Trust in Internal or External Knowledge? Generative Multi-Modal Entity Linking with Knowledge Retriever

Xinwei Long[1*], **Jiali Zeng**[2], **Fandong Meng**[2], **Jie Zhou**[2] **Bowen Zhou**[1†]
[1] Department of Electronic Engineering, Tsinghua University, Beijing, China,
[2] Pattern Recognition Center, WeChat AI, Tencent Inc, China
`longxw22@mails.tsinghua.edu.cn`, `zhoubowen@tsinghua.edu.cn`
`{lemonzeng,fandongmeng,withtomzhou}@tencent.com`

Abstract

Multi-modal entity linking (MEL) is a challenging task that requires accurate prediction of entities within extensive search spaces, utilizing multi-modal contexts. Existing generative approaches struggle with the knowledge gap between visual entity information and the intrinsic parametric knowledge of LLMs. To address this knowledge gap, we introduce a novel approach called GELR, which incorporates a knowledge retriever to enhance visual entity information by leveraging external sources. Additionally, we devise a prioritization scheme that effectively handles noisy retrieval results and manages conflicts arising from the integration of external and internal knowledge. Moreover, we propose a noise-aware instruction tuning technique during training to finely adjust the model's ability to leverage retrieved information effectively. Through extensive experiments conducted on three benchmarks, our approach showcases remarkable improvements, ranging from 2.0% to 6.5%, across all evaluation metrics compared to strong baselines. These results demonstrate the effectiveness and superiority of our proposed method in tackling the complexities of multi-modal entity linking.

1 Introduction

Entity linking is a crucial component of information extraction that plays a fundamental role in various knowledge applications (Hoffmann et al., 2011; Zeng et al., 2018, 2019; Long et al., 2020, 2021; Gao et al., 2021). It serves as a foundational tool to support tasks like visual question answering (Si et al., 2022; Ma et al., 2023) and semantic search by enabling the acquisition of entity-specific knowledge. Over the years, there have been significant advancements in textual entity linking (Zhu et al., 2023b; Chen et al., 2023), which involves establishing connections between mentions in unstructured text and entities in knowledge bases (KBs). However, the challenges arise when applying entity linking to social media contexts (Wang et al., 2022c; Zhao et al., 2024), where the understanding of multi-modal content becomes paramount. In these scenarios, the ability to link ambiguous mentions to their corresponding entities requires a comprehensive comprehension of multi-modal information, as illustrated in Fig. 1. Consequently, the field of entity linking has extended its focus to encompass multi-modal scenarios.

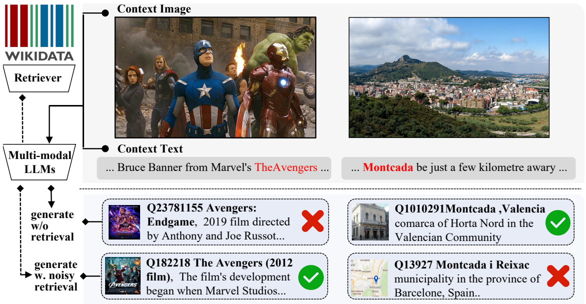
WIKIDATA
Retriever
Multi-modal LLMs
Context Image
Context Text
... Bruce Banner from Marvel's TheAvengers ...
... Montcada be just a few kilometre awary ...
generate w/o retrieval
generate w. noisy retrieval
Q23781155 Avengers: Endgame, 2019 film directed by Anthony and Joe Russot...
Q1010291Montcada ,Valencia comarca of Horta Nord in the Valencian Community
Q182218 The Avengers (2012 film), The film's development began when Marvel Studios...
Q13927 Montcada i Reixac municipality in the province of Barcelone, Spain..

Figure 1: Multi-Modal Entity Linking.

For MEL, two main types of approaches are recognized: 1) Retrieval-based methods (Xing et al., 2023) excel at gathering potentially relevant contexts but often struggle to precisely identify the correct one. 2) Generative approaches (Cao et al., 2021) generate entities using the intrinsic knowledge present in Large Language Models (LLMs). LLMs achieve better performance in top-ranked predictions, especially for common entities, thanks to their larger parameter size and stronger comprehension abilities. Nevertheless, LLMs face challenges in associating unseen visual inputs with their internal entity knowledge. This gap makes it difficult to establish connections between fine-grained visual patterns and the corresponding entities. In such cases, combining retrieval-based and LLM-

[*] The work was done when Xinwei worked as intern at Pattern Recognition Center, Wechat AI, Tencent Inc, China.
[†] Corresponding author

Findings of the Association for Computational Linguistics: ACL 2024, pages 7559–7569
August 11-16, 2024 ©2024 Association for Computational Linguistics

based methods becomes beneficial. By obtaining relevant information through matching, the model can access its internal knowledge and make predictions, leveraging the strengths of both approaches. This integration enhances the overall performance of the entity-linking system.

To this end, we introduce GELR, which bridges the knowledge gap between visual patterns and the parametric knowledge of LLMs by incorporating knowledge retrievers, achieving a balanced integration of extrinsic and intrinsic knowledge. However, utilizing external knowledge via retrievers can introduce additional issues when conflicts arise between parametric and retrieved knowledge (Li et al., 2023a). As depicted in Fig. 1, the linker adopts noisy retrieval results rather than relying on its own knowledge, leading to incorrect predictions “Montcada i Reixac”. To discern noisy retrieval results and manage knowledge conflicts, we devise a prioritization scheme: GELR makes predictions conditioned on the retrieved entities if the retrieval results contain accurate clues; otherwise, the model relies exclusively on intrinsic parametric knowledge for predictions. Towards this, we propose noise-aware instruction tuning to fine-tune the model in utilizing retrieved results by adjusting the predicted distribution for relevant and irrelevant knowledge, respectively.

