A Visual Annotation-Free Method That Rivals Fully Supervised Methods for Grounded Multimodal Named Entity Recognition

Jia Yang, Jianfei Yu, Zilin Du, Wenya Wang, Li Yang, Rui Xia, Boyang Li

***Abstract*—Grounded Multimodal Named Entity Recognition (GMNER) aims to extract named entities, their types, and corresponding visual objects from image-text pairs. However, existing GMNER methods rely on costly multimodal annotations, limiting their scalability in real applications. To address this issue, we propose a visual annotation-free framework that leverages text-only NER data and a Zero-shot Entity Visual Grounding (ZeroEVG) approach. ZeroEVG consists of three modules: (1) Candidate Object Generation, which pre-selects visual object candidates; (2) Entity-Object Matching, which determines whether an entity has a visual presence; and (3) Entity Visual Localization, which employs a variant of GradCAM to identify bounding boxes for groundable entities. Experimental results on two benchmark datasets show that our visual annotation-free framework achieves competitive performance with fully supervised multimodal approaches, and even surpasses some of them under the same backbone on both GMNER and EVG tasks.**

***Index Terms*—Multimodal Named Entity Recognition, Entity Visual Grounding, Pretrained Vision-Language Models**

I. Introduction

As an emerging task in multimodal information extraction, Grounded Multimodal Named Entity Recognition (GMNER) requires jointly extracting named entities, their associated types, and corresponding visual objects from image-text pairs [1]. For instance, given the image-text pair in Fig. 1(a), the goal of GMNER is to extract five entity-type-object triplets, i.e., *US-Location-None*, *Ted Cruz-Person-Box-1*, *Heidi Cruz-Person-Box-2*, *Donald Trump-Person-None*, and *Republican-Organization-None*. Since these extracted entity-object pairs are crucial for many applications such as multimodal knowledge graph construction [2], [3] and knowledge-intensive VQA [4], [5], GMNER has recently attracted considerable attention [6]–[9].

With the recent advancements in pre-trained models and Large Language Models (LLMs), many supervised deep learning approaches have been proposed for GMNER, achieving promising results on benchmark datasets. These include sequence labeling-based methods [8], pointer or paragraph generation methods [1], [10], and LLM-based methods [6]. Despite their success, these methods rely heavily on large-scale labeled training data, making them difficult to scale to new domains.

A major challenge in GMNER is the need for fine-grained multimodal annotation, which involves both textual and visual supervision. While text annotation for named entities is relatively accessible due to the abundance of NER datasets across various domains, annotating images for GMNER is significantly more challenging. Grounding named entities to their corresponding visual objects demands extensive prior visual knowledge and meticulous inspection of all visual elements within an image. For instance, as illustrated in Fig. 1(a), accurately annotating the entity *Ted Cruz* involves two key steps: (1) identifying all potential visual objects in the image and determining whether any correspond to the entity type of *Ted Cruz* (i.e., *Person*), and (2) recognizing *Ted Cruz* based on prior knowledge of his appearance to correctly assign the bounding box. This intricate process is both labor-intensive and costly, limiting the scalability of existing supervised methods.

To alleviate the costly image annotation, in this paper, we explore a new visual annotation-free paradigm for GMNER, which leverages text-only NER data and zero-shot entity visual grounding to extract the entity-type-object triplets, as illustrated in Fig. 1. Since many pre-trained vision-language models (VLMs) such as CLIP [11] and object detectors like VinVL [12] have demonstrated strong performance in image-text matching [13] and open-vocabulary object detection [14], several recent studies have explored their application for zero-shot visual grounding [15], [16]. However, applying these models to entity visual grounding in GMNER presents two unique challenges: (1) Not all named entities in GMNER have corresponding visual objects in the image, whereas visual grounding assumes that every text query is visually present. (2) The textual queries in GMNER are named entities, which are often domain-specific and highly personalized, rather than simple and generic phrases used in visual grounding.

To address these challenges, we propose a novel visual annotation-free framework for the GMNER task, comprising a base NER model and a Zero-shot Entity Visual Grounding (ZeroEVG) approach. First, we utilize an NER model trained solely on text-only datasets to extract named entities. Then, ZeroEVG predicts whether each entity has a visual presence and, if so, locates its corresponding visual object. ZeroEVG consists of three key modules: (1) Candidate Object Generation, which employs a category filtering strategy to pre-select candidate visual objects that match predefined entity categories, avoiding an exhaustive search on the entire

This work was supported by the National Natural Science Fouandation of China under Grant No. 62476132 and No. 62476134.

J. Yang and J. Yu are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China. (e-mail: jyang7@njust.edu.cn; jfyu@njust.edu.cn)

Z. Du, W. Wang and B. Li are with the College of Computing and Data Science, Nanyang Technological University, Singapore. (e-mail: zilin003@ntu.edu.sg; wangwy@ntu.edu.sg; boyang.li@ntu.edu.sg)

L. Yang is with the School of Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore. (e-mail: YANG0666@e.ntu.edu.sg)

R. Xia is with the School of Intelligence Science and Technology, Nanjing University, Suzhou 215163, China. (e-mail: rxia@nju.edu.cn)

Corresponding Author: *Jianfei Yu*.

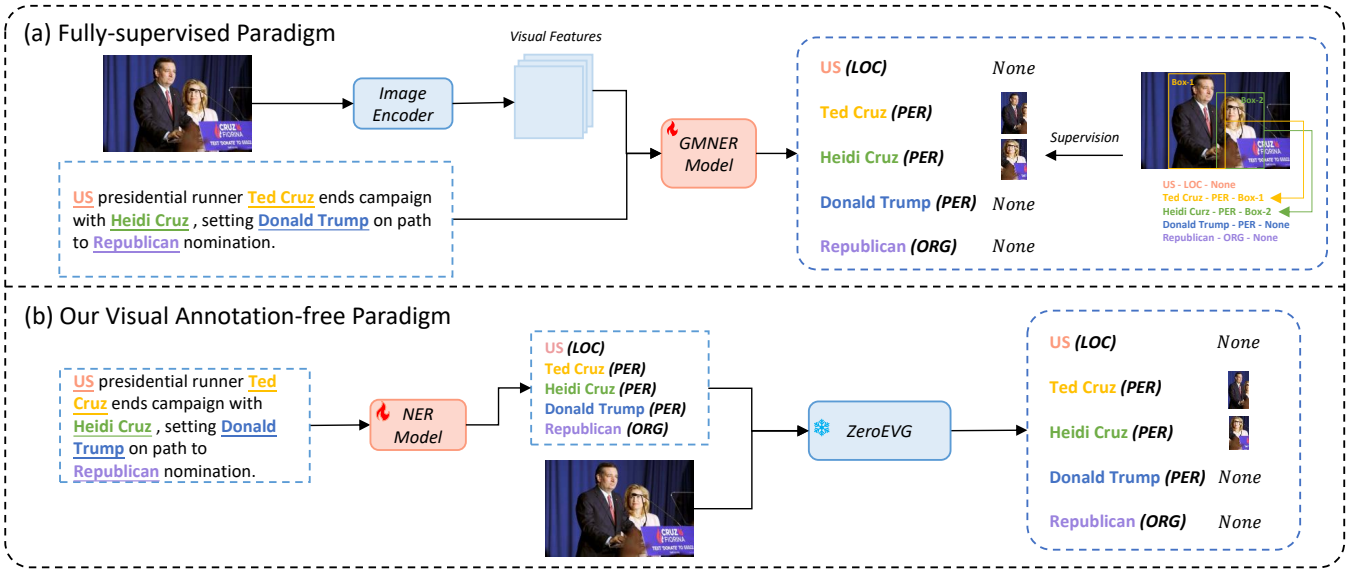


Fig. 1. A comparison between the previous fully-supervised paradigm and our visual annotation-free paradigm for the GMNER task. *None* indicates that the entity has no corresponding visual grounding boxes in the image.

image; (2) Entity-Object Matching, which computes fine-grained semantic similarity scores between detected entities and candidate objects to determine visual presence of each entity; and (3) Entity Visual Localization, which applies a variant of GradCAM [17] over cross-attention maps in VLMs to identify the most semantically relevant image regions for each groundable entity. These regions are then integrated with entity-object matching scores to predict the final bounding boxes.

