VP-MEL: Visual Prompts Guided Multimodal Entity Linking

Anonymous ACL submission

Abstract

Multimodal entity linking (MEL), a task aimed at linking mentions within multimodal contexts to their corresponding entities in a knowledge base (KB), has attracted much attention due to its wide applications in recent years. However, existing MEL methods often rely heavily on mention words as retrieval cues, which limits their ability to effectively utilize information from both images and text. This reliance poses significant challenges in scenarios where mention words are absent, as current MEL approaches struggle to leverage image-text pairs for accurate entity linking. To solve these issues, we introduce a **V**isual **P**rompts guided **M**ultimodal **E**ntity **L**inking (VP-MEL) task. Given a text-image pair, VP-MEL aims to link a marked region (*i.e.*, visual prompt) in an image to its corresponding entities in the knowledge base. To facilitate this task, we present a new dataset, VPWiki, specifically designed for VP-MEL. Furthermore, we propose a framework named FBMEL, which enhances visual feature extraction using visual prompts and leverages the pretrained Detective-VLM model to capture latent information. Experimental results on the VPWiki dataset demonstrate that FBMEL outperforms baseline methods across multiple benchmarks for the VP-MEL task. Code and datasets will be released at `https://anonymous.4open.science/r/VP-MEL-26E2`.

1 Introduction

Linking ambiguous mentions with multimodal contexts to the referent unambiguous entities in a knowledge base (KB), known as Multimodal Entity Linking (MEL) (Moon et al., 2018), is an essential task for many multimodal applications. Most MEL works (Gan et al., 2021; Wang et al., 2022a; Dongjie and Huang, 2022; Luo et al., 2023; Zhang et al., 2023a; Xing et al., 2023; Shi et al., 2024) mainly focus on improving the interaction of multimodal information and achieve promising performance. However, existing methods typically represent mentions in the form of mention words and assume that each mention is associated with a high-quality image, which results in two limitations for MEL.

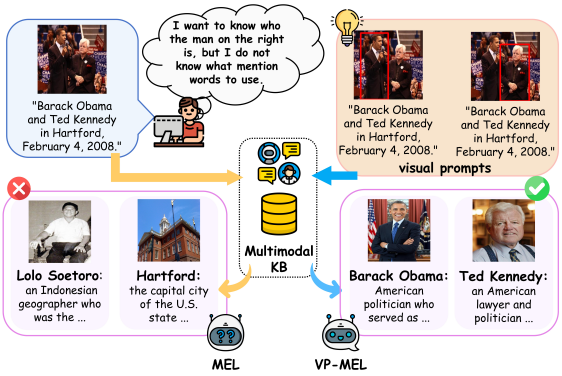


Figure 1: Comparison between MEL and VP-MEL tasks. MEL struggles with entity linking for image content when mention words are unavailable. In contrast, VP-MEL addresses this limitation by using visual prompts to link specific regions in the images to the correct entities in the knowledge base.

Text information dependency: MEL heavily relies on mention words for entity linking, which often exhibit substantial overlap with entity names in practical applications. This inherent similarity enables mention words to serve as strong cues for identifying specific entities within the knowledge base (KB). Although some ambiguity remains, this overlap substantially simplifies the entity linking process. However, MEL struggles to perform effectively and often produces incorrect outputs when mention words are unavailable. As shown in Figure 1, the expected mention is *Ted Kennedy* (on the right), but MEL generates incorrect responses, like *Lolo Soetoro* and *Hartford*. Many results, such as *Lolo Soetoro*, are entirely unrelated to both the image and text. This excessive reliance on mention words confines MEL to a text-focused

task. It is necessary to introduce a new MEL task to overcome such constraints, allowing MEL to broaden its application scenarios and adapt to the multimodal era.

Image modality impurity: Compared to the text modality, the image modality often contains more noise. As illustrated in Figure 1, only the person on the right is relevant, while the one on the left represents noise. Misinterpreting or misusing such image information can seriously affect the results. Most existing coarse-grained methods (Yang et al., 2023; Luo et al., 2023; Wang et al., 2022a) encode the entire image directly, making it difficult to eliminate noise interference. Song et al. (2024) attempts to enhance visual features by utilizing fine-grained image attributes, but they heavily depends on high-quality images and often increase model size and training costs. To better leverage image information, it is necessary for MEL to find a way to overcome noise interference in images and focus on the regions related to the mentions.

Existing methods demonstrate limited effectiveness in utilizing image information and are overly dependent on mention words as textual cues. In many application scenarios, the challenge of MEL is considerably worsened by the inability of users to provide relevant mention words. So we ask: *Is it possible to introduce a MEL task that does not depend on mention words, thereby avoiding excessive reliance on text and enhancing the performance of image-based reasoning?*

In this paper, we define a new **V**isual **P**rompts guided **M**ultimodal **E**ntity **L**inking (VP-MEL) task. As shown in Figure 1, VP-MEL is designed to address the entity linking task for image-text pairs. VP-MEL employs a straightforward method to annotate mentions directly onto images using visual prompts, eliminating the need for mention words or any additional complex operations. Free from the constraints of mention words, VP-MEL supports a broader range of applications. It enables effective entity linking even in situations where users are unfamiliar with the textual language or place greater emphasis on image-based information. To support this new task, we construct a VPWiki dataset based on the existing MEL public datasets, where visual prompts are annotated for each mention within the associated images.

We then propose a **F**eature **B**alanced **M**ultimodal **E**ntity **L**inking (FBMEL) framework. Extensive research (Cai et al., 2024; Shtedritski et al., 2023) demonstrates that pre-trained CLIP visual encoders can effectively interpret visual markers. In FBMEL, the CLIP visual encoder is employed to extract both global image features and local features guided by visual prompts. Additionally, a Vision-Language Model (VLM) equipped with CLIP visual encoder is pre-trained to generate useful text information from visual prompts, serving as an additional supplement to the text information. FBMEL enables visual prompts to deliver both supplementary visual and textual information, ensuring a balanced utilization of image and text while avoiding excessive reliance on a single modality.

Main contributions are summarized as follows:

(i) We introduce VP-MEL, a new entity linking task that replaces traditional mention words with visual prompts to annotate mentions directly in the images.

(ii) We construct a high-quality annotated dataset, VPWiki, to enhance the benchmark evaluation of the task. An automated annotation pipeline is proposed to enhance the generation efficiency of the VPWiki dataset.

(iii) We propose the FBMEL framework to address VP-MEL task by effectively leveraging multimodal information and reducing reliance on a single modality. Compared to prior methods, FBMEL achieves a 20% performance improvement in the VP-MEL task and maintains competitive results in the MEL task.

2 Related Work

2.1 Multimodal Entity Linking

Given the widespread use of image-text content in social media, the integration of both modalities for entity linking is essential. For example, Moon et al. (2018) pioneer the use of images to aid entity linking. Building on this, Adjali et al. (2020) and Gan et al. (2021) construct MEL datasets from Twitter and long movie reviews. Expanding the scope of MEL datasets, Wang et al. (2022c) present a high-quality MEL dataset from Wikinews, featuring diversified contextual topics and entity types. To achieve better performance on these datasets, a multitude of outstanding works in the MEL field (Wang et al., 2022a; Yang et al., 2023; Luo et al., 2023; Shi et al., 2024; Song et al., 2024) emerge, focusing on extracting and interacting with multimodal information. Although multimodal information can enhance entity linking performance, in these methods, text consistently dominates over images.

