

Multi-Grained Query-Guided Set Prediction Network for Grounded Multimodal Named Entity Recognition

Jielong Tang¹, Zhenxing Wang⁴, Ziyang Gong², Jianxing Yu^{1,5}, Xiangwei Zhu³, and Jian Yin^{1*}

¹School of Artificial Intelligence, Sun Yat-sen University, Zhuhai, China

²School of Atmospheric Sciences, Sun Yat-sen University, Zhuhai, China

³School of Electronics and Communication Engineering, Sun Yat-sen University, Guangzhou, China

⁴State Key Laboratory of Intelligent Game, Institute of Software, Chinese Academy of Sciences, Beijing, China

⁵Pazhou Lab, Guangzhou, 510330, China

{tangjlong3, gongzy23}@mail2.sysu.edu.cn, {issjyin, yujx26, zhuxw666}@mail.sysu.edu.cn, wangzhenxing@iscas.ac.cn

Abstract

Grounded Multimodal Named Entity Recognition (GMNER) is an emerging information extraction (IE) task, aiming to simultaneously extract entity spans, types, and corresponding visual regions of entities from given sentence-image pairs data. Recent unified methods employing machine reading comprehension or sequence generation-based frameworks show limitations in this difficult task. The former, utilizing human-designed type queries, struggles to differentiate ambiguous entities, such as *Jordan (Person)* and *off-White x Jordan (Shoes)*. The latter, following the one-by-one decoding order, suffers from exposure bias issues. We maintain that these works misunderstand the relationships of multimodal entities. To tackle these, we propose a novel unified framework named **Multi-grained Query-guided Set Prediction Network (MQSPN)** to learn appropriate relationships at intra-entity and inter-entity levels. Specifically, MQSPN explicitly aligns textual entities with visual regions by employing a set of learnable queries to strengthen intra-entity connections. Based on distinct intra-entity modeling, MQSPN reformulates GMNER as a set prediction, guiding models to establish appropriate inter-entity relationships from a optimal global matching perspective. Additionally, we incorporate a query-guided Fusion Net (QFNet) as a glue network to boost better alignment of two-level relationships. Extensive experiments demonstrate that our approach achieves state-of-the-art performances in widely used benchmarks.

Code — <https://github.com/tangjielong928/mqspn>

Introduction

To effectively comprehend and manage vast amounts of multimodal content from social media, recent research (Yu et al. 2023a) proposes a nascent multimodal information extraction task named Grounded Multimodal Named Entity Recognition (GMNER). It aims to extract multimodal entity information, including entity spans, types, and corresponding regions of entities, from image-text pairs. Prior studies decompose the GMNER task into several subtasks such as

Multimodal Named Entity Recognition (Zhang et al. 2021a) and Visual Grounding (Redmon and Farhadi 2018), adopting the pipeline approach to solve it, which leads to serious error propagation. To address this problem, recent research paradigm has transferred to detect span-type-region triplets with *unified* model by formulating GMNER as machine reading comprehension (MRC) (Jia et al. 2023) or sequence generation (Yu et al. 2023a; Wang et al. 2023).

Despite their remarkable performance, limitations still exist. The MRC-based frameworks utilize human-designed type-specific queries as prior instructions to simultaneously guide entity recognition and entity grounding, struggling to distinguish different ambiguous entities. For example in Figure 1 (a), with the input of multiple fixed *person* queries like “Please extract person: People’s name...”, the model incorrectly detects *off-White x Jordan (Shoes)* as *Jordan (Person)* and assigns the wrong region originally belonging to other entity (*Kevin Durant*). On the other hand, sequence generation-based methods suffer from exposure bias. They autoregressively decode span-type-region triples one by one in predefined sequence order, resulting in the prediction of *off-White x Jordan* region highly sensitive to errors in preceding *Kevin Durant* detection in Figure 1 (b).

In our perspective, the essence of these errors is the inappropriate modeling of multimodal entity relationships. (1) Manually building specific query for each entity is labor-intensive and impractical. Existing MRC-based frameworks overlook the distinctions of intra-entity connection due to their reliance on duplicate and untrainable type queries that cannot distinguish intrinsic features of ambiguous entities. (2) Sequence generation-based methods excessively rely on inter-entity relationships between different multimodal entities elements, where the current output is vulnerable to previous predictions. Therefore, we propose a novel unified framework named **MQSPN**, mainly consisting of learnable Multi-grained Query Set (MQS) and Multimodal Set Prediction Network (MSP) to focus on modeling appropriate relationships at intra-entity and inter-entity levels.

We maintain that modeling intra-entity connections is foundational; only when a model is capable of distinguishing individual entities can it further model inter-entity re-

*Corresponding author.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

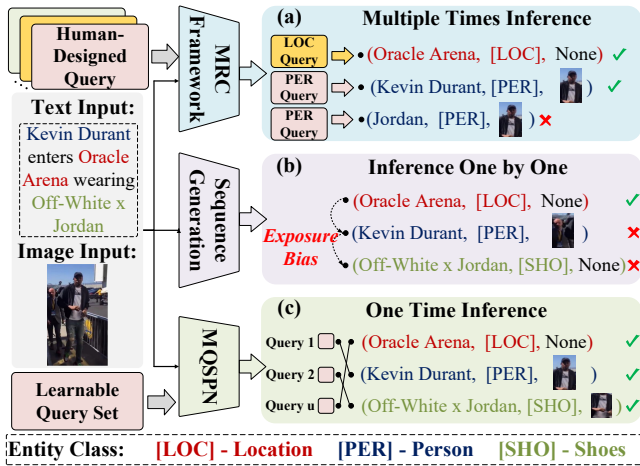


Figure 1: The comparison of existing approaches and our MQSPN. (a) The MRC-based methods trap in entity-ambiguous issue due to intra-entity misunderstanding. (b) The sequence generation-based methods suffer from exposure bias issue due to inter-entity overreliance. (c) Our MQSPN model appropriate two-level relationships with learnable query set and set prediction.

