# E2E-GMNER: End-to-End Generative Grounded Multimodal Named Entity Recognition

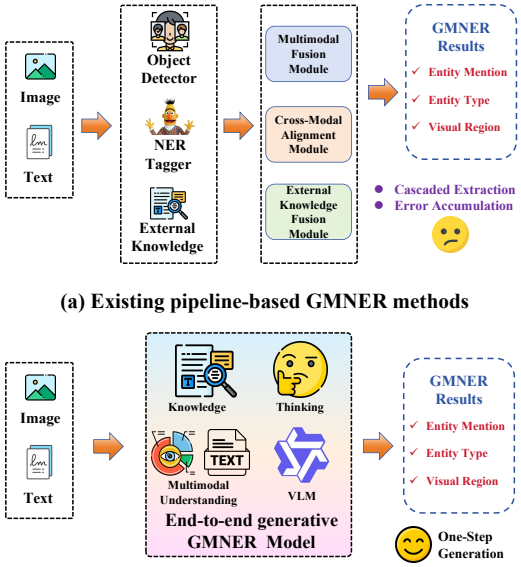
**Meng Zhang[1], Jinzhong Ning[1]*, Xiaolong Wu[1], Hongfei Lin[2], Yijia Zhang[1]***
[1]Dalian Maritime University
[2]Dalian University of Technology

## Abstract

Grounded Multimodal Named Entity Recognition (GMNER) aims to jointly identify named entity mentions in text, predict their semantic types, and ground each entity to a corresponding visual region in an associated image. Existing approaches predominantly adopt pipeline-based architectures that decouple textual entity recognition and visual grounding, leading to error accumulation and suboptimal joint optimization. In this paper, we propose **E2E-GMNER**, a fully end-to-end generative framework that unifies entity recognition, semantic typing, visual grounding, and implicit knowledge reasoning within a single multimodal large language model. We formulate GMNER as an instruction-tuned conditional generation task and incorporate chain-of-thought reasoning to enable the model to adaptively determine when visual evidence or background knowledge is informative, reducing reliance on noisy cues. To further address the instability of generative bounding box prediction, we introduce Gaussian Risk-Aware Box Perturbation (GRBP), which replaces hard box supervision with probabilistically perturbed soft targets to improve robustness against annotation noise and discretization errors. Extensive experiments on the Twitter-GMNER and Twitter-FMNERG benchmarks demonstrate that E2E-GMNER achieves highly competitive performance compared with state of the art methods, validating the effectiveness of unified end-to-end optimization and noise-aware grounding supervision. Code is available at: https://github.com/Finch-coder/E2E-GMNER

## 1 Introduction

Grounded Multimodal Named Entity Recognition (GMNER) is a vision–language task that jointly identifies named entity mentions in text, predicts their semantic types, and grounds each entity to its corresponding visual region in the associated image. By providing explicit entity-level alignments between text and visual evidence, GMNER enables structured multimodal understanding that supports downstream applications such as multimodal knowledge graph construction (Liu et al., 2019) and visually grounded question answering (Li et al., 2025).




**(a) Existing pipeline-based GMNER methods**

**(b) Our proposed end-to-end generative GMNER approach**

Figure 1: Comparison between existing pipeline-based GMNER methods and our end-to-end generative GMNER approach.

Most existing GMNER approaches adopt a pipeline-based architecture to perform cascaded extraction of named entities and their corresponding visual targets. Specifically, they typically rely on a separate BERT-based sequence labeling NER model (Lin et al., 2025; Li et al., 2024a) to identify entity mentions in text, or employ external object detectors (Yu et al., 2023; Wang et al., 2023; Tang et al., 2025b) to extract region proposals from images, followed by a fusion module that aggregates textual and visual representations to produce final GMNER predictions. In addition, some methods

*Corresponding authors.



*Findings of the Association for Computational Linguistics: ACL 2026*, pages 22460–22469

incorporate external knowledge sources, such as knowledge bases (Ok et al., 2024; Lin et al., 2025) or knowledge generated by prompting large language models (Liu et al., 2024; Wang et al., 2024), to assist joint text–image understanding.

Despite recent advances, GMNER poses several fundamental challenges. **Issue1:** Most methods adopt a pipeline architecture that decouples textual entity recognition and visual grounding, relying on separate modules such as standalone NER taggers or external object detectors, whose predictions are cascaded to produce the final GMNER outputs. This design leads to error accumulation across stages and prevents joint optimization. **Issue2:** Although most approaches attempt to resolve text–vision ambiguity through implicit cross-modal alignment or by incorporating external knowledge cues, they often lack an explicit mechanism to determine when visual evidence or external knowledge is truly informative. As a result, disambiguation can be suboptimal when visual cues are noisy, irrelevant, or misleading. **Issue3:** Under emerging generative vision–language grounding paradigms (Bai et al., 2025; Peng et al., 2023), directly predicting bounding boxes as sequences of discrete coordinate tokens introduces additional training challenges. Supervision based on a single hard target sequence makes learning sensitive to annotation noise and discretization errors, thereby reducing robustness and optimization stability.

