electronics

MDPI

Article

Joint Fine-Grained Entity-Type Disambiguation and Robust Grounding via Multi-Level Knowledge Integration for FMNERG

Jiakai Geng [1,2], Chenyang Zhang [1,*], Xufeng Zhou [1,2], Linjing Li [1,2] and Daniel Zeng [1,2]

[1] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China; gengjiakai2020@ia.ac.cn (J.G.); zhouxufeng2019@ia.ac.cn (X.Z.); linjing.li@ia.ac.cn (L.L.); dajun.zeng@ia.ac.cn (D.Z.)
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: chenyang.zhang@ia.ac.cn

Abstract

Fine-grained Multimodal Named-Entity Recognition and Grounding (FMNERG) aims to extract entity–type–object triples from image–text pairs. However, existing methods still struggle with two major challenges: distinguishing semantically similar labels in an expanded fine-grained label space and achieving robust entity–object alignment under complex image–text relations in social media. To address these issues, we propose a **M**ulti-level **K**nowledge **I**ntegration **F**ramework (MKIF) for FMNERG, which leverages multi-level knowledge to jointly improve fine-grained entity-type disambiguation and robust entity grounding in a unified manner. Specifically, we employ multimodal large language models (MLLMs) to obtain global-level knowledge, enriching the semantics of raw multimodal inputs and serving as a shared semantic anchor across subtasks. To improve fine-grained entity-type discrimination, we further derive conceptual and discriminative descriptions for each type as label-level knowledge and utilize both global- and label-level knowledge as supervision via contrastive learning. In addition, we discover potential visual concepts in the dataset as object-level knowledge and design a dynamic gating mechanism to incorporate the global- and object-level knowledge into object embeddings for robust entity–object alignment. Comprehensive experiments demonstrate that MKIF outperforms state-of-the-art methods. Furthermore, detailed analyses validate the effectiveness of each core component.

Keywords: fine-grained multimodal named-entity recognition and grounding; multi-level knowledge; visual concept; contrastive learning; multimodal large language model

Check for updates

Academic Editors: Donglin Zhang and Zhen Liu

Received: 1 September 2026
Revised: 21 October 2026
Accepted: 3 November 2026
Published: 6 November 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) license.

1. Introduction

Multimodal named-entity recognition (MNER) has attracted increasing attention in information extraction and multimodal learning. It aims to extract named entities from unstructured text with auxiliary images [1–9]. However, existing MNER studies divide entities into coarse-grained entity types (i.e., *person*, *location*, *organization*, and *miscellaneous*) and only extract entities from text without explicitly linking them to their corresponding regions in the image, restricting the applications of MNER in real-world scenarios. To address these limitations, Fine-grained Multimodal Named-Entity Recognition and Grounding (FMNERG) [10] has recently been proposed. FMNERG includes two subtasks of MNER and Entity Extraction and Grounding (EEG), aiming to extract triples that contain named entities, their fine-grained types and grounded bounding boxes in the image. Compared to MNER, FMNERG has wider applications, including multimodal knowledge graph construction [11], human–machine dialogue [12], visual question answering [13],

and privacy-aware multimodal analysis [14]. Despite the rapid progress of multimodal large language models (MLLMs), directly applying them to FMNERG remains nontrivial [5,15,16], especially when structured multimodal extraction requires both fine-grained semantic discrimination and reliable entity–object grounding.

The major challenges of FMNERG stem from the entity–type and entity–object gaps in entity–type–object triple extraction. The first gap arises from the difficulty of associating entities with their fine-grained types as the number of candidate types increases. Fine-grained type hierarchies contain semantically similar subtypes under the same coarse-grained category, resulting in blurred type boundaries that make the subtypes difficult to distinguish [17]. As shown in Figure 1a, the entity *Francona* can be easily classified into the *person* category under the coarse-grained setting. However, in FMNERG, the model needs to further recognize it as a *coach* rather than other semantically similar fine-grained categories, such as *athlete*, because the coaches are often seen on the sidelines of a field analyzing performance, while the athletes are seen in action like running or jumping in a stadium or track field. This requires the model not only to effectively leverage the multimodal cues in the tweet but also to capture the conceptual and discriminative features of fine-grained entity types for accurate entity-type extraction. The second gap lies in the complex and diverse relationships between textual entities and visual content on social media, where image–text pairs may exhibit varying degrees and forms of semantic relevance [18,19], making it difficult to reliably localize the visual regions corresponding to the entities. Figure 1b demonstrates this challenge. Given several entities which belong to the *band* category, the visual concepts associated with their corresponding regions include logos, group photos, posters and text, which are diverse in the feature space. Addressing such complex entity–object relations requires robust entity–object alignment in EEG.

In response to these challenges, two-stage methods [5,20] first extract entity–type pairs and then perform entity-aware visual grounding in a pipeline paradigm, which may lead to error propagation since MNER and EEG are treated as separate tasks. Therefore, unified methods have recently attracted attention. Sequence-to-sequence-based methods H-Index [11] and TIGER [10] directly generate the target entity–type–objects sequence from the multimodal input via pre-trained sequence-to-sequence models. MQSPN [6] further reformulates the FMNERG task as set prediction to learn intra- and inter-entity relationships of multimodal entities, which addresses the exposure bias of sequence-to-sequence-based methods. However, these methods focus on predicting fine-grained labels through autoregressive generation of entity–type–region sequences or set prediction guided by type-grained queries, without explicitly modeling the semantic differences among similar fine-grained types. GEM [21] then exploits external knowledge and multimodal large language model (MLLM)-based multimodal reranking for fine-grained type disambiguation, but it still establishes the associations between entities and visual regions only in feature space, which fails to handle the complex entity–object relations on social media.

To address both issues of distinguishing semantically similar entity types and robustly modeling entity–object relations, we draw inspiration from MLLMs. Recently, MLLMs [22–25] have demonstrated remarkable capabilities of multimodal understanding [26,27], making it possible to obtain semantically rich knowledge for multimodal inputs and fine-grained entity-type discrimination. Building on this insight, we leverage MLLMs to acquire both global- and label-level knowledge, providing sample-specific context and type-specific semantic cues, respectively. To further improve entity–object alignment, we discover dataset-level visual concepts as object-level knowledge to augment object representations. In this way, multi-level knowledge can be incorporated to support different subtasks and jointly bridge the two gaps of FMNERG.

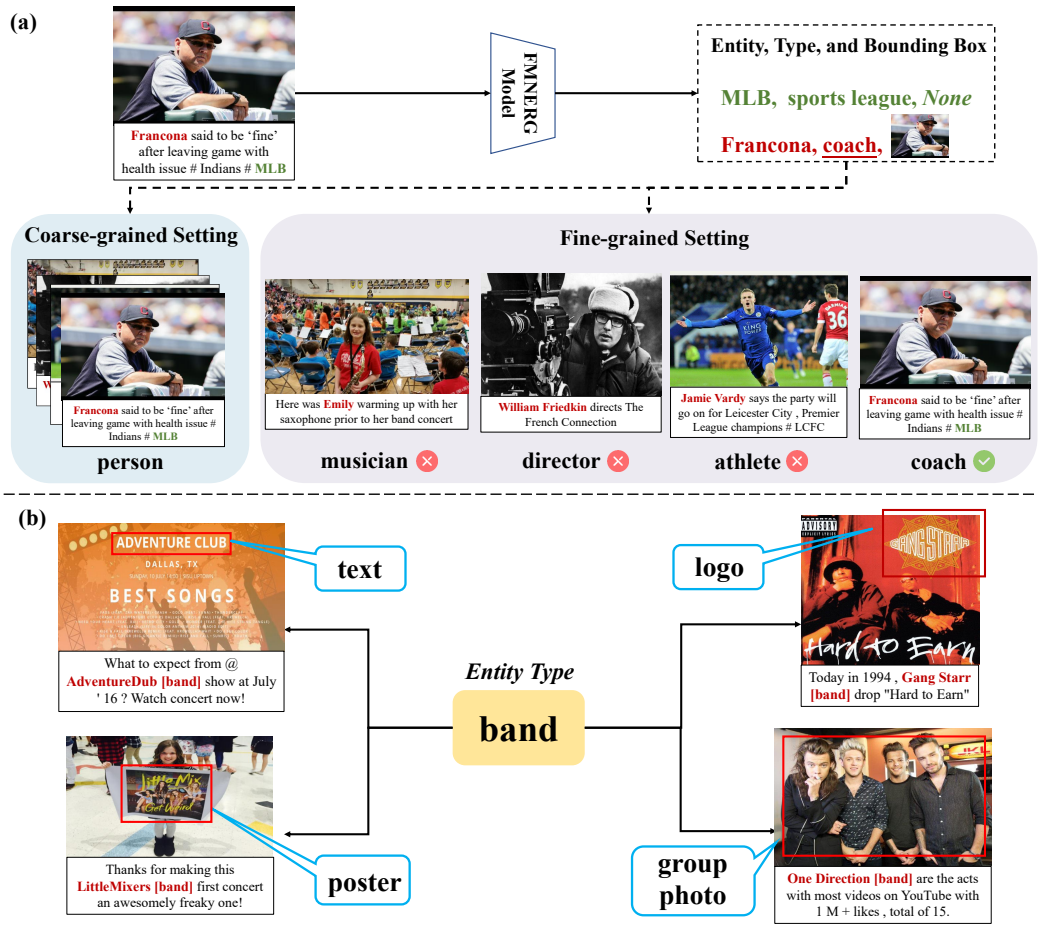


Figure 1. (**a**) An example of FMNERG. *None* indicates that the entity *MLB* has no corresponding bounding box in the image. In the coarse-grained setting, categories including *person*, *location*, *organization* and *miscellaneous* are semantically well separated, making it easy to recognize *Francona* as *person*. By contrast, in the fine-grained setting, the model must further distinguish among semantically similar candidate subtypes such as *coach*, *athlete*, *director*, and *musician*, and correctly classify *Francona* as *coach*. (**b**) Illustration of complex entity–object relations in FMNERG. The same fine-grained entity type may correspond to diverse visual concepts on social media, such as *text*, *logo*, *poster*, and *group photo* for the type *band*, which increases the difficulty of aligning textual entities with their corresponding visual objects.

In this paper, we propose a novel Multi-level Knowledge Integration Framework (MKIF) for FMNERG, as shown in Figure 2. This framework aims to jointly improve fine-grained entity-type discrimination and robust entity–object alignment by leveraging multi-level knowledge in a unified manner. Specifically, we first utilize the global-level knowledge acquired from MLLMs to enrich the semantics of multimodal inputs and serve as a shared semantic anchor across subsequent subtasks. We further design a knowledge-enhanced span representation module with multimodal interaction to obtain the entity span embeddings. Next, to address the first issue of disambiguating fine-grained entity types, we acquire label-level knowledge by generating conceptual and discriminative descriptions for entity types via MLLMs and use both global- and label-level knowledge as supervision through contrastive learning to enhance the model's ability to discriminate between fine-grained entity types. Then, to tackle the second challenge of robust entity–object alignment, we discover potential visual concepts in the dataset as object-level knowledge and enhance the object representations by dynamically fusing global- and object-level knowledge through a learnable gating mechanism, thereby enabling robust entity–object association. We conduct experiments on the widely used *Twitter-FMNERG* dataset, and the

results demonstrate that our proposed MKIF outperforms both unimodal and multimodal state-of-the-art baselines. Our main contributions are summarized as follows:

- We propose a novel Multi-level Knowledge Integration Framework (MKIF), which incorporates multi-level knowledge to jointly improve multimodal semantic understanding, fine-grained type discrimination, and entity–object alignment for FMNERG in a unified manner. This framework offers a general paradigm for incorporating heterogeneous knowledge in FMNERG.
- We design dual-level knowledge-guided alignment and knowledge discovery-enhanced fusion to operationalize the proposed framework, where the former enhances fine-grained entity-type discrimination with dual-level knowledge-guided supervision and the latter strengthens entity grounding through knowledge-enhanced object representation.
- Extensive experiments demonstrate the superior performance of MKIF compared to both unimodal and multimodal state-of-the-art methods. Furthermore, detailed ablation studies and analyses validate the effectiveness of each component of our framework and provide key insights into its advantages.

