

Learning to Better Exploit Visual Evidence for Multimodal Entity Linking

Yingyao Ma^{1*}, Han Wang^{1*}, Yuning Wu¹, Yifan Xue¹,
Jiasong Wu^{1†}, Lotfi Senhadji², Huazhong Shu¹,

¹Key Laboratory of New Generation Artificial Intelligence
Technology and Its Interdisciplinary Applications,
Ministry of Education, Southeast University, Nanjing, China

²Univ Rennes, INSERM, LTSI - UMR 1099, France

{yyma, 220255064, 220256491, 230238541, jswu, shu.list}@seu.edu.cn
lotfi.senhadji@univ-rennes1.fr

Abstract

Multimodal Entity Linking (MEL) aims to link mentions in a given context to their corresponding entities in a knowledge base by leveraging both textual and visual information. Existing MEL methods still underexploit visual information in real-world scenarios. Most approaches directly encode raw image-text inputs or apply uniform enhancement, while lacking adaptive modeling of heterogeneous mention-relevant evidence. Moreover, they typically rely on single-layer visual representations or simple cross-modal interactions, failing to capture hierarchical visual semantics required for fine-grained disambiguation. To address these limitations, we propose a novel MEL framework termed EVEMEL. First, we introduce an adaptive evidence augmentation mechanism, which analyzes the visual condition of each sample and strengthens mention-relevant discriminative clues, providing more reliable contextual evidence for entity disambiguation. Subsequently, we develop a cross-modal modeling mechanism that treats multi-layer visual representations as different semantic granularities and dynamically fuses them through token-level routing, enabling cross-modal matching. In addition, a bidirectional gated intra-modal mechanism is designed to capture multi-granularity textual and visual matching signals. Experimental results on public benchmark datasets demonstrate that the proposed EVEMEL outperforms all state-of-the-art baselines. The code is publicly available at <https://github.com/hanwang86/EVEMEL>.

1 Introduction

Entity linking (EL) maps mentions in unstructured sources (e.g., social media, news, and web content) to the correct entities in a structured knowledge base (KB), which benefits numerous downstream tasks, including information extraction (Hoffart

et al., 2011), question answering (Yih et al., 2015) and semantic search (Hasibi et al., 2016). Traditional EL methods mainly rely on textual context for disambiguation. However, with the rapid growth of image-based online content, textual context alone often fails to resolve ambiguity. This limitation has motivated increasing research interest in Multimodal Entity Linking (MEL). Despite notable progress, existing MEL methods remain limited in their utilization of visual information, leading to two key challenges.

In real-world MEL scenarios, the quality of visual and textual inputs varies substantially across samples, and the discriminative clues associated with a mention may take different forms. For example, as illustrated in Figure 1, linking “Trump” requires focusing on the person region, linking “Bible” requires capturing fine-grained visual cues from the handheld object, and linking “St. John’s Episcopal Church” requires combining textual signs in the image with contextual information. Existing MEL methods typically encode the original text and image directly or apply uniform input enhancement, leaving the extraction of discriminative clues from heterogeneous image-text inputs to the subsequent multimodal interaction and matching process. This often results in suboptimal disambiguation performance. Therefore, it is necessary to introduce an adaptive mention evidence augmentation mechanism, which selectively strengthens mention-relevant discriminative information before multimodal encoding, providing more reliable contextual evidence for downstream disambiguation.

In addition, existing MEL methods typically rely on a single-layer visual representation or simple cross-modal attention for multimodal interaction, which limits their ability to fully exploit the semantic information embedded across different layers of the visual encoder. In MEL tasks, mention disambiguation may require visual evidence at different granularities: the disambiguation process is often

*These authors contribute equally to this work.

†Corresponding author.



Figure 1: One example of Multimodal Entity Linking.

mention-centered and depends simultaneously on local details, regional structures, and global concepts. Models need to capture both the overall semantic category of the image and subtle differences in local regions to distinguish among similar candidates. Therefore, multimodal interaction modules for MEL should be capable of modeling hierarchical visual semantics and selectively attending to the visual evidence most informative for distinguishing the target mention.

To address the above issues, we propose a novel MEL framework EVEMEL. First, we design an Agent-Driven Mention Augmenter (AMA), which performs mention-centered enhancement on visual inputs before multimodal encoding. Through sample analysis and adaptive tool invocation, AMA provides more reliable contextual evidence for the key discriminative clues in different samples. Second, we propose a Hierarchical Visual MoE based Cross-modal Matching module (HVMCM), which models multi-layer ViT representations as hierarchical visual experts and learns multi-layer fused multimodal semantic features through token-level MoE, thereby performing cross-modal mention-entity matching. In addition, the Bidirectional Gated Intra-modal Matching module (BiGIM) captures bidirectional multi-granularity matching signals within both textual and visual modalities. Finally, hard negative mining and a joint training strategy are employed to further enhance the model’s disambiguation capability. Our main contributions are summarized as follows:

- We design an Agent-Driven Mention Augmenter to adaptively enhance mention-centered evidence from heterogeneous multimodal inputs.
- We propose a hierarchical visual expert routing module that exploits multi-layer visual semantics for cross-modal mention-entity matching.
- Extensive experiments on public MEL datasets demonstrate the effectiveness of EVEMEL.

2 Related Work

Text-based Entity Linking maps mentions to KB entities by matching textual context with entity descriptions and is typically divided into local and global methods. Local models (Eshel et al., 2017; Francis-Landau et al., 2016; Yamada et al., 2016; Ran et al., 2023) focus on mention-candidate semantic similarity, evolving from CNN/RNN-based encoders to BERT-style encoders. Global models (Cao et al., 2018; Wu et al., 2020; Shen et al., 2022; Jin et al., 2022) jointly resolve multiple mentions within a document, enforcing entity coherence via relational or graph-based inference, but at higher computational cost.

Multimodal Entity Linking enhances text-based linking by incorporating visual signals to resolve ambiguities beyond textual context (Ma et al., 2025). Existing methods generally fall into two categories according to how they model multimodal interactions. Early models, such as DZMNED (Moon et al., 2018), JMEL (Adjali et al., 2020), and GHMFC (Wang et al., 2022a), encode global joint image-text representations and perform candidate scoring in a shared space, often yielding limited gains. Multi-level approaches, including MIMIC (Luo et al., 2023), MELOV (Sui et al., 2024), OT-MEL (Zhang et al., 2024), M³EL (Hu et al., 2025a), DCMS (Ma et al., 2026a), and VP-MEL (Lu et al., 2025), further model both local and global features, leading to more substantial improvements. Beyond representation-based matching, recent studies, such as UniMEL (Liu et al., 2024), FissFuse (Luo et al., 2024), ThinkLinker (Ma et al., 2026b) and MMoE (Hu et al., 2025b), have begun to leverage LLMs to enrich mention or entity semantics and perform candidate-level verification or reasoning. However, they still fall short of fully exploiting visual evidence, especially when such evidence is heterogeneous, noisy, or requires hierarchical semantic understanding.

3 Methodology

This section presents our proposed EVEMEL framework, as illustrated in Figure 2.

3.1 Problem Formulation

We formulate MEL as a multimodal mention-entity matching problem. Given an entity set $\mathcal{E}$ from a multimodal knowledge base, each entity is represented as $E = (e^n, e^a, e^v)$, where e^n , e^a , and e^v denote the entity name, textual attribute, and visual

