Prompt-Guided Structured Multimodal NER with SVG and ChatGPT

Yuzhou Ma , Haolong Qian , Wei Li *

Posted Date: 13 February 2026

doi: 10.20944/preprints202602.1129.v1

Keywords: Multimodal Named Entity Recognition; Scalable Vector Graphics; ChatGPT; Graph Attention Networks

 doi:10.20944/preprints202602.1129.v1

Article

Prompt-Guided Structured Multimodal NER with SVG and ChatGPT

Yuzhou Ma [1], Haolong Qian [2] and Wei Li [2,*]

[1] School of Cyber Security and Defense, Anhui Police College, Hefei 238076, China

[2] College of Computer Science, Sichuan Normal University, Chengdu 610101, China

* Correspondence: liw@sicnu.edu.cn

Abstract

Multimodal Named Entity Recognition (MNER) leverages both textual and visual information to improve entity recognition, particularly in unstructured scenarios such as social media. While existing approaches predominantly rely on raster images (e.g., JPEG, PNG), Scalable Vector Graphics (SVG) offer unique advantages in resolution independence and structured semantic representation—an underexplored potential in multimodal learning. To fill this gap, we propose MNER-SVG, the first framework that incorporates SVG as a visual modality and enhances it with ChatGPT-generated auxiliary knowledge. Specifically, we introduce a Multimodal Similar Instance Perception Module that retrieves semantically relevant examples and prompts ChatGPT to generate contextual explanations. We further construct a Full Text Graph and a Multimodal Interaction Graph, which are processed via Graph Attention Networks (GATs) to achieve fine-grained cross-modal alignment and feature fusion. Finally, a Conditional Random Field (CRF) layer is employed for structured decoding. To support evaluation, we present SvgNER, the first MNER dataset annotated with SVG-specific visual content. Extensive experiments demonstrate that MNER-SVG achieves state-of-the-art performance with an F1 score of 82.23%, significantly outperforming both text-only and existing multimodal baselines. This work validates the feasibility and potential of integrating vector graphics and large language model–generated knowledge into multimodal NER, opening a new research direction for structured visual semantics in fine-grained multimodal understanding.

Keywords: Multimodal Named Entity Recognition; Scalable Vector Graphics; ChatGPT; Graph Attention Networks

1. Introduction

Named Entity Recognition (NER) is a fundamental NLP task that identifies entities such as persons, locations, and organizations [1,2]. With social media's rise, unstructured and multimodal user-generated content challenges traditional NER, promoting Multimodal Named Entity Recognition (MNER). By integrating visual and linguistic information, MNER has emerged as a key paradigm driving artificial intelligence toward a deeper understanding of the complex real world. Leveraging visual information, MNER enhances recognition performance and proves effective in tasks like multimodal sentiment analysis [3], vehicle Engineering[4], machine translation [5], and visual dialogue [6]. Its importance in practical applications continues to grow.

However, existing MNER approaches largely rely on raster images (e.g., JPEG, PNG), which are resolution-dependent and often suffer from information loss during compression or scaling. This limits the accuracy and interpretability of cross-modal understanding, especially in scenarios such as human-computer interaction, interface understanding, and information visualization. Scalable Vector Graphics (SVG), in contrast, describe images using mathematical primitives rather than pixels, offering resolution independence, structural clarity, and editability. These properties make SVGs particularly suitable for structured visual content, such as diagrams, logos, infographics, and interface icons. SVGs are widely adopted in modern digital platforms, including educational tools,

 doi:10.20944/preprints202602.1129.v1

scientific visualizations, mobile icons, and web interfaces. Tools like Iconfont, Figma, and AdobeXD use SVG as a standard format, highlighting its relevance for emerging multimodal understanding scenarios. Despite this, the potential of SVG as a visual modality in multimodal learning remains unexplored.

In this work, we explore the value of SVG in a downstream task, namely multimodal named entity recognition (MNER) [7]. MNER aims to identify and categorize entities in text with the help of visual context, improving recognition accuracy in noisy and informal user-generated content. As shown in Figure 1, text alone may misclassify "Charlie" as a person, while visual cues help correctly label it as "MISC". By using SVG instead of raster images, we can provide more structured and informative visual signals, facilitating precise cross-modal semantic alignment.

Figure 1. An example of multimodal named entity recognition. Relying on text alone may lead to incorrect entity classification.

Moreover, large language models (LLMs) such as ChatGPT have demonstrated strong contextual learning and generalization capabilities across NLP and multimodal tasks[8,9]. Leveraging this, we generate auxiliary knowledge using ChatGPT to further enhance MNER performance, providing semantic cues that complement the SVG visual information and improving cross-modal understanding.

We propose *MNER-SVG*, a framework that incorporates vector graphics and ChatGPT-generated auxiliary knowledge to enhance cross-modal understanding. A Multimodal Similar Instance Perception Module selects semantically related examples and combines ChatGPT knowledge with textual inputs for downstream modeling. To better align text and visual semantics, we construct a Full Text Graph to capture global contextual relations and a Multimodal Interaction Graph for fine-grained visual-textual alignment. Graph Attention Networks (GATs) enable adaptive cross-modal fusion, followed by a Conditional Random Field (CRF) decoder for entity prediction. To support evaluation, we construct **SvgNER**, the first MNER dataset with SVG-specific annotations.

Our contributions are summarized as follows:

- We are the first to introduce Scalable Vector Graphics (SVG) as a visual modality in multimodal learning, emphasizing its resolution independence and structural semantics.
- We propose an integrated framework *MNER-SVG* combining SVG representations, ChatGPT-generated auxiliary knowledge, and graph-based cross-modal interactions via Graph Attention Networks (GATs), enhancing both performance and interpretability.
- We present SvgNER, the first SVG-annotated dataset for MNER. Experimental results demonstrate the effectiveness of our approach, validating SVG as a promising modality for fine-grained multimodal understanding.