relationships effectively. Hence, we first propose MQS to adaptively learn intra-entity connections. Instead of human-designed queries in the MRC-based framework, MQS adopts a set of learnable queries (Li et al. 2023; Gong et al. 2024) (denoted as *entity-grained query*) to perform joint learning in span localization, region grounding, and entity classification for different entities, which enforces queries to learn distinguishable features and automatically establish explicit intra-entity connections. However, simply learnable queries are insufficient to detect regions and spans due to the lack of semantics information. To tackle this, we feed a prompt with masked type statement into vanilla *BERT* to build *type-grained queries* with type-specific semantics. Finally, each multi-grained query is constructed by integrating a learnable entity-grained query with a type-grained query.

Based on distinct intra-entity modeling, we further apply MSP to explore suitable inter-entity relationships. Different from previous sequence generation-based methods, MSP reformulates GMNER as *set predictions* (Tan et al. 2021; Shen et al. 2022). As shown in Figure 1 (c), with one-time input of learnable MQS, MSP parallelly predicts a set of multimodal entities in the non-autoregressive manner without the need for a preceding sequence. The training objective of MSP is to find the optimal bipartite matching with minimal global matching cost, which can be efficiently solved by the off-the-shelf Hungarian Algorithm (Kuhn 1955). In this manner, the inference of MSP will not depend on redundant dependencies dictated by a predefined decoding order, thereby guiding models to establish suitable inter-entity relationships from a global matching perspective.

Besides, since directly fusing textual features with irrelevant visual features will impair model performance (Chen et al. 2021), we further propose a QFNet between MQS and

MSP to filter this noisy information, thereby boosting better alignment of two-level relationships. Unlike direct fusion methods (Zhang et al. 2021b; Wu et al. 2023), QFNet employs queries as intermediaries to facilitate the separate integration of textual and visual region representations. Our contributions could be summarized as follows:

- We delve into the essence of existing unified GMNER methods’ weaknesses from a new perspective, two-level relationships (intra-entity and inter-entity), and propose a unified framework **MQSPN** to adaptively learn intra-entity relationships and establish inter-entity relationships from global optimal matching view.
- To the best of our knowledge, our MSP is the first attempt to apply the set prediction paradigm to the GMNER task.
- Extensive experiments on two Twitter benchmarks illustrate that our method outperforms existing state-of-the-art (SOTA) methods. The ablation study also validates the effectiveness of each designed module.

Related Work

Grounded Multimodal Named Entity Recognition. Previous multimodal named entity recognition (MNER) models (Yu et al. 2020; Chen et al. 2022a; Zhang et al. 2021a) primarily focused on how to utilize visual information to assist textual models in entity extraction. Grounded MNER (GMNER) (Yu et al. 2023a) is proposed to additionally output the bounding box coordinates of named entities within the image, which has great potential in various downstream tasks, such as knowledge base construction (Liu et al. 2019) and QA systems (Yu et al. 2021; Yu, Zha, and Yin 2019; Yu et al. 2023b). The taxonomy of previous GMNER works encompasses two branches, namely pipeline manner and unified manner. Pipeline methods (Li et al. 2024a; Ok et al. 2024) decompose the GMNER task into several subtasks like MNER, entity entailment and entity grounding. To tackle the error propagation issue, unified methods formulate GMNER as an end-to-end machine reading comprehension (Jia et al. 2023, 2022) or sequence generation task (Wang et al. 2023; Yu et al. 2023a; Li et al. 2024b). Different from them, our MQSPN reformulate GMNER as set prediction to learn appropriate intra-entity and inter-entity relationships.

Set Prediction. Set prediction is a well-known machine learning technique where the goal is to predict an unordered set of elements. It is widely used in various applications such as object detection (Carion et al. 2020). In the field of NER, Seq2Set (Tan et al. 2021) first reformulates the nested NER task as set prediction to eliminate order bias from the autoregressive model. Subsequently, PIQN (Shen et al. 2022) associates each entity with a learnable instance query to capture the location and type semantics. Recently, DQSetGen (Chen et al. 2024) extended set prediction to other sequence labeling tasks such as slot filling and part-of-speech tagging. Different from these methods, we introduce the MQSPN, which exploits the set prediction paradigm in a new GMNER area to model the two-level relationship of multimodal entities.

Visual Grounding. Visual Grounding (VG) aims to detect the most relevant visual region based on a natural language query, i.e., phrase, category, or sentence. Most ex-

