Visual Knowledge-Enhanced LLaVA for Fine-Grained Multimodal Named Entity Recognition and Grounding

Li Yuan, Yi Cai, *Member, IEEE*, Bingshan Zhu, Zhenghao Liu, *Student Member, IEEE*, Zikun Deng, Qing Li, *Fellow, IEEE*, and Tao Wang

***Abstract*—The rapid growth of multimodal data has highlighted the significance of Fine-Grained Multimodal Named Entity Recognition and Grounding (FMNERG), which focuses on extracting entities and their corresponding groundings from image-text pairs. Existing approaches typically extract entities and then associate them with entity groundings using object detection methods. However, these approaches face challenges, including the use of diverse multimodal feature representations and insufficient visual knowledge, which hinder their ability to effectively link entities to images and limit overall performance. To address these limitations, we propose the visual knowledge-enhanced LLaVA (VKEL) framework, a two-stage model designed to integrate visual knowledge with multimodal learning. In the first stage, VKEL improves entity recognition by augmenting datasets with synthetic image-text pairs and optimizing alignment through lightweight fine-tuning. In the second stage, VKEL overcomes grounding limitations by incorporating consistent and accurate visual knowledge from large language models and utilizing object annotations to guide entity identification within images. This stage enhances the model's ability to disambiguate similar entities and improve the precision of entity grounding. Extensive experiments on the FMNERG benchmark demonstrate that VKEL surpasses the SOTA by 10.24% in F1 score, with significant improvements in both fine-grained entity recognition and entity grounding performance.**

***Index Terms*—Few-shot learning, multimodal learning, entity-relation extraction, low-rank adaptation.**

I. Introduction

With the increase in multimodal data on the internet (such as Twitter or Facebook) over recent years [1], [2], multimodal named entity recognition (MNER) tasks have received widespread attention [3], [4], [5]. As one of the most important sub-tasks in multimodal information extraction [6], [7], MNER aims to extract named entities and their corresponding types from image-text pairs. As shown in Fig. 1, given the text and corresponding image, the model aims to identify the entities mentioned in text and their corresponding entity types, e.g., (*Kevin Love*, **Person**) and (*J. R. Smith*, **Person**).

However, extracting only the entity-type pair is insufficient for multimodal knowledge graph construction, as it lacks the associated visual elements in images corresponding to the entity. This hinders the application of tasks such as scene understanding [8], which require textual and visual data to fully comprehend the context. Inspired by the visual grounding task [9], [10], [11], [12], [13], which establishes explicit associations between images and text. Yu et al. [14] proposed the grounded named entity recognition (GMNER) task [15] to extract textual named entities, their corresponding entity types, and entity grounding (EG) from image-text pairs. As shown in Fig. 1, the EG task requires identifying and linking *Kevin Love* and *J. R. Smith* to their corresponding *Object 2* and *Object 1* from the given image. For entities not present in the image, **None** should be used to indicate their absence.

Besides, both MNER and GMNER have focused on assigning coarse-grained types to an entity, which often lack the specificity needed to adequately describe the entities. This limitation can lead to inefficiencies and inaccuracies in downstream tasks such as information retrieval. For example, in a knowledge graph constructed using such coarse-grained types, searching for (*Thomas Lee*, **Person**) can yield mixed information related

Received 27 July 2025; revised 29 November 2025; accepted 30 December 2025. Date of current version 28 January 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62476097 and Grant 62402184, in part by the Fundamental Research Funds for the Central Universities, South China University of Technology, under Grant x2rjD2250190, in part by the Science and Technology Planning Project of Guangdong Province under Grant 2025B0101120003, in part by the Guangdong Provincial Fund for Basic and Applied Basic Research—Regional Joint Fund Project (Key Project) under Grant 2023B1515120078, in part by the Guangdong Provincial Natural Science Foundation for Outstanding Youth Team Project under Grant 2024B1515040010, and in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515010162. The work of Tao Wang was supported in part by NIHR Maudsley Biomedical Research Centre, in part by Maudsley Charity, in part by King's Together, in part by MHaPS Early Career Researcher Award, and in part by MRC DATAMIND. The associate editor coordinating the review of this article and approving it for publication was Dr. Boyang Li. *(Corresponding author: Yi Cai.)*

Li Yuan, Yi Cai, Zhenghao Liu, and Zikun Deng are with the School of Software Engineering, South China University of Technology, Guangzhou 510640, China, and also with the Key Laboratory of Big Data and Intelligent Robot, Ministry of Education, South China University of Technology, Guangzhou 510640, China (e-mail: seyuanli@mail.scut.edu.cn; ycai@scut.edu.cn; zkdeng@scut.edu.cn).

Bingshan Zhu is with the School of Information Science, Guangdong University of Finance and Economics, Guangzhou 510320, China (e-mail: bsz@gdufe.edu.cn).

Qing Li is with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong (e-mail: qing-prof.li@polyu.edu.hk).

Tao Wang is with the Department of Biostatistics and Health Informatics, King's College London, WC2R 2LS London, U.K. (e-mail: tao.wang@kcl.ac.uk).

Digital Object Identifier 10.1109/TASLPRO.2026.3651977

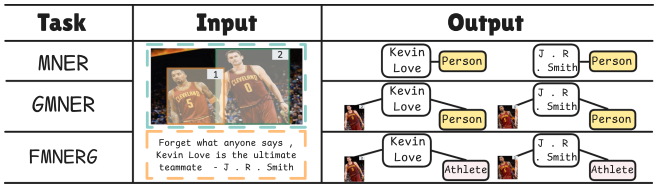

Fig. 1. The difference between MNER, GMNER, and FMNERG tasks.

TABLE I
THE ENTITY GROUNDING RESULT OF TWO WIDELY ADOPTED OBJECT DETECTION METHODS

Methods	Top-2	Top-5	Top-10	Top-15	Top-20
Anderson et al. [12]	30.47	55.74	75.7	82.93	86.60
Lyu et al. [10]	65.57	74.25	77.92	80.97	84.33

to both **musician** and **businessman**[1] entities. Thus, fine-grained entity classification is required to enhance the precision of information retrieval. To address this, Wang et al. [16] proposed the fine-grained multimodal named entity recognition and grounding (FMNERG) task, which refines coarse-grained entity types into a broader set of fine-grained categories and has garnered significant attention. A typical method for FMNERG involves using language models like T5 [17] to obtain entity-type pairs from the input text and image. These pairs are then matched to a relevant object from pre-extracted candidate objects using object detection methods [12]. However, this approach has three major limitations:

1) **Inconsistency between Visual and Textual Representations.** Visual and textual representations are derived from different pre-trained models and datasets. Research has shown that blindly injecting visual objects into a language model can disrupt textual feature representations, negatively affecting MNER performance [18], [19] and ultimately reducing the accuracy of the FMNERG task. This is particularly evident in types with sparse data, as the model struggles to effectively capture the latent features between text and images from the limited samples. Imbalanced entity distribution is common in fine-grained entity extraction, where entity types frequently exhibit a long-tail distribution. Besides, the EG task requires the model to primarily focus on image annotations, which poses a challenge to a model's visual representation capabilities [20], [21], [22]. Therefore, the text-image representations create a significant semantic gap between MNER and EG tasks, making cross-modal semantic alignment more challenging, and affecting the overall performance of FMNERG.

2) *Error Propagation from Object Detection.* Existing methods typically treat object detection as a preprocessing step to extract potential visual objects from images, which are then matched to textual entities. However, this approach is prone to error propagation: if the correct entity object is missing from the pre-extracted candidate objects, subsequent entity grounding results are compromised. To illustrate this, we analyze the matching between ground truth objects and candidate objects

[1]Wikipedia: Thomas Lee (born Thomas Lee Bass; October 3, 1962) is an American musician, while Thomas Haskell Lee (March 27, 1944 – February 23, 2023) was an American businessman, financier, and investor.

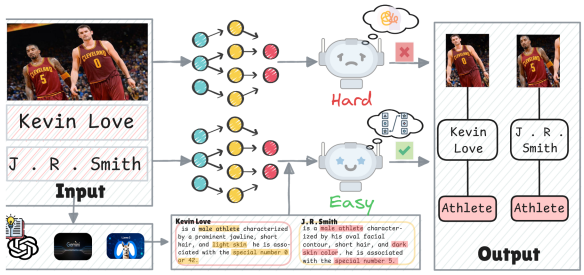

Fig. 2. Illustration of visual knowledge for given example.

and evaluated performance across top-K entities in Table I. Even when using the state-of-the-art object detection models [10], [12], the matching accuracy is only 77.92% for the top-10 candidates and improves to 86.6% for the top-20. Such detection errors inevitably propagate, impacting downstream tasks.

3) *Lack of Knowledge Bridging between Textual Entities and Visual Objects.* While object detection methods can effectively distinguish different types of entities (e.g., a *person* playing *basketball*), they often fail to distinguish multiple entities of the same type (e.g., two *persons* running together) and ground each entity to corresponding visual objects in an image. This is partly because of the limited availability of large-scale training data that captures fine-grained visual features of such entities. To address this, we consider utilizing visual knowledge from LLMs like GPT-3.5 Turbo [23], Gemini Pro [24], and LLaMA-3 70B [25] as a knowledge bridge. For example, as shown in Fig. 2, knowing specific attributes of entities (e.g., *Kevin Love's characteristics and jersey numbers*) could aid the model in solving the EG task. However, hallucinations in LLMs [26] may result in missing or inaccurate visual knowledge, as seen in Fig. 5, where LLaMA-3 misidentified *J.R. Smith as a female actor*, with similar errors from GPT-3.5 Turbo and Gemini Pro.

Besides, with the recent advances in multimodal pretraining, some language model-based multimodal models, such as LLaVA [27], have demonstrated the ability to effectively integrate image features while inheriting the cognitive capabilities of language models. Although LLaVA has made progress in addressing the limitations mentioned above, it primarily focuses on image understanding [27], [28]. However, a gap remains between image understanding and entity grounding [22]. Specifically, image understanding emphasizes comprehending the overall content, context, and semantics of an image based on salient features (e.g., identifying a "sunset on the beach"), entity grounding requires precise mapping between fine-grained objects and linguistic entities (e.g., mapping "sun" and "beach" to specific elements in the image). This limitation restricts LLaVA's ability to advance in the FMNERG task.