For entity generation during inference, our fine-tuned model directly utilizes the weighted difference between retrieval-aware and retrieval-free logits as the output logits. Note that the relevance of retrieved entities is not given as prior information to the model, and our GELR solely determines its output based on the semantic features of multi-modal inputs and retrieved results.

We carry out primary experiments on three benchmarks with knowledge bases ranging from 112K to 6M. The experimental results show significant improvements of 2.2-6.5% across all metrics compared to strong baselines. Additionally, the experimental analysis highlights our method’s effectiveness in leveraging both internal and external knowledge and alleviating the popularity bias in generative entity-linking methods.

2 Related Work

The traditional Entity Linking (EL) task (Hoffmann et al., 2011; Daiber et al., 2013), which involves identifying mentions in text and linking them to corresponding entities in a Knowledge Base (KB), has been extensively researched. Existing methods for EL can be categorized into two types. The first type (Huang et al., 2022) utilizes retrieval techniques to select the most relevant entities from an external knowledge base. The second type (Cao et al., 2021) leverages the intrinsic knowledge within language models to directly generate entities associated with mentions. While textual entity linking has achieved significant advancements (Zhu et al., 2023b; Xiao et al., 2023) in accuracy on traditional benchmarks such as AIDA (Hoffmann et al., 2011), challenges arise in social media applications (Wang et al., 2022c), where the understanding of multi-modal contexts and entities is required.

Multi-modal Entity Linking (MEL) is an extended version of the textual entity linking task, which focuses on linking textual or visual mentions with multi-modal contexts to their corresponding entities in the KB (Sun et al., 2022; Wang et al., 2022c). To address these challenges, most methods adopt a two-stage retrieval framework (Wang et al., 2022b; Luo et al., 2023; Zhang et al., 2023), which involves recalling candidate entities and re-ranking the candidate set to select the best match. However, these straightforward methods heavily rely on high-quality candidate sets and suffer from cascading errors (Shi et al., 2023a). Moreover, existing retrievers excel in acquiring relevant contexts but struggle to precisely identify the correct one, leading to performance bottlenecks in entity-related tasks.

Recently, Large Language Models (LLMs) have achieved remarkable success due to their impressive generative capabilities. Notably, studies (Li et al., 2023b) have emerged as an efficient fine-tuning paradigm where most parameters are frozen, and only newly added alignment modules are trained, surpassing the performance of the CLIP model. Building on this success, subsequent research (Zhu et al., 2023a) has extended the application of LLMs to other multi-modal tasks. In addition to efficient fine-tuning and leveraging the intrinsic knowledge of LLMs, there has been research (Chen et al., 2022, 2023) focusing on retrieval-augmented multi-modal methods that incorporate external knowledge using retrievers.

There have been preliminary attempts to use LLMs as virtual knowledge bases to address multi-modal entity linking tasks (Wang et al., 2023; Shi et al., 2023a; Long et al., 2024). Different from them, we leverage retrieved visual contexts as external potential knowledge to bridge the gap between different modalities, rather than simply using them as direct output. Furthermore, we focus on learning the integration of both noisy external knowledge and intrinsic knowledge to make accurate predictions.

3 Methodology

We propose GELR, a generative framework with knowledge retrievers for multi-modal entity linking, which learns to utilize knowledge from both external sources and intrinsic parameters of LLMs. GELR consists of three components, as depicted in Fig. 2: Multi-modal knowledge retrieval, Noise-aware instruction tuning, and Inference.

3.1 Problem Formulation

Formally, let $\mathcal{E} = \{E_1, E_2, ..., E_i\}_{i=1}^{N}$ denotes a multi-modal knowledge graph used for the multi-modal entity linking task, and each entity is characterized with rich visual and textual descriptions. Our model takes the multi-modal context $C_i = \{T_i, V_i, m_i\}$ as input, where m_i represents a mention surrounding by the textual content T_i and visual content V_i, and our goal is to generate the corresponding entity E_j in $\mathcal{E}$, which is the most relevant to the mention m_i.

Specifically, For a given input C_i, we first encode the image and its captions. Then, we utilize their features to retrieve Top-K potentially useful entries from the knowledge graph $\mathcal{E}$. Finally, we condition on the input C_i and retrieved entries $K = \{E_k\}_{k=1}^{N}$ to generate the corresponding entity E_j with an auto-regressive score, as Eq 1.

$$\mathrm{score}(E_j, C_i) = \prod_{j=1} \mathcal{F}_{\mathbf{\Theta}}(y_j | \boldsymbol{y}_{<j}, C_i, K). \quad (1)$$

where $\mathbf{\Theta}$ denotes parameters of GELR, and y_j is the j_{th} token of the entity identifier of E. During inference, our model employs a constrained strategy to guide the generation (Cao et al., 2021), which ensures that the predicted E_j indeed exists in the KB.

3.2 Multi-modal Knowledge Retrieval

The multi-modal knowledge retrieval is performed to acquire multi-modal knowledge from external sources to mitigate the gap between the intrinsic knowledge of LLMs and visual patterns. The retrieval system consists of two encoders, a vision encoder $\mathcal{F}_V$ and a language encoder $\mathcal{F}_L$, which encodes visual and textual features, respectively.

