

Article

Collaborative Multimodal Entity Linking via Multi-Channel Neural Cross-Modal Interaction with Synergistic Consistency Optimization

Huayu Li ¹, Xiaotong He ¹ , Hongjuan Pei ^{2,*} , Yujie Yuan ³ , Kai Liu ^{4,5}  and Peiying Zhang ¹ 

- ¹ Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China; lhyzj@upc.edu.cn (H.L.); s24070040@s.upc.edu.cn (X.H.); zhangpeiying@upc.edu.cn (P.Z.)
- ² School of Engineering Science, University of Chinese Academy of Sciences, Beijing 100049, China
- ³ School of Air Traffic Management, Civil Aviation University of China, Tianjin 300300, China; yjyuan@cauc.edu.cn
- ⁴ Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China; liukaiv@tsinghua.edu.cn
- ⁵ State Key Laboratory of Space Network and Communications, Tsinghua University, Beijing 100084, China
- * Correspondence: peihongjuan@ucas.ac.cn

Abstract

Multimodal entity linking (MEL) grounds entity mentions in text-image contexts to entries in a structured knowledge base; however, most existing systems still decompose a multimodal document into independent mention-level decisions. This formulation overlooks a central tension of real-world MEL: the evidence needed to disambiguate an ambiguous mention is often distributed across co-occurring mentions, visual context, and cross-modal consistency, rather than being contained in the mention itself. To address this limitation, we propose Collaborative Entity Linking through Multichannel Interaction (CELMI), a four-channel framework that jointly models textual semantics, visual perception, cross-modal alignment, and inter-mention collaboration. CELMI employs dual-level textual alignment, text-guided visual gating, learned cross-modal projection, and attention-based entity graph propagation. To stabilize joint optimization, we further introduce a multi-channel consistency objective that combines per-channel contrastive losses with an overall ranking loss, reducing channel dominance and representation collapse. Among conventional non-LLM/VLM MEL models, CELMI achieves the strongest MRR and Hits@1 performance on WikiMEL, RichpediaMEL, and WikiDiverse, with Hits@1 scores of 89.03% on WikiMEL and 83.02% on RichpediaMEL; LLM/VLM systems remain stronger on WikiDiverse, positioning CELMI as a lightweight complement. Progressive stress ablation shows a 38.50 percentage-point absolute Hits@1 drop when the architecture is reduced to a single unimodal endpoint, confirming that the gains arise from synergistic channel interaction rather than isolated module effects.



Academic Editor: Alexander Gegov

Received: 31 May 2026

Revised: 7 July 2026

Accepted: 30 July 2026

Published: 2 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: multimodal entity linking; cross-modal interaction; mention collaboration; feature fusion; consistency loss

1. Introduction

Entity linking is a core operation for transforming unstructured content into structured knowledge, supporting downstream tasks such as question answering, recommendation, and knowledge graph population [1,2]. As web documents increasingly combine natural

language with images, tables, and social media screenshots, this operation has moved from text-only entity linking to multimodal entity linking (MEL), where each mention must be grounded to a knowledge base entity by jointly interpreting textual and visual evidence [3,4]. The standard formulation, however, hides a deeper contradiction. MEL is usually optimized as a set of independent mention-to-entity retrieval decisions, while the evidence that resolves a difficult mention is often not local to that mention. It may be distributed across nearby mentions, visual co-occurrence patterns, and cross-modal constraints that only become reliable when considered together.

Figure 1 illustrates this contradiction. The text states that “Microsoft CEO Satya” and “Microsoft founder Bill” are interviewed together, while the image shows two people in the same event context. If “Satya” is linked independently, the model must choose between entities such as “Satya Nadella” and the surface-form confounder “Satya Jewelry” using only local textual and visual cues. The decisive evidence is instead collaborative: “Bill” anchors the Microsoft context, the CEO/founder relation constrains the entity pair, and the image supports a joint interpretation of the two people. The target entity is therefore not merely the best match for an isolated mention; it is the entity that remains consistent with the other mentions and modalities in the document. This observation motivates our central view: MEL should be treated as a collaborative multimodal reasoning problem rather than a collection of independent matching problems [5–8].

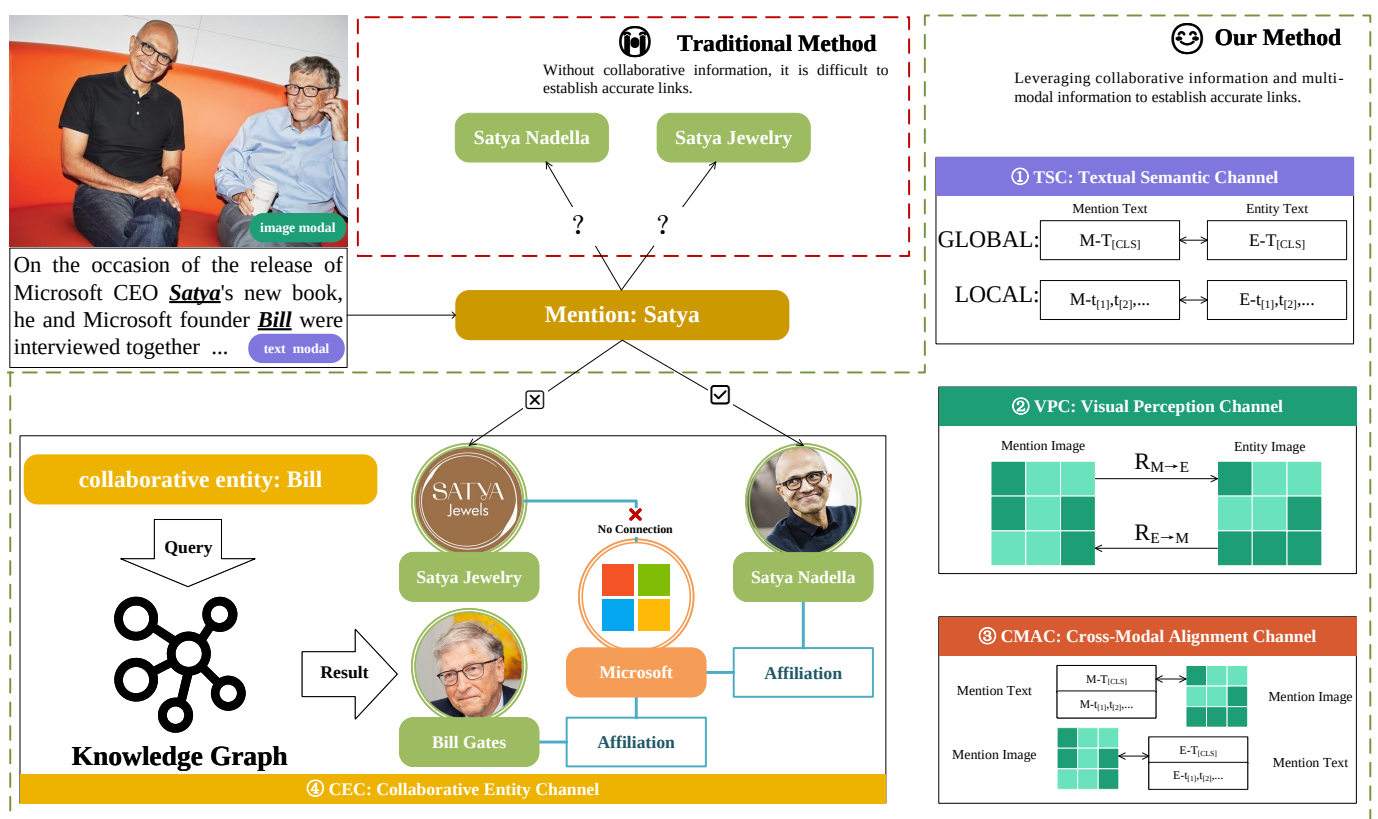


Figure 1. Multimodal entity linking example demonstrating the necessity of collaborative modeling. Different colors indicate different modalities, arrows denote information flow, and the symbols “?”, “✓”, and “×” represent ambiguity, correct linking, and incorrect linking, respectively. The ellipses indicate additional candidate entities omitted for visual simplicity. Independent mention processing leaves “Satya” ambiguous between “Satya Nadella” and “Satya Jewelry.” Collaborative modeling exploits the co-occurring mention “Bill,” the shared Microsoft context, and visual co-occurrence to establish a consistent entity assignment.

The development of MEL has largely followed this independent-matching paradigm. Early systems adopted candidate retrieval followed by reranking [9], which performed well when the correct entity was retrieved but suffered from cascading errors when recall failed. End-to-end methods later reduced this dependency by directly aligning multimodal mention and entity representations through contrastive learning [10–12]. More recently, generative frameworks such as GEMEL [13] and KGEMEL [14] have used foundation models to improve entity name generation and knowledge-guided reasoning. These advances strengthen the local mention–entity matcher, yet they do not resolve the structural mismatch above: a stronger isolated predictor still cannot use the document-level evidence that emerges only through interactions among mentions and modalities.

Addressing this limitation requires an architecture that simultaneously handles fine-grained textual semantics, noisy visual inputs, cross-modal alignment, and inter-mention structural dependencies. We identify three fundamental challenges that continue to impede accurate MEL in complex real-world scenarios:

Challenge 1: Independent mention processing forfeits collaborative disambiguation signals. Current approaches treat each mention as a separate prediction instance, preventing the model from using semantic dependencies among co-occurring mentions. Even when the context implies entity relations—such as shared affiliations, event co-participation, or temporal roles—independent processing cannot perform cross-mention verification, leading to failures in cases where entities are jointly identifiable but individually ambiguous.

Challenge 2: Heterogeneous modality representations suffer from intra-modal noise and inter-modal semantic gaps. Real-world images often contain background clutter, decorative objects, or low-quality visual regions that dilute entity-relevant cues under uniform pooling. At the same time, textual and visual features occupy different representation spaces; without explicit cross-modal projection, their complementary information cannot be integrated into a coherent disambiguation signal.

Challenge 3: Joint multi-channel optimization is susceptible to gradient imbalance and representation collapse. A collaborative MEL model must combine several interaction mechanisms but optimizing them with a single objective can allow an easy or dominant channel to absorb most of the gradient signal. The remaining channels may then underfit, reducing a nominally multi-channel architecture to a near-unimodal predictor.

To address these challenges, we propose CELMI (Collaborative Entity Linking through Multichannel Interaction), a framework that converts the above motivation into four complementary interaction channels. The Textual Semantics Channel captures both lexical-level and contextual-level correspondences. The Visual Perception Channel uses dynamic gating to suppress irrelevant regions and emphasize discriminative visual cues. The Cross-Modal Alignment Channel projects textual and visual features into a shared semantic space so that cross-modal evidence can be compared directly. The Collaborative Entity Channel builds an interaction graph over co-occurring mentions and candidate entities, allowing structural consistency signals to propagate through attention-based message passing. Finally, a multi-channel consistency loss assigns supervision to each channel while preserving an overall ranking objective, preventing a single channel from dominating training.

The main contributions of this work are as follows.

- Targeting Challenge 1, we formulate MEL as collaborative multimodal reasoning and propose CELMI, a four-channel framework that jointly models textual semantics, visual perception, cross-modal alignment, and inter-mention entity collaboration.
- Targeting Challenge 2, we design specialized textual, visual, and cross-modal channels that separate fine-grained semantic matching, noise-aware visual perception, and modality-space alignment instead of relying on a single fused representation.

- Targeting Challenge 3, we introduce a multi-channel consistency loss composed of per-channel contrastive objectives and an overall ranking loss. Ablation experiments show that removing $\mathcal{L}_O$ causes a 6.59 percentage-point absolute Hits@1 drop (9.67% relative degradation) on WikiDiverse, confirming its role in stable multi-channel optimization.
- Comprehensive empirical validation across WikiMEL, RichpediaMEL, and WikiDiverse shows that CELMI achieves the strongest MRR and Hits@1 performance among conventional non-LLM/VLM MEL models. Progressive stress ablation further reveals a 38.50 percentage-point absolute Hits@1 drop when the architecture is reduced to a single unimodal endpoint, demonstrating that the four channels contribute through synergy rather than isolated additive gains.

2. Related Work

2.1. Multimodal Entity Linking

Unlike text-only entity linking, MEL must reconcile textual semantics with visual evidence. Existing work can be broadly organized into retrieval-augmented and end-to-end paradigms, both of which have improved mention–entity matching but still largely preserve the independent-decision assumption.