To address the above limitations, we propose **E2E-GMNER**, a novel end-to-end framework for GMNER that enables unified optimization of entity recognition, visual grounding, semantic alignment, and background knowledge reasoning within a single generative vision-language model, as illustrated in **Figure** 1. For **Issue 1**, our framework unifies entity recognition and visual grounding within a single generative formulation, enabling the model to jointly predict entity mentions, their semantic types, and corresponding visual regions. By eliminating intermediate pipeline components, our approach mitigates error accumulation inherent in cascaded GMNER systems and allows joint optimization of entity recognition and grounding. For **Issue 2**, we adopt Chain-of-Thought (CoT) instruction tuning to adapt the planning, multimodal semantic reasoning, and knowledge utilization capabilities of vision–language models to the GMNER task. This design enables the model to autonomously decide when visual evidence or external knowledge cues are informative, thereby reducing noise introduced by irrelevant visual or knowledge signals while avoiding explicit reliance on external knowledge sources. For **Issue 3**, we introduce Gaussian Risk-Aware Box Perturbation (GRBP) during training. Specifically, we probabilistically perturb the center coordinates and scale of ground-truth bounding boxes, replacing a single hard supervision target with Gaussian-based soft supervision, where larger perturbations are assigned lower probabilities. This strategy improves the robustness of generative box prediction by tolerating small geometric deviations in bounding box coordinates, thereby stabilizing training under annotation noise and discretization effects.

Our contributions are summarized as follows:

- We propose E2E-GMNER, an end-to-end generative framework that unifies entity recognition and visual grounding within a single formulation, enabling joint optimization and avoiding error accumulation from pipeline-based GMNER methods.
- We incorporate Chain-of-Thought instruction tuning to adaptively decide when visual evidence or external knowledge is informative, and introduce Gaussian Risk-Aware Box Perturbation to stabilize generative box prediction under annotation noise and discretization errors.
- Extensive experiments on standard GMNER benchmarks show that E2E-GMNER achieves highly competitive results compared with state-of-the-art methods.

## 2 Related Work

**Grounded Multimodal Named Entity Recognition (GMNER)** aims to extract textual entities, their semantic types, and the corresponding visual regions from image–text pairs. Early approaches, such as H-Index (Yu et al., 2023) and TIGER (Wang et al., 2023), rely on external object detectors to generate candidate regions and then employ generative models to predict entity mentions and align them with the extracted visual proposals. With the advent of multimodal large language models (MLLMs), methods such as RiVEG (Li et al., 2024a) and GEM (Wang et al., 2024) integrate MLLMs to leverage their strong multimodal semantic understanding and knowledge capabilities for grounding entities to visual regions. Recent works including SCANNER (Ok et al., 2024) and

UnCo (Tang et al., 2025b) further improve generalization by incorporating external knowledge, particularly for handling unseen entities, while MQSPN (Tang et al., 2025a) adopts a set prediction paradigm to alleviate exposure bias in generative GMNER settings. In addition, MAKAR (Lin et al., 2025) employs an MLLM-based multi-agent system to resolve semantic ambiguity and enhance alignment between textual entities and visual regions.

It is worth noting that nearly all existing GMNER methods follow a pipeline-based architecture, which decouples textual entity recognition and visual grounding into separate modules such as standalone NER taggers or external object detectors, with their predictions cascaded to produce final GMNER outputs. This design inevitably leads to error accumulation and hinders joint optimization across tasks. **To the best of our knowledge, our proposed approach is the first fully end-to-end GMNER framework, enabling unified optimization of entity recognition, visual grounding, semantic alignment, and background knowledge reasoning within a single generative model.**

## 3 Method

### 3.1 Task Formulation

Grounded Multimodal Named Entity Recognition (GMNER) aims to jointly identify named entity mentions in text, predict their semantic types, and ground each entity to a corresponding visual region in the associated image. Formally, given an image–text pair $(I, T)$, where $I$ denotes the image and $T$ denotes the textual sequence, the goal of GMNER is to produce a set of structured entity records:

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

where $e_i$ is span of the $i$-th entity mentioned in $T$, $c_i$ is its semantic type, and $b_i = (x_i^1, y_i^1, x_i^2, y_i^2)$ denotes the bounding box grounding the entity in the image $I$, specified by the top-left and bottom-right coordinates. The output set $\mathcal{Y}$ is unordered, and the number of entities $N$ varies across instances.

### 3.2 End-to-End Generative Framework for GMNER

We propose an end-to-end generative framework for Grounded Multimodal Named Entity Recognition (GMNER) based on a multimodal large language model (MLLM) (Bai et al., 2025). Instead of decomposing the task into separate stages such as textual NER and Object Detection, our framework formulates GMNER as a structured generation problem conditioned jointly on the image and text. The MLLM encodes multimodal inputs into a shared representation and directly generates complete entity records in a single pass, enabling joint optimization of entity recognition, semantic typing, and visual grounding while avoiding error propagation in pipeline-based methods. An overview of the framework is illustrated in **Figure** 2.

