Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.085456

ARTICLE

A Two-Stage Decoupled Matching Network for Multimodal Entity Linking

Huayu Li[1], **Xiang Wang**[1], **Jia Luo**[2,3,4,*], **Xiaotong He**[1] **and Peiying Zhang**[1]

[1]Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao, China
[2]Interdisciplinary Faculty of Science and Engineering, Shimane University, Shimane, Japan
[3]College of Economics and Management, Beijing University of Technology, Beijing, China
[4]Chongqing Research Institute, Beijing University of Technology, Chongqing, China

*Corresponding Author: Jia Luo. Email: jialuo@riko.shimane-u.ac.jp

Received: 11 May 2026; Accepted: 01 July 2026; Published: 13 August 2026

ABSTRACT: Multimodal Entity Linking (MEL) aims to map ambiguous mentions in multimodal contexts to their corresponding entities in a multimodal knowledge base. However, existing methods still face limitations in terms of feature extraction granularity, the depth of cross-modal interaction, and architectural coupling. To address these issues, we propose a Two-stage Decoupled Matching Network (TDMN) for multimodal entity linking. The matching process is divided into two stages: intra-modal matching and cross-modal interaction. In the intra-modal stage, textual and visual inputs are processed independently. The framework then proceeds to the cross-modal interaction stage, following the principle of "enhancement prior to interaction." Specifically, unimodal features are first refined through a parallel dual-attention network consisting of Global Relational Attention and Adaptive Sharpening Attention, together with a multi-granularity calibration fusion module. Based on the refined representations, cross-modal alignment is subsequently performed within a symmetric bidirectional interaction architecture, in which a gated residual mechanism is introduced to facilitate information fusion. Experiments conducted on the public benchmark datasets WikiMEL and WikiDiverse demonstrate the effectiveness of TDMN. Compared with the M^3EL baseline, TDMN achieves absolute improvements of 1.39% and 1.88% in MRR and Hits@1, respectively, on the WikiDiverse dataset. In addition, compared with MIMIC, TDMN improves MRR and Hits@1 by 0.8% and 1.21%, respectively, on the WikiMEL dataset. These results support the effectiveness of the proposed approach.

KEYWORDS: Multimodal entity linking; multi-granularity feature fusion; attention mechanism; feature enhancement; multimodal representation learning

1 Introduction

Driven by the exponential proliferation of information, Entity Linking (EL) seeks to map ambiguous entity mentions within textual data to their corresponding unique entities in a knowledge base, thereby constituting a critical prerequisite for downstream applications such as dialogue systems [1]. In a parallel context, recent investigations into large-scale multi-label text classification have examined the alignment between textual inputs and expansive, dynamic semantic spaces [2].

However, when mentions are accompanied by both textual and visual modalities, conventional unimodal text-based approaches frequently prove inadequate for effective disambiguation. As illustrated in Fig. 1, reliance exclusively upon the textual descriptor "Jaguar" fails to ascertain whether the mention denotes the animal, the military aircraft, or the automotive manufacturer; conversely, visual cues within the

Copyright © 2026 The Authors. Published by Tech Science Press. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

accompanying image, such as vehicular characteristics, can unambiguously identify the correct target entity. This cross-modal complementarity has catalyzed the advancement of Multimodal Entity Linking (MEL). Moreover, parallel investigations in vision-language pre-training and visual matching have demonstrated that feature fusion and attention-based interaction mechanisms substantially augment the performance of downstream tasks [3,4]. These findings collectively reinforce the premise that high-fidelity modality-specific representations and robust cross-modal alignment are paramount to achieving precise disambiguation within the MEL paradigm.

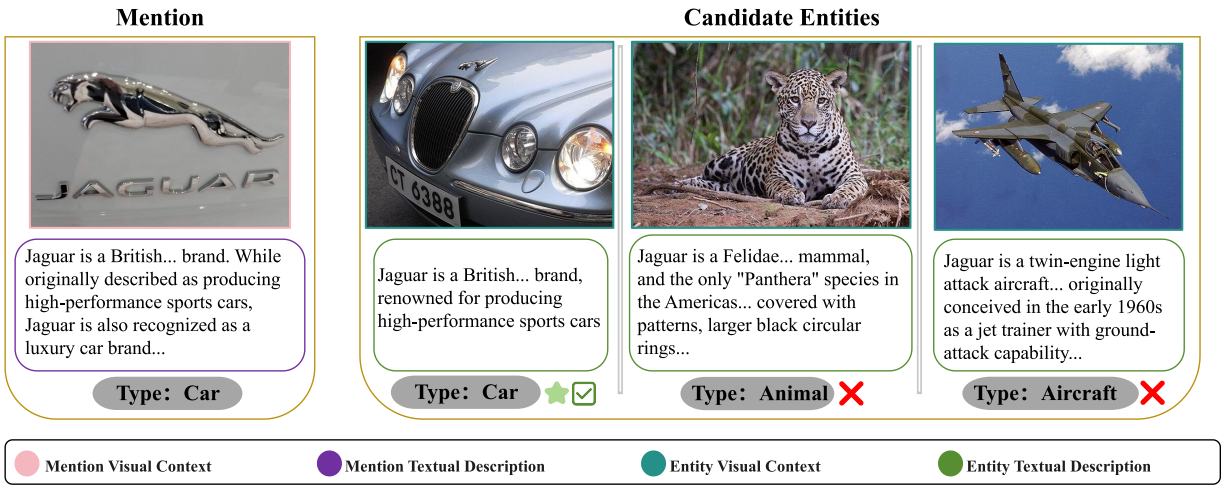

Figure 1: An illustrative example of multimodal entity linking for the mention "Jaguar". The mention is accompanied by visual and textual context, while each candidate entity is represented by visual and textual information. The task is to identify the correct entity from multiple candidates.

In the context of Multimodal Entity Linking, rigorous unimodal semantic modeling is indispensable for achieving precise disambiguation. However, contemporary unimodal feature extraction methodologies frequently exhibit an imbalance in representational granularity; specifically, reliance upon singular attention mechanisms tends to either overemphasize local details [5] or neglect critical local alignment dependencies [6]. To mitigate these deficiencies, cross-modal interaction mechanisms have been extensively incorporated. Nonetheless, earlier late-fusion strategies [7] exhibit insufficient inter-modal synergy, whereas early interaction paradigms [8] initiate cross-modal integration prior to the adequate optimization of unimodal features, thereby inducing noise propagation. While recent methodologies have sought to jointly optimize intra-modal feature enhancement and inter-modal alignment within a unified architectural framework [9–11], such coupled designs not only exacerbate optimization complexities but also constrain the theoretical performance ceiling of the model.

Specifically, the principal contributions of this work are articulated as follows:

- We propose a decoupled architecture for multimodal interaction. To circumvent the limitations of existing methodologies that inextricably couple feature extraction with cross-modal alignment, the proposed framework decouples the matching procedure into two discrete stages: intra-modal enhancement and inter-modal interaction. This architectural paradigm alleviates the complexities associated with multi-objective joint optimization and elevates the theoretical performance ceiling. Empirical ablation studies substantiate this design; reverting the framework to a conventional coupled co-attention architecture precipitates a decline in Mean Reciprocal Rank (MRR) of 2.23 and 2.87 percentage points on the WikiMEL and WikiDiverse datasets, respectively.

