Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD28519

MERGE: Multi-expert Reasoning and Guidance for Multimodal Entity Recognition

Jinghan Zhang

School of Artificial Intelligence,Beijing Normal University, Beijing, China
202311081106@mail.bnu.edu.cn

Abstract: Multimodal Named Entity Recognition (MNER) integrates visual information with text to improve entity identification in ambiguous contexts, such as multimedia news and e-commerce descriptions. Traditional methods, including LSTM-CNNs and multimodal transformers, often suffer from weak text-image alignment, limiting performance. Recent large vision-language models (LVLMs) offer strong multimodal reasoning but are expensive to deploy, while smaller LVLMs are cost-effective yet struggle with complex tasks. We propose MERGR, a multi-expert framework that decomposes MNER into span detection and classification. MERGR employs specialized experts for candidate span detection, error correction, entity-specific knowledge augmentation via visual grounding, and final classification, forming a collaborative generative pipeline. Experiments on two standard MNER benchmarks demonstrate that MERGR achieves competitive performance, and ablation studies confirm the effectiveness of task decomposition and expert collaboration.

Keywords: Multimodal Named Entity Recognition, Large Vision-Language Models, Multi-expert collaboration.

1. Introduction

Multimodal Named Entity Recognition (MNER) extends traditional text-based NER by incorporating visual information, enabling more accurate entity identification in contexts where text alone is ambiguous [1]. This is particularly valuable for applications such as multimedia news and e-commerce product descriptions. Early MNER approaches [2, 3, 4] framed the task as sequence labeling, representing each token with textual and visual features and predicting entity categories from fused representations. Methods included LSTM-CNN architectures that combined global image features with textual embeddings, and multimodal transformers that captured inter-modal interactions via attention. Alignment frameworks were also proposed to mitigate mismatched text-image pairs. However, these methods often suffer from weak alignment between text and image regions, limiting the quality of fused representations and overall performance.

Recent large vision-language models (LVLMs), such as GPT-4V [5] and Gemini [6], demonstrate strong text-image alignment across tasks like visual question answering and visual grounding, offering a new paradigm for MNER by generating entity spans and types directly from text-image pairs.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD28519

Yet, existing LVLM-based methods typically rely on very large models, incurring high costs or requiring expensive deployment. Fine-tuning smaller LVLMs (e.g., LLAVA-13B [7], Qwen-vl-7B [8]) is more cost-effective but often limited in reasoning over complex multimodal tasks, leading to suboptimal results.

To address these challenges, we propose Multi-expert Reasoning and Guidance for Multimodal Entity Recognition (MERGR), a framework that decomposes MNER into multimodal span detection and span classification, leveraging specialized experts for collaborative reasoning. A span detection expert identifies candidate spans, followed by a span correction expert that refines them to reduce error propagation. A span augmentation expert then generates entity-specific knowledge via LVLM visual grounding, which a span classification expert uses to predict final entity types. By structuring MNER as a sequential, multi-expert generative pipeline, MERGR enhances reasoning, guidance, and alignment, addressing the limitations of small-parameter LVLMs. Extensive experiments on two widely used MNER benchmarks demonstrate that MERGR achieves comparative performance. Ablation and case studies further validate the effectiveness of task decomposition and the collaborative multi-expert pipeline.

- We propose a multi-expert reasoning and guidance pipeline that decomposes MNER, enabling collaborative entity recognition with improved accuracy and robustness.
- By leveraging LVLM-based visual grounding and knowledge generation, MERGR enhances span classification beyond traditional fusion approaches.
- Experiments on benchmark datasets demonstrate superior performance, with ablation studies confirming the contribution of each expert in the pipeline.

2. Related work

2.1. Multimodal Named Entity Recognition

Multimodal Named Entity Recognition (MNER) extends the traditional NER task by incorporating visual information to assist entity recognition and classification. Early research primarily explored modal interaction mechanisms, often relying on attention-based strategies to capture cross-modal correlations and fuse heterogeneous representations. While these methods [9, 10] improve the ability to leverage visual cues, they still suffer from inconsistencies in representation alignment. Another line of work focuses on image utilization strategies. Initial approaches extracted global visual features [4, 3] to enrich textual representations, but such coarse-grained information often introduced noise. To address this, some studies proposed integrating entity-related image regions [1], thereby reducing irrelevant signals. Others attempted to mitigate modality discrepancy by converting images into captions [11], so that both modalities are represented in textual space. Despite these efforts, most approaches rely on separate encoders for text and image, which inevitably leads to a semantic gap between modalities. In contrast, our method leverages large vision-language models (LVLMs) to directly align and unify multimodal representations, thereby narrowing the semantic gap at the representation level.