#### 3.2.1 Instruction-Tuned Generative Formulation with Chain-of-Thought

We formulate Grounded Multimodal Named Entity Recognition (GMNER) as an instruction-tuned conditional generation task based on a multimodal large language model (MLLM). Given an image–text pair $(I, T)$, the model input is constructed by prepending a task-specific instruction to the multimodal input:

$$\text{Input} = [\text{Instruction}; (I, T)] \tag{2}$$

The MLLM generates outputs autoregressively in a single sequence composed of two parts: a reasoning sequence $R$ and a sequence of structured entity records corresponding to the entity set $\mathcal{Y} = (e_i, c_i, b_i)_{i=1}^{N}$. Each entity record is serialized using the following output schema:

$$e_i \mid c_i \mid [x_i^1, , y_i^1, , x_i^2, , y_i^2] \tag{3}$$

where $e_i$ denotes the entity span, $c_i$ its semantic type, and $[x_i^1, y_i^1, x_i^2, y_i^2]$ the grounding bounding box. All entity records are concatenated sequentially to form the final structured prediction.

To enable the model to adaptively determine whether visual evidence or implicit knowledge is informative for accurate grounding, we incorporate chain-of-thought (CoT) reasoning into the generation process. This design allows the model to avoid indiscriminate reliance on potentially noisy visual cues, while facilitating the joint optimization of entity recognition, visual grounding, and implicit knowledge reasoning within a unified generative framework.

Formally, the generation process is expressed as:

$$[R; \mathcal{Y}] = f_\theta([\text{Instruction}; (I, T)]) \tag{4}$$

where $f_\theta$ denotes the MLLM.

During training, reasoning sequences $R$ are provided as supervision signals generated by a stronger external large language model via

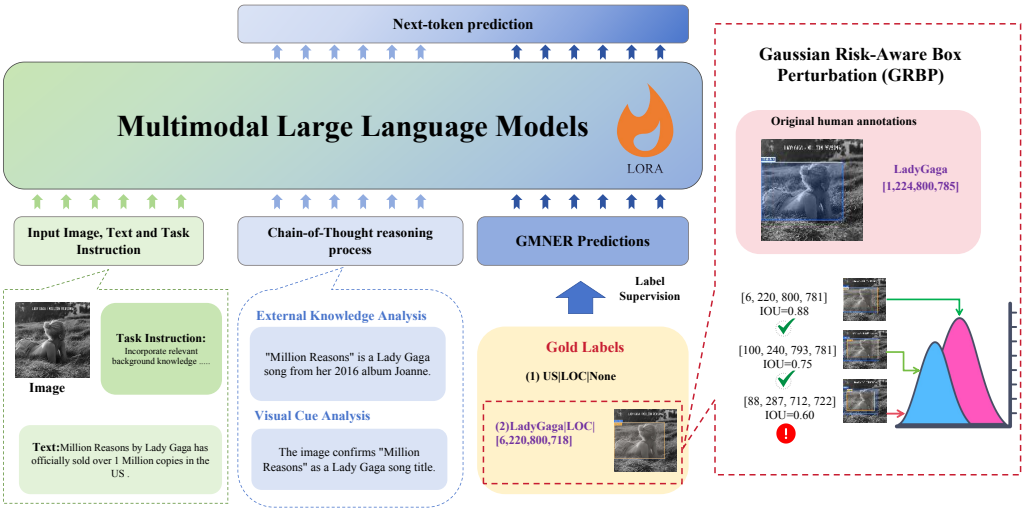



Figure 2: Overview of **E2E-GMNER**. Given an image, text, and task instruction, a LoRA-adapted multimodal large language model performs chain-of-thought multimodal reasoning (visual cue analysis and background knowledge analysis) and generates GMNER outputs—entity mentions, semantic types, and grounded bounding boxes—via next-token prediction. During training, gold entity-type and box annotations provide supervision, and we further introduce **Gaussian Risk-Aware Box Perturbation (GRBP)** to replace hard box targets with Gaussian-weighted soft supervision by probabilistically perturbing the ground-truth box (larger perturbations receive lower probability), improving robustness to annotation noise and discretization effects.

API-based inference, enabling effective chain-of-thought instruction tuning. At inference time, no external models are used, and the MLLM autonomously generates both reasoning and structured predictions in an end-to-end manner.

#### 3.2.2 Gaussian Risk-Aware Box Perturbation (GRBP)

Recent generative vision–language grounding paradigms (Bai et al., 2025; Peng et al., 2023) commonly formulate grounding as a sequence generation problem, where bounding box coordinates are directly predicted as discrete tokens. While this formulation enables end-to-end training, it also introduces significant challenges due to annotation noise and coordinate discretization errors. Under hard supervision, even small geometric deviations between predicted and ground-truth boxes can incur disproportionately large training penalties, leading to unstable optimization and reduced robustness.