The main contributions of this work are summarized as follows:

- We introduce a new visual annotation-free paradigm for GMNER, eliminating the need for costly image annotations by leveraging text-only NER data and zero-shot entity visual grounding.
- We propose Zero-Shot Entity Visual Grounding (ZeroEVG), a method that harnesses pre-trained VLMs and object detectors to perform entity-object matching and entity localization without any visual supervision.
- Our framework achieves strong and competitive performance compared with fully supervised multimodal methods across various NER models with the same backbone on two benchmark GMNER datasets, and even outperforms several of them, highlighting the potential of visual annotation-free methods for scalable and domain-agnostic GMNER applications.

II. Related Work

A. Multimodal Named Entity Recognition

Traditional information extraction studies have primarily focused on Named Entity Recognition (NER) [18]–[22], which identifies recognizing named entities in text and classifying them into predefined categories. Multimodal Named Entity Recognition (MNER) extends this task by incorporating image information to enhance recognition accuracy. Early studies [23]–[25] explored visual feature fusion to enhance textual representation learning. With the emergence of multimodal transformers, various attention-based mechanisms [26]–[33] have been designed to model cross-modal interactions for improved entity recognition. Moreover, image-to-text conversion techniques [34] and external knowledge retrieval methods [35], [36] have been proposed to augment textual information with visual and external knowledge. Beyond traditional sequence labeling-based approaches, recent works have introduced machine reading comprehension-based [37], [38], in-context learning-based [39], [40], generation-based [1], [10], and information bottleneck-based [41] methods for MNER.

B. Grounded Multimodal Named Entity Recognition

Grounded Multimodal Named Entity Recognition (GMNER) extends the MNER task by not only recognizing named entities and their categories but also grounding them to corresponding visual objects via bounding box annotations. Existing approaches primarily follow an end-to-end paradigm, where GMNER is formulated as sequence labeling tasks [8], multimodal index or paraphrase generation tasks [1], [10], or set prediction tasks [9]. Some recent LLM-based methods attempted to decompose the task into MNER, visual entailment, and visual grounding modules, and use pre-trained LLMs and visual grounding models to solve them in a pipeline mannner [6]. In addition, Wang et al. [42] proposed GEM, which combines multi-granularity entity recognition, MLLM-based reranking, and LVLM-based implicit grounding to improve fine-grained MNERG.

Despite their effectiveness, these existing GMNER methods rely heavily on large-scale annotated datasets, requiring fine-grained annotations of named entities, their types, and corresponding bounding boxes. To alleviate the costly image annotation, in this paper, we aim to explore a new visual annotation-free method for GMNER.

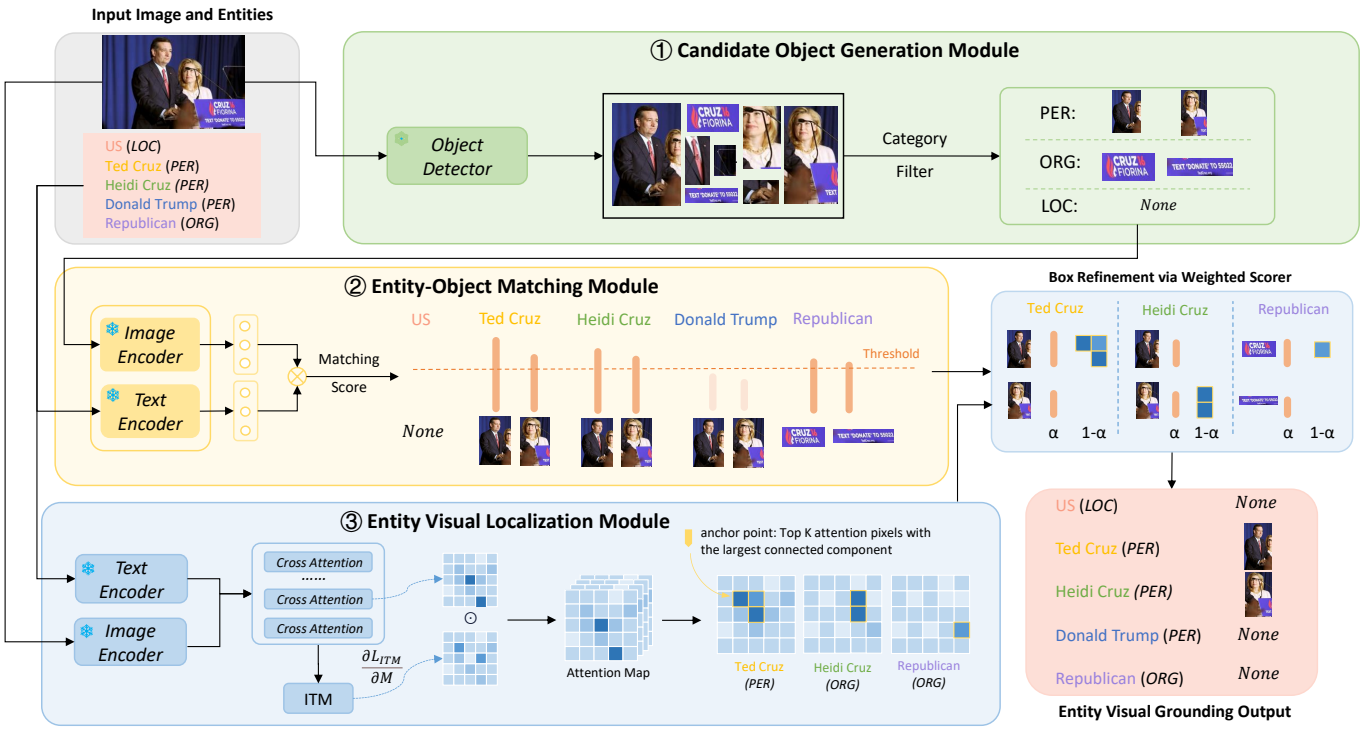


Fig. 2. The overall architecture of the proposed Zero-shot Entity Visual Grounding (ZeroEVG) approach.

C. Visual Grounding

Visual Grounding aims to localize objects in images based on referring expressions with annotated bounding boxes. With the adoption of transformer architectures, cross-modal grounding models [43]–[47] were developed to jointly encode visual and textual features. Subsequent research leveraged vision-language pre-trained models [48] to improve grounding accuracy via large-scale cross-modal alignment. Furthermore, grounding-oriented pre-training [49]–[55] and multimodal large language models [56]–[59] have pushed the field towards more general and open-world grounding scenarios.

However, VG methods typically assume the existence of matching visual objects and mainly handle short referring expressions, which limits their applicability to GMNER tasks. RIVEG [6] introduced a visual entailment module to handle ungroundable entities. In contrast, our method utilizes category filtering and score thresholding to determine entity presence, achieving superior performance over fully supervised baselines even without visual supervision.

III. Methodology

A. Task Definition

In this paper, we focus on the visual annotation-free setting of the GMNER task. The goal is to extract named entities, their associated types, and their corresponding visual objects without relying on human-annotated image supervision.

Formally, given a text-image pair ($\boldsymbol{T}$, $\boldsymbol{I}$), where $\boldsymbol{T} = (w_1, \ldots, w_n)$ is a sequence of n words and $\boldsymbol{I}$ is the associated image, the objective is to extract a set of entity-type-object triplets:

$$S_{\text{GMNER}} = \{\ldots, (e_i, c_i, o_i), \ldots\}, \tag{1}$$

where e_i is a named entity span in $\boldsymbol{T}$, c_i is the entity type from a predefined entity type set, $o_i = (o_i^{x_1}, o_i^{y_1}, o_i^{x_2}, o_i^{y_2})$ is the bounding box of the visual object corresponding to e_i, or *None* if e_i is not visually grounded.

Note that unlike traditional GMNER methods that require both text and image annotations for training, our approach is trained solely on a text-only NER corpus $\mathcal{D}$. During training, the model learns entity recognition from textual supervision only, without relying on any aligned visual annotations. At inference time, however, the model is provided with both the input text and the corresponding image. The visual information is then incorporated through the designed grounding module to enhance entity disambiguation and visual grounding.

To achieve Visual Annotation-Free GMNER, we propose a framework consisting of a base NER model trained on $\mathcal{D}$ and a Zero-shot Entity Visual Grounding (ZeroEVG) approach, which are detailed in the next two subsections.