2.2 Vision Prompt

Region-specific comprehension in complex visual scenes has become a key research topic in the field of Multimodal Computer Vision. Existing methods typically utilize textual coordinate representations (Zhu et al., 2024; Zhao et al., 2023), learned positional embeddings (Peng et al., 2024; Zhang et al., 2023b; Zhou et al., 2023), or Region of Interest (ROI) features (Zhang et al., 2023b) to anchor language to specific image regions. More recently, Cai et al. (2024) propose a coarse-grained visual prompting solution that directly overlays visual prompts onto the image canvas. In contrast, our VP-MEL provides a fine-grained entity linking method based on visual prompts to reduce reliance on text.

3 Dataset

As there is no existing MEL dataset that incorporates visual prompts, annotating a high-quality dataset is crucial to enhance the benchmarking of the VP-MEL task.

Data Collection. Our dataset is built on two benchmark MEL datasets, *i.e.*, WikiDiverse (Wang et al., 2022c) and WikiMEL (Wang et al., 2022a). Appendix A.7 provides detailed information.

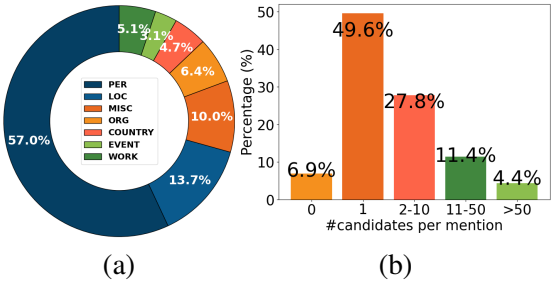
Annotation Design. Given an image-text pair and the corresponding mention words, annotators need to: 1) annotate relevant visual prompts in the image based on the mention words; 2) remove samples where the image and mention words are unrelated; 3) re-annotate samples with inaccurate pipeline annotations; 4) annotate the entity type for each example (*i.e.*, Person, Organization, Location, Country, Event, Works, Misc).

Annotation Procedure. To improve data annotation efficiency, we develop a pipeline that automatically annotates visual prompts in images based on mention words inspired by Li et al. (2024). In the pipeline, the Visual Entailment Module is employed to evaluate and filter out the highly relevant data. Subsequently, the Visual Grounding Module annotates the visual prompts in the images. The details of modules in the pipeline are provided in the Appendix A.6. The annotation team consists of 10 annotators and 2 experienced experts. All annotators have linguistic knowledge and are instructed with detailed annotation principles. Fleiss Kappa score (Fleiss, 1971) of annotators is 0.83, indicating strong agreement among them. We use the Intersection over Union (IoU) metric to assess annotation quality and discard samples with an IoU score below 0.5.

	Train	Dev.	Test	Total
pairs	8,000	1,035	1,052	10,087
ment. per pair	1.18	1.16	1.27	1.19
words per pair	9.89	9.80	10.32	9.92

Table 1: Statistics of VPWiki. ment. denotes Mentions.

Figure 2: An example from VPWiki and WikiDiverse. The red box in the left image represents the visual prompt annotated for the VP-MEL task. The red text in the right image shows the annotated mention words.



(a) (b)

Figure 3: More statistics of VPWiki. (a) Distribution of entity types. (b) Distribution of the number of candidate entities per mention.

Dataset Analysis. Figure 2 illustrates an example from the VPWiki dataset and highlights its differences from traditional MEL data. The VPWiki dataset comprises a total of 12,720 samples, which are randomly split into training, validation, and test sets with an 8:1:1 ratio. Detailed statistics of the VPWiki dataset are provided in Table 1. Additionally, Figure 3 presents the distribution of entity types and the number of candidate entities per mention in the dataset. In Figure 3(a), abbreviations are used to represent each entity type. Meanwhile, Figure 3(b) shows that as the number of candidate entities per mention increases, the task becomes increasingly challenging.

4 Task Formulation

The multimodal knowledge base is constructed by a set of entities $\mathcal{E} = \{E_i\}_{i=1}^{N}$, where each entity is denoted as $E_i = (e_{n_i}, e_{v_i}, e_{d_i}, e_{a_i})$. Here, e_{n_i} denotes the entity name, e_{v_i} represents the entity images, e_{d_i} corresponds to the entity description, and e_{a_i} captures the entity attributes. A mention is denoted as $M_j = (m_{s_j}, m_{v_j})$, where m_{s_j} represents the sentence and m_{v_j} denotes the corresponding image. The related entity of mention M_j in the knowledge base is E_i. The objective of the VP-MEL task is to retrieve the ground truth entity E_i from the entity set $\mathcal{E}$ in the knowledge base, based on M_j.

5 Methodology

In this section, we describe the proposed FBMEL framework for the VP-MEL task. As illustrated in Figure 4, FBMEL utilizes visual encoder to extract both deep semantic features and shallow texture features, which are enhanced by visual prompts (§5.1). To avoid excessive reliance on visual features, the Detective-VLM module is designed to generate supplementary textual information guided by visual prompts(§5.2), which is then combined with the original text and processed by the text encoder (§5.3). Finally, a similarity score is computed after integrating the visual and textual features(§5.4).

5.1 Visual Encoder

To extract image features, we choose pre-trained CLIP model(Dosovitskiy et al., 2021) as our visual encoder. The image m_{v_j} of M_j is reshaped into n 2D patches. After this, image patches are processed through visual encoder to extract features. The hidden states extracted from m_{v_j} by the CLIP visual encoder are represented as $V^l_{M_j} = \left[v^0_{[CLS]}; v^1_{M_j}; v^2_{M_j}; ...; v^n_{M_j}\right] \in \mathbb{R}^{(n+1)\times d_c}$, where d_c denotes the dimension of the hidden state and l denotes the number of layers in the encoder.

Since CLIP focuses on aligning deep features between images and text and may overlook some low-level visual details (Zhou et al., 2022), we selectively extract features from both the deep and shallow layers of CLIP. Specifically, a shallow feature ($V^3_{M_j}$) is used to represent the textures and geometric shapes in the image, while deep features ($V^{10}_{M_j}$, $V^{11}_{M_j}$, $V^{12}_{M_j}$) are used to represent abstract semantic information. We take the hidden states corresponding to the special [CLS] token ($v^0_{[CLS]} \in \mathbb{R}^{d_c}$) from these layers as the respective visual features F^l. These features are concatenated and normalized using LayerNorm, and then passed through a MLP layer to transform the dimensions to d_v, with the output representing the global features of the image $V^G_{M_j} \in \mathbb{R}^{d_v}$.

$$\mathrm{F}^{\mathrm{l}} = v^0_{[CLS]} \in V^l_{M_j},$$

$$V^{G\,\prime}_{M_j} = \mathrm{LN}\left(\mathrm{Concat}(F^3, F^{10}, F^{11}, F^{12})\right),$$

$$V^G_{M_j} = \mathrm{MLP}\left(V^{G\,\prime}_{M_j}\right).$$

Then, hidden states from the output layer of encoder $V^l_{M_j}$ are passed through a fully connected layer, which also transforms the dimensions to d_v, yielding the local features of the image $V^L_{M_j} \in \mathbb{R}^{(n+1)\times d_v}$:

$$V^L_{M_j} = \mathrm{FC}\left(V^l_{M_j}\right).$$

For the image e_{v_i} of entity E_i, the global feature $V^G_{E_i}$ and local feature $V^L_{E_i}$ are obtained using the same method described above.