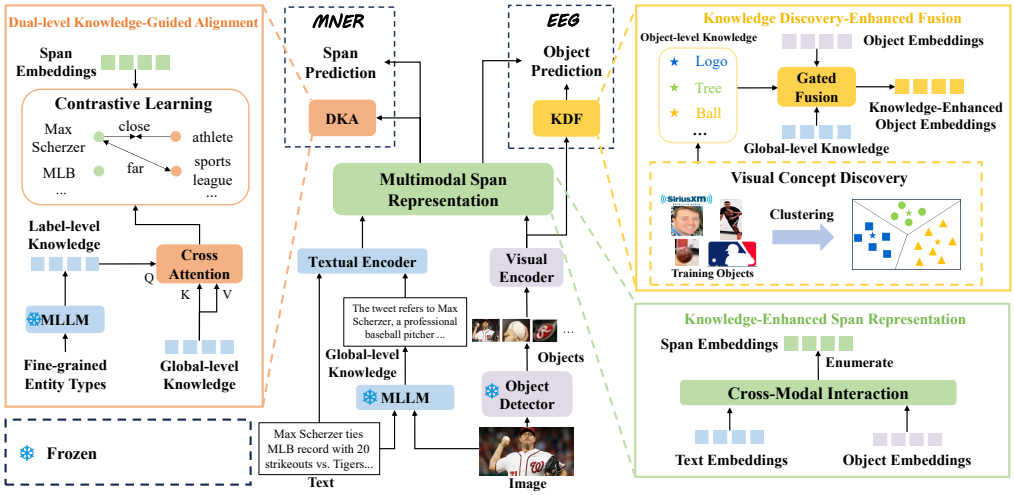


Figure 2. Overview of our proposed MKIF. MKIF first obtains entity span embeddings through the knowledge-enhanced span representation module. For MNER, label-level knowledge is incorporated through the dual-level knowledge-guided alignment (DKA) module to improve fine-grained entity-type disambiguation. For EEG, potential visual concepts are discovered as object-level knowledge. Both object- and global-level knowledge are fused with object embeddings via the knowledge discovery-enhanced fusion (KDF) module, enabling robust associations between entities and objects. The heterogeneous knowledge sources are functionally coordinated within MKIF to bridge the entity–type and entity–object gaps in FMNERG.

2. Related Works

In this section, we review representative MNER studies beginning in 2018 and FMNERG studies beginning in 2023, together with subsequent developments in both tasks.

2.1. Multimodal Named-Entity Recognition

Conventional named-entity recognition (NER) extracts entities and predicts their types based solely on textual context. However, in multimodal social media posts, short text may provide insufficient evidence for these predictions [2]. To address this limitation, MNER emerged in 2018, aiming to extract textual entities and their types with the help of images [1,2]. Since more complementary information from images is integrated, MNER has been shown to outperform unimodal NER [1–5,9,15,28]. Existing MNER methods

mainly focus on exploiting visual information and integrating it into text through modality interaction. To better utilize the image modality, prior studies have explored various forms of visual representations, including visual objects [29], object labels [4], image captions [30,31], OCR texts [31], and scene graphs [32]. For modality interaction, UMT [3] and CAT-MNER [33] propose multimodal interaction modules based on Transformer [34] to capture inter-modality interactions, showing the effectiveness of attention-based fusion for multimodal semantic understanding. UMGF [35] represents the input text and image using a multi-modal graph and proposes a graph-based multimodal fusion module to perform semantic interactions. MAF [36] proposes a cross-modal matching method to determine the proportion of image information that should be retained and exploits contrastive learning [37] to align the representations between the two modalities. However, social media posts are not always informative enough for MNER due to their short textual content and weakly relevant images [38–41]. As a result, relying solely on in-sample information may lead to suboptimal MNER performance. To tackle this challenge, knowledge-based methods acquire auxiliary external knowledge from multimodal knowledge graphs [42], knowledge corpus [43], or large language models (LLMs) [44] to enrich the semantics of raw samples. In this work, we utilize MLLMs instead of knowledge bases or unimodal LLMs to obtain high-quality external knowledge, thereby enhancing the comprehension of multimodal inputs.

2.2. Fine-Grained Multimodal Named-Entity Recognition and Grounding

FMNERG [10] extends MNER by introducing fine-grained entity types and entity-aware visual grounding. It simultaneously requires extracting entities, predicting their fine-grained types and grounded bounding boxes in the image, which makes the task more challenging than MNER and better aligned with real-world applications. Existing FMNERG methods can be generally divided into two categories. The two-stage methods [5,20] separately process MNER and entity-aware visual grounding, resulting in error propagation. Therefore, unified methods have gained attention. Yu et al. [11] and Wang et al. [10] first propose the sequence-to-sequence-based methods H-Index and TIGER, respectively. They use pre-trained sequence-to-sequence models [45,46] to encode the multimodal input and decode the indices of entities, types, and groundable or ungroundable indicators, and introduce a visual output layer to predict the distribution over candidate visual regions for entity grounding. Bao et al. [47] further propose a multi-level alignment pre-training method based on H-Index to improve congruence between multimodal data. To address the exposure bias issue of sequence-to-sequence-based methods, MQSPN [6] introduces the set prediction paradigm into FMNERG to adaptively learn intra-entity relationships and establish inter-entity relationships from a global optimal matching view. Despite their advancements, the above methods neglect to learn the discrimination of fine-grained entity types explicitly. Recently, GEM [21] exploited LLMs to obtain external knowledge and employed MLLMs to rerank the predicted fine-grained entity categories for disambiguation, followed by a large visual language model (LVLM)-based implicit grounder for EEG. However, GEM still models entity–object relations in feature space, which is insufficient for handling the complex entity–object relationships on social media. In contrast, our approach focuses on leveraging multi-level knowledge to bridge both the entity–type and entity–object gaps.

3. Methodology

In this section, we first formulate the FMNERG task and then present the details of our proposed MKIF. Figure 2 shows the architecture of MKIF, which contains three key components: (1) The *knowledge-enhanced span representation* module acquires global-level

knowledge and generates shared multimodal span representations. (2) The *dual-level knowledge-guided alignment* (DKA) module aligns the entity span representations with dual-level knowledge by contrastive learning to improve fine-grained entity-type disambiguation in MNER. (3) The *knowledge discovery-enhanced fusion* (KDF) module discovers the potential visual concepts in the dataset as object-level knowledge and integrates global- and object-level knowledge into object embeddings via a dynamic gating mechanism to facilitate robust entity–object alignment in EEG.

3.1. Problem Formulation

Given a text sequence $T = (t_1, t_2, \ldots, t_{N_t})$ of N_t tokens and its associated image I, the goal of FMNERG is to identify all named entities in T, and to predict both their fine-grained entity types from the entity-type set $\mathcal{Y}$ and their corresponding grounded bounding boxes in I. In *Twitter-FMNERG*, $\mathcal{Y}$ comprises 51 predefined fine-grained entity types organized under 8 coarse-grained categories. The detailed category hierarchy and label distribution are provided in Appendix A. The output Y is a set of triples:

$$Y = \{(e_i, c_i, b_i)\}_{i=1}^{m}, \tag{1}$$

where m denotes the number of named entities in T, e_i denotes the ith entity, $c_i \in \mathcal{Y}$ denotes the fine-grained entity type of e_i, and b_i denotes the grounded bounding box of e_i in I. If there is no grounded region of e_i, b_i is *None*, otherwise it is a 4-dimensional vector containing the top-left and bottom-right coordinates of the bounding box:

$$b_i = \begin{cases} (x_i^1, y_i^1, x_i^2, y_i^2), & \text{if } e_i \text{ is grounded,} \\ None, & \text{otherwise,} \end{cases} \tag{2}$$

where (x_i^1, y_i^1) and (x_i^2, y_i^2) denote the top-left and bottom-right coordinates of b_i, respectively.

3.2. Knowledge-Enhanced Span Representation

To enrich the semantics of multimodal inputs and enhance the model's understanding of them, we first design a knowledge-enhanced span representation module with multimodal interaction. Motivated by the impressive capabilities of MLLMs in multimodal understanding [26,27], we employ MLLMs instead of unimodal LLMs and knowledge corpora to obtain global-level knowledge for each multimodal input. Specifically, we selected GPT-4o-mini [22] as a representative MLLM for global-level knowledge acquisition, given its strong performance in multimodal understanding and cost efficiency. We input the text T and image I into GPT-4o-mini, using the instruction template "`Here is the text of a tweet:` T`, and the image of this tweet` I `is uploaded above. Please provide brief background knowledge of this tweet, tell me the named entities in the text, and tell me the visual concepts in the image briefly.`" to obtain the global-level knowledge as:

$$K = (k_1, k_2, \ldots, k_{N_k}) = \text{MLLM}(T, I), \tag{3}$$

where k_i denotes the ith knowledge token, and N_k is the number of tokens in K. We concatenate the global-level knowledge K with text T as the final textual input:

$$T_k = (t_1, t_2, \ldots, t_{N_t}, [\text{SEP}], k_1, k_2, \ldots, k_{N_k}), \tag{4}$$

where [SEP] is the separator token (taking BERT [48] for example).

We use pre-trained BERT as the text encoder. Given the textual input T_k, we obtain its contextualized representation as $\boldsymbol{T} = \text{BERT}(T_k) = (\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_{N_x})$, where N_x is the total length of $\boldsymbol{T}$ and $\boldsymbol{t}_i \in \mathbb{R}^d$ is the contextualized embedding of the ith token. For the image

modality, we follow previous works [10,11,47] to detect all visual objects of each image via the pre-trained VinVL [49] model, and rank them based on the detection probabilities, obtaining the top-N_o object candidates as $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_{N_o}) \in \mathbb{R}^{N_o \times 2048}$, where $\boldsymbol{o}_i \in \mathbb{R}^{2048}$ is the VinVL representation of the ith object. We then project $\boldsymbol{O}$ into the same dimension as $\boldsymbol{T}$ using a trainable linear layer $\boldsymbol{W}_R \in \mathbb{R}^{2048 \times d}$ as the visual encoder, obtaining $\boldsymbol{R} = \boldsymbol{O}\boldsymbol{W}_R = (\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_{N_o}) \in \mathbb{R}^{N_o \times d}$.

Next, to effectively integrate visual information with textual representations, we design a Transformer [34,50]-based multimodal interaction module for fine-grained token-object interactions in FMNERG. Specifically, we feed $\boldsymbol{T}$ and $\boldsymbol{R}$ into cross-attention transformers (CATs) to calculate the interactions, including text-guided visual attention and image-guided textual attention. In this way, the model can dynamically select complementary visual cues for each token while preserving the contextualized textual semantics. $\mathrm{CAT}(\boldsymbol{H}_x, \boldsymbol{H}_y)$ is calculated as:

$$\begin{cases} \boldsymbol{H}'_x = \mathrm{LN}\big(\mathrm{MHA}(\boldsymbol{H}_x, \boldsymbol{H}_y, \boldsymbol{H}_y) + \boldsymbol{H}_x\big), \\ \mathrm{CAT}\big(\boldsymbol{H}_x, \boldsymbol{H}_y\big) = \mathrm{LN}\big(\mathrm{FFN}(\boldsymbol{H}'_x) + \boldsymbol{H}'_x\big). \end{cases} \tag{5}$$

where MHA, LN and FFN are multi-head attention, layer normalization and feed-forward network, respectively. We calculate the text-guided visual interaction and image-guided textual interaction as $\boldsymbol{R}' = \mathrm{CAT}(\boldsymbol{R}, \boldsymbol{T})$ and $\boldsymbol{T}' = \mathrm{CAT}(\boldsymbol{T}, \boldsymbol{R})$.

We then concatenate $\boldsymbol{T}'$ and $\boldsymbol{R}'$ and feed them into a self-attention transformer (SAT). $\mathrm{SAT}(\boldsymbol{H}_x)$ is calculated as:

$$\begin{cases} \boldsymbol{H}'_x = \mathrm{LN}(\mathrm{MHA}(\boldsymbol{H}_x, \boldsymbol{H}_x, \boldsymbol{H}_x) + \boldsymbol{H}_x), \\ \mathrm{SAT}(\boldsymbol{H}_x) = \mathrm{LN}\big(\mathrm{FFN}(\boldsymbol{H}'_x) + \boldsymbol{H}'_x\big). \end{cases} \tag{6}$$

We denote the combination of the above bidirectional cross-attention and subsequent self-attention as the cross-modal interaction (CMI) module. The output of CMI is:

$$\boldsymbol{H}_{\mathrm{CMI}} = \mathrm{SAT}([\boldsymbol{T}'; \boldsymbol{R}']) \in \mathbb{R}^{(N_x + N_o) \times d}. \tag{7}$$

The resulting matrix $\boldsymbol{H}_{\mathrm{CMI}}$ jointly encodes the textual sequence and visual objects after cross-modal interaction.