B. Named Entity Recognition

In our framework, we employ a standard NER model to extract named entities and their corresponding types from text. Since our framework is agnostic to the choice of NER models, any fine-tuned NER model can be adopted. Specifically, we consider three representative models trained on the text-only NER corpus $\mathcal{D}$: (1) **BARTNER-None** [60], which formulates NER as a span-based index generation task using a BART encoder-decoder architecture to directly

generate entity span boundaries along with their types; (2) **T5-Paraphrase** [10], a text-only baseline adapted from TIGER that converts structured entity–type–object triples into natural language sentences and leverages the generative capacity of T5 to produce paraphrased outputs; (3) **RoBERTa-CRF** [61], a robust sequence labeling baseline that encodes sentences with RoBERTa and applies a CRF decoding layer to capture label dependencies and ensure valid tag transitions, thereby enhancing the robustness of token-level predictions.

C. Zero-Shot Entity Visual Grounding

After detecting named entities using the trained NER model, we propose a Zero-shot Entity Visual Grounding (ZeroEVG) approach to identify the visual presence of each detected entity and localizes the corresponding visual object for each groundable entity. As illustrated in Fig. 2, ZeroEVG consists of three modules: (1) Candidate Object Generation, which filters candidate visual objects based on entity types; (2) Entity-Object Matching, which computes the semantic similarity between entities and candidate objects; and (3) Entity Visual Localization, which leverages a variant of GradCAM over cross-attention maps of VLMs to identify relevant image regions and predict bounding boxes.

1) Candidate Object Generation Module: **Object Detection.** Given an input image $\boldsymbol{I}$, we follow previous studies [1], [10] by using pre-trained object detectors like VinVL [12] to generate object proposals within the image. We then rank these object proposals based on their prediction confidence and retain the top-K objects, along with their predicted categories (e.g., *girl*, *dog*, *church*).

Category Filter. We further apply a category filtering strategy to refine the candidate visual objects. Specifically, given each predefined entity type (e.g., *Person*, *Location*), we leverage the widely-used GPT-4o model [62] once for each dataset–detector combination in an offline step, independent of training and inference, to identify all matched object categories pre-defined by object detectors (the exact prompt template is provided in the Appendix). Object proposals are retained if their object categories match the predefined entity types:

$$\{t_k, o_k\}_{k=1}^{n} = \text{Object-Detector}(\boldsymbol{I}, K), \tag{2}$$

$$t_k \in \boldsymbol{S}(c), \tag{3}$$

$$o_k = (o_k^{x_1}, o_k^{y_1}, o_k^{x_2}, o_k^{y_2}), \tag{4}$$

where o_k is one of the top-K visual objects, t_k is the predicted category of o_k, and $\boldsymbol{S}(c)$ refers to the set of object categories matching predefined entity types. This category-based filtering significantly narrows down the search space, ensuring that only relevant candidates are kept, without the need for an exhaustive search over all object proposals.

2) Entity-Object Matching Module: Based on the detected named entities and candidate objects, the entity-object matching module aims to identify whether a named entity is matched with any candidate object in the image.

To achieve this, we compute the semantic similarity between each entity-object pair using a pre-trained VLM named CLIP [11]. For each detected entity e_i, we generate a text prompt $\boldsymbol{x}_i$ "a picture of *named entity*, a/an *entity type*", which includes the entity and its corresponding type. $\boldsymbol{x}_i$ is then fed into the CLIP text encoder to obtain the text embedding $\mathbf{x}_i^t$. Simultaneously, we crop the image region corresponding to each candidate object o_j and pass it through the CLIP image encoder to obtain the image embedding $\mathbf{x}_j^v$. The similarity between the entity e_i and the object o_j is computed as their dot product below:

$$\text{clip-score}(e_i, o_j) = \mathbf{x}_i^t \cdot \mathbf{x}_j^v, \tag{5}$$

where $\cdot$ denotes the dot product operation. If the CLIP similarity scores between a detected entity and all candidate objects are lower than a predefined threshold β, the entity is considered ungroundable; otherwise, it is deemed groundable.

3) Entity Visual Localization Module: After detecting the groundable entities, the entity visual localization module aims to localize their corresponding visual objects. To accomplish this without relying on fine-grained image annotations, we leverage the large-scale pre-trained VLM named BLIP [63] and apply a variant of GradCAM [17] to its image-text matching (ITM) loss to identify the most relevant regions in the image for each groundable entity.

Specifically, for each groundable entity e_k, we first feed its text prompt $\boldsymbol{x}_k$ and the input image $\boldsymbol{I}$ to the ITM network of BLIP. We then apply a variant of the GradCAM interpretability method [64] to highlight the most important image regions by computing the gradient of ITM loss with respect to the attention maps. Formally, let $\mathbf{X}^v \in \mathbb{R}^{p \times \boldsymbol{D}^v}$ and $\mathbf{X}^t \in \mathbb{R}^{q \times \boldsymbol{D}^t}$ represent the image feature map and the embedding of the text prompt, respectively, where p is the number of image patches, q is the number of textual tokens, and $\boldsymbol{D}^v$ and $\boldsymbol{D}^t$ are the dimensionalities of the image feature map and the textual embedding. The attention map $\mathbf{A} \in \mathbb{R}^{p \times q}$ at a given cross-attention layer can be computed by:

$$\mathbf{A} = \frac{\mathbf{X}^t \mathbf{W}_q \mathbf{W}_k^\top \mathbf{X}^v}{\sqrt{\boldsymbol{D}^t}}, \tag{6}$$

where $\mathbf{W}_q$ and $\mathbf{W}_k$ are the weight matrices for the query and key projections. Inspired by GradCAM, we use the *matching* label of the ITM network to compute the gradient of the ITM loss with respect to $\mathbf{A}$. This gradient indicates how much each attention score contributes to the entity-object match. For each groundable entity e_k, we compute the relevance score of the i-th image patch by averaging the gradients over multiple attention heads and summing them across all the q textual tokens:

$$\text{rel}_{\text{patch}_i} = \frac{1}{H} \sum_{j=1}^{q} \sum_{h=1}^{H} \max(0, \frac{\partial \mathcal{L}_{\text{ITM}}}{\partial \mathbf{A}_{ji}^h}) \mathbf{A}_{ji}^h, \tag{7}$$

where H denotes the number of attention heads and we only consider positive gradients.

Largest Connected Component. Next, to identify the most relevant region for each entity, we rank all p image patches by their relevance scores in Eq. (7) and retain the top-P patches. We then construct a binary mask matrix with the same dimensions as the input image, where the top-P patches are assigned a value of 1 and the remaining patches are set to 0. Using the fast connected-component labeling algorithm [65],

TABLE I
STATISTICS OF THE TWITTER-GMNER AND TWITTER-FMNERG DATASETS. NOTE THAT TWITTER-FMNERG CONTAINS 51 FINE-GRAINED ENTITY TYPES, WITH THEIR CORRESPONDING 8 COARSE-GRAINED TYPES PRESENTED HERE.

Split	#Tweet	#Entity	#Groundable Entity	#Box	Twitter-GMNER				Twitter-FMNERG							
					PER	LOC	ORG	Other	PER	LOC	Building	ORG	Product	Art	Event	Other
Train	7,000	11,779	4,733	5,723	5,019	1,918	3,035	1,807	5,019	1,553	365	3,035	355	495	614	343
Dev	1,500	2,450	991	1,171	1,072	407	595	376	1,072	345	62	595	82	103	126	65
Test	1,500	2,543	1,046	1,254	1,104	404	638	397	1,104	327	77	638	88	106	129	74
Total	10,000	16,772	6,770	8,148	7,195	2,729	4,268	2,580	7,195	2,225	504	4,268	525	704	869	482

we scan through the binary mask matrix and group adjacent patches (with a value of 1) into connected components based on eight-connected connectivity. Each connected component is a set of patches that are directly connected either horizontally, vertically, or diagonally (in any of the eight directions). Finally, we select the largest connected component as the most relevant region in the image. For example, in Fig. 2, the largest connected component of *Ted Cruz* is the three dark blue boxes in the center.