Given a multi-modal input $C_i = \{T_i, V_i, m_i\}$, We utilize two types of visual features: one comprises text-based vision representations acquired through image-to-text captioning tools, while the other encompasses feature-based vision representations, including holistic-level and object-level representations. For each image V_i, we obtain a textual description and then append the sequence behind the textual input T_i to form the text-based query q_t^i. We encode the text-based query through the encoder $\mathcal{F}_L$, and obtain the textual representation, as $\boldsymbol{h}_t^i$. For feature-based vision representations, region-of-interest (RoI) features are crucial to represent visual entities and their relationships. To obtain regional features, we crop the objects from the images and transform them into a fixed resolution. we employ the vision encoder $\mathcal{F}_V$ to extract both global and regional image features as $\boldsymbol{h}_v^i$ and $\{\boldsymbol{h}_r^j\}_{j=1}^{N}$.

Compared with previous image-to-text retrievers (Radford et al., 2021), GeLR utilizes richer information by employing multi-grained representations to improve knowledge retrieval. To align the multi-grained visual query features and textual features, we train an alignment network with a two-layer feed-forward layer $\mathcal{F}_A$ to project the query features into a unified feature space. Formally, the final query embeddings $\boldsymbol{Q}$, as Eq. 2,

$$\boldsymbol{Q} = [\mathcal{F}_A(\boldsymbol{h}_t^i); \mathcal{F}_A(\boldsymbol{h}_v^i); \mathcal{F}_A(\boldsymbol{h}_r^1); ...; \mathcal{F}_A(\boldsymbol{h}_r^N)]. \quad (2)$$

Given an entity E_i with its visual content V_i^e and textual description T_i^e, we first extract the crucial information from lengthy entity text descriptions, such as entity names, summaries, and keywords. Subsequently, we adopt the same backbones, $\mathcal{F}_L$ and $\mathcal{F}_V$, to encode the image and text, respectively. The final entity embeddings are formalized as $\boldsymbol{E}_i = [\boldsymbol{h}_t^e; \boldsymbol{h}_v^e]$.

We compute the similarity score between the multi-modal query $\boldsymbol{Q}$ and each entity $\boldsymbol{E}$ by a late interaction paradigm, as Eq. 3,

$$\mathrm{score(Q, E)} = \sum_{i=1}^{N+2} \max_{j=1}^{2} \boldsymbol{q}_i \boldsymbol{e}_j^T. \quad (3)$$

where $\boldsymbol{q}_i$ and $\boldsymbol{e}_j$ denote the row vectors in $\boldsymbol{Q}$ and $\boldsymbol{E}$, respectively. We train the retriever with the contrastive loss (Chen et al., 2020). We treat the ground-truth entities in the downstream entity linking task as supervision signals and adopt

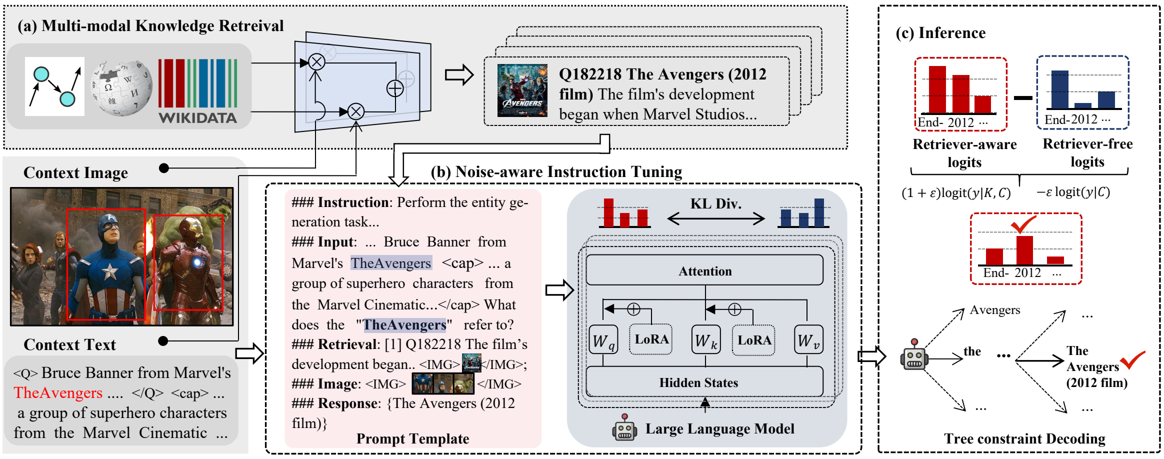
(a) Multi-modal Knowledge Retreival
WIKIDATA
Q182218 The Avengers (2012 film) The film's development began when Marvel Studios...
Context Image
Context Text
<Q> Bruce Banner from Marvel's TheAvengers .... </Q> <cap> ... a group of superhero characters from the Marvel Cinematic ...
(b) Noise-aware Instruction Tuning
### Instruction: Perform the entity generation task...
### Input: ... Bruce Banner from Marvel's TheAvengers <cap> ... a group of superhero characters from the Marvel Cinematic...</cap> What does the "TheAvengers" refer to?
### Retrieval: [1] Q182218 The film's development began.. <IMG></IMG>;
### Image: <IMG> </IMG>
### Response: {The Avengers (2012 film)}
Prompt Template
KL Div.
Attention
W_q
LoRA
W_k
LoRA
W_v
Hidden States
Large Language Model
(c) Inference
End- 2012 ···
End-2012 ···
Retriever-aware logits
Retriever-free logits
$(1+\varepsilon)\text{logit}(y|K,C)$
$-\varepsilon\ \text{logit}(y|C)$
End- 2012 ...
Avengers
the
...
The Avengers (2012 film)
Tree constraint Decoding

Figure 2: The overview of our GELR framework.

in-batch negative sampling for training, where all entities in a training batch other than the $\boldsymbol{E}^*$ are regarded negative for the current $\boldsymbol{Q}$. During inference, all entities are indexed using the FAISS vector database (Douze et al., 2024) for efficient retrieval.