(a) Overview of MQSPN

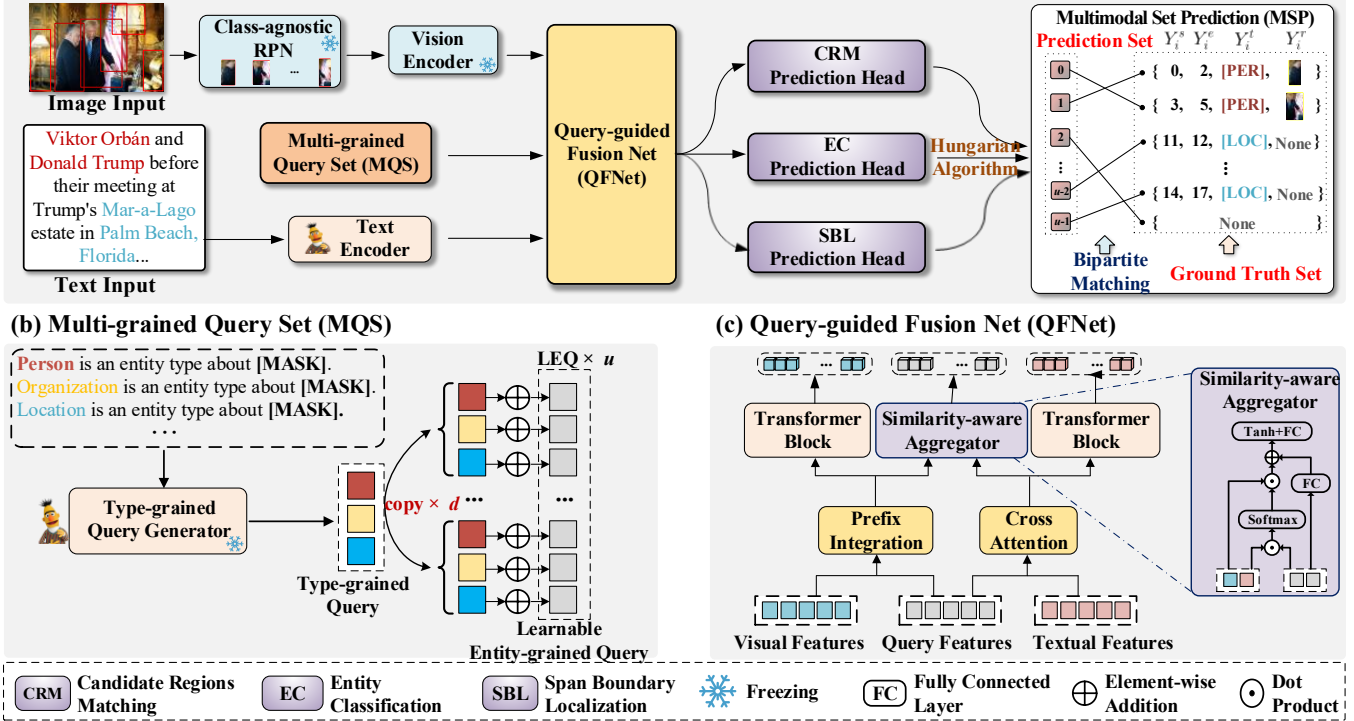


Figure 2: (a). Overview of our MQSPN. (b). The construction of Multi-grained Query Set (MQS) consists of Type-grained Queries and Learnable Entity-grained Queries (LEQ). (c). The detailed architecture of Query-guided Fusion Net (QFNet).

isting works can be divided into two branches. The first branch utilizes the one-stage detector, such as DETR (Carion et al. 2020) and YOLO (Redmon and Farhadi 2018), to directly output object bounding boxes in an end-to-end manner. The second branch first generates candidate regions with some region proposal methods, such as Region Proposal Network (RPN) (Girshick 2015) and selective search (Uijlings et al. 2013), and then selects the best-matching region based on language query. In this work, we follow the two-stage paradigm to construct an entity grounding model.

Our Method

Overview

As illustrated in Figure 2, we present a set prediction-based method named MQSPN with four different components, consisting of the Feature Extraction Module, Multi-grained Query Set (MQS), Query-guided Fusion Net (QFNet) and Multimodal Set Prediction Network (MSP). The objective of our MQSPN is to predict a set of multimodal quadruples which can be represented as:

$$Y = \{(Y_1^s, Y_1^e, Y_1^t, Y_1^r), \dots, (Y_m^s, Y_m^e, Y_m^t, Y_m^r)\}, \quad (1)$$

where $(Y_i^s, Y_i^e, Y_i^t, Y_i^r)$ denote the i -th quadruple, $Y_i^s \in [0, n-1]$ and $Y_i^e \in [0, n-1]$ are the start and end boundary indices of the i -th target entity span. Y_i^t refers to its corresponding entity type, and Y_i^r denotes grounded region. Note that if the target entity cannot be grounded in the given image, Y_i^r is *None*; otherwise, Y_i^r consists of a 4-D spatial

feature including the top-left and bottom-right coordinates of the grounded region.

Feature Extraction Module

Text Representation. Given the input sentence X , textual encoder BERT (Devlin et al. 2019) is used to tokenize it into a sequence of word embeddings $H_T = ([CLS], e_1, \dots, e_n, [SEP])$, where $e_i \in \mathbb{R}^h$, h is the hidden dimension, $[CLS]$ and $[SEP]$ are special tokens of the beginning and end positions in word embeddings.

Visual Representation. Given the input image I , we utilize VinVL (Zhang et al. 2021c) as the class-agnostic region proposal network (RPN) to obtain all candidate regions. Following the work of (Yu et al. 2023a), we also retain the top- k region proposals as our candidate regions, using the ViT-B/32 (Dosovitskiy et al. 2020) from pre-training CLIP (Radford et al. 2021) as vision encoder. The initial visual representation for the candidate regions are denoted as $\mathbf{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_k\}$. To match those entities that are ungroundable in the images, we construct a special visual token embedding $\mathbf{v}_{[ug]}$ by feeding a blank image \mathcal{T} into vision encoder. Finally, $\mathbf{v}_{[ug]}$ and \mathbf{V} are concatenated to serve as the final visual representation $H_V \in \mathbb{R}^{(k+1) \times h}$, where k is the number of candidate regions.

Multi-grained Query Set Construction

Previous manually constructed query statements struggle to learn distinguishable features for different entities, hindering

intra-entity connections modeling. In this section, we propose a learnable multi-grained query set to overcome this.

