder gains the text word embedding $T_m^{(0)}$ and $T_e^{(0)}$ by tokenizing the textual input,

$$T_m^{(0)} = [tc_m^{(0)}, t_{m;1}^{(0)}, \cdots, t_{m;L_m}^{(0)}] \in \mathbb{R}^{(L_m+1)\times d_t} \tag{2}$$

Next, the word embedding processed through K transformer blocks, the i-th layer always accepts the output of the $(i-1)$-th layer,

$$T_m^{(i)} = TL_i(T_m^{(i-1)}) \quad i = 1, \ldots, K. \tag{3}$$

finally we obtain the text sequence outputs $T_m^{(K)}$ and $T_e^{(K)}$, and then a text projection converts the last layer cls token $tc^{(K)}$ to a d_z dimensional projection space to obtain $T_{m;z} = TProj(tc_m^{(K)}) \in \mathbb{R}^{d_z}$ and $T_{e;z} = TProj(tc_e^{(K)}) \in \mathbb{R}^{d_z}$.

3.2.2 Visual-only Encoding Module

In this module, we use pre-trained vision transformer (ViT) to extract visual features, for the input original RGB image $m_v, e_v \in \mathbb{R}^{H\times W\times 3}$, H and W denote the height and width of the image. We first split the image into $N_m = (H \times W)/P^2$ patches, where the size of each patch is $P \times P$, and then flatten each patch to get a N_m-dimensional patch embedding. Finally the patch embedding is attached with a cls token as the initial inputs,

$$V_m^{(0)} = [vc_m^{(0)}, v_{e;1}^{(0)}, \cdots, v_{m;N_m}^{(0)}] \in \mathbb{R}^{(N_m+1)\times d_v} \tag{4}$$

and the input vector passes through K vision transformer blocks in turn,

$$V_m^{(i)} = VL_i(V_m^{(i-1)}) \quad i = 1, \ldots, K. \tag{5}$$

then we gain the visiual outputs $V_m^{(K)}$, $V_e^{(K)}$ and use a visual projection to convert the cls token $vc^{(K)}$ to the d_z dimensional projection space to get the final representation, $V_{m;z} = VProj(vc_m^{(K)}) \in \mathbb{R}^{d_z}$ and $V_{e;z} = VProj(vc_e^{(K)}) \in \mathbb{R}^{d_z}$.

3.2.3 Prompt-based Multimodal Interaction Encoding Module

In this module, unlike the previous methods of extracting features separately with the text and visual encoders mentioned above and then performing simple concatenation to get the final multi-modal representation, inspired by (Khattak et al., 2023), we use a coupling function to build a bridge between the textual and visual prompts in order to interactive information from the two modalities in the process of transformer processing.

Text-side prompt learning: Based on the original text word embedding, we add two sets of learnable tokens $P_m^{(0)}$ and $P_e^{(0)}$ with length r as text prompts for mention and entity respectively,

$$P_m^{(0)} = [p_{m;1}^{(0)}, p_{m;2}^{(0)}, \cdots, p_{m;r}^{(0)}] \in \mathbb{R}^{r \times d_t} \quad (6)$$

new text inputs with prompt is generated, which can be denoted as $[P_m^{(0)}, T_m^{(0)}]$ and $[P_e^{(0)}, T_e^{(0)}]$.

In the transformer learning process, for the first J layers, new prompt tokens are imported to each layer, and for the (J+1) to K layers, no new prompt tokens will be imported while the prompts from previous layer will be processed instead,

$$\begin{aligned} [-, T_m^{(i)}] &= TL_i([\divideontimes {P_m}^{(i-1)}, T_m^{(i-1)}]) \\ & \quad i = 1, \ldots, J. \\ [P_m^{(i)}, T_m^{(i)}] &= TL_i([{P_m}^{(i-1)}, T_m^{(i-1)}]) \\ & \quad i = J+1, \ldots, K. \end{aligned} \quad (7)$$

then we get the output of the last layer $[P_m^{(K)}, T_m^{(K)}]$, $[P_e^{(K)}, T_e^{(K)}]$. Same as the text-only module, $T'_{m;z}$ and $T'_{e;z}$ are obtained from a text projection.