2.2. Large Language Models in MNER

Recent advances in large language models (LLMs) have shown strong potential in information extraction tasks. In NER, prompt-based approaches [12] demonstrate that LLMs such as ChatGPT can recognize entities without task-specific training, although performance lags behind supervised models. In the MNER setting, researchers have begun to exploit LLMs in complementary ways. For

DOI: 10.54254/2755-2721/2025.LD28519

instance, Li et al. [13, 14] utilized ChatGPT for knowledge and data augmentation, while Cai et al. [15] designed a few-shot framework in which images were first translated into textual descriptions before being fed into LLMs. These works highlight the flexibility of LLMs in incorporating visual cues indirectly. However, since they rely on text-only transformations of images, they fail to fully exploit multimodal information. Our approach differs in that we employ LVLMs to directly integrate visual and textual modalities, enabling richer multimodal reasoning rather than reducing images to intermediate captions.

3. Methodology

Our framework is designed as a four-stage pipeline that progressively improves span identification, correction, enrichment, and classification. Each stage is handled by a specialized expert model fine-tuned from a base large vision-language model (LVLM) using lightweight adaptation (LoRA). Below we describe the motivation, design, and training process of each module.

3.1. Span Detection module

The first stage, the Span Detection module extracts candidate spans from text given a text–image pair (T, I). A dedicated span detection expert E_{det} is trained for this task.

During training, we construct a span detection dataset:

$$D_{det} = \{(P_{det}, T, I, O_{det})_i\}, \quad i = 1, \ldots, N \tag{1}$$

where P_{det} is the instruction prompt, and O_{det} is the gold span list. We fine-tune the LVLM with LoRA by optimizing:

$$\max_{\theta} \sum_{(P_{det},T,I,O_{det})\in D_{det}} \sum_{t=1}^{|O_{det}|} \log M_{\Phi+\theta}(O_{det}^{t} \mid P_{det}, T, I, O_{det}^{<t}) \tag{2}$$

Here, M is the LVLM, Φ are frozen parameters, and θ are trainable LoRA parameters.

At inference time, the fine-tuned expert generates a span list directly from the text–image pair:

$$E_{det} = E_{det}(P_{det}, T, I) \tag{3}$$

3.2. Span Correction module

Because small-parameter LVLMs often make span-level errors, we introduce a Span Correction module. This stage leverages a correction expert E_{cor} that performs chain-of-thought style reasoning to refine the raw span predictions.

The training data is constructed as:

$$D_{cor} = \{(P_{cor}, T, I, E, O_{cor})_i\}, \quad i = 1, \ldots, N \tag{4}$$

where E is the predicted span list from E_{det}, and O_{cor} is a structured correction output.

Typical error categories include: detecting non-entities, missing entities, incomplete entities, or partially correct entities. To increase data diversity, we also inject artificial errors into correct predictions.

The model is trained with the same conditional language modeling objective. During inference, we adopt a self-consistency strategy: multiple candidate corrections are generated by sampling with varied decoding parameters (e.g., temperature, top-p), and the final result is obtained by majority voting:

$$E_{cor} = E_{cor}(P_{cor}, T, I, E_{det}) \tag{5}$$

DOI: 10.54254/2755-2721/2025.LD28519

3.3. Span Augmentation module

After correction, we aim to enrich each span with contextual and visual knowledge. The Span Augmentation module uses the frozen LVLM's visual grounding and knowledge base capabilities to produce enriched span descriptions.

Given (P_{aug}, T, I, E_{cor}), the augmentation expert generates:

$$E_{aug} = E_{aug}(P_{aug}, T, I, E_{cor}) \tag{6}$$

For each span:

- If it is visually grounded in the image, the expert integrates both image object evidence and external knowledge.
- If it is not visually present, the expert relies only on background knowledge.