- We design a multi-granularity feature enhancement scheme. To mitigate the granularity imbalance inherent in conventional feature extraction, we construct a parallel network comprising Global Relational Attention (GRA) and Adaptive Sharpening Attention (ASA) during the intra-modal enhancement phase, which is subsequently followed by a Multi-granularity Calibration Fusion (MCF) module for feature refinement. Ablation studies conducted on the WikiDiverse dataset indicate that the removal of either the GRA or ASA branch induces an MRR reduction of 2.60 and 4.08 percentage points, respectively. Furthermore, substituting the MCF module with a rudimentary concatenation operation results in a further MRR degradation of 2.59 percentage points. These empirical findings corroborate that this functionally complementary architecture concurrently captures global semantic structures and salient local details, thereby guaranteeing high-fidelity unimodal representations prior to cross-modal alignment.
- We achieve state-of-the-art performance on mainstream benchmark datasets. Extensive comparative experiments were executed on the public WikiMEL and WikiDiverse datasets. The empirical results demonstrate that TDMN significantly outperforms existing state-of-the-art methodologies. Relative to the baseline model M^3EL, TDMN yields improvements of 1.39 and 1.88 percentage points in MRR and Hits@1, respectively, on the WikiDiverse dataset. These results robustly validate the efficacy of the proposed approach in navigating complex multimodal disambiguation scenarios.

2 Related Work

2.1 Interaction Architectures

Early Multimodal Entity Linking methodologies predominantly employed late-fusion strategies, wherein the textual and visual modalities associated with mentions and candidate entities were independently encoded, projected into a unified embedding space, and aligned via similarity estimation [7]. Although computationally straightforward, such architectures are inherently constrained by a paucity of cross-modal interaction during the representation learning phase. Consequently, in scenarios where the textual context is insufficient, the complementary utility of visual information for entity disambiguation remains suboptimally exploited.

To augment cross-modal representation learning, subsequent investigations have integrated vision-language pre-trained models, such as CLIP [12] and ViLT [13], or developed explicit modality interaction mechanisms. For instance, MIMIC [10], DRIN [11], and GHMFC [8] optimize multimodal fusion by modeling intra-modal and cross-modal interactions, facilitating dynamic feature updates, and employing co-attention mechanisms, respectively. Despite the progressive evolution of contemporary MEL methodologies from late-fusion paradigms to explicit interaction modeling, intra-modal feature enhancement and cross-modal alignment remain inadequately decoupled.

2.2 Feature Extraction

Existing multimodal interaction methods predominantly employ attention mechanisms for single-granularity modeling, thereby limiting their ability to capture global semantics and fine-grained local details simultaneously. Prior studies demonstrate that multi-granularity modeling can effectively improve representational completeness [14], whereas single-granularity modeling often produces an imbalance between localized details and global semantics.

This limitation partly arises from local weighting mechanisms. As demonstrated in OT-MEL [5], excessive emphasis on local associations can compromise a model's capacity for global semantic matching. Moreover, methods such as Zheng et al. [6] employ separate encoders to extract and aggregate

multimodal features. Such architectural separation restricts the model's ability to capture localized cross-modal information during feature extraction, thereby impeding the sufficient modeling of fine-grained cross-modal associations.

To address these limitations, recent studies have pursued multi-granularity alignment through hierarchical contrastive learning to jointly capture semantics at global and local levels [9]. Nevertheless, these approaches remain primarily dependent on auxiliary alignment objectives or supplementary attention modules, leaving fine-grained cross-modal perception during feature extraction insufficiently developed and warranting further improvement.

3 Method

This chapter presents a detailed exposition of our proposed multimodal entity linking framework, as illustrated in Fig. 2. The framework employs a decoupled matching network to systematically process both intra-modal and cross-modal information.

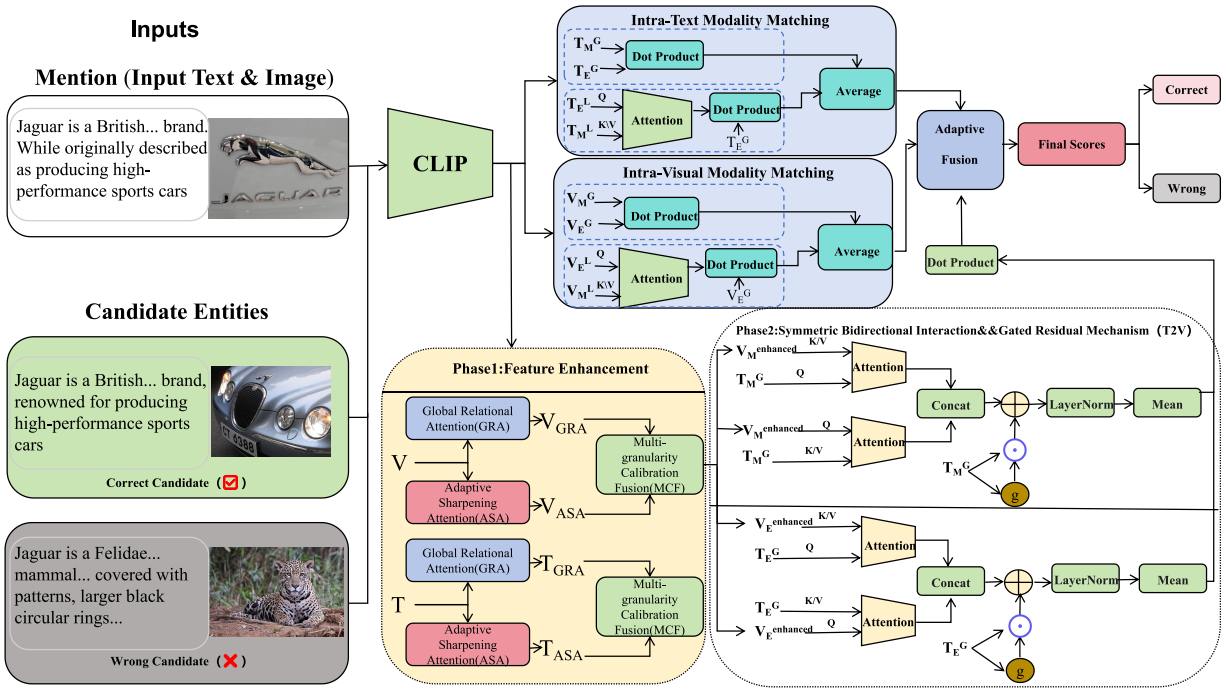

Figure 2: The overall architecture of the TDMN model. It primarily comprises a feature extraction module, an intra-modal matching module, and a bidirectional cross-modal interaction module. The bidirectional cross-modal interaction module consists of two components: a feature enhancement phase (Phase 1) and a symmetric bidirectional interaction and gated residual mechanism phase (Phase 2). The architectural diagram illustrates Phase 2 using T2V as an example. The features processed by each module are ultimately fed into an adaptive fusion module to generate the final score. The symbols ⊕ and ⊙ denote element-wise addition and multiplication of features, respectively, and g represents the gating mechanism.