Visual-side prompt learning: To ensure a more profound interaction between two modalities, rather than learning the visual prompt independently, we project the text prompt token to the visual space through a coupling function $\mathcal{F}$, to obtain the visual prompt in the first J layers, and then remaining layers process the prompt from the previous layers like text-side,

$$\begin{aligned} [-, V_m^{(i)}] &= VL_i([\mathcal{F}(\divideontimes {P_m}^{(i-1)}), V_m^{(i-1)}]) \\ & \quad i = 1, \ldots, J. \\ [\hat{P_m}^{(i)}, V_m^{(i)}] &= VL_i([\hat{P_m}^{(i-1)}, V_m^{(i-1)}]) \\ & \quad i = J+1, \ldots, K. \end{aligned} \quad (8)$$

during this process, the two sides achieve full interaction through the mutual propagation of gradients. We eventually obtain the output representations $[\hat{P_m}^{(i)}, V_m^{(i)}]$, $[\hat{P_e}^{(i)}, V_e^{(i)}]$ for mention and entity and the visual projection mapped $V'_{m;z}$, $V'_{e;z}$.

3.2.4 Similarity Matching Score Calculation

Upon acquiring the three sets of features from the previous modules, we calculate the similarity matching scores between the mention and each candidate entity using the three feature interaction score units as proposed in (Luo et al., 2023), which has been proved effective in exploring both inter-modal and intra-modal similarity at different granularities. In our model, similarity matching scores for mention-entity pairs are computed separately for each of the three expert networks. For Text Expert and Visual Expert, we use the hidden features learned from the last transformer layer as local feature and the cls token transformed by a projection function as global feature, then fed them into the text scoring unit and visual scoring unit separately,

$$S^{text} = U_T(T_m, T_{m;z}, T_e, T_{e;z}) \quad (9)$$

$$S^{visual} = U_V(V_m, V_{m;z}, V_e, V_{e;z}) \quad (10)$$

here, we omit the state symbol (K). As for the Multi-modal Expert, we input the features obtained from Section 3.2.3 into all three units, calculate three scores, and take the average value as the final score according to the following formula.

$$S_T = U_T([P_m, T_m], T'_{m;z}, [P_e, T_e], T'_{e;z}) \quad (11)$$

$$S_V = U_V([\hat{P}_m, V_m], V'_{m;z}, [\hat{P}_e, V_e], V'_{e;z}) \quad (12)$$

$$\begin{aligned} S_C = U_C([[P_m, T_m], T'_{m;z}, [\hat{P}_m, V_m], V'_{m;z} \\ [P_e, T_e], T'_{e;z}, [\hat{P}_e, V_e], V'_{e;z}) \end{aligned} \quad (13)$$

$$S^{mm} = \frac{S_T + S_V + S_C}{3} \quad (14)$$

3.3 Modality Selection Gating Network

In this subsection, for input data with O modalities, samples of either mentions or entities can be represented as $x = (x_1, ..., x_O)$, inspired by the Mixture of Experts (MoE) (Masoudnia and Ebrahimpour, 2014), specialized expert networks can be designed for each subset of M modalities to accomplish the task. In our multi-modal entity linking (MEL), since there are two modalities, we construct three expert networks, $Expert(x_t)$, $Expert(x_v)$, and $Expert(x_t, x_v)$, which use the textual, visual, and multi-modal information of the entity-mention pairs respectively to solve the entity linking task. The number of expert networks in the selectable

list is set to B. To determine the selection of the appropriate expert network, we devise a Multimodal Selection Gating Network, which ultimately outputs a one-hot B-dimensional vector to guide the choice of the expert network.