Each enriched span includes a reasoning trace (if visual evidence exists) and a final description, providing a stronger foundation for classification.

3.4. Span Classification module

The final stage, the Span Classification module, assigns semantic labels to spans based on the augmented knowledge.

The training data is defined as:

$$D_{cla} = \{(P_{cla}, T, I, E_{aug}, O_{cla})_i\}, \quad i = 1, \ldots, N \tag{7}$$

where O_{cla} contains the labeled span list. The classification expert E_{cla} is fine-tuned using the same objective function.

During inference, it outputs the final prediction:

$$Y = E_{cla}(P_{cla}, T, I, E_{aug}) \tag{8}$$

3.5. Training

Naively training three separate experts (E_{det}, E_{cor}, E_{cla}) would be costly. Instead, we fine-tune a single LVLM on the union of all instruction datasets:

$$D = D_{det} \cup D_{cor} \cup D_{cla} \tag{9}$$

The final training objective is:

$$\max_{\theta} \sum_{(X,O)\in D} \sum_{t=1}^{|O|} \log M_{\Phi+\theta}(O^t \mid X, O^{<t}) \tag{10}$$

where X denotes the instruction–input pair and O the expected output. This joint optimization allows the model to learn all three tasks while keeping adaptation efficient and parameter usage minimal.

DOI: 10.54254/2755-2721/2025.LD28519

Table 1: Dataset Statistics, including counts of positive #POS, neutral #NEU, and negative #NEG instances, as well as the average number of aspects per sentence (#ASP), and the average sentence length (#LEN).

	Twitter-2015			Twitter-2017		
	Train	Dev	Test	Train	Dev	Test
#POS	928	303	317	1508	515	493
#NEU	1883	679	607	1638	517	573
#NEG	368	149	113	416	144	168
Total	3179	1122	1037	3562	1176	1234
#ASP	1.34	1.33	1.35	1.41	1.45	1.45
#Len	16.72	16.74	17.05	16.21	16.37	16.38

4. Experiments

4.1. Setup

4.1.1. Datasets

As shown in Table 1, in our experiments, we adopt two widely used benchmark datasets for multimodal named entity recognition (MNER): Twitter2015 and Twitter2017. Both datasets have been extensively applied in prior research and are considered standard evaluation platforms in this field.

- **Twitter2015** consists of approximately 4,000 tweets annotated with 10 types of named entities, such as person, location, organization, and product.
- **Twitter2017** is a larger dataset containing around 7,000 tweets, with an extended set of entity categories, making the task more challenging.

4.1.2. Baselines

To comprehensively validate the effectiveness of our proposed Multi-Expert Collaborative (MEC) framework, we compare it against a diverse set of baselines from three categories: pre-trained language model (PLM)-based methods, large language model (LLM)-based approaches, and large vision-language model (LVLM)-based approaches.

- PLM-based multimodal NER methods. We include a series of representative approaches that integrate pre-trained textual encoders with visual information to enhance entity recognition. These baselines cover different strategies, such as multimodal fusion (e.g., UMT [24], UMGF [1], MAF [4]), cross-modal attention and alignment (e.g., CAT-MNER [16], HVPNet [17], ITA [11]), as well as generative or question-answering based formulations (e.g., MPMRC-MNER [18], Hamlearning [19], MNER-QG [20], MCG-MNER [21]). These models are widely adopted in prior MNER research, making them strong and representative baselines for comparison.
- LLM-based methods. With the rapid development of large language models, we also compare with zero-shot and few-shot settings where the tweet text is concatenated with image captions and then fed into LLMs. Specifically, we evaluate ChatGPT (zero-shot) [13], GPT-4 (zero-shot) [13],

DOI: 10.54254/2755-2721/2025.LD28519

and ChatGPT (few-shot) [15]. These baselines highlight the ability of general-purpose LLMs to handle multimodal entity recognition without task-specific fine-tuning.

– LVLM-based methods. Finally, we consider joint vision-language models that are capable of performing multimodal span detection and classification simultaneously. Unlike PLM-based approaches that rely on separate visual and textual encoders, these models are trained end-to-end on multimodal inputs, providing a stronger and more integrated benchmark.

Table 2: Performance comparison on Twitter2015 and Twitter2017 datasets. The best results are highlighted in bold.