Finally, we adopt a span-based prediction approach [51,52] to extract entities and their corresponding types. We enumerate all possible n_s spans (up to a maximum width of L_{max}) in the text as $\mathcal{S} = \{s_i\}_{i=1}^{n_s}$ and calculate their representations as:

$$\boldsymbol{h}_e(s_i) = \big[\boldsymbol{h}_{\mathrm{START(i)}}; \boldsymbol{h}_{\mathrm{END(i)}}; \phi_e(s_i)\big], \tag{8}$$

where $\boldsymbol{h}_{\mathrm{START(i)}}$ and $\boldsymbol{h}_{\mathrm{END(i)}}$ are the representations of the start and end tokens of s_i from the first N_t rows of $\boldsymbol{H}_{\mathrm{CMI}}$, and $\phi_e(s_i)$ is the span length embedding obtained from a learnable look-up table. Following prior span-based entity extraction methods [51,52], we formulate span typing as single-label multiclass classification and adopt the cross-entropy loss as the training objective for MNER:

$$\mathcal{L}_{span} = -\sum_{s_i \in \mathcal{S}} \log P_e(y_e^*(s_i) \mid s_i), \tag{9}$$

where $y_e^*(s_i)$ denotes the gold span label from set $\mathcal{Y}' = \mathcal{Y} \cup \{\epsilon\}$, and ϵ denotes a non-entity span.

3.3. Dual-Level Knowledge-Guided Alignment

To address the first gap of insufficient discrimination of semantically similar fine-grained entity types, we propose a dual-level knowledge-guided alignment (DKA) module.

DKA captures the conceptual and discriminative features of fine-grained entity types as label-level knowledge and then integrates them with global-level knowledge to enhance the model's discriminative capability through contrastive learning.

Inspired by Zhang et al. [53], we first generate both conceptual and discriminative descriptions for each fine-grained entity type using the MLLM. To be specific, given an FMNERG dataset containing M entity types, we query GPT-4o-mini [22] to generate two kinds of descriptions for each category, as shown in Table 1.

Table 1. Details of conceptual and discriminative descriptions for entity types.

	Conceptual	**Discriminative**
Definition	Descriptions that summarize the concept of each entity type.	Descriptions that distinguish semantically similar entity types
Prompt	Please describe both the textual and visual concepts of [class].	Please provide useful features for distinguishing the original class [class_1] from the similar class [class_2].

Note: Bold text identifies the column and row headings.

Conceptual Description. For each entity type c, we obtain one conceptual description d^c_{con} using the following query: "`Please describe both the textual and visual concept of each given word: [athlete, musician, …]`". This kind of description summarizes the concept of each entity type.

Discriminative Descriptions. For each entity type c, we define the other entity types that share the same coarse-grained category with it as its similar label set $\mathcal{S}_c$. Formally, let $C_{coarse}(c)$ denote the coarse-grained category of c. Then, $\mathcal{S}_c = \{c' \mid c' \neq c,\ C_{coarse}(c') = C_{coarse}(c)\}$. We then generate the discriminative description set $\mathcal{D}^c_{dis}$ for c using the following prompt: "`[fine_class_1] and [fine_class_2] belong to the category [coarse_class]. Please provide useful features for distinguishing the original class [fine_class_1] from the similar class [fine_class_2].`". $\mathcal{D}^c_{dis}$ contains $|\mathcal{S}_c|$ discriminative descriptions, with each description characterizing the distinguishing features between c and one of its similar types in $\mathcal{S}_c$. These discriminative descriptions provide informative cues for distinguishing semantically similar fine-grained entity types.

In this way, given an entity type c, we obtain its label-level knowledge $\mathcal{D}^c = \{d^c_{con}\} \cup \mathcal{D}^c_{dis}$, which contains one conceptual description and $|\mathcal{S}_c|$ discriminative descriptions. All descriptions are generated offline before model training. We then encode all descriptions using BERT to obtain $|\mathcal{D}^c| = 1 + |\mathcal{S}_c|$ label embeddings, denoted as $\boldsymbol{E}^c_{\mathcal{D}} \in \mathbb{R}^{|\mathcal{D}^c| \times d_l}$, where $d_l = d$ is the embedding dimension. The final label-level knowledge embedding $\boldsymbol{e}^c_{lk} \in \mathbb{R}^{d_l}$ is obtained by averaging all embeddings as:

$$\boldsymbol{e}^c_{lk} = \text{adaptive_avgpool}(\boldsymbol{E}^c_{\mathcal{D}}). \tag{10}$$

However, the obtained label embeddings mainly encode category-level general characteristics, which are independent of the input sample. Therefore, we employ cross-attention to incorporate global-level knowledge as complementary contextual information, enabling the model to enrich the label semantics with sample-related cues and obtain the dual-level knowledge embedding of c as:

$$\boldsymbol{e}_c = \text{CAT}(\boldsymbol{e}^c_{lk}, \boldsymbol{e}_{gk}), \tag{11}$$

where $\boldsymbol{e}_{gk}$ is the global-level knowledge embedding obtained from BERT. Notably, the BERT model used for knowledge encoding shares its parameters with the BERT model described in Section 3.2, and the parameters are frozen in this step.

Finally, motivated by Zhang et al. [53], we design a contrastive loss to improve fine-grained entity type discrimination under the supervision of dual-level knowledge. For each entity span s, the dual-level knowledge embedding of its gold label y_s serves as the positive embedding, while the dual-level knowledge embeddings of all the other entity types are treated as negative embeddings. Following normalized temperature-scaled contrastive formulations for representation learning [37], we define the loss as follows:

$$\mathcal{L}_{dka} = -\sum_{s \in \mathcal{S}_{gold}} \log \frac{\exp\left(\text{sim}\left(\boldsymbol{h}_e(s), \boldsymbol{e}_{y_s}\right)/\tau\right)}{\sum_{c=1}^{M} \exp(\text{sim}(\boldsymbol{h}_e(s), \boldsymbol{e}_c)/\tau)}, \tag{12}$$

where $\text{sim}(\cdot,\cdot)$ denotes the normalized cosine similarity, τ is a temperature parameter, $\mathcal{S}_{gold}$ is the set of gold entity spans in the input text, y_s is the gold fine-grained type of span s, and $\boldsymbol{e}_c$ represents the dual-level knowledge embedding for the cth entity type. Accordingly, the denominator contrasts the span representation against the embeddings of all entity types, encouraging the span to be close to the embedding of its gold label and far from those of incorrect labels, thereby facilitating fine-grained entity-type disambiguation.

3.4. Knowledge Discovery-Enhanced Fusion

To robustly model entity–object relations, we introduce a knowledge discovery-enhanced fusion (KDF) module, which discovers potential visual concepts from the dataset as object-level knowledge and adaptively integrates the object-level and global-level knowledge into object embeddings through a learnable gating mechanism, thereby facilitating robust entity–object alignment in EEG.

The first step is visual concept discovery. After detecting objects in all training images using VinVL (as described in Section 3.2), we employ the Vision Transformer (ViT) [54] encoder of the pre-trained CLIP model [55] to extract their feature embeddings. Benefiting from large-scale vision-language pre-training, CLIP captures broad visual semantics and exhibits strong transferability across diverse visual recognition tasks, enabling efficient feature extraction from a large number of heterogeneous objects in social media images. Based on the extracted embeddings, we then apply an unsupervised clustering algorithm to divide these objects into V clusters and treat the cluster centers as potential visual concept embeddings. For each object, we take the ID of its nearest cluster center as its corresponding visual concept v and introduce a learnable lookup table ϕ_v to obtain the object-level knowledge embedding $\boldsymbol{e}_{ok} = \phi_v(v)$, where the parameters of ϕ_v are initialized with the ViT features of the cluster centers.

Next, we introduce a learnable gating mechanism to adaptively control the contribution of global- and object-level knowledge to object representations as:

$$\boldsymbol{g} = \sigma(\boldsymbol{W}_h \boldsymbol{r}_i + \tilde{\boldsymbol{e}}), \tag{13}$$

where $\tilde{\boldsymbol{e}} = \boldsymbol{W}_e[\boldsymbol{e}_{gk}; \boldsymbol{e}_{ok}]$ denotes the projected knowledge embedding, $\boldsymbol{W}_h$ and $\boldsymbol{W}_e$ are learnable projection matrices, and $\sigma(\cdot)$ denotes the Sigmoid function. The final output of KDF is calculated as:

$$\boldsymbol{r}_i^k = \boldsymbol{r}_i + \boldsymbol{g} \odot \tilde{\boldsymbol{e}}, \tag{14}$$

where $\odot$ is the element-wise product operation. We denote the knowledge-enhanced object embedding set as: $\boldsymbol{R}^k = \left(\boldsymbol{r}_1^k, \boldsymbol{r}_2^k, \ldots, \boldsymbol{r}_{N_o}^k\right)$.

During EEG, for each span, we first determine whether it is groundable in the image, which is formulated as a binary classification task:

$$\mathcal{L}_{groundable} = -\sum_{s_i \in \mathcal{S}} \log P_g\left(y_g^*(s_i) \mid s_i\right), \tag{15}$$

where $y_g^*(s_i)$ denotes the gold label indicating whether s_i is groundable. For each groundable span s, we calculate its similarity distribution $p(\boldsymbol{z})$ over N_o knowledge-enhanced object embeddings as:

$$p(\boldsymbol{z}) = \text{Softmax}(\boldsymbol{R}^k \boldsymbol{h}_e(s)). \tag{16}$$

Following prior work [10], we use the Intersection over Union (IoU) scores as the supervision signals. For each visual region, we compute its IoU scores with regard to all ground-truth bounding boxes of a given entity and take the maximum value as its final IoU score. Notably, if the final IoU score of a visual region is lower than 0.5, we follow the setting in [56] and set it to 0. We take the normalized IoU score distribution $g(\boldsymbol{z})$ as the object supervision of each groundable entity. The training objective is to minimize the Kullback–Leibler Divergence (KLD) [57] loss between the similarity distribution $p(\boldsymbol{z})$ and object supervision $g(\boldsymbol{z})$, formulated as:

$$\mathcal{L}_{eeg} = \frac{1}{NG} \sum_{j=1}^{N} \sum_{k=1}^{G} g\left(\boldsymbol{z}_k^j\right) \log \frac{g\left(\boldsymbol{z}_k^j\right)}{p\left(\boldsymbol{z}_k^j\right)}, \tag{17}$$

where N is the number of samples, and G is the number of groundable entities.

3.5. Model Training and Inference

We train MKIF by jointly optimizing the following losses:

$$\mathcal{L} = \mathcal{L}_{span} + \mathcal{L}_{dka} + \mathcal{L}_{groundable} + \mathcal{L}_{eeg}. \tag{18}$$

During inference, we first extract entities and predict their types using $y_e(s) = \arg\max_{c \in \mathcal{Y}'} P_e(c \mid s)$, and we select $\mathcal{S}_{ent} = \{s : y_e(s) \neq \epsilon\}$ as the set of entity spans. Then, we predict whether the entities in $\mathcal{S}_{ent}$ are groundable by computing $y_g(s) = \arg\max_{b \in \{0,1\}} P_g(b \mid s)$. For the ith ungroundable entity s_i, we predict its triple as $(s_i, y_e(s_i), None)$. For the jth groundable entity s_j, we select the visual object with the highest probability in the similarity distribution as its predicted region $b_{pred} = \left(x^1_{pred}, y^1_{pred}, x^2_{pred}, y^2_{pred}\right)$, and its predicted triple is $(s_j, y_e(s_j), b_{pred})$.

4. Experiments

In this section, we evaluate the overall effectiveness of MKIF and the contributions of its proposed components. Following the experimental setup in Section 4.1, Section 4.2 compares MKIF with representative baselines to assess its performance in FMNERG, MNER, and EEG. Section 4.3 examines the contributions of DKA, KDF, and the individual knowledge sources through ablation studies. We also evaluate the contribution of each information request in the global-level knowledge generation prompt through ablation studies. Section 4.4 further investigates fine-grained type disambiguation, visual concept discovery, and the sensitivity of MKIF to the choice of knowledge-generating MLLM. Finally, Section 4.5 compares model predictions on representative samples selected from the test set to illustrate the effectiveness of MKIF in fine-grained type discrimination and robust entity–object alignment.