Box Refinement via Weighted Scorer. However, our preliminary studies show that GradCAM-identified regions often cover only small patches within the candidate object, failing to capture the full visual extent of the entity. To address this, we refine these regions using candidate objects from Section III-C2 to determine the final visual grounding. Specifically, for a given entity e_k and candidate object o_j, we compute their final relevance as a weighted sum of the CLIP similarity score and the spatial coverage of GradCAM-highlighted regions:

$$\text{rel}(e_k, o_j) = \alpha * \text{clip-score}(e_k, o_j) + (1 - \alpha) * z,$$

where α is a trade-off hyper-parameter and z is the number of image patches in the largest connected component within the object o_j. We then rank all candidate objects based on their relevance scores and select the highest-scoring object as the predicted grounding.

To handle cases where GradCAM regions are scattered or minimal (e.g., a single patch), which suggests an ungroundable entity (e.g., *Republican* in Fig. 2), we introduce a threshold γ. If the highest relevance score is lower than γ, the entity is classified as ungroundable.

IV. EXPERIMENTS

A. Experimental Settings

Datasets. We conduct experiments on two benchmark datasets, i.e., Twitter-GMNER [1] and Twitter-FMNERG [10]. Table I shows the basic statistics of these datasets.

Evaluation Metrics. For the GMNER task, we follow previous studies [1], [10] by evaluating entity-type-object triplets using precision, recall, and F1 score. Specifically, for entity and type evaluation, predictions are considered correct only if they exactly match the ground truth. For object evaluation, if an entity is ungroundable, the prediction is correct if it is labeled as *None*. If the entity is groundable, the prediction is correct if the intersection over union (IoU) with the ground-truth bounding box exceeds 0.5. The correctness of each element is computed as follows:

$$C_e/C_t = \begin{cases} 1, & p_e/p_t = g_e/g_t; \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

$$C_o = \begin{cases} 1, & p_o = g_o = \text{None}; \\ 1, & \max(\text{IoU}_1, \ldots, \text{IoU}_j) > 0.5; \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

where C_e, C_t, and C_o refer to the correctness of entity, type, and object predictions, p_e, p_t, and p_o refer to the predicted entity, type, and object, g_e, g_t, and g_o refer to the gold entity, type, and object, and IoU_j means the IoU score between the predicted object p_o with the j-th annotated bounding box $g_{o,j}$.

Based on this, we adopt precision (Pre.), recall (Rec.), and F1 score as the evaluation metrics of the GMNER task:

$$\textit{correct} = \begin{cases} 1, & \text{if } C_e \text{ and } C_t \text{ and } C_o; \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

$$Pre = \frac{\textit{\#correct}}{\textit{\#predict}}, \quad Rec = \frac{\textit{\#correct}}{\textit{\#gold}} \tag{11}$$

$$F1 = \frac{2 \times Pre \times Rec}{Pre + Rec} \tag{12}$$

where *correct* refers to the number of predicted triples that match the gold triples, and *#predict* and *#gold* denote the number of predicted and gold triples.

For the EVG task, we focus on the accuracy of object predictions. Given the ground-truth entity and type, we consider the prediction correct if the bounding box has an IoU greater than 0.5 for groundable entities. For ungroundable entities, a *None* prediction is considered accurate.

Implementation Details. For our proposed visual annotation-free framework, we adopt three NER models, i.e., BARTNER [60], T5-Paraphrase [10], and RoBERTa-CRF [61] to extract named entities and their types. In our proposed ZeroEVG method, we employ three object detectors, i.e., VinVL [12], Detic [14], and YOLO11 [66] to detect the top-K objects, with K set to 36. For the CLIP model used in Section III-C2, we utilize the `clip-vit-base-patch32` model released by [11]. For the BLIP model used in Section III-C3, we follow [67] by using the `PNP-VQA-base` model and extracting the attention maps from its 8th layer. Regarding hyper-parameters, we perform a grid search over several combinations of the CLIP threshold β, top-P patches, relevance threshold γ, and trade-off weight α. The search space for β, P, γ, and α is in the range $[19, 24]$, $[16, 24]$, $[15, 19]$, and $[0.3, 0.9]$, respectively. After conducting the grid search on the validation set, we apply the best-found parameters to evaluate the performance on the test set. We run all the experiments on an NVIDIA RTX 3090 GPU.

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT METHODS FOR THE GMNER TASK ON TWITTER-GMNER AND TWITTER-FMNERG. NOTE THAT OUR ZEROEVG ALSO EMPLOYS LLMS IN THE CATEGORY FILTER STAGE WITH VERY LIMITED RESOURCE CONSUMPTION.

	Methods	Language Model	Object Detector	Twitter-GMNER			Twitter-FMNERG		
				Pre.	Rec.	F1	Pre.	Rec.	F1
Fully-Supervised	ITA-VinVL-EVG [1]	$BERT_{base}$	VinVL	52.37	50.77	51.56	43.05	42.51	42.78
	MNER-QG [37]	$BERT_{base}$	VinVL	53.02	54.84	53.91	–	–	–
	MQSPN [9]	$BERT_{base}$	VinVL	59.03	58.49	58.76	–	–	48.57
	GEM [42]	$BERT_{base}$+LLMs	VinVL	-	-	59.83	-	-	50.54
	RiVEG [6] (SOTA)	$XLMR_{large}$+LLMs	OFA_{large}	67.02	67.10	67.06	–	–	–
	H-Index [1]	$BART_{base}$	VinVL	56.16	56.67	56.41	46.83	46.28	46.55
	H-Index-YOLO [1]	$BART_{base}$	YOLO	54.35	53.96	54.15	45.72	45.20	45.45
	TIGER [10]	$T5_{base}$	VinVL	55.52	59.58	57.48	47.57	46.85	47.20
	TIGER-YOLO [10]	$T5_{base}$	YOLO	53.28	57.20	55.17	46.84	45.03	45.92
	MQSPN-RoBERTa	$XLMR_{large}$	VinVL	–	–	60.88	–	–	50.39
Text-Supervised	BARTNER-None [60]	$BART_{base}$	N.A.	44.58	44.75	44.66	36.83	37.28	37.05
	BARTNER-LLaVA	$BART_{base}$	LLaVA-7B	32.35	32.48	32.42	23.89	24.18	24.04
	BARTNER-GPT-4o	$BART_{base}$	GPT-4o	44.46	44.63	44.54	36.83	37.28	37.05
	BARTNER-OV-VG	$BART_{base}$	OV-VG	38.19	38.34	38.27	26.03	26.35	26.19
	BERTNER-ZeroEVG (Ours)	$BERT_{base}$	VinVL	57.84	56.98	57.41	46.95	45.65	46.29
	△ MQSPN			-1.19	-1.51	-1.35	-	-	-2.28
	BERTNER-ZeroEVG (Ours)	$BERT_{base}$	Detic	58.88	58.00	58.44	48.00	46.68	47.33
	BERTNER-ZeroEVG (Ours)	$BERT_{base}$	YOLO	60.92	60.01	60.46	48.28	46.95	47.61
	BARTNER-ZeroEVG (Ours)	$BART_{base}$	VinVL	58.25	58.47	58.36	49.03	49.63	49.33
	△ H-Index			+2.09	+1.80	+1.95	+2.20	+3.35	+2.78
	BARTNER-ZeroEVG (Ours)	$BART_{base}$	Detic	59.54	59.77	59.65	50.35	50.96	50.65
	BARTNER-ZeroEVG (Ours)	$BART_{base}$	YOLO	61.73	61.97	61.85	50.78	51.40	51.08
	T5-Paraphrase-ZeroEVG (Ours)	$T5_{base}$	VinVL	59.36	58.71	59.04	49.17	48.68	48.92
	△ TIGER			+3.84	-0.87	+1.56	+1.60	+1.83	+1.72
	T5-Paraphrase-ZeroEVG (Ours)	$T5_{base}$	Detic	60.28	59.61	59.94	50.00	49.51	49.75
	T5-Paraphrase-ZeroEVG (Ours)	$T5_{base}$	YOLO	62.43	61.74	62.08	50.75	50.26	50.50
	RoBERTaCRF-ZeroEVG (Ours)	$XLMR_{large}$	VinVL	60.93	61.07	61.00	50.62	51.36	50.99
	△ MQSPN-RoBERTa			-	-	+0.12	-	-	+0.60
	RoBERTaCRF-ZeroEVG (Ours)	$XLMR_{large}$	Detic	62.02	62.17	62.10	51.63	52.38	52.00
	RoBERTaCRF-ZeroEVG (Ours)	$XLMR_{large}$	YOLO	**64.18**	**64.33**	**64.26**	**52.17**	**52.93**	**52.55**

B. Comparison Systems

GMNER Methods. To show the effectiveness of our proposed visual annotation-free framework, we adopt a series of visual annotation-free (i.e., text-supervised) and fully-supervised methods as our comparison systems. First, we consider the following text-supervised methods: (1) *BARTNER-None* employs an index generation method named BARTNER [60] to identify entity-type pairs for the NER task and then sets the visual object prediction to the majority class *None*. (2) *BARTNER-LLAVA*, *BARTNER-GPT-4o*, and *BARTNER-OV-VG* are three variants of *BARTNER-None*, where the entity visual grounding model is replaced by *LLaVA [68]*, *GPT-4o* [62], and an open-vocabulary visual grounding model *OV-VG [69]*, respectively.