Retrieval-augmented MEL. Retrieval-augmented methods decompose MEL into candidate generation and entity disambiguation. BLINK [9] established a strong retrieval reranking pipeline by combining sparse and dense retrievers with cross-encoder scoring, but its final prediction remained constrained by candidate recall. UniMEL [15] strengthened this paradigm by combining surface-form statistics with dense retrieval, while KGME [14] incorporated vision language models and structured triples to enrich multimodal entity representations before retrieval. These methods benefit from explicit candidate control, yet their performance is bounded by retrieval quality and their cost increases with knowledge base scale. More importantly for this work, retrieval is usually performed for each mention separately, so document-level consistency signals are not directly involved in representation learning.

End-to-end MEL. End-to-end methods avoid explicit retrieval by directly aligning multimodal mentions with knowledge base entities [16,17]. Wang et al. [10] proposed GHMFC, which uses hierarchical multimodal co-attention and gated fusion to extract fine-grained cross-modal features under a contrastive objective. Xing et al. [11] proposed DRIN to model multiple mention–entity alignment relations and select among them with dynamic graph convolutional networks. Sui et al. [18] developed MELOV to address noisy social media images by mining text-guided visual features and aggregating neighbourhood visual information. Generative MEL has also been explored: Shi et al. [13] showed that large language models can generate target entity names using a frozen visual encoder and a lightweight feature mapper. These methods substantially improve local multimodal matching, but the prediction unit remains the individual mention. As a result, they do not explicitly exploit co-occurring mentions as collaborative evidence during training.

A comparison between CELMI and representative entity linking methods is provided in Table 1.

The present work is most directly related to MIMIC [12], which demonstrates the effectiveness of CLIP-based text vision encoding under cross-modal contrastive objectives. CELMI adopts a similar encoder foundation but changes the interaction structure: rather than learning one global fused representation for each mention, it decomposes matching into four specialized channels and introduces inter-mention graph propagation as part of the jointly optimized representation.

2.2. Collaborative Linking Methods

Recognizing that independent mention processing discards valuable document-level consistency signals, a growing body of research has investigated collaborative entity linking strategies that jointly disambiguate all mentions within a document.

Graph-based collective linking. Collective entity linking has long recognized that mentions within a document constrain one another. NCEL [19] constructs a graph over mentions and candidate entities, applying graph convolutional networks to combine local contextual features with global coherence signals. Deng et al. [20] further introduced multi-hop attention and graph search-based sequential reasoning to fuse contextual and knowledge graph information more efficiently. These graph-based methods demonstrate the value of inter-mention influence, but they are usually designed for text-only settings, incur non-trivial graph construction cost, and are sensitive to graph quality.

Reinforcement learning-based linking. A second line treats collective linking as a sequential decision process. Fang et al. [21] trained a reinforcement learning agent to assign entities from a global perspective, allowing later decisions to condition on previously resolved mentions. DyMen [22] further learned the order in which mentions should be disambiguated, prioritizing high-confidence mentions that can constrain subsequent predictions. These approaches provide flexible linking strategies, but their optimization is complex and their large action spaces make them difficult to integrate into lightweight multimodal retrieval systems.

Multimodal collective linking. MMEL [23] extends collective reasoning to multimodal settings by extracting textual and visual context representations and applying multi-mention collaborative ranking. Its results indicate that collaborative signals can improve MEL when multiple entities appear in the same context. Because publicly reported MMEL results do not cover all benchmarks used in this study, we report available numbers and leave unavailable entries blank rather than extrapolating performance.

Despite these advances, existing collaborative methods leave an important gap for MEL. Graph-based and reinforcement learning methods expose inter-mention dependencies but are costly or unstable; multimodal collective methods validate collaboration but do not integrate it with fine-grained textual matching, noise-aware visual perception, and explicit cross-modal alignment in a unified objective. CELMI addresses this gap by embedding collaborative entity propagation as one of four jointly trained channels, allowing document-level consistency to influence representation learning rather than appearing only as a post-hoc ranking signal.

2.3. Lightweight and Robust Multimodal Deployment

Recent multimodal systems increasingly rely on large foundation models, knowledge-enhanced generators, or compact models supervised by larger teachers. This trend improves representational capacity, but it also introduces deployment constraints. Model compression and distillation can reduce inference cost, yet compact students may suffer from training-inference distribution mismatch when they are trained only on teacher-forced or static teacher outputs [24,25]. On-policy and imitation-learning distillation methods address this issue by exposing the student to its own generated states during training, thereby reducing exposure bias and improving robustness under realistic inference conditions [25]. Related robustness concerns also appear in vision-driven physics-informed control, where adaptive smooth control mechanisms are used to maintain reliable behavior under aerodynamic interference and actuator faults [26,27]. These studies motivate lightweight MEL models that avoid unnecessary full generative decoding while retaining robustness to noisy modalities, missing visual evidence, and ambiguous co-occurring mentions.

Table 1. Comparison between CELMI and closely related entity linking methods.

Method	Input Assumption	Interaction Granularity	Graph or Structure	Training Objective	Relation to CELMI
MIMIC [12]	Multimodal mention and entity representations	Mention–entity pair	No explicit inter-mention graph	Cross-modal contrastive learning	CELMI adopts the CLIP-based encoder foundation but replaces single fused matching with four specialized channels.
MMEL [23]	Multiple mentions with textual and visual context	Multi-mention ranking	Collaborative ranking among mentions	Ranking-oriented MEL objective	CELMI incorporates collaboration into representation learning through CEC rather than using collaboration only as a ranking signal.
NCEL [19]	Text-only mentions and candidate entities	Document-level candidates	Candidate graph for collective linking	Local and global coherence learning	CELMI adapts the collective-linking idea to multimodal MEL with shared visual-textual encoders and channel-specific supervision.
KGMEI [14]	Multimodal inputs plus knowledge graph triples	Retrieval and generation aided by structured knowledge	External KG relations are used	Knowledge-guided MEL objective	CELMI does not use external KB relation edges in CEC; it focuses on lightweight collaboration among co-occurring mentions and their candidates.
CELMI	Annotated MEL mentions, images, and KB entities	Textual, visual, cross-modal, and inter-mention channels	Similarity threshold mention-candidate graph without external KB relation edges	Per-channel contrastive losses plus an overall ranking loss	TSC, VPC, and CMAC adapt established matching ideas; CEC is introduced as an original collaborative channel and optimized jointly with the other channels.

3. Methodology

The overall architecture of CELMI is illustrated in Figure 2.

3.1. Problem Formulation

Let $D = (S, I)$ denote a multimodal document containing textual context S and image context I . The MEL benchmark provides a gold mention set $\mathcal{M}_D = \{m_1, \dots, m_n\}$ for each document; therefore, this work evaluates entity linking under gold mention boundaries rather than joint mention detection and linking. For each mention m_i , a candidate set $\mathcal{C}_i = \{e_{i,1}, \dots, e_{i,k_i}\}$ is drawn from the knowledge base $\mathcal{E}$, where each entity e is represented by its name, attributes, description, and available image evidence. The document-level assignment space is

$$\mathcal{Y}_D = \mathcal{C}_1 \times \mathcal{C}_2 \times \dots \times \mathcal{C}_n, \quad (1)$$

where an assignment $\mathbf{y} = (e_{1,a_1}, \dots, e_{n,a_n}) \in \mathcal{Y}_D$ links every annotated mention in D to one candidate entity.

Independent MEL models estimate a local score $s_\theta(m_i, e_{i,j})$ and predict each entity separately, as follows.

$$\hat{e}_i^{\text{ind}} = \arg \max_{e \in \mathcal{C}_i} s_\theta(m_i, e). \quad (2)$$

This factorized inference ignores the fact that the best assignment for one mention may depend on entities assigned to other mentions in the same document. Collaborative

MEL instead scores a document-level assignment by combining local multimodal evidence with cross-mention consistency:

$$S_{\theta}(D, \mathbf{y}) = \sum_{i=1}^n U_{\theta}(m_i, e_i) + \sum_{i \neq j} \phi_{\theta}(m_i, e_i, m_j, e_j), \quad (3)$$

where U_{θ} is the multi-channel mention–entity score and ϕ_{θ} measures compatibility induced by co-occurring mentions and their candidates. Exact maximization over $\mathcal{Y}_D$ can be expensive when many mentions and candidates are present. CELMI therefore uses a tractable graph-enhanced approximation: it constructs a document graph $G_D = (V_D, E_D)$ containing mention nodes and candidate entity nodes, propagates evidence over this graph, and then ranks each candidate with a collaborative score that has already absorbed document-level signals.

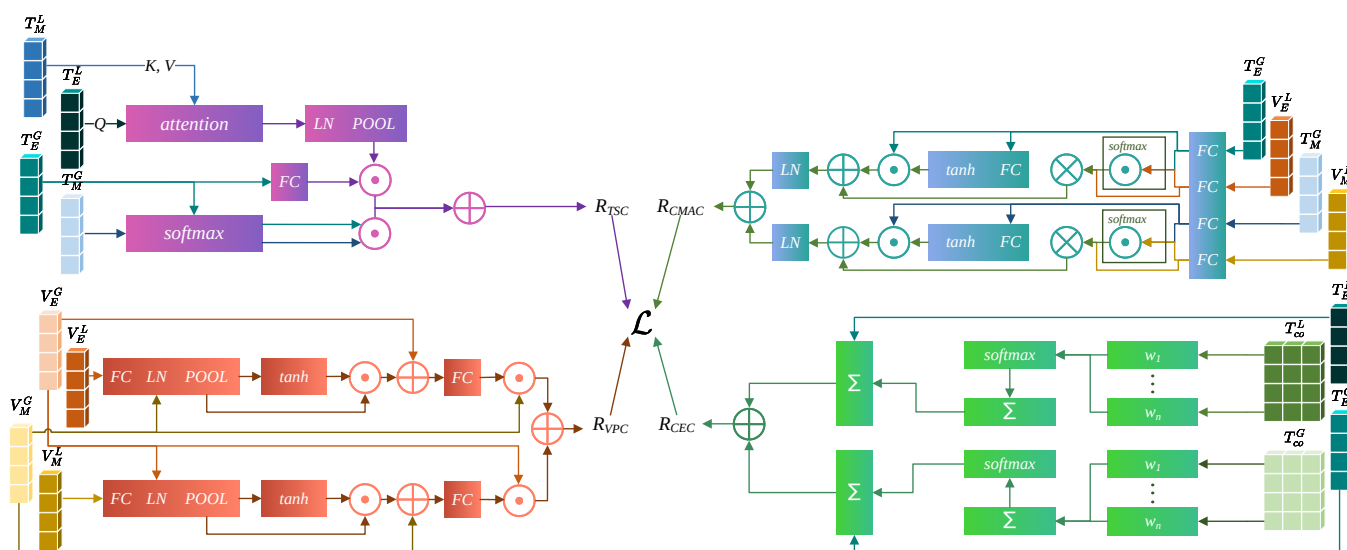


Figure 2. The overall architecture of the Collaborative Entity Linking through Multichannel Interactions (CELMI) framework.

Under this formulation, the model is trained to increase the probability of the gold entity for each mention while conditioning the score on the multimodal document graph:

$$\theta^* = \arg \max_{\theta} \sum_D \sum_{m_i \in \mathcal{M}_D} \log p_{\theta}(e_i^* | m_i, D, G_D, \mathcal{E}), \quad (4)$$

where e_i^* is the gold entity. This objective differs from independent ranking because p_{θ} is computed from representations updated by co-occurring mentions and candidate entities. The final prediction for mention m_i is

$$\hat{e}_i = \arg \max_{e \in \mathcal{C}_i} U_{\theta}^{\text{collab}}(m_i, e; D, G_D), \quad (5)$$

where $U_{\theta}^{\text{collab}}$ is instantiated by the four-channel CELMI score described below.

3.2. Encoder

The encoder provides the shared textual and visual representations consumed by all interaction channels. This design ensures that the proposed channels differ in how they model evidence rather than in the backbone features they receive.

3.2.1. Text Encoding

For each sample, we employ a pre-trained CLIP [28] text encoder to obtain semantic representations. The entity name and entity attributes extracted from the knowledge base are first concatenated using the special token [SEP] to construct the input sequence, which is then encoded as follows.

$$Input_E = e^{name} \oplus [SEP] \oplus e^{attr}, \quad (6)$$

$$T_E = \text{Encoder}^T(Input_E), \quad (7)$$

where $T_E \in \mathbb{R}^{d_T}$ represents the text feature vector of the entity, and d_T is the text feature dimension. Following CLIP's text encoder convention, the hidden state of the end-of-text token is treated as the global text feature T_E^G , and all hidden states collectively constitute the local text feature T_E^L .