Prompt-based Type-grained Query Generator. The Entity type can provide effective information for entity span extraction and candidate region matching. We designed a prompt template: *Prompt = [TYPE] is an entity type about [MASK]*, where *[TYPE]* refers to the entity type name, such as *Person, Location, Organization, and Others*. Then the prompt template is fed into a vanilla BERT model. The type-grained query embedding is calculated as the output embedding of the *[MASK]* position:

$$H_Q^o = \text{BERT}(\text{Prompt})_{[\text{MASK}]} \quad (2)$$

where $H_Q^o \in \mathbb{R}^{p \times h}$ is the type-grained query embedding, p denotes the number of entity types.

Learnable Entity-grained Query Set. Entity-grained queries are randomly initialized as learnable input embeddings $H_Q^e \in \mathbb{R}^{u \times h}$. During training, the entity-grained semantics and corresponding relationships between candidate regions and entity spans can be learned automatically by these embeddings. To ensure that type-grained and entity-grained query embeddings have the same dimensions, we replicated the former d times. Then the multi-grained query embedding is given by:

$$H_Q = H_Q^e \oplus [H_Q^o]^d \quad (3)$$

where $H_Q \in \mathbb{R}^{u \times h}$ refers to the multi-grained queries set, $p < u$, and $d = u/p$. u is the number of queries, we use the token-wise addition operation \oplus to fuse the multi-grained queries. $[\cdot]^d$ denotes repeating d times.

Query-guided Fusion Net

Previous multimodal fusion approaches suffered from direct fusion of textual and visual information due to semantic discrepancies and noisy information between multimodal data. Different from them, we use queries as intermediaries to guide the integration of textual representations and visual region representations respectively. This module includes three interaction mechanisms.

Query-text Cross-attention Interaction. As shown in Figure 2 (b), the query set H_Q and the textual sequence H_T are fed into the transformer-based architecture (Vaswani et al. 2017). The cross-attention mechanism is used to fuse these unimodal features.

Query-region Prefix Integration. Following prefix tuning (Li and Liang 2021) and its successful applications (Chen et al. 2022a,b) in multimodal fusion, we propose query-region prefix integration to reduce the semantic bias between multimodal data. Details of the Query-region Prefix Integration are provided in the Appendix.

Similarity-aware Aggregator. To mitigate the noise caused by misaligned candidate regions and entity spans, we propose a similarity-aware aggregator to learn fine-grained token-wise alignment between query tokens and regional features/textual tokens. We denote the query set, textual, and visual features as \tilde{H}_Q , \tilde{H}_T , and \tilde{H}_V , respectively. We compute the token-wise similarity matrix of the i -th query token

as follows:

$$\alpha_j^\phi = \frac{\exp(\tilde{H}_{\phi_j} \cdot \tilde{H}_Q)}{\sum_j \exp(\tilde{H}_{\phi_j} \cdot \tilde{H}_Q)} \quad (4)$$

where $\phi \in \{V, T\}$ represents the visual or textual feature. \tilde{H}_{ϕ_j} refers to the representation of the j -th visual regions or textual tokens. Finally, the fine-grained fusion for integrating similarity-aware visual or textual hidden states into the query hidden states can be represented as:

$$F(\tilde{H}_Q) = \text{Tanh}(\tilde{H}_Q W_1 + \sum_\phi \sum_j \lambda_\phi \alpha_j^\phi \tilde{H}_{\phi_j}) W_2 + \mathbf{b} \quad (5)$$

where λ_ϕ is the trade-off coefficient to balance the visual similarity and textual similarity $\sum_\phi \lambda_\phi = 1$. W_1 and W_2 are linear transformations parameters and \mathbf{b} is the bias term.

Multimodal Set Prediction Network

Instead of relying on the predefined decoding order in previous methods, we propose a Multimodal Set Prediction Network to maintain suitable inter-entity relationships in a global matching perspective.

Span Boundary Localization. Given the output textual representation $\hat{H}_T \in \mathbb{R}^{n \times h}$ and the query set representation $\hat{H}_Q \in \mathbb{R}^{u \times h}$ from QFNet, we first expand their dimensions as $\hat{H}_T \in \mathbb{R}^{1 \times n \times h}$ and $\hat{H}_Q \in \mathbb{R}^{u \times 1 \times h}$, and then integrate them into a joint representation:

$$S^b = \text{ReLU}(\hat{H}_Q W_Q^b + \hat{H}_T W_T^b) \quad (6)$$

where $S^b \in \mathbb{R}^{u \times n \times h}$ is the joint representation for span boundary localization, W_Q^b and $W_T^b \in \mathbb{R}^{h \times h}$ are learnable parameters. Thus the probability matrix of each textual token being a start index can be calculated as below:

$$P^s = \text{sigmoid}(S_b W_s^b) \quad (7)$$

where $P^s \in \mathbb{R}^{u \times n}$, W_s^b is learnable parameter. Similarly, we can simply replace W_s^b with a new parameter W_e^b to obtain the probability matrix of the end index $P^e \in \mathbb{R}^{u \times n}$.

Candidate Regions Matching. Given the output visual region representation $\hat{H}_V \in \mathbb{R}^{(k+1) \times h}$ and the query set representation $\hat{H}_Q \in \mathbb{R}^{u \times h}$, the candidate regions matching task is quite similar to boundary localization, which also uses a query set to predict the corresponding visual index in the candidate regions proposed by the class-agnostic RPN. The process can be formalized as follows:

$$\begin{aligned} S^r &= \text{ReLU}(\hat{H}_Q W_Q^r + \hat{H}_V W_V^r), \\ P^r &= \text{sigmoid}(S^r W_s^r) \end{aligned} \quad (8)$$

where $P^r \in \mathbb{R}^{u \times (k+1)}$ represents the matching probability matrix. W_s^r , W_Q^r and W_V^r are learnable parameters.