5.2 Detective-VLM

To prevent the model from becoming overly dependent on visual information, it is necessary to extract additional textual information to balance contributions of both modalities.

Most VLMs (Liu et al., 2024; Zhu et al., 2024; Ye et al., 2023; Li et al., 2023) choose to use the CLIP visual encoder, which means they have a better ability to focus on the markers in images compared to other visual methods (Cai et al., 2024; Shtedritski et al., 2023). Therefore, we instruction fine-tune a VLM, designed to extract effective information from images. The VLM follows template designed below to further mine potential information from image m_{v_j} and sentence m_{s_j} of mention M_j, assisting in subsequent feature extraction:

Background: $\{Image\}$
Text: $\{Sentence\}$
Question: In the red box of the image, tell me briefly what is the $\{Entity\ Type\}$, and which $\{Entity\ Name\}$ in the text corresponds to the $\{Entity\ Type\}$?
Answer: $\{Entity\ Name\}$ $\{Entity\ Type\}$

We utilize VPWiki dataset to design the fine-tuning dataset, where $\{Image\}$ and $\{Sentence\}$ correspond to m_{v_j} and m_{s_j} in M_j, respectively. During the inference process, $\{Entity\ Name\}$ and $\{Entity\ Type\}$ need to be generated by VLM. Details of the dataset and Detective-VLM can be found in Appendix A.4.

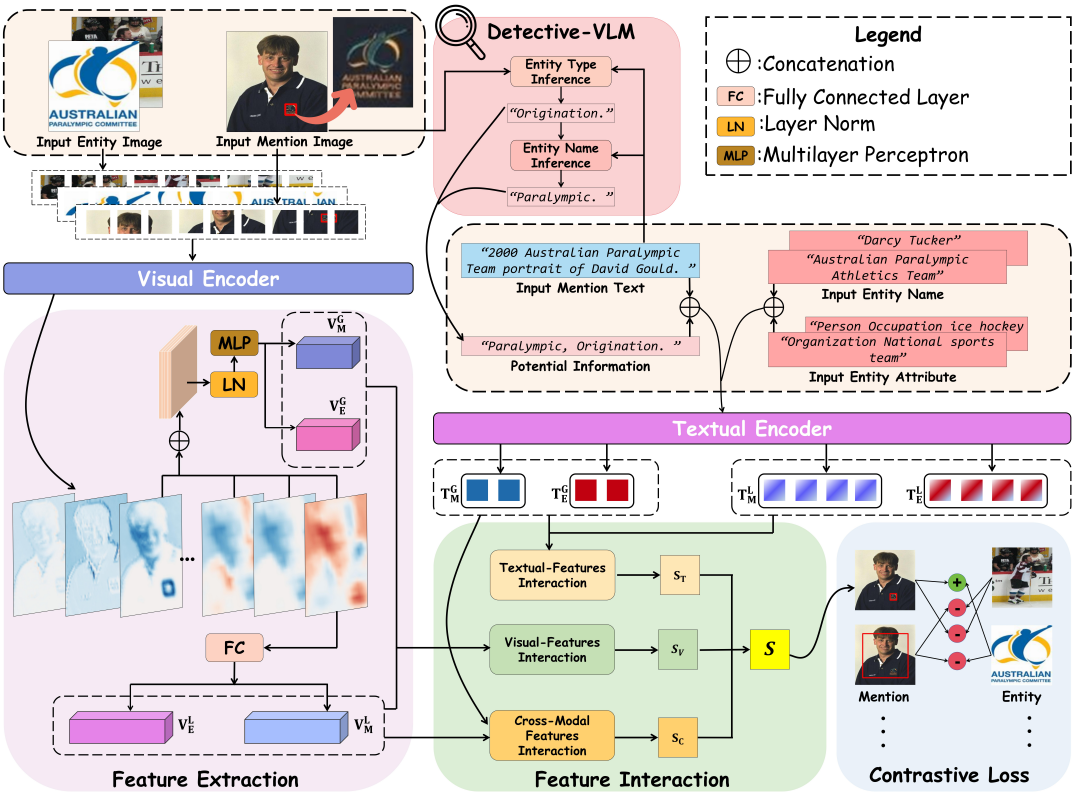


Figure 4: The overall architecture of Feature Balanced Multimodal Entity Linking (FBMEL) framework. The image-text pairs of the Mention and Entity are used together as input. Specifically, Mention Text is the sentence corresponding to Mention Image, while Entity Text consists of Entity Name and Entity Attribute corresponding to the Entity Image in the Knowledge Base.

The objective formula for instruction fine-tuning Detective-VLM is expressed as follows:

$$\min_{\theta} \sum_{i=1}^{N} \mathcal{L}(f_\theta(x_i), y_i),$$

where f represents the pre-trained VLM, and θ denotes the model parameters. N represents the number of instruction-output pairs, x_i is the i-th instruction, and y_i is the corresponding desired output. $\mathcal{L}$ is defined as:

$$\mathcal{L}(f_\theta(x_i), y_i) = -\sum_{t=1}^{T} \log P_\theta(y_i^{(t)} | x_i),$$

where T is the length of output sequence, $y_i(t)$ is the t-th word of the expected output y_i at time step t, and P is conditional probability that the model generates the output $y_i(t)$ at time step t.

Detective-VLM aims to ensure that the output is both accurate and relevant, minimizing the likelihood of generating irrelevant information. Notably, we represent the **Answer** output by VLM as m_{w_j}.

5.3 Textual Encoder

For the mention M_j, after concatenating mention sentence m_{s_j} with m_{w_j}, they form the input sequence, with different parts separated by [CLS] and [SEP] tokens:

$$I_{M_j} = [CLS]\, m_{w_j}\, [SEP]\, m_{s_j}\, [SEP]\,.$$

Hidden states of output layer after the input sequence passes through text encoder are represented as $T_{M_j} = \left[t^0_{[CLS]}; t^1_{M_j}; ...; t^{l_t}_{M_j}\right] \in \mathbb{R}^{(l_t+1)\times d_t}$, where d_t represents the dimension of output layer features, and l_t denotes the length of input. We use the hidden state corresponding to [CLS] as the global feature of the text $T^G_{M_j} \in \mathbb{R}^{d_t}$, and the entire hidden states as the local features of the text $T^L_{M_j} \in \mathbb{R}^{(l_t+1)\times d_t}$.

The input sequence for entity E_i consists of the entity name e_{n_i} and entity attributes e_{a_i}, which can be represented as:

$$I_{E_i} = [CLS]\, e_{n_i}\, [SEP]\, e_{a_i}\, [SEP]\,.$$

Then, using the above method, we obtain the text features $T^G_{E_i}$ and $T^G_{E_i}$ for the entity.