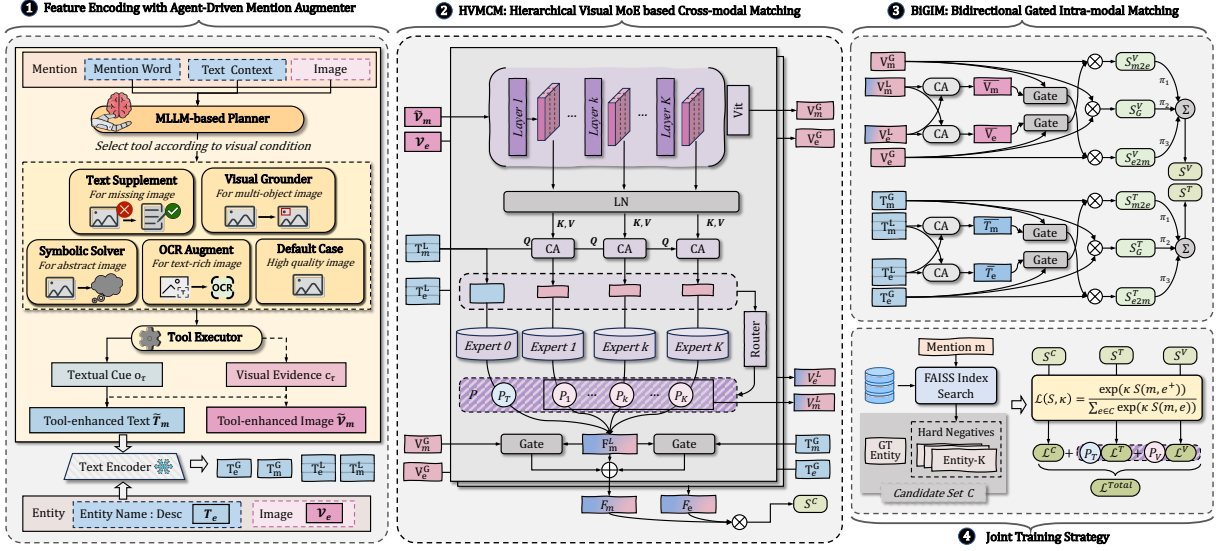


Figure 2: Overview of our proposed EVEMEL framework for Multimodal Entity Linking.

image, respectively. Each mention is represented as $M = (m^w, m^t, m^v)$, where m^w , m^t , and m^v denote the mention word, textual context, and associated image, respectively.

For each mention M , the MEL task aims to identify the ground-truth entity E^* by comparing the mention with all candidates and selecting the one with the highest similarity score $\text{Score}(\cdot)$:

$$E^* = \arg \max_{E^j \in \mathcal{E}} \text{Score}(M, E^j). \quad (1)$$

3.2 AMA: Agent-Driven Mention Augmenter

Images in MEL often provide heterogeneous and unreliable evidence: they may be missing, noisy, text-dependent, multi-object, or symbolic. Therefore, we formulate mention augmentation as a lightweight tool-using agent. For each mention, the agent first diagnoses the visual condition of the sample, then invokes a suitable specialist tool, and finally normalizes the tool output into mention-centered evidence for downstream matching.

Condition-Aware Planner. Given the mention word w_m , its textual context $\mathcal{T}_m$, and the associated image $\mathcal{V}_m$, the planner acts as a routing policy over a predefined tool set:

$$\tau = \mathcal{P}_{\text{MLLM}}(w_m, \mathcal{T}_m, \mathcal{V}_m), \quad \tau \in \mathcal{A}, \quad (2)$$

where $\mathcal{A} = \{TS, VG, SS, OCR, DC\}$ denotes the tool set, corresponding to **Text Supplement**, **Visual Grounder**, **Symbolic Solver**, **OCR Augment**, and **Default Case**, respectively. The planner first diagnoses the visual condition of the sample and

then determines how the visual evidence should be accessed or compensated for the current mention.

Tool-Specific Executor. After selecting the tool τ , the executor applies the corresponding operation and returns a textual cue, optionally with tool-specific visual evidence:

$$(o_\tau, c_\tau) = \text{Tool}_\tau(w_m, \mathcal{T}_m, \mathcal{V}_m), \quad (3)$$

where o_τ is a concise visual semantic cue, and c_τ denotes optional auxiliary visual evidence.

The executor performs the corresponding operation as follows: (1) *TS* is invoked for missing images, where the executor generates a brief visual description from the mention word and textual context. (2) *VG* is invoked for cluttered or multi-object images, where *GroundingDINO*¹ localizes the mention-related target and produces a mention-aware visual cue. (3) *OCR* is invoked for text-dependent images, where *PaddleOCR* (v3.4.1)² extracts image text and an LLM refines it with the mention context into an entity-oriented cue. (4) *SS* is invoked for abstract or symbolic images, where the executor interprets image-level symbolic visual semantics. (5) *DC* is invoked for clear single-subject images, where the executor summarizes the mention-related visual semantics directly.

The tool output is combined with the original mention word and textual context to obtain the enhanced textual input $\tilde{\mathcal{T}}_m$. If auxiliary visual evidence is produced, it is used to form the enhanced

¹<https://github.com/idea-research/groundingdino>

²<https://pypi.org/project/paddleocr/3.4.1>

visual input $\tilde{\mathcal{V}}_m$; otherwise, the original image is retained. This keeps the downstream matching model unchanged while injecting explicit visual semantics into the mention input.

3.3 HVMCM: Hierarchical Visual MoE based Cross-modal Matching

After constructing the tool-augmented mention input, we encode mentions and entities. For a mention, the textual input is the enhanced text $\tilde{T}_m$ from Section 3.2, and the visual input is the enhanced image $\tilde{\mathcal{V}}_m$ if auxiliary visual evidence is available; otherwise, the original image is used. For an entity, we concatenate its name and textual attributes as the textual input $\mathcal{T}_e$, and use its associated image as the visual input $\mathcal{V}_e$. The following procedure is described using the mention side as an example.

Specifically, for textual features, we use the CLIP text encoder to produce the local and global representations, denoted as T_m^L and T_m^G , respectively. For visual features, we extract hidden states from multiple intermediate ViT layers and index them as $\{V_m^k\}_{k=1}^{K_v}$, where K_v denotes the number of selected visual layers. The global visual representation is denoted as V_m^G . These selected multi-layer visual features are regarded as hierarchical visual experts, where lower layers mainly preserve local visual patterns, while higher layers provide stronger semantic cues.

Hierarchical Visual Expert Routing. HVMCM dynamically selects visual semantics from different ViT layers. We first apply layer normalization to the textual tokens and visual expert features. Then, each visual expert interacts with the textual tokens through text-guided multi-head cross-attention, where T_m^L serve as the query, and the V_m^k from each layer serve as the key and value.

$$C_m^k = \text{MHCA}(T_m^L, V_m^k, V_m^k), \quad (4)$$

where $k = 1, \dots, K_v$ and the resulting C_m^k denote the visual-enhanced textual representation produced by the k -th visual expert.

For each textual token, a linear router followed by a softmax is applied to the concatenated features of T_m^L and $\{C_m^k\}_{k=1}^{K_v}$, producing a token-level distribution for dynamically selecting between the textual residual expert and all visual experts:

$$\mathbf{P}_m = [P_m^T, P_m^1, \dots, P_m^k, \dots, P_m^{K_v}], \quad (5)$$

where P_m^T denotes the weight of the textual expert, and P_m^k denotes the weight of the k -th visual expert. The full routing weights are denoted as $\mathbf{P}_m$.

Token-level MoE Fusion. Each branch first produces a branch-wise fusion representation. The textual residual branch gives $F_m^T = \text{MLP}_T(T_m^L)$, while the k -th visual expert gives $F_m^k = \text{MLP}_k([T_m^L; C_m^k])$. For each textual token, the local fusion representation is computed by weighting the textual residual expert and all visual experts according to the token-level routing distribution:

$$F_m^L = P_m^T \odot F_m^T + \sum_{k=1}^{K_v} P_m^k \odot F_m^k. \quad (6)$$

This yields the cross-modal fused feature F_m^L .

We further reuse the visual-expert routing weights to obtain routed visual tokens for BiGIM by averaging them over textual tokens, i.e., $\bar{P}_m^k = 1/N_t \sum_{i=1}^{N_t} P_{m,i}^k$, where $P_{m,i}^k$ is the routing weight of the k -th visual expert for the i -th textual token.

Then, the routed visual tokens are computed as

$$V_m^L = \sum_{k=1}^{K_v} \frac{\bar{P}_m^k}{\sum_{k'=1}^{K_v} \bar{P}_m^{k'} + \epsilon} V_m^k. \quad (7)$$

The entity-side routed visual tokens V_e^L are obtained in the same manner.