4.1. Setup

4.1.1. Dataset

We conducted experiments on the widely used *Twitter-FMNERG* dataset [10]. It contains 10,000 tweet–image pairs and 16,772 entities, including 6770 visually groundable entities associated with 8148 bounding boxes. Table 2 shows the statistics of *Twitter-FMNERG*,

while Appendix A presents the detailed entity–type hierarchy and coarse-grained category distribution.

Table 2. Statistics of *Twitter-FMNERG* dataset. "#" denotes the number of items.

Statistics	Train	Valid	Test	Total
#Tweet	7000	1500	1500	10,000
#Entity	11,779	2450	2543	16,772
#Groundable Entity	4733	991	1046	6770
#Box	5723	1171	1254	8148

4.1.2. Baselines

To evaluate the effectiveness of our method, we compared it with several representative unimodal and multimodal baselines.

First, we selected the following unimodal methods which only exploit the text modality: *HBiLSTM-CRF-None* [2], *BERT-None* [48], *BERT-CRF-None* [10] and *T5-Paraphrase-None* [10]. *None* denotes that we only detected the entity–type pairs for the MNER subtask and set the visual object prediction to the majority class *None*.

Next, we selected several competitive multimodal baselines, including *GVATT-VinVL-EVG* [2], *UMT-VinVL-EVG* [3], *UMGF-VinVL-EVG* [35], *ITA-VinVL-EVG* [31], *H-Index* [11], *MMT5-VinVL-EVG* [10], *TIGER* [10], *MQSPN* [6] and *GEM* [21]. *EVG* means that we used MNER methods to extract entity–type pairs and then predicted the bounding box for each pair via entity-aware visual grounding (EVG) proposed in [10].

4.1.3. Implementation Details

We conducted experiments on a single NVIDIA A100 GPU with PyTorch 1.10.1. In the knowledge-enhanced span representation module, we used the pre-trained bert-base-uncased model to encode the textual inputs. The maximum length of spans L_{max} was set to eight as in [52]. Following previous work [10], the number of object candidates N_o detected by VinVL was 32. The temperature parameter τ in the contrastive loss was set to 0.1, informed by [37]. These settings were kept fixed during hyperparameter selection. In the KDF module, we used the pre-trained CLIP ViT-B/32 model for object encoding, and used the k-means algorithm [58] to discover the potential visual concepts. The number of potential visual concepts V was 100.

We selected task-specific hyperparameters using validation FMNERG F1, varying one parameter group at a time while keeping the others fixed. Starting with a batch size of eight, one attention layer in each of CMI and DKA, and a maximum of 50 epochs, we sequentially evaluated $V \in \{50, 100, 200, 300, 400, 500\}$, five joint CMI/DKA attention layer configurations $\{(1,1), (2,2), (4,4), (1,2), (2,1)\}$, and batch sizes $\{4, 8, 16, 32\}$. Each configuration was compared using its best validation checkpoint. The resulting settings were $V = 100$, one attention layer in each of the CMI and DKA modules, and a batch size of 8. After selecting the final configuration, we set its training limit to 40 epochs based on the validation learning curve from a separate run using the same configuration but a different random seed. Detailed results and the learning curve are provided in Appendix E. The test set was excluded from configuration and checkpoint selection.

We used the AdamW optimizer [59] to minimize the training loss function. The learning rate was set to 2×10^{-5} for the pre-trained BERT, as evaluated in its original fine-tuning study [48], and 2×10^{-4} for the other trainable parameters, consistent with the learning rate ranges considered in [60]. Both learning rates were fixed rather than included in our hyperparameter search.

4.1.4. Metrics

For a fair comparison, we followed previous works [6,10,21] and took the F1 score as the metric to evaluate FMNERG performance. To be specific, in the MNER subtask, if the span boundary and entity type of a predicted entity were both correct, we considered it a correct prediction. In the EEG subtask, if an entity was not groundable, it was a correct prediction only if the EEG prediction was *None*. In contrast, for a groundable entity, it was a correct prediction if the IoU score between the predicted visual object and any of the ground-truth bounding boxes was greater than 0.5 [61]. This process is demonstrated as follows:

$$C_e/C_t = \begin{cases} 1, & p_e/p_t = g_e/g_t, \\ 0, & \text{otherwise}, \end{cases} \tag{19}$$

$$C_r = \begin{cases} 1, & p_r = g_r = \text{ None}, \\ 1, & \max\left(\text{IoU}_1, \ldots, \text{IoU}_j\right) > 0.5, \\ 0, & \text{otherwise}, \end{cases} \tag{20}$$

where C_e, C_t, and C_r denote whether the entity, type, and grounding result are correctly predicted, respectively. p_e, p_t, and p_r represent the predicted entity, type, and visual region, while g_e, g_t, and g_r denote their corresponding gold annotations. IoU_j is the IoU score between the predicted region p_r and the jth gold bounding box.

In this way, we can calculate the Precision (Pre), Recall (Rec) and F1 score (F1) as:

$$correct = \begin{cases} 1, & C_e \text{ and } C_t \text{ and } C_r, \\ 0, & \text{otherwise}, \end{cases} \tag{21}$$

$$Pre = \frac{\#correct}{\#predict}, \quad Rec = \frac{\#correct}{\#gold}, \quad F1 = \frac{2 \times Pre \times Rec}{Pre + Rec}, \tag{22}$$

where #*correct* is the number of predicted triples that match the gold triples, #*predict* and #*gold* are the numbers of predicted and gold triples, respectively.

4.2. Main Results

To assess the overall effectiveness of MKIF, we compared it with representative unimodal and multimodal baselines on *Twitter-FMNERG*. Table 3 reports the results on FMNERG and its two subtasks, MNER and EEG.

The main results show that multimodal approaches outperform unimodal ones, highlighting the necessity of visual information for joint entity recognition and grounding. The unified methods are generally superior to the EVG-based pipeline methods due to the alleviation of error propagation. Compared with state-of-the-art methods, MKIF achieves a relative F1 improvement of 4.99% over GEM, as well as gains of 2.76% and 2.67% on the MNER and EEG subtasks, respectively. Our method also achieves higher precision and recall than the baseline on FMNERG, MNER and EEG.

To further assess the robustness of MKIF's gains over TIGER, we evaluated both models across five random seeds. For each of FMNERG, MNER, and EEG, we applied a two-sided paired t-test to the seed-wise F1 differences, testing whether their mean was zero. We used a Bonferroni-adjusted significance threshold of $0.05/3 = 0.0167$ for the three comparisons. As shown in Figure 3, MKIF outperforms TIGER on FMNERG under every seed, with a mean gain of 5.30 F1 points (95% CI: [4.59, 6.00]; $p = 3.13 \times 10^{-5}$). The mean gains on MNER and EEG are 3.93 and 3.53 F1 points, respectively, and both are significant under the adjusted threshold $p = 1.06 \times 10^{-4}$ and $p = 1.51 \times 10^{-3}$. These results support the consistency of MKIF's performance gains across the evaluated seeds. Appendix B provides the detailed results for each seed and the full statistical report.

Table 3. Comparison of MKIF with baselines on *Twitter-FMNERG*. "-" indicates that the corresponding results are not reported in the original paper.

Modality	Methods	FMNERG			MNER			EEG		
		Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
Text	HBiLSTM-CRF-None	34.86	32.38	33.57	62.31	56.55	59.29	49.25	43.27	46.07
	BERT-None	33.28	34.28	33.77	58.91	60.05	59.47	46.27	47.65	46.94
	BERT-CRF-None	34.41	35.51	34.95	60.06	61.38	60.72	46.93	48.44	47.67
	T5-Paraphrase-None	37.38	37.29	37.33	64.83	65.32	65.07	49.03	48.91	48.97
Text + Image	GVATT-VinVL-EVG	42.02	38.75	40.32	63.08	57.85	60.35	55.27	53.46	54.35
	UMT-VinVL-EVG	40.67	41.99	41.32	61.24	62.01	61.63	53.58	55.32	54.43
	UMGF-VinVL-EVG	41.73	42.11	41.92	61.68	61.90	61.79	54.51	55.00	54.75
	ITA-VinVL-EVG	43.05	42.51	42.78	63.80	62.64	63.21	57.63	56.90	57.26
	H-Index	46.83	46.28	46.55	65.25	64.45	64.84	60.82	60.10	60.46
	MMT5-VinVL-EVG	45.35	45.08	45.21	66.46	66.77	66.61	58.35	58.01	58.18
	TIGER	47.57	46.85	47.20	64.43	65.40	64.91	62.44	61.49	61.96
	MQSPN	-	-	48.57	-	-	67.09	-	-	62.64
	GEM	-	-	50.54	-	-	68.09	-	-	63.59
	Ours	**53.97**	**52.18**	**53.06**	**71.13**	**68.86**	**69.97**	**66.41**	**64.22**	**65.29**

Note: Bold and underlined values indicate the best and second-best results, respectively.

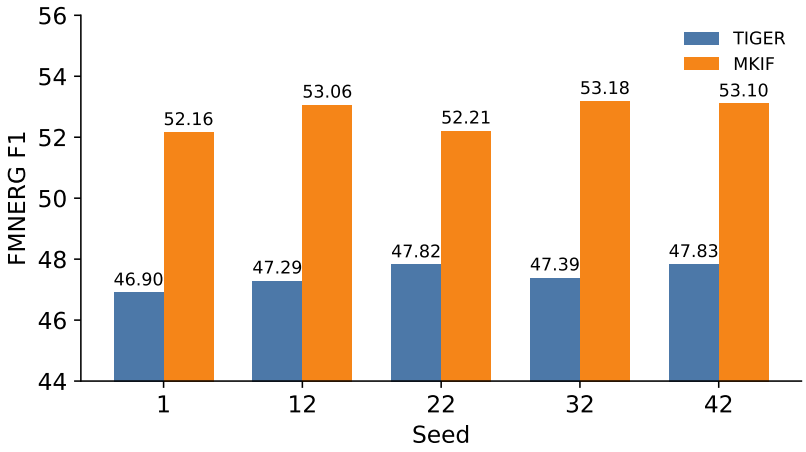


Figure 3. FMNERG F1 scores of MKIF and TIGER across five matched random seeds.

We further evaluated our method on the grounded multimodal named entity recognition (GMNER) [11] task, which is the coarse-grained version of FMNERG and includes only four entity types in the MNER subtask (i.e., *person*, *location*, *organization* and *miscellaneous*). Note that we removed DKA as its role in enhancing fine-grained entity-type discrimination is not required in this setting. We supplemented the evaluation with two additional GMNER baselines, MNER-QG [62] and Bao et al.'s method [47], and report the results on *Twitter-GMNER* in Table 4. From the table, we can observe that MKIF achieves the highest F1 scores on both GMNER and EEG, and the second-highest F1 on MNER, further confirming its effectiveness.

All these results demonstrate the effectiveness of MKIF. We attribute its superior performance over the baselines to the following reasons: (1) MKIF acquires high-quality global-level knowledge from the MLLM to enrich the semantics of raw samples, enabling a better understanding of multimodal inputs than non-knowledge-based methods such as TIGER and MQSPN. (2) MKIF leverages both global- and label-level knowledge to explicitly address the insufficient disambiguation of semantically similar entity types, leading to substantially improved MNER performance. Moreover, our method does not require additional instruction tuning of MLLM as in GEM, which improves computational efficiency. (3) By

discovering visual concepts and performing gated fusion between knowledge and objects, MKIF effectively captures complex entity–object relations, outperforming the baselines that capture such relationships only in the feature space during EEG.

Table 4. Results on *Twitter-GMNER*. "-" indicates that the corresponding results are not reported in the original paper.

	GMNER			MNER			EEG		
Methods	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**
MNER-QG	53.02	54.84	53.91	78.16	78.59	78.37	58.48	56.59	57.52
H-Index	56.16	56.67	56.41	79.37	80.10	79.73	60.90	61.46	61.18
TIGER	55.84	57.45	56.63	79.88	80.70	80.28	60.72	61.81	61.26
Bao et al.	57.93	58.56	58.24	79.65	80.55	80.10	-	-	-
MQSPN	59.03	58.49	58.76	81.22	79.66	80.43	61.86	62.94	62.40
GEM	-	-	<u>59.83</u>	-	-	**83.15**	-	-	<u>63.16</u>
Ours	**62.29**	**61.58**	**61.93**	**82.22**	**81.38**	<u>81.80</u>	**65.91**	**65.16**	**65.53**

Note: Bold and underlined values indicate the best and second-best results, respectively.