3.3 Noise-aware Instruction Tuning

With multi-modal context as input, our model leverages internal knowledge from model parameters and external knowledge from external sources to predict relevant entities. Specifically, we first choose the Top-K entities with their key information as possibly useful knowledge through retrievers discussed in Sec 3.2. Then, we adopt two types of visual representations, including feature-based visual features encoded by the visual encoder $\mathcal{F}_V$, and text-based image description which will be directly fed into the LLM. Following previous studies (Zhu et al., 2023a), we utilize a simple linear layer to map the visual features into the same latent space of LLMs. To efficiently fine-tune the LLM, we create the instruction template with predefined slots for filling the multi-modal contexts and retrieved entities as depicted in Fig.2. The template also contains a task description to prompt the model generation of entity identifiers (e.g. entity name and entity URL). In this stage, we freeze the parameters of the backbones and only keep the parameters of LORA adapters (Hu et al., 2021) and linear layers between each backbone updated.

However, utilizing external knowledge via retrievers can introduce additional challenges when conflicts arise between parametric and retrieved knowledge (Li et al., 2023a; Shi et al., 2023b). In practice, irrelevant and misleading content within retrieved entities can lead the model to struggle to determine whether to prioritize external or internal knowledge for final predictions, consequently leading to performance degradation.

To make a good trade-off between external and internal knowledge, we aim to train an ideal model capable of distinguishing the relevance of retrieved entities. In other words, assuming the retrieved entities contain correct clues, the model makes predictions conditioned on the retrieved results; otherwise, the model relies solely on internal knowledge for predictions, as Eq. 4,

$$\mathcal{F}_\Theta(E|C_i,K) = \begin{cases} \mathcal{F}_\Theta(E|C_i,K) & E^* \in \{E\}_{j=1}^{N} \\ \mathcal{F}_\Theta(E|C_i) & otherwise \end{cases} \tag{4}$$

To achieve this goal, we aim for the model to learn the prioritization scheme. Specifically, if the retrieved set of entities contains the ground-truth entity $E*$, the model should utilize the retrieved results rather than internal parametric knowledge. To establish the priority order, We obtain the final probability y_t through the weighted difference of logits from retrieval-aware outputs $\mathcal{F}_\Theta(\boldsymbol{E}|C_i,K)$ and retrieval-free outputs $\mathcal{F}_\Theta(\boldsymbol{E}|C_i)$ as Eq. 5, and optimize them using auto-regressive loss as $\mathcal{L}_{\mathrm{AR}}$.

$$\begin{aligned} y_t = \sigma[&(1+\epsilon)\mathcal{F}_\Theta(y_t|\boldsymbol{y}_{<t},C_i,K) \\ &-\epsilon\cdot\mathcal{F}_\Theta(y_t|\boldsymbol{y}_{<t},C_i)]. \end{aligned} \tag{5}$$

Conversely, if the retrieved entities E are irrelevant to the context, the model relies on internal parametric knowledge for predictions. This means that the model's output should remain consistent with the output when there is no retrieved result. To capture this consistency relationship, we employ KL divergence to model the difference between the distributions of the two outputs, when retrieved results are irrelevant to the contexts, as Eq. 6,

$$\mathcal{L}_{\text{KL}} = D_{KL}(\sigma(\mathcal{F}_{\Theta}(y_t|\boldsymbol{y}_{<t}, C_i, K)) \\ || \; \sigma(\mathcal{F}_{\Theta}(y_t|\boldsymbol{y}_{<t}, C_i))) \quad (6)$$

where σ denotes the Softmax function in Eq. 5 and Eq. 6. Although we differentiate between relevant and irrelevant retrieval results to compute loss, the model is not provided with the relevance of retrieval results. It has to learn to determine relevance based on the semantics of both multi-modal contexts and retrieved entities.

3.4 Inference

To ensure accurate entity generation utilizing both internal and external knowledge, we persist in utilizing Eq. 5, which represents the weighted difference between retrieval-aware and retrieval-free logits, as the output logits of our model. Despite Eq. 5 being formulated under the assumption that retrieved entities are relevant, it remains applicable during inference even when the retrieval results are uncertain. This is because when the retrieval results are irrelevant, the distribution of retrieval-aware logits converges to that of retrieval-free logits, owing to the constraint imposed by Eq. 6 in training. The Adjustment of Logits can enhance the capacity of GELR to extract useful information from the retrieved entities.

During inference, our model applies the tree-based constraint decoding strategy (Cao et al., 2021) to guide the decoder in searching within a limited token space at each step, so as to generate a valid entity identifier (e.g. entity name, and URLs) that can be mapped to one entry within the knowledge graph.

4 Experiments

4.1 Experimental Settings

Datasets and Evaluation Metrics. We select two widely used MEL datasets, WikiDiverse (Wang et al., 2022c) and WikiMEL (Wang et al., 2022b), along with a visual entity-linking dataset WikiPerson (Sun et al., 2022).[1] These benchmark datasets rely on Wikidata[2] or its subsets as their primary knowledge sources. Detailed statistics of the datasets can be found in Tab. 5. For the WikiDiverse, we employ F1 as the evaluation metric. It is worth mentioning that in GMEL (Shi et al., 2023a), they filter out "NIL" type data samples from the WikiDiverse and use Accuracy@K as an alternative evaluation metric. Therefore, we report both F1 (Wang et al., 2022c) and Accuracy@K (Shi et al., 2023a) for the WikiDiverse. Additionally, we use Accuracy@K to evaluate the performance on WikiMEL and Recall@K to evaluate the performance on WikiPerson.