3.1 Problem Formulation

The Multimodal Entity Linking task aims to disambiguate a given mention M occurring within a multimodal context by mapping it to its unique ground-truth entity E^* within a multimodal knowledge base K.

Formally, a mention M is defined as a triplet $M = (N_M, T_M, V_M)$, comprising its surface name N_M, textual context T_M, and visual context V_M. The knowledge base K encompasses a set of entities $\{E_1, E_2, \ldots, E_N\}$, wherein each entity E_i is analogously structured as $E_i = (N_{E_i}, T_{E_i}, V_{E_i})$, representing its entity name, textual description, and visual representation, respectively. The primary objective is to learn a scoring function $S(M, E_i)$ that quantifies the semantic compatibility between the mention M and each candidate entity $E_i \in \mathcal{E}_c$, thereby identifying the entity yielding the maximum score as the final linking prediction:

$$E^* = \arg\max_{E_i \in \mathcal{E}_c} S(M, E_i) \tag{1}$$

3.2 Initial Feature Extraction and Preprocessing

The proposed framework leverages the pre-trained CLIP model as a unified feature extractor, exploiting its dual-stream architecture to concurrently extract dual-granularity representations, encompassing both global and local features, across the textual and visual modalities. For notational brevity, the subscript $X \in \{M, E\}$ is employed hereinafter to denote a mention and an entity, respectively.

Textual Features: The surface name N_X and the textual description T_X are concatenated to construct the input sequence. Upon propagation through the CLIP text encoder, the output vector corresponding to the special token appended at the terminal position of the sequence is designated as the global textual feature T_X^G, whereas the output vectors corresponding to the remaining tokens constitute the local textual features T_X^L.

Visual Features: The input image is partitioned into a sequence of patches, with a [CLS] token prepended to the sequence. Following processing by the CLIP visual encoder (ViT), the output vector corresponding to the [CLS] token serves as the global visual feature V_X^G, whereas the output vectors corresponding to the remaining patches constitute the local visual features V_X^L.

3.3 Decoupled Matching Network

3.3.1 Intra-Modal Matching

Leveraging these multi-granularity features, we compute the intra-modal matching scores between the mention and each candidate entity separately within the textual and visual modalities. This design simultaneously captures global semantic consistency and fine-grained local alignment.

1. Textual Intra-modal Matching
   The textual intra-modal matching score S_{text} is derived by fusing two sub-scores: $S_{\text{text}}^{\text{G2G}}$ and $S_{\text{text}}^{\text{G2L}}$.
   Global-to-Global Matching. This computes the dot product between the mention's global textual feature T_M^G and the candidate entity's global textual feature T_E^G, formulated as:

   $$S_{\text{text}}^{\text{G2G}} = T_M^G \cdot (T_E^G)^\top \tag{2}$$

   Global-to-Local Matching. To capture token-level alignment, we employ an attention mechanism. We treat the candidate entity's local features T_E^L as the query (Q), while the mention's local textual features T_M^L serve as the key (K) and value (V). For each candidate entity, a context-aware aggregated representation, denoted as $Cagg_{\text{text}}$, is extracted from the mention's local textual features:

   $$Cagg_{\text{text}} = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{3}$$

Subsequently, the dot product between this aggregated vector and the entity's global feature is computed as $S_{\text{text}}^{\text{G2L}} = \hat{T}_E^G \cdot (Cagg_{\text{text}})^\top$. The final textual intra-modal matching score is the average of these two sub-scores: $S_{\text{text}} = (S_{\text{text}}^{\text{G2G}} + S_{\text{text}}^{\text{G2L}})/2$.

2. Visual Intra-modal Matching
The matching computation for the visual modality is completely analogous to that of the textual modality. Utilizing the visual global features (V_M^G, V_E^G) and local patch features (V_M^L, V_E^L) of the mention and candidate entity, the final visual intra-modal matching score $S_{visual} = (S_{visual}^{\text{G2G}} + S_{visual}^{\text{G2L}})/2$ is derived through equivalent G2G dot-product computations and G2L cross-attention aggregation.

3.3.2 Bidirectional Cross-Modal Interaction

1. Feature Enhancement within the Interaction Module
To mitigate limitations in feature granularity, we perform deep feature enhancement for each modality prior to cross-modal interaction. Specifically, the global representations, initially of dimension $\mathbb{R}^{B\times D}$, are reshaped to $\mathbb{R}^{B\times 1\times D}$ and concatenated with the local feature sequences of dimension $\mathbb{R}^{B\times L\times D}$ to form a unified contextual sequence X. This sequence is processed by a parallel dual-attention network to capture complementary feature perspectives. The resulting outputs are subsequently integrated via a multi-granularity calibration fusion mechanism, yielding enhanced unimodal representations for the subsequent cross-modal interaction stage.
Global Relational Attention (GRA). This mechanism quantifies global feature correlations by computing angular relationships within a normalized space, thereby focusing on macroscopic semantic structures. Internally, the unified input sequence X undergoes a linear transformation to generate a joint representation, which is then split and reshaped along the channel dimension to produce the query (q), key (k), and value (v) tensors. The core design involves applying L2 normalization to q and k, denoted as $\hat{q}$ and $\hat{k}$, and computing the attention output as $\hat{q}(\hat{k}^\top v)$. To ensure numerical stability, the final output is formulated as a fixed-weighted sum of the original value tensor v and the attention output:

$$X_{\text{GRA}} = \frac{1}{2}v + \frac{1}{\pi}\hat{q}\cdot(\hat{k}^\top \cdot v) \tag{4}$$

Adaptive Sharpening Attention (ASA). While sharing the same input and q, k, v generation pipeline as GRA, ASA addresses the rigidity of fixed scaling factors in conventional attention mechanisms by introducing a learnable temperature parameter τ. This design enables the model to dynamically adjust the sharpness of the attention distribution based on input characteristics, thereby enhancing its capacity to extract highly discriminative information. The output is computed as:

$$X_{\text{ASA}} = \left[\text{Softmax}\left(\tau\cdot\hat{q}\hat{k}^\top\right)v\right]^\top, \quad \text{where} \quad \hat{q} = \frac{q}{\|q\|}, \quad \hat{k} = \frac{k}{\|k\|} \tag{5}$$

Multi-granularity Calibration Fusion (MCF). To effectively integrate the features generated by the parallel attention networks, we design a cross-fusion module. This module dynamically calibrates and enhances the two feature streams by computing their global compatibility.
Specifically, for a given modality, the module receives X_{GRA} and X_{ASA}. For each feature stream $X_i \in \{X_{\text{GRA}}, X_{\text{ASA}}\}$, global contextual information is extracted via average pooling and max pooling. These pooled vectors are independently processed through bottleneck MLP networks and combined via element-wise addition to yield a compact global descriptor d_i:

$$d_i = \text{MLP}_{\text{avg}}(\text{AvgPool}(X_i)) + \text{MLP}_{\text{max}}(\text{MaxPool}(X_i)) \tag{6}$$

Subsequently, the module projects the two global descriptors, d_{GRA} and d_{ASA}, into the query and key spaces, respectively, and computes their element-wise product to derive the global compatibility vector α_{compat}, as formulated below:

$$\alpha_{\mathrm{compat}} = \frac{\mathrm{MLP}_Q(d_{\mathrm{GRA}}) \odot \mathrm{MLP}_K(d_{\mathrm{ASA}})}{\sqrt{D}} \tag{7}$$

Finally, the global compatibility vector is transformed into normalized attention weights via the Softmax function; these weights are then element-wise multiplied with the global relation attention X_{GRA} to dynamically modulate the latter.