This paper proposes a visual knowledge-enhanced LLaVA (VKEL) model to tackle the FMNERG task, as shown in Fig. 3. This model aims to improve LLaVA's grounding capabilities by incorporating external visual knowledge about entities. Specifically, we divide the FMNERG task into two fundamental stages. In the first stage, we focus on achieving better coarse-grained entity extraction to enhance entity grounding performance and improve the quality of visual knowledge related to entities,

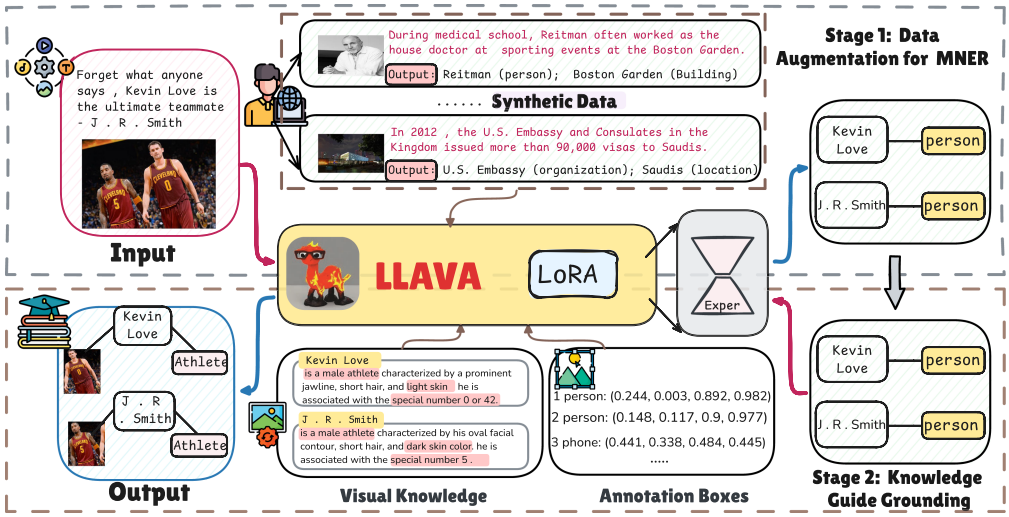

Fig. 3. The proposed VKEL architecture for FMNERG consists of two stages designed to activate and enhance the LLaVA model's abilities in MNER and EG, thereby improving the handling of FMNERG: (a) Data Augmentation for MNER, which introduces a low-cost synthetic data generation method to boost MNER performance; and (b) Knowledge-Guided Grounding, which aims to enhance LLaVA's entity grounding capabilities by incorporating visual entity knowledge and object annotations, ultimately addressing the challenges of FMNERG.

particularly for minority entity types that have a small number of samples, such as *Product*, *Art*, and *Event* in the FMNERG dataset. However, existing MNER datasets lack fine-grained annotations linking textual entities to their corresponding visual objects. Collecting high-quality MNER data requires image-text matching, a time-consuming and labor-intensive process. Thus, we propose a low-cost data augmentation method to enhance MNER performance by constructing synthetic image-text pairs from existing text-only fine-grained NER datasets. We retrieve images from the Internet and filter out irrelevant ones using a CLIP [29], resulting in effective and relevant image-text pairs.

In the second stage, we introduce a multimodal consistency method to enhance the accuracy and robustness of visual knowledge extracted across multiple LLMs, thereby improving the model's performance in the EG task. By leveraging knowledge from multiple LLMs, we aggregate consistent visual information associated with entities, effectively reducing hallucinations that may occur when relying solely on a single model [30]. This refined visual knowledge is subsequently incorporated into the downstream fine-tuning process, further bolstering the model's EG capabilities. Comprehensive experiments conducted on the FMNERG dataset demonstrate the superiority of our approach. The main contributions of this study are summarized as follows:

- We propose a visual knowledge-enhanced LLaVA framework, a two-stage model that integrates visual knowledge with multimodal learning to enhance performance on the FMNERG task. The first stage introduces a cost-effective synthetic data augmentation method to improve the extraction performance for minority entity types, which have limited sample sizes. The second stage features a knowledge-guided grounding approach that bridges the gap between textual entities and visual objects by incorporating visual knowledge from LLMs.
- We propose a multimodal consistency mechanism to obtain accurate and consistent visual knowledge about entities using multiple LLMs. This mechanism enhances visual grounding by integrating visual prior knowledge with object annotation data, significantly improving entity disambiguation and grounding precision.
- Experiments on the FMNERG benchmark dataset demonstrate that our proposed method achieves state-of-the-art performance on the FMNERG task and its two subtasks.

The remaining sections of this article are organized as follows: Section II reviews related work, Section III gives a detailed description of the proposed model, Section IV presents our experimental results, and Section V concludes the article.

II. Related Work

A. Multimodal Name Entity Recognition

Recently, the rapid growth of multimodal content on social media has led to a significant increase in attention towards multimodal named entity recognition (MNER). MNER enables the extraction of potential entity cues from visual information. Existing MNER approaches can generally be categorized into three main groups: (1) Designing various attention-based mechanisms to model interaction modalities [3], [4], [5], [31], [32], [33]. For instance, Lu et al. [3] utilize a BiLSTM network [34] to model contextual information, followed by a CRF layer [35] for entity-type identification. Yu et al. [4] introduce a multimodal interaction module based on a Transformer architecture, which generates word representations that incorporate visual information, alongside a text-based entity span detection module to improve prediction accuracy. Additionally, Zhao et al. [33] propose a relation-enhanced graph convolutional network for MNER, constructing inter-modal and intra-modal relation graphs to capture relevant image information for each entity-name pair effectively. (2) Using image captions to bridge the cross-modal semantic gap [36], [37], [38]. For example, Wang et al. [37] extract image captions, optical characters, and object names from images as auxiliary visual context, thereby facilitating the alignment of image features with the textual representation.

(3) Incorporating pretrained object detectors to enhance textual representations [39], [40]. For instance, Zheng et al. [39] propose a bilinear attention network with adversarial learning to align visual objects with entities, employing a bilinear attention module to exclude irrelevant objects. Chen et al. [40] introduce a method called hierarchical visual prefix network, which integrates hierarchical multi-scale visual features as visual prefixes to enhance textual representations and improve robustness against irrelevant visual inputs in addressing the MNER task.

The MNER task often seeks to improve the accuracy of NER models, which assign entities in text to predefined types or categories, by incorporating visual information. However, most MNER methods have limited capacities for visual understanding and fail to establish a fine-grained alignment between linguistic/textual entities and visual objects. This limitation hinders their applications in tasks requiring precise and context-aware multimodal understanding and reasoning, such as visual question answering, autonomous and virtual reality systems.

B. Multimodal Named Entity Recognition and Grounding

1) Coarse-Grained Multimodal Named Entity Recognition and Grounding: To establish explicit associations between text and visual information, Yu et al. [14] propose the grounded multimodal named entity recognition (GMNER) task to ground textual entities to their corresponding visual objects. Given its early stage of development, most GMNER methods draw heavily from visual understanding techniques [12], [13], [15], [41], [42], which aim to establish explicit associations between images and text. Existing visual understanding methods can be categorized into two types: single-stage and two-stage approaches. The former focuses on leveraging recent end-to-end object detection models [9], [10], [11], [43]. For example, Tang et al. [42] propose an end-to-end, query-guided network that employs learnable queries and a query-guided fusion module to model multimodal entity relationships and improve alignment in GMNER. While the latter first uses object detection models to generate region candidates and then ranks them based on their relevance to a textual query [12], [13]. Wang et al. [43] introduce a granular entity-mapping framework that integrates multi-granularity entity recognition with an LVLM-based reranking module. The approach further leverages a pretrained LVLM as an implicit visual grounder to enhance multimodal entity extraction.

However, visual grounding tasks primarily focus on the relationship between entity types (e.g., *person* and *organization*) and image objects, while entity grounding further emphasizes the direct connection between entity descriptions and image objects. As shown in Fig. 1, the task requires identifying the image objects *Object 2* and *Object 1* that correspond to *Kevin Love* and *J. R. Smith* in a given image. Therefore, traditional visual grounding methods struggle to establish more connections between entity descriptions and image objects.

2) Fine-Grained Multimodal Named Entity Recognition and Grounding: To obtain more precise information about entities, Wang et al. [16] extended the GMNER task by introducing fine-grained entity types. They further proposed a new task, Fine-grained Multimodal Named Entity Recognition and Grounding (FMNERG), which jointly addresses entity recognition and visual grounding at a fine-grained level. This task has recently attracted increasing attention from the research community [42], [43].

Existing methods typically convert images into feature representations and use language models, such as T5 [17] or BERT [44], to obtain entity-type pairs. These are followed by object detection or visual grounding techniques, such as Mask-RCNN [12], to extract a candidate object set from the image. Finally, sorting is applied to identify the object most relevant to the entity within the candidate set, yielding the final grounding box result. However, this approach faces several inherent limitations: (1) inconsistencies between visual and textual representations; (2) error propagation from object detection methods; (3) the challenge of effectively grounding entities to visual objects, when entities of the same type are involved.

To address these challenges, we introduce three key improvements. First, we build a visual knowledge-enhanced model based on LLaVA [45], a large multimodal model designed for general-purpose visual and language understanding, to achieve a unified representation of text and images. Second, we adopt a generative approach for producing the final object grounding results, effectively mitigating the error propagation issue seen in traditional methods. Third, we enhance LLaVA's grounding capabilities by incorporating a visual knowledge-enhanced method, which integrates visual knowledge from multiple LLM outputs via a multimodal consistency mechanism, thereby achieving accurate and fine-grained visual representations.

