ThinkLinker: From Low-Rank Interaction to Knowledge-Aware Verification for Multimodal Entity Linking

Anonymous ACL submission

Abstract

Recent advances in Multimodal Entity Linking (MEL) exploit textual and visual information to disambiguate mentions and align them with entities in a knowledge base. Existing methods typically design separate and complex network modules for each type of interaction among multi-granular and multimodal features, while lacking explicit modeling of the joint dependencies among these features. Moreover, most approaches rely on unidirectional retrieval-based matching and lack knowledge-driven verification, leading to unreliable disambiguation in weak-context scenarios. To address these challenges, we propose a novel two-stage MEL framework termed ThinkLinker. First, we introduce a low-rank fusion mechanism to model the joint dependencies among multi-granular and multimodal features, enabling comprehensive and explicit interactions while learning task-relevant discriminative information for candidate ranking in a lower-dimensional space. Subsequently, we develop a bidirectional retrieval-verification paradigm, where the ranked candidate entities guide an LLM-based multi-turn, dialogue-style verification process to generate mention-specific contextual augmentation. The augmented context is then adaptively fused with the original representation to further refine the linking model. Experimental results on public benchmark datasets demonstrate that the proposed ThinkLinker outperforms all state-of-the-art baselines. The code is publicly available at https://anonymous.4open.science/r/ThinkLinker-D443.

1 Introduction

Entity linking (EL) maps mentions in unstructured sources (e.g., social media, news, and web content) to the correct entities in 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 still exhibit several notable limitations.

From the perspective of feature interaction, most existing methods rely on simple feature concatenation, while some approaches employ attention mechanisms to weight the importance of different features. In MEL, entity disambiguation often depends on the joint consistency of evidence from multiple modalities and granularities across different levels. During feature interaction, existing methods typically require separately designed complex modules for each type of interaction and fail to explicitly capture the joint dependency among features, which makes them difficult to scale to multi-level feature interaction modeling. Directly modeling multiple features through high-order tensor multiplication provides strong expressive power and scalability, but it incurs prohibitive parameter growth and computational overhead, which limits its practical applicability. Therefore, it is necessary to develop a structured interaction framework that enables explicit multi-level feature interactions while maintaining controlled model complexity.

In terms of reasoning paradigms, most existing MEL methods are limited to "retrieval and matching", encoding mentions in the semantic space and retrieving unidirectionally from the knowledge space, without support for knowledge-based verification. By contrast, humans leverage candidate entity information to revisit the original context for targeted analysis. As illustrated in Figure 1, when contextual information is limited and multiple mentions share overlapping context, retrieval-

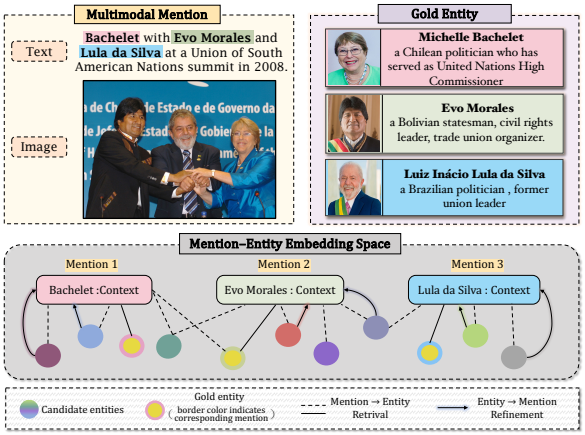

Figure 1: One example of Multimodal Entity Linking.

based methods struggle to distinguish candidates and fail to extract discriminative semantic cues for each mention. Therefore, a framework capable of bidirectional reasoning between the semantic and knowledge spaces is essential to better approximate human-like reasoning over mentions in weak-context scenarios.

To address the above issues, we propose ThinkLinker, a novel two-stage MEL framework. In the first stage, we design a reusable low-rank fusion mechanism to support multi-level compact interactions among features of different modalities and granularities, thereby more precisely computing the similarity score between mention and candidate entities. Subsequently, to address the second challenge, we leverage the obtained candidate ranking to conduct LLM-based multi-turn, dialogue-style verification in the semantic space, generating mention-specific responses as contextual augmentation. The augmented context is then fused with the original context via an aggregation-based adaptive fusion strategy, and the first-stage model is fine-tuned with these enhanced representations to further improve its disambiguation ability. Our contributions are summarized as follows:

- We propose a low-rank-based interaction strategy that enables multi-level explicit interactions, enhancing the accuracy and robustness of MEL.
- We develop a bidirectional retrieval-verification reasoning framework with an LLM-based multi-turn dialogue module, enabling targeted disambiguation under weak-context conditions.
- Extensive experiments on two public MEL benchmarks demonstrate the effectiveness and scalability of the proposed ThinkLinker framework.

2 Related Work

2.1 Entity Linking

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., 2020a; 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.

2.2 Multimodal Entity Linking

Traditional MEL 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. Single-level models, such as DZMNED (Moon et al., 2018), JMEL (Adjali et al., 2020), and M3EL (Hu et al., 2025a), encode global joint image–text representations and perform candidate scoring in a shared space, often yielding limited gains. In contrast, multi-level approaches, including MIMIC (Luo et al., 2023), FissFuse (Luo et al., 2024), and MMoE (Hu et al., 2025b), jointly model local and global features and achieve more substantial improvements. However, most multi-level methods rely on simple concatenation or aggregation for cross-level fusion, limiting their ability to fully exploit complementary multimodal information.

LLM-based MEL has recently attracted increasing attention, with large language models being incorporated into MEL pipelines in various ways. One line of research treats the LLM as the core inference module and determines the link by generating the target entity name, exemplified by GELR (Wang et al., 2023) and GEMEL (Shi et al., 2024). Another uses LLMs to expand context for the mention and refine entity descriptions, such as UniMEL (Liu et al., 2024). A third line employs an LLM-based rethinking module to perform further verification or reasoning over retrieved candidates, as in FissFuse and KGMEL (Kim et al., 2025). While these methods still rely on unidirectional retrieval and cannot revisit context for targeted verification, highlighting the need for bidirectional reasoning between semantic space and knowledge spaces.

3 Methodology

This section presents the two-stage ThinkLinker framework, illustrated in Figure 2.

3.1 Problem Formulation

To formalize the notations and objectives in this study, we define the MEL task as follows. Each mention is represented as $m = (m_w, m_t, m_v)$, where m_w is the mention token, m_t is the sentence containing the mention, and m_v is the associated visual context. The initial candidate set for each mention m is $C^0(m) = \{e^j = (e_n^j, e_d^j, e_v^j)\}_{j=1}^N$, where e_n^j is the entity name, e_d^j is the textual description, and e_v^j is the visual information.