Specifically, we concat the text token embedding and the 4096-dimensional visual feature extract by VGG to obtain $f_m = concat(f_{m_t}, f_{m_v})$ and $f_e = concat(f_{e_t}, f_{e_v})$ as the original input of the two identical MLPs. As is showen in Figure 2, each MLP network has two-layers with each layer consisting of an FC layer, a LeakyRelu layer, and a Dropout layer. After obtaining h_e and h_m by two MLPs, an expansion operation and a concatenation operation are performed before entering the final MLP, which consists of a FC layer and a RELU layer. This final MLP ultimately outputs a B-dimensional categorical vector $\pi \in \mathbb{R}^B$, a continuous distribution vector for modal selection. Since we want to generate a discrete decision, this selection process, however, is not directly microscopic. To solve this issue, we introduce the Categorical Reparameterization with Gumbel-Softmax technique (Jang et al., 2016) (Xue and Marculescu, 2023) for training. Specifically, we firstly apply the Gumbel-Max trick, which samples a series of interpolated noises, $g_1, ..., g_B$, from $Gumbel(0, 1)$ to enhance the robustness of the model, and with the following formula to draw samples $\mathbf{z}$ from a categorical distribution with class probabilities π,

$$\mathbf{z} = Onehot(\arg\max_i [g_i + \log \pi_i]) \quad (15)$$

$[z_1, \dots z_B]$ obtained by this formulation should be used for modal selection, however, because of the discrete non-fracturable problem, we use the softmax function as the continuous differentiable approximation of an $argmax$ during the training process and generate a B-dimensional continuous vector $\tilde{\mathbf{z}}$ which can be computed in back-propagation,

$$\tilde{z}_i = \frac{exp(\log(\pi_i) + g_i)/\tau}{\sum_{j=1}^{B} exp(\log(\pi_j) + g_j)/\tau} \quad i = 1, \dots, B \quad (16)$$

τ denotes the temperature of the softmax, when τ is closer to 0, Gumbel-softmax approximates to the discrete distribution, while as τ increases, Gumbel-softmax is closer to the uniform distribution. We use $\tilde{\mathbf{z}}$ in the back-propagation process of training and $\mathbf{z}$ in the forward-propagation process and inference process. Since $\bigtriangledown\tilde{\mathbf{z}} \approx \bigtriangledown\mathbf{z}$, we can not only achieve modal selection during inference, but also update the model during training with this method.

3.4 Optimization and Training Strategy

Due to the potential issue that early-stage training optimization of our modality selection gating network could lead to some expert networks being less likely chosen, we adopt a two-stage training strategy. This strategy, along with the test stage, will be detailed as follows.

Expert Network Training Stage: In this stage, B expert networks are trained independently using the dataset to make them reach their optimal performance. After obtaining the similarity scores, a contrastive training loss function is applied to each batch, which aims to bring positive mention-entity pairs closer and distance negative pairs,

$$\mathcal{L}(S(\cdot)) = -\log \frac{exp(S(m, e))}{\sum_i exp(S(m, \bar{e}_i))} \quad (17)$$

$S(\cdot)$ indicates the used score function, e and $\bar{e}$ respectively denotes the positive and negative entities in current batch.

Gating Selector Learning and Expert Network Fine-Tuning Stage: In this stage, we freeze the feature encoding module with relatively high parameter cost, which have been trained in the first stage, only fine-tune the similarity computation module. In addition, the selection gating network is integrated into the optimization process, for each mention-entity pair, the one-hot selection vector $\mathbf{z}$ obtained from the gating network is used to determine the score function to be utilized,

$$S = [z_1, \dots, z_B] \cdot [S_1, \dots, S_B]^T \quad (18)$$

S_i indicates the score function of Expert Network i, then we apply Equation 17 to derive final loss $\mathcal{L}$.

Testing Stage: During the testing stage, we use $\mathbf{z}$ to select one of the branches to compute the final score S. The entity with the highest score is then chosen as the final selection for the mention.

4 Experiment

4.1 Experimental Setup

Datasets: We validate the effectiveness of our proposed method on three datasets: RichpediaMEL and WikiMEL which were proposed by Wang et al. (2022a), and WikiDiverse which is proposed by Wang et al. (2022b) The statistical data, and the division methods for the three datasets are presented in Appendix A.1.