Similarly, mentions and co-occurring mentions are encoded as follows.

$$T_{M_i} = \text{Encoder}^T(m^{name} \oplus [SEP] \oplus m^{sentence}), \quad (8)$$

where T_M^G and T_M^L denote the global and local text features of the mention sample, respectively, and T_{co}^G and T_{co}^L denote the corresponding features of co-occurring mentions.

3.2.2. Visual Encoding

The CLIP visual encoder is employed to extract visual semantic information. For the image set $\{img_1^E, img_2^E, \dots, img_k^E\}$ associated with entity E , one image is sampled at random during training, while the first image is selected during evaluation.

Image preprocessing rescales the input image to $W \times H$ pixels using the Lanczos interpolation algorithm to preserve image quality, as follows.

$$img_E = \text{Resize}(img_{\text{original}}, (W, H), \text{Lanczos}). \quad (9)$$

The preprocessed image is then passed through the CLIP visual encoder to obtain visual features, as follows.

$$V_E = \text{Encoder}^V(img_E), \quad (10)$$

where $V_E \in \mathbb{R}^{C \times W \times H}$ represents the image tensor and C is the number of image channels.

The visual encoder partitions the image into $[(W \times H) / P^2 + 1]$ patches, denoted as $(n + 1)$ chunks including a [CLS] token for global feature representation. Each chunk is projected to d_v dimensions via a fully connected layer, yielding the entity image feature $V_E \in \mathbb{R}^{(n+1) \times d_v}$. All patch features are treated as local features V_E^L , and the [CLS] token feature serves as the global feature V_E^G [12]. The identical processing pipeline is applied to mention images to obtain V_M .

3.3. Multi-Channel Fusion

The multi-channel fusion module decomposes MEL evidence into four complementary channels. The Textual Semantics Channel (TSC) captures surface and contextual semantic associations between mentions and entities; the Visual Perception Channel (VPC) extracts task-relevant visual correspondences; the Cross-Modal Alignment Channel (CMAC) establishes bidirectional alignment between text and image representations; and the Collaborative Entity Channel (CEC) models inter-entity dependencies so that co-occurring mentions can provide mutual disambiguation signals.

3.3.1. Textual Semantics Channel

Textual entity representations often rely on a global embedding such as the [CLS] token, which compresses useful token-level evidence. Since ambiguous mentions may differ only in fine-grained lexical context, TSC employs a dual-level alignment mechanism that captures both coarse semantic similarity and local token-level correspondence.

Shallow semantics, encoded in the global representation T^G , captures overall sentence-level similarity suitable for rapid coarse-grained matching. Deep semantics, extracted from local token sequences T^L through cross-attention, captures fine-grained lexical alignments and contextual dependencies that are crucial for disambiguating semantically similar entities. The dynamic fusion of these two levels enables the model to adaptively balance coarse-grained and fine-grained matching strategies according to the degree of entity ambiguity. The hierarchical attention mechanism establishes correspondences between surface and deep text representations through learnable query, key, and value transformations, as follows.

$$Q, K, V = T_E^L W_Q^{TSC}, T_M^L W_K^{TSC}, T_M^L W_V^{TSC}, \quad (11)$$

$$H^{TSC} = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_{TSC}}}\right)V, \quad (12)$$

where $W_Q^{TSC}, W_K^{TSC}, W_V^{TSC} \in \mathbb{R}^{d_T \times d_{TSC}}$ are learnable projection matrices and d_{TSC} denotes the TSC channel dimension.

Mean pooling and layer normalization are then applied to H^{TSC} to obtain the context vector, which is combined with the projected global entity representation to yield the deep alignment score, as follows.

$$h^{TSC} = \text{LN}(\text{MeanPool}(H^{TSC})), \quad (13)$$

$$R_{Text}^{deep} = \text{FC}(T_E^G) \cdot h^{TSC}, \quad (14)$$

where the fully connected layer is parameterized by $W_T^{TSC} \in \mathbb{R}^{d_T \times d_{TSC}}$ and bias $b_T^{TSC} \in \mathbb{R}^{d_{TSC}}$.

The surface-level alignment score is obtained by computing the dot product between the softmax-normalized global features of the mention and the entity, as follows.

$$R_{Text}^{sur} = \text{softmax}(T_M^G) \cdot \text{softmax}(T_E^G). \quad (15)$$

The final TSC matching score is obtained through dynamically learned weight fusion, as follows.

$$R_{TSC} = \alpha_{TSC} \cdot R_{Text}^{sur} + (1 - \alpha_{TSC}) \cdot R_{Text}^{deep}, \quad (16)$$

where $\alpha_{TSC} \in [0, 1]$ is an adaptively learned parameter. This dynamic weighting mechanism enables the model to adjust the relative contribution of surface and deep alignment, yielding more discriminative matching across diverse entity linking scenarios.

3.3.2. Visual Perception Channel

Real-world images, particularly those from social media contexts, often contain background clutter, irrelevant objects, or decorative elements that dominate the visual field. Uniform pooling treats all visual patches as equally informative and can dilute the cues that are actually useful for linking. VPC therefore introduces a dynamic gating mechanism that amplifies task-relevant visual interactions while suppressing noise-dominated regions. This design is also useful because different entity types rely on visual information to different

degrees: landmarks may be visually distinctive, whereas abstract or weakly visual entities can be harmed by noisy images.

The dynamic gate parameter λ^{VPC} adaptively modulates the contribution of visual features on a per-sample basis, amplifying informative visual signals and attenuating noise-dominated ones. The channel begins by fusing the mean-pooled local image features V_E^L of the entity with the corresponding global image representation V_M^G of the mention, followed by layer normalization and linear projection to obtain a deep visual entity representation, as follows.

$$h_{chunk} = \text{Proj}^{VPC} \left(\text{LN} \left(\text{MP} \left(V_E^L + V_M^G \right) \right) \right), \quad (17)$$

where the projection layer Proj^{VPC} contains learnable parameters $W_V^{VPC} \in \mathbb{R}^{d_V \times d_{VPC}}$ and bias $b_V^{VPC} \in \mathbb{R}^{d_{VPC}}$, and MP denotes mean pooling.

Gate weights are generated through a fully connected layer and a Tanh activation to control the degree of feature fusion, as follows.

$$\lambda^{VPC} = \tanh(\text{FC}(h_{chunk})), \quad (18)$$

$$h_{fused} = \text{LN}(\lambda^{VPC} \odot h_{chunk} + v_E^G), \quad (19)$$

where the gating layer parameters are $W_{gate}^{VPC} \in \mathbb{R}^{d_{VPC} \times 1}$ and $b_{gate}^{VPC} \in \mathbb{R}$, and $\odot$ denotes element-wise multiplication. The gating mechanism enables the model to selectively retain informative feature interactions while suppressing the influence of visual noise.

The entity-to-mention alignment score is then computed via the inner product between the fused entity features and the global mention representation, as follows.

$$R_{E \rightarrow M}^V = h_{fused} \cdot v_M^G. \quad (20)$$

The mention-to-entity direction follows an identical computational process with the roles of the inputs exchanged.

The final visual perception score is obtained by adaptively fusing both directional scores, as follows.

$$R_{VPC} = \alpha_{VPC} \cdot R_{E \rightarrow M}^V + (1 - \alpha_{VPC}) \cdot R_{M \rightarrow E}^V, \quad (21)$$

where $\alpha_{VPC} \in [0, 1]$ is a learnable parameter.

3.3.3. Cross-Modal Alignment Channel

The Cross-Modal Alignment Channel addresses the representation gap between textual semantics and visual content. Rather than fusing raw text and image features directly, CMAC first maps them into a common semantic subspace, allowing visual regions to be weighted according to their textual relevance and enabling mention–entity matching through cross-modal mutual guidance.

$$h_E^T, h_M^T = \text{Proj}_T^{\text{CMAC}}(T_E^G), \text{Proj}_T^{\text{CMAC}}(T_M^G), \quad (22)$$

$$h_E^V, h_M^V = \text{Proj}_V^{\text{CMAC}}(V_E^L), \text{Proj}_V^{\text{CMAC}}(V_M^L), \quad (23)$$

where $\text{Proj}^{\text{CMAC}}$ is parameterized by weight matrix $W^{\text{CMAC}} \in \mathbb{R}^{d \times d_{\text{CMAC}}}$ and bias $b^{\text{CMAC}} \in \mathbb{R}^{d_{\text{CMAC}}}$, with $\cdot \in \{T, V\}$.

The importance weight of each visual region block with respect to the entity text representation is computed as follows.

$$\omega_i^{CMAC} = \frac{\exp(h_E^T \odot h_{E_i}^V)}{\sum_{j=1}^{n+1} \exp(h_E^T \odot h_{E_j}^V)}, \quad (24)$$

where $h_{E_i}^V$ denotes the i -th projected visual patch feature.

Based on the above weight distribution, the text-guided visual context vector is constructed as follows.

$$c_E^V = \sum_{i=1}^{n+1} \omega_i^{CMAC} \times h_{E_i}^V. \quad (25)$$

A gating mechanism is introduced to regulate the activation intensity of the textual features, as follows.

$$\lambda_E^{CMAC} = \tanh(\text{Proj}_\lambda^{CMAC}(h_E^T)), \quad (26)$$

where the projection layer is parameterized by $W_\lambda \in \mathbb{R}^{d_T \times d_{CMAC}}$ and bias $b_\lambda \in \mathbb{R}^{d_{CMAC}}$.

The entity alignment representation is obtained under the modulation of the gate factor, as follows.

$$z_E = \text{LN}(\lambda_E^{CMAC} \odot h_E^T + c_E^V). \quad (27)$$

The mention-side alignment representation z_M is obtained through the identical computational pipeline, and the final cross-modal matching score is measured by their vector inner product, as follows.

$$R_{CMAC} = z_E \cdot z_M. \quad (28)$$

3.3.4. Collaborative Entity Channel

Traditional entity linking approaches process each mention independently, neglecting the semantic relationships among co-occurring entities. When an image and sentence jointly mention “Satya Nadella” and “Bill Gates,” for instance, their shared association with “Microsoft” provides mutual verification evidence. CEC captures such collaborative signals by modelling mention–entity dependencies through a graph-based interaction mechanism.

In the benchmark setting, co-occurring mentions are obtained from the mention annotations provided by the MEL datasets, so the evaluation isolates entity linking from mention detection. In a raw document deployment setting, these mentions can be supplied by an external named entity recognizer; errors from that upstream detector would affect CEC graph quality and are therefore treated as a deployment limitation rather than a benchmark variable. For a document D , CEC selects all annotated mentions that occur in the same multimodal context as m_i . Each selected mention contributes one mention node and its associated candidate entity nodes. No external knowledge base relation edges are used, which keeps the channel lightweight and prevents the comparison from depending on additional relational resources unavailable to all baselines.

The pseudocode of CEC graph construction and message passing is presented in Table 2.

The node set is defined as

$$V_D = \mathcal{M}_D \cup \left(\bigcup_{m_i \in \mathcal{M}_D} \mathcal{C}_i \right), \quad (29)$$

where mention nodes are initialized from mention text–image representations and candidate nodes are initialized from entity text–image representations. Edges connect mention–

candidate pairs and semantically compatible nodes in the same document context. For any two nodes u and v , the edge weight is defined as follows:

$$w_{uv} = \begin{cases} \text{sim}(h_u, h_v) & \text{if } \text{sim}(h_u, h_v) \geq \tau, \\ 0 & \text{otherwise,} \end{cases} \quad (30)$$

where $\text{sim}(\cdot)$ denotes cosine similarity and τ is a similarity threshold set to 0.5 based on validation performance. Noisy or incorrectly retrieved candidates can enter the graph, but threshold filtering and attention normalization reduce their influence when they are weakly compatible with the document context. If noise remains highly similar to the context, CEC may propagate misleading evidence; this limitation motivates the error analysis in Section 4.3.

Table 2. Pseudocode for CEC graph construction and message passing.