To improve the robustness of generative box prediction, we introduce **Gaussian Risk-Aware Box Perturbation (GRBP)** as a supervision strategy during training. Instead of supervising the model with a single deterministic bounding box target, GRBP applies probabilistic perturbations to the ground-truth box coordinates. Intuitively, larger perturbations correspond to a higher risk of deviating from valid object regions, and vice versa. Based on this observation, we adopt a Gaussian prior to model distributions over box center locations and scales, subject to an intersection-over-union (IoU) constraint with the original box. This formulation produces a set of soft supervision targets that assign higher probability to boxes closer to the ground truth while tolerating small geometric deviations, thereby preserving empirical risk minimization while improving robustness.

Formally, for each ground-truth bounding box $b$, we generate a perturbed box $\tilde{b}$ through the following steps: (i) **Center Perturbation**: the box center is shifted by Gaussian noise proportional to the box width and height, controlled by a hyperparameter $\beta$; (ii) **Scale Perturbation**: the box width and height are perturbed using multiplicative Gaussian noise, controlled by $\gamma$, with minimum box size and scale bounds applied to avoid degeneracy; (iii) **IoU Guard**: if the perturbed box satisfies $\mathrm{IoU}(\tilde{b}, b) \geq \tau$, it is accepted; otherwise, the perturbation process is resampled up to $T$ times, and the original box $b$ is used as a fallback. All perturbed boxes are clipped to the image boundaries. The overall procedure is summarized in Algorithm 1.

### 3.3 Training Objective and Inference

**Training Objective.** We train E2E-GMNER using a standard autoregressive maximum likeli-

**Algorithm 1:** GRBP: IoU-Guarded Gaussian Box Perturbation

**Input:** Ground-truth box $b = (x_1, y_1, x_2, y_2)$, image size $(W, H)$, jitter $\beta, \gamma$, IoU threshold $\tau$, max tries $T$, min size $m$, scale bounds $[s_{\min}, s_{\max}]$.

**Output:** Perturbed box $\tilde{b}$.

$(c_x, c_y, w, h) \leftarrow \mathrm{ToCenterSize}(b)$

**if** $w < m$ ***or*** $h < m$ **then**

  **return** $b$

**for** $t \leftarrow 1$ **to** $T$ **do**

  `// Center jitter (relative to box size)`

  draw $\delta_x, \delta_y \sim \mathcal{N}(0, \beta^2)$

  $c'_x \leftarrow c_x + \delta_x \cdot w$

  $c'_y \leftarrow c_y + \delta_y \cdot h$

  `// Scale jitter (multiplicative, bounded)`

  draw $\epsilon_w, \epsilon_h \sim \mathcal{N}(0, \gamma^2)$

  $a_w \leftarrow \min(s_{\max}, \max(s_{\min}, 1 + \epsilon_w))$

  $a_h \leftarrow \min(s_{\max}, \max(s_{\min}, 1 + \epsilon_h))$

  $w' \leftarrow \max(m, w \cdot a_w)$

  $h' \leftarrow \max(m, h \cdot a_h)$

  $\tilde{b} \leftarrow \mathrm{ToBox}(c'_x, c'_y, w', h')$

  **if** $\mathrm{IoU}(\tilde{b}, b) \geq \tau$ **then**

    **return** $\tilde{b}$

**return** $b$

hood objective over the generated output sequence. Given an image–text pair $(I, T)$ and its corresponding supervision consisting of a reasoning sequence $R$ and structured entity records $\mathcal{Y}$, the training objective is defined as:

$$\mathcal{L} = -\sum_t \log p_\theta(y_t \mid y_{<t}, \text{Instruction}, I, T), \tag{5}$$

where $\{y_t\}$ denotes the tokens in the concatenated output sequence $[R; \mathcal{Y}]$, and $p_\theta$ is parameterized by the multimodal large language model.

For entity grounding, bounding box coordinates are serialized as discrete tokens following the output schema in Eq. 3. During training, Gaussian Risk-Aware Box Perturbation (GRBP) is applied to generate perturbed bounding boxes, which serve as soft supervision targets for coordinate generation. This noise-aware supervision reduces the sensitivity of the loss to minor geometric deviations while remaining compatible with standard token-level likelihood optimization.

**Inference.** At inference time, E2E-GMNER operates in a fully end-to-end manner without relying on any external teacher models or API-based reasoning supervision. Given a test image–text pair $(I, T)$ and the task instruction, the model autoregressively generates a reasoning sequence followed by structured entity predictions, including entity mentions, semantic types, and grounding bounding boxes.

Importantly, although chain-of-thought reasoning is used as an auxiliary supervision signal during training, no teacher-generated reasoning is required at inference. The model autonomously determines when visual evidence or implicit knowledge is informative and produces final GMNER outputs in a single forward generation pass. This design ensures that inference remains efficient and self-contained, while retaining the benefits of adaptive reasoning learned during instruction tuning.