Taking the textual branch as an example, the enhanced textual representation is formulated as:

$$T^{\mathrm{enhanced}} = \mathrm{Softmax}(\alpha_{\mathrm{compat\text{-}T}}) \odot X_{\mathrm{GRA\text{-}T}} \tag{8}$$

Analogously, the enhanced visual representation is computed as:

$$V^{\mathrm{enhanced}} = \mathrm{Softmax}(\alpha_{\mathrm{compat\text{-}V}}) \odot X_{\mathrm{GRA\text{-}V}} \tag{9}$$

2. Symmetric Bidirectional Interaction Pathway and Gated Residual Mechanism

Based on the multi-granularity enhanced uni-modal features, the model devises symmetric bi-directional interaction pathways comprising text-guided (T2V) and vision-guided (V2T) branches. Within each pathway, the model executes independent bi-directional query operations and incorporates a gated residual mechanism to generate the final representation.

Text-guided Visual Interaction Pathway (T2V). Using the mention as an example, the global textual feature T_M^G is first linearly projected to form the query (Q), while the enhanced visual feature sequence V_M^{enhanced} is projected to generate the key (K) and value (V). Scaled dot-product attention is then applied to compute a text-centric contextual vector $C_{T\leftarrow V}$:

$$C_{T\leftarrow V} = \mathrm{LayerNorm}\left(\mathrm{Softmax}\left(\frac{Q_T K_V^{\top}}{\sqrt{d_k}}\right) V_V\right) \tag{10}$$

Conversely, to compute the vision-centric contextual vector $C_{V\leftarrow T}$, the enhanced visual sequence V_M^{enhanced} serves as Q, while T_M^G is projected as K and V. These two context vectors, derived from complementary interaction perspectives, are aligned and concatenated to form a visually augmented contextual representation, denoted as C_M^{T2V}.

Subsequently, a gated residual mechanism is employed to adaptively regulate the retention of the original global textual feature T_M^G. This feature is first broadcast along the sequence dimension and element-wise modulated by a learnable gating score g, before being integrated with the newly generated context vector C_M^{T2V} via weighted fusion. The gating score is computed as $g = \tanh(\mathrm{Linear}(T_M^G))$. The final representation for the mention under the T2V pathway, F_M^{T2V}, is obtained as:

$$F_M^{\mathrm{T2V}} = \mathrm{LayerNorm}(C_M^{\mathrm{T2V}} + T_M^G \odot g) \tag{11}$$

Analogously, the entity's final representation F_E^{T2V} is derived. The matching score for this pathway is calculated via the dot product of the mean-pooled mention and entity representations:

$$S_{\mathrm{T2V}} = \mathrm{Mean}(F_M^{\mathrm{T2V}}) \cdot (\mathrm{Mean}(F_E^{\mathrm{T2V}}))^{\top} \tag{12}$$

Vision-guided Textual Interaction Pathway (V2T). This pathway operates symmetrically to the T2V pathway. It performs bidirectional interaction between the global visual features and the enhanced

textual features, applying the identical gated residual mechanism to yield the final representations F_M^{V2T} and F_E^{V2T} for the mention and entity, respectively, along with the corresponding matching score S_{V2T}.

3.4 Adaptive Score Fusion and Training Objective

3.4.1 Adaptive Score Fusion

Our framework introduces an adaptive score fusion module that concatenates the global textual and visual features, processes the concatenated vector via a multi-layer perceptron (MLP) to generate three unnormalized logits, and applies a Softmax normalization subject to a minimum weight constraint to derive the adaptive weights $w = (w_{\mathrm{text}}, w_{\mathrm{visual}}, w_{\mathrm{cross}})$. The final matching score S_{final} is computed as a weighted aggregation of the individual pathway scores, modulated by the predicted weights w:

$$S_{\mathrm{final}} = w_{\mathrm{text}} \cdot S_{\mathrm{text}} + w_{\mathrm{visual}} \cdot S_{\mathrm{visual}} + w_{\mathrm{cross}} \cdot \frac{S_{\mathrm{T2V}} + S_{\mathrm{V2T}}}{2} \tag{13}$$

3.4.2 Multi-Level Joint Training Objective

We construct a multi-level joint training objective designed to synergistically optimize the initial feature representations, the intermediate matching modules, and the final entity linking objective.

- Matching Loss: Both the primary matching loss and the four independent auxiliary matching losses employ the standard cross-entropy loss function. The primary matching loss optimizes the final output following adaptive fusion, formulated as:

$$\mathcal{L}_{\mathrm{match}} = -\log(P(E_{\mathrm{gt}}|M)) \tag{14}$$

  The independent auxiliary losses operate on the output scores of the intermediate matching modules. These include the textual intra-modal score S_{text}, the visual intra-modal score S_{visual}, and the bidirectional cross-modal matching scores S_{T2V} and S_{V2T}, corresponding to the auxiliary losses $\mathcal{L}_T$, $\mathcal{L}_V$, $\mathcal{L}_{\mathrm{T2V}}$, and $\mathcal{L}_{\mathrm{V2T}}$, respectively.
- Intra-modal Contrastive Learning Loss: To enhance the discriminability of intra-modal feature representations, we construct intra-class negative samples (mismatched samples from the same entity set or the same mention set) and inter-class negative samples (mismatched samples from different sets, i.e., entity vs. mention) for a given anchor. For a positive sample pair comprising an entity text embedding T_e^i and its corresponding mention text embedding T_m^i, the intra-modal contrastive loss $\mathcal{L}(T_e^i, T_m^i)$ is defined as:

$$\mathcal{L}(T_e^i, T_m^i) = -\log \frac{\theta(T_e^i, T_m^i)}{\theta(T_e^i, T_m^i) + \beta \cdot \Phi_{\mathrm{intraclass}} + \gamma \cdot \Phi_{\mathrm{interclass}}} \tag{15}$$

  where $\Phi_{\mathrm{intraclass}}$ and $\Phi_{\mathrm{interclass}}$ denote the similarity aggregation terms for the intra-class and inter-class negative samples, respectively:

$$\begin{cases} \Phi_{\mathrm{intraclass}} = \sum\limits_{\mathbf{T}_e^j \in N_e} \theta(\mathbf{T}_e^i, \mathbf{T}_e^j) \\ \Phi_{\mathrm{interclass}} = \sum\limits_{\mathbf{T}_m^j \in N_m} \theta(\mathbf{T}_e^i, \mathbf{T}_m^j) \end{cases} \tag{16}$$