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 $\mathrm{Score}(\cdot)$:

$$e^* = \underset{e^j \in C^0(m)}{\arg\max}\, \mathrm{Score}\left(m, e^j\right). \tag{1}$$

3.2 Dual-Stream Feature Encoding

Given a mention m and its candidate set C^0, ThinkLinker first extracts local textual and visual features using the CLIP text encoder $\mathrm{Enc}^T(\cdot)$ and visual encoder $\mathrm{Enc}^V(\cdot)$:

$$\begin{aligned} T_m^L, V_m^L &= \mathrm{Enc}^T(m_w; m_t),\ \mathrm{Enc}^V(m_v), \\ T_e^L, V_e^L &= \mathrm{Enc}^T(e_n; e_d),\ \mathrm{Enc}^V(e_v). \end{aligned} \tag{2}$$

Here $T_m^L, T_e^L \in \mathbb{R}^{l \times d_T}$ denote local textual sequence features and $V_m^L, V_e^L \in \mathbb{R}^{P \times d_V}$ are visual patch-level features, where l and P correspond to the sequence length and patch count, and d_T and d_V denote the hidden dimensions.

To capture holistic semantics, global representations $T_m^G, V_m^G, T_e^G, V_e^G$ are obtained further derived by applying the CLIP encoder's internal aggregation operations, producing compact yet context-rich multimodal embeddings.

3.3 Low-Rank Multi-level Modal Fusion Framework

This section describes the first stage of ThinkLinker, detailing its two key components, the feature fusion process, and the joint training strategy.

3.3.1 Hybrid Pooling Gate

To aggregate local sequence features into a compact, discriminative representation that shares the same dimensionality as the global vector to facilitate subsequent low-rank fusion, we design the Hybrid Pooling Gate (HPG) module.

The mention and candidate representations from Section 3.2 are first projected into a unified embedding space $\mathbb{R}^d$. The local feature is denoted as $X^L = [x_1, x_2, \ldots, x_l]$. The HPG module aggregates local vectors via three pooling operations:

$$\begin{cases} h_{\max} = \mathrm{Max}(X^L), \\ h_{\text{mean}} = \mathrm{Mean}(X^L), \\ h_{\text{attn}} = \mathrm{AttnPool}(X^G, X^L). \end{cases} \tag{3}$$

Here, max pooling $h_{\max}$ highlights the most salient local components; mean pooling h_{mean} captures overall statistics, and h_{attn} is an attention-weighted local vector conditioned on the global vector X^G.

The three vectors are concatenated as $p = [h_{\max}; h_{\text{mean}}; h_{\text{attn}}]$, and fed into a linear layer with sigmoid activation to produce a dimension-wise gating vector g. We then split g into $[g_1, g_2, g_3]$with each $g_i \in \mathbb{R}^d$, the final representation is computed via channel-wise weighting and concatenation: $\overline{X}^L = \mathrm{concat}\big(g_1 \odot h_{\max}, g_2 \odot h_{\text{mean}}, g_3 \odot h_{\text{attn}}\big)$, where $\odot$ denotes element-wise multiplication. The adaptive gates balance the three pooling paths, enhancing robustness and discriminative power.

3.3.2 Scalable Low-rank Fusion Tower

We propose the scalable Low-rank Fusion Tower (LFT), which follows Liu et al. (2018) by approximating the interaction tensor as a sum of outer products of low-dimensional factors. This design enables comprehensive and explicit interactions among features while learning a compact low-rank space to suppress redundancy and noise, substantially reducing parameter count and computational cost. Moreover, it naturally scales to an arbitrary number of feature hierarchies and modalities for efficient high-order fusion. The detailed derivation of LFT is provided in Appendix A. Expanding Equation (15), we obtain:

$$F = \left(\sum_{i=1}^{k} w_i^1 \cdot z^1\right) \circ \cdots \circ \left(\sum_{i=1}^{k} w_i^M \cdot z^M\right). \tag{4}$$

Here, for each feature index $m \in 1, \ldots, M$, z^m is the feature representation, w_i^m the corresponding low-rank factor, and $\circ$ denotes element-wise multiplication. This formulation relies solely on low-dimensional operations to preserve key information while suppressing noise.

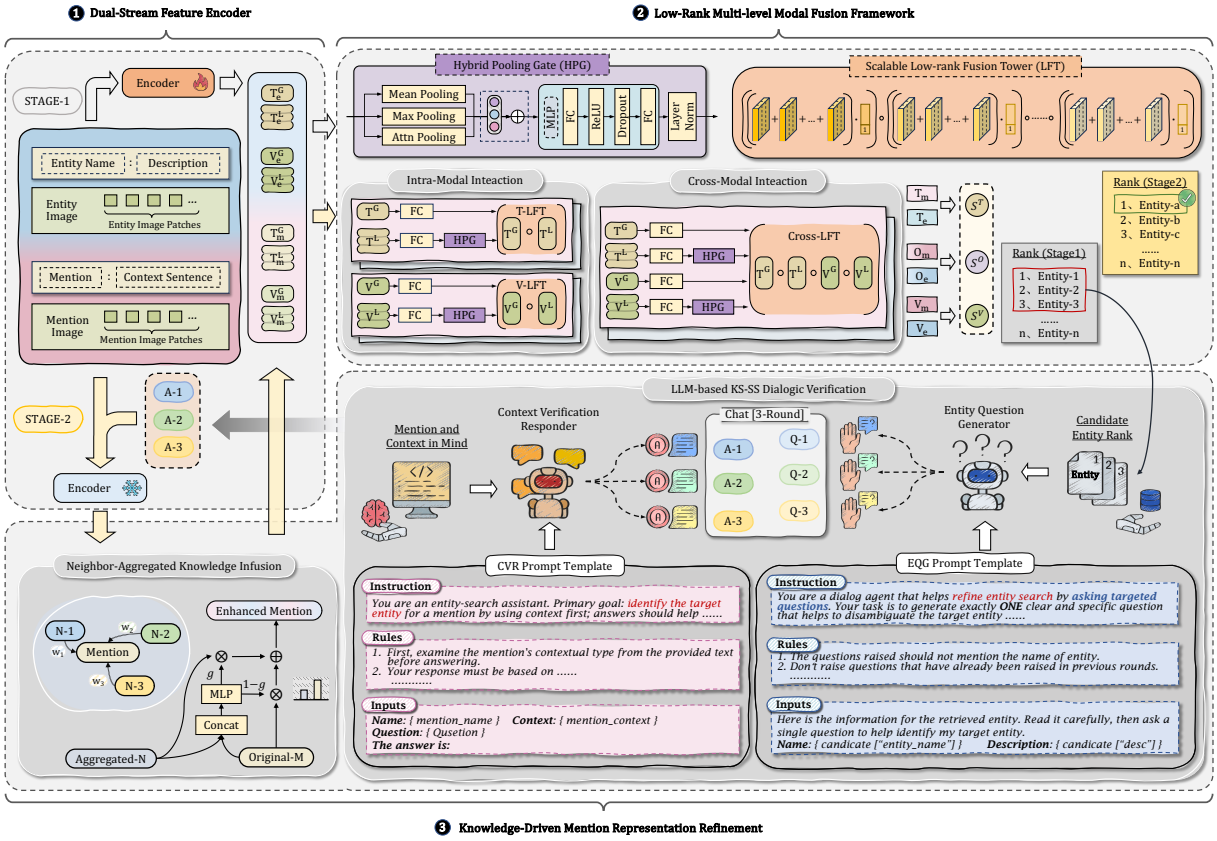

Figure 2: Overview of our proposed ThinkLinker framework for entity linking.