C. Multimodal Pre-Trained Models

One of the most promising areas in MENR is using multimodal pre-trained models to get aligned representations across multi-modal data. These models can be divided into two main paradigms: discriminative and generative approaches [29], [46], [47], [48], [49]: (1) Discriminative models [29], [46], like CLIP [29], align visual and textual modalities using contrastive learning, projecting them into a shared representation space for tasks like image classification and retrieval. These models, however, struggle with complex reasoning due to their reliance on multi-modal direct associations. (2) Generative models [47], [48], [49], such as OFA [49], unify multimodal tasks within a sequence-to-sequence framework, offering flexibility in tasks like image captioning and visual question answering. While generative models can handle a broad range of tasks, they may not excel in reasoning. Recent advances in LLMs [25], [50], [51], [52] suggest that using LLMs as backbone can provide a potential solution to the reasoning limitations of generative models, which leads to the development of multimodal large language models (MLLMs) [23], [24], [53], [54]. These models effectively leverage the pre-trained knowledge of each modality, avoiding the computational costs associated with training from scratch while inheriting the cognitive capabilities of LLMs, and have demonstrated impressive performance in tasks such as visual understanding and visual question answering. A typical example of MLLMs is LLaVA [53], an open-access MLLM built on LLaMA-3 as its language backbone, with CLIP serving as the

image encoder. LLaVA achieves state-of-the-art performance across various visual-language understanding tasks [55], [56], including visual question answering [55], all while maintaining a comparable number of parameters.

Although LLaVA demonstrates strong visual understanding capabilities, its limitations in entity grounding [22] hinders its performance in the FMNERG task. To overcome this, we propose a two-stage framework built on LLaVA. In the first stage, we introduce a low-cost synthetic data augmentation method to improve MNER performance. This method generates high-quality synthetic image-text pairs using existing fine-grained NER datasets and applies CLIP-based filtering to ensure robust alignment between the visual and textual modalities. In particular, it enhances the model's ability to learn from sparse types. In the second stage, we propose a multimodal consistency mechanism to enhance LLaVA's entity grounding capabilities. This mechanism refines visual knowledge by integrating insights from multiple LLMs, improving visual grounding accuracy by combining consistent visual priors and object annotation data.

III. Method

Our proposed VKEL model is a two-stage, multi-LoRA framework designed to address the FMNERG task, as illustrated in Fig. 3. We first provide a brief overview of the FMNERG task formulation and introduce the main backbone, LLaVA, followed by a detailed description of our method, VKEL, which consists of two fundamental stages. In Stage 1, a dedicated LoRA module fine-tunes LLaVA on augmented multimodal data, generated by pairing text-only NER datasets with synthetic images, enabling the extraction of coarse-grained entities and their types from both text and images. In Stage 2, a second LoRA module refines entity grounding by incorporating visual knowledge. This knowledge is obtained through multi-LLM consistency, guided by carefully designed templates, and further enhanced with object annotations, allowing for more precise alignment between entities and their corresponding visual representations.

A. Task Definition

The FMNERG task is defined as follows: given an input text with n words $w = (w_1, w_2, \ldots, w_n)$ and a corresponding image v, the goal is to extract a set of entity-type-object triples $y = \{e_c, t_c, b_c\}_{c=1}^{C}$, where e_c, t_c, and b_c represent the c-th triple. The entity denotes one of the named entities in the text w and t_c is its corresponding type, selected from predefined categories such as *Politician*, *Sports Team*, or *City*. The object b_c represents the grounded visual object corresponding to e_c in the image. If there is no corresponding object, b_c is assigned as *None*. Otherwise, b_c is defined as the bounding box of the visual object, consisting of the top-left and bottom-right coordinates.

B. Llava

Since our proposed VKEL model is built upon LLaVA as a key backbone, we briefly introduce it. LLaVA [45] is a state-of-the-art multimodal model designed to bridge visual and textual modalities. It achieves this by aligning a pretrained vision encoder, such as CLIP [29], with a large language model, such as LLaMA [25], using high-quality image-text alignment data, represented as follows:

$$\begin{aligned} Y &= \text{LLaMA}([\text{H}_V, \text{H}_T]) \\ \text{H}_V &= W \cdot \text{CLIP}(X_V) \end{aligned} \tag{1}$$

where H_T and X_V denote the input text embedding and image, respectively, while W represents a trainable projection layer used to align visual features with text embeddings. LLaVA excels in tasks requiring cross-modal reasoning and contextual understanding by leveraging its multimodal pretraining to generate intelligent responses to visual and textual queries. However, it struggles with FMNERG tasks, which require fine-grained differentiation of visually similar entities or the integration of domain-specific visual knowledge.

C. Visual Knowledge-Enhanced LLaVA

The proposed VKEL model is a two-stage framework that incorporates multiple Low-Rank Adaptation (LoRA) modules to enhance LLaVA's task-specific capabilities, leading to substantial improvements on the FMNERG benchmark. Specifically, a dedicated LoRA module fine-tunes LLaVA on augmented multimodal data, generated by pairing text-only NER datasets with synthetic images, to extract coarse-grained entities and their types from both text and images. These preliminary outputs provide natural-language descriptions of each entity, forming the foundation for the next stage. In the second stage, a second LoRA module incorporates visual knowledge, obtained through multi-LLM consistency and guided by designed templates, along with object annotations to refine entity grounding. This stage produces fine-grained entity types and absolute bounding boxes for each detected entity.

1) Data Augmentation for MNER: In the first stage, the objective is to extract potential entities and their corresponding coarse-grained types from the given input text and image. As shown in Fig. 3, we identify the entities *Rosie Donnell* and *Michelle Obama* with the corresponding type **person** from the sentence: *Rosie Donnell lets loose on "s * * t stain" Michelle Obama*. In a data-driven deep learning method, a common approach is to augment the training dataset [57], [58], [59]. However, obtaining paired MNER data is time-consuming and labor-intensive, as both image and text pairs are required. Thus, we propose a low-cost data augmentation method that enhances MNER performance by constructing high-quality synthetic image-text pairs from existing text-only fine-grained NER datasets [60]. The overall process is illustrated in Fig. 4.

We first map the fine-grained labels of the dataset to the predefined coarse-grained categories, aligning them with the MNERG dataset. To generate synthetic images, we input both the text $\tilde{w}$ and the target entities into the Google API,[2] a cost-effective approach to retrieve reasonably relevant images. We obtain a set of candidate images $\widetilde{v} = \{\widetilde{v}^1, \widetilde{v}^2, \ldots, \widetilde{v}^5\}$ that are potentially related to the input text. To filter out candidate images with low relevance, we employ CLIP [61] to compute the similarity

[2]https://images.google.com/

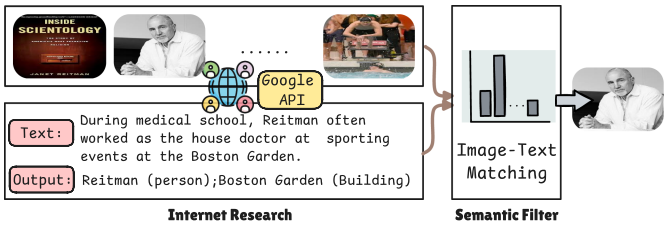

Fig. 4. The process of constructing low-cost synthetic MNER datasets.

between the text and synthetic images. The process is described as follows:

$$q = \text{CLIP}_T(\tilde{w})$$
$$z^i = \text{CLIP}_V(\tilde{v}^i)$$
$$\alpha = \underset{m\in\{1,2,\cdots 5\}}{\arg\text{Top1}} \frac{(q)^T z^i}{\|q\|_2 \|z^i\|_2} \tag{2}$$

where CLIP_T and CLIP_V represent the text and image encoders of CLIP, respectively. Additionally, we exclude samples with $\alpha < 0.3$ to further filter out those with lesser correlation between text and synthetic image. Thus, we obtain a large set of synthetic text-image pairs.

We conjunct the original training data D_o^1 with synthetic data D_s as the final training data for stage 1 $D^1 = \{D_o^1; D_s\}$. Thus, we obtain numerous synthetic text-image pairs with a certain degree of correlation. We perform LoRA fine-tuning on these synthetic pairs in conjunction with the original training data:

$$O^1 = \text{LVLMs}(D^1, \text{LoRA}^1) \tag{3}$$

where O^1 denotes the target output of the first stage, expressed in natural languages, such as *Kevin Love is a person* in Fig. 3. The LoRA represents the LoRA, an efficient method for fine-tuning. The LoRA module introduces a low-rank decomposition of a linear layer in LVLMs. The original linear transformation, $h = W_0 x$, is modified as follows:

$$h = \text{LoRA}(x) = W_0 x + BAx \tag{4}$$

where $A \in \mathbb{R}^{r\times d_{in}}$ and $B \in \mathbb{R}^{d_{out}\times r}$ are low-rank matrices, with $r \ll \min(d_{out}, d_{in})$. The number of trainable parameters in A and B is much smaller than that in W_0. During training, only A and B are updated.

D. Knowledge Guided Grounding

In the second stage, we leverage visual knowledge to enhance LLaVA's entity grounding ability for the FMNERG task. Interestingly, some studies suggest that large language models possess visual knowledge related to keywords, even without the input of relevant images [62], [63]. Therefore, we first focus on customizing prompts to encourage the large language models to generate the most distinguishable visual knowledge for each coarse-grained entity type. Then, we introduce the multi-LLM consistency visual knowledge approach, which utilizes consistent visual knowledge outputs from different models to improve the accuracy and robustness of the knowledge. Finally, we use this knowledge and the outputs from the first stage to fine-tune

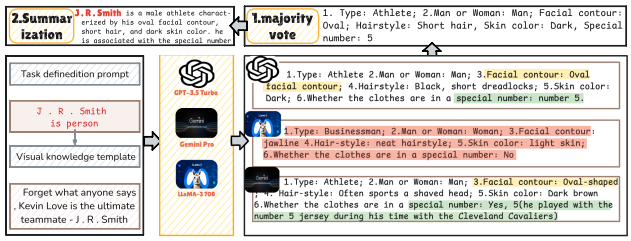

Fig. 5. The process of obtaining visual knowledge by using the proposed multi-LLM consistency.