Baselines. We adopt several baseline methods for comparison, categorized as follows: (1) retrieval-based methods: CLIP (Radford et al., 2021), VNEL (Sun et al., 2022), LXMERT (Wang et al., 2022c), REAVL (Rao et al., 2023), and DRIN (Xing et al., 2023); (2) Generative methods: the freely available multi-modal LLM, LLaVA-13B (Liu et al., 2023), GENRE (Cao et al., 2021), GDMM (Wang et al., 2023), and GMEL (Shi et al., 2023a). We also adopt the results of GPT-3.5-Turbo from GMEL as a strong baseline.

Implementation Details. In our experiments, we use ViT-L/14 from pre-trained CLIP (Radford et al., 2021) as the image encoder and LLaMA-7B (Touvron et al., 2023) as the text backbone. We employ the checkpoint of Minigpt-4 (Zhu et al., 2023a) to initialize the parameters of the linear layer. We use YOLOv7 (Wang et al., 2022a) to obtain the bounding boxes for the datasets that do not provide the ground-truth boxes and keep up to 3 boxes as the ROIs. Our model is implemented by Pytorch and trained using the Adam optimizer with a warmup strategy. Training is performed on 4 Nvidia A6000 48G GPUs and completed within two hours. We choose the Top-8 entities from the retrieved results and set the value of ϵ as 0.2. To generate valid entity identifiers, we use a constrained beam search with 5 beams. For More experimental details and hyper-parameters please refer to the appendix A, and the code will be released in this repository[3].

4.2 Main Results

We compare our GELR with the aforementioned baselines for both multi-modal and visual entity linking tasks. The experimental results illustrate that GELR achieves significant improvements over the competitive baselines on three publicly available datasets.

[1] Due to a lack of maintenance or licensing restrictions, certain MEL datasets are inaccessible.

[2] https://www.wikidata.org/

[3] https://github.com/xinwei666/GELR

Method	WikiDiverse F1	WikiDiverse ACC@1	WikiMEL ACC@1	WikiPerson R@1	WikiPerson R@3	WikiPerson R@5
CLIP (Radford et al., 2021)	45.4	48.5	30.8	74.6	-	84.4
VNEL (Sun et al., 2022)	-	-	-	73.0	<u>82.4</u>	<u>85.1</u>
VNEL (Sun et al., 2022)	40.2	41.9	-	73.7	81.8	83.5
LXMERT (Wang et al., 2022c)	71.1	-	-	-	-	-
REAVL (Rao et al., 2023)	-	-	-	<u>77.5</u>	-	84.4
DRIN (Xing et al., 2023)	-	51.1	65.5	-	-	-
GPT-3.5-Turbo	-	72.7	73.8	-	-	-
LLaVA-13B (Liu et al., 2023)	75.4	77.6	76.1	63.6	-	-
GENER (Cao et al., 2021)	76.8	78.0	60.1	-	-	-
GDMM (Wang et al., 2023)	79.1	-	-	-	-	-
GMEL (Shi et al., 2023a)	<u>82.5</u>	<u>86.1</u>	<u>82.6</u>	35.5	50.7	57.7
GELR (Ours)	**87.4**	**89.5**	**84.8**	**81.7**	**88.9**	**90.5**

Table 1: Results on the multi-modal entity linking benchmarks WikiDiverse and WikiMEL, and visual entity linking benchmark WikiPerson. For each dataset, **bold** indicates the best model, and <u>underline</u> indicates the second best.

Multi-modal Entity Linking. As shown in Tab. 1, GELR outperforms the competitive baseline, GMEL, on both Wikidiverse and WikiMEL, which demonstrates the effectiveness of our method in leveraging both external and internal knowledge for entity linking tasks. In particular, GELR achieves the state-of-the-art micro F1 on Wikidiverse, surpassing the previous best generative method by more than 4.9% and outperforming discriminative retrieval-based methods by 16.3%. Wikidiverse poses a challenging issue of identifying mentions that lack corresponding entities in the knowledge base and categorizing them into the “NIL” category. GELR incorporates retriever results and seamlessly integrates them with multi-modal context during inference. It excels at distinguishing samples as “NIL” when the correlation between the retrieved information and the context is low. In contrast, other generative methods, such as LLaVA-13B and GMEL, often generate entity names similar to the mentioned text, leading to less accurate results.

Additionally, we consistently observe that generative baselines (e.g., GMEL) outperform retrieval-based baselines (e.g., DRIN). This finding can be attributed to several factors. Retrieval baselines struggle to precisely identify the best entity from a vast search space, while LLMs tend to predict common entities based solely on the text input when textual context dominates. We refer to this phenomenon as the popularity bias of LLMs, which we will discuss further in Sec. 5.1.

Visual Entity Linking. Back to Tab. 1, GELR demonstrates more significant performance improvements on WikiPerson, with Recall@K increasing by 4.2% to 6.5%. Notably, other generative baselines show poor performance. This can be attributed to the dominant influence of the visual modality, where language models face challenges in comprehending the intricate association between fine-grained visual patterns and entity knowledge. In contrast, our GELR effectively bridges this gap by leveraging multi-modal knowledge retrievers to incorporate potentially useful content.

4.3 Ablation Study

We conducted a series of ablation studies by gradually removing each module from the bottom to the top layer of our framework, and the corresponding results are presented in Tab. 2.

Initially, we can observe a decrease of w/o Retriever in F1 score by 3.2% and Acc@ by 3.1%. This indicates that the retrieval results play a crucial role in improving entity prediction. We will delve into the specific impact of the retriever on the final prediction in Sec. 5.2. Next, we removed the Visual Input while retaining the multi-modal information obtained through retrieval. It results in a marginal decrease in performance, suggesting that the retrieval results can partially compensate for the loss caused by the absence of visual input. Furthermore, removing both the Knowledge Retriever and Visual Input modules led to a significant drop of over 7%. This finding underscores the indispensability of entity knowledge embedded in visual information, which cannot be replaced by image captioning tools.