The remainder of this paper is organized as follows: Section 2 reviews related work on multimodal named entity recognition. Section 3 describes the proposed MNER-SVG model in detail.

 doi:10.20944/preprints202602.1129.v1

Section 4 presents experimental results and analyses. Section 5 discusses the theoretical and practical implications, followed by conclusions and future work in Section 6.

2. Related Works

2.1. Multimodal Named Entity Recognition

In recent years, some progress has been made in the research on multimodal named entity recognition. Zhang et al. [10] defines the novel task of named entity recognition (NER) for multimodal tweets by incorporating visual information, and proposes an Adaptive Co-attention Network (ACN) to fuse textual and visual features, with gated fusion and filtration gates for noise reduction. They build a large-scale manually annotated multimodal tweet dataset, and validate experimentally that the method outperforms text-only SOTA models, demonstrating visual information's effectiveness for tweet NER. Yu et al. [11] proposes a Unified Multimodal Transformer (UMT) for Multimodal Named Entity Recognition (MNER) in social media posts. It introduces a multimodal interaction module to generate image-aware word representations and an auxiliary text-based entity span detection module to mitigate visual bias. The model achieves state-of-the-art performance on two benchmark Twitter datasets. Zhang et al. [12] proposes a unified multimodal graph fusion (UMGF) model for named entity recognition with images. It constructs a graph linking words and visual objects, uses stacked fusion layers for cross-modal interaction, and achieves state-of-the-art results on Twitter datasets. Xu et al. [13] enhances multimodal named entity recognition by integrating topic prompts from images and using multi-curriculum denoising to mitigate noise. This approach strengthens model reliability and performance in complex multimodal settings. Li et al. [14] proposes the AMLR method that advances multimodal NER through entity-level fusion, expanding text to multi-scale representations before cross-modal attention. This entity-level reinforcement outperforms token-level methods on Twitter datasets. Zhang et al. [15] proposes PGMNER, a framework integrating object detection and prompts to enhance multimodal NER and co-reference resolution. It achieves state-of-the-art results on the CIN dataset by improving cross-modal alignment. Mu et al. [16] proposes MCIRP, a model for multi-image MNER that employs relation propagation to filter irrelevant images and multi-granularity fusion for better cross-modal interaction, achieving state-of-the-art results.

2.2. Prompt Engineering for NER with ChatGPT

Since its release in November 2022, ChatGPT (Chat Generative Pre-trained Transformer) has demonstrated strong generalization across various tasks, including Information Extraction (IE). Despite its closed-source nature, recent studies have explored its zero-shot capabilities. For example, ChatIE [17] formulates IE as a prompting task for ChatGPT and conducts comprehensive performance evaluations. Chen et al. [18] proposes a framework addressing medical NER challenges by creating a large multi-scenario dataset, and robust prompt training, achieving significant performance gains. De et al. [19] proposes AgNER-BERTa and AgRE-BERTa models with advanced data augmentation for agricultural NER and RE. Further investigations [20,21] assess ChatGPT's interpretability, calibration, and robustness, highlighting its potential while acknowledging limitations such as overconfidence and occasional misclassification.

As a core subtask of IE, Named Entity Recognition (NER) benefits significantly from prompt-based interaction with large language models (LLMs). Unlike traditional fine-tuning approaches (e.g., BERT [22]), LLMs like ChatGPT support In-Context Learning (ICL), enabling few-shot or zero-shot NER without parameter updates [23]. This paradigm offers faster adaptation and greater flexibility, particularly for domain-specific or multimodal scenarios.

With ongoing advances in Multimodal NER (MNER), integrating visual information with LLM-driven prompt engineering opens new possibilities. The fusion of textual and visual cues, combined with ChatGPT's generative reasoning, presents an efficient and scalable approach to entity recognition in complex multimodal settings.

 doi:10.20944/preprints202602.1129.v1

2.3. Summary

Although numerous achievements have been made in the fields of multimodal named entity recognition and named entity recognition based on large language models, several issues remain. Firstly, no researcher has yet conducted multimodal named entity recognition studies on SVG format images, failing to fully leverage the advantages of the SVG vector graphics format to enhance the accuracy of named entity recognition. Secondly, there is still limited research on multimodal named entity recognition based on the large language model ChatGPT.

3. Methods

MNER generally consists of three steps: *text processing*—recognizing named entities; *image processing*—extracting visual cues via object detection and segmentation; and *joint modeling*—fusing textual and visual features to improve recognition.

Multimodal Named Entity Recognition (MNER) extracts named entities from a text sequence $X = \{x_1, x_2, \ldots, x_n\}$ and visual information I to improve classification, treated as a sequence labeling task. The goal is to assign an entity label $Y = \{y_1, y_2, \ldots, y_n\}$ to each token, based on the BIO2 annotation scheme.

Figure 2 illustrates the MNER-SVG framework, comprising four key modules: (1) text and SVG feature extraction, (2) auxiliary knowledge generation via ChatGPT, (3) multimodal interaction and fusion, and (4) CRF decoding. The process begins with extracting features from both modalities, followed by ChatGPT-generated knowledge integration into a Transformer encoder. To align text and visual features, we construct text and multimodal graphs, fused via Graph Attention Networks (GAT). A final CRF layer predicts entity labels, improving recognition accuracy. Importantly, MNER-SVG leverages SVG structural information and ChatGPT-generated knowledge to enhance cross-modal understanding beyond pixel-level visual features.