  Here, $N_e = \{T_e^j | i \neq j\}$ and $N_m = \{T_m^j | i \neq j\}$. The function $\theta(x, y) = e^{\frac{\delta(x,y)}{\tau}}$ represents the exponential cosine similarity regulated by the temperature hyperparameter τ. The hyperparameters β and γ

regulate the relative contributions of intra-set and cross-set negative samples in the denominator, respectively. Specifically, β governs the influence of intra-set negative samples, enabling the model to better discriminate among samples within the same set. Conversely, γ modulates the impact of cross-set negative samples, directing the model to pay closer attention to the interference caused by erroneous cross-set matches in entity linking predictions. A sensitivity analysis of β and γ is further provided in the experimental section. Given the asymmetry of the contrastive loss, the loss for the inverse pair (T_m^i, T_e^i), denoted as $\mathcal{L}(T_m^i, T_e^i)$, is computed independently. The contrastive losses for the visual modality, $\mathcal{L}(V_e^i, V_m^i)$ and $\mathcal{L}(V_m^i, V_e^i)$, are calculated in the same manner. The total intra-modal contrastive learning loss is the average of these individual components:

$$\mathcal{L}_{\text{cl}} = \text{avg}\left(\sum_i \left[\mathcal{L}(T_e^i, T_m^i) + \mathcal{L}(T_m^i, T_e^i) + \mathcal{L}(V_e^i, V_m^i) + \mathcal{L}(V_m^i, V_e^i)\right]\right) \tag{17}$$

- Total Loss Function: The overall loss of the model is the sum of the aforementioned components, facilitating end-to-end joint optimization across different levels of the model with equal weights:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{match}} + \mathcal{L}_T + \mathcal{L}_V + \mathcal{L}_{\text{T2V}} + \mathcal{L}_{\text{V2T}} + \mathcal{L}_{\text{cl}} \tag{18}$$

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets

We evaluate our method on two widely adopted public benchmark datasets for Multimodal Entity Linking: WikiMEL [8] and WikiDiverse [15]. Detailed statistical information for both datasets is presented in Table 1.

Table 1: Statistics of the two datasets. "Num." and "Img." denote number and image, respectively.

Statistic	WikiMEL	WikiDiverse
# Num. of sentences	22,070	7405
# Num. of entities	109,976	132,460
# entities with Img.	67,195	67,309
# Num. of mentions	25,846	15,093
# Img. of mentions	22,136	6697
# mentions in train	18,092	11,351
# mentions in valid	2585	1664
# mentions in test	5169	2078

4.1.2 Evaluation Metrics

Model performance is evaluated using two standard metrics: Mean Reciprocal Rank (MRR) and Hits@N (where $N \in \{1, 3, 5\}$). Their formal definitions are provided below:

MRR assesses the model's overall capability to rank the correct entity at the top of the candidate list. It is computed as the average of the reciprocal ranks of the correct target entities across all queries:

$$\text{MRR} = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{\text{rank}_i} \tag{19}$$

Hits@N measures the proportion of queries in which the correct target entity appears within the top-N positions, directly reflecting the model's hit accuracy at various cutoff depths. Its formal definition is given by:

$$\text{Hits@N} = \frac{1}{Q}\sum_{i=1}^{Q}\mathbb{I}(\text{rank}_i \le N) \tag{20}$$

4.2 Comparison with Baselines

To validate the superiority of our proposed approach, we compare TDMN against state-of-the-art methods. To ensure a fair and rigorous comparison, we reproduced the M^{3}EL model under identical experimental conditions, whereas the results for other baselines, such as MIMIC, are directly cited from their original papers. The detailed comparison results are presented in Table 2.

1. On the WikiMEL dataset, TDMN achieves absolute improvements of 0.80 and 1.21 percentage points in MRR and Hits@1, respectively, over the strong MIMIC baseline. On the more challenging WikiDiverse dataset, TDMN exhibits a substantially more pronounced advantage over MIMIC, with Hits@1 improving by a remarkable 11.71 percentage points, underscoring the model's maximal performance gain in top-rank retrieval accuracy. Furthermore, compared to another robust baseline, M^{3}EL, TDMN achieves gains of 0.81 and 1.22 percentage points in MRR and Hits@1, respectively, on WikiMEL. On WikiDiverse, it outperforms M^{3}EL by 1.39 and 1.88 percentage points in MRR and Hits@1, respectively. Furthermore, we independently trained and evaluated both models across three different random seeds. A subsequent paired t-test yielded $p < 0.05$, ruling out the possibility of experimental error or random fluctuations.
2. These performance gains strongly validate the core design philosophy articulated in our introduction. TDMN's "enhance-before-interact" decoupled paradigm systematically addresses the limitations of shallow and coupled cross-modal interactions. Concurrently, its multi-granularity feature extraction and fusion scheme effectively mitigates granularity constraints, enabling robust and precise entity linking in complex multimodal scenarios.

4.3 Ablation Study

In this section, we conduct ablation studies to systematically analyze the impact of each key module on model performance.

To validate the effectiveness of the decoupled architecture, we replace the core module of TDMN with a coupled co-attention baseline for comparison. This baseline abandons the "enhancement before interaction" decoupled strategy, instead tightly coupling independent textual and visual features into a unified representation at an early stage, and reducing the original bidirectional cross-modal interaction to a post-fusion co-attention computation.

The results in Table 3 show that replacing the proposed decoupled architecture with the coupled co-attention baseline reduces model performance. Specifically, MRR and Hits@1 decrease by 2.23% and 2.79%, respectively, on WikiMEL, and by 2.87% and 4.05% on WikiDiverse. These results support the effectiveness of the proposed architecture: separating the "enhancement" and "interaction" phases into two stages reduces optimization complexity, enables more targeted optimization of each subtask, and improves the framework's overall matching performance.

To evaluate the multi-granularity feature extraction and fusion scheme, we conduct ablation studies on the parallel attention networks. Removing either GRA or ASA leads to performance declines. On WikiMEL, removing GRA decreases MRR and Hits@1 by 0.81% and 1.36%, respectively, whereas removing ASA yields reductions of 0.20% and 0.29%. On WikiDiverse, the corresponding decreases are 2.60% and 3.61% for GRA and 4.08% and 4.62% for ASA. These results indicate that the two complementary mechanisms contribute

to unimodal representation learning and cross-modal interaction. Moreover, replacing MCF with simple concatenation decreases MRR and Hits@1 by 1.13% and 1.86% on WikiMEL and by 2.59% and 4.19% on WikiDiverse. This result suggests that MCF integrates multi-granularity information by dynamically aligning features across representation streams.

Table 2: Performance comparison of entity linking on the WikiMEL and WikiDiverse datasets. All metrics are reported in percentages (%). The highest scores are highlighted in **bold**.