Delete Module	F1	ACC@1
Full Model	87.4	89.5
w/o Retriever	84.2	86.4
w/o Visual Input	85.2	87.1
w/o Retriever & Visual Input	80.1	82.0
w/o Noise-aware Tuning	84.1	87.1
w/o Adjustment of Logits	86.8	88.7
w/o Lora Adapter	83.5	86.3

Table 2: Results of Ablation Studies.

To explore the model's capacity to handle noise in retrieval results, we removed the Noise-aware Tuning module. Interestingly, we observed that the model's performance was similar to that of removing the retrievers. This indicates that without the noise-aware tuning module, the model may be susceptible to noise in the retrieval results, hindering its ability to leverage external knowledge effectively.

Finally, when we removed the LORA adapter and froze all parameters of the LLM, the model's performance decreased by 3.2%. However, even without fine-tuning the LLM, utilizing visual inputs and retrieval results as prompts enabled the model to access the parametric knowledge of the LLM, resulting in decent performance that surpassed other baselines.

5 Analysis

5.1 Popularity Bias of LLMs

Previous research indicates that generative methods exhibit varying predictive performance for head and tail entities. Common entities can often be inferred correctly by generative models based on limited textual features alone, without relying heavily on visual features. This phenomenon is commonly referred to as the popularity bias issue (Chen et al., 2021). Here, we analyzed the distribution and predictive performance of head and tail entities in Wikidiverse. Following previous work (Cao et al., 2021), we categorize entities based on their frequency of redirection through hyperlinks and aliases. Entities ranking within the top 20% in terms of frequency were classified as head entities, while the remaining entities were considered tail entities. Our analysis, as shown in Figure 3, reveals that 89% of the entities in the test set are classified as head entities, while only 11% are categorized as tail entities. This distribution helps explain why generative baselines tend to perform better on Wikidiverse, as they are more proficient in handling the prevalent head entities.

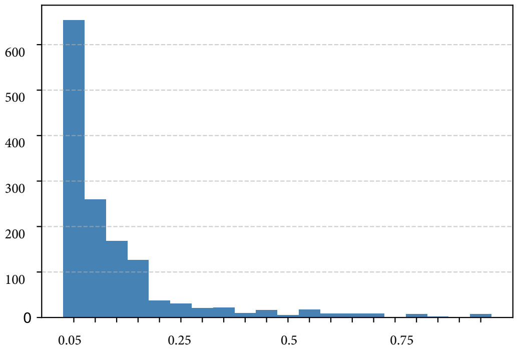

Figure 3: Popularity distribution on Wikidiverse.

Methods	Head Ent.	Tail Ent.
LLAVA-13B	81.2	48.5
GMEL	88.6	65.7
GELR (w/o Retriever)	88.5	69.5
GELR (Full Model)	89.9	80.8

Table 3: Effects of Popularity Bias.

Furthermore, we analyze the performance of our method on head entities and tail entities, respectively, comparing it with the best generative baselines, GMEL and LLaVA-13B. The results in Tab. 3 demonstrate that GELR surpasses the baseline performance on both head and tail entities, with a notably larger margin of 15.1% observed for tail entities. Moreover, we find that there is a performance gap between the tail and head entities on two generative baselines. The impressive results on head entities may be attributed to the internal knowledge of LLMs. Head entities appear more frequently during the pre-training phase, leading to a popularity bias in LLM generation towards head entities.

To investigate the gain of knowledge retrieval on predicting tail entities, we remove the knowledge retriever module. We find that GELR exhibited a significant decrease of 11.3% in performance for tail entities. This indicates the effectiveness of our framework in mitigating the effects of LM's popularity bias issue towards tail entity prediction, thereby improving the overall performance.

5.2 Effects of Retrieval Performance

Although GELR consistently demonstrates improvements across all metrics compared to previous generative and discriminative baselines, there is curiosity about the impact of the knowledge re-

Method	Top-K	Recal@K	Acc@K
GELR w/o R.	0	0	86.4
GELR w. R.	1	15.9	86.5
	3	42.6	87.5
	5	59.9	87.9
	8	73.0	89.5
	10	79.6	88.4

Table 4: Retriever coverage and performance impact of GELR on the WikiDiverse test set.

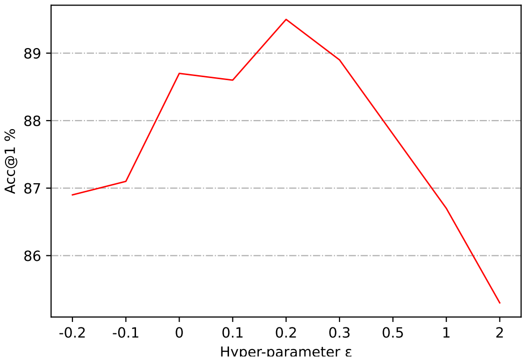

Figure 4: Effect of the hyper-parameter ϵ.

triever on the final prediction results. Moreover, it is worthwhile to explore whether knowledge retrieval can enhance the prediction of hard samples which requires fine-grained visual understanding.

Retriever Coverage. Tab. 4 reveals the relationship among the number of top retrieved entities K, the corresponding ground-truth entity Recall@K, and the Accuracy@K evaluated on WikiDiverse. It is observed that the performance of GELR generally improves as K increases, up to 8, which aligns with the intuition as with increasing K, retrieved results have more chance to cover gold entities, thereby increasing the likelihood for GELR to link the correct entity. However, as Recall@K increases, the number of incorrect entities also increases, posing a challenge for GELR to distinguish noisy information from the set of candidates. Therefore, we choose $K = 8$ as a compromise between the recall rate and the number of entities.