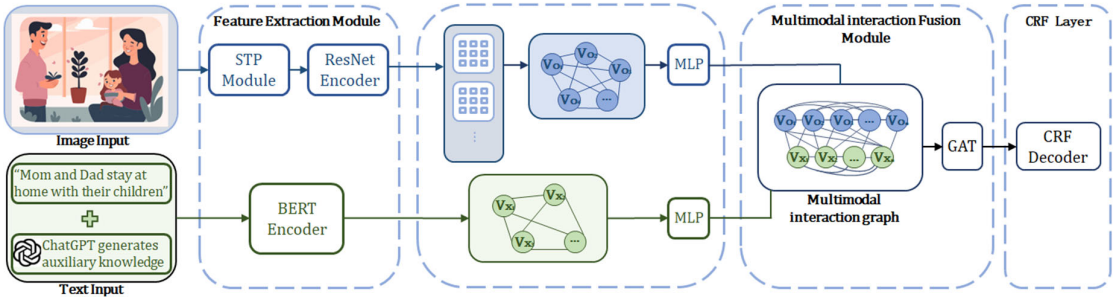

Figure 2. Architecture overview of MNER-SVG.

3.1. Feature Extraction Module

MNER-SVG captures both textual and visual features for multimodal representation.

3.1.1. Textual Representation

For a sentence S, BERT provides contextual embeddings $\hat{H} \in \mathbb{R}^{n \times d}$, where n is the sequence length and d is the hidden dimension. A self-attention mechanism further refines these representations:

$$\mathrm{M} = \mathrm{softmax}\left(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\right), \quad \mathrm{H} = \mathrm{LN}(\hat{\mathrm{H}} + \mathrm{MV}) \tag{1}$$

where LN denotes layer normalization. This enables the model to capture long-range dependencies among tokens.

3.1.2. Visual Representation

 doi:10.20944/preprints202602.1129.v1

To utilize vector-based SVG images, we convert them to raster format using a lightweight SVG2PNG module (Algorithm 1), which applies cairosvg.svg2png and loads results via PIL.Image.

Leveraging SVG Structural and Semantic Information. Although rasterized PNG images are used for CNN processing, MNER-SVG explicitly preserves the inherent structural and semantic information of SVGs. These representations are not limited to pixel-level features, but encode vector-specific relations that support downstream fusion and reasoning. The Full Text Graph captures global text relations, and the Multimodal Interaction Graph models fine-grained text-SVG correspondences. In addition, ChatGPT-generated auxiliary knowledge derived from SVG descriptions provides vector-aware semantic cues that complement visual features extracted by the CNN. This ensures that the advantages of SVG are utilized beyond pixel-level representations.

The conversion pipeline consists of three stages:

- Vector Parsing: Parse XML-based primitives, line segments $L(x_1, y_1, x_2, y_2)$, Bézier curves $B(t)$, and region fills.
- Rasterization: Convert continuous vector coordinates to discrete pixels with resolution $R = (w, h)$: $(x_p, y_p) = (\lfloor x \cdot w \rfloor, \lfloor y \cdot h \rfloor)$.
- Encoding: PNG images $I \in \mathbb{R}^{w \times h \times 4}$ with RGBA channels are compressed using the DEFLATE algorithm: $C = \text{Compress}(I)$.

Algorithm 1. Convert SVG to PNG.

Input: Path to an SVG file

Output: PNG image in RGB format, or None if conversion fails

```
if the SVG file does not exist then
    Raise an error indicating missing file
end if
Initialize an in-memory buffer for output
Attempt to convert the SVG file to PNG format and store in the buffer
Load the PNG image from the buffer and convert it to RGB
return The RGB image
If any error occurs during conversion, return None
```

3.1.3. Visual Feature Extraction

We preprocess SVG images using SVG2PNG, then extract visual features with the pre-trained ResNet-152 model. The images are resized to 224×224 pixels, and features are extracted as $U = (u_1, u_2, \ldots, u_{49})$, where each $u_i \in \mathbb{R}^{2048}$. To align visual features with text features, we apply a linear transformation: $V = W_u^\top U$, where $V = (v_1, v_2, \ldots, v_{49})$ and each $v_i \in \mathbb{R}^d$, used in downstream multimodal fusion tasks.

3.2. ChatGPT-Generated Auxiliary Knowledge Module

To enhance MNER with external knowledge, we leverage ChatGPT to generate auxiliary explanations based on multimodal context (Figure 3). Importantly, the ChatGPT-generated knowledge is derived from SVG descriptions and associated textual annotations, rather than from rasterized PNG images. This ensures that vector-specific structural and semantic information is preserved and provides semantic cues that complement the visual features extracted from CNNs. This process includes three steps: (1) manually annotating a small training subset as high-quality reference data, (2) selecting similar examples via multimodal similarity, and (3) prompting ChatGPT to generate auxiliary knowledge integrated into the input for better entity recognition.

 doi:10.20944/preprints202602.1129.v1

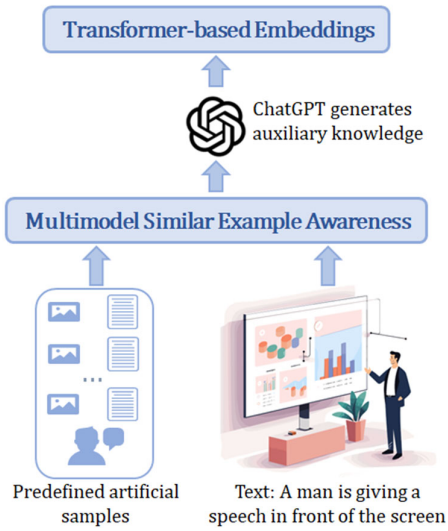

Figure 3. Enhancing MNER via ChatGPT-generated knowledge from SVG inputs.