Local-Global Gated Aggregation. After obtaining F_m^L , we take its EOT token as the local fused summary and denote it as $\bar{F}_m^L$. We then compute a local-text gate and a global-visual gate:

$$\begin{cases} \lambda_m^T = \sigma(W_T[T_m^G; \text{Proj}_L(\bar{F}_m^L)]), \\ \lambda_m^V = \sigma(W_V[T_m^G; \text{Proj}_V(V_m^G)]). \end{cases} \quad (8)$$

The two gated branches are defined as

$$\begin{cases} \tilde{F}_{m,T} = \lambda_m^T \odot \text{Proj}_L(\bar{F}_m^L) + (1 - \lambda_m^T) \odot T_m^G, \\ \tilde{F}_{m,V} = \lambda_m^V \odot \text{Proj}_V(V_m^G) + (1 - \lambda_m^V) \odot T_m^G. \end{cases} \quad (9)$$

Finally, we obtain the global mention representation

$$F_m = \text{Norm}\left(\frac{\tilde{F}_{m,T}^G + \tilde{F}_{m,V}^G}{2}\right), \quad (10)$$

where $\text{Norm}(\cdot)$ denotes ℓ_2 normalization. The same gated fusion is applied to candidate entities and hard negative entities, producing F_e and the corresponding hard-negative representations.

The cross-modal matching score between a mention and a candidate entity is computed as $S^C(m, e) = F_m \cdot F_e$, where F_m and F_e are normalized global fusion representations. HVMCM outputs S^C for main cross-modal alignment, together with the routed visual tokens V^L for the subsequent BiGIM module.

3.4 BiGIM: Bidirectional Gated Intra-modal Matching

While HVMCM performs cross-modal fusion within each mention or entity instance, BiGIM further models fine-grained mention–entity matching within the same modality. After HVMCM, each mention–entity pair contains textual features T_m^G , T_m^L , T_e^G , and T_e^L , and visual features V_m^G , V_m^L , V_e^G , and V_e^L . Since the textual and visual branches share the same formulation, we use the visual branch as an example.

Bidirectional Intra-modal Matching. We first project the global visual representations into a shared matching space and normalize them, denoted as $\hat{V}_m^G$ and $\hat{V}_e^G$. BiGIM then computes two directional local matching signals. In the entity-to-mention direction, entity visual tokens query mention visual tokens through cross-attention. We first pool the cross-attention output into a local interaction representation:

$$V_{e \rightarrow m}^L = \text{Pool}(\text{CrossAttn}(V_e^L, V_m^L, V_m^L)), \quad (11)$$

where V_e^L serves as the query, and V_m^L serves as the key and value. Then, we inject the mention-side global visual signal through a learnable residual gate:

$$V_{e \rightarrow m}^G = \text{Norm}(V_{e \rightarrow m}^L + \lambda_{e \rightarrow m}^V \hat{V}_m^G). \quad (12)$$

We compute the entity-to-mention score as $S_{e2m}^V = V_{e \rightarrow m}^G (\hat{V}_e^G)^\top$. The mention-to-entity score S_{m2e}^V is obtained analogously by swapping the mention and entity inputs. In addition, BiGIM computes a coarse global-to-global score $S_G^V = \hat{V}_m^G (\hat{V}_e^G)^\top$.

Gated Score Fusion. The final visual intra-modal matching score is obtained by a learned weighted fusion of the three scores:

$$S^V = \pi_G^V S_G^V + \pi_{e2m}^V S_{e2m}^V + \pi_{m2e}^V S_{m2e}^V, \quad (13)$$

where $\pi^V = [\pi_G^V, \pi_{e2m}^V, \pi_{m2e}^V]$ is produced by a softmax over learnable fusion parameters. The textual matching score S^T is computed in the same way using T_m^G , T_m^L , T_e^G , and T_e^L . The two modality-specific scores are used as fine-grained auxiliary signals during training, complementing the global cross-modal alignment learned by HVMCM.

3.5 Training Strategy

The previous modules produce three types of matching scores. HVMCM computes the global cross-modal score $S^C(m, e)$ between a mention

and a candidate entity. BiGIM further produces two fine-grained intra-modal scores, namely the textual score $S^T(m, e)$ and the visual score $S^V(m, e)$. During training, S^C is used for the main cross-modal alignment objective, while S^T and S^V are used as auxiliary supervision signals to strengthen textual and visual discrimination.

Hard Negative Mining. In-batch negatives are often easy negatives determined by batch sampling, which provides limited supervision for distinguishing semantically similar entities. We therefore construct an offline hard negative pool using CLIP-based text retrieval. Each mention text and entity text is encoded by the CLIP text encoder, and normalized entity embeddings are indexed with FAISS. For each mention, we retrieve the nearest entities and remove the gold entity. During training, we sample K_h hard negatives from this pool and combine them with in-batch entities for both the main and BiGIM-based auxiliary objectives.

Dynamic Auxiliary Weights. Different instances may rely on textual or visual information to different degrees. We therefore derive auxiliary weights from the token-level routing distribution $\mathbf{P} = [P^T, P^1, \dots, P^{K_v}]$ produced by HVMCM. After averaging $\mathbf{P}$ over the batch and textual tokens, we clip the averaged textual residual weight as ω_T and the summed visual expert weights as ω_V .

Joint Training Objective. For the i -th mention, the candidate set $\mathcal{C}_i$ contains its gold entity, in-batch entities, and sampled hard negatives. Given a matching score function $S(\cdot, \cdot)$ and a scaling factor κ , we define a generic contrastive objective:

$$\mathcal{L}(S, \kappa) = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(\kappa S(m_i, e_i^+))}{\sum_{e \in \mathcal{C}_i} \exp(\kappa S(m_i, e))}. \quad (14)$$

We instantiate this objective with different matching scores. The main loss is computed as $\mathcal{L}_{\text{main}} = \mathcal{L}(S^C, \gamma)$, where γ is a learnable scaling parameter. The textual and visual auxiliary losses are computed as $\mathcal{L}_T = \mathcal{L}(S^T, 1/\tau)$ and $\mathcal{L}_V = \mathcal{L}(S^V, 1/\tau)$, where τ is a fixed temperature. The final training objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{main}} + \omega_T \mathcal{L}_T + \omega_V \mathcal{L}_V. \quad (15)$$

Overall, the main loss learns global cross-modal mention–entity alignment, while the two auxiliary losses improve fine-grained textual and visual matching.

Methods	WikiDiverse				WikiMEL				RichpediaMEL			
	MRR	H@1	H@3	H@5	MRR	H@1	H@3	H@5	MRR	H@1	H@3	H@5
CLIP (Radford et al., 2021)	71.69	61.21	79.63	85.18	88.23	83.23	92.10	94.51	77.57	67.78	85.22	90.04
ViLT (Kim et al., 2021)	45.22	34.39	51.07	57.83	79.46	72.64	84.51	87.86	56.63	45.85	62.96	69.80
ALBEF (Li et al., 2021)	69.93	60.59	75.59	81.30	84.56	78.64	88.93	91.75	75.29	65.17	82.84	88.28
METER (Dou et al., 2022)	63.71	53.14	70.93	77.59	79.49	72.46	84.41	88.17	74.15	63.96	82.24	87.08
DZMNED (Moon et al., 2018)	67.59	56.90	75.34	81.41	84.97	78.82	90.02	92.62	76.63	68.16	82.94	87.33
GHMFC (Wang et al., 2022a)	70.99	60.27	79.40	84.74	83.36	76.55	88.40	92.01	80.76	72.92	86.85	90.60
MELOV (Sui et al., 2024)	76.57	67.32	83.69	87.54	92.32	88.91	95.61	96.58	88.80	84.14	92.81	94.89
OT-MEL (Zhang et al., 2024)	75.43	66.07	82.82	87.39	92.59	88.97	95.63	96.96	88.27	83.30	92.39	94.83
MIMIC [†] (Luo et al., 2023)	79.19	70.79	85.76	89.75	91.79	87.89	95.13	96.34	88.42	83.64	92.62	94.69
M ³ EL [†] (Hu et al., 2025a)	81.29	74.06	86.57	90.04	92.30	88.84	95.20	96.71	88.26	82.82	92.73	95.34
VP-MEL [†] (Lu et al., 2025)	77.81	68.62	85.03	89.27	90.61	86.63	93.65	95.53	86.78	81.50	90.79	93.57
FissFuse (Luo et al., 2024)	–	–	–	–	92.02	87.89	95.36	–	–	–	–	–
MMoE [†] (Hu et al., 2025b)	84.23	77.57	89.12	92.49	<u>93.75</u>	<u>90.77</u>	<u>96.27</u>	<u>97.41</u>	89.86	85.51	<u>93.43</u>	95.31
EVEMEL	83.49	77.82	88.31	90.52	93.55	90.71	95.94	96.90	89.89	85.43	93.35	<u>95.59</u>
EVEMEL [†] (w/o AMA)	86.44	81.33	90.86	92.88	93.18	89.88	95.94	97.14	89.96	85.74	93.37	95.31
EVEMEL [†]	89.48	85.66	92.54	94.47	94.59	92.22	96.40	97.50	91.17	87.56	93.99	95.73
Δ	+3.04	+4.33	+1.68	+1.59	+1.41	+2.34	+0.46	+0.36	+1.21	+1.82	+0.62	+0.42

Table 1: Performance comparison of all baselines on three MEL datasets. The best and second-best results are highlighted in **bold** and underline. Bold method names indicate LLM-enhanced methods. [†] indicates that entity-side descriptions are concatenated into the entity input. Δ denotes the improvement brought by module AMA. Except for MIMIC and VP-MEL, which are reproduced by ourselves, all other baseline results are taken from MMoE.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method on three public MEL benchmarks: WikiDiverse (Wang et al., 2022b), WikiMEL, and RichpediaMEL (Wang et al., 2022a). For fair comparison, we follow the original data splits provided by Luo et al. (2023). Details are given in Appendix A.