Input: document mentions $\mathcal{M}_D$, candidate sets $\{\mathcal{C}_i\}$, node features $h_v^{(0)}$, threshold τ , propagation layers L .
1. Create one mention node for each $m_i \in \mathcal{M}_D$ and one candidate node for each $e \in \mathcal{C}_i$.
2. Initialize mention nodes with encoded mention features and candidate nodes with encoded entity features.
3. For every node pair (u, v) in the same document graph, compute $\text{sim}(h_u^{(0)}, h_v^{(0)})$.
4. Retain an edge if the similarity is at least τ ; otherwise suppress the edge by assigning weight zero.
5. For each layer $l = 0, \dots, L - 1$, compute attention weights over the retained neighbors and update every node by weighted aggregation.
6. Score each mention-candidate pair using the propagated mention and candidate representations.

Information propagates through the graph via L layers of attention-based message passing. At each layer, node representations are updated by aggregating information from neighboring nodes as follows.

$$h_v^{l+1} = \sigma \left(\sum_{u \in \mathcal{N}(v)} \alpha_{uv}^{(l)} W^{(l)} h_u^{(l)} \right), \quad (31)$$

where $\mathcal{N}(v)$ denotes the neighborhood of node v , $W^{(l)}$ is a learnable transformation matrix, σ is a non-linear activation function, and $\alpha_{uv}^{(l)}$ represents the attention weight computed via scaled dot-product attention as follows.

$$\alpha_{uv}^{(l)} = \frac{\exp(S_{uv}^{(l)})}{\sum_{k \in \mathcal{N}(v)} \exp(S_{kv}^{(l)})}, \quad S_{uv}^{(l)} = \frac{(z_u^{(l)})^\top z_v^{(l)}}{\sqrt{d_{\text{CEC}}}}, \quad (32)$$

where $z_u^{(l)} = W^{(l)} h_u^{(l)}$ and d_{CEC} is the channel dimension.

To maintain computational efficiency while preserving expressive power, threshold-based control dynamically regulates the neighborhood scope during propagation. At layer l , the neighborhood of node v is restricted as follows.

$$\mathcal{N}_v^{(l)} = \{u \mid \text{sim}(h_v^{(l)}, h_u^{(l)}) \geq \tau\}, \quad (33)$$

where nodes failing to satisfy this criterion are excluded, and propagation for isolated nodes terminates automatically. This filtering reduces unnecessary message passing and limits the influence of candidate noise.

The graph construction requires $O(N^2d)$ similarity operations for N graph nodes and each propagation layer incurs $O(L \cdot N \cdot |\mathcal{N}| \cdot d^2)$ operations. After L layers ($L = 2$ in all experiments), the final CEC score is computed as the inner product between the updated mention and candidate representations, as follows.

$$R_{CEC}(m_i, e_{i,j}) = h_{m_i}^{(L)} \cdot h_{e_{i,j}}^{(L)}, \quad (34)$$

where $h_{m_i}^{(L)}$ and $h_{e_{i,j}}^{(L)}$ denote the L -layer propagated representations of mention m_i and candidate entity $e_{i,j}$, respectively. This score is incorporated directly into the joint optimization objective, enabling collaborative reasoning signals to contribute to gradient backpropagation in a manner consistent with all other channels.

3.4. Multi-Channel Consistency Objective Function

Based on the matching scores computed by the four channels above, CELMI first forms a composite mention–entity score. Let $X \in \{\text{TSC}, \text{VPC}, \text{CMAC}, \text{CEC}\}$ denote a channel and $R_X(M, E)$ denote its matching score. The overall score is defined as follows.

$$U(M, E) = \sum_X \beta_X R_X(M, E), \quad (35)$$

where β_X is the channel aggregation coefficient. In all experiments, we set $\beta_X = 1/4$ for the four channels, using an equal-weight composite to avoid introducing dataset-specific tuning beyond the channel-internal adaptive gates. The coefficient β_X controls inference-time score aggregation, whereas λ_X in the training objective below controls the strength of auxiliary channel supervision; the two sets of coefficients therefore affect different parts of the optimization and inference pipeline.

The encoding module and multi-channel fusion module are then jointly trained by minimizing a contrastive learning objective. This loss encourages the model to assign higher scores to correct mention–entity pairs while reducing the scores of incorrect pairings, as follows.

$$\mathcal{L}_O = -\log \frac{\exp(U(M, E))}{\sum_i \exp(U(M, E'_i))}, \quad (36)$$

where U denotes the composite channel score, and E'_i denotes negative sample entities in the knowledge base. In practice, an intra-batch negative sampling strategy is adopted: given a positive pair (M, E) , all other entities within the same batch serve as negatives, maintaining training efficiency without exhaustive comparison against the full knowledge base.

However, when $U(M, E)$ averages scores across four channels, a dominant channel may monopolize gradient flow and cause other channels to underfit. To resolve this, an independent contrastive loss is assigned to each channel by applying the same contrastive form to the channel-specific score $R_X(M, E)$, as follows.

$$\mathcal{L}_X = -\log \frac{\exp(R_X(M, E))}{\sum_i \exp(R_X(M, E'_i))}, \quad (37)$$

where $X \in \{\text{TSC}, \text{VPC}, \text{CMAC}, \text{CEC}\}$ denotes the channel index.

The final multi-channel consistency objective is defined as follows.

$$\mathcal{L} = \lambda_O \mathcal{L}_O + \sum_X \lambda_X \mathcal{L}_X. \quad (38)$$

In the reported experiments, all loss weights are set to 1. This equal-weight setting preserves the intended auxiliary supervision role of each channel loss without introducing additional validation-tuned hyperparameters. This design ensures that each channel main-

tains discriminative feature learning capability while contributing coherently to the overall framework, mitigating both single-channel dominance and representation collapse.

Evaluation phase. During inference, the model computes similarity scores between each mention and all entities in the knowledge base. Entity representations are pre-encoded offline by the text and vision encoders. At test time, only the mention representation needs to be computed, and its similarity with all entities is obtained by a single matrix multiplication. Given that the entity set size is on the order of 10^5 , this operation is computationally tractable on modern GPUs. The final entity ranking is determined by

$$R = U(M, E), \quad (39)$$

which is the equal-weight composite score defined above.

The introduction of per-channel losses alongside $\mathcal{L}_O$ is theoretically motivated by gradient allocation in multi-branch models. If training uses only the composite score $U(M, E)$, the derivative of $\mathcal{L}_O$ with respect to the parameters of channel X is scaled by that channel's contribution to the fused score. A channel that learns a high-confidence shortcut early in training can therefore dominate the fused probability, while weaker but complementary channels receive smaller effective gradients. This imbalance can produce channel dominance: the fused model improves locally, but visual, cross-modal, or collaborative branches remain under-optimized.

The auxiliary term $\mathcal{L}_X$ changes this optimization geometry. Because every channel must independently distinguish the gold entity from negatives, each branch receives a direct discriminative signal even when its current fused contribution is small. At the same time, $\mathcal{L}_O$ prevents the branches from drifting into unrelated objectives by requiring their aggregated score to support the final ranking decision. The resulting objective can be interpreted as a constrained form of multi-objective learning: channel-specific losses preserve branch-level separability, whereas the overall loss aligns those branches into a coherent mention–entity ranking function. This does not mathematically eliminate all collapse modes, but it reduces the incentive for a single easy channel to absorb the full supervision signal and leave the remaining channels inactive.

4. Experiments

4.1. Experimental Settings

Datasets. Experiments are conducted on three publicly available MEL benchmarks: WikiMEL [10], RichpediaMEL [10], and WikiDiverse [29]. WikiMEL is collected from Wikipedia entity pages and contains over 22k multimodal sentences, with person entities as the predominant type. RichpediaMEL is derived from the multimodal knowledge graph Richpedia [30], where entities are sourced from Richpedia and multimodal information is supplemented from Wikidata [31]; it also primarily covers person entities. WikiDiverse is constructed from Wikinews and encompasses diverse topical domains including sports, technology, and economics, with Wikidata serving as the knowledge base after filtering mentions absent from Wikidata. Statistical details of the three datasets are presented in Table 3.

Metrics. Model performance is assessed using three standard metrics.

- Hits@ n measures the proportion of test samples for which the correct entity appears within the top- n ranked candidates, defined as follows.

$$\text{Hits@}n = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(r_i \leq n), \quad (40)$$

where N denotes the total number of test samples, r_i is the rank of the correct entity for the i -th sample, and $\mathbb{I}(\cdot)$ is the indicator function.

- Mean Rank (MR) is the average rank of correct entities across all test samples, defined as follows.

$$\text{MR} = \frac{1}{N} \sum_{i=1}^N r_i. \quad (41)$$

- Mean Reciprocal Rank (MRR) is the average reciprocal rank across all test samples, defined as follows.

$$\text{MRR} = \frac{1}{N} \sum_{i=1}^N \frac{1}{r_i}. \quad (42)$$

Baselines. We compare CELMI with three categories of methods: (1) off-the-shelf pre-trained models, including BLINK [9], BERT [32], RoBERTa [33], CLIP [28], ViLT [34], ALBEF [35], and METER [36]; (2) representative conventional MEL models, including GHMFC [10], MIMIC [12], OTMEL [37], DIM [38], DWE [39], DWE+ [40], DRIN [11], and MMEL [23]; and (3) LLM/VLM-based MEL models, including GELR [41], GPT-3.5-Turbo [42], GEMEL [13], and KGEMEL [14]. To improve traceability, the results for BLINK, BERT, RoBERTa, CLIP, ViLT, ALBEF, METER, GHMFC, and MIMIC are reused from the MIMIC paper after confirming that the same benchmark datasets and metrics are used. MIMIC also reports earlier traditional baselines such as DZMNED, JMEL, and VELML; these rows are omitted from the main table to keep the comparison focused on methods discussed in the analysis. For later methods, we report public numbers from the corresponding papers only when the dataset and metric match exactly. When public results or reproducible configurations are unavailable for a specific dataset, the entry is left blank rather than inferred.

Table 3. Statistics of the three benchmark datasets. “#” denotes the number of items, “ent.” stands for entity/entities, and “ment.” stands for mention(s).

Statistic	WikiMEL	RichpediaMEL	WikiDiverse
# sentences	22,070	17,724	7405
# mentions	25,846	17,805	15,093
# img. of ment.	22,136	15,853	6697
# ment. in train	18,092	12,463	11,351
# ment. in valid	2585	1780	1664
# ment. in test	5169	3562	2078
# ent. of KB	109,976	160,935	132,460
# ent. with img.	67,195	86,769	67,309

Implementation Details. Model weights are initialized from pre-trained CLIP-ViT-Base-Patch32, which employs the ViT-B/32 Transformer architecture as the image encoder with a patch size of 32. All images are resized to 224×224 pixels, and zero padding is applied to mentions and entities that lack associated images. The maximum text input length is set to 40 tokens, and the text feature dimension d_T is set to 512. For all three datasets, the channel dimensions d_{TSC} , d_{VPC} , d_{CMAC} , and d_{CEC} are uniformly set to 96. The CEC propagation depth is set to $L = 2$, the graph similarity threshold is set to $\tau = 0.5$ based on validation performance, and all channel aggregation and loss weights are set uniformly as described in Section 3.4. During evaluation, entity representations are compared in chunks of 6000 entities to keep GPU memory usage bounded without changing the ranking scores. The method is implemented in PyTorch2.4.1 [43] and trained on an NVIDIA RTX A5000

GPU. The AdamW [44] optimizer is employed with a batch size of 128, $\beta = (0.9, 0.999)$, a learning rate of 1×10^{-5} , and a weight decay of 1×10^{-4} . Training proceeds for a maximum of 20 epochs. For negative sampling, an intra-batch negative strategy is adopted, training each positive pair (M, E) against 127 in-batch negatives. An early stopping strategy with a patience of 5 epochs is applied based on validation MRR, and the best checkpoint on the validation set is retained for testing.

4.2. Main Results

Table 4 reports the performance of CELMI against existing MEL methods on three benchmark datasets. The results should be interpreted along two axes: comparison with conventional MEL models, where CELMI is directly comparable in architecture and deployment cost, and comparison with LLM/VLM-based systems, which may use substantially larger models and external generative capacity.

Table 4. Performance comparison on WikiMEL, RichpediaMEL, and WikiDiverse (% for all metrics). The best score in each column is highlighted in bold with an asterisk; the second-best is underlined.