Entity Classification. Since the query set has been endowed with type-level information during its construction, the entity classification task can be regarded as an existence detection for type-grained queries. Besides, considering that boundary localization and region-matching information are useful for query existence detection, we concatenate them

Category	Methods	GMNER			MNER			EEG		
		Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
Pipeline Methods	GVATT-RCNN-EVG	49.36	47.80	48.57	78.21	74.39	76.26	54.19	52.48	53.32
	UMT-RCNN-EVG	49.16	51.48	50.29	77.89	79.28	78.58	53.55	56.08	54.78
	UMT-VinVL-EVG	50.15	52.52	51.31	77.89	79.28	78.58	54.35	56.91	55.60
	UMGF-VinVL-EVG	51.62	51.72	51.67	79.02	78.64	78.83	55.68	55.80	55.74
	ITA-VinVL-EVG	52.37	50.77	51.56	80.40	78.37	79.37	56.57	54.84	55.69
	BARTMNER-VinVL-EVG	52.47	52.43	52.45	<u>80.65</u>	<u>80.14</u>	<u>80.39</u>	55.68	55.63	55.66
Unified Methods	MNER-QG♣	53.02	54.84	53.91	78.16	78.59	78.37	58.48	56.59	57.52
	H-Index	<u>56.16</u>	56.67	56.41	79.37	80.10	79.73	<u>60.90</u>	61.46	61.18
	TIGER♣	55.84	<u>57.45</u>	<u>56.63</u>	79.88	80.70	80.28	60.72	<u>61.81</u>	<u>61.26</u>
	MQSPN (Ours)	59.03	58.49	58.76	81.22	79.66	80.43	61.86	62.94	62.40

Table 1: Performance comparison of different competitive approaches on Twitter-GMNER datasets. Bold represents the optimal result, and underlined represents the suboptimal result. For the baseline methods, ♣ indicates that the results are reproduced according to the corresponding papers, and others are from (Yu et al. 2023a).

with queries. The representation of query existence detection can be formalized as:

$$S^c = \text{ReLU} \left(\left[\hat{H}_Q W_Q^c; P^s \hat{H}_T; P^e \hat{H}_T; P^r \hat{H}_V \right] \right) \quad (9)$$

$$P^c = \text{sigmoid}(S^c W_s^c) \quad (10)$$

where $P^c \in \mathbb{R}^u$ denotes the existence probability of type-grained query. W_Q^c and W_s^c are learnable parameters. Finally, the multimodal entity quadruple predicted by the i -th query can be represented as $\hat{Y}_i = (\hat{Y}_i^s, \hat{Y}_i^e, \hat{Y}_i^t, \hat{Y}_i^r)$, and the predicted quadruple set is $\hat{Y} = \{\hat{Y}_i\}_{i=1}^u$.

Bipartite Matching Loss. Following previous works (Shen et al. 2022; Tan et al. 2021; Chen et al. 2024), we construct a loss function based on optimal bipartite matching. However, since the number of queries u is greater than the total quantity of gold samples m , we define a label \emptyset for those unassignable predictions and then add the label \emptyset to the gold set Y , replicating Y until it reaches the size of u . Our training objective is to find the optimal assignment π that minimizes the global cost. Searching for a permutation of u elements $\pi \in \mathbb{O}_u$ can be formalized as follows:

$$\hat{\pi} = \arg \min_{\pi \in \mathbb{O}_u} \sum_i^u \mathcal{L}_{\text{cost}} \left(Y_i, \hat{Y}_{\pi(i)} \right), \quad (11)$$

$$\mathcal{L}_{\text{cost}} \left(Y_i, \hat{Y}_{\pi(i)} \right) = -1_{\{Y_i^t \neq \emptyset\}} \left[p_{\pi(i)}^c \left(Y_i^t \right) \right. \\ \left. + p_{\pi(i)}^s \left(Y_i^s \right) + p_{\pi(i)}^e \left(Y_i^e \right) + p_{\pi(i)}^r \left(Y_i^r \right) \right].$$

where $\mathcal{L}_{\text{cost}}(\cdot)$ denotes the pair matching cost between the gold quadruple Y_i and a prediction with index $\pi(i)$. $1_{\{\delta\}}$ represents the indicator function that takes 1 when δ is true and 0 otherwise. The optimal assignment π can then be efficiently solved by the off-the-shelf Hungarian Algorithm (Kuhn 1955). To jointly train the model, the final loss function is defined as follows:

$$\mathcal{L}(Y, \hat{Y}) = \sum_{i=1}^u \left\{ -\log p_{\hat{\pi}(i)}^c \left(Y_i^t \right) \right. \\ \left. + 1_{\{Y_i^t \neq \emptyset\}} \left[-\log p_{\hat{\pi}(i)}^s \left(Y_i^s \right) \right. \right. \\ \left. \left. - \log p_{\hat{\pi}(i)}^e \left(Y_i^e \right) - \log p_{\hat{\pi}(i)}^r \left(Y_i^r \right) \right] \right\}. \quad (12)$$

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

Table 2: F1 scores of previous unified methods and our MQSPN on Twitter-FMNERG datasets for each fine-grained entity type in GMNER task.

Experiments

Experiment Settings

GMNER Datasets. We conduct experiments on the Twitter-GMNER (Yu et al. 2023a) and Twitter-FMNERG (Wang et al. 2023). Please refer to Appendix for detailed information about these datasets and evaluation metrics.