Baselines: We compare our method against two categories of models: (1) Entity Linking based

Method	Dataset	RichpediaMEL				WikiMEL				WikiDiverse			
	Metric	H@1↑	H@3↑	H@5↑	MR↓	H@1↑	H@3↑	H@5↑	MR↓	H@1↑	H@3↑	H@5↑	MR↓
Pre(T)	BERT (Devlin et al., 2019)	59.55	81.12	87.16	278.08	74.82	86.79	90.47	51.23	55.77	75.73	83.11	373.96
Pre(V+T)	CLIP (Radford et al., 2021)	67.78	85.22	90.04	107.16	83.23	92.1	94.51	17.6	61.21	79.63	85.18	313.35
Pre(V+T)	ALBEF (Li et al., 2021)	65.17	82.84	88.28	122.3	78.64	88.93	91.75	47.95	60.59	75.59	81.3	291.17
EL(T)	BLINK (Wu et al., 2020)	58.47	81.51	88.09	178.57	74.66	86.63	90.57	51.48	57.14	78.04	85.32	332.03
MEL(V+T)	DZMNED (Moon et al., 2018)	68.16	82.94	87.33	313.85	78.82	90.02	92.62	152.58	56.9	75.34	81.41	563.26
MEL(V+T)	JMEL (Adjali et al., 2020)	48.82	66.77	73.99	470.9	64.65	79.99	84.34	285.14	37.38	54.23	61	996.63
MEL(V+T)	GHMFC (Wang et al., 2022a)	72.92	86.85	90.6	214.64	76.55	88.4	92.01	54.75	60.27	79.4	84.74	628.87
MEL(V+T)	MIMIC (Luo et al., 2023)	81.02	91.77	94.38	55.11	87.98	95.07	96.37	11.02	**63.51**	**81.04**	86.43	227.08
ours*	DSMIP	**85.03**	**92.89**	**94.97**	**21.24**	**88.97**	**95.7**	**96.86**	**10.37**	62.46	80.6	**86.92**	**222.9**

Table 1: Performance comparison of different methods on three MEL datasets.(T: using textual modality, V: using visual modality, Pre: pre-training model, EL: entity linking method, MEL: multi-modal entity liking method).

Methods: This includes EL methods that utilize solely text information, such as BLINK, and MEL methods DZMNED, JMEL, GHMFC, and MIMIC. (2) Pre-trained Model based Methods: This encompasses models that are purely text-based pretrained, such as BERT, as well as V-L pretrained models including CLIP and ALBEF. A detailed description of the baselines is provided in Appendix A.2.

Metrics: We evaluate above methods using H@k and MR. The specific computation for each metric is provided in Appendix A.3. H@k denotes the hit rate of the ground-truth entity within the top-k ranked entities. MR represents the average rank of the ground-truth entity across all entities.

Implementation Details: Our proposed method is implemented using the PyTorch framework and trained on NVIDIA GeForce GTX4090 GPU, setting up Ubuntu 16.04 operating system and has an Intel(R) Xeon(R) Gold 6226R CPU. We find the optimal hyperparameter through grid search. The maximum length for text inputs is set to 77. The dimensions for the text and visual hidden layers, as well as the projection space, are set to 512, 768, and 512, respectively. The prompt length is set to 2, with a prompt depth update of 9. The gated network's hidden layer dimension is set to 512, with a softmax temperature of 0.1. Training is conducted using the Adam optimizer with a fixed number of epochs set at 20, a batch size of 128, and a learning rate of $5e^{-4}$. We deploy three expert networks with $B = 3$ on WikiMEL. Due to the underperformance of the visual expert network, we utilize only the text and multimodal expert networks on RichpediaMEL and WikiDiverse with $B = 2$.

4.2 Main Experimental Results

Our comparative experiment is conducted to evaluate the effectiveness of our model. Table 1 presents the results of all models in terms of H@1, H@3, H@5, and MR metrics, where a higher value signifies better performance for all metrics except for MR.

Firstly, the pre-trained model based methods outperform most specialized MEL approaches, and those employing large-scale pretraining models for feature extraction, such as GHMFC and MIMIC, also surpass the others, fully demonstrating the efficacy of large-scale pretrained models in MEL tasks. Notably, CLIP performs exceptionally well, which substantiates our selection of CLIP as the original pretrained encoder for feature extraction. Surprisingly, text-based approaches achieve favorable results on these datasets, surpassing several MEL methods across numerous metrics, inadvertently reflects the fact that incorporating visual information can sometimes contribute noise and in some cases even negatively impact the performance. Ultimately, the experimental results confirm that our DSMIP achieves state-of-the-art performance. On RichpediaMEL and WikiMEL, DSMIP surpasses all baseline methods in every metric, with the most notable gains in Hit@1—the most demanding metric. While our method lags slightly behind MIMIC on some metrics on WikiDiverse, it achieves better performance in low-resource settings, which are more reflective of real-world scenarios, as will be discussed in Section 4.3.