Methods	WikiMEL				RichpediaMEL				WikiDiverse			
	MRR	Hits@1	Hits@3	Hits@5	MRR	Hits@1	Hits@3	Hits@5	MRR	Hits@1	Hits@3	Hits@5
BLINK	81.72	74.66	86.63	90.57	71.39	58.47	81.51	88.09	69.15	57.14	78.04	85.32
BERT	81.78	74.82	86.79	90.47	71.67	59.55	81.12	87.16	67.38	55.77	75.73	83.11
RoBERTa	80.86	73.75	85.85	89.80	72.80	61.34	81.56	87.15	70.52	59.46	78.54	85.08
CLIP	88.23	83.23	92.10	94.51	77.57	67.78	85.22	90.04	71.69	61.21	79.63	85.18
ViLT	79.46	72.64	84.51	87.86	56.63	45.85	62.96	69.80	45.22	34.39	51.07	57.83
ALBEF	84.56	78.64	88.93	91.75	75.29	65.17	82.84	88.28	69.93	60.59	75.59	81.30
METER	79.49	72.46	84.41	88.17	74.15	63.96	82.24	87.08	63.71	53.14	70.93	77.59
GHMFC	83.36	76.55	88.40	92.01	80.76	72.92	86.85	90.60	70.99	60.27	79.40	84.74
MIMIC	91.82	87.98	<u>95.07</u>	96.37	<u>86.95</u>	<u>81.02</u>	<u>91.77</u>	94.38	73.44	63.51	81.04	86.43
OTMEL	<u>91.93</u>	88.20	94.87	96.52	85.28	78.33	91.16	94.33	73.82	63.38	81.95	97.25 *
DIM	-	64.60	-	97.60 *	-	66.10	-	97.50 *	-	52.70	-	<u>94.50</u>
DWE	-	71.10	-	97.10	-	66.10	-	96.80	-	51.20	-	91.00
DWE+	-	73.00	-	<u>97.40</u>	-	67.60	-	<u>97.30</u>	-	52.50	-	91.30
DRIN	-	65.50	-	91.30	-	-	-	-	-	51.10	77.90	89.30
MMEL	-	71.50	-	91.70	-	-	-	-	-	-	-	-
GELR	-	84.80	-	-	-	-	-	-	-	89.50 *	-	-
GPT-3.5-Turbo	-	73.80	-	-	-	-	-	-	-	-	-	-
GEMEL	-	82.60	-	-	-	-	-	-	-	<u>86.30</u>	-	-
KGMEL	89.99	87.29	92.47	93.94	80.94	76.40	85.92	88.82	86.00 *	82.12	90.28 *	92.07
CELMI	92.41 *	89.03 *	95.07 *	96.54	88.31 *	83.02 *	92.67 *	94.83	<u>76.72</u>	68.14	<u>83.25</u>	87.63

On WikiMEL, CELMI achieves the best MRR and Hits@1 performance in Table 4, with an MRR of 92.41% and Hits@1 of 89.03%. Compared with the strongest conventional baseline OTMEL, CELMI improves MRR by 0.48% and Hits@1 by 0.83%. The gain is moderate but consistent with the dataset's relatively entity-centric structure, where strong textual and visual cues already support high baseline performance.

On RichpediaMEL, CELMI attains an MRR of 88.31% and Hits@1 of 83.02%, surpassing the second-best conventional model MIMIC by 1.36% in MRR and 2.00% in Hits@1. Relative to the CLIP baseline, CELMI improves MRR by 10.74% and Hits@1 by 15.24%, indicating that multi-channel interaction is especially useful when entity descriptions and images exhibit varied semantic correspondence.

On WikiDiverse, CELMI achieves the highest performance among conventional MEL models, with an MRR of 76.72% and Hits@1 of 68.14%, improving upon MIMIC by 3.28% in MRR and 4.63% in Hits@1. LLM-based methods such as KGMEL obtain higher absolute scores on this dataset, including a 13.98% Hits@1 advantage over CELMI. This gap indicates that generative entity-name prediction and knowledge-guided reasoning remain highly effective for diverse news-domain entities. CELMI therefore should not be viewed as

replacing large generative systems; rather, it provides a lightweight and task-specialized alternative with substantially lower inference cost and a complementary modelling direction for collaborative multimodal evidence.

Across all three datasets, CELMI consistently leads among conventional multimodal entity linking methods, with MRR improvements ranging from 0.48% to 10.74% over competitive baselines. These results support the paper’s central claim: explicitly modelling inter-mention collaboration and separating complementary interaction channels improves MEL beyond independent processing and single-fusion designs.

4.3. Statistical Reliability and Error Analysis

To examine whether the observed gains are reliable under stochastic training, we repeat CELMI and MIMIC over five random seeds and report mean and standard deviation in Table 5. The statistical scope is deliberately limited to CELMI and MIMIC, because many other baselines are available only as aggregate results from public papers and do not provide matched random seeds, training logs, or per-instance predictions. Under this boundary, CELMI maintains stable improvements over MIMIC on MRR and Hits@1 across all three datasets.

Table 5. Five-seed reliability analysis for CELMI and MIMIC. Values are mean $\pm$ standard deviation in percentage points.

Dataset	Model	MRR	Hits@1	Hits@3	Hits@5
WikiMEL	CELMI	92.41 $\pm$ 0.10	89.03 $\pm$ 0.15	95.07 $\pm$ 0.12	96.54 $\pm$ 0.08
WikiMEL	MIMIC	91.82 $\pm$ 0.10	87.98 $\pm$ 0.15	95.07 $\pm$ 0.10	96.37 $\pm$ 0.09
RichpediaMEL	CELMI	88.31 $\pm$ 0.20	83.02 $\pm$ 0.25	92.67 $\pm$ 0.19	94.83 $\pm$ 0.18
RichpediaMEL	MIMIC	86.95 $\pm$ 0.18	81.02 $\pm$ 0.23	91.77 $\pm$ 0.20	94.38 $\pm$ 0.17
WikiDiverse	CELMI	76.72 $\pm$ 0.26	68.14 $\pm$ 0.32	83.25 $\pm$ 0.28	87.63 $\pm$ 0.29
WikiDiverse	MIMIC	73.44 $\pm$ 0.31	63.51 $\pm$ 0.38	81.04 $\pm$ 0.34	86.43 $\pm$ 0.32

The paired significance test results comparing CELMI with MIMIC are summarized in Table 6.

Table 6. Paired significance tests comparing CELMI with MIMIC. MRR uses a Wilcoxon signed-rank test; Hits@ n uses McNemar’s test over test instances. Holm-adjusted p values correct the twelve primary CELMI–MIMIC comparisons.

Dataset	Metric	CELMI	MIMIC	Δ	p	Holm p	Effect
WikiMEL	MRR	92.41	91.82	+0.59	0.0039	0.0234	0.040
WikiMEL	Hits@1	89.03	87.98	+1.05	0.0033	0.0231	0.064
WikiMEL	Hits@3	95.07	95.07	0.00	1.0000	1.0000	0.000
WikiMEL	Hits@5	96.54	96.37	+0.17	0.2710	0.5420	0.015
RichpediaMEL	MRR	88.31	86.95	+1.36	0.00003	0.0003	0.070
RichpediaMEL	Hits@1	83.02	81.02	+2.00	0.00016	0.00144	0.097
RichpediaMEL	Hits@3	92.67	91.77	+0.90	0.0135	0.0675	0.052
RichpediaMEL	Hits@5	94.83	94.38	+0.45	0.0442	0.1326	0.042
WikiDiverse	MRR	76.72	73.44	+3.28	0.000001	0.000011	0.126
WikiDiverse	Hits@1	68.14	63.51	+4.63	0.0000003	0.0000036	0.164
WikiDiverse	Hits@3	83.25	81.04	+2.21	0.0010	0.0080	0.095
WikiDiverse	Hits@5	87.63	86.43	+1.20	0.0221	0.0884	0.066

After Holm correction, CELMI remains significantly better than MIMIC for WikiMEL MRR and Hits@1, RichpediaMEL MRR and Hits@1, and WikiDiverse MRR, Hits@1, and Hits@3. The corrected tests do not support significance claims for WikiMEL Hits@3 or Hits@5, RichpediaMEL Hits@3 or Hits@5, or WikiDiverse Hits@5. Paired five-seed confidence intervals for the CELMI–MIMIC deltas are positive for MRR and Hits@1 on all

datasets: 0.581–0.599 and 1.027–1.073 percentage points on WikiMEL, 1.334–1.386 and 1.964–2.036 on RichpediaMEL, and 3.221–3.339 and 4.543–4.717 on WikiDiverse. These values should be read as evidence for the closest reproducible comparison, not as paired statistical evidence against every published baseline in Table 4.

We further analyze WikiMEL errors by rank bucket and visual-evidence condition. CELMI ranks the gold entity first for 89.03% of test samples and keeps it within the top five for another 7.51%, leaving only 0.91% of cases outside the top 20. Errors are concentrated in visually under-supported settings: Hits@1 is 92.10% when both mention and entity images are available but drops to 80.00% for noisy or weak mention images, 78.64% when the gold entity image is missing, and 75.85% when the mention image is missing.

Finally, we evaluate whether CEC is more useful when more co-occurring mentions are present. Table 7 shows that the full model consistently outperforms w/o CEC, with the Hits@1 gain increasing from 1.08 percentage points for 0–1 co-occurring mentions to 3.37 percentage points for 4+ co-occurring mentions. A sample-level rank-change audit further shows that CEC improves the gold rank in 812 cases (15.71%), leaves the rank bucket unchanged in 4057 cases (78.49%), and worsens it in 300 cases (5.80%). These results indicate that CEC provides useful collaborative evidence while also revealing a boundary condition: misleading co-occurrence can occasionally propagate noise.

Table 7. CEC analysis by co-occurring mention count and rank-change category on WikiMEL.

Co-Mentions	Samples	Full Hits@1	w/o CEC Hits@1	Δ Hits@1	Rank Change	Samples	%	Mean Rank Gain
0–1	1850	85.19	84.11	+1.08	Improved	812	15.71	6.84
2–3	2250	90.13	87.87	+2.26	Unchanged	4057	78.49	0.00
4+	1069	93.36	89.99	+3.37	Worsened	300	5.80	−2.37

The error distribution under different visual conditions is shown in Table 8.

Table 8. WikiMEL error analysis by rank bucket and visual condition.

Rank Bucket	Samples	%	Visual Condition	Samples	Hits@1	Hits@5	MRR
Hit@1	4602	89.03	Both mention and entity images available	3950	92.10	97.97	94.50
Hit@2–5	388	7.51	Mention image noisy or weak	880	80.00	94.66	87.00
Hit@6–10	87	1.68	Gold entity image missing	103	78.64	89.32	84.00
Hit@11–20	45	0.87	Mention image missing	236	75.85	82.63	81.00
Miss@20+	47	0.91	—	—	—	—	—

4.4. Ablation Study

To validate the contribution of each component in CELMI, we conduct ablation experiments across two dimensions. The first dimension separates training efficacy from *architectural necessity*: removing only a channel loss tests whether auxiliary supervision is needed, whereas removing both the loss and the corresponding module tests whether the functional channel itself is needed. The second dimension performs progressive stress ablation to reveal whether channel effects are additive or synergistic under increasingly restricted interaction capacity.

4.4.1. Primary Ablation: Module and Loss Function Analysis

We evaluate the framework through two complementary ablation strategies. Loss function ablation removes specific loss terms ($\mathcal{L}_{TSC}$, $\mathcal{L}_{VPC}$, $\mathcal{L}_{CMAC}$, $\mathcal{L}_{CEC}$, $\mathcal{L}_O$) while retaining all modules, thereby measuring the contribution of the multi-channel consistency objective to optimization. Module ablation removes both the channel-specific loss and

the corresponding functional unit, thereby measuring the architectural contribution of the channel itself. Table 9 presents quantitative results across all three datasets.

Table 9. Ablation study results on WikiMEL, RichpediaMEL, and WikiDiverse. Rows denoted “w/o $\mathcal{L}_X$ ” remove loss functions while preserving modules; rows denoted “w/o $\mathcal{L}_X + X$ ” remove both the loss function and the corresponding module. Bold indicates the best value in each column.