Metrics. We evaluate all methods using H@k and MRR. H@k measures whether the gold entity appears in the top-k candidates ($k \in \{1, 3, 5\}$), MRR denotes the average reciprocal rank of the gold entity. Higher values indicate better performance.

Baselines. We compare our EVEMEL with two groups of baselines. The first group includes vision-language pre-trained methods, including CLIP (Radford et al., 2021), ViLT (Kim et al., 2021), ALBEF (Li et al., 2021), and METER (Dou et al., 2022). The second group includes representative MEL methods, including DZMNED (Moon et al., 2018), GHMFC (Wang et al., 2022a), FissFuse (Luo et al., 2024), MELOV (Sui et al., 2024), OT-MEL (Zhang et al., 2024), MIMIC (Luo et al., 2023), VP-MEL (Lu et al., 2025), M³EL (Hu et al., 2025a), and MMoE (Hu et al., 2025b). We provide detailed descriptions of baselines in Appendix B.

Implementation Details. Our proposed method is implemented using the PyTorch framework and trained on a NVIDIA GTX A6000 GPU. We use "clip-vit-base-patch32" as the backbone encoder

and fine-tune it during training. The selected visual expert layers are $\{0, 3, 6, 9, 12\}$. The batch size is 128, the maximum number of epochs is 20, and early stopping is applied with patience 3 based on validation MRR. We use a fixed random seed of 43 for all experiments. Additional implementation details are provided in Appendix D.

4.2 Main Results

Table 1 summarizes the performance of all baselines on three datasets. Overall, MEL-specific methods generally outperform generic vision-language pretrained models. This indicates that general cross-modal representations alone are insufficient for accurate entity disambiguation, and MEL requires more explicit modeling of the correspondence between mentions and entities. Recent methods, such as MIMIC, M³EL, and MMoE, achieve stronger performance by introducing task-oriented multimodal interaction, fine-grained matching, or richer entity-side semantics. Interestingly, after incorporating entity description information, models generally achieve substantial performance improvements on the more challenging WikiDiverse dataset, suggesting that richer entity semantics are beneficial for handling diverse contexts.

Our final model achieves the best results across all datasets. The results show that even without the AMA module, EVEMEL already obtains competitive performance. This can be attributed to its effective exploitation of multi-granularity visual

Methods Metrics	WikiDiverse			WikiMEL			RichpediaMEL		
	MRR	H@1	H@3	MRR	H@1	H@3	MRR	H@1	H@3
w/o HVMCM	73.15 $\downarrow 16.33$	66.07 $\downarrow 19.59$	78.01 $\downarrow 14.53$	91.18 $\downarrow 3.41$	87.42 $\downarrow 4.80$	94.35 $\downarrow 2.05$	86.33 $\downarrow 4.84$	82.06 $\downarrow 5.50$	89.02 $\downarrow 4.97$
w/o Weight P	82.41 $\downarrow 7.07$	75.60 $\downarrow 10.06$	88.16 $\downarrow 4.38$	93.92 $\downarrow 0.67$	91.02 $\downarrow 1.20$	96.28 $\downarrow 0.12$	88.69 $\downarrow 2.48$	84.50 $\downarrow 3.06$	91.89 $\downarrow 2.10$
w/o Hard Negatives	85.43 $\downarrow 4.05$	79.56 $\downarrow 6.10$	90.47 $\downarrow 2.07$	93.88 $\downarrow 0.71$	91.00 $\downarrow 1.22$	96.17 $\downarrow 0.23$	88.16 $\downarrow 3.01$	83.02 $\downarrow 4.54$	92.17 $\downarrow 1.82$
w/o Vision-BiGIM	89.73 $\uparrow 0.25$	85.85 $\uparrow 0.19$	92.64 $\uparrow 0.10$	92.23 $\downarrow 2.36$	88.93 $\downarrow 3.29$	94.78 $\downarrow 1.62$	89.43 $\downarrow 1.74$	85.09 $\downarrow 2.47$	92.90 $\downarrow 1.09$
w/o Text-BiGIM	89.43 $\downarrow 0.05$	85.23 $\downarrow 0.43$	92.28 $\downarrow 0.26$	95.25 $\uparrow 0.66$	93.15 $\uparrow 0.93$	97.16 $\uparrow 0.76$	91.64 $\uparrow 0.47$	87.84 $\uparrow 0.28$	94.64 $\uparrow 0.65$
w/o BiGIM	88.82 $\downarrow 0.66$	84.22 $\downarrow 1.44$	92.69 $\downarrow 0.15$	93.23 $\downarrow 1.36$	90.19 $\downarrow 2.03$	95.88 $\downarrow 0.52$	89.90 $\downarrow 1.27$	85.65 $\downarrow 1.91$	93.12 $\downarrow 0.87$
EVEMEL	89.48	85.66	92.54	94.59	92.22	96.40	91.17	87.56	93.99

Table 2: Ablation study on key components of EVEMEL.

information and the carefully designed hierarchical cross-modal interaction mechanism. After further incorporating AMA, the model achieves additional improvements. Compared with the strongest baseline MMoE, EVEMEL improves H@1 by 8.09, 1.45, and 2.05 points on three datasets respectively. The larger improvement on WikiDiverse can be explained by the fact that, when visual inputs are more diverse, noisy, or require additional interpretation, AMA provides useful mention-side semantic cues that complement the hierarchical multimodal matching mechanism.

4.3 Ablation Study

Model Component Ablation. Table 2 reports the component ablation results. First, removing HVMCM causes substantial performance drops on all datasets, indicating that this mechanism is the core component of our method. The more significant drop on WikiDiverse further suggests that hierarchical visual expert routing is particularly important for diverse multimodal contexts. Removing the dynamic auxiliary weights **P** also leads to notable performance degradation, demonstrating the importance of adaptive loss weighting. Second, removing the entire BiGIM module decreases performance across all three datasets, showing that bidirectional intra-modal matching serves as a useful complement. However, its textual and visual branches provide effective but dataset-dependent supervision signals. When textual semantics are already sufficiently informative, additional textual supervision may become redundant. For the visual branch, it improves discrimination when visual clues are relatively clear, whereas on WikiDiverse, the visual supervision signal is less stable. Finally, removing hard negatives leads to performance degradation, indicating that challenging negative samples help distinguish semantically similar entities, especially for datasets with larger entity sets.

Tool Ablation. Table 3 reports the ablation results of the AMA. Removing Default Case leads to relatively large drops on both datasets, indicating

Dataset	Method	MRR	H@1	H@3
WikiDiverse	w/o Default Case	87.56 $\downarrow 1.92$	82.68 $\downarrow 2.98$	91.58 $\downarrow 0.96$
	w/o Symbolic Solver	89.10 $\downarrow 0.38$	85.23 $\downarrow 0.43$	92.25 $\downarrow 0.29$
	w/o Visual Grounder	88.72 $\downarrow 0.76$	84.79 $\downarrow 0.87$	91.48 $\downarrow 1.06$
	w/o OCR Augment	89.38 $\downarrow 0.10$	85.43 $\downarrow 0.23$	92.45 $\downarrow 0.09$
	w/o Text Supplement	88.42 $\downarrow 1.06$	83.97 $\downarrow 1.69$	91.96 $\downarrow 0.58$
	EVEMEL	89.48	85.66	92.54
WikiMEL	w/o Default Case	93.53 $\downarrow 1.06$	90.65 $\downarrow 1.57$	95.99 $\downarrow 0.41$
	w/o Symbolic Solver	94.39 $\downarrow 0.20$	91.89 $\downarrow 0.33$	96.21 $\downarrow 0.19$
	w/o Visual Grounder	94.36 $\downarrow 0.23$	91.78 $\downarrow 0.44$	96.18 $\downarrow 0.22$
	w/o OCR Augment	94.51 $\downarrow 0.08$	92.02 $\downarrow 0.20$	96.34 $\downarrow 0.06$
	w/o Text Supplement	94.59 $\downarrow 0.00$	92.22 $\downarrow 0.00$	96.40 $\downarrow 0.00$
	EVEMEL	94.59	92.22	96.40

Table 3: Ablation study of different AMA tools.