3.3.3 Intra-Modal Interaction

To capture multi-level information within the textual and visual modalities, we apply scalable low-rank fusion modules to independently enhance modality-specific semantic interactions, denoted as T-LFT and V-LFT.

For both modalities, we apply HPG to obtain aggregated local features $\overline{X}^L$. We then fuse $\overline{X}^L$ with the corresponding global features X^G through T-LFT and V-LFT, instantiating Equation (4) as:

$$F = \left(\sum_{i=1}^{k} w_i^{L_X} \cdot \overline{X}^L\right) \circ \left(\sum_{i=1}^{k} w_i^{G_X} \cdot X^G\right), \quad (5)$$

where $X \in T, V$. This produces mention and entity representations T_m and T_e for the textual modality and V_m and V_e for the visual modality. Similarity within each modality is computed via inner product: $S^T = T_e \cdot T_m$, $S^V = V_e \cdot V_m$.

3.3.4 Cross-Modal Low-Rank Fusion

In this module, we model the cross-modal interactions between text and vision. Since mentions and candidate entities are processed symmetrically, we omit this distinction below. We first apply HPG to obtain the aggregated local textual features $\overline{T}^L$ and visual features $\overline{V}^L$. These, together with their corresponding global features, are jointly fused within a low-rank fusion tower Cross-LFT. Concretely, we expand Equation (4) into four terms as follows:

$$F = \left(\sum_{i=1}^{k} w_i^{L_T} \overline{T}^L\right) \circ \left(\sum_{i=1}^{k} w_i^{G_T} T^G\right) \circ \left(\sum_{i=1}^{k} w_i^{L_V} \overline{V}^L\right) \circ \left(\sum_{i=1}^{k} w_i^{G_V} V^G\right). \quad (6)$$

After this cross-modal low-rank fusion, the mention and candidate entity are represented as O_m and O_e, respectively. In the fused space, we use the inner product as the similarity function: $S^O = O_e \cdot O_m$.

3.3.5 Joint Training

Intra-modal and cross-modal similarities provide complementary views of mention-entity matching. We average the three similarity scores to obtain the ranking score: $S^* = (S^T + S^V + S^O)/3$.

To enhance the discriminability of unimodal and joint representations, we apply cross-entropy loss

to each similarity score and minimize their sum:

$$\mathcal{L}_{final} = \mathcal{L}_T + \mathcal{L}_V + \mathcal{L}_O + \mathcal{L}_*. \tag{7}$$

After training, each mention m ranks its candidate entities using S^*, and the top R entities are retained as the updated candidate list: $C^1(m) = \left[e^1, e^2, \ldots, e^R\right]$.

3.4 Knowledge-Driven Mention Refinement

In the second stage, we leverage candidate entity list C^1 produced in the first satge and design LLM-based dialogic verification and neighborhood aggregation to enhance semantic-space (SS) mention representations with the knowledge space (KS).

3.4.1 LLM-based KS-SS Dialogic Verification

Mention context is often short and co-occurring mentions receive similar representations, weakening discrimination. We therefore adopt knowledge-driven question-answer generation to produce candidate-aware cues.

Concretely, we design the **Entity Question Generator (EQG)** to construct a question prompt $T_{EQG}(\cdot)$ from the key attributes of a candidate e. At each round r, the question-generation model L_{EQG} produces a candidate-specific question:

$$q_r = L_{EQG}\left(T_{EQG}\left(e^r\right)\right). \tag{8}$$

Here, $1 \leq r \leq R$ means we select top-R candidate entities from C_1, because higher-ranked candidates tend to contain more reliable and distinctive semantic cues, they are more likely to yield highly discriminative questions.

Then, the **Context Verification Responder (CVR)** combines q_r with the original context of mention m^i to form an answer prompt $T_{CVR}(\cdot)$, and the answer-generation model L_{CVR} produces a mention-conditioned answer:

$$a_r = L_{CVR}\left(T_{CVR}\left(m^i, q_r\right)\right). \tag{9}$$

Each answer a_r is jointly constrained by the mention context and candidate semantics, enriching the mention representation with candidate-aware discriminative information. All answers for m^i are aggregated as $\mathcal{A}^i = \{a_r\}_{r=1}^R$, forming an expanded semantic neighborhood for subsequent encoding and disambiguation. The algorithmic pipeline is detailed in Appendix B, and the complete templates for EQG and CVR are provided in Appendix C.

3.4.2 Neighbor-Aggregated Knowledge Infusion

We model the mention as a center node with the generated answers $\mathcal{A}$ as neighbors, and perform ordered center-neighbor aggregation to obtain a multi-perspective neighborhood representation.

Specifically, each answer $a_r \in \mathcal{A}$ is encoded into global and local features a_r^G and a_r^L, while the mention is represented by T^G and T^L.

▷**Adaptive Neighbor Weighting (ANW).** Each neighbor is assigned a learned weight ω_r, estimated from the mention and neighbor representations, and the weighted sum of neighbor embeddings forms the aggregated neighborhood representations at both global and local levels:

$$G = \sum_{r=1}^{R} \omega_r a_r^G, \quad L_j = \sum_{r=1}^{R} \omega_r a_{r,j}^L, \tag{10}$$

where the same weights ω_r are shared across global and local features to preserve semantic consistency and reduce parameter overhead.

▷**Residual Gating Aggregation (RGA).** To avoid neighbor information overriding the original context, we employ a dimension-wise gating mechanism to fuse aggregated and original representations. Concretely, the aggregated feature $U \in [G, L_j]$ and $T \in [T^G, T_j^L]$ are concatenated and passed through a linear layer to compute the gating vector, and the final fused representation is obtained in a residual form:

$$T_{final} = \alpha \odot U + ((1-\alpha) \odot T). \tag{11}$$

3.5 Training Strategy

The training strategy consists of two stages. In the first stage, the text-visual encoder and the low-rank interaction module are trained jointly on the original data. In the second stage, the encoder is frozen, and the enhanced mention representations are fed, together with the remaining unchanged vectors, back into the first-stage matching architecture to further fine-tune the low-rank fusion network using the augmented data. This two-stage scheme allows the high-level matching module to adapt to the updated semantic distributions while keeping the underlying representations stable, thereby more effectively leveraging neighborhood information and contextual cues for entity disambiguation.

Table 1: Performance comparison of different methods on two MEL datasets. All baseline results are taken from the FissFuse paper. The best scores are highlighted in **bold** and the second best in underlined. △ denotes the change of our method from the first stage to the second stage. FissFuse† denotes FissFuse with LLM re-ranking, and ThinkLinker† denotes ThinkLinker with the second stage.