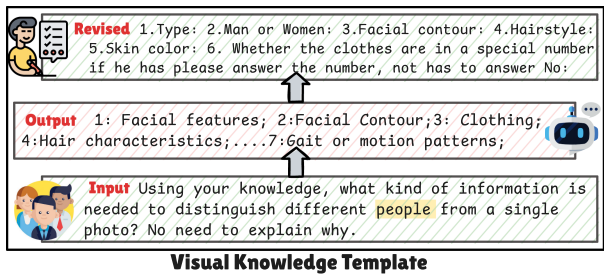

Fig. 6. The process of the designed visual knowledge template for **person** entity type.

LLaVA, encouraging the model to effectively integrate visual knowledge for the FMNERG task.

1) Design of Visual Knowledge Templates: Since the relevance and usefulness of visual knowledge obtained from large language models are highly dependent on the input prompt, we have tailored visual knowledge templates, denoted as $\text{KT}_{(i)}$, for the i-th coarse-grained entity type. These templates are designed to obtain more effective visual knowledge from LLMs for linking entities to images, based on the output O^1 from Stage 1. The visual knowledge consists of two components: the first encourages the large language model to predict the fine-grained types of entities, while the second incorporates explicit appearance knowledge to bridge the gap between entities and their grounding. The process is as follows: First, we use ChatGPT to find the most relevant visual attributes that distinguish different coarse-grained entity types. These attributes are then manually revised to create the final prompt template. This process is illustrated in Fig. 6, which presents the template used to extract visual knowledge for the **Person** entity. Visual knowledge templates for other coarse-grained types follow a similar approach. For instance, the **Building** template emphasizes *color* and *appearance*, while the **Product** template focuses on *trademark* and *appearance features*.

2) Multi-LLM Consistency Visual Knowledge: Since hallucinations in LLMs [26] may cause the omission or generation of inaccurate visual knowledge associated with entities, we draw inspiration from the use of self-consistency to enhance LLM reasoning abilities [64]. Building on this idea, we propose a multi-LLM consistency approach to generate relevant and stable visual knowledge by leveraging multiple language models. The process is shown in Fig. 5. First, we use customized coarse-grained visual templates $\text{Entity}_{(i)}$ to extract knowledge from different LLMs for the input entity. Therefore, the final prompt

is:

$$I_i = [\text{TD}; \text{Entity}_{(i)}; \text{KT}_{(i)}; Sentence]$$

where TD denotes Task Definition prompt, and $\text{Entity}_{(i)}$ refers to the required entity terms along with their associated coarse-grained types, such as *Rosie Donnel is a person*. Thus, the visual knowledge V_i^n produced by each LLM can be expressed as follows:

$$V_i^n = \text{LLMs}^n(I_i) \tag{5}$$

where n represents the n-th large language model. Drawing inspiration from self-consistency, we aim to enhance knowledge consistency across models, thereby facilitating the generation of more robust and accurate visual knowledge. However, the outputs of different models can convey the same meaning using different expressions [65]. For example, as shown in Fig. 5, both GPT-3.5 Turbo and LLaMA-3 70B offer synonymous meanings for *facial contour* and *special number* (highlighted in yellow and green), although the expression is not identical. These variations in expression complicate the effective application of traditional voting methods, such as XGBoost [66]. To address this challenge, we use large models as semantic voters to perform soft voting and obtain the final visual knowledge for each attribute:

$$k_i = \text{LLM}_{\text{Voter}}\left(v_i^1, v_i^2, \ldots, v_i^n\right) \tag{6}$$

Since the visual knowledge may contain irrelevant information, we use LLMs to summarize the key information related to the entity's appearance. Consequently, the final visual knowledge output is denoted as:

$$\widetilde{K} = \{\text{Summary}(k_i),\ i \in M\} \tag{7}$$

where *M* denotes the number of entities predicted in Stage 1.

3) Object Box Annotation: Although we encourage the model to directly generate the final bounding box rather than selecting from candidate boxes to reduce error propagation caused by object annotations [16], Zhang et al. [22] demonstrate that providing prior object annotation information to LLaVA can improve its visual grounding ability. Therefore, we incorporate object annotation bounding box information into our approach. We use the toolkit [10][3] to extract object bounding boxes from the images. As shown in Fig. 3, we concatenated the extracted boxes with their corresponding categories to obtain the final object annotation information, denoted as A, which is expressed as follows:

$$A = \text{argTOP}_K\left(\text{Annotation}(v)\right) \tag{8}$$

where K represents the number of objects and their corresponding bounding boxes.

4) Knowledge Enhanced Fine-Tuning: At this stage, we integrate the extracted knowledge and prior entities with their corresponding coarse-grained types. We then combine these elements with the original input text w and image v to generate fine-grained types and bounding boxes, thereby addressing the grounding problem in FMNERG. The final formalized input is defined as follows:

$$I_{Know} = [\text{TaskPrompt}; \text{Knowledge}; O^1]$$

where the Knowledge consists of the visual knowledge $\widetilde{K}$, and object annotation information A. Following Wang et al. [16], incorporating coarse-grained representations into the final output has been shown to improve performance. To improve readability and interpretability, each component of a triple, **entity**, **coarse-grained type**, **fine-grained type**, and **normalized bounding boxes**, is converted into a natural language sketch. Rather than selecting from a set of candidate objects, the model directly generates normalized bounding-box coordinates for each entity. Accordingly, the ground-truth output for the second stage, Y, is defined as:

"Entity is a(an) coarse-grained type and a(an) fine-grained type, whose bounding boxes are boxes".

When multiple entities are present in a sentence, their corresponding descriptions are concatenated using semicolons (;). As illustrated in the example shown in Fig. 3:

"Kevin Love is a(an) person and a(an) athlete, whose bounding boxes are (0.448, 0.02, 0.913, 1.0); J. R. Smith is a(an) person and a(an) athlete, whose bounding boxes are (0.107, 0.268, 0.51, 1.0)".

Thus, the second-stage training process can therefore be formulated as:

$$O^2 = \text{LVLMs}(D, I_{Know}, \text{LoRA}^2) \tag{9}$$

where D denotes the original input data, including text w and image v. To supervise this generative process, we apply a token-level cross-entropy loss with respect to the ground-truth sequence Y:

$$\mathcal{L}_{\text{CE}} = -\log P(Y \mid I_{\text{Know}}, D) \tag{10}$$

$$= -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, I_{\text{Know}}, D) \tag{11}$$

where y_t denotes the ground-truth token at step t. To reduce resource consumption and improve model maintainability, we adopt LLaVA as in stage 1, while applying distinct LoRA modules to activate different model capabilities. It is worth noting that the results from Stage 1 are not always accurate. Therefore, we introduce a controlled amount of noise into the Stage 2 training data to enhance model robustness — for example, by perturbing the grounding of coarse-grained entity types obtained in Stage 1. The detailed setup is discussed in Section IV-L.

IV. Experiment

A. Datasets

To evaluate the proposed VKEL, we utilized the FMNERG benchmark dataset, the only fine-grained benchmark annotated by Wang et al. [16]. from multimodal tweets. This dataset comprises 8 coarse-grained types and 51 fine-grained types. The statistics of the dataset are presented in Table II.

[3]https://github.com/open-mmlab

TABLE II
THE STATISTICS OF THE FMNERG DATASET ARE SUMMARIZED AS FOLLOWS: #GE DENOTES THE GROUNDABLE ENTITIES, WHILE #PER, #LOC, #BUI, #ORG, #PRO, AND #EVE REPRESENT THE COARSE TYPES OF PERSON, LOCATION, BUILDING, ORGANIZATION, PRODUCT, AND EVENT, RESPECTIVELY

Split	Basic Statistics				Coarse Type Statistics							
	#Tweet	Entity	#GE	BoX	#Per	#Loc	#Bui	#Org	#Pro	Art	#Eve	Other
Train	7,000	11,779	4,733	5,723	5,019	1,553	365	3,035	355	495	614	343
Dev	1,500	2,450	991	1,171	1,072	345	62	595	82	103	126	65
Test	1,500	2,543	1046	1,254	1,104	327	77	638	88	106	129	74
Total	10,000	16,772	6,770	8,148	7,195	2,225	504	4,268	525	704	869	482

B. Baselines

Our baseline methods include the unimodal and multimodal methods to address the FMNERG. The implementation details for each method are described below.

1) Unimodal Methods: Unimodal methods refer to using NER methods to extract entity-type pairs, followed by setting the visual object prediction to **None**.

- **HBiLSTM-CRF** [3] originally addresses the MNER task using a sequence labeling approach. It employs a BiLSTM to model contextual information, followed by a CRF layer to identify entity types.
- **BERT and BERT-CRF** [44] are two variants of HBILSTM-CRF. Using BERT, which is pre-trained on a large dataset to learn general representation capabilities, Wang et al. [16] replaced the BiLSTM as the in-context encoder to enhance the model's incontext ability.
- **T5-Generation** [17] converts the NER task as a paraphrase generation problem and employs the sequence-to-sequence model T5 [17] to generate paraphrase sentences containing entity-type pairs.

2) Multimodal Methods: These methods first extract entity-type pairs using an MNER approach and then predict the visual objects using a visual grounding model based on objects detected by Faster R-CNN [12].