Additionally, we also consider a series of existing fully-supervised methods for comparison: (3) *ITA-VinVL-EVG* employs an image translation-based sequence labeling method ITA [34] for MNER and use VinVL-EVG [1] for visual grounding. (4) *MNER-QG* [37] is an end-to-end Machine Reading comprehension framework with Query Grounding for GMNER. (5) *H-index* [1] is a hierarchical index generation model that combines text with VinVL features to generate entity–type–object triplets in a hierarchical decoding process. (6) *H-index-YOLO* [1] is a variant of H-index where the original object detector VinVL is replaced with YOLO11 [66]. (7) *TIGER* [10] is a T5-based framework that formulates GMNER as a paraphrase generation task, taking both the text and VinVL-extracted visual features as input to generate the paraphrased entity–type–object sequences. (8) *TIGER-YOLO* [10] is a TIGER variant that replaces the original VinVL detector with YOLO11 [66]. (9) *MQSPN-RoBERTa* [9] is a unified framework which adaptively learns intra-entity relationships and establish inter-entity relationships from global optimal matching view. (10) *RiVEG* [6] is a pipeline method that uses LLMs as bridges to reformulate the GMNER task into a unified MNER-Visual Entailment-Visual Grounding task.

EVG Methods. The Entity Visual Grounding (EVG) task is a sub-task of GMNER. Given an entity name and its type, the goal of EVG is to determine whether the entity is visually present in the corresponding image and, if so, to provide its bounding box. First, we consider four zero-shot EVG methods as baselines, i.e., None, LLaVA [68], GPT-4o, and OV-VG [69], which are ablated models of the text-supervised baseline methods for GMNER. To show the advantage of our ZeroEVG method, we also compare it with a fully-supervised method VinVL-EVG [1], which encodes the input named entities and types with BERT [70] and integrates it with visual objects from VinVL [12] via a Cross-Modal Transformer to predict the entity presence in each object.

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT METHODS FOR THE ENTITY VISUAL GROUNDING (EVG) TASK.

Methods	Twitter-GMNER			Twitter-FMNERG		
	withbox	nobox	overall	withbox	nobox	overall
Fully-Supervised						
VinVL-EVG$_{\text{BERT}}$	36.23	89.71	67.71	37.95	83.86	67.95
Zero-Shot						
None	0.00	**100.00**	58.81	0.00	**100.00**	58.81
LLaVA	**54.49**	29.46	39.76	**53.92**	26.65	37.87
GPT-4o	0.96	96.19	57.12	0.96	96.19	57.12
OV-VG	3.06	85.10	51.36	6.21	66.93	41.96
ZeroEVG$_{\text{VinVL}}$	46.94	91.92	73.42	51.53	91.25	74.91
ZeroEVG$_{\text{Detic}}$	46.46	93.79	74.32	52.77	92.85	76.37
ZeroEVG$_{\text{Yolo}}$	52.68	93.19	**76.52**	53.25	94.12	**77.31**

C. Main Results

Results on the GMNER Task. We report the results of different GMNER methods in Table II. First, it is clear that our method significantly outperforms other text-supervised methods. Specifically, BARTNER-ZeroEVG with VinVL consistently outperforms the four BARTNER-based comparison methods by more than 10% in terms of F1 score across the two datasets. Second, when adopting the same language model and object detector, it is surprising that our proposed method performs better than the corresponding fully-supervised methods, i.e., H-index, TIGER, and MQSPN-RoBERTa. For instance, our BARTNER-ZeroEVG method outperforms H-index by 1.95% and 2.78% points in F1 score on Twitter-GMNER and Twitter-FMNERG datasets, respectively. Lastly, replacing the object detector VinVL with Detic or YOLO leads to further performance gains. Using YOLO as the object detector in RoBERTaCRF-ZeroEVG leads to an F1 score of 64.26% on Twitter-GMNER and 52.55% on Twitter-FMNERG, showing performance comparable to the state-of-the-art results achieved by RiVEG.It is worth noting that RiVEG relies on both text and visual supervisions and substantial training resource requirements, whereas our method only relies on the text supervision. In contrast, when H-index and TIGER are re-implemented with YOLO, their performance drops to 54.15% and 45.92% in F1 score on Twitter-GMNER and Twitter-FMNERG, respectively. This degradation mainly stems from their strong reliance on detector-specific visual features, while our ZeroEVG only leverages bounding boxes and labels, making it more robust to different detectors.

Impact of NER Backbones. We further analyze the impact of different textual NER models. RoBERTaCRF-ZeroEVG achieves the best overall results (64.26% and 52.55% F1 scores on the two datasets with YOLO), while T5-Paraphrase-ZeroEVG also surpasses TIGER by clear margins. Although BERTNER-ZeroEVG and BARTNER-ZeroEVG yield slightly lower absolute scores, they still outperform the fully-supervised counterparts under the same backbone and detector, such as ITA-VinVL-EVG, MNER-QG, and H-index. These findings verify the backbone-agnostic nature of ZeroEVG and highlight that more powerful NER backbones can further enhance GMNER performance.

TABLE IV
ABLATION STUDY ON THE EVG TASK WITH VINVL AS THE OBJECT DETECTOR.

Ablation	Twitter-GMNER			Twitter-FMNERG		
	withbox	nobox	overall	withbox	nobox	overall
Ours	46.94	**91.92**	**73.42**	51.53	**91.25**	**74.91**
w/o Category Filter	34.13	79.16	60.64	38.72	73.15	58.99
w/o Matching	**68.07**	32.06	46.87	**67.69**	42.42	52.81
w/o Localization	66.54	35.40	48.21	67.21	48.23	56.04
w/o Weighted Scorer	67.59	32.06	46.68	67.02	42.42	52.54

Results on the EVG Task. Table III shows the results of different EVG methods. First, it is evident that our ZeroEVG method outperforms all zero-shot baseline methods across both datasets. Second, ZeroEVG surpasses the fully-supervised method VinVL-EVG by 5.71% and 6.96% on the Twitter-GMNER and Twitter-FMNERG datasets, respectively. Finally, we split the test set into two subsets: one for groundable entities (i.e., withbox) and one for ungroundable entities (i.e., nobox). The results for each subset are reported separately. Notably, while most methods tend to predict entities as ungroundable, our ZeroEVG method achieves a balanced trade-off between the prediction of groundable and ungroundable entities.

V. IN-DEPTH ANALYSIS

A. Ablation Study

To assess the impact of each component of our proposed ZeroEVG method, we conduct ablation study and report the results in Table IV. First, we remove the Category Filter in Candidate Object Generation Module, which leads to a decline of 12.78% and 15.92% points on the two datasets. This shows that the Category Filter effectively removes irrelevant objects from unrelated categories, reducing noise in the subsequent Matching and Localization modules. Second, we remove the Matching Module and Localization Module separately. Specifically, after applying the Category Filter (if applicable), we directly select the bounding box with the largest GradCAM-highlighted region or the highest CLIP similarity. We observe that while this change improves accuracy on the groundable subset, it harms accuracy on the ungroundable subset, resulting in a decrease in overall performance. Finally, we remove the Weighted Scorer, which also results in a performance degradation. This demonstrates the contribution of the Weighted Scorer in refining the model's decision-making process.

B. Results on Different Entity Types

We further evaluate the performance across 8 coarse-grained entity types of the Twitter-FMNERG dataset, as shown in Table V. We find that our method, which operates without visual supervision, outperforms the fully supervised TIGER model on four of the entity types, including the Person type, which appears most frequently in this dataset. Across all entity types, our ZeroEVG method consistently performs better than any variant with an ablated module, demonstrating the effectiveness of the proposed approach.