Configuration	WikiMEL				RichpediaMEL				WikiDiverse			
	MRR	Hits@1	Hits@3	Hits@5	MRR	Hits@1	Hits@3	Hits@5	MRR	Hits@1	Hits@3	Hits@5
CELMi (Full)	0.9241	0.8903	0.9507	0.9654	0.8831	0.8302	0.9267	0.9483	0.7672	0.6814	0.8325	0.8763
w/o $\mathcal{L}_{TSC}$	0.8945	0.8484	0.9311	0.9547	0.8546	0.7911	0.9040	0.9293	0.7594	0.6651	0.8330	0.8802
w/o $\mathcal{L}_{VPC}$	0.8702	0.8155	0.9083	0.9451	0.8269	0.7490	0.8908	0.9270	0.7425	0.6449	0.8138	0.8681
w/o $\mathcal{L}_{CMAC}$	0.8967	0.8487	0.9358	0.9563	0.8389	0.7810	0.8835	0.9127	0.7556	0.6631	0.8272	0.8744
w/o $\mathcal{L}_{CEC}$	0.9101	0.8723	0.9393	0.9571	0.8617	0.8007	0.9107	0.9368	0.7574	0.6660	0.8282	0.8797
w/o $\mathcal{L}_O$	0.9155	0.8783	0.9427	0.9602	0.8347	0.7590	0.8994	0.9303	0.7192	0.6155	0.7955	0.8662
w/o $\mathcal{L}_{TSC} + TSC$	0.8953	0.8483	0.9307	0.9559	0.8518	0.7858	0.9043	0.9340	0.7514	0.6569	0.8277	0.8725
w/o $\mathcal{L}_{VPC} + VPC$	0.8977	0.8522	0.9342	0.9555	0.8155	0.7327	0.8829	0.9189	0.7650	0.6747	0.8330	0.8816
w/o $\mathcal{L}_{CMAC} + CMAC$	0.9120	0.8719	0.9453	0.9609	0.7855	0.7398	0.8209	0.8414	0.6994	0.6011	0.7714	0.8234
w/o $\mathcal{L}_{CEC} + CEC$	0.9202	0.8839	0.9495	0.9646	0.8567	0.7994	0.8972	0.9320	0.7359	0.6410	0.9104	0.8595

Figure 3 visualizes the performance gaps induced by module and loss ablations across datasets, quantifying the contribution of each component.

Cross-dataset patterns. Three consistent observations emerge. First, the Cross-Modal Alignment Channel (CMAC) constitutes the architectural lynchpin: its removal induces the most severe degradation across all datasets, with MRR drops of 2.74% on WikiMEL, 9.76% on RichpediaMEL, and 6.78% on WikiDiverse. This disproportionate impact reflects CMAC’s role in projecting heterogeneous textual and visual representations into a unified semantic space; without cross-modal alignment, unimodal channels operate in incompatible feature spaces, preventing effective integration. Accordingly, CELMI should be interpreted as a multi-channel architecture in which CMAC acts as a necessary bridge rather than a sufficient single driver: TSC, VPC, and CEC supply complementary lexical, visual, and collaborative constraints that improve robustness beyond a CMAC-dominated matcher.

Second, the overall loss $\mathcal{L}_O$ provides critical gradient coordination. While removing individual channel losses incurs 2–5% performance reductions, removing $\mathcal{L}_O$ triggers comparatively larger degradations of 0.86% on WikiMEL, 4.84% on RichpediaMEL, and 4.80% on WikiDiverse. This pattern validates the multi-channel consistency mechanism: channel-specific losses ensure discriminative feature learning within individual modules, while $\mathcal{L}_O$ prevents gradient conflicts during joint optimization.

Third, unimodal channels (TSC, VPC) exhibit graceful degradation of 2–4% when individually removed, reflecting architectural fault tolerance: when one modality provides unreliable signals, the remaining channels and cross-modal alignment partially compensate. This characteristic is particularly valuable in real-world scenarios where image quality varies substantially.

The VPC results on WikiDiverse further indicate that visual evidence is not uniformly beneficial across all ranking depths. Removing VPC slightly reduces Hits@1 (68.14% to 67.47%) but yields small improvements in Hits@3 and Hits@5, suggesting that noisy or weakly discriminative visual cues may perturb lower-rank ordering even when they help top-1 disambiguation. This pattern supports the need for adaptive visual gating and also motivates future work on stronger image selection and gate calibration for heterogeneous news-domain images.

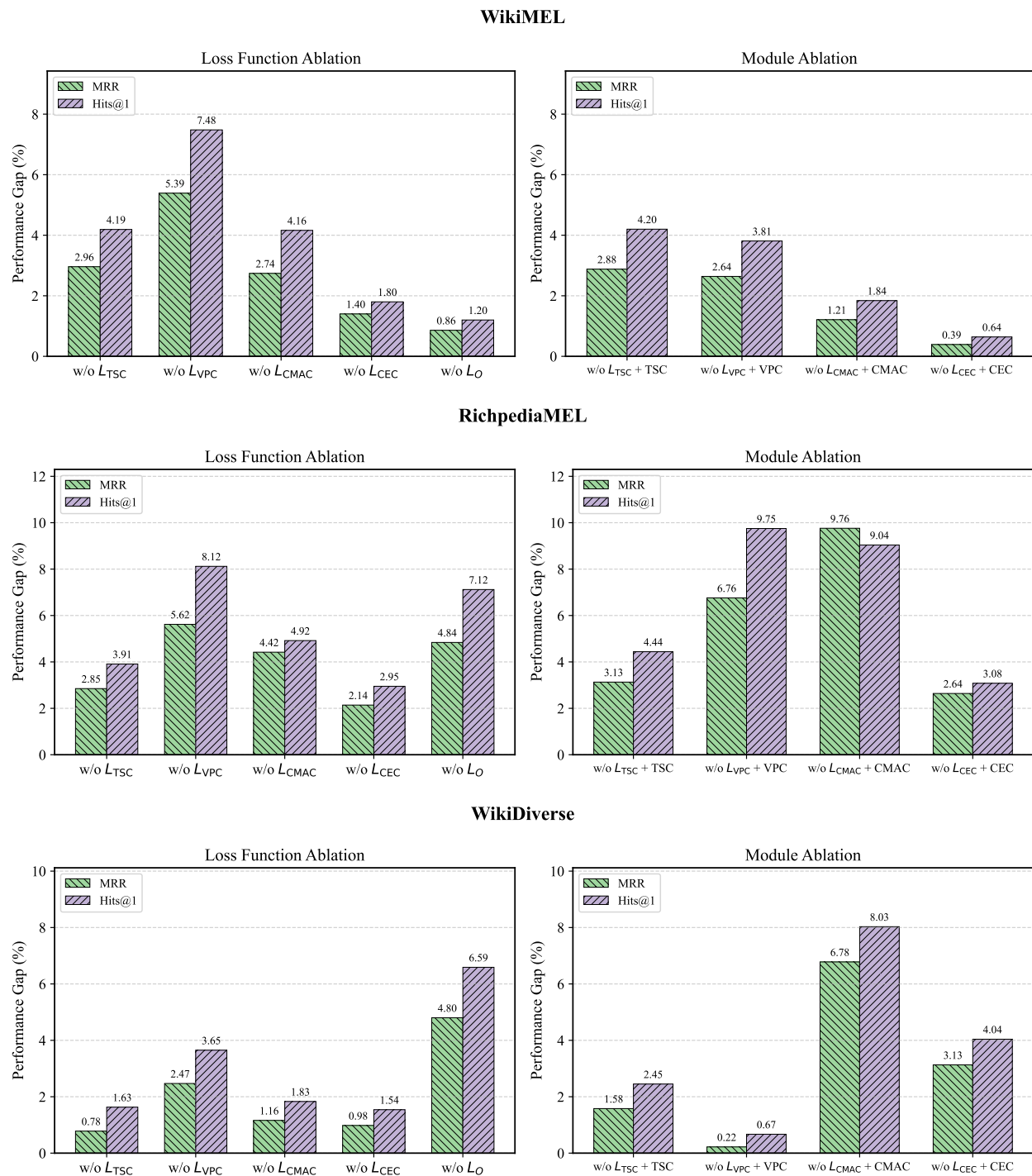


Figure 3. Module and loss function ablation performance gaps across three datasets. **Left** panels show performance degradation when functional modules are removed; **Right** panels show degradation when individual loss components are removed. CMAC consistently exhibits the largest single-module impact, validating its role as the architectural bridge between modalities.

Dataset-specific characteristics. WikiMEL demonstrates the smallest performance variance across ablations (0.86–2.96% MRR range), indicating balanced reliance on all four channels; its entity-centric structure with clear textual descriptions and high-quality images facilitates stable multimodal fusion. RichpediaMEL exhibits the greatest sensitivity to CMAC removal (9.76% MRR drop), reflecting its dependence on cross-modal alignment to reconcile diverse semantic categories with varying text-image correlation patterns. WikiDiverse shows pronounced sensitivity to L_O removal (4.80% MRR drop versus

0.86% on WikiMEL), suggesting that its news-domain mentions with complex contextual dependencies benefit most from global gradient coordination.

The functional necessity of each component is further confirmed through coupled ablations: removing both the loss and the corresponding module consistently causes larger degradation than removing the loss alone, confirming that architectural design and training mechanisms contribute independently to overall performance.

4.4.2. Channel Weight Sensitivity

To further test whether CELMI's gains can be explained by a CMAC-dominated aggregation, we conduct a channel-weight sensitivity experiment on WikiMEL and RichpediaMEL. The default setting uses equal fusion weights (1, 1, 1, 1) for TSC, VPC, CMAC, and CEC. We compare it with a CMAC-heavy fixed setting (0.1, 0.1, 0.7, 0.1), a CEC-disabled setting under equal weights, and a learnable fusion setting initialized from equal weights. Table 10 reports the results.

Table 10. Channel weight sensitivity results. The CMAC-heavy setting increases the CMAC fusion weight while retaining all channels; w/o CEC disables the collaborative entity channel under equal weights. Δ MRR is measured in percentage points relative to equal fusion on the same dataset. Bold values indicate the best performance within each dataset.

Dataset	Configuration	Weights (TSC, VPC, CMAC, CEC)	MRR	Hits@1	Hits@3	Hits@5	Δ MRR
WikiMEL	Equal fusion	(1, 1, 1, 1)	0.9241	0.8903	0.9507	0.9654	—
WikiMEL	CMAC-heavy fixed	(0.1, 0.1, 0.7, 0.1)	0.9237	0.8932	0.9469	0.9642	−0.04
WikiMEL	w/o CEC	(1, 1, 1, 1)	0.8946	0.8474	0.9319	0.9533	−2.95
WikiMEL	Learnable fusion	initialized as (1, 1, 1, 1)	0.8859	0.8391	0.9188	0.9463	−3.82
RichpediaMEL	Equal fusion	(1, 1, 1, 1)	0.8831	0.8302	0.9267	0.9483	—
RichpediaMEL	CMAC-heavy fixed	(0.1, 0.1, 0.7, 0.1)	0.8627	0.8009	0.9101	0.9368	−2.04
RichpediaMEL	w/o CEC	(1, 1, 1, 1)	0.8039	0.7102	0.8718	0.9107	−7.92
RichpediaMEL	Learnable fusion	initialized as (1, 1, 1, 1)	0.8508	0.7865	0.9043	0.9291	−3.23

The sensitivity results support three conclusions. First, increasing the CMAC weight does not consistently improve performance: on WikiMEL it yields only a small Hits@1 gain (+0.29 percentage points) while slightly reducing MRR, Hits@3, and Hits@5; on RichpediaMEL it lowers MRR by 2.04 percentage points and Hits@1 by 2.93 percentage points. Thus, CMAC is essential for cross-modal alignment, but over-weighting it weakens the complementary constraints supplied by the other channels. Second, disabling CEC causes clear degradation, especially on RichpediaMEL, where MRR drops by 7.92 percentage points and Hits@1 drops by 12.00 percentage points. This confirms that collaborative entity interaction is not merely decorative even when CMAC remains active. Third, unconstrained learnable fusion underperforms equal fusion on both datasets, suggesting that the balanced fixed aggregation acts as a useful regularizer against channel dominance under the current training regime.

4.4.3. Progressive Stress Ablation

To address whether small individual performance drops indicate architectural redundancy, we conduct progressive stress ablation on WikiDiverse—the dataset exhibiting the highest entity ambiguity and contextual complexity. The ablation progressively restricts interaction capacity from the full model to collaborative-free, cross-modal-free, and finally single-unimodal endpoints: Full $\rightarrow$ w/o CEC $\rightarrow$ w/o CEC + CMAC $\rightarrow$ TSC-only $\rightarrow$ VPC-only.

Figure 4 illustrates the degradation trajectory, and Table 11 provides detailed metrics at each removal step.