- **GVATT-VinVL-EVG** uses a visual attention-based BiLSTM-CRF model [3] to extract the entity-type pairs and then uses the EVG model to detect their corresponding visual objects.
- **UMT-VinVL-EVG** utilizes the unified multimodal transformer [4] to extract entity-type pairs. The UMT employs a unified framework that combines a multimodal transformer with a span-based detection module to mitigate biases arising from visual information. It then uses the visual grounding model to detect the corresponding visual objects.
- **ITA-VinVL-EVG** utilizes the image-text alignment framework [37] to extract entity-type pairs. This framework transforms images into regional tags and captions, facilitating the alignment of image features with the textual space. This alignment enhances the utilization of attention mechanisms in pretrained textual models and supports effective interaction modeling. It then uses the EVG model to detect their corresponding visual objects.
- **MMT5-VinVL-EVG** utilizes the T5 model to extract entity-type pairs and then predicts the visual objects using the EVG model.
- **TIGER** [16] is a T5-based multimodal generation framework that formulates FMNERG as a generation problem. It differs from MMT5-VinVL-EVG in that it jointly trains to extract entity-type pairs and predict the corresponding visual objects.
- $\mathbf{LLaVA_{direct}}$ is an end-to-end baseline that directly fine-tunes the original LLaVA model using LoRA on the FMNERG dataset, without synthetic augmented data, to evaluate LLaVA's intrinsic performance on this task.
- $\mathbf{VKEL_{LoRAMoe}}$ replaces the LoRA fine-tuning strategy in VKEL with the LoRAMoe structure. LoRAMoe [67], a multi-expert model built upon LoRA, has shown superior performance compared to LoRA in multi-task training.

C. Evaluation Metrics

The FMNERG task requires extracting a set of entity-type-object triples and includes two corresponding subtasks: fine-grained multimodal named entity recognition (FMNER), which extracts entity-type pairs, and entity extraction and grounding (EEG), which extracts entity-object pairs. Following the previous setup, for both entity and entity type, only the predicted entity p_e and entity type p_t that are consistent with the ground-truth values g_e and g_t are considered correct, as shown in (12):

$$C_e/C_t = \begin{cases} 1, & p_e/p_t = g_e/g_t; \\ 0, & \text{otherwise}\ ; \end{cases} \tag{12}$$

where C_e and C_t refer to the correctness of the entity and type. For object extraction, if the visual object is not grounded in the image, the prediction is considered correct only if it is labeled as *None*. Otherwise, if the visual object is groundable, the prediction is deemed correct when the IoU score between the predicted bounding box p_o and one of the truth bounding boxes $[g_{o,1}, \ldots, g_{o,J}]$ exceeds 0.5 [16], as shown in (13):

$$C_o = \begin{cases} 1, & p_o = g_o = None \\ 1, & Max(\text{IoU}_0, \ldots, \text{IoU}_J) > 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

where C_o refers to the correctness of the bounding box, and IoU_j denotes the IoU score between the predicted box p_o and the j-th annotated bounding box $g_{o,j}$. The IoU score is computed as the ratio of the area of intersection between the two boxes to the area of their union, which is denoted as:

$$\text{IoU}_j = \frac{\text{Intersection}(p_o, g_{o,j})}{\text{Union}(p_o, g_{o,j})} \tag{14}$$

TABLE III
THE STATISTICS OF THE SYNTHETIC DATASET ARE PRESENTED BELOW. *AVG* DENOTES THE AVERAGE SENTENCE LENGTH.

Base Information			Coarse Type Statistics							
Number	Entity	Avg	#Per	#Loc	#Bui	#Org	#Pro	#Art	#Eve	#Other
21950	34150	16.5	13173	3233	2846	5949	2184	2159	1791	2815

Thus, we define the correct sample for the FMNERG, FMNER, and EEG tasks, as shown in (15):

$$\begin{aligned} C_{FMNERG} &= C_e \cap C_t \cap C_o \\ C_{FMNER} &= C_e \cap C_t \\ C_{EEG} &= C_e \cap C_o \end{aligned} \tag{15}$$

Based on this, we evaluate each task using precision, recall, and F1 score as the evaluation metrics.

D. Implementation Details

To facilitate result reproduction, we provide the following detailed settings. The LLaVA-1.5 model is initialized from a pre-trained Vicuna-7B language model, with the CLIP-ViT-Large-Patch14 used as the visual encoder. The number of training epochs, batch size, and maximum sentence length are set to 10, 30, and 2048, respectively. The learning rate is 2e-5, and the LoRA rank r and alpha are set to 8 and 16, respectively. We implement our proposed model using PyTorch, adopt the AdamW optimizer during training, and run all experiments on an RTX 8000 GPU. For the synthetic data used in the first stage, we employed existing fine-grained textual NER datasets [60] and mapped their fine-grained labels to the 8 coarse-grained categories defined by FMNER. To ensure consistency with FMNERG, sequences longer than 26 tokens were discarded. This process resulted in 21,950 synthetic samples. The statistics of the synthetic data are presented in Table III. The adapted datasets and the PyTorch implementation are publicly available at: https://github.com/YuanLi95/VKEL.

E. Comparison Results

Table IV presents a comprehensive summary of the comparison results between the proposed model and previous methods, explicitly focusing on precision, recall, and F1 scores in FMNERG and two subtasks: FMNER and EEG.

1) Results on FMNERG: First, we observe that T5-paragraph-None achieves the best performance in text-only methods, likely due to its exceptional performance on the FMNER task. This suggests that language models with more comprehensive pre-training knowledge are crucial for improving FMNER and FMNERG performance. Additionally, most multimodal models generally outperform their corresponding text-only counterparts, indicating that FMNERG relies more heavily on image information than FMNER to complete entity grounding. Furthermore, TIGER achieves the best performance, benefiting from the strong MNER capabilities of T5 and its joint training approach for extracting entity-type pairs and predicting corresponding visual objects.

Additionally, directly fine-tuning LLaVA with LoRA ($\mathrm{LLaVA}_{\mathrm{direct}}$) achieves competitive performance on FMNERG, benefiting from its strong FMNER capabilities. Furthermore, we incorporated the data generated in the first stage into the original FMNERG dataset and trained the model using the same procedure as $\mathrm{LLaVA}_{\mathrm{direct}}$, denoted as $\mathrm{LLaVA}_{\mathrm{argument}}$. Although the generated data are not entirely consistent with the original FMNERG samples, the enhanced entity extraction ability (an improvement of 2.64% in FMNER F1-score) leads to a performance gain on FMNERG as well.

Compared with these baseline models, our approach achieves a substantial improvement, outperforming TIGER by 10.24% and our main baseline $\mathrm{LLaVA}_{\mathrm{direct}}$ by 7.07%. This performance gain primarily stems from the two-stage optimization framework. In the first stage, the language understanding capabilities of the MLLM are activated, while in the second stage, LLaVA's entity grounding ability is strengthened through refined visual knowledge integration.

Finally, we find that $\mathrm{VKEL}_{\mathrm{LoRAMoe}}$ does not yield performance improvements, particularly in the FMNERG and FMNER tasks. We believe that $\mathrm{VKEL}_{\mathrm{LoRAMoe}}$ leverages multi-expert learning to capture task differences. However, in FMNERG, the task differences are not substantial enough, and the additional parameters lead to training instability, causing a decline in performance.

2) Results on FMNER and EEG: The FMNER and EEG columns in Table IV indicate that multimodal models do not significantly improve performance over their unimodal counterparts in FMNER. For instance, TIGER shows no notable advantage over T5-Paraphrase-None in FMNER.

In contrast, our proposed model achieves notable improvements of 5.99% and 6.83% in the FMNER and EEG subtasks, respectively. These gains are attributed to several factors: (1) Low-cost MNER data synthesis: This approach significantly enhances the model's performance in MNER by generating effective training data; (2) Integration of refined auxiliary visual knowledge: Incorporating relevant visual information provides critical context, boosting the model's EEG performance; (3) Consistency across multiple large language models: Leveraging consistency across multiple LLMs improves the accuracy and relevance of the retrieved visual knowledge.

3) Results on Each Coarse-Grained Entity Type: To better understand the performance differences across methods, we further present the coarse-grained entity type results of our VKEL model compared to baseline models in Table V.

First, the proposed low-cost MNER data synthesis method substantially enhances LLaVA's capability to recognize most entity types, with the exception of the **Product** category. In particular, for the rare categories **Event**, **Building**, and **Art**, $\mathrm{LLaVA}_{\mathrm{augment}}$ achieves significant performance gains over $\mathrm{LLaVA}_{\mathrm{direct}}$. Moreover, our proposed VKEL model achieves performance comparable to the state of the art, except for the **Building** and **Location** categories. The most notable improvement is observed in the **Person** category, where VKEL attains an 11.21% performance gain. This category also constitutes the largest proportion of the dataset. The improvement can be attributed to the fact that the **Person** category benefits more from

TABLE IV
THE MAIN RESULTS OF VARIOUS METHODS ON FMNERG AND ITS TWO SUBTASKS ARE PRESENTED. THE BEST RESULTS ARE SHOWN IN BOLD, AND THE SECOND-BEST RESULTS ARE UNDERLINED. PRE., REC., AND F1 REPRESENT PRECISION, RECALL, AND F1-SCORE, RESPECTIVELY.