4.3 Low Resource Experiment

In this subsection, we evaluate our model in low-resource settings, which is crucial for real-world application scenarios where data labeling is resource-intensive. We maintain the validation and test sets unchanged and use only 10% of training data. Figure 3 and 4 illustrate the performance of multimodal methods in these settings and their relative decline from high-resource setting, respectively.

Most methods show a significant decline in per-

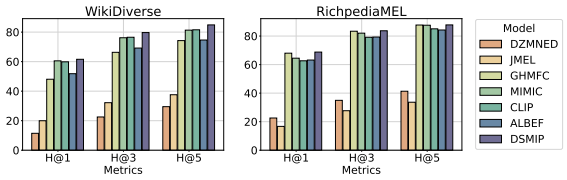

Figure 3: Low-Resource Performance Comparison.

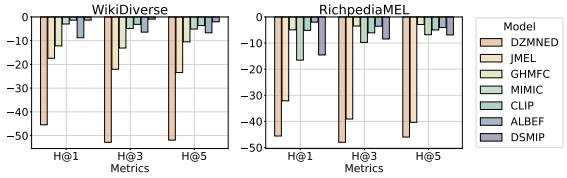

Figure 4: Performance Decline in Low vs. High Resource Settings.

formance with fewer training resources. However, large-scale pre-trained V-L models like CLIP and ALBEF show less degradation, especially on RichpediaMEL, proving that abundant knowledge and language comprehension can be gained from large-scale pretraining, where even a limited amount of target task data can effectively guide model fine-tuning for MEL. However, their overall efficacy is still limited. Notably, our approach slightly trails MIMIC on WikiDiverse in high-resource setting, mainly due to its effective inter-modal and intra-modal interaction. However, we surpasses it in low-resource settings and achieves the least performance decline on this dataset. In summary, our method outperforms all baseline methods and achieves a small performance decline, fully demonstrating DPMIS's adaptability and superiority in low-resource scenarios.

4.4 Ablation Experiments

To investigate the effectiveness of our proposed Prompt-based Multimodal Interaction Encoding Module and the impact of the three expert networks, we conducted ablation studies on the WikiMEL, with the results shown in Table 2.

Firstly, it was observed that all single expert networks exhibited a performance gap compared to the complete model, validating our modality selection network. The multimodal network saw the least drop, with a 3% decrease in Hit@1. For the single modality expert networks, the Text Expert Network did not show much decline, but the Visual Expert Network underperformed significantly, indicating the potential for visual noise to misguide the model. Further experiment with a non-interactive CLIP model, in place of interactive prompts, led to a notable drop in performance, affirming the utility of prompts and the coupling function for modality integration, thus justifying our adoption of prompt-based learning.

Model	H@1↑	H@3↑	H@5↑
DSMIP	88.97	95.7	96.86
w/ $Expert_{text}$	65.31	79.35	84.13
w/ $Expert_{image}$	31.51	37.57	40.64
w/ $Expert_{multimodal}$	86.26	94	95.85
w/o $Interactive\ Prompt$	78.6	88.93	91.97

Table 2: Ablation studies on WikiMEL.

	Trainable params (M)
w/ CLIP encoder	153
w/ Interacticve prompt encoder	7.6

Table 3: Trainable parameters Comparison: Multimodal Expert Network across two encoder configurations.

4.5 Computational Cost Analysis

To validate the computational savings of our prompt interactive expert network, we evaluated the number of trainable parameters of the multimodal expert network in two manners: (a) fine-tuning the CLIP pre-trained encoder and (b) using the interactive prompt learning encoder proposed in this study.

As shown in Table 3, it is evident that our prompt learning encoder reduces trainable parameters from 153M to 7.6M. This reduction confirms the module's effectiveness in facilitating modal interactions with substantial computational savings.