Category	Methods	WikiMEL					WikiDiverse					Avg.	
		H@1↑	H@2↑	H@3↑	MR↓	MRR↑	H@1↑	H@2↑	H@3↑	MR↓	MRR↑	H@1↑	MRR↑
EL	BERT (Devlin et al., 2019)	39.95	53.68	61.31	6.36	54.07	57.08	74.57	84.32	2.12	72.03	48.52	63.05
	BLINK (Wu et al., 2020b)	36.00	49.54	57.52	7.54	50.36	56.30	73.40	82.69	2.19	71.19	46.15	60.78
MEL	DZMNED (Moon et al., 2018)	39.41	50.97	57.90	7.77	52.13	29.11	47.37	61.16	3.53	49.53	34.26	50.83
	JMEL (Adjali et al., 2020)	47.99	63.60	71.68	4.33	62.42	51.55	68.08	78.49	2.47	67.15	49.77	64.79
	MEL-HI (Zhang et al., 2021)	30.86	45.26	54.73	6.22	47.18	53.88	70.59	80.00	2.36	69.01	42.37	58.10
	ViLT (Kim et al., 2021)	79.40	84.08	85.65	3.41	83.80	40.27	58.17	68.49	2.91	58.38	59.84	71.09
	CLIP (Radford et al., 2021)	81.53	89.97	93.15	1.78	87.89	61.12	79.70	89.16	1.88	75.61	71.33	81.75
	GHMFC (Wang et al., 2022a)	56.69	72.99	80.61	2.91	70.45	55.71	72.35	80.94	2.30	70.31	56.20	70.38
	MIMIC (Luo et al., 2023)	81.62	90.29	93.58	1.77	88.05	67.90	85.14	92.63	1.62	80.57	74.76	84.31
	DRIN (Xing et al., 2023a)	66.05	79.81	85.39	2.11	80.84	49.43	66.90	77.17	1.83	57.21	57.74	69.02
	FissFuse (Luo et al., 2024)	84.80	92.37	95.05	1.61	90.26	80.30	91.42	95.34	1.39	88.11	82.55	89.18
MEL+LLMs	GPT-3.5	73.80	-	-	-	-	72.70	-	-	-	-	73.25	-
	GEMEL (Shi et al., 2024)	75.20	-	-	-	-	80.20	-	-	-	-	77.70	-
	FissFuse† (Luo et al., 2024)	87.89	93.42	95.36	1.54	92.02	83.29	92.53	95.89	1.35	89.81	85.59	90.92
	UniMEL (Liu et al., 2024)	88.19	93.81	95.80	1.55	92.34	81.01	92.25	95.48	1.37	88.61	84.60	90.48
Ours	ThinkLinker	90.44	95.51	97.23	1.35	93.98	83.34	92.80	**96.30**	1.34	89.93	86.89	91.96
	ThinkLinker†	**91.74**	**96.44**	**97.66**	**1.29**	**94.86**	**87.05**	**93.97**	95.64	**1.27**	**92.07**	**89.40**	**93.47**
	△	+1.30	+0.93	+0.43	-0.06	+0.88	+3.71	+1.17	-0.66	-0.07	+2.14	+2.51	+1.51

4 Experiments

4.1 Experimental Settings

Datasets: We evaluate our method on two public MEL benchmarks, which were proposed by Wang et al. (2022a), and WikiDiverse which is proposed by Wang et al. (2022b). For fair comparison, we follow the original data splits and candidate sets provided by Xing et al. (2023b). Detailed dataset statistics and split settings are given in Appendix D.1.

Baselines: To comprehensively assess the effectiveness of ThinkLinker, we compare it with three groups of baselines. (1) **Text-only EL** methods: BERT (Devlin et al., 2019) and BLINK (Wu et al., 2020b). (2) **Multimodal MEL** models: DZMNED (Moon et al., 2018), JMEL (Adjali et al., 2020), MEL-HI (Zhang et al., 2021), ViLT (Kim et al., 2021), CLIP (Radford et al., 2021), GHMFC (Wang et al., 2022a), MIMIC (Luo et al., 2023), DRIN (Xing et al., 2023a), and FissFuse (Luo et al., 2024). (3) **MEL+LLMs** methods: GPT-3.5, GEMEL (Shi et al., 2024), FissFuse† (Luo et al., 2024), and UniMEL (Liu et al., 2024). For details, we provide extensive descriptions of baselines in Appendix D.2.

Implementation Details: We initialize the multimodal encoder with the pre-trained CLIP-ViT-B/32 model. The textual and visual features are projected into a shared 256 dimensional hidden space. In the low-rank fusion module, the rank is fixed to 4. The dialog rounds R is set to 3 on WikiMEL and 2 on WikiDiverse. DeepSeek-V3.1 is used as the underlying LLM for both EQG and CVR. Training follows a two-stage scheme with up to 30 epochs per stage and early stopping based on development performance. We optimize all parameters with AdamW. Learning rates are set per dataset and stage as follows: the first stage uses 1×10^{-6} for WikiDiverse and 1×10^{-5} for WikiMEL, while the second stage uses 1×10^{-4} for WikiDiverse and 3×10^{-5} for WikiMEL.

4.2 Main Result

We conduct comparative experiments on two public benchmarks, WikiMEL and WikiDiverse, to evaluate the effectiveness of different models on MEL tasks. Table 1 reports the results in terms of H@k, MR, and MRR. The formal definitions of these evaluation metrics are provided in Appendix D.3.

Firstly, multimodal methods consistently outperform text-only baselines, confirming the effectiveness of visual information in MEL. The gains are substantially larger on WikiMEL than on WikiDiverse, likely because the latter contains noisier and less complete images, which weaken the contribution of visual signals.

Table 2: Ablation study on key components of ThinkLinker. * indicates the corresponding second-stage results. The full ablations of the second-stage are reported in the Appendix F.

Dataset		WikiMEL						WikiDiverse					
Metric		H@1	H@1*	H@2	H@3	MR	MRR	H@1	H@1*	H@2	H@3	MR	MRR
ThinkLinker		**90.44**	**91.74**	**95.51**	**97.23**	**1.35**	**93.98**	**83.34**	**86.57**	**92.80**	**96.30**	**1.34**	**89.93**
Modality-level	(1) w/o Text	56.61	59.01	68.33	74.62	4.16	67.98	46.20	45.17	59.01	66.14	3.23	60.45
	(2) w/o Image	83.63	85.97	90.97	93.69	1.75	89.17	69.91	84.78	84.92	90.27	1.71	81.06
Module-level	(3) w/o Cross-Modal	88.88	91.06	94.35	96.60	1.41	92.89	74.30	84.65	86.36	91.02	1.65	83.54
	(4) w/o Intra-Modal	84.66	88.04	91.37	94.04	1.73	92.89	72.79	86.02	86.15	91.36	1.65	82.83
Feature-level	(5) w/o Local	88.76	91.00	94.93	96.98	1.39	92.99	73.47	86.36	87.32	91.98	1.63	83.41
	(6) w/o Global	88.41	90.23	94.68	96.71	1.40	92.75	73.00	84.51	85.81	91.84	1.62	83.00
(7) w/o Low Rank		88.76	91.06	94.93	96.98	1.39	92.99	73.82	86.43	89.95	90.95	1.66	83.24

Secondly, among MEL methods, architectural choices lead to clear performance gaps. Models built on vision and language pre-training, particularly CLIP-based approaches, show consistent advantages thanks to well-aligned cross-modal representations and strong generalization. In contrast, earlier MEL models fail to fully exploit multimodal cues. Recent methods such as MIMIC and FissFuse further improve performance by incorporating multi-granularity feature alignment and modality-aware fusion. In the generative MEL paradigm, LLMs provide rich external knowledge, while approaches such as GEMEL that directly generate target entities are constrained by hallucination and high computational cost. The results of FissFuse† and UniMEL demonstrate that, under proper constraints, leveraging LLMs for result verification or data augmentation is both feasible and effective.