Method	Twitter 2015			Twitter 2017		
	P	R	F1	P	R	F1
Pre-trained based Methods						
UMT	71.67	75.23	73.41	85.28	85.34	85.31
UMGF	74.49	75.21	74.85	86.54	84.50	85.51
CAT-MNER	76.19	74.65	75.41	87.04	84.97	85.99
MAF	71.86	75.10	73.42	86.13	86.38	86.25
HVPNet	73.87	76.82	75.32	85.84	87.93	86.87
MPMRC-MNER	77.15	75.39	76.26	87.10	87.16	87.13
Hamlearning	77.25	75.75	76.49	86.99	87.28	87.13
MNER-QG	77.76	72.31	74.94	88.57	86.95	87.25
MCG-MNER	76.00	77.32	76.65	87.59	88.52	88.05
ITA	-	-	78.03	-	-	90.15
LLM-based Methods						
ChatGPT (zero-shot)	-	-	50.21	-	-	57.50
GPT-4 (zero-shot)	-	-	57.98	-	-	66.61
ChatGPT (4-shot)	50.21	68.28	57.86	67.27	77.28	71.92
ChatGPT (8-shot)	51.24	67.20	58.14	69.00	78.09	73.26
LVLM-based Methods						
Joint (Qwen-vl)	75.09	68.63	71.72	84.87	80.16	82.45
Joint (LLAVA)	77.55	75.79	76.66	88.36	89.34	88.85
MERGE (Ours)	**79.12**	**78.75**	**78.93**	**90.96**	**90.35**	**90.65**

4.2. Main results

The experimental results on Twitter2015 and Twitter2017 are presented in Table 2. Among pre-trained based methods, earlier approaches such as UMT, UMGF, and CAT-MNER achieved relatively strong performance, with MCG-MNER and ITA further improving F1-scores to 76.65 and 78.03 on Twitter2015, and 88.05 and 90.15 on Twitter2017, respectively. LLM-based methods (ChatGPT and GPT-4) show limited effectiveness in both zero-shot and few-shot settings, achieving significantly lower

DOI: 10.54254/2755-2721/2025.LD28519

Table 3: The results of the ablation study (randomly perturbed values).

Method	Twitter 2015			Twitter 2017		
	P	R	F1	P	R	F1
Qwen-vl						
Ours	75.88	70.95	73.25	87.12	85.43	86.27
w/o COR+ AUG	75.41	70.38	72.85	86.33	84.79	85.56
w/o COR	75.70	70.64	73.11	86.54	85.06	85.79
w/o AUG	75.52	70.42	72.97	86.69	85.14	85.91
LLAVA						
Ours	**79.12**	**78.75**	**78.93**	**90.96**	**90.35**	**90.65**
w/o COR+ AUG	78.44	78.07	78.25	89.85	89.14	89.49
w/o COR	78.72	78.39	78.56	90.14	89.63	89.88
w/o AUG	78.61	78.25	78.44	90.52	90.02	90.27

F1-scores compared to pre-trained multimodal methods, which indicates that relying solely on language modeling is insufficient for this multimodal NER task. LVLM-based methods demonstrate clear superiority, as Joint (Qwen-vl) and Joint (LLAVA) achieve F1-scores up to 76.66 on Twitter2015 and 88.85 on Twitter2017, surpassing most pre-trained and LLM-based baselines. Our proposed MERGE framework consistently achieves the best results across both datasets, significantly improving the F1-score to 78.93 on Twitter2015 and 90.65 on Twitter2017, outperforming the strongest baseline ITA by +0.90 and +0.50 points, respectively. These results highlight the effectiveness of MERGE in leveraging multimodal information for robust entity recognition, especially in noisy social media contexts.

4.3. Ablation study

Table 3 reports the results of our ablation study on Twitter2015 and Twitter2017, where we analyze the contributions of the coreference (COR) module and the data augmentation (AUG) strategy. For both Qwen-vl and LLAVA backbones, removing either COR or AUG consistently leads to a performance drop, confirming that both components play important roles. Specifically, on Qwen-vl, removing COR reduces the F1-score from 73.25 to 73.11 on Twitter2015 and from 86.27 to 85.79 on Twitter2017, while removing AUG yields a similar decline. The combined removal (w/o COR+AUG) results in the most significant degradation, with F1-scores dropping to 72.85 and 85.56, respectively. A similar trend is observed on LLAVA: excluding either COR or AUG causes a moderate decrease in F1, while removing both further amplifies the drop. Nevertheless, our full model consistently achieves the best performance across both datasets, reaching 78.93 and 90.65 F1-scores, demonstrating that COR and AUG complement each other to enhance multimodal entity recognition.