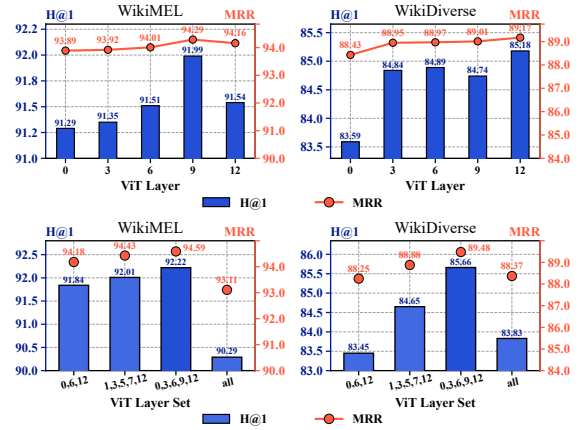


Figure 3: Sensitivity analysis of ViT layer selection.

that explicit visual semantic descriptions are still useful even for clear single-subject images. The effects of other tools are more dataset-dependent. Text Supplement has no effect on WikiMEL because there are no missing images, but its removal causes a notable performance drop on WikiDiverse. Visual Grounder also plays an important role in localizing mention-related regions in multi-object or noisy images. Symbolic Solver and OCR Augment bring smaller but consistent gains due to the limited number of corresponding samples. Overall, AMA improves MEL task by selecting suitable tools across visual conditions, rather than relying on a single tool to enhance all samples.

4.4 Discussion and Analysis

Sensitivity to ViT Layers. Figure 3 analyzes the impact of different ViT layer selections. For single-

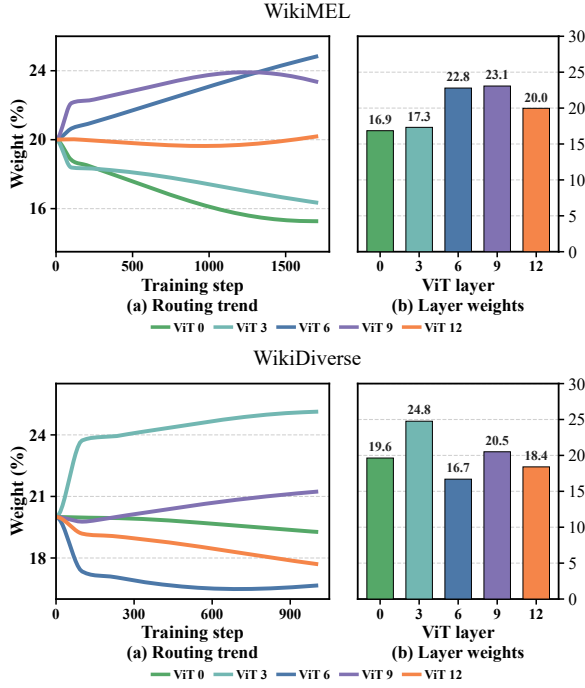


Figure 4: Training Dynamics of Visual Expert Weights.

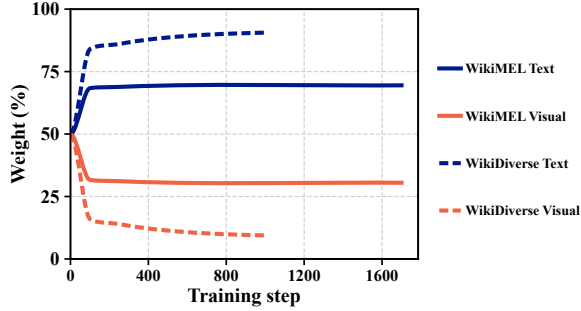


Figure 5: Training Dynamics of ω_T and ω_V .

layer, both datasets benefit more from high-level visual representations than from the lower layer, with the best results achieved at layer 9 on WikiMEL and layer 12 on WikiDiverse. Multi-layer generally outperforms single-layer, but using all hidden states does not yield the best performance. Instead, the five-layer setting $\{0, 3, 6, 9, 12\}$ achieves the best results on both datasets. These results indicate that selecting representative layers from different depths provides a better balance between low-level visual details and high-level semantic cues.

Router Behavior. Figure 4 shows that the routing patterns change during training and gradually become stable. Moreover, HVMCM learns different routing patterns according to the characteristics of different datasets. The router assigns higher weights to middle and high ViT layers on WikiMEL, suggesting that semantic-level visual

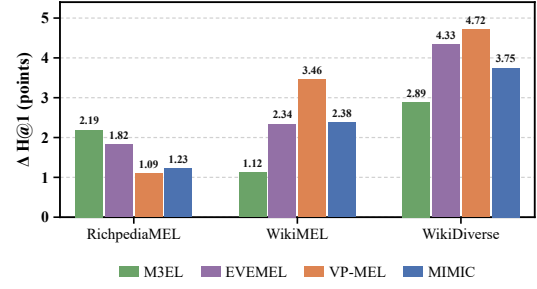


Figure 6: Transferability of AMA across different MEL backbones.

cues are more useful for this dataset. In contrast, WikiDiverse assigns the largest weight to layer 3, indicating a stronger reliance on relatively fine-grained visual features. Figure 5 further illustrates the evolution of the loss weights ω_T and ω_V during training. Overall, the router assigns higher weights to the textual expert, indicating that text remains the dominant signal in MEL. In addition, WikiMEL maintains relatively higher visual weights due to the higher quality of its mention images, while WikiDiverse reduces its reliance on visual signals. Therefore, the learned weights serve as a bridge between HVMCM and BiGIM, allowing the routing preferences to guide modality-specific auxiliary supervision.

Transferability of AMA To examine whether AMA can generalize to different MEL methods, we apply it to three representative existing backbones, including M3EL, VP-MEL, and MIMIC. As shown in Figure 6, AMA consistently improves all models across the three datasets, demonstrating that it can serve as a plug-and-play mention augmentation module. On average, AMA improves H@1 by 2.61 points across the combinations of models and datasets. The gains are particularly notable on WikiDiverse, suggesting that AMA is especially beneficial under visually diverse or noisy scenarios.

5 Conclusion

In this work, we propose EVEMEL, a novel framework for MEL. EVEMEL introduces an adaptive mention evidence augmentation mechanism to strengthen mention-relevant visual clues before encoding, and further exploits hierarchical visual semantics through multi-layer expert routing for cross-modal matching. Extensive experiments on public MEL datasets verify the effectiveness and interpretability of EVEMEL.

Limitations

EVEMEL improves MEL by adaptively enhancing mention-relevant visual evidence and exploiting hierarchical visual semantics. However, AMA relies on external MLLMs and visual tools to construct enhanced mention evidence. Although this process provides richer semantic complements, it also introduces additional offline preprocessing costs. Therefore, improving the processing efficiency while maintaining the reliability of augmented evidence will be an important direction for future optimization. In addition, the current method mainly focuses on textual and visual modalities, and has not fully considered the varying modality requirements across different samples. Future work can explore more flexible modality selection mechanisms, allowing the model to adaptively decide whether to invoke external tools or use specific modalities according to sample complexity and modality quality. Meanwhile, extending MEL to more complex scenarios involving video, audio, and other modalities is also a promising direction for further research.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2025YFE0150500, in part by the National Natural Science Foundation of China under Grant 62476055, in part by the innovation project of Jiangsu Province under Grants BZ2023042, BY2022564, and in part by the Fundamental Research Funds for the Central Universities under Grant 2242024k30035. The authors would like to thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations in this paper.