3.2.1. Reference Sample Annotation

We annotated 200 training samples with entity labels, multimodal explanations, and textual-visual alignment. This curated set serves as few-shot examples to guide ChatGPT in generating relevant auxiliary knowledge.

3.2.2. Multimodal Similarity-Based Example Selection

Since GPT's performance depends heavily on example selection [24], we design a Multimodal Similarity-Based Example Selection (MSEA) module to retrieve contextually relevant pairs from the reference set.

Feature Extraction and Similarity Evaluation: We use existing MNER models to extract fused multimodal features from both test input and reference samples, denoted as $D = \{(t_i, p_i, y_i)\}_{i=1}^{M}$, $G = \{(t_j, p_j, y_j)\}_{j=1}^{N}$. Features $H = M_b(t, p)$ are computed via a shared encoder M_b, where higher cosine similarity indicates greater contextual alignment.

Example Selection Mechanism: We compute cosine similarity between the test input's feature and each annotated sample, selecting the top N most similar examples for ChatGPT's contextual references: $C = \{(t_j, p_j, y_j) \mid j \in I\}$ Precomputing all fusion features allows efficient retrieval during inference.

Prompt Construction: Each prompt follows a fixed format consisting of three parts: (1) a concise task definition ("Extract named entities from the following multimodal input"), (2) few-shot context examples selected based on multimodal similarity, and (3) the current test input (text + SVG).

This structure ensures consistent guidance, improving accuracy and reproducibility.

To illustrate our prompt design, Table 1 provides a real sample of the prompt input and ChatGPT's output auxiliary knowledge.

Table 1. Example prompt and ChatGPT-generated auxiliary knowledge.

Prompt Input (Few-Shot + Test)	ChatGPT Output (Auxiliary Knowledge)
Task: Extract named entities from the following text and SVG description. Example1: Text: "Tesla opens a factory in Berlin." SVG: factory icon. Entities: ["Tesla": ORG, "Berlin": LOC]. Input: Text: "Beijing University organizes a seminar in Shanghai." SVG: university logo icon. Entities: [?] — to be identified.	"The SVG shows a stylized educational institution logo. So 'Beijing University' should be ORG and 'Shanghai' LOC." This auxiliary text helps disambiguate that "Beijing University" is an organization entity in the educational context.

 doi:10.20944/preprints202602.1129.v1

3.3. Multimodal Interaction Fusion

This module includes: (1) *Multimodal interaction graph construction*; and (2) *GAT-based fusion* to propagate and integrate node features, suppressing noise and enhancing cross-modal understanding.

3.3.1. Multimodal Interaction Graph Construction

To model fine-grained text-image relations, we build a graph with two types of nodes: textual words and visual objects. Edges are divided into intra-modal and inter-modal connections.

Intra-modal edges fully connect nodes within the same modality. For inter-modal edges: (1) *Text-to-visual:* Noun phrases (via Stanford Parser) link to detected visual objects; (2) *Cross-modal:* Links connect phrases to matched objects; (3) *Global alignment:* Unmatched words connect to all visual nodes to capture latent semantics.

The resulting graph is denoted as $G_S = (V_S, E_S)$, with adjacency matrix A_S.

3.3.2. Cross-Modal Feature Fusion with GAT

We employ a two-layer Graph Attention Network (GAT) for cross-modal feature fusion (see Figure 4). The attention mechanism adaptively weights neighbor nodes to mitigate visual noise and enhance integration.

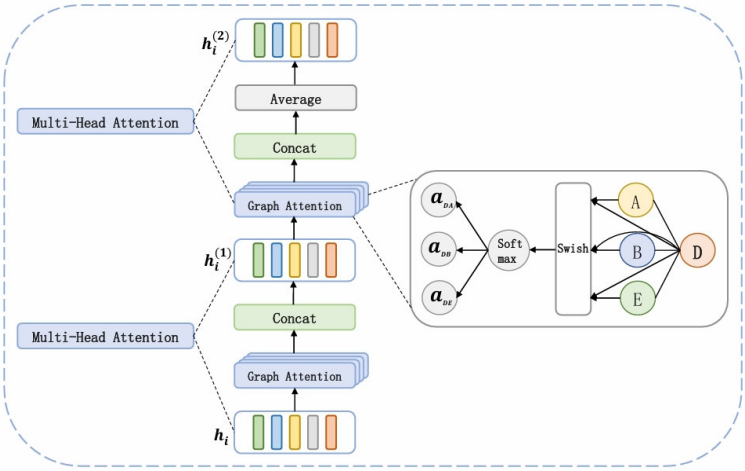

Figure 4. Multimodal interaction graph construction and GAT-based feature fusion.

GAT Fusion. Each layer computes multi-head attention to capture cross-modal interactions. Node features are aggregated from neighbors using attention scores via Softmax. Swish activation improves expressiveness and avoids issues like dying neurons. Outputs from all heads are concatenated and averaged to obtain fused node features.

Node Representation. The final fused representation is:

$$H_S = \left[\widehat{H_X}, \widehat{H_{O_t}}\right] = [\hat{x}_1, \ldots, \hat{x}_n, \hat{o}_1, \ldots, \hat{o}_m] \tag{2}$$

where $\widehat{H_X}$ and $\widehat{H_{O_t}}$ denote fused textual and visual features.

The attention score between nodes i and j is:

$$\mathrm{a_{ij}} = \frac{\exp(\mathrm{Swish}(\mathrm{a^T}[\mathrm{Wh_i} \parallel \mathrm{Wh_j}]))}{\sum_{\mathrm{k}\in\mathrm{N_i}} \exp\left(\mathrm{Swish}(\mathrm{a^T}[\mathrm{Wh_i} \parallel \mathrm{Wh_k}])\right)} \tag{3}$$

with learnable parameters W and a. These adaptive scores guide fine-grained multimodal fusion.