TABLE V
F1 SCORES OF OUR T5-PARAPHRASE-ZEROEVG METHOD AND TIGER [10] ON 8 COARSE-GRAINED ENTITY TYPES WITH VINVL AS THE OBJECT DETECTOR ON TWITTER-FMNERG.

Methods	Art	Build.	Event	LOC	ORG	Other	PER	Prod.
TIGER	43.27	40.00	**48.39**	**67.69**	**46.75**	**48.28**	43.78	27.38
ZeroEVG	**45.45**	**40.76**	46.64	65.54	44.43	46.04	**48.78**	**37.80**
w/o Category filter	27.27	35.67	38.74	59.72	34.76	38.85	35.14	29.27
w/o Macthing	35.35	31.85	25.30	39.82	20.11	34.53	44.63	24.39
w/o Localization	37.37	35.67	30.04	43.19	21.04	34.53	45.98	25.61
w/o W-Scorer	35.35	31.85	25.30	39.82	20.11	34.53	44.17	24.39

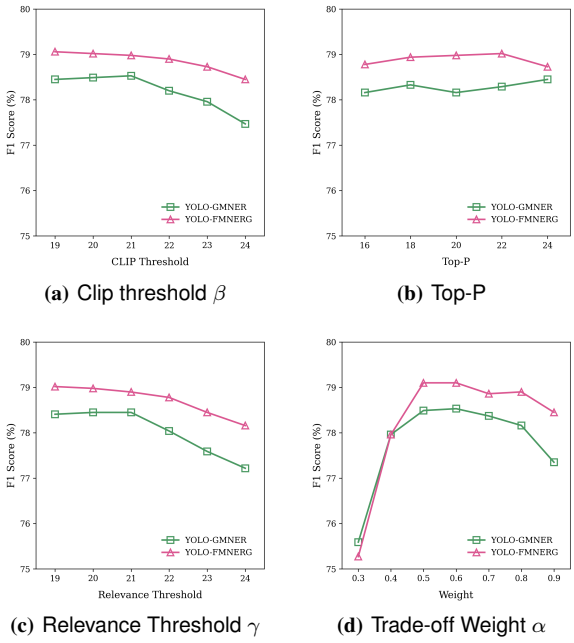
C. Sensitivity of Hyper-parameters

As shown in Fig. 3, we conduct a sensitivity analysis on the four hyper-parameters across the development set of two datasets. Note that since the performance trend of YOLO and that of the other two object detectors are similar, we only report the results of using YOLO as the object detector.

We observe that as the CLIP threshold increases beyond a certain point, the performance decreases. This can be attributed to the fact that when β is set too high, the model becomes overly restrictive, rigidly classifying most candidate boxes as mismatched, which negatively impacts performance. Similarly, as the relevance threshold γ increases, performance initially improves before slightly declining. When γ is too small, the entity matching criteria become overly lenient, potentially leading to the misclassification of ungroundable entities as groundable. Conversely, an excessively large γ may cause the model to skip objects that meet the matching requirements. For the top-P parameter, performance follows a similar trend, initially increasing before declining, though the overall fluctuation remains relatively small. Regarding the weight parameter α, performance also exhibits an increasing-then-decreasing pattern, with peak performance attained at 0.6 and 0.5 for the two datasets, respectively.

D. Case Study

To provide a clearer understanding of our method's decision process and highlight its advantages and the necessity of each module of ZeroEVG, we analyze four test cases from the Twitter-FMNERG dataset in Table VI. In case (a), the Matching module mistakenly identifies the man in the image *Tiger Woods* as *Lindsey Vonn*. However, the Localization module correctly highlights the relevant region, allowing our method to make an accurate prediction. In case (b), neither the Matching nor the Localization module alone can generate the correct result. However, our ZeroEVG method effectively integrates their predictions, enabling our method to reach the correct prediction. In case (c), while the Localization module fails to attend to the correct area, the Matching module assigns the highest score to *Hermione*, ultimately leading to the correct prediction. In case (d), while the Localization module correctly identifies the image as a game interface, its ability to highlight the relevant area is constrained by the top-P patches selected in the attention map. In this case, the Matching module compensates for this limitation, guiding the model for the right prediction.



(a) Clip threshold β (b) Top-P

(c) Relevance Threshold γ (d) Trade-off Weight α

Fig. 3. The sensitivity of different hyper-parameters on the development set of the EVG task with YOLO as the object detector.

VI. CONCLUSION

In this paper, we introduced a visual annotation-free framework for GMNER by leveraging text-only NER data and zero-shot entity visual grounding, eliminating the need for costly multimodal annotations. We proposed a novel method ZeroEVG, which integrates pre-trained VLMs and object detectors to achieve zero-shot visual grounding of entities. Experiments on two datasets show that our framework outperforms other text-supervised approaches and even surpasses fully-supervised methods using the same backbone models. These results demonstrate the potential of visual annotation-free paradigm for scalable GMNER applications.

While our proposed visual annotation-free framework demonstrates substantial improvements over existing methods, it does have several limitations. First, we did not incorporate contextual information in our ZeroEVG method, which may hinder the localization accuracy in more complex scenarios. Additionally, the threshold and weight parameters in our approach require grid search for optimal selection on the development set. We plan to propose an automatic parameter selection method without the need of development set in the future. Furthermore, our experiments were conducted solely on two benchmark datasets, and the method's effectiveness in other domains remains to be explored.

VII. APPENDIX

A. LLM Prompt Template for Category Filtering

For category filtering, we provide all object categories from the object detector along with entity types to GPT-4o, requesting the model to associate each entity type with the relevant object categories. We then manually review and refine the results to ensure accurate alignments.

Table VII presents the prompt template we used to perform category filtering with GPT-4o.

TABLE VI
CASE STUDY ON FOUR TEST SAMPLES OF TWITTER-FMNERG BASED ON VINVL. ✓ AND × ARE CORRECT AND INCORRECT PREDICTIONS.

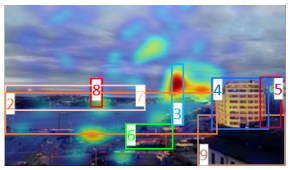
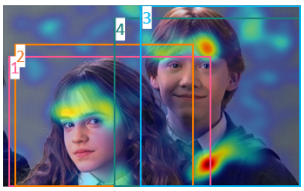
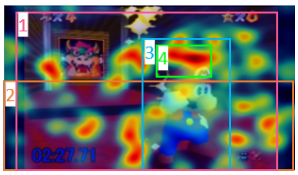
Impact of Localization Module	Impact of Weighted Scorer	Impact of Entity-Object Matching Module	
RT @ ReutersIndia : **Tiger Woods**$_{\text{athlete}}$ and Olympic skier **Lindsey Vonn**$_{\text{athlete}}$ break up	**Stockholm Sluice Area**$_{\text{park}}$ to be reconstructed : start 2016 , finish 2022 , cost 1beuro ! # Slussen # **Stockholm** $_{\text{city}}$ # DN @ SwedeninHR	(**Radio Times**$_{\text{news\_agency}}$) :Did you notice that # **Hermione**$_{\text{character}}$ waited six # **Harry Potter** books to give **Ron**$_{\text{character}}$ a . .	RT @ Kotaku : A new world record for beating **Super Mario 64**$_{\text{game}}$ with no stars
			
(a) Lindsey Vonn, athlete, Box-1	(b) Stockholm Sluice Area, park, None	(c) Hermione, character, Box-1	(d) Super Mario 64, game, Box-1
Filtered boxes: Box-2, 3, 4 ZeroEVG: Box-3 ✓ w/o Localization: Box-2 × w/o Matching: Box-3 ✓ w/o Weighted Scorer: Box-3 ✓	Filtered boxes: Box-2, 3, 4, 5, 6, 7, 8, 9 ZeroEVG: None ✓ w/o Localization: Box-2 × w/o Matching: Box-3 × w/o Weighted Scorer: Box-3 ×	Filtered boxes: Box-2, 3, 4 ZeroEVG: Box-2 ✓ w/o Localization: Box-2 ✓ w/o Matching: Box-3 × w/o Weighted Scorer: Box-3 ×	Filtered boxes: Box-2, 3, 4 ZeroEVG: Box-2 ✓ w/o Localization: Box-2 ✓ w/o Matching: Box-3 × w/o Weighted Scorer: Box-3 ×

TABLE VII
PROMPT TEMPLATE FOR CATEGORY FILTERING WITH GPT-4O

Instruction: Given a list of entity types (PER, LOC, ORG...) and the set of object categories detected by an open-vocabulary detector, assign each entity type to its most relevant object categories. The output should be in JSON format.
Objects: person, man, shirt, hair, letter, face, wall, tree, woman, window, ...