References

- Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020. Multimodal entity linking for tweets. In *Advances in Information Retrieval*, pages 463–478.
- Yixin Cao, Lei Hou, Juanzi Li, and Zhiyuan Liu. 2018. Neural collective entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 675–686.
- Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. 2022. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18145–18155.
- Yotam Eshel, Noam Cohen, Kira Radinsky, Shaul Markovitch, Ikuya Yamada, and Omer Levy. 2017. Named entity disambiguation for noisy text. In *Proceedings of the 21st Conference on Computational Natural Language Learning*, pages 58–68.
- Matthew Francis-Landau, Greg Durrett, and Dan Klein. 2016. Capturing semantic similarity for entity linking with convolutional neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1256–1261.
- Faegheh Hasibi, Krisztian Balog, and Svein Erik Bratsberg. 2016. Exploiting entity linking in queries for entity retrieval. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, pages 209–218.
- Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstena, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792.
- Zhiwei Hu, Victor Gutiérrez-Basulto, Ru Li, and Jeff Z. Pan. 2025a. Multi-level matching network for multimodal entity linking. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Volume 1*, pages 508–519.
- Zhiwei Hu, Víctor Gutiérrez-Basulto, Zhiliang Xiang, Ru Li, and Jeff Z. Pan. 2025b. Multi-level mixture of experts for multimodal entity linking. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Volume 2*, pages 979–990.
- Fengmei Jin, Wen Hua, Thomas Zhou, Jiajie Xu, Matteo Francia, Maria E. Orlowska, and Xiaofang Zhou. 2022. Trajectory-based spatiotemporal entity linking. *IEEE Transactions on Knowledge and Data Engineering*, 34(9):4499–4513.
- Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5583–5594.
- Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*.

- Qi Liu, Yongyi He, Tong Xu, Defu Lian, Che Liu, Zhi Zheng, and Enhong Chen. 2024. Unimel: A unified framework for multimodal entity linking with large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1909–1919.
- Jinghui Lu, Yang Xu, Bin Liang, Ruifeng Xu, Lin Gui, and Yulan He. 2025. Vp-mel: Visual prompts guided multimodal entity linking. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17122–17137.
- Pengfei Luo, Tong Xu, Che Liu, Suojuan Zhang, Linli Xu, Minglei Li, and Enhong Chen. 2024. Bridging gaps in content and knowledge for multimodal entity linking. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9311–9320.
- 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*, pages 1583–1594.
- Yingyao Ma, Yifan Xue, Jiasong Wu, Lotfi Senhadji, Huazhong Shu, and Jian Yang. 2025. Multimodal entity linking with dynamic modality selection and interactive prompt learning. *IEEE Transactions on Knowledge and Data Engineering*, 37(9):5467–5480.
- Yingyao Ma, Yuanyuan Zhou, Yifan Xue, Wanqiang Cai, Lotfi Senhadji, Huazhong Shu, and Jiasong Wu. 2026a. Video entity linking in complex multimodal scenarios: High-quality benchmark datasets and a general baseline framework. *Information Fusion*, page 104610.
- Yingyao Ma, Yuanyuan Zhou, Congyu Zhang, Yi Yuan, Jiasong Wu, Lotfi Senhadji, and Huazhong Shu. 2026b. ThinkLinker: From low-rank interaction to knowledge-aware verification for multimodal entity linking. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 24926–24942, San Diego, California, United States. Association for Computational Linguistics.
- Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity disambiguation for noisy social media posts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2000–2008.
- 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*, pages 8748–8763.
- Chenwei Ran, Wei Shen, Jianbo Gao, Yuhan Li, Jianyong Wang, and Yantao Jia. 2023. Learning entity linking features for emerging entities. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7088–7102.
- Wei Shen, Yuwei Yin, Yang Yang, Jiawei Han, Jianyong Wang, and Xiaojie Yuan. 2022. Toward tweet entity linking with heterogeneous information networks. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):6003–6017.
- 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 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 816–826.
- Peng Wang, Jiangheng Wu, and Xiaohang Chen. 2022a. Multimodal entity linking with gated hierarchical fusion and contrastive training. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 938–948.
- Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. 2022b. WikiDiverse: A multimodal entity linking dataset with diversified contextual topics and entity types. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 4785–4797.
- Junshuang Wu, Richong Zhang, Yongyi Mao, Hongyu Guo, Masoumeh Soflaei, and Jinpeng Huai. 2020. Dynamic graph convolutional networks for entity linking. In *Proceedings of The Web Conference 2020*, pages 1149–1159.
- Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. Joint learning of the embedding of words and entities for named entity disambiguation. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 250–259.
- Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. 2015. Semantic parsing via staged query graph generation: Question answering with knowledge base. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1321–1331.
- Zefeng Zhang, Jiawei Sheng, Chuang Zhang, Liangyunzhi Liangyunzhi, Wenyuan Zhang, Siqi Wang, and Tingwen Liu. 2024. Optimal transport guided correlation assignment for multimodal entity linking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 4103–4117.

A Statistics of Datasets

Table 4: Statistics of the three datasets. “With images” denotes mentions or entities associated with images.

Statistic		Wiki Diverse	Wiki MEL	Richpedia MEL
Sentences		7,405	22,070	17,724
Mentions	Total	15,093	25,846	17,805
	With images	6,697	22,136	15,853
	Train	11,351	18,092	12,463
	Valid	1,664	2,585	1,780
	Test	2,078	5,169	3,562
KB entities	Total	132,460	109,976	160,935
	With images	67,309	67,195	86,769

In the experiments, we validate the effectiveness of our proposed method on three datasets: RichpediaMEL, WikiMEL, and WikiDiverse. RichpediaMEL (Wang et al., 2022a) contains more than 17,000 multimodal samples, with entities gathered from Richpedia and corresponding multimodal information retrieved from Wikipedia. WikiMEL (Wang et al., 2022a) comprises over 22,000 multimodal samples, constructed by collecting entities from Wikidata and subsequently extracting textual and visual descriptions for each entity from Wikipedia. WikiDiverse (Wang et al., 2022b) includes over 8,000 diverse context topics and entity types sourced from Wikinews, using Wikipedia, which hosts more than 16 million entities, as the corresponding knowledge base.

Table 4 summarizes the statistics of the three datasets. Their differences in scale, KB size, and image availability provide diverse evaluation scenarios for MEL. For fair comparison, we follow the original train/validation/test splits of all datasets.

B Descriptions of Baselines

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

CLIP (Radford et al., 2021) is a contrastive vision–language model trained on large-scale image–text pairs, providing a shared embedding space for text–image similarity.

VILT (Kim et al., 2021) is a vision-language pretrained model that directly feeds image patch embeddings and text embeddings into Transformer, without using convolutional or region features.

ALBEF (Li et al., 2021) aligns visual and textual representations before cross-modal fusion. It combines contrastive learning, image–text match-

ing, masked language modeling, and momentum distillation to learn robust representations.

METER (Dou et al., 2022) is an end-to-end vision–language Transformer framework that systematically studies key design choices in multimodal pretraining.

DZMNED (Moon et al., 2018) is an early MEL model that integrates short textual context and associated images via attention mechanisms for entity disambiguation in noisy social media posts.

GHMFC (Wang et al., 2022a) is a MEL model that mines fine-grained cross-modal relations with Transformer-based encoders and employs gated fusion and contrastive training to obtain more informative multimodal entity representations.

MELOV (Sui et al., 2024) improves multimodal entity linking by optimizing visual features in a latent space, aiming to reduce the noise introduced by raw visual representations and enhance visual evidence for entity disambiguation.

OT-MEL (Zhang et al., 2024) introduces optimal transport to guide correlation assignment between multimodal mention and entity features, enabling more structured cross-modal matching.

MIMIC (Luo et al., 2023) is a multi-granularity interaction framework that explicitly models feature interactions between mentions and entities across modalities using several cross-modal units at local and global levels.

MMoE (Hu et al., 2025b) proposes a multi-level mixture-of-experts architecture, utilizing dynamically routed experts to capture complex heterogeneous cross-modal interactions.

M³EL (Hu et al., 2025a) is a multi-level matching network that extracts modality-specific representations and employs intra-modal contrastive learning to align visual and textual features.

FissFuse (Luo et al., 2024) adopts a fission-fusion dual-branch architecture to model modality-specific features, optionally augmented with an LLM-based re-ranking stage.