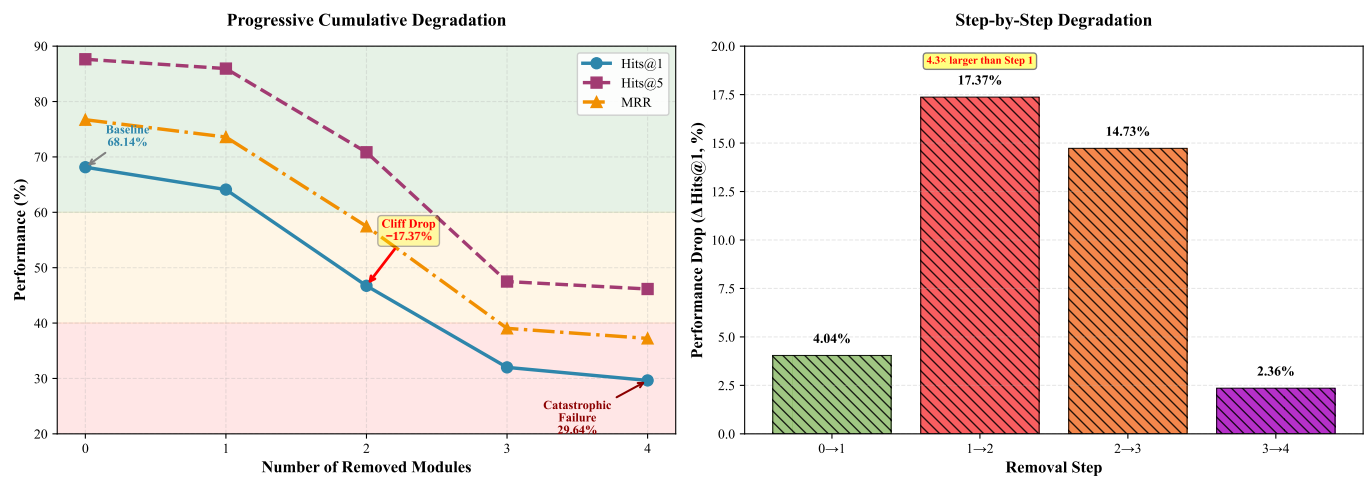


Figure 4. Progressive stress ablation on WikiDiverse. Left: Performance degradation curve showing Hits@1, Hits@5, and MRR as interaction capacity is reduced toward single-channel endpoints. The cliff drop at Step 2 demonstrates CMAC’s critical role in maintaining cross-modal semantic coherence. Right: Step-by-step degradation magnitude. The Step 1→2 drop (17.37%) is 4.3× larger than the Step 1 drop (4.04%), revealing non-linear synergistic dependencies between CEC and CMAC.

Table 11. Progressive stress ablation results on WikiDiverse. Steps 0–2 progressively remove collaborative and cross-modal interaction capacity; Steps 3–4 compare the two single-unimodal endpoints. The non-linear degradation—particularly the cliff drop at Step 2—demonstrates that cumulative absolute drops substantially exceed the sum of individual contributions.

Step	Configuration	Hits@1	Hits@5	MRR	Δ Hits@1
0	Full Model	68.14%	87.63%	0.7672	—
1	w/o CEC	64.10%	85.95%	0.7359	−4.04%
2	w/o CEC + CMAC	46.73%	70.84%	0.5747	−17.37%
3	TSC-only (w/o CEC + CMAC + VPC)	32.00%	47.50%	0.3905	−14.73%
4	VPC-only (w/o CEC + CMAC + TSC)	29.64%	46.15%	0.3723	−2.36%

The progressive stress ablation reveals three distinct degradation regimes. In the graceful degradation regime (Step 0→1), removing CEC causes a modest 4.04% drop as the three remaining channels maintain core matching functionality. At Step 2, however, a cliff drop is observed: removing CMAC on top of CEC produces an additional 17.37% degradation—4.3× larger than the Step 1 drop. This non-linearity demonstrates synergistic dependency: collaborative reasoning in CEC requires cross-modal alignment to project entity representations into a unified space; without CMAC, graph-based propagation in CEC operates on incompatible feature spaces, collapsing structural consistency mechanisms.

Steps 3–4 enter the catastrophic failure regime, where performance approaches the random baseline. These two steps should be read as single-unimodal endpoints rather than a strict continuation of Step 3 into Step 4: TSC-only retains only textual matching, whereas VPC-only retains only visual matching. The TSC-only endpoint (32.00% Hits@1) marginally outperforms the VPC-only endpoint (29.64% Hits@1), confirming the text-centric nature of knowledge base entity descriptions while showing that neither unimodal endpoint can recover the collaborative and cross-modal evidence used by the full model.

The total cumulative degradation is a 38.50 percentage-point absolute Hits@1 drop, which substantially exceeds the sum of individual module contributions predicted from primary ablation. Additive prediction based on individual absolute drops would yield approximately 2.55 (VPC) + 17.18 (CMAC) + 4.04 (CEC) = 23.77 percentage points; the observed 38.50 percentage-point drop is 1.62× larger, demonstrating multiplicative inter-channel interaction effects. This analysis refutes the redundancy hypothesis: individual ro-

business stems from complementary compensation mechanisms that are eliminated under cumulative removal, exposing the true dependency structure of the four-channel design.

4.5. Computational Efficiency and Scalability Analysis

We further evaluate the computational efficiency of CELMI under increasing candidate set sizes and co-mention densities. All measurements are conducted on an NVIDIA RTX A5000 GPU with a batch size of 16. Entity representations are precomputed offline, while online inference comprises multimodal mention encoding, chunked transfer of cached entity representations from CPU to GPU, four-channel candidate matching, and complete candidate sorting. The current system does not employ a separate candidate retrieval stage; it performs exhaustive ranking over the evaluated entity set as a stress test rather than as the recommended deployment design. Image loading, preprocessing, and tokenization are excluded from the GPU inference timing. Because CEC-related computation is integrated into the matcher in the current implementation, it is included in the matching latency rather than reported as a separate graph construction stage.

Table 12 reports steady-state online latency as the candidate set increases from 1000 entities to the complete WikiMEL knowledge base of 109,976 entities. Each result is the mean and standard deviation over 20 timing measurements, obtained after three warm-up iterations. End-to-end latency increases approximately linearly with candidate set size, from 12.60 ms per mention for 1000 candidates to 1046.53 ms for exhaustive full-KB ranking. The 1000–20,000 candidate rows approximate the reranking cost after a practical retrieval stage has already reduced the candidate set; they are not measurements of retrieval latency itself. At full scale, transfer of cached entity representations accounts for approximately 88.1% of online latency, while matching accounts for 11.7%; complete sorting contributes less than 0.01%. Thus, the principal deployment bottleneck in the exhaustive stress test is entity-cache movement rather than the ranking operation itself.

Table 12. Online runtime and memory scaling with the number of candidate entities.

Candidates	Encoding (ms)	Transfer (ms)	Matching (ms)	Ranking (ms)	End-to-End (ms)	Peak Memory (MB)
1000	2.598 ± 0.022	8.888 ± 1.528	1.107 ± 0.016	0.007 ± 0.000	12.601 ± 1.561	1482.7
6000	2.620 ± 0.012	56.120 ± 0.272	6.714 ± 0.035	0.012 ± 0.001	65.466 ± 0.277	5865.5
20,000	2.593 ± 0.022	172.178 ± 0.245	22.209 ± 0.083	0.015 ± 0.001	196.994 ± 0.280	5866.3
109,976	2.591 ± 0.018	921.488 ± 0.551	122.412 ± 0.213	0.041 ± 0.002	1046.532 ± 0.502	5871.8

The memory results demonstrate the effect of chunked evaluation. Peak allocated GPU memory increases from 1.48 GB at 1000 candidates to 5.87 GB at a chunk size of 6000 but remains nearly constant as the total candidate set grows from 6000 to 109,976 entities. The corresponding peak reserved memory is 6.78 GB for full-KB evaluation. The model-and-input baseline before candidate scoring allocates approximately 596.8 MB. Consequently, chunking bounds GPU memory independently of the total knowledge base size, although repeated CPU-to-GPU transfer increases latency.

To estimate the computational contribution of CEC, we additionally compare full-matcher latency with inference-time CEC disabled. At full-KB scale, the estimated CEC increment is 56.32 ± 0.18 ms per mention, corresponding to approximately 46.0% of matcher time but only 5.4% of end-to-end latency because entity transfer dominates the current exhaustive-ranking pipeline.

Table 13 compares stable full-KB measurements across naturally occurring co-mention groups. Increasing the number of co-mentions from 2–3 to 4+ produces no material increase in encoding or end-to-end latency. This behavior is consistent with the implementation,

which concatenates co-mention text and pads or truncates it to a fixed input length of 40 tokens.

Table 13. Full-KB latency under increasing co-mention density.

Co-Mentions	Encoding (ms)	Transfer (ms)	Matching (ms)	Ranking (ms)	End-to-End (ms)	Peak Memory (MB)
2–3	2.589 ± 0.018	921.508 ± 0.601	122.335 ± 0.261	0.041 ± 0.002	1046.472 ± 0.522	5871.8
4+	2.593 ± 0.019	921.468 ± 0.528	122.490 ± 0.121	0.041 ± 0.002	1046.592 ± 0.501	5871.8

Entity representations are encoded once and cached before online evaluation. Encoding all 109,976 WikiMEL entities requires 1535.63 s (25.59 min), corresponding to 13.96 ms per entity, with peak allocated and reserved GPU memory of 2.17 GB and 3.06 GB, respectively. This one-time offline cost is excluded from the online latency values in Tables 12 and 13.

Finally, CELMI contains 151.913 M parameters in total. The shared encoder accounts for 151.400 M parameters (99.663%), whereas the complete four-channel matcher contains 0.513 M parameters (0.337%). TSC, VPC, CMAC, and CEC contain 0.197 M, 0.057 M, 0.059 M, and 0.197 M parameters, respectively. These results confirm that the additional matching architecture is parameter-efficient. For deployment, CELMI is better suited to retrieval plus reranking than to exhaustive full-KB scoring; the present full-KB measurement identifies the upper-bound cost and shows that entity-representation transfer is the dominant bottleneck under CPU-cached exhaustive evaluation. The parameter distribution of CELMI is reported in Table 14.

Table 14. Parameter distribution of CELMI.

Component	Parameters	Percentage of Model
Shared encoder	151.400 M	99.663%
Complete matcher	0.513 M	0.337%
TSC	0.197 M	0.130%
VPC	0.057 M	0.037%
CMAC	0.059 M	0.039%
CEC	0.197 M	0.130%
Total	151.913 M	100.000%

5. Conclusions

This paper presents CELMI, a collaborative multimodal entity linking framework motivated by a structural limitation of existing MEL systems: difficult mentions are often resolved by evidence distributed across co-occurring mentions and modalities, yet most models still make independent mention-level decisions. CELMI addresses this mismatch through four specialized channels. The Textual Semantics Channel captures both lexical and contextual correspondences; the Visual Perception Channel suppresses noisy visual regions through dynamic gating; the Cross-Modal Alignment Channel maps text and image features into a shared semantic subspace; and the Collaborative Entity Channel propagates inter-mention consistency signals through attention-based graph interaction. A multi-channel consistency objective further stabilizes joint training by supervising both individual channels and the final ranking score.

Effectiveness and necessity. Experiments on WikiMEL, RichpediaMEL, and WikiDiverse demonstrate that CELMI achieves the strongest MRR and Hits@1 performance among conventional non-LLM/VLM MEL models, with MRR scores of 92.41%, 88.31%, and 76.72%, respectively. The gains are most pronounced on settings where multimodal and contextual ambiguity require more than local mention matching. Ablation studies further confirm

that the four channels are not redundant: CMAC produces the largest single-module degradation when removed, while progressive stress ablation on WikiDiverse reveals a 38.50 percentage-point absolute Hits@1 drop when the architecture is reduced to a single unimodal endpoint. Channel-weight sensitivity also shows that a CMAC-heavy fusion does not consistently outperform equal fusion and that disabling CEC causes substantial degradation, supporting the central claim that CELMI's improvement arises from inter-channel synergy and collaborative evidence propagation.