Finally, our proposed ThinkLinker achieves the best overall performance on both datasets. In the first stage, it surpasses the strongest MEL baseline by 5.64 H@1 points, demonstrating the effectiveness of low-rank interaction with multi-granular hierarchical modeling. In the second stage, multi-round LLM-based mention enrichment brings an additional 3.23-point gain on WikiDiverse, indicating that knowledge-driven verification improves ranking precision. It is worth noting that the slight drop in H@3 on WikiDiverse suggests that the verification-oriented context refinement sharpens the model's discrimination, prioritizes identifying the most accurate entity rather than expanding recall over multiple candidates.

4.3 Ablation Study

We conduct ablation studies to assess each component's contribution, as shown in Table 2.

First, we examine modality-level inputs and module-level interaction components. Removing text or visual input, or ablating cross-modal or intra-modal modules, consistently degrades performance, confirming that both unimodal signals and interaction mechanisms are essential, demonstrating that our model is particularly effective at leveraging the noisy visual cues in WikiDiverse.

Next, discarding either local or global representations consistently harms performance, demonstrating that fine-grained local cues and coarse-grained semantic coherence are complementary. This effect is stronger on WikiDiverse, highlighting its importance in noisy, complex scenarios. Replacing low-rank fusion also leads to degradation, indicating that the low-rank constraint helps suppress redundancy and noise while encouraging the model to focus on a more discriminative subspace.

Finally, H@1* reports ablations for the second stage. Across all settings, this stage consistently overpass the first stage, indicating that LLM-based mention enrichment injects richer information and stabilizes performance, and that the key components remain important across both stages.

4.4 Discussion

Analysis of HPG Mechanism. We analyze the HPG design by ablating its three pooling branches (Table 3). On WikiMEL, removing any branch

Table 3: Ablation on HPG structure.

Dataset	Pool Method	H@1	H@2	H@3	MR	MRR
WikiMEL	ThinkLinker	**90.44**	**95.51**	**97.23**	**1.35**	**93.98**
	w/o AvgPool	89.20	94.54	96.32	1.42	93.07
	w/o AttnPool	89.79	95.41	97.12	1.39	93.62
	w/o MaxPool	89.36	94.47	96.98	1.37	93.40
WikiDiverse	ThinkLinker	**83.34**	92.80	**96.30**	**1.34**	**89.93**
	w/o AvgPool	82.86	92.53	95.96	**1.34**	89.64
	w/o AttnPool	82.88	**92.87**	**96.30**	**1.34**	89.80
	w/o MaxPool	82.04	91.71	95.61	1.37	89.04

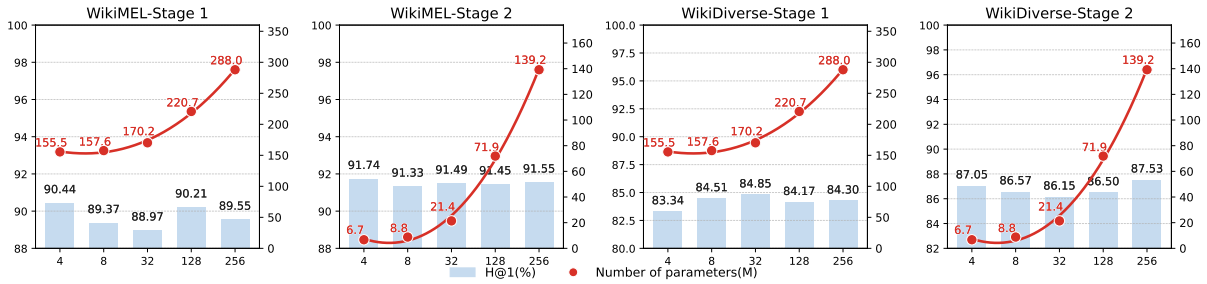

Figure 3: Effect of rank on performance and computation cost in low-rank fusion.

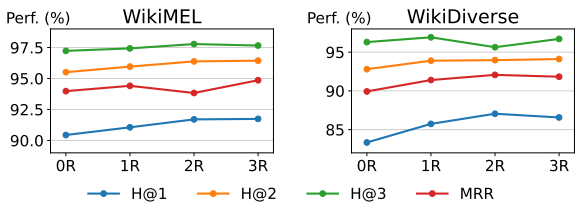

Figure 4: Effect of dialogue rounds on performance.

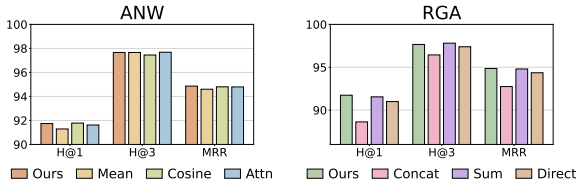

Figure 5: Effect of neighbor-aggregated knowledge infusion strategies.

degrades performance, confirming the complementarity of modeling salient, statistical, and global-conditioned local cues. Notably, removing attention pooling on WikiDiverse is nearly comparable to our full model, indicating that the additional benefit of global attention pooling is limited in more complex contexts. Overall, the full HPG achieves the best performance, showing that our HPG effectively balances different pooling paths and produces compact, discriminative local representations compatible with subsequent low-rank fusion.

Influence of rank on Low-Rank Fusion Performance. We vary the rank k of the low-rank fusion module on both stages (Figure 3). As k increases, parameters grow rapidly, while H@1 changes by only 1-2 points; small ranks often achieve near-optimal performance. We thus fix $k = 4$ as a practical trade-off between accuracy and model size. This indicates that the discriminative features for MEL task are relatively sparse and low-dimensional. Low-rank representations capture most cross-modal correlations, enabling multi-granular interactions in a compact subspace, reducing noise and overfitting while preserving the most informative signals.

Impact of KS-SS Dialogue Rounds. We evaluate different dialogue-round settings (Figure 4). Performance generally increases with more rounds, but gains diminish over time. The largest improvement occurs from 0 to 1 round, showing that a small interaction suffices to provide key disambiguation context. Additional rounds yield marginal gains, particularly on WikiDiverse, where short mention contexts make extra dialogue prone to introducing noise or hallucinations. We therefore use three rounds for WikiMEL and two for WikiDiverse to balance improvements with stability.

Effect of Neighbor-Aggregated Knowledge Infusion. We evaluate the contributions of the ANW and RGA components on WikiMEL (Figure 5). For ANW, our method outperform uniform mean weighting, cosine-softmax weighting, and learnable attention weighting, indicating that our adaptive weighting better captures high-order nonlinear interactions between the mention and its neighbors. For RGA, we compare it against direct concatenation, summation, and direct use the aggregated neighbor vector. RGA consistently performs best, as the residual path preserves essential mention cues while the gating mechanism adaptively balances mention and neighbor features, improving robustness to noisy neighbors.