Implementation Details All experiments are implemented on 4 NVIDIA RTX3090 GPUs with Pytorch 1.9.1. For a fair comparison, we use the pre-trained BERT-based model¹ as the textual encoder, ViT-B/32 from pre-training CLIP² as the visual encoder, and VinVL³ as a class-agnostic RPN. For model’s efficiency, we freeze the visual encoder and RPN and assign the layer of the QFNet as $L = 3$. We initialize the learnable query part with a normal distribution $\mathcal{N}(0, 0.02)$. For training, we set the batch size to 16, the learning rate to 2×10^{-5} , and the training epoch to 50. Our model uses an Adam optimizer with a linear warmup of ratio 0.05. To allow the multi-grained queries to learn query semantics initially, we first train the model for 5 epochs with

¹<https://huggingface.co/google-bert/bert-base-uncased>

²<https://huggingface.co/openai/clip-vit-base-patch32>

³<https://github.com/pzzhang/VinVL>

Module	Settings	GMNER	FMNERG
MQSPN	-	58.76	48.57
MQS	w/o PTQ	56.95 (↓1.81)	47.23 (↓1.34)
	w/o LEQ	56.02 (↓2.74)	45.39(↓3.18)
QFNet	w/o QCT	57.21 (↓1.55)	47.61(↓0.96)
	w/o QPI	57.86 (↓0.90)	48.33(↓0.24)
	w/o SAG	58.13 (↓0.63)	47.96(↓0.61)
MSP	w/o BML	56.58 (↓2.18)	46.81(↓1.76)

Table 3: Ablation study of each component on overall F1 score of Twitter-GMNER and Twitter-FMNERG.

the freezing parameters of the pre-training model. Please refer to Appendix for baseline methods.

Overall Performance

Performance on GMNER, MNER, and EEG. Following (Yu et al. 2023a), we also report two subtasks of GMNER, i.e., Multimodal Named Entity Recognition (MNER) and Entity Extraction & Grounding (EEG). MNER aims to identify entity span and entity type pairs, while EEG aims to extract entity span and entity region pairs. Table 1 shows the performance comparison of our method with competitive baseline models on Twitter-GMNER benchmarks.

First, unified models are significantly superior to pipeline methods due to the joint training of MNER and EEG to mitigate error propagation. Second, MQSPN significantly outperforms the MRC-based method MNER-QG by +4.85% F1 scores. Furthermore, compared with the previous state-of-the-art (SOTA) generative model TIGER, our method MQSPN exhibits superior performance, achieving +0.15%, +1.14%, and +2.13% F1 scores improvements on the MNER, EEG, and GMNER tasks, respectively.

We attribute the performance improvements of MQSPN to the following reasons: (1) Compared with the sequence generation-based approach H-Index and TIGER, MQSPN eliminates the dependencies on predefined decoding order and establishes suitable inter-entity relationships from a global matching view. (2) Compared with the MRC-based approach MNER-QG, MQSPN can learn fine-grained entity semantics and model distinct intra-entity connections between regions and entities.

Performance on fine-grained GMNER. We validate the fine-grained GMNER ability of different methods on Twitter-FMNERG datasets, and the experimental results are shown in Table 2. It reveals that our proposed MQSPN achieves the best results on all fine-grained entity types except *Other* and *Art* among the state-of-the-art models. Specifically, we achieve a great improvement of +3.51%, +3.34%, +4.78%, +2.28%, +1.27%, +3.49%, and +1.37% F1 scores on the *Person*, *Location*, *Building*, *Organization*, *Product*, *Event*, and *All* entity types respectively. The experimental results demonstrate that through robust modeling of intra-entity and inter-entity relationships, MQSPN exhibits superior capability in fine-grained multimodal entity understanding.

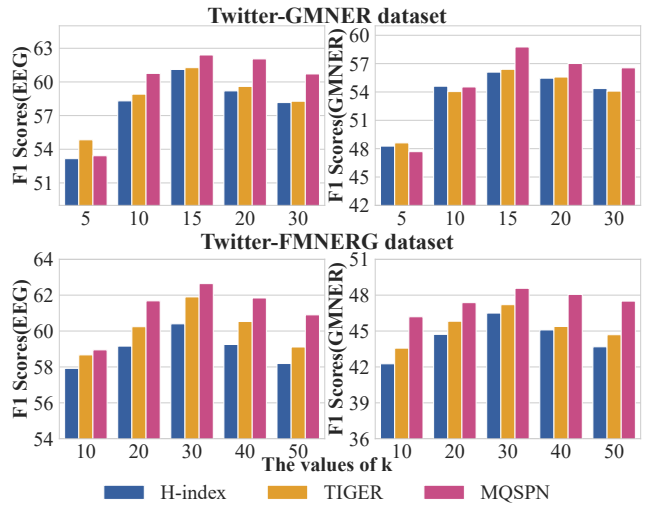


Figure 3: Analysis of candidate visual regions number k for H-Index, TIGER, and our MQSPN.

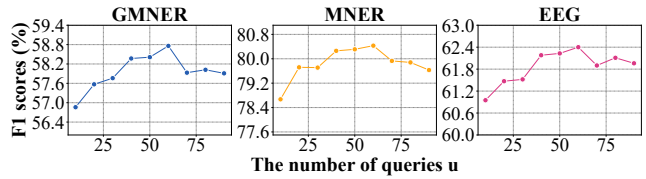


Figure 4: Analysis of the multi-grained queries quantity u on GMNER, MNER and EEG tasks.