3.4. CRF Decoding Module

 doi:10.20944/preprints202602.1129.v1

To model label dependencies and ensure coherent predictions, we apply a CRF layer over the fused features H_S. The probability of label sequence y is:

$$p(y|H_S) = \frac{\prod_{i=1}^{N} F_i(y_{i-1}, y_i, H_S)}{\sum_{y' \in Y} \prod_{i=1}^{N} F_i(y'_{i-1}, y'_i, H_S)} \quad (4)$$

where F_i models compatibility between adjacent labels and features.

We optimize the log-likelihood:

$$\mathcal{L} = -\sum_{i=1}^{M} \log p(y^{(i)}|H_S^{(i)}) \quad (5)$$

At inference, the most probable label sequence is:

$$\hat{y} = \underset{y' \in Y}{\operatorname{argmax}} p(y'|H_S) \quad (6)$$

The CRF decoder enforces structured output consistency in MNER.

4. Experimental Evaluation

4.1. SvgNER Dataset

Existing vector graphic datasets focus on font generation [25], line drawing [26], or synthesis, but lack multimodal NER integration. We propose SvgNER, a dataset of 2,000+ high-quality SVG images, covering diverse shapes, styles, and semantics. Samples are shown in Figure 5.

Figure 5. Samples from the SvgNER dataset in vector graphics format, which have rich semantic meanings, shapes, and paths.

4.2. Data Structure

To effectively integrate SVG graphics into MNER using deep neural networks, we adopt the BIO tagging scheme (e.g., B-PER, I-LOC, O) , **O**) and follow [27] to split the dataset into train/val/test (70%/15%/15%) . Statistics are shown in Table 2.

Table 2. Statistics of the SvgNER dataset.

Total entities	Beginning of Entity	Inside of Entity	Outside of Entity
2130	903	462	765

4.3. Experimental Details

We conduct experiments on our SvgNER dataset. All models are implemented in PyTorch 2.0 and trained on a single NVIDIA RTX 4060 GPU. We use a max sequence length of 256, batch size 4, and train for 25 epochs. The checkpoint with the highest dev F1 is used for evaluation.

Our MNER-SVG framework adopts BERT-base and a frozen ResNet-152 for textual and visual features, respectively. Cross-modal attention heads are set to 12. Learning rate, dropout, and balancing factor λ are set to 5×10^{-6}, 0.1, and 0.5 via grid search. All results are averaged over three random seeds for robustness.

4.4. Evaluation Metrics

 doi:10.20944/preprints202602.1129.v1

We evaluate models using Precision (P), Recall (R), and F1-score (F1), defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

4.5. Comparison Experiments

We compare MNER-SVG with representative baselines in both text-only and multimodal NER. Baseline models are shown in Table 3.

Table 3. Main technical information of the baseline methods.

Methods	Year	Text Encoder	Image Encoder	Multimodal Fusion	Decoder
Text-based Methods					
BiLSTM-CRF [28]	2015	BiLSTM	N/A	N/A	CRF
BERT-CRF [22]	2018	BERT	N/A	N/A	CRF
RoBERTa-span [29]	2020	RoBERTa	N/A	N/A	N/A
Multimodal Methods (Text + Image)					
UMT [11]	2020	BERT	ResNet (Local)	Transformer + Cross-attention	CRF
MNER-QG [30]	2022	BERT	ResNet-101	Query Grounding + Concatenation	CRF
CAT-MNER [31]	2022	BERT	ResNet-50	Cross-Modal Refinement	CRF
MCIRP[16]	2026	BERT	ResNet	Relational Propagation + Multi-granularity Cross-modal Interaction	CRF

For text-based models, we include BiLSTM-CRF [28], a classic sequence labeling model using BiLSTM with CRF decoding; BERT-CRF [22], which uses BERT for contextual embeddings with a CRF layer; and RoBERTa-span [29], which adopts span-based modeling with auxiliary tasks for entity recognition.

Among multimodal baselines, UMT [11] enhances MNER via entity span detection using visual and textual features; MNER-QG [30] reformulates MNER as a question answering task with query-image-text encoding; and CAT-MNER [31] employs cross-modal attention with knowledge-enhanced feature fusion. MCIRP employs relation propagation to filter irrelevant images and is state-of-the-art result.

Our proposed model, MNER-SVG, is the first designed for Scalable Vector Graphics (SVG) datasets. By leveraging the structured and semantic of SVGs, it enables more precise visual feature extraction. Additionally, auxiliary knowledge from ChatGPT enriches contextual representation. Experimental results on the SvgNER dataset show consistent improvements over all baselines, particularly in complex scenarios.

All baseline models are retrained on the SvgNER dataset under the same training/validation split as MNER-SVG for fair comparison.

 doi:10.20944/preprints202602.1129.v1

4.6. Results and Analysis

Table 4 presents the precision (P), recall (R), and F1-score (F1) of various methods on the SvgNER dataset.

Table 4. Performance comparison of NER and MNER models on the SvgNER dataset.