Methods	WikiMEL				WikiDiverse			
	MRR	Hits@1	Hits@3	Hits@5	MRR	Hits@1	Hits@3	Hits@5
BLINK [16]	81.72	74.66	86.63	90.57	69.15	57.14	78.04	85.32
BERT [17]	81.78	74.82	86.79	90.47	67.38	55.77	75.73	83.11
RoBERTa [18]	80.86	73.75	85.85	89.80	70.52	59.46	78.54	85.08
DZMNED [19]	84.97	78.82	90.02	92.62	67.59	56.90	75.34	81.41
JMEL [7]	73.39	64.65	79.99	84.34	48.19	37.38	54.23	61.00
VELML [6]	83.42	76.62	88.75	91.96	66.13	54.56	74.43	81.15
GHMFC [8]	83.36	76.55	88.40	92.01	70.99	60.27	79.40	84.74
CLIP [12]	88.23	83.23	92.10	94.51	71.69	61.21	79.63	85.18
ViLT [13]	79.46	72.64	84.51	87.86	45.22	34.39	51.07	57.83
ALBEF [20]	84.56	78.64	88.93	91.75	69.93	60.59	75.59	81.30
METER [21]	79.49	72.46	84.41	88.17	63.71	53.14	70.93	77.59
MIMIC [10]	91.82	87.98	95.07	96.37	73.44	63.51	81.04	86.43
M^3EL [22]	91.81	87.97	95.01	96.48	80.82	73.34	86.38	90.28
TDMN (Ours)	**92.62**	**89.19**	**95.34**	**96.92**	**82.21**	**75.22**	**87.87**	**91.15**

Table 3: Ablation study results for different components of TDMN on the WikiMEL and WikiDiverse datasets.The highest scores are highlighted in **bold**.

Configurations	WikiMEL				WikiDiverse			
	MRR	Hits@1	Hits@3	Hits@5	MRR	Hits@1	Hits@3	Hits@5
TDMN (Full Model)	**92.62**	**89.19**	95.34	96.92	**82.21**	**82.21**	**82.21**	**82.21**
w/o GRA	91.81	87.83	95.20	96.75	79.61	71.61	85.51	89.51
w/o ASA	92.42	88.90	95.24	**96.94**	78.13	70.60	83.88	87.34
w/o MCF	91.49	87.33	94.99	96.90	79.62	71.03	86.67	90.52
w/o Decoupled Arch	90.39	86.40	93.31	95.22	79.34	71.17	85.76	89.70
w/o CL	91.57	87.87	94.56	96.13	80.34	72.81	86.14	89.94
w/o E2M-CL	91.30	87.31	94.49	96.38	81.34	73.92	87.10	90.76
w/o M2E-CL	91.92	88.06	95.05	96.54	79.44	71.61	85.37	89.27
w/o Text-CL	92.34	88.74	**95.38**	96.65	81.66	74.35	87.30	91.05
w/o Visual-CL	91.11	87.25	94.37	96.00	80.17	72.86	85.51	89.51

To assess the bidirectional and bimodal formulation of the intra-modal contrastive learning objective, we conduct ablation studies on its individual components. Specifically, w/o CL removes the complete intra-modal contrastive loss; w/o E2M-CL and w/o M2E-CL remove the entity-to-mention and mention-to-entity terms, respectively; and w/o Text-CL and w/o Visual-CL remove the objectives for the textual and visual modalities. Removing the complete loss reduces performance on both datasets: MRR and Hits@1 decline by 1.05 and 1.32 percentage points on the first dataset and by 1.87 and 2.41 percentage points on the second. These results indicate that the loss improves the discriminability of entity and mention representations. The performance decreases observed for both w/o E2M-CL and w/o M2E-CL further suggest that the two directions provide complementary supervision. Removing Text-CL produces smaller declines, whereas removing Visual-CL has a greater effect, particularly on the second dataset, where MRR and Hits@1 decrease by 2.04 and 2.36 percentage points, respectively. This finding indicates that visual intra-modal constraints contribute more strongly to multimodal entity linking. Overall, the results support the effectiveness of the bidirectional and bimodal design and justify its added complexity.

4.4 Parameter Sensitivity Analysis

To assess the model's sensitivity to core hyperparameters and identify the optimal configuration, we conducted one-factor-at-a-time sensitivity experiments on seven key parameters using the WikiMEL and WikiDiverse datasets, with MRR serving as the evaluation metric. In each experiment, only a single target hyperparameter was varied, whereas all of the remaining hyperparameters were fixed at their default values.

Because WikiMEL and WikiDiverse exhibit distinct data distributions, the same hyperparameter may have different effective search ranges across the two datasets. Accordingly, the candidate ranges of selected hyperparameters were adjusted in accordance with dataset-specific characteristics to cover the principal regions of performance variation. Figs. 3 and 4 present the sensitivity curves for each parameter, with the x-axis uniformly denoting the relative Variation Level to eliminate dimensional discrepancies. Based on the observed trends, the optimal configurations and underlying mechanisms for each parameter are analyzed as follows: (1) Learning Rate (lr): To balance convergence speed and training stability, the optimal values for WikiMEL and WikiDiverse are 2e−5 and 1e−5, respectively; (2) Dropout Rate (dropout_rate): Both datasets achieve optimal performance at a rate of 0.2, effectively balancing regularization constraints with feature representation capability; (3) Number of Attention Heads (head_num): The optimal numbers of heads are 6 and 8 for WikiMEL and WikiDiverse, respectively. An appropriate number of heads facilitates the capture of multidimensional semantics while suppressing redundant computational noise; (4) Maximum Text Sequence Length (text_max_length): Both datasets reach peak performance at a length of 40, indicating that a locally compact context is sufficient to encompass key semantics while effectively filtering out background noise; (5) Loss Temperature Coefficient (loss_temperature): At a setting of 0.03, the model achieves an optimal balance between optimization stability and feature discriminability; (6) Intra-class negative sample weight (β): On both datasets, β achieves the best performance at 0.9, indicating that appropriately controlling the weight of intra-class negative samples helps prevent samples from the same class from being excessively separated, thereby preserving the semantic consistency of the representation space; (7) Inter-class negative sample weight (γ): On both datasets, γ achieves the best performance at 1.2, indicating that a stronger inter-class negative sample constraint helps improve the discriminability between entity and mention representations, whereas an excessively large weight may weaken positive sample alignment.