VP-MEL (Lu et al., 2025) introduces a visual-prompt-guided MEL framework that leverages visual prompts to guide textual cues and uses vision–language models to supplement latent textual information for precise entity linking.

C Detailed Prompts for AMA

The detailed AMA prompts for the planner and tool executors are shown in Figure 7a and Figure 7b, respectively.

Role: You are the Visual Strategy Controller (Planner) of a multimodal entity linking system. Your core task is to select the best visual feature extraction strategy for downstream models based on the input Mention, Sentence, and Image Metadata, using internal reasoning, and output a standardized JSON tool-call instruction.

Tools Description: Select tools strictly according to the definitions and scenarios below.

1. Text_Supplement

- **Scenario:** Missing Image.
- **Trigger:** has_image is False.
- **Function:** Use the LLM's world knowledge to infer the entity's standard visual appearance from the text context.
- **Constraint:** If has_image is True, never call Text_Supplement.

2. Visual_Grounder

- **Scenario:** Multi-Entity Grounding and concrete reference.
- **Trigger:** The Mention refers to a specific physical entity, such as a person or object in the image, and the image has a cluttered background, multiple similar objects, or requires precise localization.
 1. Cluttered background needs noise removal.
 2. Multi-subject disambiguation: If the image contains multiple salient subjects, such as a group photo, two people, or many objects, and the Mention refers to only one specific individual/entity, crop the target to remove interference from other subjects.
- **Constraint:** Never use this tool for place names, cities, countries, administrative regions, or abstract locations. If the Mention is a place name, skip this tool.
- **Function:** Call a detection model to physically crop the target and remove background noise.

3. Symbolic_Solver

- **Scenario:** Abstract concepts, modifiers, and non-literal correspondence.
- **Trigger:**
 1. The Mention is an abstract noun, such as "Justice" or "Economy".
 2. The Mention is an adjective, place-based adjective, or style, such as "American" or "Pichileminian".
 3. Containment and non-iconic cases: The Mention is a broad region or place name, such as "Westminster" or "Japan", while the image only shows an ordinary object, building, or map from that region, such as a bank, a bowl of ramen, or a city/country map, and the image cannot strongly represent the place. If the image is not an iconic landmark of the place, such as the Eiffel Tower for Paris, treat it as symbolic/contextual rather than the entity itself.
- **Function:** Preserve the full-image context and mark it as symbolic mode, guiding the downstream model to focus on overall atmosphere or cultural context instead of forcing alignment with specific pixels.

4. OCR_Augment

- **Scenario:** Text-rich image.
- **Trigger:** The Mention is a book title, movie title, plaque, or organization name, and the image clearly contains text, such as a poster, cover, or document.
- **Function:** Extract text from the image as enhanced features.
- **Constraint:** If the image, Mention, and Sentence are strongly related, and OCR_Augment is selected due to visible text, preserve useful visual information as well by setting "auxiliary_info": "image reserved", because visual shape is usually more robust than OCR. If the relationship is weak, leave "auxiliary_info" empty and do not preserve visual information.

5. Default_Case

- **Scenario:** Single-object close-up and dual augmentation.
- **Trigger:** The image is clear, contains only the target subject, and has no obvious ambiguity.
- **Function:** Preserve the original visual features and generate a detailed semantic description, providing dual-modal redundancy.
- **Constraint:** If the image contains two or more salient subjects, such as two people sitting together, and the Mention refers to only one of them, never use this tool. Use Visual_Grounder instead.

Decision Logic Priority

1. Check image availability: if no image -> Text_Supplement.
2. Check abstractness: if the Mention is abstract and the image-text relation is non-literal -> Symbolic_Solver.
3. Check text dependency: if entity recognition heavily depends on text in the image, such as a book title -> OCR_Augment.
4. Check background interference: if the background is cluttered or ambiguous -> Visual_Grounder.
5. Default: if the image is clear and has a single subject -> Default_Case.

Important: You must select exactly one tool. Do not output no-tool calls.

Output Format

Output must be strict JSON containing:

- "reasoning": an empty string
- "tool_call": the selected tool and its parameters

Although "reasoning" must be empty, you should still perform the reasoning internally.

(a) Planner prompt.

# Role Your task is to identify a specific entity in a multimodal context, using both image and text, and provide a precise encyclopedic definition for disambiguation. # Input Data Entity Name: "{mention_name}" Context: "{mention_context}" Image: [Attached Image] # Task Based on the visual clues and the provided text context, generate a single-sentence description defining who or what the entity is. # Constraints - Focus only on the target entity itself. Do not describe the whole image, background, environment, or camera angle. - The description must be factual, like the opening sentence of a Wikipedia entry. - Use internal world knowledge to add distinguishing background information, such as nationality, occupation, location, or major achievements. - Do not mention meaningless appearance details, such as clothing color. Focus on semantic identity or real-world status. - The output must be purely objective. Do not use any wording that implies the existence of an image, such as "as seen in the image," "shown in the picture," or "the photo displays." - Use English. - Keep the summary under 25 words. # Output Format [Your description] Default_Case	# Role Your task is to identify an abstract entity or broad concept in a multimodal context and provide a precise encyclopedic definition for disambiguation. # Input Data Entity Name: "{mention_name}" Context: "{mention_context}" Image: [Attached Image] # Task Generate a single-sentence description defining what the entity is. # Constraints - Treat the image only as a symbol or partial representation of the entity, such as a chart for "economy" or a street for "city." Do not equate the entity with specific objects in the image. - Infer the deeper meaning, cultural context, or overall atmosphere conveyed by the image, and combine it with the text context and world knowledge to define the entity. - Focus only on the entity itself. Do not describe the literal image content. - The output must be purely objective. Do not use any wording that implies the existence of an image, such as "as seen in the image," "shown in the picture," or "the photo displays." - Use English. - Keep the summary under 25 words. # Output Format [Your description] Symbolic_Solver	# Role Your task is to precisely identify a target entity in a complex multimodal context with cluttered backgrounds or multiple subjects, and provide an encyclopedic definition for disambiguation. # Input Data Entity Name: "{mention_name}" Context: "{mention_context}" Image: [Attached Image] # Task Use the visual clues and context to generate a single-sentence description defining who or what the entity is. # Constraints - Ignore background noise and unrelated subjects. Align strictly with the text context and focus only on the target entity. - Do not describe the surroundings, irrelevant people, occlusions, or the image's overall composition. - Use a factual, encyclopedic style, like the first sentence of a Wikipedia entry. - Use internal knowledge to provide key background information, such as social identity or functional role. - Do not mention meaningless appearance details, such as clothing color. Focus on semantic identity or real-world status. - The output must be purely objective. Do not use wording that implies the existence of an image, such as "as seen in the image" - Use English. - Keep the summary under 25 words. # Output Format [Your description] Visual_Grounder	# Role Your task is to identify the target entity using context, extracted OCR text, and visual clues, then provide a precise text-only encyclopedic definition for entity linking. # Input Data Entity Name: "{mention_name}" Context: "{mention_context}" Raw OCR Text: "{ocr_extracted_text}" Image: [Attached Image] # Task Treat OCR text and visual features, such as logos or distinctive typography, only as clues to confirm the entity's real-world identity. Then use world knowledge to generate a single objective sentence defining what or who the Entity Name refers to. # Constraints - Focus only on the input Entity Name. OCR text is only supporting evidence. For example, if the entity is "TED", define TED itself, not a specific TEDx event found in OCR. - The output must be a purely objective factual definition. Do not describe the image scene, text carrier, layout, or position. - Filter OCR noise such as typos, fragments, irrelevant background text. - Use English. - Keep the summary under 25 words. # Output Format [Your description] OCR_Augment	# Role Your task is to provide an objective encyclopedic or typical visual description of the target entity based only on text context, helping downstream entity linking when no image is available. # Input Data Entity Name: "{mention_name}" Context: "{mention_context}" # Task Infer the real identity of the Entity Name in the sentence, then use world knowledge to provide one objective encyclopedic definition. Emphasize typical static features or visual appearance. # Constraints - The output must be a general encyclopedic definition. For example, if the entity is a college sports team, describe it as a team representing a higher education institution, usually wearing uniforms with the university logo and colors. - Do not describe specific events, actions, or states from the current context. - Ignore irrelevant entities and details. Focus only on the essential nature of the Entity Name. - Use English. - Keep the summary under 30 words. # Output Format [Your description] Text_Supplement
--	---	---	---	---

(b) Tool executor prompt.