5.4 Multimodal Feature Interaction

Inspired by the multi-grained multimodal interaction approach (Luo et al., 2023), we build the feature interaction part. The multimodal feature interaction section consists of three different units. Notably, this section focuses only on introducing the functions of each unit, detailed mathematical derivations are provided in Appendix A.5.

Visual-Features Interaction (VFI). Image features of the mention M_j and the entity E_i interact separately. For feature interaction from M_j to E_i, after passing through VFI:

$$S_V^{M2E} = \mathrm{VFI}_{\mathrm{M2E}}(V_{M_j}^G, V_{E_i}^G, V_{E_i}^L).$$

The three input features are sufficiently interacted and integrated, resulting in the similarity matching score S_V^{M2E}. Similarly, for the feature interaction from E_i to M_j, the similarity score S_V^{E2M} can be obtained through VFI:

$$S_V^{E2M} = \mathrm{VFI}_{\mathrm{E2M}}(V_{E_i}^G, V_{M_j}^G, V_{M_j}^L).$$

Based on this, the final visual similarity score S_V can be obtained:

$$S_V = (S_V^{M2E} + S_V^{E2M})/2.$$

Textual-Features Interaction (TFI). TFI computes the dot product of the normalized global features $T_{M_j}^G$ and $T_{E_i}^G$, yielding the text global-to-global similarity score S_T^{G2G}:

$$S_T^{G2G} = T_{M_j}^G \cdot T_{E_i}^G.$$

To further uncover fine-grained clues within local features, TFI applies attention mechanism to capture context vector from the local features $T_{M_j}^L$ and $T_{E_i}^L$, producing the global-to-local similarity score S_T^{G2L} between the global feature $T_{E_i}^G$ and the context vector:

$$S_T^{G2L} = \mathrm{TFI}_{\mathrm{G2L}}(T_{E_i}^G, T_{M_j}^L, T_{E_i}^L).$$

Based on this, the final textual similarity score S_T can be obtained:

$$S_T = (S_T^{G2G} + S_T^{G2L})/2.$$

Cross-Modal Features Interaction (CMFI). CMFI performs a fine-grained fusion of features across modalities. It integrates visual and textual features to generate a new context vector, h_e:

$$h_e = \mathrm{CMFI}(T_{E_i}^G, V_{E_i}^L).$$

The mention is processed similarly to produce the new context vector h_m:

$$h_m = \mathrm{CMFI}(T_{M_j}^G, V_{M_j}^L).$$

Based on this, the final multimodal similarity score S_C can be obtained:

$$S_C = h_e \cdot h_m.$$

5.5 Contrastive Learning

Based on the three similarity scores S_V, S_T, and S_C, the model is trained using contrastive loss function. For a mention M and entity E, the combined similarity score is the average of the similarity scores from the three independent units:

$$S(M, E) = (S_V + S_T + S_C)/3.$$

This loss function can be formulated as:

$$\mathcal{L}_O = -log\frac{\exp(S(M_j, E_i))}{\sum_i \exp(S(M_j, E_i'))},$$

where E_i represents the positive entity corresponding to M_j, while E_i' denotes negative entity from the knowledge base $\mathcal{E}$. It is expected to assign higher evaluation to positive mention-entity pairs and lower evaluation to negative ones.

Similarly, the three independent units are trained separately using contrastive loss function:

$$\mathcal{L}_X = -log\frac{\exp(S_X(M_j, E_i))}{\sum_i \exp(S_X(M_j, E_i'))}, X \in \{V, T, C\}.$$

The final optimization objective function is expressed as:

$$\mathcal{L} = \mathcal{L}_O + \lambda(\mathcal{L}_V + \mathcal{L}_T + \mathcal{L}_C),$$

where λ is the hyperparameter to control the loss.

6 Experiments

6.1 Experimental Settings

All the training and testing are conducted on a device equipped with 4 Intel(R) Xeon(R) Platinum 8380 CPUs and 8 NVIDIA A800-SXM4-80GB GPUs. Detailed experimental settings are provided in Appendix A.1. To comprehensively evaluate the effectiveness of our approach, we compare FBMEL with various competitive MEL baselines and VLM baselines. A detailed introduction of these baselines is provided in the Appendix A.2.

For the VP-MEL task experiments, all approaches are evaluated on the VPWiki dataset. And for the MEL task experiments, all approaches are evaluated on the WikiDiverse (Wang et al., 2022c) dataset. Additional experiments and detailed explanations are provided in the Appendix A.8.

Table 2: Performance comparison on the VP-MEL(a) and MEL(b) tasks. Baseline results marked with "∗" are based on Sui et al. (2024). Each method is run three times with different random seeds, and the mean value of each metric is reported. The model marked with "†" indicates the version without VLM. The best score is highlighted in bold. Detailed evaluation metrics can be found in Appendix A.3.

Methods	VP-MEL		
	H@1	H@3	H@5
BLIP-2-xl (Li et al., 2023)	15.86	35.41	45.32
BLIP-2-xxl (Li et al., 2023)	21.90	37.31	49.70
mPLUG-Owl3-7b (Ye et al., 2023)	29.46	30.45	48.94
LLaVA-1.5-7b (Liu et al., 2024)	43.20	64.35	65.71
LLaVA-1.5-13b (Liu et al., 2024)	32.93	65.56	66.92
MiniGPT-4-7b (Zhu et al., 2024)	28.10	33.53	37.31
MiniGPT-4-13b (Zhu et al., 2024)	37.61	37.61	40.03
VELML (Zheng et al., 2022)	22.51	37.61	43.35
GHMFC (Wang et al., 2022a)	25.53	41.39	48.94
MIMIC (Luo et al., 2023)	24.62	42.35	49.25
MELOV (Song et al., 2024)	26.44	42.75	51.51
FBMEL(ours)	**48.34**	**67.53**	**77.50**

(a)

Methods	MEL		
	H@1	H@3	H@5
ViLT* (Kim et al., 2021)	34.39	51.07	57.83
ALBEF* (Li et al., 2021)	60.59	75.59	83.30
CLIP* (Radford et al., 2021)	61.21	79.63	85.18
METER* (Dou et al., 2022)	53.14	70.93	77.59
BERT* (Devlin et al., 2019)	55.77	75.73	83.11
BLINK* (Wu et al., 2020)	57.14	78.04	85.32
JMEL* (Adjali et al., 2020)	37.38	54.23	61.00
VELML (Zheng et al., 2022)	55.53	78.11	84.61
GHMFC (Wang et al., 2022a)	61.17	80.53	86.21
MIMIC (Luo et al., 2023)	63.51	81.04	86.43
MELOV* (Sui et al., 2024)	**67.32**	**83.69**	**87.54**
FBMEL†(ours)	63.24	79.76	86.36

(b)

6.2 Main Results

Results on VP-MEL. As shown in Table 2a, FBMEL significantly outperforms all other methods on VP-MEL task. First, among the VLM methods, LLaVA-1.5 has the smallest performance gap compared to our method, with differences of 5.14%, 1.97%, and 10.58% from FBMEL across the three metrics, respectively. Even so, considering the substantial data and computational resources required to train LLaVA, FBMEL still demonstrates an obvious advantage. Second, there is a notable performance gap between MEL methods and FBMEL. MEL methods struggle with effective entity linking in scenarios where mention words are absent, underscoring their limitations and the robustness of our approach.