## 4 Experiment

### 4.1 Datasets

We evaluate on two social-media Grounded Multimodal NER benchmarks: **Twitter-GMNER** (Yu et al., 2023) and **Twitter-FMNERG** (Wang et al., 2023). Both datasets annotate image–text pairs with entity spans, types, and grounding regions in the image. Twitter-GMNER provides four coarse-grained entity types, while Twitter-FMNERG expands the label space to eight coarse-grained types and 51 fine-grained subtypes for more detailed evaluation.

### 4.2 Evaluation Metrics

Following (Yu et al., 2023), we report F1 for the overall **GMNER** task and its two subtasks: **Multimodal Named Entity Recognition (MNER)** and **Entity Extraction and Grounding (EEG)**. MNER evaluates span and type correctness, while EEG evaluates span extraction with grounding correctness under the standard intersection-over-union criterion. The overall GMNER score requires both correct recognition and correct grounding. All formal definitions and formulas are deferred to Appendix A.

### 4.3 Baselines

To evaluate our framework, we compare against three categories of baselines. The first category

| Methods | Twitter-GMNER | | | Twitter-FMNERG | | |
|---|---|---|---|---|---|---|
| | GMNER | MNER | EEG | GMNER | MNER | EEG |
| GMDA† (Li et al., 2024b) | 58.61 | - | - | 47.37 | - | - |
| GEM† (Wang et al., 2024) | 59.83 | 83.15 | 63.19 | 50.54 | 68.09 | 63.59 |
| RiVEG† (Li et al., 2024a) | 63.80 | 82.89 | 66.92 | - | - | - |
| MAKAR† (Lin et al., 2025) | 71.88 | 86.38 | 74.64 | 60.54 | 71.24 | 75.66 |
| GPT4o (Hurst et al., 2024) | 41.29 | 65.07 | 44.95 | 32.37 | 52.26 | 41.60 |
| GVATT-OD-EVG (Lu et al., 2018) | 48.57 | 76.26 | 53.32 | 40.32 | 60.35 | 54.35 |
| UMT-OD-EVG (YU et al.) | 50.29 | 78.58 | 54.78 | 41.32 | 61.63 | 54.43 |
| UMGF-OD-EVG (Zhang et al., 2021) | 51.67 | 78.83 | 55.74 | 41.92 | 61.79 | 54.75 |
| ITA-OD-EVG (Wang et al., 2022) | 51.56 | 79.37 | 55.69 | 42.78 | 63.21 | 57.26 |
| MMT5 / BARTMNER-OD-EVG (Yu et al., 2023) | 52.45 | 80.39 | 55.66 | 45.21 | 66.61 | 58.18 |
| H-Index (Yu et al., 2023) | 56.41 | 79.73 | 61.18 | 46.55 | 64.84 | 60.46 |
| TIGER (Wang et al., 2023) | 57.48 | - | - | 47.20 | 64.91 | 61.96 |
| MQSPN (Tang et al., 2025a) | 58.76 | 80.43 | 62.40 | 47.86 | 66.83 | 61.95 |
| UnCo (Tang et al., 2025b) | 58.83 | 79.55 | 63.49 | 48.17 | 65.06 | 62.73 |
| E2E-GMNER-7B(Ours) | 63.94 | 77.65 | 66.12 | 54.32 | 65.67 | 66.78 |

Table 1: Performance comparison of different methods on Twitter-GMNER and Twitter-FMNERG datasets. † indicates the methods using additional data or knowledge augmentation.

includes methods that leverage external knowledge or additional data beyond the standard training set. Specifically, GMDA† (Li et al., 2024b), GEM† (Wang et al., 2024), RiVEG† (Li et al., 2024a), and MAKAR† (Lin et al., 2025) incorporate auxiliary resources—such as pre-trained vision-language models, extra datasets, or structured knowledge—to enhance multimodal entity recognition and grounding.

The second category consists of pipeline-based approaches, and we further divide it into two subgroups. The first subgroup (GPT4o (Hurst et al., 2024), GVATT-OD-EVG (Lu et al., 2018), UMT-OD-EVG (YU et al.), UMGF-OD-EVG (Zhang et al., 2021), ITA-OD-EVG (Wang et al., 2022), and MMT5/BARTMNER-OD-EVG (Yu et al., 2023)) follows an MNER-first paradigm: they first extract textual entities using a named entity recognition model and then ground these entities to visual regions via object detectors or alignment modules. The second subgroup (H-Index (Yu et al., 2023), TIGER (Wang et al., 2023), MQSPN (Tang et al., 2025a), and UnCo (Tang et al., 2025b)) adopts a feature-fusion strategy: they independently encode text with a language encoder and extract salient visual regions using an object detector, then fuse these representations through cross-modal interaction layers to jointly predict entities and their grounded spans. Despite their strong performance, both subgroups suffer from limited end-to-end optimization and error propagation between stages.

Finally, we include our proposed models, which are fully end-to-end generative frameworks. Unlike prior work, they jointly generate textual entities and their corresponding visual bounding boxes in a single sequence without relying on external resources or intermediate pipelines, enabling tighter cross-modal alignment and more robust grounding.