Methods	PER/ F1	LOC/ F1	ORG/ F1	MISC/ F1	Overall Pre.	Overall Rec.	Overall F1
BiLSTM-CRF (text only)	72.34	76.83	51.59	32.52	60.32	58.05	59.17
BERT-CRF (text only)	74.74	70.51	60.27	37.29	59.22	64.59	61.81
RoBERTa-span (text only)	77.20	73.58	66.33	50.66	67.48	67.43	67.45
UMT	85.24	81.58	73.03	49.45	71.67	75.23	73.41
MNER-QG	85.68	81.42	73.62	41.53	77.76	72.31	74.94
CAT-MNER	88.04	84.70	68.04	52.33	78.75	78.69	78.72
MCIRP	89.23	86.72	81.61	79.49	81.39	78.45	79.89
MNER-SVG (ours)	89.37	88.29	83.59	85.56	80.27	79.84	82.23

In comparative experiments, we selected four entity types for named entity recognition: person names (PER), locations (LOC), organizations (ORG), and miscellaneous (MISC). The recognition performance of different models on these four entity types is presented in Figure 6. The experimental results demonstrate the effectiveness of various models on specific entity types in the context of SVG multimodal named entity recognition. Notably, our proposed model achieves significant improvements on particular entity types in this task.

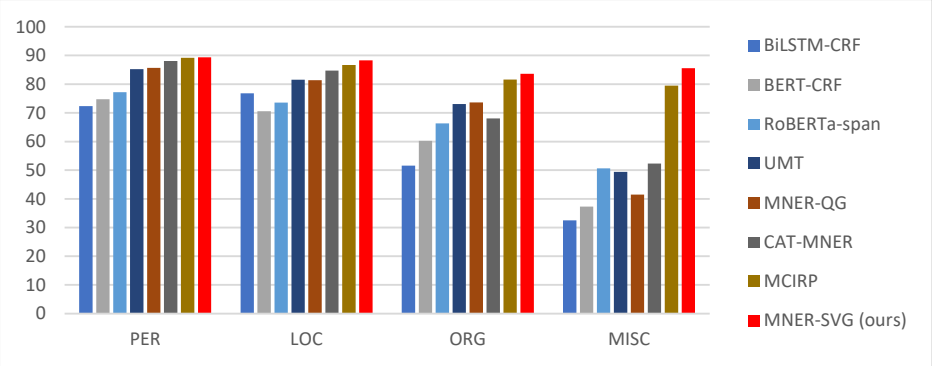

Figure 6. The recognition performance of various models on the PER, LOC, ORG, MISC entity types.

In comparative experiments, the overall precision, recall, and F1-score of the recognition results for the four entity types—person names (PER), locations (LOC), organizations (ORG), and miscellaneous (MISC)—are shown in Figure 7. The experimental results demonstrate that our proposed model achieves significant effectiveness in the task of SVG multimodal named entity recognition.

 doi:10.20944/preprints202602.1129.v1

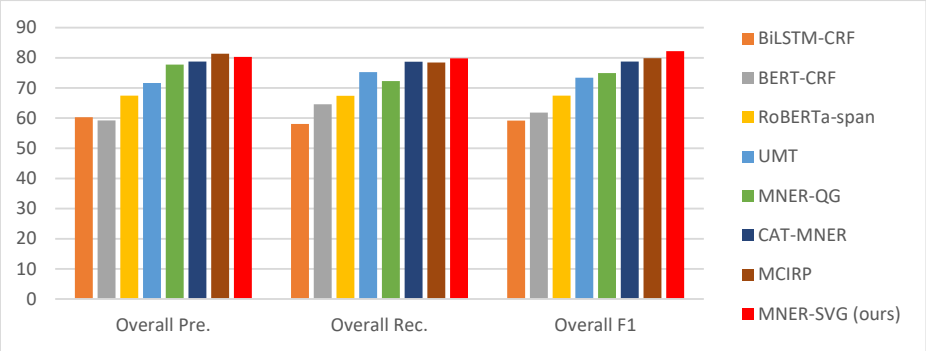

Figure 7. The accuracy, recall, and F1-score of the recognition results from various baseline models.

When only text-based named entity recognition (NER) methods are considered, BERT-CRF and RoBERTa-span both outperform BiLSTM-CRF, validating the effectiveness of pre-trained language models and conditional random fields (CRF) in NER tasks. Under the multimodal NER setting, however, text-only approaches significantly underperform their multimodal counterparts, indicating that multimodal modeling confers substantial advantages over unimodal text modeling in specific scenarios involving the integration of visual and audio information alongside text. This finding aligns with conclusions drawn in existing related studies.

Multimodal named entity recognition (NER) models that integrate both textual and visual information—such as CAT-MNER, MCIRP, and our proposed MNER-SVG—enhance NER performance by fusing image and text features. MNER-SVG achieves an F1-score of 82.23% on the SvgNER dataset, substantially outperforming existing methods, and is the first model to introduce Scalable Vector Graphics (SVG) into the NER task. Further experiments show that incorporating auxiliary knowledge generated by ChatGPT can effectively improve model performance. By fully leveraging the structured representations inherent to SVG and exploiting external semantic knowledge from ChatGPT, MNER-SVG offers a novel perspective for multimodal NER. Experimental results confirm that SVG structural information complements traditional visual features, demonstrating the feasibility and potential value of integrating vector graphic information into multimodal entity recognition.

4.7. Ablation Study

We conducted ablation experiments to assess the impact of ChatGPT-generated auxiliary knowledge and the Graph Neural Network (GNN) module.

4.7.1. Impact of ChatGPT-Generated Knowledge

Table 5 compares performance across varying sample sizes and prompt strategies. "VanillaGPT" uses no prompts, while "PromptGPT" selects top-N similar samples for context.

Prompted ChatGPT significantly improves performance in low-resource settings, with F1 scores increasing steadily as more relevant samples are included. These results highlight the auxiliary module's effectiveness in enhancing MNER-SVG, especially under few-shot conditions.

Table 5. Comparison of ChatGPT and PGIM. VanillaGPT and PromptGPT represent direct predictions using ChatGPT.