Results on MEL. Table 2b presents the experimental results comparing FBMEL with other methods on MEL dataset. During the testing process, the Detective-VLM is removed from FBMEL, and its output is replaced with mention words from WikiDiverse. The performance of FBMEL is slightly behind MIMIC and MELOV, with differences of only 0.27%, 0.28%, and 0.07% across the three metrics. This is expected, as FBMEL places greater emphasis on image features while reducing reliance on text features. Furthermore, the absence of visual prompts in the images limits the ability of FBMEL to fully utilize visual information.

Methods	WikiDiverse			WikiDiverse*		
	H@1	H@3	H@5	H@1	H@3	H@5
VELML	55.53	78.11	84.61	15.35	26.32	31.38
GHMFC	61.17	80.53	86.21	17.37	28.97	34.36
MIMIC	63.51	81.04	86.43	17.23	29.60	34.84
MELOV	**67.32**	**83.69**	**87.54**	17.66	30.03	36.43
FBMEL†	63.24	79.76	86.36	**23.87**	**38.37**	**45.14**

Table 3: Performance comparison in the absence of mention words on the WikiDiverse dataset. The symbol "∗" represents the dataset with mention words removed, while "†" denotes models without VLM.

6.3 Detailed Analysis

Influence of Mention Words on MEL Methods. For a fair comparison, we remove the Detective-VLM module from FBMEL. As shown in Table 3, the performance of MEL methods drop significantly across all three metrics in the absence of mention words. The average performance decline is 72.65%, 64.48%, and 60.28%, respectively. This indicates that MEL methods fail to extract meaningful information from visual and textual data, making them unsuitable for tasks without mention words. In contrast, even without Detective-VLM, visual prompts, or mention words, FBMEL can still achieve the best metrics. This demonstrates that FBMEL in the VP-MEL task possesses a stronger

Methods	VP-MEL				
	H@1	**H@3**	**H@5**	**H@10**	**H@20**
MiniGPT-4-7b	28.55	43.66	52.27	62.99	70.70
MiniGPT-4-13b	27.04	43.96	53.02	63.44	70.72
BLIP-2-xl	37.16	54.38	59.52	66.62	72.81
BLIP-2-xxl	40.63	54.53	61.78	68.73	74.62
LLaVA-1.5-7b	42.45	63.14	69.03	76.74	82.33
LLaVA-1.5-13b	41.54	59.37	66.92	73.11	77.80
Detective-VLM(ours)	**48.34**	**67.53**	**77.50**	**82.63**	**87.92**

Table 4: Performance comparison in different VLMs.

V^G**-Layer**	VP-MEL				
	H@1	**H@3**	**H@5**	**H@10**	**H@20**
Single Shallow Layer	39.88	60.88	71.00	79.31	86.56
Single Deep Layer	39.73	58.91	69.94	80.82	**88.07**
(Shallow+Deep) Layers	43.66	60.88	68.58	78.70	84.29
(3 Shallow+Deep) Layers	40.33	59.22	67.37	75.38	83.23
FBMEL	**48.34**	**67.53**	**77.50**	**82.63**	87.92

Table 5: Performance comparison across different feature layers in V^G.

capability to leverage both image and text information effectively.

Effect Analysis of Detective-VLM. As shown in Table 4, we evaluate the effectiveness of Detective-VLM by replacing it with various VLMs and analyzing the results. Our method achieves the best performance across all metrics. In particular, Detective-VLM shows an absolute improvement of 5.89% in Hit@1 compared to the second-best approach. In contrast, non-fine-tuned VLMs often produce a large amount of irrelevant information, which hampers subsequent processing.

Contributions of Visual Features from Different Layers. As shown in Table 5, we combine visual features from different layers during the extraction of V^G to compare the effects of various combinations. In the deeper layers of CLIP visual encoder, the model tends to focus more on abstract, high-level concepts. VP-MEL focuses on aligning high-level concepts between images and text, facilitating the capture of their semantic correspondence. This explains why using a single deep layer feature achieves the highest H@20 score of 88.07%. However, in the VP-MEL task, low-level texture details are equally important. Shallow texture features need to be extracted to help the model focus

Methods	VP-MEL				
	H@1	**H@3**	**H@5**	**H@10**	**H@20**
FBMEL	**48.34**	**67.53**	**77.50**	**82.63**	**87.92**
FBMEL†	35.65	53.93	65.26	73.57	80.51
FBMEL*	35.03	53.80	65.01	73.26	80.39

Table 6: The model marked "†" without VLM. The model marked "$*$" without VLM and Visual Prompts.

on the presence of visual prompts. Based on this, we choose to concatenate the deep features with the shallow features. Experimental results show that the best performance is achieved when the proportion of deep features is larger.

Ablation Study. In Table 6, we conduct ablation study on the FBMEL framework. First, we remove the Detective-VLM module from FBMEL, which results in a decline across all metrics. Notably, even without Detective-VLM, FBMEL shows robust entity linking performance, outperforming MEL methods as shown in Table 2a. This highlights the ability of FBMEL to efficiently leverage multimodal information from both images and text. To further evaluate the impact of Visual Prompts, we remove the visual prompts from the images in the previous setup. This leads to a noticeable drop in all metrics, underscoring the importance of visual prompts in guiding the model to focus on relevant regions within images.

7 Conclusion

In this paper, we introduce a new task, VP-MEL, aiming to link visual regions in image-text pairs to the corresponding entities in knowledge base, guided by visual prompts. To support this task, we construct a high-quality dataset, VPWiki, leveraging an automated annotation pipeline to enhance the efficiency of data annotation. To address the VP-MEL task, we propose FBMEL, a framework that can effectively utilize visual prompts to extract enhanced local visual features and generate supplementary textual information. FBMEL maintains a balance between visual and textual features, preventing excessive reliance on a single modality. Extensive experimental results demonstrate that FBMEL surpasses state-of-the-art methods. Moreover, VP-MEL significantly alleviates the constraints of mention words and expands the applicability of MEL to real-world scenarios.

Limitations

VP-MEL expands the application scenarios of MEL, allowing users to directly annotate areas of interest within images. However, this requires the image and text to be correlated to some extent. When the image and text are not correlated, the performance of VP-MEL can be limited. We choose to use grounding box as the format for visual prompts. In real-world application scenarios, users may use any irregular shape to make markings. In future work, the design of visual prompts will be enhanced. We hope this work can encourage more research to apply the recent advanced techniques from both natural language processing and computer vision fields to improve its performance.

Ethics Statement

The datasets employed in this paper, WikiDiverse, WikiMEL, and RichpediaMEL, are all publicly accessible. As such, the images, texts, and knowledge bases referenced in this study do not infringe upon the privacy rights of any individual.

References

Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020. Multimodal entity linking for tweets. In *ECIR*, pages 463–478. Springer.

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. 2024. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In *CVPR*, pages 12914–12923.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *HLT-NAACL*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhang Dongjie and Longtao Huang. 2022. Multimodal knowledge learning for named entity disambiguation. In *Findings of ACL: EMNLP*, pages 3160–3169, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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*. OpenReview.net.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. 2022. An empirical study of training end-to-end vision-and-language transformers. In *CVPR*, pages 18166–18176.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. In *Psychological bulletin*, volume 76, page 378. American Psychological Association.