5 Conclusion

In this work, we propose ThinkLinker, a two-stage framework that combines low-rank multi-level interaction with LLM-based bidirectional reasoning. It explicitly captures the joint dependencies among features across different granularities and modalities, and revisits candidates to enhance and verify mention semantics. Experiments show that ThinkLinker consistently outperforms sota methods.

Limitations

ThinkLinker adopts a two-stage framework, where the second stage relies on an LLM to perform multi-round dialogue-style generation, and most limitations stem from this stage. First, hallucinations in LLMs remain a non-negligible issue. Second, engaging an LLM in multiple rounds of interaction introduces additional computational and latency overhead, with overall cost typically scaling linearly or even super-linearly with dataset size. Third, the verification process currently operates only on textual information and does not incorporate visual modalities, which may limit the model's ability to fully exploit multimodal cues. Future work will explore lighter-weight generation or retrieval mechanisms, more effective hallucination mitigation strategies, and ways to incorporate visual signals into the verification stage to further reduce cost and improve applicability to larger-scale and more diverse MEL scenarios.

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.

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

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 Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792.

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.

Juyeon Kim, Geon Lee, Taeuk Kim, and Kijung Shin. 2025. Kgmel: Knowledge graph-enhanced multimodal entity linking. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3015–3019.

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.

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.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2247–2256.

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.

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.

Senbao Shi, Zhenran Xu, Baotian Hu, and Min Zhang. 2024. Generative multimodal entity linking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 7654–7665.

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.

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*, pages 7841–7857.

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. 2020a. Dynamic graph convolutional networks for entity linking. In *Proceedings of The Web Conference 2020*, pages 1149–1159.

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

Shangyu Xing, Fei Zhao, Zhen Wu, Chunhui Li, Jianbing Zhang, and Xinyu Dai. 2023a. Drin: Dynamic relation interactive network for multimodal entity linking. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3599–3608.

Shangyu Xing, Fei Zhao, Zhen Wu, Chunhui Li, Jianbing Zhang, and Xinyu Dai. 2023b. Drin: Dynamic relation interactive network for multimodal entity linking. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3599–3608.

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.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, pages 5634–5641.

Li Zhang, Zhixu Li, and Qiang Yang. 2021. Attention-based multimodal entity linking with high-quality images. In *Proceedings of the 26th International Conference on Database Systems for Advanced Applications*, pages 533–548.

A Derivation of the LFT Module

A detailed mathematical derivation of the LFT module is provided below.

The original linear fusion can be written as:

$$F = W \cdot Z + b, \tag{12}$$

where Z is a joint representation of all features, and b is a bias. In order to capture interactions among arbitrary subsets of the M features, we adopt the approach of Zadeh et al. (2018), which augments each feature vector with a constant and then takes the outer product of all such augmented vectors. Consequently, the bias term in Equation (12) is implicitly encoded and may be omitted. Based on this, Z can be expressed as the outer product of the augmented feature vectors, $Z = \otimes_{m=1}^{M} z^m$, with its computation presented as follows:

$$\begin{aligned} Z &= \begin{bmatrix} z^1 \\ 1 \end{bmatrix} \otimes \begin{bmatrix} z^2 \\ 1 \end{bmatrix} \otimes \cdots \otimes \begin{bmatrix} z^M \\ 1 \end{bmatrix} \\ &= 1 + \sum_{i=1}^{M} z^i + \sum (z^i \otimes z^j) \\ &\quad + \sum (z^i \otimes z^j \otimes z^k) \\ &\quad + \cdots + z^1 \otimes z^2 \otimes \cdots \otimes z^M. \end{aligned} \tag{13}$$

For brevity, the summations in the expression above omit index ranges; unless otherwise specified, all indices range over $\{1, \ldots, M\}$ in strictly increasing order.

Equation (13) expands into interaction terms ranging from zero-order to M-th order. The zero-order term implements the bias; first-order terms capture individual feature information or marginal effects; second-order terms explicitly model multiplicative pairwise interactions between features and thus express pairwise joint semantics; higher-order terms encode multi-way interactions to represent more complex joint dependencies.

To further reduce computational complexity, we replace the full weight tensor W with a low-rank weight $\widetilde{\mathcal{W}}$:

$$\widetilde{\mathcal{W}} = \sum_{i=1}^{k} \bigotimes_{m=1}^{M} w_m^{(i)}, \tag{14}$$

where the minimal k is the tensor rank and the vectors $w_m^{(i)}$ are the decomposition factors. Consequently, the fused tensor is:

$$\begin{aligned} F &= \left(\sum_{i=1}^{k} \bigotimes_{m=1}^{M} w_m^{(i)} \right) \cdot Z \\ &= \sum_{i=1}^{k} \left(\bigotimes_{m=1}^{M} w_m^{(i)} \cdot Z \right) \\ &= \sum_{i=1}^{k} \left(\bigotimes_{m=1}^{M} w_m^{(i)} \cdot \bigotimes_{m=1}^{M} z^m \right) \\ &= \Lambda_{m=1}^{M} \left[\sum_{i=1}^{k} w_m^{(i)} \cdot z^m \right]. \end{aligned} \tag{15}$$

B Pipeline of Knowledge-Driven QA Generation Paradigm

The detailed pipeline of LLM-based KS-SS Dialogic Verification is specified in Algorithm 1 in pseudocode form. The algorithm defines how to construct the question template from candidate entities and how to construct the answer template from the mention and the generated question; it then applies L_{EQG} and L_{CVR} iteratively to generate questions and answers, yielding the augmented text set A_i.

Algorithm 1: Knowledge-Driven Question-Answer Generation Paradigm

Inputs: mention $m \in M$, a trained retrieval model $Model$, total question–answer round R, encoder Enc^T, question-generation large model L_{EQG} and answer-generation large model L_{CVR}.
Output: an augmented text set $\mathcal{A}^i = \{a_r\}_{r=1}^{R}$.

Retrieve $C^1(m) = Model(m) = [e^1, \ldots, e^R]$.
Encode $e = \mathrm{Enc}^T(c)$ for $c \in C^1$.
Initialize text set $\mathcal{A}^i = []$.
for $r = 1$ **to** R **do**:
    Construct question template $T'_{EQG} = T_{EQG}(e^r)$.
    Generate question $q_r = L_{EQG}(T'_{EQG})$.
    Construct answer template $T'_{CVR} = T_{CVR}(m, q_r)$.
    Generate answer $a_r = L_{CVR}(T'_{CVR})$.
    Append a_r to $\mathcal{A}^i$.
end for
return $\mathcal{A}^i$.

C Question–Answer Prompt Templates

The prompt templates used to query the question-generation LLM L_{EQG} are shown in Table 4, and those used for the answer-generation LLM L_{CVR} are shown in Table 5.

Table 4: Prompt template for Entity Question Generator (EQG).

Prompt Template for Entity Question Generator (EQG)

System prompt

You are a dialog agent that helps refine entity search by asking targeted questions.
You will be given an entity from a knowledge base (called the "anchor entity"): the entity's name and a short description.
However, this anchor entity may not be the exact entity I'm actually looking for.
Your task is to generate exactly **ONE** clear and specific question that helps to disambiguate the target entity.