	SvgNER Pre.	SvgNER Rec.	SvgNER F1
fs-50	48.72	50.38	49.51
fs-100	58.63	67.58	62.82
fs-200	67.48	71.34	69.44
full-shot	82.34	82.09	82.23
VanillaGPT	51.83	73.17	59.76

 doi:10.20944/preprints202602.1129.v1

PromptGPTN=1	53.62	71.26	62.20
PromptGPTN=5	68.37	72.45	71.81
PromptGPTN=10	71.86	75.50	74.38

4.7.2. Effect of Removing ChatGPT Auxiliary Knowledge

We evaluate the impact of removing ChatGPT-generated auxiliary knowledge by using only plain SVG features without ChatGPT guidance. Results are reported in Table 6.

Table 6. Ablation on ChatGPT auxiliary knowledge on SvgNER dataset.

	SvgNER Pre.	**SvgNER Rec.**	**SvgNER F1**
w/o ChatGPT (plain SVG)	78.90	79.55	79.22
MNER-SVG (full)	82.34	82.09	82.23

The results show that removing ChatGPT-generated knowledge leads to a significant drop in performance, confirming the effectiveness of our auxiliary knowledge module.

4.7.3. Effectiveness of the GNN Module

We evaluate the impact of different GNN architectures through ablation experiments, with results in Table 7. "w/o" denotes removal of components, "X^{GCN}" and "I^{GCN}" refer to GCN processing of the full-text and visual object graphs, respectively, while "S^{GAT}" indicates GAT-based multimodal fusion.

Table 7. Ablation results of MNER-SVG on SvgNER dataset.

	SvgNER Pre.	**SvgNER Rec.**	**SvgNER F1**
w/o X^{GCN}	80.42	80.91	80.16
w/o I^{GCN}	81.59	80.27	81.40
w/o $X\&I^{GCN}$	80.19	83.14	80.22
w/o S^{GAT}	79.58	81.89	80.25
MNER-SVG	82.34	82.09	82.23

Key findings:

- **GCN for Text and Visual Representation:** Removing GCN from either graph severely impacts performance, demonstrating its role in enhancing representation by aggregating topological and node features.
- **GAT for Multimodal Fusion:** GAT outperforms GCN in fusion, adaptively weighting features and improving cross-modal relationships for better integration.

These results highlight the essential role of each module in the MNER-SVG framework.

5. Conclusions

We propose MNER-SVG, a novel multimodal named entity recognition framework that integrates scalable vector graphics (SVG) with auxiliary knowledge generated by ChatGPT. By combining multimodal example selection, text-image alignment, and Graph Attention Networks (GAT), the framework enhances cross-modal interaction and improves recognition performance. Experiments on the SvgNER dataset demonstrate that the resolution independence and structural clarity of SVG provide valuable visual context, resulting in significant accuracy gains.

Future work will focus on integrating richer auxiliary knowledge to enhance reasoning capabilities and expanding datasets for cross-domain evaluation. As SVG gains traction in digital media, its combination with other modalities holds promising potential to advance multimodal NER and cross-modal information processing.

 doi:10.20944/preprints202602.1129.v1

Author Contributions: Conceptualization, Y.M. and W.L.; methodology Y.M. and W.L.; software, H.Q.; validation, Y.M., H.Q. and W.L.; data curation, Y.M.; writing—original draft preparation, Y.M.; writing—review and editing, Y.M. and W.L.; visualization, H.Q.; supervision, W.L.; project administration, W.L.; funding acquisition, W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key Research Project of Anhui Provincial Department of Education, China (2022AH053089). Outstanding Scientific Research and Innovation Team of Anhui Police College (2023GADT06).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The source code presented in the study are openly available in GitHub https://github.com/TeacherWLee/MNER_SVG/ (accessed on 28 January 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Seow, W.L.; Chaturvedi, I.; Hogarth, A.; Mao, R.; Cambria, E. A Review of Named Entity Recognition: From Learning Methods to Modelling Paradigms and Tasks. *Artif Intell Rev* **2025**, *58*, 315, doi:10.1007/s10462-025-11321-8.
2. Li, J.; Sun, A.; Han, J.; Li, C. A Survey on Deep Learning for Named Entity Recognition. *IEEE transactions on knowledge and data engineering* **2020**, *34*, 50–70.
3. Zhu, C.; Chen, M.; Zhang, S.; Sun, C.; Liang, H.; Liu, Y.; Chen, J. SKEAFN: Sentiment Knowledge Enhanced Attention Fusion Network for Multimodal Sentiment Analysis. *Information Fusion* **2023**, *100*, 101958.
4. Geng, H.; Qing, H.; Hu, J.; Huang, W.; Kang, H. A Named Entity Recognition Method for Chinese Vehicle Fault Repair Cases Based on a Combined Model. *Electronics* **2025**, *14*, 1361, doi:10.3390/electronics14071361.
5. Shi, X.; Yu, Z. Adding Visual Information to Improve Multimodal Machine Translation for Low-Resource Language. *Mathematical Problems in Engineering* **2022**, *2022*, 1–9, doi:10.1155/2022/5483535.
6. Chen, F.; Chen, X.; Xu, S.; Xu, B. Improving Cross-Modal Understanding in Visual Dialog Via Contrastive Learning. In Proceedings of the ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 2022; pp. 7937–7941.
7. Xu, B.; Huang, S.; Sha, C.; Wang, H. MAF: A General Matching and Alignment Framework for Multimodal Named Entity Recognition. In Proceedings of the Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining; ACM: Virtual Event AZ USA, February 11 2022; pp. 1215–1223.
8. Chai, Y.; Xie, H.; Qin, J.S. Text Data Augmentation for Large Language Models: A Comprehensive Survey of Methods, Challenges, and Opportunities. *Artif Intell Rev* **2025**, *59*, 35, doi:10.1007/s10462-025-11405-5.
9. Wang, Z.; Chen, H.; Xu, G.; Ren, M. A Novel Large-Language-Model-Driven Framework for Named Entity Recognition. *Information Processing & Management* **2025**, *62*, 104054, doi:10.1016/j.ipm.2024.104054.
10. Zhang, Q.; Fu, J.; Liu, X.; Huang, X. Adaptive Co-Attention Network for Named Entity Recognition in Tweets. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence; 2018; Vol. 32.
11. Yu, J.; Jiang, J.; Yang, L.; Xia, R. Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer.; Association for Computational Linguistics, 2020.
12. Zhang, D.; Wei, S.; Li, S.; Wu, H.; Zhu, Q.; Zhou, G. Multi-Modal Graph Fusion for Named Entity Recognition with Targeted Visual Guidance. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence; 2021; Vol. 35, pp. 14347–14355.
13. Xu, M.; Peng, K.; Liu, J.; Zhang, Q.; Song, L.; Li, Y. Multimodal Named Entity Recognition Based on Topic Prompt and Multi-Curriculum Denoising. *Information Fusion* **2025**, *124*, 103405, doi:10.1016/j.inffus.2025.103405.
14. Li, E.; Li, T.; Luo, H.; Chu, J.; Duan, L.; Lv, F. Adaptive Multi-Scale Language Reinforcement for Multimodal Named Entity Recognition. *IEEE Transactions on Multimedia* **2025**.