Jingru Gan, Jinchang Luo, Haiwei Wang, Shuhui Wang, Wei He, and Qingming Huang. 2021. Multimodal entity linking: A new dataset and a baseline. In *ACM Multimedia*, page 993–1001, New York, NY, USA. Association for Computing Machinery.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

Jinyuan Li, Han Li, Di Sun, Jiahao Wang, Wenkun Zhang, Zan Wang, and Gang Pan. 2024. LLMs as bridges: Reformulating grounded multimodal named entity recognition. In *Findings of ACL: ACL*, pages 1302–1318, Bangkok, Thailand. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

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 *NeurIPS*, volume 34, pages 9694–9705.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306.

Pengfei Luo, Tong Xu, Shiwei Wu, Chen Zhu, Linli Xu, and Enhong Chen. 2023. Multi-grained multimodal interaction network for entity linking. In *KDD*, page 1583–1594, New York, NY, USA. Association for Computing Machinery.

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

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. 2024. Grounding multimodal large language models to the world. In *ICLR*. OpenReview.net.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR.

Senbao Shi, Zhenran Xu, Baotian Hu, and Min Zhang. 2024. Generative multimodal entity linking. In *LREC-COLING*, pages 7654–7665, Torino, Italia. ELRA and ICCL.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. 2023. What does clip know about a red circle? visual prompt engineering for vlms. In *ICCV*, pages 11987–11997.

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*, volume 38, pages 19008–19016.

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 *Findings of ACL: ACL*, pages 816–826, Bangkok, Thailand. Association for Computational Linguistics.

Peng Wang, Jiangheng Wu, and Xiaohang Chen. 2022a. Multimodal entity linking with gated hierarchical fusion and contrastive training. In *SIGIR*, page 938–948, New York, NY, USA. Association for Computing Machinery.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022b. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, pages 23318–23340. PMLR.

Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. 2022c. WikiDiverse: A multimodal entity linking dataset with diversified contextual topics and entity types. In *ACL (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 *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 *ACM Multimedia*, page 3599–3608, New York, NY, USA. Association for Computing Machinery.

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 *UAI*, volume 216 of *Proceedings of Machine Learning Research*, pages 2411–2421. PMLR.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *CVPR*, pages 13040–13051.

Gongrui Zhang, Chenghuan Jiang, Zhongheng Guan, and Peng Wang. 2023a. Multimodal entity linking with mixed fusion mechanism. In *DASFAA*, pages 607–622. Springer.

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. 2023b. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*.

Liang Zhao, En Yu, Zheng Ge, Jinrong Yang, Haoran Wei, Hongyu Zhou, Jianjian Sun, Yuang Peng, Runpei Dong, Chunrui Han, et al. 2023. Chatspot: Bootstrapping multimodal llms via precise referring instruction tuning. *arXiv preprint arXiv:2307.09474*.

Qiushuo Zheng, Hao Wen, Meng Wang, and Guilin Qi. 2022. Visual entity linking via multi-modal learning. In *Data Intel*, volume 4, pages 1–19.

Chong Zhou, Chen Change Loy, and Bo Dai. 2022. Extract free dense labels from clip. In *ECCV*, pages 696–712. Springer.

Qiang Zhou, Chaohui Yu, Shaofeng Zhang, Sitong Wu, Zhibing Wang, and Fan Wang. 2023. Regionblip: A unified multi-modal pre-training framework for holistic and regional comprehension. *arXiv preprint arXiv:2308.02299*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*. OpenReview.net.

A Appendix

A.1 Experimental Settings

For our proposed model framework, we use pre-trained ViT-B/32 (Dosovitskiy et al., 2021) as the visual encoder, initialized with weights from CLIP-ViT-Base-Patch32[1], with d_v and d_c set to 96. The

[1] https://huggingface.co/openai/clip-vit-base-patch32

batch size is set to 128. For the text encoder, we select pre-trained BERT model (Devlin et al., 2019), setting the maximum input length for text to 40 and the output feature dimension d_t to 512. We train and test on a device equipped with 4 Intel(R) Xeon(R) Platinum 8380 CPUs and 8 NVIDIA A800-SXM4-80GB GPUs.

A.2 Descriptions of Baselines

To thoroughly evaluate the performance of our method, we compare it against strong MEL baselines, including **BERT** (Devlin et al., 2019), **BLINK** (Wu et al., 2020), **JMEL** (Adjali et al., 2020), **VELML** (Zheng et al., 2022), **GHMFC** (Wang et al., 2022a), **MIMIC** (Luo et al., 2023) and **MELOV** (Sui et al., 2024).

Additionally, we select robust VLMs for comparison, including **BLIP-2-xl** [2], **BLIP-2-xxl** [3] (Li et al., 2023), **mPLUG-Owl3-7b** [4] (Ye et al., 2023), **LLaVA-1.5-7b** [5], **LLaVA-1.5-13b** [6] (Liu et al., 2024), **MiniGPT-4-7b** [7], **MiniGPT-4-13b** [8] (Zhu et al., 2024), **ViLT** (Kim et al., 2021), **ALBEF** (Li et al., 2021), **CLIP** (Radford et al., 2021), and **METER** (Dou et al., 2022). We reimplemented JMEL, VELML and MELOV according to the original literature due to they did not release the code. We ran the official implementations of the other baselines with their default settings.

•**BERT** (Devlin et al., 2019) is a pre-trained language model based on the Transformer architecture, designed to deeply model contextual information from both directions of a text, generating general-purpose word representations.

•**BLINK** (Wu et al., 2020) present a two-stage zero-shot linking algorithm, where each entity is defined only by a short textual description.

•**JMEL** (Adjali et al., 2020) extracts both unigram and bigram embeddings as textual features. Different features are fused by concatenation and a fully connected layer.

•**VELML** (Zheng et al., 2022) utilizes VGG-16 network to obtain object-level visual features. The two modalities are fused with additional attention mechanism.

•**GHMFC** (Wang et al., 2022a) extracts hierarchical features of text and visual co-attention through the multi-modal co-attention mechanism.

•**MIMIC** (Luo et al., 2023) devise three interaction units to sufficiently explore and extract diverse multimodal interactions and patterns for entity linking.

•**MELOV** (Sui et al., 2024) incorporates inter-modality generation and intra-modality aggregation.

•**BLIP-2** (Li et al., 2023) effectively utilizes the noisy web data by bootstrapping the captions, where a captioner generates synthetic captions and a filter removes the noisy ones.

•**mPLUG-Owl3** (Ye et al., 2023) propose novel hyper attention blocks to efficiently integrate vision and language into a common language-guided semantic space, thereby facilitating the processing of extended multi-image scenarios.

•**LLaVA-1.5** (Liu et al., 2024) is an end-to-end trained large multimodal model that connects a vision encoder and an LLM for general purpose visual and language understanding.

•**MiniGPT-4** (Zhu et al., 2024) aligns a frozen visual encoder with a frozen LLM, Vicuna, using just one projection layer.

•**ViLT** (Kim et al., 2021) commissions the transformer module to extract and process visual features in place of a separate deep visual embedder.

•**ALBEF** (Li et al., 2021) introduce a contrastive loss to align the image and text representations before fusing them through cross-modal attention, which enables more grounded vision and language representation learning.