Prioritize useful attributes such as:

– entity category and time range
– field/industry, intended use, or function
– geographic information (country, city, coverage area)
– alternative names / full name / abbreviations
– key characteristics (size/model/color/capacity/version/license type, etc.)

Rules:

1. The questions raised should not mention the name of entity.
2. Do not raise questions that have already been raised in previous rounds.
3. Always output exactly ONE question. If information is limited, still ask the best possible disambiguating question.
4. Do not output empty text. If you cannot find a strong question, ask a weaker but still relevant question.
5. Do not ask yes/no questions.
6. Do not answer the question yourself.
7. You have 3 rounds and you can only ask one question at a time.

User prompt

Here is the information for the retrieved entity. Read it carefully, then ask a single question to help identify my target entity.
Note: sometimes the description may be missing or vague.
Please do not raise questions that have already been raised in previous rounds.

Name: anchor_candidate["entity_name"]
Description: anchor_candidate["desc"]

Question:

Table 5: Prompt template for Context Verification Responder (CVR).

Prompt Template for Context Verification Responder (CVR)

System prompt

You are an entity-search assistant. Primary goal: identify the target entity for a mention by using context first; answers should help entity linking, not primarily serve as free-form answers.

Your answer must strictly follow the following rules:

1. First and foremost, examine the mention's contextual type from the provided text before answering.
2. Your response must be based on this contextual type. If the context-implied type does NOT match the question-implied type, output in this form: "<Mention> is a <CONTEXTUAL TYPE>." Then provide affirmative facts about the mention specifically in its capacity as that <CONTEXTUAL TYPE> using reliable external knowledge (avoid repeating the provided context). CRITICAL: IGNORE the question's implied type entirely in this case. Do not answer the original question if it is about the wrong type. Do not provide facts related to entities mentioned in questions of different types. Do not use negative sentences such as "It is not ..." in your response.
3. If the context type matches the type implied by the question, there is no need to specify the entity type in the answer. Please answer the question directly in the form of an affirmative sentence.
4. Keep your answer to one or two sentences.
5. Always respond with affirmative statements, and avoid using negative words like "not" or "no" in your answers.

Your responses must remain factual, precise, and concise.

User prompt

Name: mention_name
Context: mention_context
Question: Question

The answer is:

Table 6: Statistics of two datasets.

Datasets		WikiMEL	WikiDiverse
Mentions	Train	18092	12268
	Valid	2585	1459
	Test	5169	1459
Candidates		100	10
Avg. Sentence Length		10.13	8.2

D Experimental Details

D.1 Statistics of Datasets

We conduct experiments on two publicly available multimodal entity linking datasets: WikiMEL (Wang et al., 2022a) and WikiDiverse (Wang et al., 2022b).

WikiMEL 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 contains over 8,000 diverse context topics and entity types sourced from Wikinews, using Wikipedia, which hosts more than 16 million entities, as the corresponding knowledge base.

Table 6 summarizes the statistics of the two datasets. Following the original data splits in Luo et al. (2024), we assign the (train, valid, test) splits as (70%, 10%, 20%) and (80%, 10%, 10%) for WikiMEL and WikiDiverse, respectively.

D.2 Descriptions of Baselines

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

BERT (Devlin et al., 2019) is a bidirectional Transformer encoder pre-trained on large-scale corpora and commonly used as a textual backbone for entity linking.

BLINK (Wu et al., 2020b) is a scalable text-based entity linking framework that combines bi-encoder retrieval with cross-encoder re-ranking based on BERT representations.

DZMNED (Moon et al., 2018) is an early MEL model that integrates short textual context and associated images via attention mechanisms.

JMEL (Adjali et al., 2020) is a joint multimodal entity linking model for Twitter that learns text and image representations with a dual-branch architecture optimized using triplet loss.

MEL-HI (Zhang et al., 2021) proposes a hierarchical interaction framework with multi-level attention to suppress noisy visual signals and enhance discriminative features.

ViLT (Kim et al., 2021) is a VLP model that employs a lightweight visual backbone to produce generic multimodal representations.

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.

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.

MIMIC (Luo et al., 2023) is a 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.

DRIN (Xing et al., 2023a) is a dual-relation interaction network that constructs multi-type mention–entity graphs and applies graph convolutional networks to capture different alignment relations.

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.

GPT-3.5 uses large language model in a zero-shot or few-shot manner for MEL by prompting it to directly predict target entities based on the textual and multimodal input.

GEMEL (Shi et al., 2024) treats a generative LLM as the central component and uses frozen encoders together with a lightweight feature mapper to enable cross-modal conditioning in generation.

UniMEL (Liu et al., 2024) is a unified framework that combines textual and visual information for representation enhancement and employs embedding-based retrieval for MEL.

D.3 Evaluation Metrics

H@k indicates whether the ground truth appears within the top k positions of the ranking list generated by the model: $\mathrm{H@k} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\big(\mathrm{rank}(i) \leq k\big)$, where N denotes the total number of samples, and $\mathbb{I}$ is an indicator function. In our experiment, we set the value of k to 1, 2, and 3. **MR** computes the average rank of the ground-truth entity: $\mathrm{MR} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{rank}(i)$, where lower values indicate better performance. **MRR** measures the average reciprocal rank of the ground-truth entity in the ranking list for all samples, which can be represented by $MRR = \frac{1}{N}\sum_{i}^{N} 1/rank(i)$.

Table 7: Ablation study on key components of ThinkLinker in the second stage. The best results are shown in **bold**.

Dataset		WikiMEL					WikiDiverse				
Metric		H@1	H@2	H@3	MR	MRR	H@1	H@2	H@3	MR	MRR
ThinkLinker†		**91.74**	**96.44**	97.66	**1.29**	**94.86**	**86.57**	94.11	96.71	**1.27**	**91.83**
Modality-level	(1) w/o Text	90.64	95.57	97.14	1.35	94.09	84.92	93.42	96.44	1.30	90.89
	(2) w/o Image	89.38	94.83	96.69	1.40	93.24	85.81	**94.17**	**97.12**	1.28	91.51
Module-level	(3) w/o Cross-Modal	91.06	96.19	97.62	1.34	94.45	84.65	93.42	96.30	1.32	90.71
	(4) w/o Intra-Modal	88.04	94.58	96.54	1.45	92.51	86.02	93.15	96.30	1.29	91.39
Feature-level	(5) w/o Local	91.00	96.01	97.56	1.30	94.41	86.36	93.76	96.64	1.28	91.68
	(6) w/o Global	90.23	95.49	97.19	1.34	93.87	84.51	93.35	96.30	1.30	90.68
(7) w/o Low Rank		91.06	94.30	**97.79**	1.31	94.52	86.43	94.17	96.64	1.28	91.77

Table 8: Performance and computational cost comparison between our low-rank fusion and tensor fusion.