### 4.4 Main result

In Table 1, we compare the performance of our method with several state-of-the-art approaches under three evaluation metrics: GMNER, MNER, and EEG, on the Twitter-GMNER and Twitter-FMNERG datasets. The compared baselines can be broadly categorized into two groups: methods that rely on external knowledge sources and those that do not. Overall, the experimental results show that our proposed E2E-GMNER consistently outperforms all state-of-the-art methods that do not leverage external knowledge, and achieves competitive performance even when compared with methods that explicitly incorporate external knowledge sources. Based on these results, we draw the following conclusions.

(1) As the first fully end-to-end framework for GMNER, E2E-GMNER outperforms all pipeline-based methods except MAKAR, which relies on external knowledge sources.[1] This demonstrates that jointly modeling entity recognition, visual grounding, semantic understanding, and knowledge reasoning in a unified generative framework effec-

[1]This method may be less cost-effective in practice, as it requires full-parameter fine-tuning on 8×A100 GPUs. In addition, its reliance on an external search engine necessitates internet access.

| Method | GMNER | FMNERG |
|---|---|---|
| E2E-GMNER | 63.94 | 54.32 |
| w/o CoT | 62.57 | 53.96 |
| w/o GRBP | 61.51 | 52.92 |

Table 2: Ablation study of E2E-GMNER on GMNER and FMNERG datasets.

| Teacher Prompt | GMNER | FMNERG |
|---|---|---|
| Qwen2.5-VL-72B | 63.94 | 54.32 |
| Qwen2.5-VL-7B | 61.23 | 52.74 |
| GPT4o | 63.77 | 53.96 |

Table 3: Performance comparison of different teacher prompts in generating CoT data on GMNER and FMNERG datasets.

tively mitigates error accumulation in cascaded architectures and enables more effective joint optimization.

(2) Compared with knowledge-augmented approaches, E2E-GMNER still achieves highly competitive performance, indicating that our method effectively exploits the implicit knowledge acquired by multimodal large language models during pretraining and successfully adapts it to the GMNER task without requiring explicit external knowledge retrieval.

### 4.5 Ablation Study

To assess the contribution of the key components in our end-to-end generative framework, we conduct an ablation study by individually removing Chain-of-Thought (CoT) reasoning and Gaussian Risk-Aware Box Perturbation (GRBP). The results are summarized in Table 2.

Removing CoT reasoning (*w/o CoT*) leads to a clear and consistent performance degradation on both GMNER and FMNERG. This indicates that CoT reasoning plays a critical role in enabling the model to perform structured multimodal reasoning, particularly in determining when visual evidence or implicit background knowledge should be leveraged. Without this reasoning mechanism, the model is more prone to relying on noisy or irrelevant visual cues, which adversely affects its ability to jointly perform entity recognition and grounding.

Eliminating GRBP (*w/o GRBP*) also results in a noticeable performance drop, though to a lesser extent compared with removing CoT. This observation suggests that GRBP effectively improves the robustness of generative grounding by alleviating the sensitivity of discrete bounding box prediction to annotation noise and discretization errors. By providing noise-aware supervision, GRBP helps stabilize training and improves generalization in multimodal grounding scenarios.

### 4.6 Impact of Teacher Prompts on CoT Data Generation

To investigate how the choice of teacher model affects the quality of chain-of-thought (CoT) supervision and the resulting GMNER performance, we compare different teacher prompts used during CoT data generation, as summarized in Table 3. The results indicate that larger teacher models generally lead to better downstream performance, suggesting that increased model capacity enables the generation of more informative and reliable reasoning traces. In addition, although GPT4o produces competitive results, it underperforms compared with the Qwen2.5-VL-72B series. A plausible explanation is that our finetuned backbone model belongs to the Qwen-2.5-VL family, whose reasoning patterns and multimodal representations are more consistent with those of the Qwen-based teachers, resulting in better alignment during instruction tuning. Finally, it is worth noting that all evaluated teacher models yield reasonably strong performance, demonstrating that our training strategy is robust and effective across different teacher choices, rather than being tightly coupled to a specific teacher model.

### 4.7 Perturbation Strength Analysis

We conduct a systematic analysis of different Gaussian perturbation strengths to investigate the impact of Gaussian Risk-Aware Box Perturbation (GRBP) on both entity grounding and overall GMNER performance, as shown in Figure 3. Specifically, we compare models trained with varying values of the perturbation factors $\beta$ and $\gamma$ in terms of three metrics: Acc@0.5, which reflects the recall of correctly grounded entities with valid bounding boxes; the overall GMNER F1 score; and the MeanIoU between predicted bounding boxes and gold annotations.