5 Conclusion

In this paper, we propose a novel DSMIP for Multimodal Entity Linking. Our method designs three expert networks for different subsets of modalities, and devises a gating network that generates a one-hot vector to select the optimal Expert Network for each modality-entity pair, thereby overcoming the mention ambiguity and modality-induced noise. Moreover, we update multimodal prompts to extract features of mentions and entities within the multimodal expert networks and establish a coupling function to enable interaction between the modalities, thereby reducing modality heterogeneity and saving the training overhead. Experimental results demonstrate that our model outperforms other state-of-the-art methods.

6 Limitations

The limitations of our method are as follows:

- Our proposed modality selection network uses the same expert network for both the mention and entity within a sample pair. While this approach can choose the most suitable modality for the current sample pair, it does not consider cross-modality interactions between the mention and entity during the selection process, potentially leading to the loss of some cross-modal information. Thus, designing a cross-modal interaction expert network for mention-entity pairs, allowing the different choice, is an important direction for our improvement.

- Our method compute contrastive training loss by comparing mentions with all entities in the knowledge base, a process that can be time-consuming. However, previous studies employing a pre-selection of candidate entities before ranking have been less time-intensive but less effective. Balancing the two and finding a more efficient method for candidate entity selection remains an area for further exploration.

References

Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020. Multimodal entity linking for tweets. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I*, page 463–478, Berlin, Heidelberg. Springer-Verlag.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jingru Gan, Jinchang Luo, Haiwei Wang, Shuhui Wang, Wei He, and Qingming Huang. 2021. Multimodal entity linking: A new dataset and a baseline. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 993–1001, New York, NY, USA. Association for Computing Machinery.

Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. 2022. Dynamic neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 7436–7456.

Faegheh Hasibi, Krisztian Balog, and Svein Erik Bratsberg. 2016. Exploiting entity linking in queries for entity retrieval. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, page 209–218, New York, NY, USA. Association for Computing Machinery.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *International Conference on Learning Representations,International Conference on Learning Representations*.

Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. 2022. Prompting visual-language models fornbsp;efficient video understanding. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, page 105–124, Berlin, Heidelberg. Springer-Verlag.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19113–19122.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705. Curran Associates, Inc.

Pengfei Luo, Tong Xu, Shiwei Wu, Chen Zhu, Linli Xu, and Enhong Chen. 2023. Multi-grained multimodal interaction network for entity linking. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 1583–1594, New York, NY, USA. Association for Computing Machinery.

Saeed Masoudnia and Reza Ebrahimpour. 2014. Mixture of experts: a literature survey. *Artif. Intell. Rev.*, 42(2):275–293.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity disambiguation for noisy social media posts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2000–2008, Melbourne, Australia. Association for Computational Linguistics.

Rameswar Panda, Chun-Fu Richard Chen, Quanfu Fan, Ximeng Sun, Kate Saenko, Aude Oliva, and Rogerio Feris. 2021. Adamml: Adaptive multi-modal learning for efficient video recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*.

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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Peng Wang, Jiangheng Wu, and Xiaohang Chen. 2022a. Multimodal entity linking with gated hierarchical fusion and contrastive training. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 938–948, New York, NY, USA. Association for Computing Machinery.

Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. 2022b. WikiDiverse: A multimodal entity linking dataset with diversified contextual topics and entity types. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4785–4797, Dublin, Ireland. Association for Computational Linguistics.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Shangyu Xing, Fei Zhao, Zhen Wu, Chunhui Li, Jianbing Zhang, and Xinyu Dai. 2023. Drin: Dynamic relation interactive network for multimodal entity linking. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 3599–3608, New York, NY, USA. Association for Computing Machinery.

Z. Xue and R. Marculescu. 2023. Dynamic multimodal fusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2575–2584, Los Alamitos, CA, USA. IEEE Computer Society.

Chengmei Yang, Bowei He, Yimeng Wu, Chao Xing, Lianghua He, and Chen Ma. 2023. MMEL: A joint learning framework for multi-mention entity linking. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 2411–2421. PMLR.

Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. 2015. Semantic parsing via staged query graph generation: Question answering with knowledge base. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1321–1331, Beijing, China. Association for Computational Linguistics.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, page 2337–2348.

Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. 2024. Multi-modal knowledge graph construction and application: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(2):715–735.

A Appendix

A.1 Datasets Details

The introductions for the three datasets are as follows:

(1) RichpediaMEL (Wang et al., 2022a): It possessing more than 17,000 multimodal samples. The entities were gathered from Richpedia and corresponding multimodal information retrieved from Wikipedia.

(2) WikiMEL (Wang et al., 2022a): This dataset comprises over 22,000 multimodal samples, constructed by collecting entities from Wikidata and subsequently extracting textual and visual descriptions for each entity from Wikipedia.

(3) WikiDiverse (Wang et al., 2022b): It contains over 8,000 diverse context topics and entity types sourced from Wikinews, using Wikipedia, which hosts more than 16 million entities, as the corresponding knowledge base.

Table 4 presents detailed data statistics for the three datasets, and Table 5 shows the partitioning of the datasets. Here, we utilize the original splits for the three datasets. For RichpediaMEL and WikiMEL, we allocate the (train, valid, test) split as (70%, 10%, 20%), while for WikiDiverse, the division is set at (80%, 10%, 10%).

Dataset	Mentions	Samples	Img of Mentions	Entites	Img of Entities
RichpediaMEL	17805	17724	15853	160935	86769
WikiMEL	25846	22070	22136	109976	67195
WikiDiverse	15093	7405	6697	132460	67309

Table 4: Summary statistics for the datasets.

Dataset	Mentions	Entity
RichpediaMEL-train	12463	
RichpediaMEL-valid	1780	160935
RichpediaMEL-test	3562	
WikiMEL-train	18092	
WikiMEL-valid	2585	109976
WikiMEL-test	2078	
WikiDiverse-train	11351	
WikiDiverse-valid	1664	132460
WikiDiverse-test	2078	

Table 5: Data Division for Three Datasets.

A.2 Baselines Details

The baseline methods employed in the experimental section are described as follows:

(1) BERT (Devlin et al., 2019): A foundational model in natural language processing that employs a series of Transformer encoders. It leverages extensive pre-training over a large textual corpus and demonstrates proficiency across diverse language understanding tasks.

(2) CLIP (Radford et al., 2021): A large pre-trained approach that learns visual concepts from textual annotations and trained on a wide variety of image-text pairs. It is proficient in recognizing and classifying images based on textual descriptions, even with limited training examples.

(3) ALBEF (Li et al., 2021): An advanced pre-trained visual-language technique that align the visual and textual modalities by contrastive loss before fuse them via a multimodal Transformer encoder. It also incorporates momentum distillation to enhance robustness against noisy datasets.

(4) BLINK (Wu et al., 2020): A two-stage zero-shot linking algorithm where entities are defined solely by brief textual descriptions. The first stage involves retrieval candidates in a dense space defined by a bi-encoder and the second stage re-ranked them using a cross-encoder.

(5) DZMNED (Moon et al., 2018): A deep zero-shot multimodal network which is the first MEL method. It extracts context from text and images and predicts the correct entity in a knowledge graph embedding space, enabling zero-shot disambiguation of entities not seen in the training set.

(6) JMEL (Adjali et al., 2020): A MEL method designed for tweets that trains a dual-branch feed forward neural network to minimize a triplet loss that defines an implicit joint feature space, and projects each modality into this space via its respective branches.

(7) GHMFC (Wang et al., 2022a): A method that incorporates a transformer to explore fine-grained cross-modal relationships for MEL task, learning meaningful multimodal mention representations through gated fusion and contrastive training.

(8) MIMIC (Luo et al., 2023): A advanced approach that proposes a multi-granularity multimodal interaction network, establishing three interaction units to comprehensively explore intra-modal interactions and inter-modal fusion between entity and mention features.

A.3 Metrics Details

The calculation formula of the metrics is defined as follows:

(1) H@k

$$H@k = \frac{1}{N}\sum_{i}^{N} I(rank(i) < k) \tag{19}$$

where N denotes the total number of samples, and I is an indicator function. I is set to 1 when the acceptance condition is met, and 0 otherwise.

(2) MR

$$MR = \frac{1}{N}\sum_{i}^{N} rank(i) \tag{20}$$