Modality	Methods	FMNERG			FMNER			EEG		
		Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
Text	HBiLSTM-CRF-None	34.86	32.38	33.57	62.31	56.55	59.29	49.25	43.27	46.07
	BERT-None	33.28	34.28	33.77	58.91	60.05	59.47	46.27	47.65	46.94
	BERT-CRF-None	34.41	35.51	34.95	60.06	61.38	60.72	46.93	48.44	47.67
	T5-Paraphrase-None	37.38	37.29	37.33	64.83	65.32	65.07	49.03	48.91	48.97
Text+Image	GVATT-VinVL-EVG	42.02	38.75	40.32	63.08	57.85	60.35	55.27	53.46	54.35
	UMT-VinVL-EVG	40.67	41.99	41.32	61.24	62.01	61.63	53.58	55.32	54.43
	UMGF-VinVL-EVG	41.73	42.11	41.92	61.68	61.90	61.79	54.51	55.00	54.75
	ITA-VinVL-EVG	43.05	42.51	42.78	63.80	62.64	63.21	57.63	56.90	57.26
	H-Index	46.83	46.28	46.55	65.25	64.45	64.84	60.82	60.10	60.46
	MMT5-VinVL-EVG	45.35	45.08	45.21	66.46	66.77	66.61	58.35	58.01	58.18
	TIGER	47.57	46.85	47.20	64.43	65.40	64.91	62.44	61.49	61.96
	$\text{LLaVA}_{\text{direct}}$	49.78	50.97	50.37	67.80	69.41	68.60	60.92	62.37	61.63
	$\text{LLaVA}_{\text{argument}}$	51.68	52.23	51.95	70.86	71.62	71.24	62.45	63.11	62.78
	VKEL	**57.12**	**57.78**	**57.44**	**72.18**	**73.01**	**72.59**	68.07	68.85	68.46
	$\text{VKEL}_{\text{LoRAMoe}}$	56.78	57.31	57.04	71.24	72.26	71.74	**68.44**	**69.24**	**68.83**

TABLE V
THE F1 SCORES OF MULTIMODAL METHODS ON FMNERG ARE REPORTED FOR EACH COARSE-GRAINED ENTITY TYPE, WITH THE PERCENTAGES BENEATH EACH TYPE INDICATING THE PROPORTION OF THIS ENTITY TYPE IN THE TEST DATASETS

Methods	Person (43.41%)	Location (12.76%)	Building (3.03%)	Organization (25.09%)	Product (3.46%)	Art (4.17%)	Event (5.17%)	Other (2.91%)
GVATT-VinVL-EVG	35.21	61.64	35.37	42.60	15.38	32.79	42.55	41.03
UMT-VinVL-EVG	37.10	63.58	35.09	42.82	18.28	36.62	46.79	38.24
UMGF-VinVL-EVG	37.04	63.16	38.51	44.71	17.39	30.70	47.81	38.89
ITA-VinVL-EVG	37.91	65.52	39.16	44.34	17.18	34.91	46.62	36.36
H-Index	45.13	62.33	32.88	46.68	28.19	38.89	45.56	41.81
MMT5-VinVL-EVG	38.61	**69.44**	37.18	46.30	16.18	41.79	50.00	46.98
TIGER	43.78	67.69	**40.00**	**46.75**	27.38	43.27	48.39	48.28
$\text{LLaVA}_{\text{direct}}$	48.67	62.28	33.33	42.71	29.05	44.81	40.46	44.63
$\text{LLaVA}_{\text{argument}}$	49.31	64.57	38.46	43.91	25.11	47.75	50.79	45.56
VKEL	**54.99**	66.39	37.20	44.39	**29.09**	**52.63**	**51.59**	**49.31**

TABLE VI
F1 SCORES FROM THE ABLATION STUDY OF VKEL ON THE FMNERG, FMNER, AND EEG TASKS

Methods	FMNERG	FMNER	EEG
VKEL	57.44	72.59	68.46
w Joint Training	56.52	71.81	67.47
w/o Synthetic Data	55.18	69.54	66.50
w/o Visual Knowledge	52.70	70.35	63.33
w/o Object Annotation	54.60	71.88	65.10

visual knowledge than other categories, as visual attributes such as *gender*, *skin tone*, and *hair color* are relatively consistent and standardized across instances, providing a distinct advantage in recognition tasks.

F. Ablation Studies

We conducted ablation studies to assess the contributions of individual components in the proposed model: synthetic data, visual knowledge, and object annotation. Additionally, we explored the joint training of two stages using a single LoRA (referred to as w Joint Training). The results of the ablation experiments are presented in Table VI, revealing varying degrees of performance decline for different ablation models. These findings highlight the unique contributions of each component to the proposed model.

Removing **visual knowledge** from the FMNERG task disrupts the effective connection between entities and images, leading to a significant drop in grounding performance (5.13% decrease in the F1 score for EEG) and, consequently, a notable decline in overall FMNERG performance (4.74% decrease in F1 score). Similarly, removing **object annotation** eliminates the prior knowledge of the bounding box, essential to improve grounding capabilities. Furthermore, excluding synthetic data used to optimize MNER performance ultimately impairs FMNERG task performance. Despite being a low-cost solution, this shows that the proposed synthetic data method contributes to FMNER performance. Furthermore, we observe that **joint training** does not lead to improvements. This is because of the different focuses of the tasks: FMNER emphasizes text-related capabilities, while entity grounding relies on visual reasoning. The indiscriminate combination of these tasks likely results in gradient conflicts between them [68].

G. Generalizability in Coarse-Grained Entity Grounding

To evaluate the performance and generalizability of our model, we compared it with several state-of-the-art grounded

TABLE VII
F1 SCORES OF GMNER AND ITS TWO ASSOCIATED SUBTASKS

Methods	GMNER	MNER	EEG
H-Index	56.41	79.73	61.18
MQSPN	58.76	80.43	64.40
GEM	61.54	84.81	64.49
RiVEG	66.02	84.19	69.18
VKEL	66.87	83.75	70.43

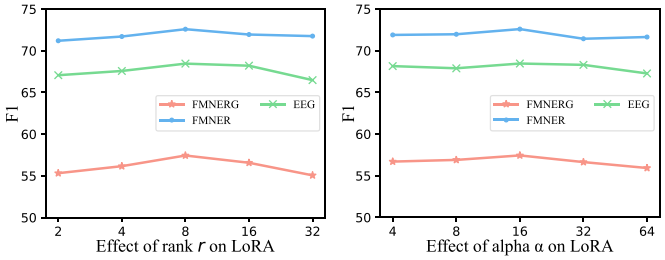

Fig. 7. Effects of different parameters.

multimodal named entity recognition (GMNER) models, including H-Index [14], GEM [43], MQSPN [42], and RiVEG [15], as summarized in Table VII. Compared with the H-Index model, MQSPN employs finer-grained cross-modal feature alignment, resulting in a 2.35% improvement in F1 score on the GMNER task.

Furthermore, our proposed model demonstrates superior performance over RiVEG, particularly on EEG tasks, achieving a relative F1 improvement of 1.25%. This gain primarily stems from our model's ability to establish robust visual–textual alignment, effectively bridging the gap between textual entities and their corresponding visual objects. In contrast, RiVEG mainly relies on background knowledge of entities (e.g., *Tim Duncan is a basketball player*). By leveraging visual grounding, our model focuses more directly on entity–object correspondence, thereby substantially enhancing performance in EEG tasks.

H. Effects of Hyperparameters

Since our model utilizes LoRA for low-parameter fine-tuning, we conducted a series of sensitivity experiments to evaluate the impact of two critical LoRA parameters, rank r and alpha α, on the performance of VKEL. As illustrated in Fig. 7, lower rank values in VKEL resulted in suboptimal performance across most tasks. This can be attributed to the insufficient parameter capacity of smaller ranks, which limits the model's ability to effectively bridge the gap between FMNER and EEG tasks through learning visual knowledge. Conversely, larger rank values introduced excessive model parameters, leading to overfitting and a subsequent decline in performance. The optimal performance was achieved at $r = 8$. We also examined the effect of α on model performance. Overall, α had a relatively minor impact, with the best performance observed at $\alpha = 16$. This suggests that VKEL exhibits a degree of robustness to variations in α.

TABLE VIII
GENERALIZATION OF VKEL ACROSS DIFFERENT LVLM BACKBONES. RESULTS ARE REPORTED AS F1 SCORES (%).

Backbone	Method	FMNERG	FMNER	EEG
MMT5	directly	39.85	64.51	51.20
	VKEL	46.28	69.15	56.43
Qwen2.5-VL	directly	51.41	68.62	62.89
	VKEL	55.69	68.82	67.91

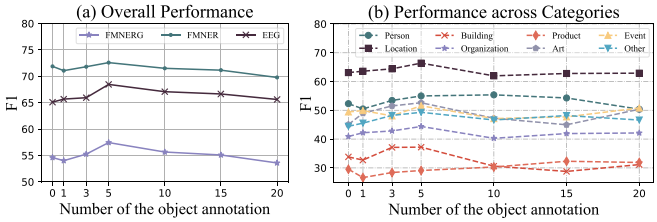

Fig. 8. Effects of different numbers of object annotations.

I. Generalization Across Different LVLMs

To further evaluate the generality of the proposed two-stage framework, comprising Data Augmentation for MNER and Knowledge-Guided Grounding, we extend our approach to additional LVLMs, including MMT5, the backbone used in TIGER [16], and Qwen2.5-VL (8B). As shown in Table VIII, the proposed VKEL consistently improves performance across all evaluation benchmarks for both model families.

Specifically, VKEL-MMT5 achieves relative F1 score improvements of 6.43%, 4.64%, and 5.23% over the directly fine-tuned MMT5 on FMNERG, FMNER, and EEG, respectively. Similarly, VKEL-Qwen2.5-VL yields gains of 4.28%, 0.20%, and 5.02% on the same tasks compared with its direct fine-tuning baseline. Notably, for MMT5, both FMNER and EEG show consistent improvements, as MMT5, being a relatively lightweight model based on T5-base, possesses limited domain knowledge for extraction-oriented tasks and thus benefits more from the data augmentation process. In contrast, Qwen2.5-VL exhibits its largest gain on EEG, indicating that VKEL effectively enhances grounding and visual reasoning for entity localization.

These results demonstrate that VKEL is model-agnostic and can be seamlessly integrated into diverse multimodal large language model architectures. By leveraging data augmentation and knowledge-guided alignment through integrated visual knowledge, VKEL enhances entity extraction while exhibiting strong generalization and scalability.