The results show that a moderate perturbation strength achieves the best overall performance, with $\beta = \gamma = 0.03$ yielding the most favorable

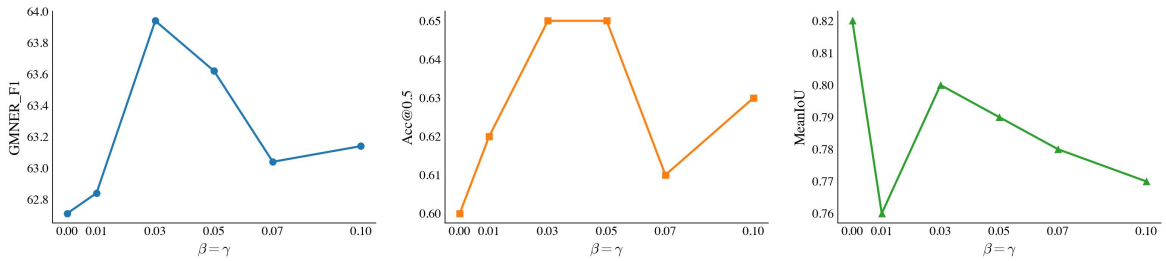



Figure 3: Impact of Gaussian perturbation strength on GMNER performance. The figure reports Acc@0.5, GMNER F1 score, and MeanIoU under different values of the Gaussian perturbation factors $\beta$ and $\gamma$, illustrating the effect of Gaussian Risk-Aware Box Perturbation (GRBP) on entity grounding recall, overall extraction performance, and localization accuracy.

trade-off across metrics. Compared with training without perturbation, the GMNER F1 score exhibits only a marginal difference, while Acc@0.5 shows a substantially larger improvement. This indicates that GRBP primarily enhances the recall of the grounding process, enabling the model to correctly associate more entities with valid visual regions, even when the predicted boxes are not perfectly aligned with the gold annotations.

Interestingly, under $\beta = \gamma = 0.03$, the MeanIoU between predicted and gold bounding boxes slightly decreases compared with the no-perturbation setting. However, this minor reduction in localization precision is accompanied by a clear improvement in grounding recall and overall extraction performance. This observation suggests that GRBP encourages the model to be more tolerant of small geometric deviations during training, which improves its ability to generalize and recover valid grounding regions, rather than overfitting to exact box coordinates.

Overall, this analysis shows that GRBP provides a favorable robustness–precision trade-off: moderate Gaussian perturbations slightly reduce localization accuracy in terms of MeanIoU, but substantially improve grounding recall, resulting in better end-to-end GMNER performance.

## 5 Conclusion

In this work, we presented E2E-GMNER, the first fully end-to-end generative framework for grounded multimodal named entity recognition that jointly models entity extraction, semantic classification, and visual grounding within a single vision–language model. By eliminating cascaded pipeline components, our approach enables unified optimization and mitigates error accumulation inherent in prior methods. The incorporation of chain-of-thought instruction tuning allows the model to adaptively leverage visual cues and implicit knowledge only when beneficial, while Gaussian Risk-Aware Box Perturbation improves the robustness and stability of generative grounding under annotation noise. Experimental results on two widely used GMNER benchmarks show that E2E-GMNER consistently outperforms or matches strong baselines without relying on external object detectors or online knowledge sources. These findings highlight the promise of end-to-end generative paradigms for structured multimodal understanding and suggest future directions for extending such frameworks to broader multimodal information extraction tasks.

## Limitations

Although E2E-GMNER demonstrates strong performance on grounded multimodal named entity recognition, our current framework is specifically designed and evaluated for the GMNER task. The model architecture, output schema, and training objectives are tailored to jointly predict entity spans, semantic types, and grounding boxes, and have not yet been adapted or validated on other multimodal information extraction tasks such as relation extraction or event extraction. In addition, our approach relies on chain-of-thought supervision during training, which introduces extra annotation or teacher-model inference cost. Exploring task-agnostic formulations and reducing the dependence on auxiliary reasoning supervision remain important directions for future work.

## Acknowledgement

This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China (NSFC) [Grant No. 62506058]

## References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Fan Li, Jianxing Yu, Jielong Tang, Wenqing Chen, Hanjiang Lai, Yanghui Rao, and Jian Yin. 2025. Answering complex geographic questions by adaptive reasoning with visual context and external commonsense knowledge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25498–25514, Vienna, Austria. Association for Computational Linguistics.

Jinyuan Li, Han Li, Di Sun, Jiahao Wang, Wenkun Zhang, Zan Wang, and Gang Pan. 2024a. Llms as bridges: Reformulating grounded multimodal named entity recognition. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1302–1318.

Ziyan Li, Jianfei Yu, Jia Yang, Wenya Wang, Li Yang, and Rui Xia. 2024b. Generative multimodal data augmentation for low-resource multimodal named entity recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7336–7345.

Xinkui Lin, Yuhui Zhang, Yongxiu Xu, Kun Huang, Hongzhang Mu, Yubin Wang, Gaopeng Gou, Li Qian, Li Peng, Wei Liu, and 1 others. 2025. Makar: a multi-agent framework based knowledge-augmented reasoning for grounded multimodal named entity recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6121–6141.