Effect of Adjustment of Predicted Logits. Tab. 2 shows a 0.8% improvement of adjustment of logits in terms of Accuracy@1 and F1 metrics, which enhance the capacity of GELR to extract useful information from the retrieved entities. In inference, we introduce a hyper-parameter ϵ to control the adjustment level of the predicted distribution, where the bigger ϵ, the farther away from the prediction based on the intrinsic knowledge of LLMs. We conduct experiments with various values of ϵ, and present the results in Fig. 4. We observe that a small value of ϵ ranging from 0.2 to 0.5 presents the best performance, since the top-K retrieved results may not always cover the gold entity.

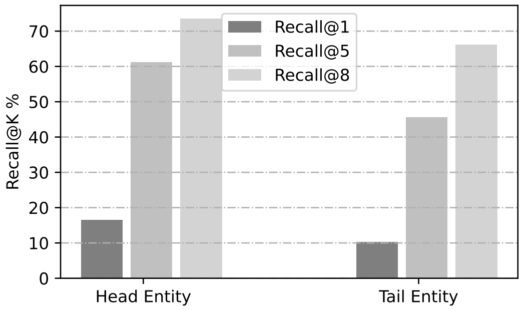

Figure 5: The Performance of Knowledge Retriever on the Head and Tail Entities. The shadow on the bar represents the proportion of correct predictions made when the gold entity is retrieved.

Retrievers on Hard Samples. As discussed in Sec. 5.1, removing the knowledge retriever leads to a greater performance decrease on tail entities compared to head entities. This suggests that the role and gain of the knowledge retriever differ for head and tail entities. Due to their lower frequency during LLM pre-training, long-tail entities are more challenging for LLM to generate, so these samples deserve separate investigations.

We present the retrieval performance in terms of Recall@K in Fig. 5, and we find that although the retrieval performance for head entities is higher than that for tail entities, the difference between the two is not significant. Moreover, the recall@8 rate for tail entities also reaches 66.2%. If GELR can effectively utilize the retrieved entity knowledge, it may be able to make accurate predictions.

Therefore, we calculate the samples where the Top-K candidate set contains the gold entity, and we find that 89.7% of them eventually received correct predictions. Additionally, upon analyzing misclassified samples by *GELR w/o R.*, We notice that 44.7% of these samples were correctly predicted due to the inclusion of the knowledge retriever.

6 Conclusion

In this paper, we propose a generative framework with knowledge retrievers for multi-modal entity linking tasks. This framework leverages knowledge retrievers to bridge the gap between multi-modal input and parametric knowledge of LLMs. We carry out primary experiments on three benchmarks. The experimental results show significant improvements of 2.2% to 6.5% across all metrics compared to strong baselines.

Limitation

One limitation of this work is that the multi-modal generative model is more computationally expensive than discriminative retrieval-based methods. This may suggest that our method is better suited for scenarios prioritizing performance over time efficiency. Also, our work is limited by the scale and domain of the dataset. Due to budget limitations and the unavailability of the GPT-4V model, our GELR has only been compared with the freely available multi-modal LLM, LLaVA-13B (Liu et al., 2023). Future research will expand to larger-scale and more open domains, integrating with downstream applications such as question-answering and search systems. Moreover, our work primarily relies on entity knowledge sourced from Wikipedia. In the future, we plan to explore tasks in specific domains, such as medical information (Zhang et al., 2024), scientific discoveries (Qi et al., 2023), and other vertical fields (Gao et al., 2024).

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2022ZD0160603).

References

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Anthony Chen, Pallavi Gudipati, Shayne Longpre, Xiao Ling, and Sameer Singh. 2021. Evaluating entity disambiguation and the role of popularity in retrieval-based NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021*, pages 4472–4485.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W. Cohen. 2022. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5558–5570. Association for Computational Linguistics.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 14948–14968. Association for Computational Linguistics.

Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 2013. Improving efficiency and accuracy in multilingual entity extraction. In *I-SEMANTICS 2013 - 9th International Conference on Semantic Systems, I-SEMANTICS '13, Graz, Austria, September 4-6, 2013*, pages 121–124. ACM.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Weibo Gao, Qi Liu, Zhenya Huang, Yu Yin, Haoyang Bi, Mu-Chun Wang, Jianhui Ma, Shijin Wang, and Yu Su. 2021. Rcd: Relation map driven cognitive diagnosis for intelligent education systems. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 501–510.

Weibo Gao, Qi Liu, Hao Wang, Linan Yue, Haoyang Bi, Yin Gu, Fangzhou Yao, Zheng Zhang, Xin Li, and Yuanjing He. 2024. Zero-1-to-3: Domain-level zero-shot cognitive diagnosis via one batch of early-bird students towards three diagnostic objectives. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8417–8426.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 541–550, Portland, Oregon, USA. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Shen Huang, Yuchen Zhai, Xinwei Long, Yong Jiang, Xiaobin Wang, Yin Zhang, and Pengjun Xie. 2022. DAMO-NLP at NLPCC-2022 task 2: Knowledge enhanced robust NER for speech entity linking. In *Natural Language Processing and Chinese Computing - 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24-25, 2022, Proceedings, Part II*, volume 13552 of *Lecture Notes in Computer Science*, pages 284–293. Springer.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix X. Yu, and Sanjiv Kumar. 2023a. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1774–1793.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023b. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Xinwei Long, Shuzi Niu, and Yucheng Li. 2020. Hierarchical region learning for nested named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 4788–4793. Association for Computational Linguistics.