Additionally, we report the F1 scores of each method on coarse-grained entity types in Table 5. The results show that for coarse-grained categories *person*, *building*, *product* and *art*, our method achieves F1 improvements of 14.80%, 2.61%, 2.99% and 6.94% over MQSPN, respectively. MKIF also achieves the second-best performance on *location*, *organization* and *other*. The results on *Twitter-FMNERG* and *Twitter-GMNER* support the effectiveness and generalizability of MKIF across fine- and coarse-grained entity-type settings.

Table 5. F1 scores of our methods and baselines on FMNERG with respect to each coarse-grained entity type.

Entity Type	**H-Index**	**TIGER**	**MQSPN**	**Ours**
Person	45.13	43.78	<u>48.64</u>	**55.84**
Location	62.33	67.69	**71.03**	<u>68.75</u>
Building	32.88	40.00	<u>44.78</u>	**45.95**
Organization	46.68	46.75	**49.03**	<u>47.33</u>
Product	28.19	27.38	<u>29.46</u>	**30.34**
Art	38.89	<u>43.27</u>	42.84	**45.92**
Event	45.56	<u>48.39</u>	**51.88**	48.28
Other	41.81	**48.28**	43.49	<u>43.80</u>
All	46.55	47.20	<u>48.57</u>	**53.06**

Note: Bold and underlined values indicate the best and second-best results, respectively. Italics indicate entity type labels.

To complement the task performance, we also compare the online inference efficiency of MKIF and TIGER. The measurement setup and detailed results for inference latency, FLOPs, and peak GPU memory are provided in Appendix F.

4.3. Ablation Study

In this section, we examine the contributions of the proposed components in MKIF and the information requested in the global-level knowledge generation prompt through ablation experiments.

First, we compared the full model with variants that removed DKA, KDF, or individual knowledge sources to assess their respective contributions to performance.

The results are presented in Table 6, which includes the following settings: (1) We removed DKA (w/o DKA). We further assessed the contribution of each component within DKA by individually removing global-level knowledge (w/o Glob.), conceptual descriptions (w/o Conc.), and discriminative descriptions (w/o Disc.). (2) We removed KDF during EEG (w/o KDF) and further removed the global-level knowledge (w/o Glob.) and object-level knowledge (w/o Obj.) individually during fusion. (3) We removed both DKA and KDF (w/o DKA and KDF).

Table 6. Ablation study of MKIF.

Methods	FMNERG			MNER			EEG		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
MKIF	**53.97**	**52.18**	**53.06**	71.13	68.86	**69.97**	**66.41**	**64.22**	**65.29**
w/o DKA	52.79	51.00	51.88 (↓ 1.18)	70.45	68.15	69.28 (↓ 0.69)	65.41	63.19	64.28 (↓ 1.01)
w/o Glob.	52.64	51.71	52.17	70.14	68.98	69.55	65.25	64.10	64.67
w/o Conc.	53.75	51.00	52.34	**71.41**	67.84	69.57	66.10	62.72	64.37
w/o Disc.	52.59	51.87	52.23	70.02	**69.13**	69.57	64.87	63.98	64.42
w/o KDF	52.46	51.51	51.98 (↓ 1.08)	70.08	68.90	69.49 (↓ 0.48)	65.16	63.98	64.56 (↓ 0.73)
w/o Glob.	53.07	51.95	52.50	70.43	69.02	69.72	65.49	64.10	64.79
w/o Obj.	53.27	51.87	52.56	70.44	68.66	69.54	65.63	63.90	64.75
w/o DKA & KDF	51.73	51.20	51.46 (↓ 1.60)	69.41	68.78	69.09 (↓ 0.88)	64.12	63.47	63.79 (↓ 1.50)

Note: Bold values indicate the best results. ↓ indicates the decrease in F1 points relative to the full MKIF model.

The detailed discussions are as follows:

- **Effectiveness of DKA**. As demonstrated in Table 6, removing DKA leads to a 2.22% relative decrease in F1 on FMNERG, indicating its contribution to distinguishing semantically similar fine-grained categories. The results also validate that global-level knowledge helps filter out sample-irrelevant information. Conceptual descriptions provide category-level semantics, while discriminative descriptions further highlight inter-class differences. Such complementary information provides more effective supervision for fine-grained type disambiguation.
- **Effectiveness of KDF**. As presented in Table 6, aligning the entities only with raw object embeddings results in a 2.04% relative decrease in F1 on FMNERG. Moreover, adding either global-level or object-level knowledge individually improves performance. This shows that global-level knowledge provides sample-specific contextual cues and object-level knowledge captures concept-level visual semantics. Meanwhile, the gated fusion mechanism enables the model to adaptively integrate the two knowledge sources into object representations, thereby learning complex entity–object relations robustly.

Second, we conducted an ablation study to evaluate how removing specific information requests from the global-level knowledge generation prompt for GPT-4o-mini affected the performance of MKIF. Specifically, based on the full prompt, we constructed three ablated variants by removing one information requirement at a time, including *background knowledge*, *named entities*, and *visual concepts*. The F1 scores of FMNERG under different prompt templates are reported in Table 7.

As shown in Table 7, the full prompt yields the highest F1 score, while all ablated versions still achieve strong performance. The results suggest that the effectiveness of MKIF is not sensitive to a particular prompt template, and MKIF can consistently benefit from MLLM-generated knowledge under different prompt formulations, verifying the robustness of our approach. Meanwhile, removing background knowledge, named entities, or visual concepts each causes a performance drop to different extents, which indicates that the

three queried information types are complementary rather than redundant in global-level knowledge construction. In particular, removing background knowledge or visual concepts from the instruction leads to a significant performance decline, indicating the importance of sample-level contextual semantics, and demonstrating that the image-side concepts provide useful cross-modal cues for enhancing entity–object alignment. By contrast, removing the named-entity-related instruction results in a relatively smaller performance drop, indicating that MKIF can effectively capture entity-related information from the original textual input and span representations, rather than relying on entity mentions provided by the MLLM.

Table 7. Ablation study of different prompt templates for obtaining global-level knowledge.

Prompt Templates for Obtaining Global-level Knowledge	**F1**
Please provide brief background knowledge of this tweet, tell me the named entities in the text, and tell me the visual concepts in the image briefly. (***full prompt***)	**53.06**
Please tell me the named entities in the text, and tell me the visual concepts in the image briefly. (***w/o background knowledge***)	50.64
Please provide brief background knowledge of this tweet, and tell me the visual concepts in the image briefly. (***w/o named entities***)	<u>52.05</u>
Please provide brief background knowledge of this tweet, and tell me the named entities in the text. (***w/o visual concepts***)	50.71
TIGER (***non-knowledge-based baseline***)	47.20

Note: Bold and underlined F1 scores indicate the best and second-best results, respectively. Bold italic text highlights the prompt settings and the baseline.

4.4. Analysis and Discussion

To better understand the effectiveness of MKIF, we evaluated the contribution of DKA to distinguishing semantically similar entity types, the semantic characteristics of the discovered visual concepts and the sensitivity of MKIF to the number of the visual concepts, and the performance of MKIF with different MLLMs as knowledge sources. These aspects were investigated through the following research questions.

▶**RQ1**: *Does DKA improve the disambiguation of semantically similar fine-grained entity types in MNER?*

A: To provide an intuitive illustration of DKA's contribution, we visualize several test instances of both MKIF and MKIF w/o DKA on the MNER subtask in Figure 4.

In case #1, the entity *LuverneMN* belongs to the *city* category, but the model without DKA incorrectly recognizes it as a *sports team*. In contrast, by leveraging the conceptual description, the full model captures the concept that a sports team is "a group of players competing in sports, with visuals including uniforms, team logos, and group huddles", which is inconsistent with the visual evidence. Consequently, the full model decreases the prediction probability of class *sports team* and correctly recognizes the entity as a *city*. In case #2, removing DKA results in ambiguous predictions between *athlete* and *artist* when extracting the entity *Gwen*, although the input image suggests an *athlete*. This occurs because the ablated model fails to capture the discrimination that athletes are characterized by physical fitness and competition, whereas artists focus on creative expression and artistic techniques. These observations indicate that DKA plays a crucial role in mitigating ambiguous predictions by exploiting dual-level knowledge supervision through contrastive learning.

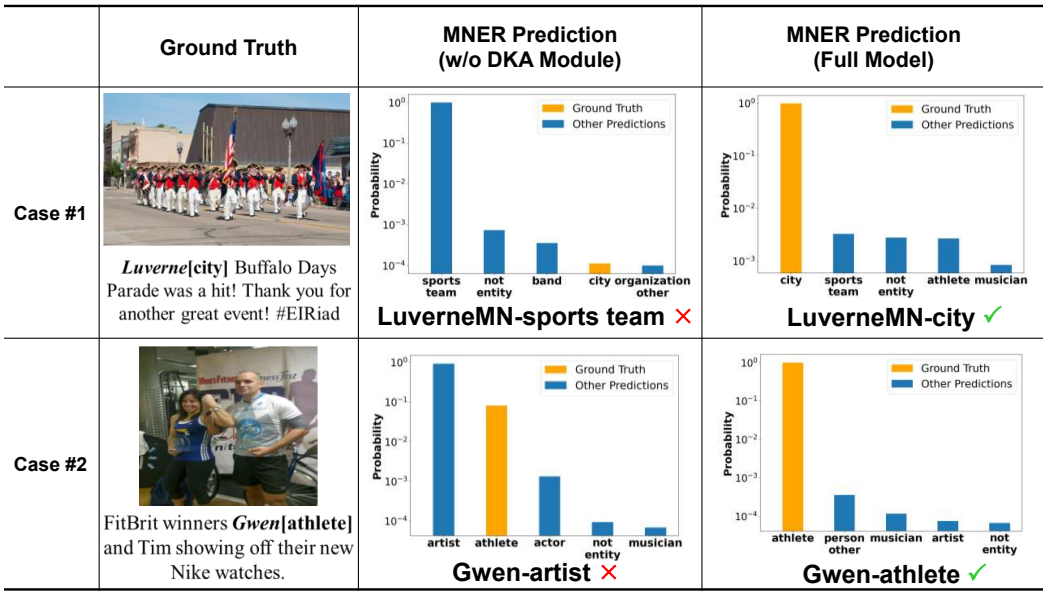


Figure 4. MNER predictions of MKIF and its ablated version. We visualize the predicted probability distributions for several test samples, showing the top-5 classes. In case #1, the conceptual descriptions enable the model to infer that the tweet content is inconsistent with *sports team*, thereby correctly predicting the entity as *city*. In case #2, the discriminative descriptions assist the model in distinguishing the semantically similar entity types *athlete* and *artist*.

▶**RQ2**: *How does visual concept discovery in KDF contribute to the effectiveness of MKIF?*

A: We first illustrate representative visual concepts discovered from the training set. As shown in Figure 5, unsupervised clustering identifies coherent visual concepts, which provide complementary semantic cues for enriching object representations during gated fusion.

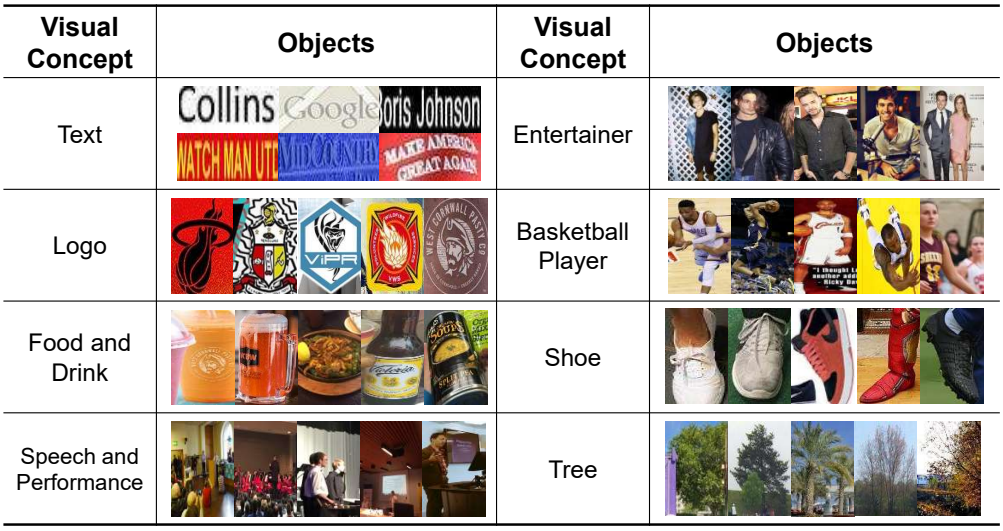


Figure 5. Representative visual concepts discovered from the training set via unsupervised clustering.