Innovation and practical utility. CELMI contributes a channelized architecture for collaborative MEL and a consistency objective that prevents channel collapse during joint optimization. Unlike methods that rely on complex global graph inference or a single fused mention representation, CELMI embeds collaborative reasoning into lightweight channel-specific matchers operating on shared CLIP representations. The four-channel matcher adds only 0.513 M parameters, accounting for 0.337% of the full model. Full-KB online inference over 109,976 WikiMEL entities requires approximately 1.047 s per mention under the measured CPU-cached exhaustive-ranking stress test, with bounded peak allocated GPU memory of about 5.87 GB. Candidate-limited reranking is substantially faster, requiring 12.60 ms for 1000 candidates and 196.99 ms for 20,000 candidates under the same measurement boundary. These results indicate that CELMI is parameter-efficient, while also identifying entity-cache transfer as the main bottleneck for future deployment optimization.

Limitations and future directions. Several limitations remain. First, although CELMI is competitive among conventional MEL models, it still trails large generative or knowledge-enhanced systems on WikiDiverse, indicating that collaborative interaction does not fully replace foundation-model-scale knowledge and generative reasoning. Second, the graph-based Collaborative Entity Channel depends on the availability and quality of co-occurring mention annotations; in raw document deployment, upstream NER errors may affect graph construction, and robustness under predicted mention boundaries remains to be evaluated. Third, CEC introduces additional overhead compared with single-stream models, particularly when many mentions co-occur in the same context. Fourth, collaborative modelling naturally provides less benefit for samples containing only one or very few mentions, suggesting that adaptive channel activation could improve efficiency. Finally, the experiments are conducted mainly on English-centric Wikipedia-derived datasets, leaving cross-lingual, low-resource, and more informal social media scenarios insufficiently explored. Random image selection during training may also be suboptimal when entity image sets contain large quality variation, and highly ambiguous entities with similar textual descriptions and appearances may require stronger external knowledge such as relational triples.

Future work will pursue three directions: (1) optimizing computational efficiency through dynamic channel pruning and adaptive mention graph construction; (2) extending collaborative modeling to exploit knowledge graph topology through deeper graph neural network architectures; and (3) evaluating cross-domain generalization on challenging datasets such as social media content with low-quality images and informal text, and assessing multilingual capabilities across diverse language families. These directions aim to address the identified limitations while broadening the applicability of collaborative multimodal entity linking to more diverse and demanding real-world scenarios.

Author Contributions: Conceptualization, H.L., H.P. and P.Z.; methodology, X.H., H.L., H.P. and P.Z.; software, X.H.; validation, X.H., H.L., P.Z., K.L. and Y.Y.; formal analysis, X.H. and K.L.; investigation, X.H. and Y.Y.; resources, H.L., P.Z. and Y.Y.; data curation, X.H.; writing—original draft preparation, X.H.; writing—review and editing, K.L., H.P., Y.Y., H.L. and P.Z.; visualization, X.H. and K.L.; supervision, H.L., H.P. and P.Z.; project administration, H.L., H.P. and P.Z.; funding acquisition, H.L., H.P. and P.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the Beijing Natural Science Foundation under Grant 4264135, partially supported by the National Natural Science Foundation of China under Grant 62341130, partially supported by the Natural Science Foundation of Tianjin under Grant 24JCQNJC00280, and partially supported by the Natural Science Foundation of Shandong Province under Grants ZR2023LZH017 and ZR2024MF066.

Data Availability Statement: The original data presented in this study are openly available on GitHub at <https://github.com/pengfei-luo/MIMIC> (accessed on 29 July 2026). Our model is available at <https://github.com/sim1n999/CELMi> (accessed on 29 July 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Shen, W.; Wang, J.; Han, J. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Trans. Knowl. Data Eng.* **2014**, *27*, 443–460.
- Li, H.; He, X.; Yue, Y.; Wang, C. Advances and Challenges in Multimodal Entity Linking: A Comprehensive Survey. *Int. J. Softw. Eng. Knowl. Eng.* **2025**, *35*, 943–961.
- Shen, W.; Li, Y.; Liu, Y.; Han, J.; Wang, J.; Yuan, X. Entity linking meets deep learning: Techniques and solutions. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 2556–2578.
- Xu, Y.; Wang, Z.; Wang, Z.; Guo, Y.; Fan, R.; Tian, H.; Wang, X. SimDCL: Dropout-based simple graph contrastive learning for recommendation. *Complex Intell. Syst.* **2023**, *9*, 4751–4763.
- De Cao, N.; Wu, L.; Papat, K.; Artetxe, M.; Goyal, N.; Plekhanov, M.; Zettlemoyer, L.; Cancedda, N.; Riedel, S.; Petroni, F. Multilingual autoregressive entity linking. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 274–290.
- Ge, C.; Wang, P.; Chen, L.; Liu, X.; Zheng, B.; Gao, Y. CollaborEM: A self-supervised entity matching framework using multi-features collaboration. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 12139–12152.
- Chen, H.; Zhang, F.; Li, Q.; Li, X.; Ding, Y.; Zhang, D.; Cheng, J.; Wang, X. Triple confidence-aware encoder–decoder model for commonsense knowledge graph completion. *Int. J. Mach. Learn. Cybern.* **2025**, *16*, 2073–2091.
- Wang, Z.; Xu, Y.; Wang, Z.; Fan, R.; Guo, Y.; Li, W. Knowledge graph-aware deep interest extraction network on sequential recommendation. *Neural Process. Lett.* **2024**, *56*, 207.
- 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)*, Online; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 6397–6407. <https://doi.org/10.18653/v1/2020.emnlp-main.519>.
- Wang, P.; Wu, J.; Chen, X. Multimodal Entity Linking with Gated Hierarchical Fusion and Contrastive Training. In *Proceedings of the SIGIR*; Association for Computing Machinery: New York, NY, USA, 2022; pp. 938–948.
- Xing, S.; Zhao, F.; Wu, Z.; Li, C.; Zhang, J.; Dai, X. Drin: Dynamic relation interactive network for multimodal entity linking. In *Proceedings of the 31st ACM International Conference on Multimedia*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 3599–3608.
- 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*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 1583–1594.
- Shi, S.; Xu, Z.; Hu, B.; Zhang, M. Generative multimodal entity linking. *arXiv* **2023**, arXiv:2306.12725.
- Kim, J.; Lee, G.; Kim, T.; Shin, K. KGME: Knowledge Graph-Enhanced Multimodal Entity Linking. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*; Association for Computing Machinery: New York, NY, USA, 2025.
- Liu, Q.; He, Y.; Lian, D.; Zheng, Z.; Xu, T.; Liu, C.; Chen, E. UniMEL: A Unified Framework for Multimodal Entity Linking with Large Language Models. *arXiv* **2024**, arXiv:2407.16160.
- Xu, Y.H.; Wang, Z.H.; Wang, Z.R.; Fan, R.; Wang, X. A recommendation algorithm based on a self-supervised learning pretrain transformer. *Neural Process. Lett.* **2023**, *55*, 4481–4497.
- Li, Y.; Chen, Y.; Zhu, K.; Bai, C.; Zhang, J. An effective federated learning verification strategy and its applications for fault diagnosis in industrial IoT systems. *IEEE Internet Things J.* **2022**, *9*, 16835–16849.
- Sui, X.; Zhang, Y.; Zhao, Y.; Song, K.; Zhou, B.; Yuan, X. MELOV: Multimodal entity linking with optimized visual features in latent space. In *Proceedings of the Findings of the Association for Computational Linguistics ACL 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 816–826.
- Cao, Y.; Hou, L.; Li, J.; Liu, Z. Neural Collective Entity Linking. *arXiv* **2018**, arXiv:1811.08603.

20. Deng, Z.; Li, Z.; Yang, Q.; Liu, Q.; Chen, Z. Improving Entity Linking with Graph Networks. In *Proceedings of the WISE (1)*; Springer: Cham, Switzerland, 2020; pp. 343–354.
21. Fang, Z.; Cao, Y.; Zhang, D.; Li, Q.; Zhang, Z.; Liu, Y. Joint Entity Linking with Deep Reinforcement Learning. *arXiv* **2019**, arXiv:1902.00330.
22. Sun, J.; Zhou, Y.; Zong, C. Learning to Select the Next Reasonable Mention for Entity Linking. *arXiv* **2021**, arXiv:2112.04104.
23. Yang, C.; He, B.; Wu, Y.; Xing, C.; He, L.; Ma, C. MMEL: A Joint Learning Framework for Multi-Mention Entity Linking. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence, 31 July–4 August 2023*; PMLR: Cambridge, MA, USA, 2023; Volume 216, pp. 2411–2421.
24. Gu, Y.; Dong, L.; Wei, F.; Huang, M. MiniLLM: Knowledge Distillation of Large Language Models. In *Proceedings of the Twelfth International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024*.
25. Agarwal, R.; Vieillard, N.; Zhou, Y.; Stanczyk, P.; Ramos, S.; Geist, M.; Bachem, O. Generalized Knowledge Distillation for Auto-regressive Sequence Models. *arXiv* **2024**, arXiv:2306.13649.
26. Zhao, H.; Wang, B.; Fu, Y.; Li, N.; Gao, Z. Fixed-time adaptive fault-tolerant control of a multi-mode VTOL UAV with variable prescribed performance boundaries under random disturbances. *ISA Trans.* **2026**, *175*, 126–138.
27. Zhao, H.; Wang, B.; Fu, Y.; Zhou, M.; Li, N.; Gao, Z. Fixed-time adaptive smooth control with relaxed performance constraints for a canard rotor/wing UAV subject to aerodynamic interferences and actuator faults. *Aerosp. Sci. Technol.* **2026**, *178*, 112959. <https://doi.org/10.1016/j.ast.2026.112959>.
28. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2021; pp. 8748–8763.
29. Wang, X.; Tian, J.; Gui, M.; Li, Z.; Wang, R.; Yan, M.; Chen, L.; Xiao, Y. WikiDiverse: A Multimodal Entity Linking Dataset with Diversified Contextual Topics and Entity Types. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 4785–4797. <https://doi.org/10.18653/v1/2022.acl-long.328>.
30. Wang, M.; Wang, H.; Qi, G.; Zheng, Q. Richpedia: A large-scale, comprehensive multi-modal knowledge graph. *Big Data Res.* **2020**, *22*, 100159.
31. Vrandečić, D.; Krötzsch, M. Wikidata: A free collaborative knowledgebase. *Commun. ACM* **2014**, *57*, 78–85.
32. Devlin, J.; Chang, M.W.; 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, Volume 1 (Long and Short Papers), Minneapolis, Minnesota*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186. <https://doi.org/10.18653/v1/N19-1423>.
33. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
34. 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*; PMLR: Cambridge, MA, USA, 2021; pp. 5583–5594.
35. Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; Hoi, S.C.H. Align before fuse: Vision and language representation learning with momentum distillation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 9694–9705.
36. Dou, Z.Y.; Xu, Y.; Gan, Z.; Wang, J.; Wang, S.; Wang, L.; Zhu, C.; Zhang, P.; Yuan, L.; Peng, N.; 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*; IEEE: New York, NY, USA, 2022; pp. 18166–18176.
37. Zhang, Z.; Sheng, J.; Zhang, C.; Liang, Y.; Zhang, W.; Wang, S.; Liu, T. Optimal Transport Guided Correlation Assignment for Multimodal Entity Linking. In *Proceedings of the Findings of the Association for Computational Linguistics ACL 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 4103–4117.
38. Song, S.; Li, S.; Yu, J.; Zhao, S.; Li, X.; Ma, J.; Liu, X.; Li, Z.; Mao, X. Dim: Dynamic integration of multimodal entity linking with large language model. In *Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*; Springer: Singapore, 2024; pp. 187–200.
39. Song, S.; Zhao, S.; Wang, C.; Yan, T.; Li, S.; Mao, X.; Wang, M. A dual-way enhanced framework from text matching point of view for multimodal entity linking. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 19008–19016.
40. Song, S.; Li, S.; Zhao, S.; Li, X.; Wang, C.; Yu, J.; Ma, J.; Yan, T.; Ji, B.; Mao, X. DWE+: Dual-Way Matching Enhanced Framework for Multimodal Entity Linking. *arXiv* **2024**, arXiv:2404.04818.
41. Long, X.; Zeng, J.; Meng, F.; Zhou, J.; Zhou, B. Trust in Internal or External Knowledge? Generative Multi-Modal Entity Linking with Knowledge Retriever. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 7559–7569. <https://doi.org/10.18653/v1/2024.findings-acl.450>.
42. OpenAI. ChatGPT. 2025. Available online: <https://chatgpt.com/> (accessed on 5 July 2025).

43. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimeshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32.
44. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representations*, New Orleans, LA, USA, 6–9 May 2019.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.