Answer: { "PER": ["man", "woman", "person", "girl", "boy", "spectator", "child", "lady", "baby", "couple", ...] }

REFERENCES

[1] J. Yu, Z. Li, J. Wang, and R. Xia, "Grounded multimodal named entity recognition on social media," in *Proc. Assoc. Comput. Linguistics*, 2023, pp. 9141–9154.

[2] X. Zhu, Z. Li, X. Wang, X. Jiang, P. Sun, X. Wang, Y. Xiao, and N. J. Yuan, "Multi-modal knowledge graph construction and application: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 2, pp. 715–735, 2022.

[3] W. Liang, P. D. Meo, Y. Tang, and J. Zhu, "A survey of multi-modal knowledge graphs: Technologies and trends," *ACM Computing Surveys*, vol. 56, no. 11, pp. 1–41, 2024.

[4] Y. Chen, H. Hu, Y. Luan, H. Sun, S. Changpinyo, A. Ritter, and M.-W. Chang, "Can pre-trained vision and language models answer visual information-seeking questions?" in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2023, pp. 14 948–14 968.

[5] H. Hu, Y. Luan, Y. Chen, U. Khandelwal, M. Joshi, K. Lee, K. Toutanova, and M.-W. Chang, "Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 12 065–12 075.

[6] J. Li, H. Li, D. Sun, J. Wang, W. Zhang, Z. Wang, and G. Pan, "Llms as bridges: Reformulating grounded multimodal named entity recognition," in *Proc. Assoc. Comput. Linguistics Findings*, 2024, p. 1302–1318.

[7] Z. Li, J. Yu, J. Yang, W. Wang, L. Yang, and R. Xia, "Generative multimodal data augmentation for low-resource multimodal named entity recognition," in *Proc. ACM Multimedia*, 2024, pp. 7336–7345.

[8] H. Ok, T. Kil, S. Seo, and J. Lee, "Scanner: Knowledge-enhanced approach for robust multi-modal named entity recognition of unseen entities," in *Proc. Assoc. Comput. Linguistics*, 2024, pp. 7718–7730.

[9] J. Tang, Z. Wang, Z. Gong, J. Yu, X. Zhu, and J. Yin, "Multi-grained query-guided set prediction network for grounded multimodal named entity recognition," *arXiv preprint arXiv:2407.21033*, 2024.

[10] J. Wang, Z. Li, J. Yu, L. Yang, and R. Xia, "Fine-grained multimodal named entity recognition and grounding with a generative framework," in *Proc. ACM Multimedia*, 2023, pp. 3934–3943.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[12] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, "Vinvl: Revisiting visual representations in vision-language models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5579–5588.

[13] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 9694–9705, 2021.

[14] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2022, pp. 350–368.

[15] S. Subramanian, W. Merrill, T. Darrell, M. Gardner, S. Singh, and A. Rohrbach, "Reclip: A strong zero-shot baseline for referring expression comprehension," in *Proc. Assoc. Comput. Linguistics*, 2022, pp. 5198–5215.

[16] H. Shen, T. Zhao, M. Zhu, and J. Yin, "Groundvlp: Harnessing zero-shot visual grounding from vision-language pre-training and open-vocabulary object detection," in *Proc. Conf. Assoc. Advancement Artif. Intell.*, 2024, pp. 4766–4775.

[17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, pp. 336–359, 2020.

[18] Y. Mo, J. Liu, H. Tang, Q. Wang, Z. Xu, J. Wang, X. Quan, W. Wu, and Z. Li, "Multi-task multi-attention transformer for generative named entity recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, 2024.

[19] J. Liu, D. Ji, J. Li, D. Xie, C. Teng, L. Zhao, and F. Li, "Toe: A grid-tagging discontinuous ner model enhanced by embedding tag/word relations and more fine-grained tags," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 177–187, 2022.

[20] Y. Fu, N. Lin, X. Yu, and S. Jiang, "Self-training with double selectors for low-resource named entity recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 1265–1275, 2023.

[21] H. Chang, H. Xu, J. van Genabith, D. Xiong, and H. Zan, "Joiner-bart: joint entity and relation extraction with constrained decoding, representation reuse and fusion," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 3603–3616, 2023.

[22] T. Qian, M. Zhang, Y. Lou, and D. Hua, "A joint model for named entity recognition with sentence-level entity type attentions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1438–1448, 2021.

[23] Q. Zhang, J. Fu, X. Liu, and X. Huang, "Adaptive co-attention network

for named entity recognition in tweets," in *Proc. Conf. Assoc. Advancement Artif. Intell.*, 2018.

[24] S. Moon, L. Neves, and V. Carvalho, "Multimodal named entity recognition for short social media posts," *arXiv preprint arXiv:1802.07862*, 2018.

[25] D. Lu, L. Neves, V. Carvalho, N. Zhang, and H. Ji, "Visual attention model for name tagging in multimodal social media," in *Proc. Assoc. Comput. Linguistics*, 2018, pp. 1990–1999.

[26] J. Yu, J. Jiang, L. Yang, and R. Xia, "Improving multimodal named entity recognition via entity span detection with unified multimodal transformer," in *Proc. Assoc. Comput. Linguistics*, 2020, pp. 3342–3352.

[27] Z. Wu, C. Zheng, Y. Cai, J. Chen, H.-f. Leung, and Q. Li, "Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts," in *Proc. ACM Multimedia*, 2020, pp. 1038–1046.

[28] M. Asgari-Chenaghlu, M. R. Feizi-Derakhshi, L. Farzinvash, M. Balafar, and C. Motamed, "Cwi: A multimodal deep learning approach for named entity recognition from social media using character, word and image features," *Neural Comput. Appl.*, pp. 1–18, 2022.

[29] B. Xu, S. Huang, C. Sha, and H. Wang, "Maf: a general matching and alignment framework for multimodal named entity recognition," in *Proc. ACM Int. Conf. Web Search Data Mining*, 2022, pp. 1215–1223.

[30] J. Wu, C. Gong, Z. Cao, and G. Fu, "Mcg-mner: A multi-granularity cross-modality generative framework for multimodal ner with instruction," in *Proc. ACM Multimedia*, 2023, pp. 3209–3218.

[31] F. Chen, J. Liu, K. Ji, W. Ren, J. Wang, and J. Chen, "Learning implicit entity-object relations by bidirectional generative alignment for multimodal ner," in *Proc. ACM Multimedia*, 2023, pp. 4555–4563.

[32] X. Chen, N. Zhang, L. Li, S. Deng, C. Tan, C. Xu, F. Huang, L. Si, and H. Chen, "Hybrid transformer with multi-level fusion for multimodal knowledge graph completion," in *Proc. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, 2022, pp. 904–915.

[33] J. Wang, Y. Yang, K. Liu, Z. Zhu, and X. Liu, "M3s: Scene graph driven multi-granularity multi-task learning for multi-modal ner," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 111–120, 2022.

[34] X. Wang, M. Gui, Y. Jiang, Z. Jia, N. Bach, T. Wang, Z. Huang, and K. Tu, "Ita: Image-text alignments for multi-modal named entity recognition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2022, pp. 3176–3189.

[35] X. Wang, J. Cai, Y. Jiang, P. Xie, K. Tu, and W. Lu, "Named entity and relation extraction with multi-modal retrieval," in *Findings of EMNLP*, 2022, pp. 5925–5936.

[36] J. Li, H. Li, Z. Pan, D. Sun, J. Wang, W. Zhang, and G. Pan, "Prompting chatgpt in mner: enhanced multimodal named entity recognition with auxiliary refined knowledge," in *Proc. EMNLP*, 2023.