J. Sensitivity Analysis

1) Number of Annotation Boxes: Our model incorporates prior knowledge via object annotation boxes, although the final bounding boxes are generated automatically. To assess the impact of visual objects, we conducted an analysis, with overall results presented in Fig. 8(a). The results show that as the value of K increases, model performance initially improves, peaking at $K = 5$, after which it declines. We posit that when $K < 5$, essential object information may be omitted from the candidate

TABLE IX
PERFORMANCE COMPARISON UNDER DIFFERENT IOU THRESHOLDS

IoU thresholds	0.1	0.3	0.5	0.7	0.9	1.0
FMNERG	59.84	58.11	57.44	55.55	52.09	38.36
EEG	71.49	69.32	68.46	66.26	62.20	46.82
Person	58.40	55.88	54.99	52.67	46.01	20.13
Location	66.95	66.38	66.39	65.81	64.96	64.96
Building	40.31	40.31	37.20	37.20	35.66	34.11
Organization	46.33	44.99	44.39	43.95	43.49	42.90
Product	30.30	30.30	29.09	27.87	25.45	18.18
Art	54.39	52.63	52.63	51.74	50.00	42.98
Event	53.21	51.59	51.59	49.59	49.59	49.59
Other	59.65	49.31	49.31	45.20	43.84	39.72

bounding boxes. Conversely, when $K > 10$, the inclusion of excessive noise outweighs the benefits of visual object information, resulting in diminished performance.

Besides, we further analyze each category's sensitivity to the number of annotation boxes, as shown in Fig. 8(b). Visually grounded categories such as **Person**, **Location**, and **Art** benefit the most, with performance peaking around $K = 5$. In contrast, **Building** and **Product** improve with a moderate number of boxes but decline when excessive boxes are added, indicating sensitivity to noise. More abstract categories, including **Event** and **Other**, show minimal variation.

2) IoU Thresholds: Following prior work [16], we use an IoU threshold of 0.5 as the default metric for evaluating grounding performance. To provide a more comprehensive assessment, we additionally evaluate our best-performing model across multiple IoU thresholds, with the results summarized in Table IX.

It should be noted that an IoU of 1 indicates a perfect overlap between the predicted and ground-truth bounding boxes, a condition rarely achieved in practice. Since FMNER results are largely insensitive to IoU thresholds, they are excluded from this analysis. Overall, the performance shows only a slight decline as the IoU threshold increases. Even under the stringent IoU $= 0.9$ setting, the proposed model attains F1 scores of 52.09% and 62.20% on FMNERG and EEG, respectively. These results demonstrate that, benefiting from the proposed adaptive grounding generation mechanism, the model can precisely align predictions with human annotations, thereby exhibiting strong robustness.

From coarse-grained entity type, the **Location** category comprises a considerable number of **None** instances, representing cases without matched entities. Conversely, the **Person** and **Organization** categories show a moderate decrease in performance with higher IoU thresholds, highlighting the challenge of more accurately localizing these entity types.

3) Proportion of Synthetic Data: Since additional synthetic data were employed in the initial stage to enhance multimodal named entity recognition and improve overall performance, we investigate the impact of progressively integrating synthetic datasets into the original training data. The corresponding results are presented in Fig. 9. We observe that when the proportion of synthetic data is below 20%, model performance deteriorates, likely due to a domain gap between the synthetic and FMNERG datasets. A small amount of low-quality synthetic data may

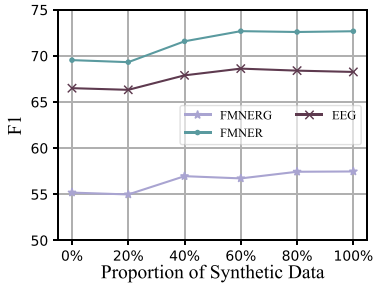

Fig. 9. Effect of synthetic data proportion.

TABLE X
F1 SCORES OF INDIVIDUAL LLMS IN VKEL ON THE FMNERG, FMNER, AND EEG TASKS

Methods	FMNERG	FMNER	EEG
VKEL	57.44	72.59	68.46
only LLaMA-3 70B	54.09	69.68	64.94
only GPT3.5-Turbo	55.18	71.07	67.55
only Gemini-Pro	54.82	71.49	66.38

interfere with the highly relevant pretraining process. However, when the proportion exceeds 40%, the model begins to benefit from the increased diversity, learning more generalized entity representations. Performance continues to improve, reaching its maximum at 80% synthetic data; beyond this point, additional synthetic data offers no substantial performance gains.

K. Effects of Individual LLMs for Visual Knowledge

Our model employs multi-consistency across LLMs to achieve more accurate and robust visual knowledge acquisition. To evaluate the performance of individual LLMs, we conducted experiments on LLaMA-3 70B, GPT-3.5 Turbo, and Gemini Pro to assess their quality. As shown in Table X, LLaMA-3 70B exhibited the weakest performance, primarily due to its frequent *N/A* responses. While these responses help reduce hallucinations, they also limit the retrieval of relevant visual information. Additionally, even the best-performing individual model, GPT-3.5 Turbo, demonstrated a significant performance gap, with a 2.26% lower F1 score in FMNER compared to our proposed VKEL model. This performance gap is largely attributed to VKEL's voting mechanism, which effectively mitigates hallucinations caused by reliance on single models, resulting in substantial improvements in the quality of retrieved visual knowledge and FMNER information.

L. Effect of Noisy Knowledge Inputs

The proposed VKEL framework consists of two stages: data enhancement for MNER and knowledge-guided grounding. As the input to the second stage may not always be accurate during generation, we aimed to enhance the model's robustness and facilitate autonomous decision-making for providing visual knowledge. To achieve this, we intentionally introduced noise into the data during the second stage of training. Specifically, we randomly removed entities or altered their coarse-grained types from the first-stage output at varying rates. For instance,

TABLE XI
EFFICIENCY AND RESOURCE COMPARISON OF VKEL WITH DIFFERENT BACKBONES

Backbone	API Token		Training Parameter	Total Parameter	Inference Time (S)	
	Visual Knowledge	Summary			Stage 1	Stage 2
LLaVA (r=8)	Input: 734 Response: 180	Input: 673 Response: 216	19.99M	6.76B	1.08	1.93
Qwen2.5 (r=4)			11.89M	8.31B	0.81	1.54
MMT5 (Full)			224.50M	224.50M	0.35	0.61

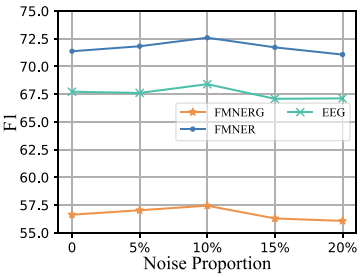

Fig. 10. Effects of different proportions of noisy samples in the second stage.

as illustrated in Fig. 3, the entity type **Person** was changed to **Product** for *Roise Donnell* when fed into the second stage.

The results, shown in Fig. 10, indicate that the model's performance remains stable when the noise ratio is below 10%. The optimal performance occurs at a noise ratio of 10%, which enhances the model's robustness during the prediction phase. However, when the noise ratio exceeds 15%, performance declines significantly. This degradation is likely due to excessive noise, which forces the model to overly rely on second-stage decisions, impairing its ability to learn the mappings between entities and their visual grounding effectively.

M. Efficient Analysis

In this section, we analyze the efficiency and computational cost of VKEL under different backbone configurations. As shown in Table XI, we report both training-related statistics (i.e., trainable and total parameters) and inference-related factors, including average API token usage and runtime per stage. All inference experiments were conducted on a single NVIDIA RTX 8000 GPU with a batch size of 4.

Since VKEL retrieves visual knowledge from multiple LLMs, we first measure the average number of input and output tokens generated by each LLM for every candidate entity, as well as the total number of tokens consumed during the summarization stage across all knowledge calls. In practice, VKEL typically requires about 900 tokens for each piece of visual knowledge processed by each LLM, and an additional 900 tokens for summarization, which is entirely feasible in practical engineering settings. Notably, different LLMs can process their respective visual knowledge in parallel, and repeated entities can be cached to prevent redundant queries, thereby improving overall efficiency.

We also examine the fine-tuning overhead in both stages. The *r* and *Full* denote the LoRA rank and full-parameter tuning, respectively. We find that the second stage introduces a slightly higher inference cost due to the integration of visual knowledge as a prompt prefix. Specifically, the total inference time is primarily determined by the backbone model size rather than the two-stage structure itself. For example, using **LLaVA (r=8)** results in 1.08 s and 1.93 s for the first and second stages, respectively. In contrast, the **Qwen2.5 (r=4)** achieves a 25.9% reduction in total inference time while maintaining comparable performance. In practical deployment scenarios, the two-stage inference process can be executed with larger batch sizes or parallelized across instances, further enhancing throughput and scalability.

N. Case Study

To provide deeper insights into our model, Fig. 11 presents four examples alongside the prediction results of different models. In example (a), both TIGER and LLaVA failed to correctly predict entity types and groundings due to inconsistencies caused by different objects in the image. In contrast, our proposed model, leveraging integrated visual knowledge such as *female artist* and *light skin*, effectively bridged the gap between entities and their groundings. Additionally, the inclusion of object annotations addressed LLaVA's limitations in grounding capabilities, enabling the VKEL model to produce more accurate entity-grounding pairs. In example (b), benefiting from the power of the visual encoder and LLMs, both our model and LLaVA successfully identified *Duke's slogan* from the image, whereas TIGER failed to match it with the pre-extracted objects. Additionally, although TIGER and LLaVA correctly identified **Grayson Allen** as an *athlete*, they struggled to map the entity to its correct grounding, mistakenly assigning *box 1* as the grounding for **Grayson Allen**. In contrast, our model effectively used the visual signal of *Grayson Allen's clothing*, which features the number *3*, to accurately identify his grounding *box 2*.

Example (c) further highlights the necessity of visual knowledge. Since the entities **Melissa Hamilton** and **Eric Underwood** have a many-to-many relationship with the image objects (*box 1* and *box 2*), both TIGER and LLaVA incorrectly associate the entities with the wrong groundings. In contrast, our model, supported by visual knowledge, successfully establishes the correct grounding for these entities. However, we also observed that all models, including TIGER, LLaVA, and our proposed VKEL model, incorrectly classify **Royal Bell** as a *company* rather than an *educational institution*. This misclassification is likely due to the scarcity of *educational institution* examples in the dataset, which limits the models' ability to effectively learn the distinguishing features of this category. Furthermore, example (d) illustrates the impact of the long-tail distribution in the data. The entity type *medical thing* is rarely present in the training

Fig. 11. Case study on four test samples. The term "LLaVA" refers to $\mathrm{LLaVA}_{\mathrm{direct}}$, where the $\surd$ and ✗ represent correct and incorrect predictions, respectively.

datasets, accounting for only 0.28% of the data (20/7000). As a result, all models struggle to effectively learn features related to *medical thing*, making it challenging to generalize to the recognition of the **DNA** entity.