Next, we varied the number of visual concepts V from 50 to 500 and report the performance of MKIF under different settings in Figure 6. From the results, we can observe that MKIF achieves the best performance when $V = 100$, while F1 decreases when V becomes higher or lower. This can be explained as follows: when V is too small, the potential visual concepts in the dataset cannot be adequately discovered, leading to semantic inconsistency among objects within the same cluster; but when V is too large, the excessively fine-grained

clustering may split semantically similar objects into different clusters, introducing redundant information.

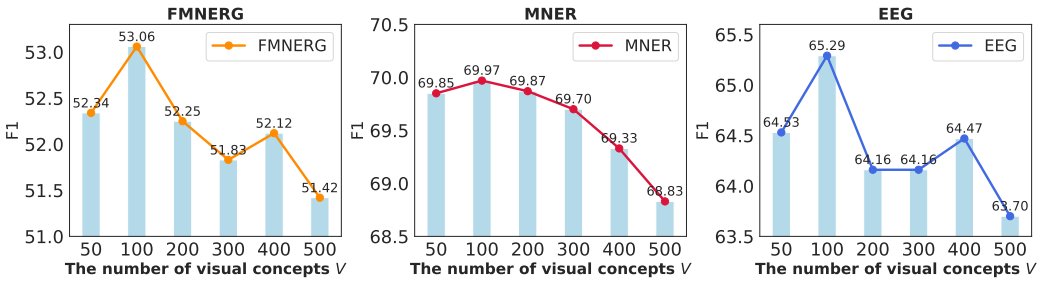


Figure 6. The performance of MKIF under different numbers of visual concepts V discovered from the dataset.

To complement our analysis of the number of visual concepts, we further analyzed the sensitivity of MKIF to different clustering algorithms in Appendix D.

▶**RQ3**: *How does the choice of MLLM affect the effectiveness of MKIF?*

A: We selected two representative open-source MLLMs as alternative sources of both global- and label-level knowledge. LLaVA-1.5-7B [24] is a widely used baseline for visual instruction tuning with a simple and reproducible design, while InternVL-2.5-8B [63,64] represents more recent advances through dynamic-resolution image processing and training on diverse vision-language tasks. As shown in Table 8, we compared all knowledge-based variants with the non-knowledge-based method TIGER.

Table 8. Performance comparison with different MLLMs in global- and label-level knowledge generation.

Model	FMNERG			MNER			EEG		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
TIGER	47.57	46.85	47.20	64.43	65.40	64.91	62.44	61.49	61.96
Ours (LLaVA-1.5-7B)	50.89	49.63	50.25	68.75	67.13	67.93	63.95	62.37	63.15
Ours (InternVL-2.5-8B)	52.21	51.12	51.66	69.04	67.68	68.35	65.14	63.78	64.46
Ours (GPT-4o-mini)	**53.97**	**52.18**	**53.06**	**71.13**	**68.86**	**69.97**	**66.41**	**64.22**	**65.29**

Note: Bold values indicate the best results.

The results show that all knowledge-based variants consistently outperform TIGER, demonstrating the effectiveness of incorporating global- and label-level knowledge. Although the performance varies with the adopted MLLM, MKIF remains effective across different knowledge sources, indicating that its gains are not tightly coupled to a specific MLLM. Compared with GPT-4o-mini, LLaVA-1.5-7B and InternVL-2.5-8B reduce FMNERG F1 by 2.81 and 1.40 percentage points, respectively. These performance gaps reflect differences in the relevance and semantic richness of global-level knowledge for each input image–text pair, as well as differences in the conceptual and discriminative quality of label-level knowledge. Therefore, we use GPT-4o-mini as the default knowledge source in our final model.

More capable MLLMs may further improve the quality of generated knowledge. However, as shown in Table 2, *Twitter-FMNERG* contains 10,000 image–text pairs, and generating sample-specific global-level knowledge across this dataset makes API cost an important consideration in model selection. GPT-4o-mini therefore offers a practical balance between the performance demonstrated above and knowledge-generation cost. Moreover, the design of MKIF allows more capable MLLMs to serve as sources of both global- and label-level knowledge without requiring architectural changes.

4.5. Case Study

To gain deeper insights into the superiority of MKIF in distinguishing semantically similar fine-grained entity types and robust entity–object alignment, we compare its predictions with those of TIGER on representative test samples in Figure 7.

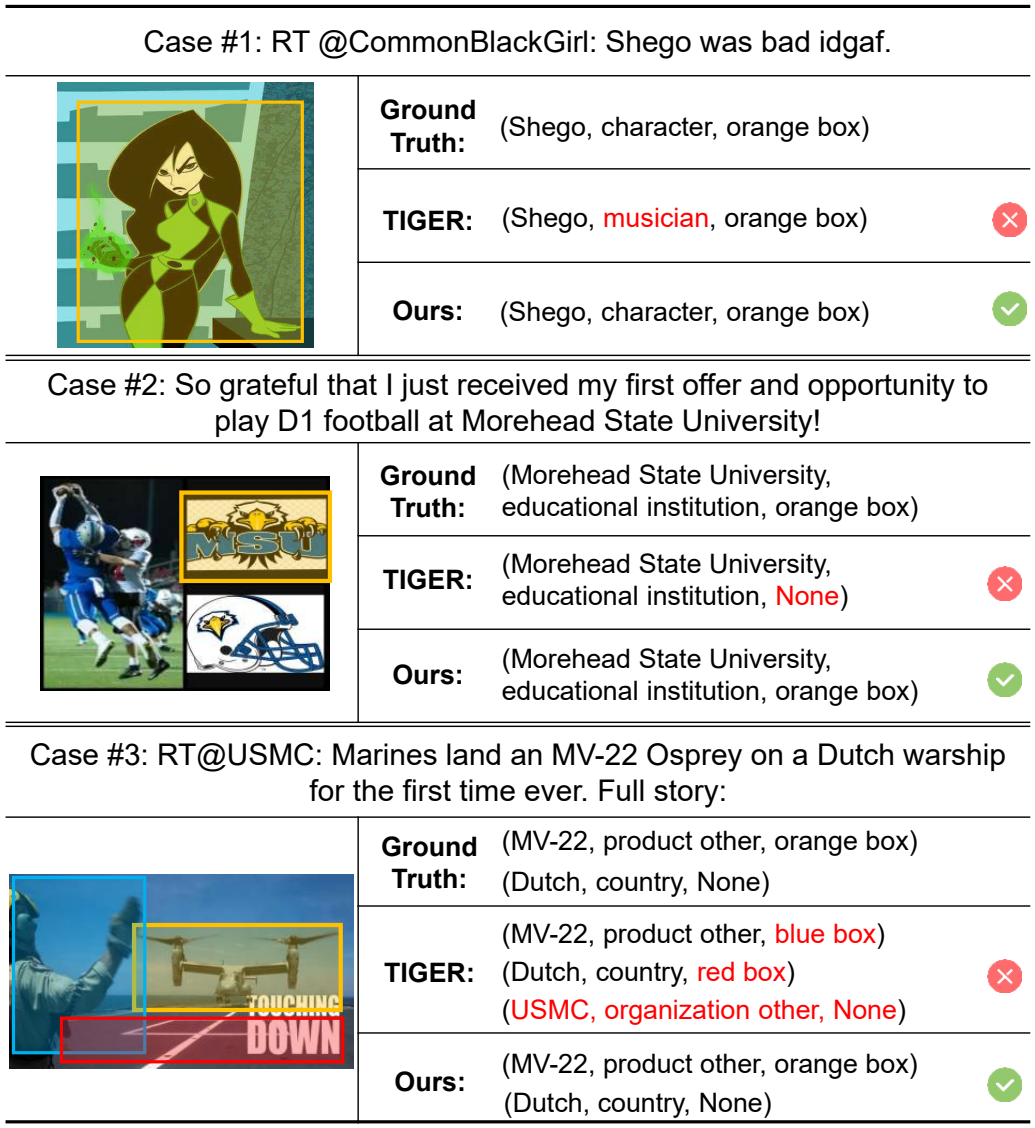


Figure 7. Cases selected from the test set of *Twitter-FMNERG*. The wrong predictions are highlighted in red.

In case #1, both MKIF and TIGER correctly detect the corresponding region of entity *Shego* in the image, but compared to TIGER, our model correctly recognizes it as a character rather than a musician. Such improvement can be attributed to DKA, which captures the distinction that characters are fictional figures in stories, animations or games, while musicians are involved in creating and performing music. Consequently, our model correctly extracts the triple from a tweet about cartoons. In case #2, although TIGER correctly recognizes the entity *Morehead State University* and its type, it fails to link the entity to its corresponding visual object. Moreover, in the more challenging case #3, TIGER incorrectly matches *MV-22* with a human and links the country *Dutch* to plain text. In contrast, our method successfully captures the implicit relation between the entity in the *educational institution* category and the visual concept *logo*, as well as the correlation between *product* and

transportation. Therefore, MKIF accurately detects the regions of *Morehead State University* and *MV-22* in the images.

These cases illustrate that MKIF is more effective than TIGER in resolving semantic ambiguities among fine-grained entity types and robustly capturing complex entity–object relations, thereby addressing the key challenges of FMNERG.

5. Conclusions

In this paper, we proposed MKIF, a novel framework that leverages multi-level knowledge to jointly address two key challenges in FMNERG: distinguishing semantically similar fine-grained entity types and robustly aligning textual entities with visual objects. We first acquired global-level knowledge and designed a knowledge-enhanced span representation module to generate knowledge-enriched entity span embeddings. To further improve the discrimination of fine-grained entity types, we proposed a dual-level knowledge-guided alignment (DKA) module, which exploits both global- and label-level knowledge as supervision via contrastive learning. Moreover, we proposed a knowledge discovery-enhanced fusion (KDF) module that integrates object-level knowledge with global-level knowledge into object embeddings through a dynamic gating mechanism, enabling robust modeling of entity–object relations. Comprehensive experiments demonstrated that MKIF consistently outperformed state-of-the-art methods, and detailed analyses validated the effectiveness of its core components. In future work, we plan to equip our framework with continual learning capabilities, enabling better adaptation to the dynamically evolving content on social media, and to explore privacy-aware FMNERG pipelines that protect the privacy of multimodal content while maintaining downstream performance.

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

Funding: This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences under Grant XDA0480301.

Data Availability Statement: The datasets used in this study are publicly available from the official repositories at https://github.com/NUSTM/FMNERG (accessed on 2 November 2026) and https://github.com/NUSTM/GMNER (accessed on 2 November 2026). Materials related to this work are available upon reasonable request.

Conflicts of Interest: The authors declare that they have no competing interests relevant to the content of this article.

Appendix A. Fine-Grained Entity Types and Distributions in the *Twitter-FMNERG* Dataset

Table A1 presents the hierarchy and distribution of the entity types in the *Twitter-FMNERG* dataset. The dataset contains 51 fine-grained entity types organized under eight coarse-grained categories.

Table A1. Fine-grained entity types and distributions of the *Twitter-FMNERG* dataset.

Coarse-Grained Category	Fine-Grained Entity Types	Train	Valid	Test	Total	Proportion
Person	Politician, musician, actor, artist, athlete, author, businessman, character, coach, director, intellectual, journalist, person other	5019	1072	1104	7195	42.90%
Location	City, country, state, continent, park, road, location other	1553	345	327	2225	13.27%
Building	Cultural place, entertainment place, sports facility, building other	365	62	77	504	3.01%
Organization	Company, educational institution, band, government agency, news agency, political party, social organization, sports league, sports team, organization other	3035	595	638	4268	25.45%
Product	Brand name products, game, software, product other	355	82	88	525	3.13%
Art	Film and television works, magazine, music, written work, art other	495	103	106	704	4.20%
Event	Festival, sports event, event other	614	126	129	869	5.18%
Other	Animal, award, medical thing, website, ordinance	343	65	74	482	2.87%
All	**51 fine-grained entity types**	**11,779**	**2450**	**2543**	**16,772**	**100%**

Note: Bold text highlights the totals and improves readability.