•**CLIP** (Radford et al., 2021) is a neural network trained on a variety of (image, text) pairs. It can be instructed in natural language to predict the most relevant text snippet, given an image.

•**METER** (Dou et al., 2022) systematically investigate how to train a fully-transformer VLP model in an end-to-end manner.

A.3 Evaluation Metrics

For evaluation, we utilize Top-k accuracy as the metric that can be calculated by the following

[2] https://huggingface.co/Salesforce/blip2-flan-t5-xl-coco

[3] https://huggingface.co/Salesforce/blip2-flan-t5-xxl

[4] https://huggingface.co/mPLUG/mPLUG-Owl3-7B-240728

[5] https://huggingface.co/liuhaotian/llava-v1.5-7b

[6] https://huggingface.co/liuhaotian/llava-v1.5-13b

[7] https://drive.google.com/file/d/1RY9jV0dyqLX-o38LrumkKRh6Jtaop58R/view?usp=sharing

[8] https://drive.google.com/file/d/1a4zLvaiDBr-36pasffmgpvH5P7CKmpze/view?usp=share_link

formula:

$$\text{Accuracy}_{top-k} = \frac{1}{N}\sum_{i}^{N} I(t_i \in y_i^k),$$

where N represents the total number of samples, and I is the indicator function. When the receiving condition is satisfied, I is set to 1, and 0 otherwise.

A.4 Detective-VLM

Detective-VLM is based on the mplug-owl2 framework (Ye et al., 2024), with instruction fine-tuning carried out using the mplug-owl2-llama2-7b model [9].

We utilize VPWiki dataset to design the fine-tuning dataset, where $\{Image\}$ and $\{Sentence\}$ correspond to m_{v_j} and m_{s_j} in M_j, respectively. In the fine-tuning dataset, the $\{Entity\ Name\}$ corresponds to the mention words in M_j that are associated with the Visual prompt, the $\{Entity\ Type\}$ is one of [$Person$, $Organization$, $Location$, $Country$, $Event$, $Works$, $Misc$].

A.5 Feature Interaction Formula

Visual-Features Interaction (VFI). The two similarity scores S_V^{M2E} and S_V^{E2M} in visual feature interaction are calculated using the same method. Here, we take S_V^{M2E} as an example.

$$\overline{h}_p = \text{MeanPooling}(V_{E_i}^L),$$

$$h_{vc} = \text{FC}(\text{LayerNorm}(\overline{h}_p + V_{M_j}^G)),$$

$$h_{vg} = \text{Tanh}(\text{FC}(h_{vc})),$$

$$h_v = \text{LayerNorm}(h_{vg} * h_{vc} + V_{E_i}^G),$$

$$S_V^{M2E} = h_v \cdot V_{M_j}^G.$$

Textual-Features Interaction (TFI). The calculation of the global-to-local similarity score S_T^{G2L} incorporates an attention mechanism as follows:

$$Q, K, V = T_{E_i}^L W_{tq}, T_{M_j}^L W_{tk}, T_{M_j}^L W_{tv},$$

$$H_t = \text{softmax}(\frac{QK^T}{\sqrt{d_T}})V,$$

where $T_{E_i}^L W_{tq}$, $T_{M_j}^L W_{tk}$, $T_{M_j}^L W_{tv}$ are learnable matrices.

$$h_t = \text{LayerNorm}(\text{MeanPooling}(H_t)),$$

$$S_T^{G2L} = \text{FC}(T_{E_i}^G) \cdot h_t.$$

[9] https://huggingface.co/MAGAer13/mplug-owl2-llama2-7b

	WIKIDiverse	WikiMEL
Sentences	7,405	22,070
M. in train	11,351	18,092
M. in valid	1,664	2,585
M. in test	2,078	5,169
Entities	132,460	109,976

Table 7: Statistics of WIKIDiverse and WikiMEL. M. denotes Mentions.

Cross-Modal Features Interaction (CMFI). CMFI performs alignment and fusion of features from different modalities.

$$h_{et}, h_{mt} = \text{FC}_{c1}(T_{E_i}^G), \text{FC}_{c1}(T_{M_j}^G),$$

$$H_{ev}, H_{mv} = \text{FC}_{c2}(V_{E_i}^L), \text{FC}_{c2}(V_{M_j}^L),$$

in which FC_{c1} is defined by $W_{c1} \in \mathbb{R}^{d_t \times d_c}$ and $b_{c1} \in \mathbb{R}^{d_c}$, FC_{c2} is defined by $W_{c2} \in \mathbb{R}^{d_v \times d_c}$ and $b_{c2} \in \mathbb{R}^{d_c}$.

$$\alpha_i = \frac{exp(h_{et} \cdot H_{ev}^i)}{\sum_1^{n+1} exp(h_{et} \cdot H_{ev}^i)},$$

$$h_{ec} = \sum_{i}^{n} \alpha_i * H_{ev}^i, i \in [1, 2, ..., (n+1)],$$

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

in which FC_{c3} is defined by $W_{c3} \in \mathbb{R}^{d_c \times d_c}$ and $b_{c3} \in \mathbb{R}^{d_c}$.

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

By replacing inputs h_{et} and H_{ev} with h_{mt} and H_{mv}, h_m can be obtained using the aforementioned formula.

A.6 Pipeline Module Settings

For the Visual Entailment Module and Visual Grounding Module, we choose $\text{OFA}_{large(VE)}$ and $\text{OFA}_{large(VG)}$ (Wang et al., 2022b), respectively.

A.7 WikiDiverse and WikiMEL

WikiDiverse is a high-quality human-annotated MEL dataset with diversified contextual topics and entity types from Wikinews, which uses Wikipedia as the corresponding knowledge base. WikiMEL is collected from Wikipedia entities pages and contains more than 22k multimodal sentences. The statistics of WIKIDiverse and WikiMEL are shown

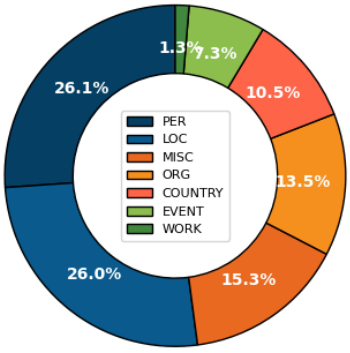


Figure 5: Entity type distribution of WIKIDiverse.

Methods	WikiMEL			RichpediaMEL		
	H@1	H@3	H@5	H@1	H@3	H@5
ViLT*	72.64	84.51	87.86	45.85	62.96	69.80
ALBEF*	78.64	88.93	91.75	65.17	82.84	88.28
CLIP*	83.23	92.10	94.51	67.78	85.22	90.04
METER*	72.46	84.41	88.17	63.96	82.24	87.08
BERT*	74.82	86.79	90.47	59.55	81.12	87.16
BLINK*	74.66	86.63	90.57	58.47	81.51	88.09
JMEL*	64.65	79.99	84.34	48.82	66.77	73.99
VELML	68.90	83.50	87.77	62.80	82.04	87.84
GHMFC	75.54	88.82	92.59	76.95	88.85	92.11
FBMEL†	81.15	92.38	93.44	78.96	91.63	94.36
MIMIC	87.98	95.07	96.37	81.02	91.77	94.38
MELOV*	**88.91**	**95.61**	**96.58**	**84.14**	**92.81**	**94.89**

Table 8: Baseline results marked with "∗" according to Sui et al. (2024). We run each method three times with different random seeds and report the mean value of every metric. The model marked '†' is the FBMEL retrained on the WikiMEL and RichpediaMEL datasets individually. The best score is highlighted in bold.

in Table 7. The entity type distribution of WIKIDiverse is illustrated in Figure 5.