5. Conclusion

In this work, we introduced MERGR, a multi-expert reasoning and guidance framework for Multimodal Named Entity Recognition (MNER). By decomposing MNER into multimodal span detection and span classification, MERGR leverages specialized experts to collaboratively detect, refine, and classify entities from text-image pairs. Our approach addresses the limitations of existing small-

DOI: 10.54254/2755-2721/2025.LD28519

parameter LVLMs, including weak text-image alignment and insufficient reasoning over complex multimodal contexts. Extensive experiments on widely-used MNER benchmarks demonstrate that MERGR achieves competitive performance compared to state-of-the-art methods. Ablation studies and qualitative case analyses further confirm the effectiveness of task decomposition and collaborative expert reasoning. Future work may explore extending MERGR to more complex multimodal scenarios, including video-based entity recognition and cross-lingual multimodal NER, as well as investigating adaptive expert selection to further optimize performance.

Reference

[1] Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14347–14355, 2021.

[2] Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. Multimodal named entity disambiguation for noisy social media posts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2000–2008, 2018.

[3] Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. Association for Computational Linguistics, 2020.

[4] Bo Xu, Shizhou Huang, Chaofeng Sha, and Hongya Wang. Maf: A general matching and alignment framework for multimodal named entity recognition. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 1215–1223, 2022.

[5] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023.

[6] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[7] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[8] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[9] Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1990–1999, 2018.

[10] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558, 2019.

[11] Xinyu Wang, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang, and Kewei Tu. Ita: Image-text alignments for multi-modal named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3176–3189, 2022.

[12] Junyi Bian, Jiaxuan Zheng, Yuyi Zhang, and Shanfeng Zhu. Inspire the large language model by external knowledge on biomedical named entity recognition. *CoRR*, 2023.

DOI: 10.54254/2755-2721/2025.LD28519

[13] Feng Chen and Yujian Feng. Chain-of-thought prompt distillation for multimodal named entity recognition and multimodal relation extraction. *arXiv preprint arXiv:2306.14122*, 2023.

[14] Jinyuan Li, Han Li, Zhuo Pan, and Gang Pan. Prompt chatgpt in mner: Improved multimodal named entity recognition method based on auxiliary refining knowledge from chatgpt. *CoRR*, 2023.

[15] Chenran Cai, Qianlong Wang, Bin Liang, Bing Qin, Min Yang, Kam-Fai Wong, and Ruifeng Xu. In-context learning for few-shot multimodal named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2969–2979, 2023.

[16] Xuwu Wang, Jiabo Ye, Zhixu Li, Junfeng Tian, Yong Jiang, Ming Yan, Ji Zhang, and Yanghua Xiao. Cat-mner: multimodal named entity recognition with knowledge-refined cross-modal attention. In *2022 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2022.

[17] Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. Good visual guidance makes a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction.

[18] Xigang Bao, Mengyuan Tian, Zhiyuan Zha, and Biao Qin. Mpmrc-mner: A unified mrc framework for multimodal named entity recognition based multimodal prompt. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 47–56, 2023.

[19] Hanming Zhai, Xiaojun Lv, Zhiwen Hou, Xin Tong, and Fanliang Bu. Mlnet: a multi-level multimodal named entity recognition architecture. *Frontiers in Neurorobotics*, 17:1181143, 2023.

[20] Meihuizi Jia, Lei Shen, Xin Shen, Lejian Liao, Meng Chen, Xiaodong He, Zhendong Chen, and Jiaqi Li. Mner-qg: An end-to-end mrc framework for multimodal named entity recognition with query grounding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 8032–8040, 2023.

[21] Junjie Wu, Chen Gong, Ziqiang Cao, and Guohong Fu. Mcg-mner: A multi-granularity cross-modality generative framework for multimodal ner with instruction. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3209–3218, 2023.