Ablation Study

Ablation setting. To verify the effectiveness of each designed component in MQSPN, we conduct a comprehensive ablation study on the Twitter-GMNER and Twitter-FMNERG dataset: **For Multi-grained Query Set (MQS)**, (1) *w/o PTQ*: we remove the prompt-based type-grained part of the query set. (2) *w/o LEQ*: we replace the learnable queries part with human-designed queries (the same as MNER-QG query construction). **For Query-guided Fusion Net (QFNet)**, (3) *w/o QCT*: we encode sentences and the query set using the original BERT without query-text cross-attention. (4) *w/o QPI*: we eliminate the query-region prefix integration. (5) *w/o SAG*: we eliminate the similarity-aware aggregator. **For Multimodal Set Prediction (MSP)**, (6) *w/o BML*: we replace the bipartite matching loss with a joint cross-entropy loss in the fixed permutation of the entities. The experimental results are shown in Table 3.

The effectiveness of Multi-grained Query Set. In Table 3, we observe a clear F1 scores drop in model performance (1.81% in Twitter-GMNER and 1.34% in Twitter-FMNERG) without the prompt-based type-grained queries. This indicates that integrating type-grained information into the query can enhance model performance. Furthermore, adding an additional learnable entity-grained part to the query substantially improves model performance: in the Twitter-GMNER and Twitter-FMNERG datasets, the F1

Input Text:	It' s almost real hockey time: # Blackhawks Bradon Saad (20) ahead of Detroit' s Johan Franzen in the first period ...			
EEG:				
	Ground Truth	w/o LEQ	w/o BML	MQSPN
MNER:	(Blackhawks, Sport Team) ✓	(Blackhawks Bradon Saad, Person) ✗	(Blackhawks, Other) ✗	(Blackhawks, Sport Team) ✓
	(Bradon Saad, Person) ✓	(Detroit, Other) ✗	(Bradon Saad, Person) ✓	(Bradon Saad, Person) ✓
	(Detroit, Sport Team) ✓	(Detroit' s Johan Franzen, Person) ✗	(Detroit, Other) ✗	(Detroit, Sport Team) ✓
	(Johan Franzen, Person) ✓		(Johan Franzen, Person) ✓	(Johan Franzen, Person) ✓

Figure 5: Predictions of MQSPN in different ablation setting.

scores increased by +2.74% and +3.18%, respectively. The main reason is that these learnable vectors can effectively learn the entity-grained semantics and model intra-entity connections between entity spans and regions, thereby boosting model performance.

The effectiveness of Query-guided Fusion Net. In Table 3, the model without three fusion modules achieves consistent drops. We find the average F1 scores of these three modules decreased by 1.03% and 0.60% on Twitter-GMNER and Twitter-FMNERG, respectively. Experimental results demonstrate that the QFNet can fuse information across different modalities and filter the noise introduced by irrelevant visual regions, which is crucial for improving the performance of two-level relationship modeling.

The effectiveness of Bipartite Matching Loss. In Table 3, excluding the bipartite matching loss from the model leads to an F1 scores decline of 2.18% and 1.76% in Twitter-GMNER and Twitter-FMNERG datasets. This illustrates that the bipartite matching loss contributes to model performance through modeling inter-entity relationships in global matching perspective.

Discussion and Analysis

Sensitivity Analysis of Irrelevant Visual Regions. The performance of EEG is largely determined by the ground truth coverage of the top- k regions proposed by the class-agnostic RPN. However, a higher value of k implies the introduction of more irrelevant visual regions. To delve into the model's error tolerance to noisy visual regions, we conduct experiments on H-Index, TIGER, and MQSPN under different values of k . As shown in Figure 3, we can observe that the F1 scores of these three methods first increase, and then decrease with the augment of k . Compared with H-Index and TIGER, our proposed MQSPN has better and more stable performance on both EEG and GMNER tasks when k becomes larger. This demonstrates MQSPN's insensitivity to irrelevant visual regions, indicating its excellent error tolerance and robustness of intra-entity connection.

Analysis of query quantity. To explore the impact of

multi-grained query quantity (i.e., u) on GMNER and its 2 subtasks, we report the F1 scores of Twitter-GMNER in Figure 4 by tuning u from 10 to 90. We observe that as the number of queries increases from 10 to 60, the model's performance gradually improves. These results indicate that query quantity plays a crucial role in intra-entity connection learning. However, a query quantity exceeding 60 does not lead to better performance. There is an optimal number of queries for MQSPN. In our experiments, we find $u = 60$ in the Twitter-GMNER dataset can achieve the optimal results.

Case Study. We conduct a comprehensive case study in different ablation setting. As shown in Figure 5, when we replace learnable queries with human-designed queries in MQS (*w/o LEQ setting*), the model cannot differentiate *Bradon Saad* and *Johan Franzen* regions, and incorrectly localize *Blackhawks* and *Detroit* as *Person* spans, highlighting deficiencies in intra-entity learning for ambiguous entities. After integrating learnable MQS, the model precisely aligns textual spans with corresponding regions for *Person* entities. However, using MQS without MSP (*w/o BML*) leads to confusion in mapping proper inter-entity relationships, resulting in incorrect classifications for *Blackhawks* and *Detroit* and more erroneous span-type-region triplet predictions. After combining MQS and MSP, MQSPN effectively differentiates entities and makes accurate predictions, demonstrating the effect of intra-entity and inter-entity modeling.

Conclusion

In this paper, we propose a novel method named MQSPN to model appropriate intra-entity and inter-entity relationships for the GMNER task. To achieve this, we propose a learnable Multi-grained Queries Set (MQS) to adaptively learn explicit intra-entity connections. Besides, we reformulate GMNER as a multimodal set prediction (MSP) to model inter-entity relationships from an optimal matching view. Experimental results demonstrate that MQSPN achieves state-of-the-art performance on GMNER and its 2 subtasks across two Twitter benchmarks.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (U2001211, U22B2060, 62276279), National Natural Science Foundation Youth Basic Research Program (623B2100), Research Foundation of Science and Technology Plan Project of Guangzhou City (2023B01J0001, 2024B01W0004), Guangdong Basic and Applied Basic Research Foundation (2024B1515020032).