Appendix B. Statistical Significance and Consistency of MKIF's Performance Gains across Random Seeds

To assess whether the improvements of MKIF over baselines (e.g., TIGER) remained consistent across multiple random seeds, we evaluated both methods on *Twitter-FMNERG* using five matched seeds (1, 12, 22, 32, 42). Table A2 reports the precision, recall, and F1 scores for FMNERG and the MNER and EEG subtasks in each run. MKIF achieves a higher F1 score than TIGER under every seed for all three tasks.

Table A2. Results of TIGER and MKIF across five random seeds on *Twitter-FMNERG*.

		FMNERG			MNER			EEG		
Method	**Seed**	**Pre.**	**Rec.**	**F1**	**Pre.**	**Rec.**	**F1**	**Pre.**	**Rec.**	**F1**
TIGER	1	47.33	46.47	46.90	64.62	65.80	65.20	61.68	60.56	61.12
	12	47.57	47.01	47.29	64.88	65.60	65.24	62.63	61.90	62.26
	22	48.28	47.36	47.82	65.26	66.47	65.86	62.91	61.71	62.31
	32	47.45	47.34	47.39	65.51	65.64	65.57	61.76	61.62	61.69
	42	48.00	47.66	47.83	65.45	65.92	65.68	61.61	61.17	61.39
MKIF	1	53.54	50.85	52.16	70.89	67.40	69.10	66.63	63.27	64.91
	12	53.97	52.18	53.06	71.13	68.86	69.97	66.41	64.22	65.29
	22	52.98	51.48	52.21	70.26	68.35	69.29	65.28	63.43	64.34
	32	53.64	52.73	53.18	70.36	69.25	69.80	66.48	65.36	65.91
	42	54.05	52.18	53.10	70.22	67.87	69.03	67.17	64.85	65.99

For each task and seed i, let d_i denote the F1 difference between MKIF and TIGER. We used a two-sided paired t-test to examine the null hypothesis $H_0 : \mathbb{E}[d_i] = 0$ against

$H_1 : \mathbb{E}[d_i] \neq 0$. Paired observations yielded four degrees of freedom ($df = 4$). We report the mean and sample standard deviation of each method's F1 scores, the mean paired difference, the test statistic, the p-value, and the 95% confidence interval for the mean difference in Table A3.

Table A3. Paired t-test of MKIF and TIGER across five seeds on *Twitter-FMNERG*.

Task	TIGER F1	MKIF F1	Mean Difference	$t(4)$	p	95% CI
FMNERG	47.45 ± 0.39	52.74 ± 0.51	+5.30	20.84	3.13×10^{-5}	[4.59, 6.00]
MNER	65.51 ± 0.28	69.44 ± 0.42	+3.93	15.31	1.06×10^{-4}	[3.22, 4.64]
EEG	61.75 ± 0.53	65.29 ± 0.69	+3.53	7.72	1.51×10^{-3}	[2.26, 4.81]

The paired comparisons indicated consistent F1 gains for MKIF over TIGER on all three tasks. All three p-values remained below the Bonferroni-adjusted significance threshold of $0.05/3 = 0.0167$, supporting the robustness of these gains across the evaluated random seeds.

Appendix C. Trainable Parameter Analysis

This appendix reports the trainable parameter counts of MKIF and TIGER and provides a component-wise breakdown of MKIF. Table A4 compares the trainable parameter counts of the two models.

Table A4. Trainable parameter counts of MKIF and TIGER. M denotes 10^6 parameters.

Model	Trainable Parameters (M)
TIGER	224.48
Ours	**152.14**

Note: Bold text highlights the model with fewer trainable parameters.

MKIF contains 152.14M trainable parameters, compared with 224.48M for TIGER, corresponding to approximately 72.34M fewer parameters (32.23%). This difference primarily reflects the backbone architectures: MKIF uses a BERT encoder with task-specific modules, whereas TIGER retains the encoder–decoder architecture of T5-base. Together with the results in Table 3, this comparison shows that the performance gains of MKIF over TIGER are achieved with fewer downstream trainable parameters, demonstrating a more efficient trade-off between task performance and parameter efficiency.

Beyond the overall parameter comparison, Table A5 provides a component-wise breakdown of the trainable parameters in MKIF. The text encoder accounts for approximately 71.57% of these parameters, while the remaining components together contain approximately 43.25M parameters. "Other" denotes the parameters associated with span representation and task prediction.

Table A5. Trainable parameter counts of MKIF by component. M denotes 10^6 parameters.

Component	Trainable Parameters (M)
Textual encoder (BERT-base)	108.89
Cross-modal interaction (CMI)	12.41
Dual-level knowledge-guided alignment (DKA)	7.67
Knowledge discovery-enhanced fusion (KDF)	18.72
Other	4.45
Total	**152.14**

Note: Bold text highlights the total parameter count.

Appendix D. Sensitivity of MKIF to Different Clustering Algorithms in KDF

To examine the sensitivity of MKIF to the choice of clustering algorithm, we compared *k*-means, spherical *k*-means [65], and a Gaussian mixture model (GMM) [66] for visual concept discovery. All three variants used the same object-feature extraction procedure and 100 clusters or mixture components, and the resulting visual concepts provided object-level knowledge for KDF. We evaluated their performance on *Twitter-FMNERG* and compared them with TIGER, which does not explicitly enrich object representations with discovered visual concepts for EEG. The results are reported in Table A6.

Table A6. Performance comparison of MKIF with different clustering algorithms in KDF and TIGER on *Twitter-FMNERG*.

Model	FMNERG			MNER			EEG		
	Pre	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**
TIGER	47.57	46.85	47.20	64.43	65.40	64.91	62.44	61.49	61.96
Ours (GMM)	50.83	49.51	50.16	69.84	68.11	68.97	63.95	62.29	63.11
Ours (spherical *k*-means)	51.94	50.53	51.23	70.25	68.43	69.33	64.55	62.80	63.66
Ours (*k*-means)	**53.97**	**52.18**	**53.06**	**71.13**	**68.86**	**69.97**	**66.41**	**64.22**	**65.29**

Note: Bold values indicate the best results.

Table A6 shows that *k*-means achieves the highest scores across all reported metrics, followed by spherical *k*-means and GMM. Compared with spherical *k*-means, *k*-means improves FMNERG, MNER, and EEG F1 by 1.83, 0.64, and 1.63 percentage points, respectively. The corresponding improvements over GMM are 2.90, 1.00, and 2.18 points. The larger differences in EEG than in MNER are consistent with the role of the discovered concepts in enriching object representations for grounding. All three MKIF variants achieve higher F1 scores than TIGER on all three tasks, supporting the effectiveness of MKIF under different choices of visual concept discovery. These results are also consistent with the role of KDF in exploiting the discovered visual concepts to enrich object representations for EEG. The performance differences may reflect the distinct clustering objectives and the resulting visual concept assignments. Among the evaluated methods, *k*-means achieves the strongest results and is therefore adopted as the default clustering algorithm in KDF.

Appendix E. Hyperparameter Selection

This appendix reports the validation results supporting the hyperparameter selection procedure described in Section 4.1.3. We varied one parameter group at a time in the order of visual concept count, joint CMI/DKA attention layer configuration, and batch size, while holding the remaining settings fixed. Each candidate was trained for up to 50 epochs, and its score was taken from the checkpoint with the highest validation FMNERG F1. The test set was not used for these comparisons.

Tables A7–A9 show that $V = 100$, one attention layer in each of CMI and DKA, and a batch size of eight yield the highest validation FMNERG F1 in their respective comparisons. This sequential search did not evaluate every combination of the candidate values.

To determine the final training duration, we examine the validation learning curve from a separate run of the selected configuration with a different random seed in Figure A1, as learning curves can inform early stopping and model selection [67]. The training loss decreases throughout training, whereas validation FMNERG F1 reaches its maximum at epoch 38 and then fluctuates around a similar level, with no sustained improvement between epochs 40 and 50. We therefore use 40 epochs as the maximum training duration for the

final configuration and select the checkpoint with the highest validation FMNERG F1 within that duration.

Table A7. FMNERG precision, recall, and F1 on the validation set with different numbers of visual concepts V.

V	Pre	Rec	F1
50	55.31	52.08	53.65
100	54.88	53.71	**54.29**
200	54.17	53.06	53.61
300	52.95	52.69	52.82
400	54.47	52.73	53.59
500	53.72	52.73	53.22

Note: Bold text indicates the highest F1 score and the corresponding hyperparameter configuration.

Table A8. FMNERG precision, recall, and F1 on the validation set with different CMI/DKA attention layer configurations.

CMI Attention Layers	DKA Attention Layers	Pre	Rec	F1
1	**1**	54.88	53.71	**54.29**
2	2	53.79	53.63	53.71
4	4	54.33	52.98	53.65
1	2	53.49	53.43	53.46
2	1	53.79	53.59	53.69

Note: Bold text indicates the highest F1 score and the corresponding hyperparameter configuration.

Table A9. FMNERG precision, recall, and F1 on the validation set with different batch sizes.

Batch size	Pre	Rec	F1
4	54.15	52.49	53.31
8	54.88	53.71	**54.29**
16	54.60	53.55	54.07
32	50.60	52.04	51.34

Note: Bold text indicates the highest F1 score and the corresponding hyperparameter configuration.

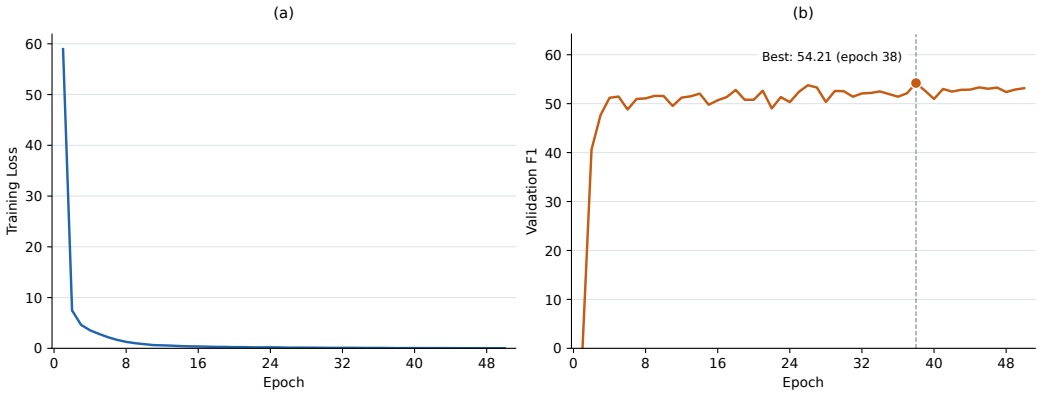


Figure A1. (**a**) Training loss and (**b**) validation FMNERG F1 over 50 epochs for a separate run of the selected configuration using a different random seed. The highest validation F1 is 54.21 at epoch 38.

Appendix F. Computational Efficiency

To assess MKIF's inference efficiency relative to the representative FMNERG baseline TIGER, we compared the online inference efficiency of MKIF and TIGER on the same

128 test samples using an NVIDIA A100 GPU. Inference latency and peak GPU memory were measured with a batch size of eight, after 10 warm-up batches and over three repetitions. FLOPs were estimated with the PyTorch profiler on the same eight samples at a batch size of one. The results are reported in Table A10.

Table A10. Inference efficiency comparison between TIGER and MKIF.

Model	Latency (ms/Sample) ↓	FLOPs (G) ↓	Peak GPU Memory (MB) ↓
TIGER	131.01	54.72	1125.69
MKIF	12.81	57.38	802.25

Table A10 shows that MKIF has lower inference latency and peak GPU memory usage than TIGER, despite its slightly higher FLOP count. This latency difference is consistent with their different prediction procedures. TIGER autoregressively generates entity–type–object triples with a T5-based encoder–decoder, whereas MKIF predicts spans largely in parallel. Together with the FMNERG performance in Table 3, these results further demonstrate that MKIF combines strong task performance with practical online inference efficiency.

For clarity, the measurements cover only the online inference stage. The global- and label-level knowledge generated by the MLLM, the object candidates detected by VinVL, the object features extracted by CLIP for visual concept discovery, and the discovered visual concepts are prepared offline before model training and inference. Their preprocessing costs are not included in the reported inference latency, FLOPs, or peak GPU memory.