Metrics		Low Rank Fusion	Tensor Fusion
H@1	WikiMEL	91.74	89.96
	WikiDiverse	87.05	83.21
Trainable Parameters (M)		4.5	86.9
GPU Memory (MB)		1017.69	3117.53
GFLOPs		6.31	24.45

E Efficiency and Effectiveness Analysis of Low-Rank Fusion

To evaluate the efficiency and effectiveness of the proposed low-rank tower for MEL, we compare the low-rank strategy adopted in the Low-Rank Multi-level Modal Fusion Framework with a variant in which the low-rank formulation is replaced by direct fusion using the original high-dimensional tensors. Evaluation is conducted from both performance and computational perspectives, using H@1 as the effectiveness metric and three efficiency indicators: the number of trainable parameters, GPU memory consumption, and computational complexity measured in GFLOPs. All experiments are performed under identical settings to ensure fair comparison.

As shown in Table 8, low-rank fusion consistently outperforms tensor fusion on both WikiMEL and WikiDiverse, yielding absolute H@1 improvements of 1.78 and 3.84 percentage points, respectively. This suggests that structured low-rank constraints help suppress redundant cross-modal interactions and emphasize more discriminative cues, leading to improved entity matching accuracy. In terms of computational efficiency, low-rank fusion requires only 4.5M parameters (over 19× fewer than tensor fusion), while reducing GPU memory from 3117.53 MB to 1017.69 MB and computation from 24.45 to 6.31 GFLOPs. This efficiency arises from low-rank decomposition, which effectively approximates high-order tensor interactions without the combinatorial overhead of full tensor fusion. Overall, low-rank fusion significantly reduces computational and memory overhead while delivering consistently superior performance across datasets, making it more suitable for large-scale multimodal entity linking and practical deployment.

F Ablation Study of The Second Stage

We conduct an ablation study in the second stage with results summarized in Table 7.

The ablation results show that removing individual components in the second stage still leads to performance degradation, exhibiting trends overall consistent with those observed in the first stage. This indicates that the modules and feature designs of our framework continue to play important roles after mention enrichment. Meanwhile, the model demonstrates stronger overall stability, as the performance drops caused by ablation are substantially smaller than those in the first stage. This can be attributed to the combination of first-stage pretraining and second-stage mention enrichment, which provides richer and more robust contextual representations and reduces the model's reliance on any single module or feature. Interestingly, for some configurations, H@1 decreases noticeably while H@2 and H@3 show slight improvements. This phenomenon arises because the second-stage verification module encourages the model to focus more on mention-centered discriminative information, thereby favoring more precise localization of the target entity.

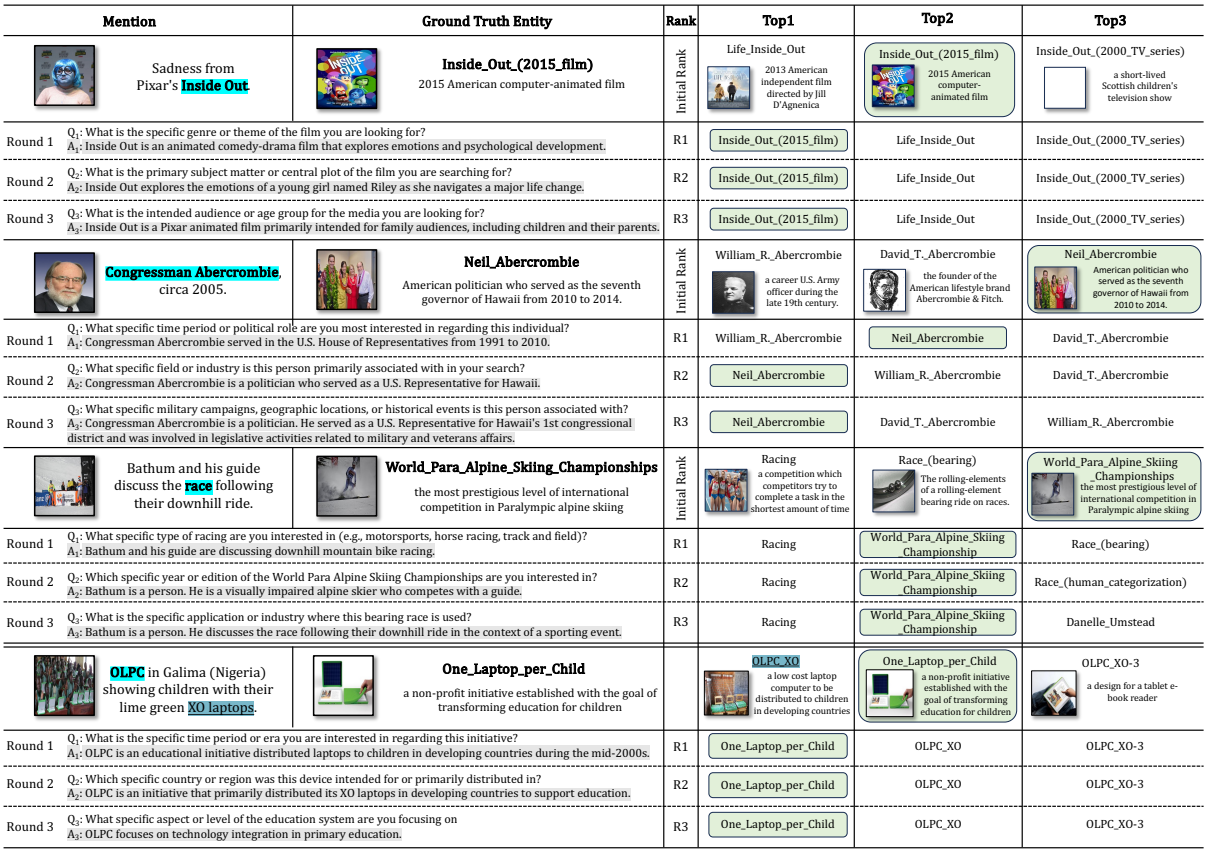

Figure 6: Case Study of ThinkLinker.

G Case Study

To qualitatively evaluate the effectiveness of the second-stage Knowledge-Driven Mention Representation Refinement, we conduct a case study with representative examples shown in Table 6.

Cases 1 and 2 show that when the original semantic-space context is short and lacks discriminative cues, the initial ranking tends to favor incorrect entities that are only surface-similar to the mention. In contrast, the multi-round LLM-based dialogue module starts from candidate entities in the knowledge space and performs iterative validation back into the semantic space, encouraging the model to generate evidence that is more discriminative for the current mention. In Case 1, a single dialogue round is sufficient to promote the correct entity to the top rank. In Case 2, the first round significantly improves the rank of the ground-truth entity, while the second round introduces the key attribute "politician", which clearly distinguishes the target entity from the previously top-ranked distractor.

Case 3 presents a failure case. Due to the abstract nature of the mention and the ambiguity of the surrounding context, the model shifts its focus to another, easier-to-resolve entity and generates detailed information about it. Although some context beneficial for disambiguation is also produced and the rank of the ground-truth entity is improved, this process simultaneously introduces hallucinations and noise. Such severely underspecified mentions are challenging even for human annotators, indicating that additional safeguards or specialized strategies are required for these cases.

Case 4 focuses on sentences containing multiple highly similar mentions. Because the baseline representation concatenates each mention with the same sentence context, co-occurring mentions often receive highly similar embeddings, which can lead to erroneous linking. In contrast, candidate-guided dialogue generates differentiated textual expansions for each mention, introducing additional discriminative cues and effectively reducing confusion in multi-mention scenarios.