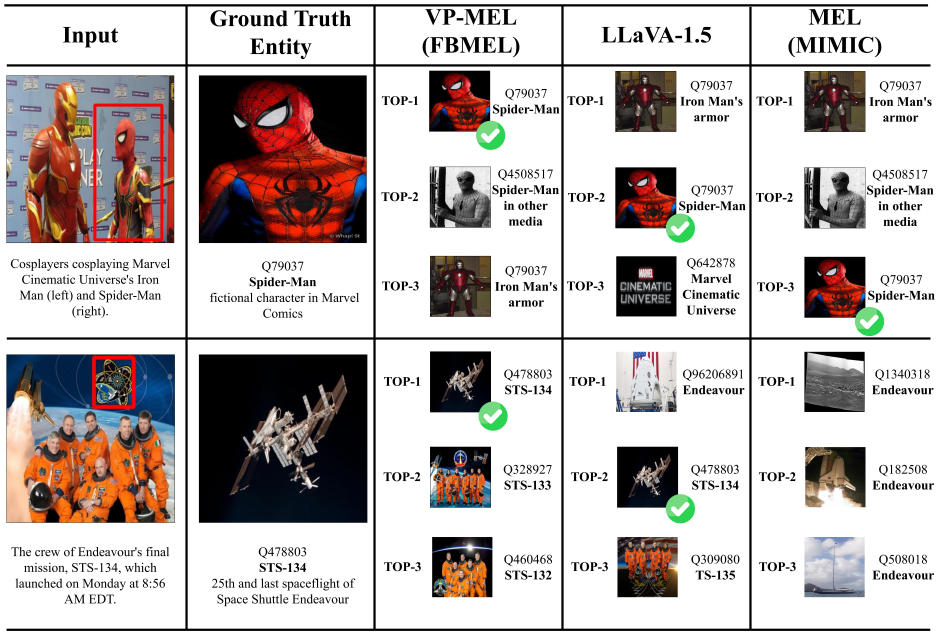
During the data collection process, we select the entire WIKIDiverse dataset along with 5,000 samples from the WikiMEL dataset. Compared to WikiMEL, WIKIDiverse features more content-rich images that better represent real-world application scenarios, making it particularly suitable for meeting the requirements of the VP-MEL task in practical contexts. Consequently, WIKIDiverse constitutes the majority of the VPWiki dataset. Additionally, we integrate the knowledge bases (KBs) from both datasets, resulting in an entity set encompassing all entities in the main namespace.

A.8 Additional Experiments and Experimental Descriptions

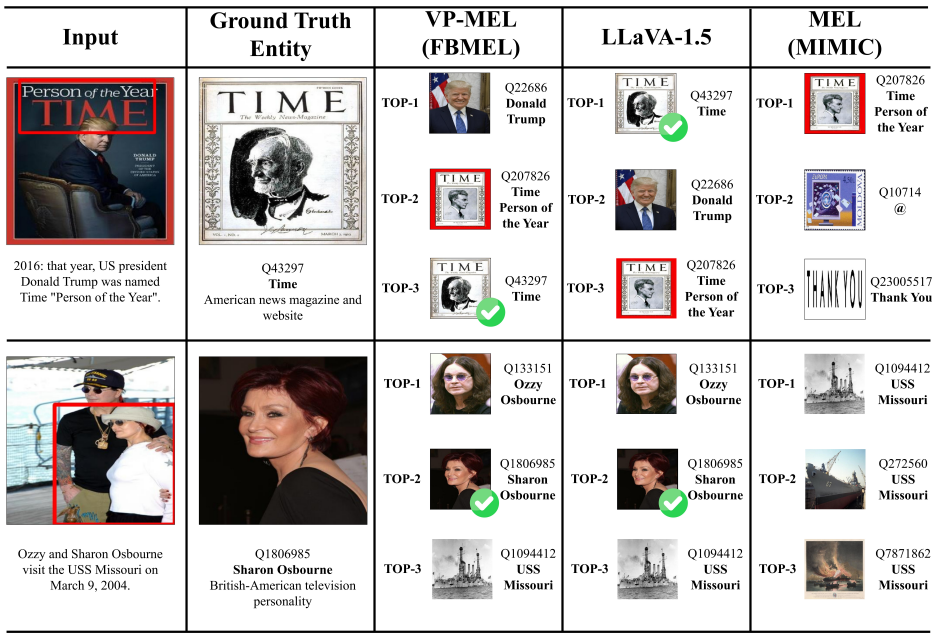
The knowledge base (KB) used in the WikiMEL and RichpediaMEL datasets differs from that of VPWiki and WIKIDiverse. To comprehensively assess the performance of the FBMEL framework in the MEL task, we retrain FBMEL on the WikiMEL and RichpediaMEL datasets. Importantly, neither visual prompts nor the Detective-VLM are employed during the retraining process. The experimental results are shown in Table 8. FBMEL is second only to MIMIC and MELOV, demonstrating its competitive performance as an MEL method.

A.9 Case Study

To clearly demonstrate the proposed VP-MEL task and the FBMEL model, we conduct case studies and compare them against two strong competitors (*i.e.*, LLaVA-1.5 and MIMIC), in Figure 6. As shown in Figure 6a, in the first case, all three methods correctly predicted the entity. FBMEL makes full use of both image and text information, allowing it to more effectively distinguish between the different individuals in the image. LLaVA-1.5 may be overwhelmed by the textual information, while MIMIC struggles to identify the correct entity when the mention words are unavailable. In the second case, both LLaVA-1.5 and MIMIC retrieve *Endeavour* as the first choice. Only FBMEL, with the guidance of Visual Prompts and integration of textual information, correctly predicts the right entity. In Figure 6b, we present the failed predictions. In the first case, when the content of the image interferes with the visual prompt, it impairs the reasoning process of FBMEL. The red box in the image bears a high similarity to the visual prompt. As a result, FBMEL incorrectly focuses on the wrong region of the image, ranking *Donald Trump* first. When FBMEL encounters difficulties in distinguishing the objects within the visual prompts, it leads to incorrect inferences. For example, in the second case, the distinguishing features of the two individuals in the image are obstructed, which causes FBMEL to struggle in differentiating between them. The image content in real-world data is often complex, which makes VP-MEL a challenging task. We hope that this task can be further refined and developed over time.



(a) Successful predictions.



(b) Failed predictions.

Figure 6: Case study for VP-MEL. Each row is a case, which contains Input, ground truth entity, and top three retrieved entities of three methods, *i.e.*, FBMEL (ours), LLaVA-1.5 (Liu et al., 2024), MIMIC (Luo et al., 2023). Each retrieved entity is described by its Wikidata QID and entity name, with the entity marked with a checkmark indicating the correct one.