Jintao Liu, Chenglong Liu, and Kaiwen Wei. 2024. Multi-view prompt for fine-grained multimodal named entity recognition and grounding. In *ECAI 2024*, pages 2693–2700. IOS Press.

Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S Rosenblum. 2019. Mmkg: multi-modal knowledge graphs. In *European Semantic Web Conference*, pages 459–474. Springer.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. 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.

Hyunjong Ok, Taeho Kil, Sukmin Seo, and Jaeho Lee. 2024. Scanner: Knowledge-enhanced approach for robust multi-modal named entity recognition of unseen entities. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7718–7730.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Jielong Tang, Zhenxing Wang, Ziyang Gong, Jianxing Yu, Xiangwei Zhu, and Jian Yin. 2025a. Multi-grained query-guided set prediction network for grounded multimodal named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25246–25254.

Jielong Tang, Yang Yang, Jianxing Yu, Zhen-Xing Wang, Haoyuan Liang, Liang Yao, and Jian Yin. 2025b. Unco: Uncertainty-driven collaborative framework of large and small models for grounded multimodal ner. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7644–7662.

Jieming Wang, Ziyan Li, Jianfei Yu, Li Yang, and Rui Xia. 2023. Fine-grained multimodal named entity recognition and grounding with a generative framework. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3934–3943.

Xinyu Wang, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang, and Kewei Tu. 2022. 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.

Ziqi Wang, Chen Zhu, Zhi Zheng, Xinhang Li, Tong Xu, Yongyi He, Qi Liu, Ying Yu, and Enhong Chen. 2024. Granular entity mapper: Advancing fine-grained multimodal named entity recognition and grounding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3211–3226.

Jianfei YU, Jing JIANG, Li YANG, and Rui XIA. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer.(2020). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3342–3352.

Jianfei Yu, Ziyan Li, Jieming Wang, and Rui Xia. 2023. Grounded multimodal named entity recognition on social media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9141–9154.

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021. 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.

# A Formulas and Metrics

$$C_{e,t} = \begin{cases} 1 & \text{if } \hat{e} = e \wedge \hat{t} = t, \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

$$C_b = \begin{cases} 1 & \text{if } \hat{b} = \varnothing \wedge b = \varnothing \vee \mathrm{IoU}(\hat{b}, b) \geq 0.5, \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

For **MNER**, a prediction is correct if and only if $C_{e,t} = 1$.

For **EEG**, a prediction is correct if and only if $\hat{e} = e$ and $C_b = 1$.

For **MNERG**, a prediction is correct if and only if

$$C = C_{e,t} \cdot C_b = 1. \tag{8}$$

Let $\#\mathrm{correct}$, $\#\mathrm{pred}$, and $\#\mathrm{gold}$ represent the number of correct predictions, total predictions, and gold records, respectively. Precision, recall, and F1 score are computed as:

$$\mathrm{Pre} = \frac{\#\mathrm{correct}}{\#\mathrm{pred}}, \qquad \mathrm{Rec} = \frac{\#\mathrm{correct}}{\#\mathrm{gold}}, \tag{9}$$

$$\mathrm{F1} = \frac{2 \cdot \mathrm{Pre} \cdot \mathrm{Rec}}{\mathrm{Pre} + \mathrm{Rec}}. \tag{10}$$

In addition to the standard metrics, we also report the following:

The accuracy at specific IoU thresholds (e.g., 0.5 and 0.75) is used to evaluate the model's performance in extracting and grounding entities. Let $N_{\text{matched}}$ denote the number of matched pairs at a given threshold. The accuracy at IoU threshold $\mathrm{IoU}_{\mathrm{thr}}$ is computed as:

$$\text{Acc@IoU}_{\text{thr}} = \frac{N_{\text{matched}}}{\#\text{gold with box}}. \tag{11}$$

The overall mean IoU is the average IoU between all gold entity boxes and their respective predicted boxes. It is computed as:

$$\text{Overall Mean IoU} = \frac{\sum_{i=1}^{N_{\text{gold}}} \mathrm{IoU}(b_i, \hat{b}_i)}{N_{\text{gold}}}. \tag{12}$$

# B Implementation Details

All experiments are conducted on a single NVIDIA RTX 5090 GPU. To construct Chain-of-Thought (CoT) training data, we leverage three strong vision-language teacher models: `Qwen2.5-VL-72B-Instruct`, `Qwen2.5-VL-7B-Instruct`, and `GPT-4o`. Our primary experiments fine-tune the `Qwen2.5-VL-7B` vision-language models as backbones. We optimize using AdamW with a learning rate of $4 \times 10^{-5}$ and weight decay of $1 \times 10^{-3}$. The per-device batch size is set to 2, and we apply gradient accumulation over 8 steps, resulting in an effective batch size of 16. The maximum input sequence length is limited to 2048 tokens. To improve training efficiency and reduce memory footprint, we employ Low-Rank Adaptation (LoRA) during fine-tuning.