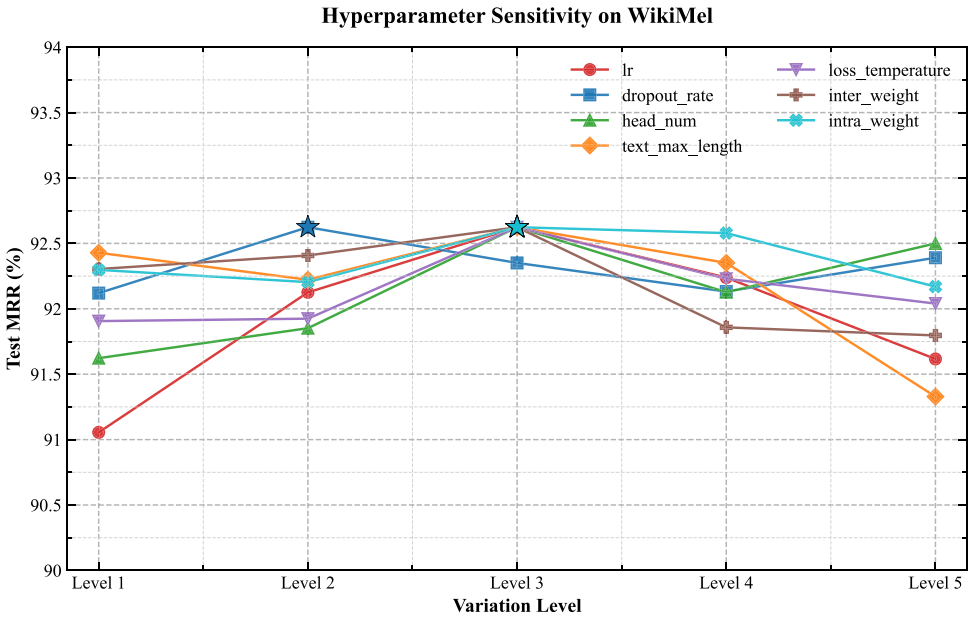

Figure 3: Comprehensive hyperparameter sensitivity analysis on the WikiMEL dataset. To eliminate dimensional discrepancies, the horizontal axis uniformly adopts relative parameter tiers (Level 1 to Level 5). The specific testing ranges for each parameter are as follows: lr $\in \{5e{-}6, 1e{-}5, 2e{-}5, 3e{-}5, 5e{-}5\}$; dropout_rate $\in \{0.1, 0.2, 0.3, 0.4, 0.5\}$; head_num $\in \{2, 4, 6, 8, 12\}$; text_max_length $\in \{32, 36, 40, 44, 48\}$; loss_temperature $\in \{0.01, 0.02, 0.03, 0.05, 0.07\}$; $\beta \in \{0.5, 0.7, 0.9, 1.1, 1.3\}$ for the intra-class negative sample weight; $\gamma \in \{0.8, 1.0, 1.2, 1.4, 1.6\}$ for the inter-class negative sample weight. The asterisks in the figure denote the points where each parameter achieves its best performance.

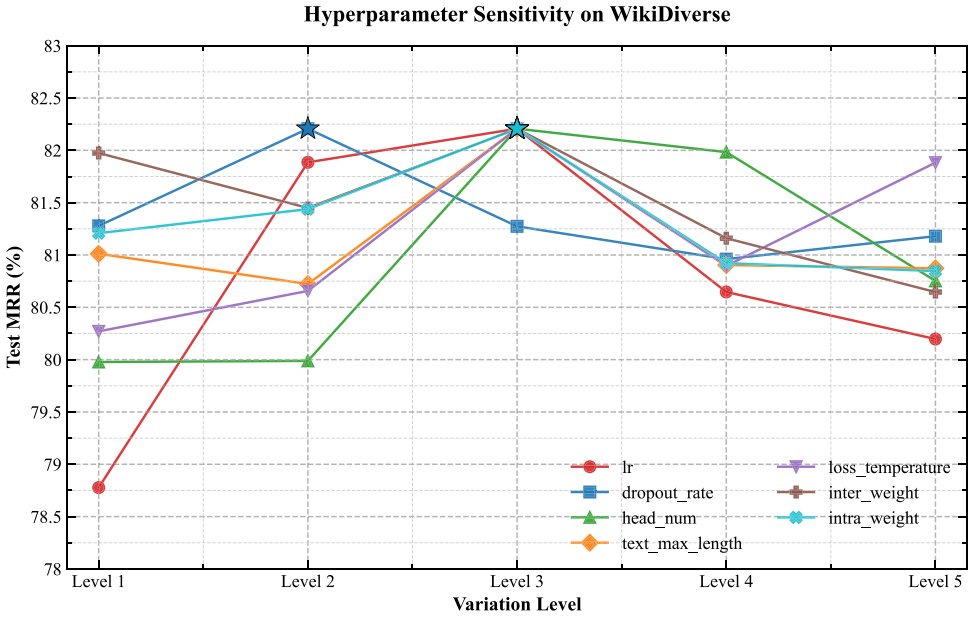

Figure 4: Comprehensive hyperparameter sensitivity analysis on the WikiDiverse dataset. The parameter settings are identical to those in Fig. 3, except for the following ranges: lr $\in \{1e{-}6, 5e{-}6, 1e{-}5, 2e{-}5, 3e{-}5\}$ and head_num $\in \{2, 4, 8, 16, 32\}$.

4.5 Qualitative Analysis

To further investigate the model's performance in multimodal entity linking, we conduct a qualitative analysis on several representative successful and failure cases selected from the WikiDiverse test set, as illustrated in Fig. 5.

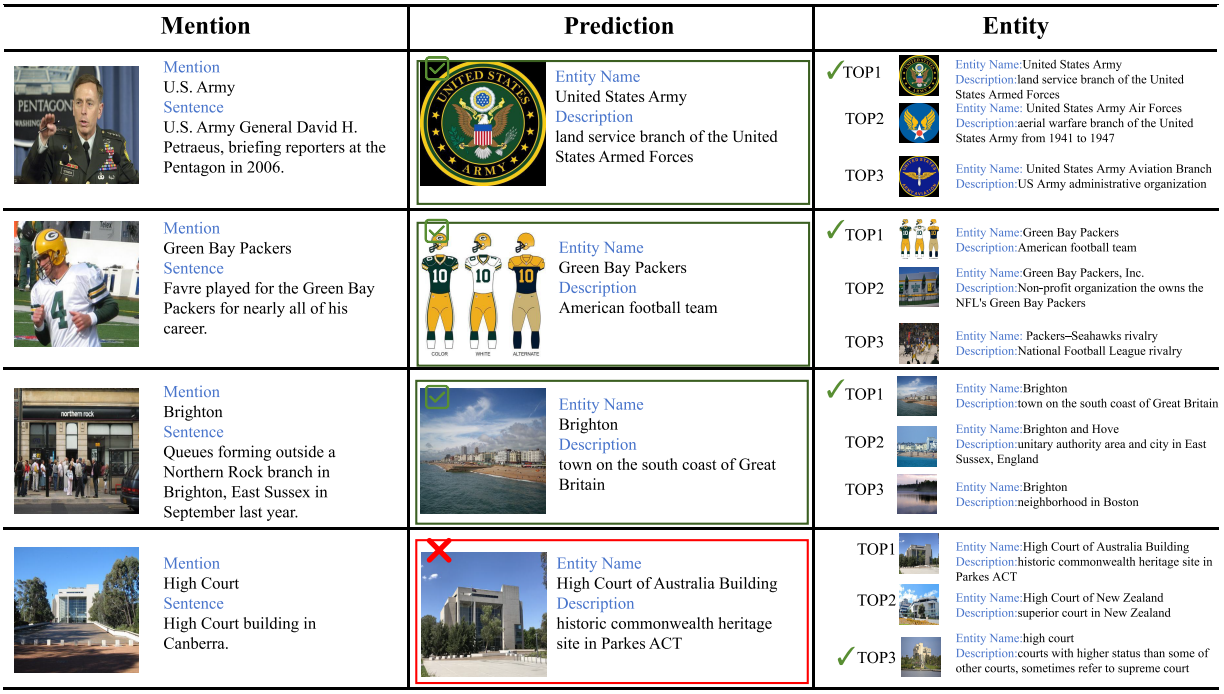

Figure 5: Representative success and failure cases on the WikiDiverse test set.