V. Conclusion

This study presents the visual knowledge-enhanced LLaVA (VKEL) framework to tackle the FMNERG task. The VKEL framework follows a two-stage design: the first stage utilizes low-cost synthetic data augmentation to enhance multimodal named entity recognition, while the second stage incorporates refined visual knowledge and object annotations to improve entity grounding precision. The framework aligns textual and visual modalities to generate robust representations by leveraging a multimodal consistency mechanism. Experimental results on the FMNERG benchmark dataset demonstrate that VKEL achieves state-of-the-art performance, surpassing existing methods by 10.24% in the F1 score. Additionally, a series of experiments were conducted to assess the contributions of synthetic data generation, multimodal knowledge integration, and object annotation to the model's performance. Ablation studies reveal the effectiveness of each component and offer valuable insights into the interplay between visual and textual modalities. In future work, we plan to explore the integration of external knowledge to address the challenge of sparse category recognition in FMNERG.

References

[1] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, Sep. 2015.

[2] K. L. O'Halloran, G. Pal, and M. Jin, "Multimodal approach to analysing big social and news media data," *Discourse, Context Media*, vol. 40, 2021, Art. no. 100467.

[3] 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*, I. Gurevych and Y. Miyao, Eds., 2018, pp. 1990–1999.

[4] 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.

[5] L. Sun, J. Wang, K. Zhang, Y. Su, and F. Weng, "RpBERT: A text-image relation propagation-based bert model for multimodal NER," in *Proc. Conf. Assoc. Advance. Artif. Intell.*, 2021, pp. 13860–13868.

[6] Y. Liu, H. Li, A. Garcia-Duran, M. Niepert, D. Onoro-Rubio, and D. S. Rosenblum, "MMKG: Multi-modal knowledge graphs," in *Proc. 16th Int. Conf. Semantic Web*, 2019, pp. 459–474.

[7] S. Cui et al., "Enhancing multimodal entity and relation extraction with variational information bottleneck," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 1274–1285, 2024.

[8] L. Xie, F. Lee, L. Liu, K. Kotani, and Q. Chen, "Scene recognition: A comprehensive survey," *Pattern Recognit.*, vol. 102, 2020, Art. no. 107205.

[9] A. Farhadi and J. Redmon, "YoLOv3: An incremental improvement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, vol. 1804, pp. 1–6.

[10] C. Lyu et al., "RTMDet: An empirical study of designing real-time object detectors," 2022, *arXiv:2212.07784*.

[11] P. Zhang et al., "VinVL: Revisiting visual representations in vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5579–5588.

[12] P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6077–6086.

[13] L. Chen, W. Ma, J. Xiao, H. Zhang, and S.-F. Chang, "Ref-NMS: Breaking proposal bottlenecks in two-stage referring expression grounding," in *Proc. Conf. Assoc. Advance. Artif. Intell.*, 2021, pp. 1036–1044.

[14] 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.

[15] J. Li et al., "LLMs as bridges: Reformulating grounded multimodal named entity recognition," in *Proc. Assoc. Comput. Linguistics*, 2024, pp. 1302–1318.

[16] 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 Int. Conf. Multimedia*, 2023, pp. 3934–3943.

[17] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[18] J. Li et al., "Prompting ChatGPT in MNER: Enhanced multimodal named entity recognition with auxiliary refined knowledge," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2023, pp. 2787–2802.

[19] l. Y. yuan Cai and J. Huang, "Few-shot joint multimodal entity-relation extraction via knowledge-enhanced cross-modal prompt model," in *Proc. ACM Int. Conf. Multimedia*, 2024, pp. 8701–8710.

[20] S. Chen and B. Li, "Multi-modal dynamic graph transformer for visual grounding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15534–15543.

[21] J. Deng et al., "TransVG++ : End-to-end visual grounding with language conditioned vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13636–13652, Nov., 2023.

[22] H. Zhang et al., "LLaVA-grounding: Grounded visual chat with large multimodal models," in *Proc. Eur. Conf. Comput. Vis.*, 2025, pp. 19–35.

[23] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[24] M. Reid et al., "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," 2024, *arXiv:2403.05530*.

[25] A. Dubey et al., "The llama 3 herd of models," 2024, *arXiv:2407.21783*.

[26] L. Huang et al., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Trans. Inf. Syst.*, vol. 43, no. 2, pp. 1–55, 2025.

[27] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023, pp. 34892–34916.

[28] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," in *Proc. Int. Conf. Learn. Representations*, 2024, pp. 1–15.

[29] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[30] P. Wang, Z. Wang, Z. Li, Y. Gao, B. Yin, and X. Ren, "SCOTT: Self-consistent chain-of-thought distillation," in *Proc. Assoc. Comput. Linguistics*, 2023, pp. 5546–5558.

[31] D. Zhang, S. Wei, S. Li, H. Wu, Q. Zhu, and G. Zhou, "Multi-modal graph fusion for named entity recognition with targeted visual guidance," in *Proc. Conf. Assoc. Advance. Artif. Intell.*, 2021, pp. 14347–14355.

[32] X. Chen et al., "Hybrid transformer with multi-level fusion for multimodal knowledge graph completion," in *Proc. ACM SIGIR Conf. Res. Develop. Informat. Retrieval*, 2022, pp. 904–915.

[33] F. Zhao, C. Li, Z. Wu, S. Xing, and X. Dai, "Learning from different text-image pairs: A relation-enhanced graph convolutional network for multimodal NER," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 3983–3992.

[34] A. Graves and A. Graves, "Long short-term memory," in *Supervised Sequence Labelling With Recurrent Neural Networks*, vol. 385, Berlin, Germany: Springer, 2012, pp. 37–45.

[35] J. Lafferty et al., "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. Int. Conf. Mach. Learn.*, 2001, pp. 282–289.

[36] S. Chen, G. Aguilar, L. Neves, and T. Solorio, "Can images help recognize entities? A study of the role of images for multimodal NER," in *Proc. 7th Workshop Noisy User-generated Text*, 2021, pp. 87–96.

[37] X. Wang et al., "ITA: Image-text alignments for multi-modal named entity recognition," in *Proc. Assoc. Comput. Linguistics*, 2022, pp. 3176–3189.

[38] X. Yang et al., "Few-shot joint multimodal aspect-sentiment analysis based on generative multimodal prompt," in *Proc. Assoc. Comput. Linguistics*, 2023, pp. 11575–11589.

[39] C. Zheng, Z. Wu, T. Wang, Y. Cai, and Q. Li, "Object-aware multimodal named entity recognition in social media posts with adversarial learning," *IEEE Trans. Multimedia*, vol. 23, pp. 2520–2532, 2021.

[40] X. Chen et al., "Good visual guidance make a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2022, pp. 1607–1618.

[41] 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 Int. Conf. Multimedia*, 2024, pp. 7336–7345.

[42] 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," in *Proc. Conf. Assoc. Advance. Artif. Intell.*, 2025, pp. 25246–25254.

[43] Z. Wang et al., "Granular entity mapper: Advancing fine-grained multimodal named entity recognition and grounding," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2024, pp. 3211–3226.

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

[45] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2023, pp. 34892–34916.

[46] L. H. Li et al., "Grounded language-image pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10965–10975.

[47] Y.-C. Chen et al., "Uniter: Universal image-text representation learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 104–120.

[48] J. Cho, J. Lei, H. Tan, and M. Bansal, "Unifying vision-and-language tasks via text generation," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 1931–1942.

[49] P. Wang et al., "OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 23318–23340.

[50] T. B. Brown, "Language models are few-shot learners," 2020, *arXiv:2005.14165*.

[51] J. Zhao, L. Q. H. Chetwin, and Y. Wang, "SinTechSVS: A singing technique controllable singing voice synthesis system," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 2641–2653, 2024.

[52] F. Jin, Y. Liu, and Y. Tan, "Derivative-free optimization for low-rank adaptation in large language models," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 4607–4616, 2024.

[53] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 26296–26306.

[54] W. Dai et al., "InstructBLIP: Towards general-purpose vision-language models with instruction tuning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2024, pp. 49250–49267.

[55] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6904–6913.

[56] Y. Liu et al., "MMBench: Is your multi-modal model an all-around player?," in *Proc. Eur. Conf. Comput. Vis.*, 2025, pp. 216–233.

[57] E. Cabrio and S. Villata, "Five years of argument mining: A data-driven analysis.," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 5427–5433.

[58] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[59] D. Oneaă and H. Cucu, "Improving multimodal speech recognition by data augmentation and speech representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4579–4588.

[60] X. Ling and D. Weld, "Fine-grained entity recognition," in *Proc. Conf. Assoc. Advance. Artif. Intell.*, 2012, pp. 94–100.

[61] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[62] S. Menon and C. Vondrick, "Visual classification via description from large language models," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 1–17.

[63] P. Jian, D. Yu, and J. Zhang, "Large language models know what is key visual entity: An LLM-assisted multimodal retrieval for VQA," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., 2024, pp. 10939–10956.

[64] X. Wang et al., "Self-consistency improves chain of thought reasoning in language models," in *Proc. Int. Conf. Mach. Learn. Representations*, 2023, pp. 1–24.

[65] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The state and fate of linguistic diversity and inclusion in the NLP world," in *Proc. Assoc. Comput. Linguistics*, 2020, pp. 6282–6293.

[66] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 785–794.

[67] S. Dou et al., "LoRAMoE: Alleviating world knowledge forgetting in large language models via MoE-style plugin," in *Proc. Assoc. Comput. Linguistics*, 2024, pp. 1932–1945.

[68] H. Yun and H. Cho, "Achievement-based training progress balancing for multi-task learning," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 16935–16944.