Figure 7: Prompt templates used in AMA.

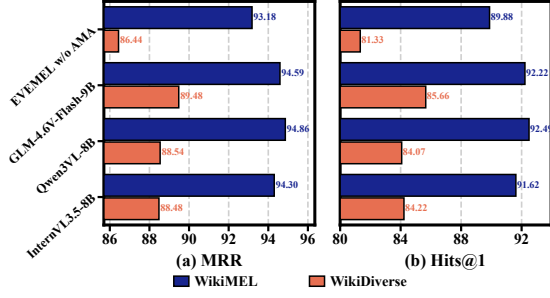


Figure 8: Effect of various MLLM backbones for AMA.

D Implementation Details

We fine-tune Hugging Face CLIP-ViT-B/32 as the backbone encoder. Images are resized to 224×224 , the maximum text length is set to 60, and missing images are replaced with a black placeholder. The hidden dimensions of the text representation, fusion representation, and BiGIM matching space are set to 512, 512, and 256, respectively. Cross-attention layers use 8 heads with a dropout rate of 0.1.

We optimize the model with AdamW, using learning rates of 3×10^{-5} for the backbone and 3×10^{-4} for newly introduced modules, with a weight decay of 1×10^{-4} . The contrastive temperature is set to 0.07, and the learnable logit scale is initialized as $\log(1/0.07)$.

Hard negatives are mined offline using a FAISS index built from CLIP text embeddings. For each mention, we retrieve the top-50 nearest entities, remove the gold entity by QID, and keep the top-10 as the hard negative pool, from which one negative is sampled during training. AMA outputs are generated offline using only mention-side information. GLM-4.6V-Flash is used for planning and cue generation, while Grounding DINO and PaddleOCR implement Visual Grounding and OCR Augment, respectively.

E Effect of MLLM Backbones.

Figure 8 compares different MLLM backbones used in AMA. As shown in the figure, GLM-4.6V-Flash achieves the best performance, demonstrating that the quality of generated visual cues can affect the final entity linking results. Moreover, incorporating AMA with different backbones consistently yields substantial performance improvements over EVEMEL without AMA. This indicates that our method is not highly sensitive to model scale, and that the AMA module can deliver stable improvements even with relatively lightweight LLMs.

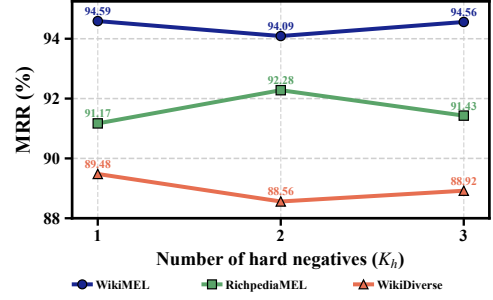


Figure 9: Effect of the number of hard negatives.

F Effect of Hard Negatives.

Figure 9 shows that increasing the number of hard negatives does not consistently improve performance. While $K_h = 2$ improves RichpediaMEL, larger K_h brings limited gains or even drops on WikiMEL and WikiDiverse. This suggests that hard negatives provide challenging supervision, but excessive hard negatives may introduce noisy candidates and increase computational cost. We therefore use a small number of hard negatives in experiments to balance effectiveness and efficiency.

G Case Study

Figure 10 presents representative cases to illustrate how AMA improves mention disambiguation under different visual conditions. For clear single-subject images, the default tool provides concise visual semantics and helps distinguish visually or semantically similar candidates, such as ranking *Bernhard Glass* above other entities sharing the same surname. For multi-object images, the visual grounder focuses on the mention-related region and filters out irrelevant visual context, enabling the model to correctly link the mention to *Barack Obama*. When the image contains informative text, OCR Augment extracts entity-related textual evidence from the image, which helps recover the correct entity *Leopold Einstein* from a lower initial rank. For abstract or symbolic images, the symbolic solver converts visual cues into higher-level semantic concepts, allowing the model to link the sports-related mention to *Germany*. In cases where the image is missing or uninformative, Text Supplement generates a reasonable visual-semantic description from the mention context, helping the model distinguish *tablet computer* from other ambiguous meanings of “tablet”. These examples show that AMA provides mention-specific complementary evidence

Mention	Used Tool	Rank	Top1	Top2	Top3	Top4
 In July 2010, Glass became a coach of the Canadian Luge Association.	Default Case Tool output: Glass is a coach for the Canadian Luge Association, having joined the organization in July 2010.	Initial Rank	 Tanner Glass Canadian professional ice hockey player	 Bernhard Glass East German luger	 Presley T. Glass American politician	 Fridolin Glass Austrian Nazi activist and SS officer
		Final Rank	Bernhard Glass	Tanner Glass	Fridolin Glass	Presley T. Glass
 May receiving the Medal of Arts award from President Obama , July 13, 2013	Visual Grounder Tool output: Barack Obama is the 44th President of the United States, who presented the National Medal of Arts in 2013.	Initial Rank	 Michelle Obama Lawyer and Former First Lady of the United States	 Barack Obama president of the United States from 2009 to 2017	 Obama Josefi species of Turbellaria	 Obama, Fukui city in Fukui prefecture, Japan
		Final Rank	Barack Obama	Michelle Obama	Barack Obama Sr.	Ronald Reagan
 When Einstein read L.	OCR Augment OCR: LEOPOLD EINSTEIN in the SIXTEENTH DIMENSION Tool output: Leopold Einstein, the nephew of Albert Einstein, an American writer and scientist (known for works like "In the Sixteenth Dimension").	Initial Rank	 Albert Einstein German-born theoretical physicist; developer of the theory of relativity	 Carl Einstein German art historian (1885-1940)	 Hermann Einstein German electrical engineer and businessman, father of Albert Einstein	 Leopold Einstein Jewish teacher, vendor and writer
		Final Rank	Leopold Einstein	Albert Einstein	Hermann Einstein	Carl Einstein
 Team Germany in the match against the USA.	Symbolic Solver Tool output: Germany is a sovereign country in Europe, represented by national teams in international competitions such as sports matches.	Initial Rank	 Germany women's national wheelchair basketball team national sports team	 Germany country in Central Europe	 Germans residents and citizens of Germany	 Germany national association football team German men's national association football team
		Final Rank	Germany	Germany national association football team	Germany women's national wheelchair basketball team	Germany women's national football team
 If RIM launches a new tablet device , it will compete against Apple's iPad.	Text Supplement Tool output: A portable electronic device with a large touchscreen display, designed for mobile computing and media consumption	Initial Rank	 tablet solid form for doses of drugs or medication often taken orally	 tablet sugary confection from Scotland	 tablet computer mobile computer with integrated display, circuitry and battery	 graphic tablet computer input device
		Final Rank	tablet computer	tablet	tablet	iPad

Figure 10: Case Study of our EVEMEL.

for different visual conditions, thereby improving the final ranking of the correct entity.

H Computational Overhead of AMA

To evaluate the computational overhead of AMA, we report the routing frequency, token consumption, and processing time for the planner and individual tool executors on the WikiDiverse test set. As shown in Table 5, the planner is applied to every sample, followed by one adaptively selected executor. Therefore, AMA requires two MLLM calls per sample on average.

Table 5: Computational overhead of AMA on the WikiDiverse test set. Tokens and time are averaged over samples routed to each component.

Route	N	Freq. (%)	Avg. Tokens	Avg. Time (s)
Planner	2078	100.00	252.70	0.81
OCR Augment	79	3.80	274.56	4.20
Symbolic Solver	889	42.78	213.33	2.05
Text Supplement	159	7.65	210.25	1.75
Visual Grounder	236	11.36	229.31	3.71
Default Case	715	34.41	218.37	1.93
Overall AMA	2078	100.00	471.67	3.07

Symbolic Solver and Default Case account for 42.78% and 34.41% of the samples, respectively, whereas OCR Augment and Visual Grounder are activated for only 3.80% and 11.36%. This shows

that AMA selectively invokes additional visual processing rather than applying expensive tools uniformly to all samples. Overall, AMA consumes 471.67 tokens and 3.07 seconds per sample on average. Since the augmented evidence is generated offline and can be cached, this preprocessing cost is incurred only once and does not add repeated overhead to subsequent mention–entity matching.