References

- Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.
- Chen, S.; Aguilar, G.; Neves, L.; and Solorio, T. 2021. 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)*, 87–96.
- Chen, X.; Li, L.; Zhu, Y.; Deng, S.; Tan, C.; Huang, F.; Si, L.; Zhang, N.; and Chen, H. 2024. Sequence Labeling as Non-autoregressive Dual-Query Set Generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.
- Chen, X.; Zhang, N.; Li, L.; Deng, S.; Tan, C.; Xu, C.; Huang, F.; Si, L.; and Chen, H. 2022a. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, 904–915.
- Chen, X.; Zhang, N.; Li, L.; Yao, Y.; Deng, S.; Tan, C.; Huang, F.; Si, L.; and Chen, H. 2022b. Good Visual Guidance Make A Better Extractor: Hierarchical Visual Prefix for Multimodal Entity and Relation Extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 1607–1618.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv*, abs/1810.04805.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissensborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houshy, N. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv*, abs/2010.11929.
- Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.
- Gong, Z.; Li, F.; Deng, Y.; Bhattacharjee, D.; Zhu, X.; and Ji, Z. 2024. CoDA: Instructive Chain-of-Domain Adaptation with Severity-Aware Visual Prompt Tuning. *arXiv preprint arXiv:2403.17369*.
- Jia, M.; Shen, L.; Shen, X.; Liao, L.; Chen, M.; He, X.; Chen, Z.; and Li, J. 2023. 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*, volume 37, 8032–8040.
- Jia, M.; Shen, X.; Shen, L.; Pang, J.; Liao, L.; Song, Y.; Chen, M.; and He, X. 2022. Query prior matters: A mrc framework for multimodal named entity recognition. In *Proceedings of the 30th ACM international conference on multimedia*, 3549–3558.
- Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2): 83–97.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.
- Li, J.; Li, H.; Sun, D.; Wang, J.; Zhang, W.; Wang, Z.; and Pan, G. 2024a. LLMs as Bridges: Reformulating Grounded Multimodal Named Entity Recognition. *arXiv preprint arXiv:2402.09989*.
- Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4582–4597.
- Li, Z.; Yu, J.; Yang, J.; Wang, W.; Yang, L.; and Xia, R. 2024b. Generative Multimodal Data Augmentation for Low-Resource Multimodal Named Entity Recognition. In *ACM Multimedia 2024*.
- Liu, Y.; Li, H.; Garcia-Duran, A.; Niepert, M.; Onoro-Rubio, D.; and Rosenblum, D. S. 2019. MMKG: Multi-modal Knowledge Graphs. In *The Semantic Web*, 459–474. Cham: Springer International Publishing.
- Ok, H.; Kil, T.; Seo, S.; and Lee, J. 2024. SCANNER: Knowledge-Enhanced Approach for Robust Multimodal Named Entity Recognition of Unseen Entities. *arXiv preprint arXiv:2404.01914*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *ICML*, 8748–8763.
- Redmon, J.; and Farhadi, A. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.
- Shen, Y.; Wang, X.; Tan, Z.; Xu, G.; Xie, P.; Huang, F.; Lu, W.; and Zhuang, Y. 2022. Parallel Instance Query Network for Named Entity Recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 947–961.
- Tan, Z.; Shen, Y.; Zhang, S.; Lu, W.; and Zhuang, Y. 2021. A sequence-to-set network for nested named entity recognition. *arXiv preprint arXiv:2105.08901*.
- Uijlings, J. R.; Van De Sande, K. E.; Gevers, T.; and Smeulders, A. W. 2013. Selective search for object recognition. *International journal of computer vision*, 104: 154–171.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, J.; Li, Z.; Yu, J.; Yang, L.; and Xia, R. 2023. Fine-grained multimodal named entity recognition and grounding

with a generative framework. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3934–3943.

Wu, J.; Gong, C.; Cao, Z.; and Fu, G. 2023. MCG-MNER: A Multi-Granularity Cross-Modality Generative Framework for Multimodal NER with Instruction. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3209–3218.

Yu, J.; Jiang, J.; Yang, L.; and Xia, R. 2020. Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer. In *ACL*, 3342–3352.

Yu, J.; Li, Z.; Wang, J.; and Xia, R. 2023a. Grounded multimodal named entity recognition on social media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9141–9154.

Yu, J.; Su, Q.; Quan, X.; and Yin, J. 2021. Multi-hop reasoning question generation and its application. *IEEE Transactions on Knowledge and Data Engineering*, 35(1): 725–740.

Yu, J.; Wang, S.; Zheng, L.; Su, Q.; Liu, W.; Zhao, B.; and Yin, J. 2023b. Generating Deep Questions with Commonsense Reasoning Ability from the Text by Disentangled Adversarial Inference. In *Findings of the Association for Computational Linguistics: ACL 2023*, 470–486.

Yu, J.; Zha, Z.; and Yin, J. 2019. Inferential machine comprehension: Answering questions by recursively deducing the evidence chain from text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2241–2251.

Zhang, D.; Wei, S.; Li, S.; Wu, H.; Zhu, Q.; and Zhou, G. 2021a. Multi-modal Graph Fusion for Named Entity Recognition with Targeted Visual Guidance. In *AAAI*, volume 35, 14347–14355.

Zhang, D.; Wei, S.; Li, S.; Wu, H.; Zhu, Q.; and Zhou, G. 2021b. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 14347–14355.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021c. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5579–5588.