[37] M. Jia, X. Shen, L. Shen, J. Pang, L. Liao, Y. Song, M. Chen, and X. He, "Query prior matters: A mrc framework for multimodal named entity recognition," in *Proc. ACM Multimedia*, 2022, pp. 3549–3558.

[38] X. Bao, M. Tian, Z. Zha, and B. Qin, "Mpmrc-mner: A unified mrc framework for multimodal named entity recognition based multimodal prompt," in *Proc. Int. Conf. Knowl. Manag.*, 2023, pp. 47–56.

[39] C. Cai, Q. Wang, B. Liang, B. Qin, M. Yang, K.-F. Wong, and R. Xu, "In-context learning for few-shot multimodal named entity recognition," in *Findings of EMNLP*, 2023, pp. 2969–2979.

[40] F. Chen and Y. Feng, "Chain-of-thought prompt distillation for multimodal named entity and multimodal relation extraction," *arXiv preprint arXiv:2306.14122*, 2023.

[41] S. Cui, J. Cao, X. Cong, J. Sheng, Q. Li, T. Liu, and J. Shi, "Enhancing multimodal entity and relation extraction with variational information bottleneck," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 1274–1285, 2024.

[42] Z. Wang, C. Zhu, Z. Zheng, X. Li, T. Xu, Y. He, Q. Liu, Y. Yu, and E. Chen, "Granular entity mapper: Advancing fine-grained multimodal named entity recognition and grounding," in *Findings of EMNLP*, 2024, pp. 3211–3226.

[43] J. Ye, J. Tian, M. Yan, X. Yang, X. Wang, J. Zhang, L. He, and X. Lin, "Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15 502–15 512.

[44] J. Deng, Z. Yang, D. Liu, T. Chen, W. Zhou, Y. Zhang, H. Li, and W. Ouyang, "Transvg++: End-to-end visual grounding with language conditioned vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13 636–13 652, 2023.

[45] L. Xiao, X. Yang, F. Peng, Y. Wang, and C. Xu, "Hivg: Hierarchical multimodal fine-grained modulation for visual grounding," in *Proc. ACM Multimedia*, 2024, pp. 5460–5469.

[46] R. Yao, S. Xiong, Y. Zhao, and Y. Rong, "Visual grounding with multi-modal conditional adaptation," in *Proc. ACM Multimedia*, 2024, pp. 3877–3886.

[47] S. Liu, S. Huang, F. Li, H. Zhang, Y. Liang, H. Su, J. Zhu, and L. Zhang, "Dq-detr: Dual query detection transformer for phrase extraction and grounding," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 2, 2023, pp. 1728–1736.

[48] L. Xiao, X. Yang, F. Peng, M. Yan, Y. Wang, and C. Xu, "Clip-vg: Self-paced curriculum adapting of clip for visual grounding," *IEEE Transactions on Multimedia*, vol. 26, pp. 4334–4347, 2023.

[49] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, "Mdetr-modulated detection for end-to-end multi-modal understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1780–1790.

[50] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang *et al.*, "Grounded language-image pre-training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10 965–10 975.

[51] Z. Yang, Z. Gan, J. Wang, X. Hu, F. Ahmed, Z. Liu, Y. Lu, and L. Wang, "Unitab: Unifying text and box outputs for grounded vision-language modeling," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2022, pp. 521–539.

[52] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang, "Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2022, pp. 23 318–23 340.

[53] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[54] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vl-bert: Pre-training of generic visual-linguistic representations," *arXiv preprint arXiv:1908.08530*, 2019.

[55] C. Li, H. Xu, J. Tian, W. Wang, M. Yan, B. Bi, J. Ye, H. Chen, G. Xu, Z. Cao *et al.*, "mplug: Effective and efficient vision-language learning by cross-modal skip-connections," *arXiv preprint arXiv:2205.12005*, 2022.

[56] K. Chen, Z. Zhang, W. Zeng, R. Zhang, F. Zhu, and R. Zhao, "Shikra: Unleashing multimodal llm's referential dialogue magic," *arXiv preprint arXiv:2306.15195*, 2023.

[57] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, Q. Ye, and F. Wei, "Grounding multimodal large language models to the world," in *Proc. 12th Int. Conf. Learn. Represent.*, 2024.

[58] G. Chen, L. Shen, R. Shao, X. Deng, and L. Nie, "Lion: Empowering multimodal large language model with dual-level visual knowledge," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2024, pp. 26 540–26 550.

[59] Z. Li, Q. Xu, D. Zhang, H. Song, Y. Cai, Q. Qi, R. Zhou, J. Pan, Z. Li, V. T. Vu *et al.*, "Groundinggpt: Language enhanced multi-modal grounding model," *arXiv preprint arXiv:2401.06071*, 2024.

[60] H. Yan, T. Gui, J. Dai, Q. Guo, Z. Zhang, and X. Qiu, "A unified generative framework for various ner subtasks," in *Proc. Assoc. Comput. Linguistics-IJCNLP*, 2021, pp. 5808–5822.

[61] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[62] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[63] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2022, pp. 12 888–12 900.

[64] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 618–626.

[65] L. He, Y. Chao, K. Suzuki, and K. Wu, "Fast connected-component labeling," *Pattern recognition*, vol. 42, no. 9, pp. 1977–1987, 2009.

[66] G. Jocher and J. Qiu, "Ultralytics yolo11," 2024. [Online]. Available: https://github.com/ultralytics/ultralytics

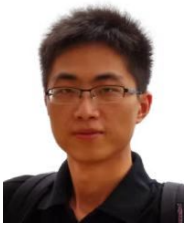
[67] A. M. H. Tiong, J. Li, B. Li, S. Savarese, and S. C. Hoi, "Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training," *arXiv preprint arXiv:2210.08773*, 2022.

[68] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Adv. Neural Inf. Process. Syst.*, vol. 36, 2024.

[69] C. Wang, W. Feng, X. Li, G. Cheng, S. Lyu, B. Liu, L. Chen, and Q. Zhao, "Ov-vg: A benchmark for open-vocabulary visual grounding," *Neurocomputing*, vol. 591, p. 127738, 2024.

[70] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 4171–4186.

Jia Yang received the B.S. degree from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2022. She is currently pursuing her M.S. degree at the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing. Her research interests include natural language processing, information extraction, and social media analysis.

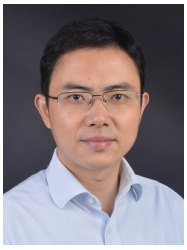


Jianfei Yu received the B.S. and M.S. degrees from Nanjing University of Science and Technology, Nanjing, China, in 2012 and 2015, respectively, and the Ph.D. degree from Singapore Management University, Singapore, in 2018. He is currently an associate professor in the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. His research interests include natural language processing, machine learning, and data mining.

Zilin Du is currently a third-year Ph.D. student at the College of Computing and Data Science, Nanyang Technological University (NTU), Singapore. His research interests include multi-modal learning, data synthesis, training with generated data, and knowledge distillation. He has published over ten papers in the areas of multi-modal learning and data mining.

Wenya Wang received the bachelor's degree from the School of Physical and Mathematical Sciences, Nanyang Technological University (NTU), Singapore, in 2014, and the Ph.D. degree from the School of Computer Science and Engineering, NTU, in 2018. She is currently an assistant professor in the School of Computer Science and Engineering, NTU. Her research interests include deep learning, logic reasoning, knowledge integration and their applications in natural language processing.

Li Yang received her Bachelor's Degree from Nanjing University of Science and Technology and her Master's Degree from Nanyang Technological University. She is currently pursuing her Ph.D. degree from Wee Kim Wee School of Communication and Information, Nanyang Technological University. Her research interests include natural language processing, multimodal sentiment analysis, and social media analysis.



Rui Xia received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences in 2011. He is currently a Professor in the School of Intelligence Science and Technology, Nanjing University, China. His research interests include natural language processing, data mining and affective computing. He has published more than 50 papers in top journals and conferences. His work on emotion-cause pair extraction has received the ACL2019 Outstanding Paper Award.

Boyang Li received the Ph.D. degree from Georgia Institute of Technology, Atlanta, GA, USA, in 2015. He is currently working as an Associate Professor with the College of Computing and Data Science, Nanyang Technological University. His research interests include multimodal learning, data-centric AI, computational narrative intelligence, and machine learning. His research work has been covered by major media outlets such as Engadget, TechCrunch, New Scientist, and National Public Radio.