Xinwei Long, Shuzi Niu, and Yucheng Li. 2021. Consistent inference for dialogue relation extraction. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3885–3891. ijcai.org.

Xinwei Long, Jiali Zeng, Fandong Meng, Zhiyuan Ma, Kaiyan Zhang, Bowen Zhou, and Jie Zhou. 2024. Generative multi-modal knowledge retrieval with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18733–18741.

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 2023, Long Beach, CA, USA, August 6-10, 2023*, pages 1583–1594. ACM.

Zhiyuan Ma, Zhihuan Yu, Jianjun Li, and Guohui Li. 2023. Hybridprompt: bridging language models and human priors in prompt tuning for visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13371–13379.

Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. 2023. Large language models are zero shot hypothesis proposers.

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,*, Proceedings of Machine Learning Research.

Jiahua Rao, Zifei Shan, Longpo Liu, Yao Zhou, and Yuedong Yang. 2023. Retrieval-based knowledge augmented vision language pre-training. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 5399–5409. ACM.

Senbao Shi, Zhenran Xu, Baotian Hu, and Min Zhang. 2023a. Generative multimodal entity linking. *CoRR*, abs/2306.12725.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen tau Yih. 2023b. Trusting your evidence: Hallucinate less with context-aware decoding.

Qingyi Si, Fandong Meng, Mingyu Zheng, Zheng Lin, Yuanxin Liu, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. 2022. Language prior is not the only shortcut: A benchmark for shortcut learning in VQA. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3698–3712. Association for Computational Linguistics.

Wen Sun, Yixing Fan, Jiafeng Guo, Ruqing Zhang, and Xueqi Cheng. 2022. Visual named entity linking: A new dataset and A baseline. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2403–2415. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. 2022a. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*.

Peng Wang, Jiangheng Wu, and Xiaohang Chen. 2022b. Multimodal entity linking with gated hierarchical fusion and contrastive training. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 938–948. ACM.

Sijia Wang, Alexander Hanbo Li, Henghui Zhu, Sheng Zhang, Pramuditha Perera, Chung-Wei Hang, Jie Ma, William Yang Wang, Zhiguo Wang, Vittorio Castelli, Bing Xiang, and Patrick Ng. 2023. Benchmarking diverse-modal entity linking with generative models. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 7841–7857. Association for Computational Linguistics.

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 *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4785–4797. Association for Computational Linguistics.

Zilin Xiao, Ming Gong, Jie Wu, Xingyao Zhang, Linjun Shou, and Daxin Jiang. 2023. Instructed language models with retrievers are powerful entity linkers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2267–2282. 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 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 3599–3608. ACM.

Jiali Zeng, Yang Liu, Jinsong Su, Yubin Ge, Yaojie Lu, Yongjing Yin, and Jiebo Luo. 2019. Iterative dual domain adaptation for neural machine translation.

Jiali Zeng, Jinsong Su, Huating Wen, Yang Liu, Jun Xie, Yongjing Yin, and Jianqiang Zhao. 2018. Multi-domain neural machine translation with word-level domain context discrimination. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 447–457, Brussels, Belgium. Association for Computational Linguistics.

Gongrui Zhang, Chenghuan Jiang, Zhongheng Guan, and Peng Wang. 2023. Multimodal entity linking with mixed fusion mechanism. In *Database Systems for Advanced Applications - 28th International Conference, DASFAA 2023, Tianjin, China, April 17-20, 2023, Proceedings, Part III*, volume 13945 of *Lecture Notes in Computer Science*, pages 607–622. Springer.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.*, 23(10):1499–1503.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, and et al. 2024. Ultramedical: Building specialized generalists in biomedicine.

Haiquan Zhao, Xuwu Wang, Shisong Chen, Zhixu Li, Xin Zheng, and Yanghua Xiao. 2024. OVEL: large language model as memory manager for online video entity linking. *CoRR*, abs/2403.01411.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592.

Tiantian Zhu, Yang Qin, Qingcai Chen, Xin Mu, Changlong Yu, and Yang Xiang. 2023b. Controllable contrastive generation for multilingual biomedical entity linking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5742–5753. Association for Computational Linguistics.

A Experimental Details

	WikiDiverse	WikiMEL	WikiPerson
Modality	VL → L	VL → L	V → L
Image Queries	7824	22136	48201
Mentions	16327	25846	52057
Text Length (avg.)	10.2	8.2	0.0
Mention (avg.)	2.1	1.2	1.1
KB Size	6.1M	5.9M	120K

Table 5: Statistics of Datasets.

For the multi-modal knowledge retriever, we set the number of layers in the alignment network as 2. The learning rate is set to 2e-6 for the visual part and 1e-5 for the textual part in CLIP and the batch size is set to 128. Images are resized to 224 × 224 pixels and the textual information is truncated to 77 tokens. For the WikiPerson dataset, we also adopt the MTCNN (Zhang et al., 2016) to extract face features following the setting of the baseline model (Sun et al., 2022). Additionally, We use Minigpt-4 (Zhu et al., 2023a) to obtain the image caption for each image.

For the generative module, we use learning rate 6e-5, batch size 12, and gradient accumulation 2 for up to 5 epochs. For the visual backbone, we adopt the CLIP and the Q-former (Li et al., 2023b) to encode the images and transform them into a short sequence. For the retrieved entities, we extract their entity name and summary as the textual entity features, and we extract the first image in their entries as the visual entity feature. During model training, we freeze the parameters of visual and textual backbones and employ low-rank adaptation (LORA) (Hu et al., 2021) for efficient adaptation. The optimization of the model is conducted using cross-entropy loss and the Kullback-Leibler Divergence in a teacher-forcing manner.