Successful cases demonstrate that the model can, to a certain extent, mitigate entity confusion arising from semantic similarity, visual resemblance, and insufficient contextual localization. In the "U.S. Army" case, the model correctly links the mention to United States Army among several closely related military organization entities, thereby demonstrating its ability to distinguish the target entity from affiliated branch entities. In the "Green Bay Packers" case, although the candidate entities share similar visual cues, including team logos, uniform colors, and game scenes, the model still correctly selects the team entity itself, thereby indicating its robustness against confusion among visually similar entities. In the "Brighton" case, the model exploits the geographic cues in "Brighton, East Sussex" to correctly link the mention to Brighton, thereby demonstrating its capacity to resolve entity confusion caused by insufficient contextual localization. Conversely, the failure cases show that the model remains susceptible to errors when mention boundaries are ambiguous. In the "High Court" case, the original context, "High Court building in Canberra," shifts the semantic representation of the mention toward a building entity, causing the model to incorrectly link it to High Court of Australia Building. This result indicates that when the mention boundary overlaps with the surrounding contextual semantics, the model still encounters difficulty distinguishing among entities at closely related levels of granularity, such as institutions, registries, and buildings.

5 Conclusion

Evaluations on WikiMEL and WikiDiverse show that the proposed TDMN decoupled matching framework consistently improves predictive performance. The architecture also exhibits generalizability

and offers methodological insights for related downstream applications, including question answering, semantic search, and cross-modal retrieval. Future work will refine the framework and systematically assess the applicability and scalability of the decoupled design across a broader range of multimodal fusion and understanding tasks.

Acknowledgement: Not applicable.

Funding Statement: This work was partially supported by the Beijing Natural Science Foundation under Grant 9242003, partially supported by the Natural Science Foundation of Chongqing, China under Grant CSTB2023NSCQ-MSX0391, partially supported by the National Natural Science Foundation of China under Grant 62471493, and partially supported by the Natural Science Foundation of Shandong Province under Grants ZR2023LZH017 and ZR2024MF066.

Author Contributions: Conceptualization: Huayu Li, Xiang Wang, Jia Luo; Methodology: Huayu Li, Xiang Wang; Validation: Huayu Li, Xiang Wang, Jia Luo, Xiaotong He, Peiying Zhang; Formal analysis: Huayu Li, Xiang Wang, Jia Luo, Xiaotong He, Peiying Zhang; Investigation: Huayu Li, Xiang Wang; Writing—original draft: Huayu Li, Xiang Wang, Jia Luo, Peiying Zhang; Writing—review and editing: Huayu Li, Xiang Wang, Jia Luo; Supervision: Huayu Li, Jia Luo, Xiaotong He, Peiying Zhang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the results of this study are openly available at https://github.com/seukgcode/MELBench (WikiMEL) and https://github.com/wangxw5/wikiDiverse (WikiDiverse).

Ethics Approval: Not applicable.

Conflicts of Interest: Given his role as Editorial Board Member of this journal, Peiying Zhang had no involvement in the peer review of this article and had no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to another journal editor. The authors declare no other conflicts of interest.

References

1. Ahmadvand A, Sahijwani H, Choi JI, Agichtein E. Concet: entity-aware topic classification for open-domain conversational agents. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management; 2019 Nov 3–7; Beijing, China. p. 1371–80.
2. Ren L, Liu Y, Ouyang C, Yu Y, Zhou S, He Y, et al. DyLas: a dynamic label alignment strategy for large-scale multi-label text classification. Inf Fusion. 2025;120(3):103081. doi:10.1016/j.inffus.2025.103081.
3. Fayou S, Ngo HC, Sek YW, Meng Z. Clustering swap prediction for image-text pre-training. Sci Rep. 2024;14(1):11879. doi:10.1038/s41598-024-60832-x.
4. Wang Z, Chen H, Yuan L, Ren Y, Tian H, Wang X. SiamMLT: siamese hybrid multi-layer transformer fusion tracker. Neural Process Lett. 2023;55(7):9651–67. doi:10.1007/s11063-023-11219-y.
5. Zhang Z, Sheng J, Zhang C, Liang Y, Zhang W, Wang S, et al. Optimal transport guided correlation assignment for multimodal entity linking. arXiv:2406.01934. 2024.
6. Zheng Q, Wen H, Wang M, Qi G. Visual entity linking via multi-modal learning. Data Intell. 2022;4(1):1–19. doi:10.1162/dint_a_00114.
7. Adjali O, Besançon R, Ferret O, Le Borgne H, Grau B. Multimodal entity linking for tweets. In: European Conference on Information Retrieval. Berlin/Heidelberg, Germany: Springer; 2020. p. 463–78.
8. Wang P, Wu J, Chen X. 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; 2022 Jul 11–15; Madrid, Spain. p. 938–48.
9. Song S, Li S, Zhao S, Li X, Wang C, Yu J, et al. DWE+: dual-way matching enhanced framework for multimodal entity linking. arXiv:2404.04818. 2024.

10. Luo P, Xu T, Wu S, Zhu C, Xu L, Chen E. Multi-grained multimodal interaction network for entity linking. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; 2023 Aug 6–10; Long Beach, CA, USA. p. 1583–94.
11. Xing S, Zhao F, Wu Z, Li C, Zhang J, Drin DX. Dynamic relation interactive network for multimodal entity linking. In: Proceedings of the 31st ACM International Conference on Multimedia; 2023 Oct 28–Nov 3; Ottawa, ON, Canada. p. 3599–608.
12. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning; 2021 Jul 18–24; Virtual. p. 8748–63.
13. Kim W, Son B, Kim I. VILT: vision-and-language transformer without convolution or region supervision. In: Proceedings of the International Conference on Machine Learning; 2021 Jul 18–24; Virtual. p. 5583–94.
14. Li J, Jie Z, Wang X, Zhou Y, Wei X, Ma L. Weakly supervised semantic segmentation via progressive patch learning. IEEE Trans Multimed. 2022;25:1686–99. doi:10.1109/tmm.2022.3152388.
15. Wang X, Tian J, Gui M, Li Z, Wang R, Yan M, et al. WikiDiverse: a multimodal entity linking dataset with diversified contextual topics and entity types. arXiv:2204.06347. 2022.
16. Wu L, Petroni F, Josifoski M, Riedel S, Zettlemoyer L. Scalable zero-shot entity linking with dense entity retrieval. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020 Nov 16–20; Virtual. p. 6397–407.
17. Devlin J, Chang MW, Lee K, Toutanova K. 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; 2019 Jun 2–7; Minneapolis, MN, USA. Vol. 1 (long and short papers). p. 4171–86.
18. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: a robustly optimized BERT pretraining approach. arXiv:1907.11692. 2019.
19. Moon S, Neves L, Carvalho V. Multimodal named entity disambiguation for noisy social media posts. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2018 Jul 15–20; Melbourne, VIC, Australia. p. 2000–8.
20. Li J, Selvaraju R, Gotmare A, Joty S, Xiong C, Hoi SCH. Align before fuse: vision and language representation learning with momentum distillation. Adv Neural Inf Process Syst. 2021;34:9694–705.
21. Dou ZY, Xu Y, Gan Z, Wang J, Wang S, Wang L, et al. 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; 2022 Jun 18–24; New Orleans, LA, USA. p. 18166–76.
22. Hu Z, Gutiérrez-Basulto V, Li R, Pan JZ. Multi-level matching network for multimodal entity linking. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1; 2025 Aug 3–7; Toronto, NA, Canada. p. 508–19.