References

1. Zhang, Q.; Fu, J.; Liu, X.; Huang, X. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Palo Alto, CA, USA, 2018; pp. 5674–5681. [CrossRef]
2. Lu, D.; Neves, L.; Carvalho, V.; Zhang, N.; Ji, H. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 1990–1999. [CrossRef]
3. Yu, J.; Jiang, J.; Yang, L.; Xia, R. Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 3342–3352. [CrossRef]
4. Xu, B.; Huang, S.; Du, M.; Wang, H.; Song, H.; Xiao, Y.; Lin, X. A unified visual prompt tuning framework with mixture-of-experts for multimodal information extraction. In *Proceedings of the International Conference on Database Systems for Advanced Applications*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 544–554. [CrossRef]
5. Li, J.; Li, H.; Sun, D.; Wang, J.; Zhang, W.; Wang, Z.; Pan, G. LLMs as Bridges: Reformulating Grounded Multimodal Named Entity Recognition. In *Proceedings of the Findings of the Association for Computational Linguistics ACL 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 1302–1318. [CrossRef]
6. Tang, J.; Wang, Z.; Gong, Z.; Yu, J.; Zhu, X.; Yin, J. Multi-Grained Query-Guided Set Prediction Network for Grounded Multimodal Named Entity Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Palo Alto, CA, USA, 2025; pp. 25246–25254. [CrossRef]
7. Lin, X.; Zhang, Y.; Xu, Y.; Huang, K.; Mu, H.; Wang, Y.; Gou, G.; Qian, L.; Peng, L.; Liu, W.; et al. MAKAR: A Multi-Agent framework based Knowledge-Augmented Reasoning for Grounded Multimodal Named Entity Recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 6110–6130. [CrossRef]
8. Yuan, L.; Cai, Y.; Zhu, B.; Liu, Z.; Deng, Z.; Li, Q.; Wang, T. Visual Knowledge-Enhanced LLaVA for Fine-Grained Multimodal Named Entity Recognition and Grounding. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2026**, *34*, 781–795. [CrossRef]
9. Zhang, M.; Ning, J.; Wu, X.; Lin, H.; Zhang, Y. E2E-GMNER: End-to-End Generative Grounded Multimodal Named Entity Recognition. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2026*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2026; pp. 22460–22469. [CrossRef]
10. Wang, J.; Li, Z.; Yu, J.; Yang, L.; Xia, R. Fine-grained multimodal named entity recognition and grounding with a generative framework. In *Proceedings of the 31st ACM International Conference on Multimedia*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 3934–3943. [CrossRef]

11. Yu, J.; Li, Z.; Wang, J.; Xia, R. Grounded multimodal named entity recognition on social media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 9141–9154. [CrossRef]
12. Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J.M.; Parikh, D.; Batra, D. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2017; pp. 326–335. [CrossRef]
13. Fukui, A.; Park, D.H.; Yang, D.; Rohrbach, A.; Darrell, T.; Rohrbach, M. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 457–468. [CrossRef]
14. Song, W.; Fu, C.; Zheng, Y.; Zhang, Y.; Chen, J.; Wang, P. Batch image encryption using cross image permutation and diffusion. *J. Inf. Secur. Appl.* **2024**, *80*, 103686. [CrossRef]
15. Tang, J.; Yang, Y.; Yu, J.; Wang, Z.X.; Liang, H.; Yao, L.; Yin, J. UnCo: Uncertainty-Driven Collaborative Framework of Large and Small Models for Grounded Multimodal NER. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 7633–7651. [CrossRef]
16. Zhang, H.; Liang, X.; Cai, W.; Chen, P.; Chen, B.; Jiang, B.; Wang, Y. Bridging demonstration and multi-source attention fusion using LLMs for grounded multimodal named entity recognition. *Inf. Fusion* **2026**, *127*, 103800. [CrossRef]
17. Zuo, X.; Liang, H.; Jing, N.; Zeng, S.; Fang, Z.; Luo, Y. Type-enriched Hierarchical Contrastive Strategy for Fine-Grained Entity Typing. In *Proceedings of the 29th International Conference on Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 2405–2417.
18. Vempala, A.; Preotiuc-Pietro, D. Categorizing and Inferring the Relationship between the Text and Image of Twitter Posts. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 2830–2840. [CrossRef]
19. Xu, N.; Wang, J.; Tian, Y.; Zhang, R.; Mao, W. AnANet: Association and Alignment Network for Modeling Implicit Relevance in Cross-Modal Correlation Classification. *IEEE Trans. Multimed.* **2023**, *25*, 7867–7880. [CrossRef]
20. Ok, H.; Kil, T.; Seo, S.; Lee, J. SCANNER: Knowledge-Enhanced Approach for Robust Multi-modal Named Entity Recognition of Unseen Entities. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 7718–7730. [CrossRef]
21. Wang, Z.; Zhu, C.; Zheng, Z.; Li, X.; Xu, T.; He, Y.; Liu, Q.; Yu, Y.; Chen, E. Granular Entity Mapper: Advancing Fine-grained Multimodal Named Entity Recognition and Grounding. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 3211–3226. [CrossRef]
22. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774. [CrossRef]
23. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2023; pp. 34892–34916.
24. Liu, H.; Li, C.; Li, Y.; Lee, Y.J. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2024; pp. 26296–26306. [CrossRef]
25. Bai, S.; Cai, Y.; Chen, R.; Chen, K.; Chen, X.; Cheng, Z.; Deng, L.; Ding, W.; Gao, C.; Ge, C.; et al. Qwen3-VL Technical Report. *arXiv* **2025**, arXiv:2511.21631. [CrossRef]
26. Zhang, D.; Yu, Y.; Dong, J.; Li, C.; Su, D.; Chu, C.; Yu, D. MM-LLMs: Recent Advances in MultiModal Large Language Models. In *Proceedings of the Findings of the Association for Computational Linguistics ACL 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 12401–12430. [CrossRef]
27. Ma, X.; Zhang, Y.; Fu, Y. Towards Unified Multimodal Large Language Models: A survey. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2026*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2026; pp. 37212–37230. [CrossRef]
28. Moon, S.; Neves, L.; Carvalho, V. Multimodal Named Entity Recognition for Short Social Media Posts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 852–860. [CrossRef]
29. Wu, Z.; Zheng, C.; Cai, Y.; Chen, J.; Leung, H.f.; Li, Q. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In *Proceedings of the 28th ACM International Conference on Multimedia*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 1038–1046. [CrossRef]
30. Chen, S.; Aguilar, G.; Neves, L.; Solorio, T. Can images help recognize entities? A study of the role of images for Multimodal NER. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 87–96. [CrossRef]
31. Wang, X.; Gui, M.; Jiang, Y.; Jia, Z.; Bach, N.; Wang, T.; Huang, Z.; Tu, K. ITA: Image-Text Alignments for Multi-Modal Named Entity Recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 3176–3189. [CrossRef]

32. Wang, J.; Yang, Y.; Liu, K.; Zhu, Z.; Liu, X. M3S: Scene graph driven multi-granularity multi-task learning for multi-modal NER. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *31*, 111–120. [CrossRef]
33. Wang, X.; Ye, J.; Li, Z.; Tian, J.; Jiang, Y.; Yan, M.; Zhang, J.; Xiao, Y. CAT-MNER: Multimodal named entity recognition with knowledge-refined cross-modal attention. In *Proceedings of the 2022 IEEE International Conference on Multimedia and Expo*; IEEE: New York, NY, USA, 2022; pp. 1–6. [CrossRef]
34. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 5998–6008.
35. Zhang, D.; Wei, S.; Li, S.; Wu, H.; Zhu, Q.; Zhou, G. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Palo Alto, CA, USA, 2021; pp. 14347–14355. [CrossRef]
36. Xu, B.; Huang, S.; Sha, C.; Wang, H. MAF: A General Matching and Alignment Framework for Multimodal Named Entity Recognition. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2022; pp. 1215–1223. [CrossRef]
37. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*; JMLR: New York, NY, USA, 2020; pp. 1597–1607.
38. Sun, L.; Wang, J.; Su, Y.; Weng, F.; Sun, Y.; Zheng, Z.; Chen, Y. RIVA: A Pre-trained Tweet Multimodal Model Based on Text-image Relation for Multimodal NER. In *Proceedings of the 28th International Conference on Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 1852–1862. [CrossRef]
39. Sun, L.; Wang, J.; Zhang, K.; Su, Y.; Weng, F. RpBERT: A text-image relation propagation-based BERT model for multimodal NER. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Palo Alto, CA, USA, 2021; pp. 13860–13868. [CrossRef]
40. Xu, B.; Huang, S.; Du, M.; Wang, H.; Song, H.; Sha, C.; Xiao, Y. Different data, different modalities! Reinforced data splitting for effective multimodal information extraction from social media posts. In *Proceedings of the 29th International Conference on Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 1855–1864.
41. Wu, S.; Fei, H.; Cao, Y.; Bing, L.; Chua, T.S. Information Screening whilst Exploiting! Multimodal Relation Extraction with Feature Denoising and Multimodal Topic Modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 14734–14751. [CrossRef]
42. Chen, D.; Li, Z.; Gu, B.; Chen, Z. Multimodal named entity recognition with image attributes and image knowledge. In *Proceedings of the Database Systems for Advanced Applications: 26th International Conference*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 186–201. [CrossRef]
43. Wang, X.; Cai, J.; Jiang, Y.; Xie, P.; Tu, K.; Lu, W. Named Entity and Relation Extraction with Multi-Modal Retrieval. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 5925–5936. [CrossRef]
44. Li, J.; Li, H.; Pan, Z.; Sun, D.; Wang, J.; Zhang, W.; Pan, G. Prompting ChatGPT in MNER: Enhanced Multimodal Named Entity Recognition with Auxiliary Refined Knowledge. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 2787–2802. [CrossRef]
45. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 7871–7880. [CrossRef]
46. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 140.
47. Bao, X.; Tian, M.; Wang, L.; Zha, Z.; Qin, B. Contrastive pre-training with multi-level alignment for grounded multimodal named entity recognition. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 795–803. [CrossRef]
48. 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*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186. [CrossRef]
49. Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; Gao, J. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2021; pp. 5579–5588. [CrossRef]
50. Li, K.; Li, J.; Guo, D.; Yang, X.; Wang, M. Transformer-based visual grounding with cross-modality interaction. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 1–19. [CrossRef]

51. Fu, J.; Huang, X.J.; Liu, P. SpanNER: Named Entity Re-/Recognition as Span Prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 7183–7195. [CrossRef]
52. Zhong, Z.; Chen, D. A Frustratingly Easy Approach for Entity and Relation Extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 50–61. [CrossRef]
53. Zhang, W.; Wan, C.; Liu, T.; Tian, X.; Shen, X.; Ye, J. Enhanced motion-text alignment for image-to-video transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2024; pp. 18504–18515. [CrossRef]
54. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the 9th International Conference on Learning Representations, Virtual, 3–7 May 2021.
55. 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*; JMLR: New York, NY, USA, 2021; pp. 8748–8763.
56. Yu, Z.; Yu, J.; Xiang, C.; Zhao, Z.; Tian, Q.; Tao, D. Rethinking diversified and discriminative proposal generation for visual grounding. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*; IJCAI: Sydney, Australia, 2018; pp. 1114–1120. [CrossRef]
57. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
58. Forgy, E. Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. *Biometrics* **1965**, *21*, 768–769.
59. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In *Proceedings of the 7th International Conference on Learning Representations*; International Society of the Learning Sciences: Boulder, CO, USA, 2019.
60. Joshi, M.; Chen, D.; Liu, Y.; Weld, D.S.; Zettlemoyer, L.; Levy, O. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 64–77. [CrossRef]
61. Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A.L.; Murphy, K. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2016; pp. 11–20. [CrossRef]
62. Jia, M.; Shen, L.; Shen, X.; Liao, L.; Chen, M.; He, X.; Chen, Z.; Li, J. MNER-QG: An end-to-end MRC framework for multimodal named entity recognition with query grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Palo Alto, CA, USA, 2023; pp. 8032–8040. [CrossRef]
63. Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2024; pp. 24185–24198. [CrossRef]
64. Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. *arXiv* **2024**, arXiv:2412.05271. [CrossRef]
65. Dhillon, I.S.; Modha, D.S. Concept Decompositions for Large Sparse Text Data Using Clustering. *Mach. Learn.* **2001**, *42*, 143–175. [CrossRef]
66. McLachlan, G.J.; Peel, D. *Finite Mixture Models*; Wiley Series in Probability and Statistics; Wiley: Hoboken, NJ, USA, 2000. [CrossRef]
67. Mohr, F.; van Rijn, J.N. Learning curves for decision making in supervised machine learning: A survey. *Mach. Learn.* **2024**, *113*, 8371–8425. [CrossRef]

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.