 doi:10.20944/preprints202602.1129.v1

15. Zhang, Q.; Song, Z.; Wang, D.; Cai, Y.; Bi, M.; Zuo, M. Named Entity Recognition and Coreference Resolution Using Prompt-Based Generative Multimodal. *Complex Intell. Syst.* **2025**, *12*, 10, doi:10.1007/s40747-025-02122-1.
16. Mu, Y.; Guo, Z.; Li, X.; Shao, L.; Liu, S.; Li, F.; Mei, G. MCIRP: A Multi-Granularity Cross-Modal Interaction Model Based on Relational Propagation for Multimodal Named Entity Recognition with Multiple Images. *Information Processing & Management* **2026**, *63*, 104384, doi:10.1016/j.ipm.2025.104384.
17. Wei, X.; Cui, X.; Cheng, N.; Wang, X.; Zhang, X.; Huang, S.; Xie, P.; Xu, J.; Chen, Y.; Zhang, M.; et al. ChatIE: Zero-Shot Information Extraction via Chatting with ChatGPT 2024.
18. Chen, Y.; Zhang, B.; Li, S.; Jin, Z.; Cai, Z.; Wang, Y.; Qiu, D.; Liu, S.; Zhao, J. Prompt Robust Large Language Model for Chinese Medical Named Entity Recognition. *Information Processing & Management* **2025**, *62*, 104189, doi:10.1016/j.ipm.2025.104189.
19. De, S.; Sanyal, D.K.; Mukherjee, I. Fine-Tuned Encoder Models with Data Augmentation Beat ChatGPT in Agricultural Named Entity Recognition and Relation Extraction. *Expert Systems with Applications* **2025**, *277*, 127126, doi:10.1016/j.eswa.2025.127126.
20. Han, R.; Peng, T.; Yang, C.; Wang, B.; Liu, L.; Wan, X. Is Information Extraction Solved by Chatgpt? An Analysis of Performance, Evaluation Criteria, Robustness and Errors. *arXiv preprint arXiv:2305.14450* **2023**, 48.
21. Li, B.; Fang, G.; Yang, Y.; Wang, Q.; Ye, W.; Zhao, W.; Zhang, S. Evaluating ChatGPT's Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness 2023.
22. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers); 2019; pp. 4171–4186.
23. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A. Language Models Are Few-Shot Learners. *Advances in neural information processing systems* **2020**, *33*, 1877–1901.
24. Yang, Z.; Gan, Z.; Wang, J.; Hu, X.; Lu, Y.; Liu, Z.; Wang, L. An Empirical Study of Gpt-3 for Few-Shot Knowledge-Based Vqa. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence; 2022; Vol. 36, pp. 3081–3089.
25. Lopes, R.G.; Ha, D.; Eck, D.; Shlens, J. A Learned Representation for Scalable Vector Graphics.; 2019; pp. 7930–7939.
26. G. C. Lab The Quick, Draw! Dataset Available online: https://github.com/googlecreativelab/quickdraw-dataset (accessed on 25 January 2026).
27. Moon, S.; Neves, L.; Carvalho, V. Multimodal Named Entity Recognition for Short Social Media Posts. In Proceedings of the Proceedings of NAACL-HLT; 2018; pp. 852–860.
28. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF Models for Sequence Tagging 2015.
29. Yamada, I.; Asai, A.; Shindo, H.; Takeda, H.; Matsumoto, Y. LUKE: Deep Contextualized Entity Representations with Entity-Aware Self-Attention 2020.
30. Jia, M.; Shen, L.; Shen, X.; Liao, L.; Chen, M.; He, X.; Chen, Z.; Li, J. Mner-Qg: An End-to-End Mrc Framework for Multimodal Named Entity Recognition with Query Grounding. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence; 2023; Vol. 37, pp. 8032–8040.
31. Wang, X.; Ye, J.; Li, Z.; Tian, J.; Jiang, Y.; Yan, M.; Zhang, J.; Xiao, Y. CAT-MNER: Multimodal Named Entity Recognition with Knowledge-Refined Cross-Modal Attention. In Proceedings of the 2022 IEEE international conference on multimedia and expo (ICME); IEEE, 2022; pp. 1–6.