Cross-Domain Named Entity Recognition with Image-aware Contexts: Leveraging Image Captions and Chain-of-Thought

Anonymous ACL submission

Abstract

Cross-Domain Named entity recognition is a crucial task in natural language processing that helps extract meaningful entities from text when transferring across different domains. However current cross-domain NER methods are often limited in leverage heterogeneous information from other modalities, which limits the ability of cross-domain knowledge discovery and data mining, thereby constraining the application potential of large-scale information systems. To address these challenges, we propose a cross-domain NER method that utilizes image-aware contexts, consisting of Domain-specific Dynamic Image Captioning(DDC) and Cross-domain Reasoning Chain(CRC). DDC generates contextualized image captions by aligning the semantics of text and captions conditioned on textual domain cues. Then CRC identifies potential entities and classifies them using captions generated by DDC and chain-of-thought. Experimental results demonstrate that our method achieves a remarkable 6.23% average F1 improvement across all tested domains. Particularly notable are the performance gains in the political and scientific domains, where our approach surpasses the best baseline model with F1-score increases of 8.22% and 9.58%.

1 Introduction

Named Entity Recognition (NER) is a core task in information extraction and knowledge discovery (Li et al., 2023b)(Esmaail et al., 2024)(Bhowmick et al., 2023)(Li et al., 2023a)(Wang et al., 2024), which is widely applied in various scenarios, including question-answering systems (Mollá et al., 2006)(He and Golub, 2016), automatic summarization (Chen et al., 2004)(Etzioni et al., 2008)(Aone et al., 1999)(Aramaki et al., 2009), and information retrieval (Sun et al., 2020)(Zeng et al., 2023)(Simonyan and Zisserman, 2015)(Guo et al., 2009)(Petkova and Croft, 2007). In recent years, increasing attention has been focused on cross-domain NER, aiming to address the challenges posed by textual data from diverse domains which, as data sources and channels expand, are particularly evident in the scarcity of high-quality annotated data (Li et al., 2023b)(Bhowmick et al., 2023). For example, in domain-specific texts like scientific literature or political reports, entity annotations for specialized terms are scarce. Annotating unlabeled data often requires significant time and human resources. Therefore, efficiently acquiring entities in these low-resource settings has become a focal point of research (Bhowmick et al., 2023)(Arora and Park, 2023)(Zhao et al., 2022). Some studies have alleviated domain differences through label alignment and domain adaptation approaches (Golde et al., 2024)(Li et al., 2020). For instance, LAR proposed a label alignment and reallocation strategy to enhance cross-domain capabilities by passing label information between source and target domains (Zhang et al., 2023). In social media streams, (Bhowmick et al., 2023) used a global context embedding aggregation strategy to enhance the coherence and accuracy of entity recognition, demonstrating high adaptability in data-scarce environments. (Li et al., 2020) explored meta-learning approaches to improve NER adaptability and performance in few-shot learning scenarios. By separating task-irrelevant and task-specific components, the model can quickly adapt to different few-shot tasks and reduce the risk of overfitting. While these cross-domain NER methods have attempted to address these challenges, they tend to focus primarily on text and lack the ability to effectively incorporate other modalities, such as images, which could provide valuable contextual information and enhance entity recognition. This limitation has hindered the progress of cross-domain NER, especially in real-world applications where multimodal data is abundant but underutilized. Recognizing this untapped potential, multimodal NER has emerged as a promising direction that synergistically combines text with visual/audio modal-

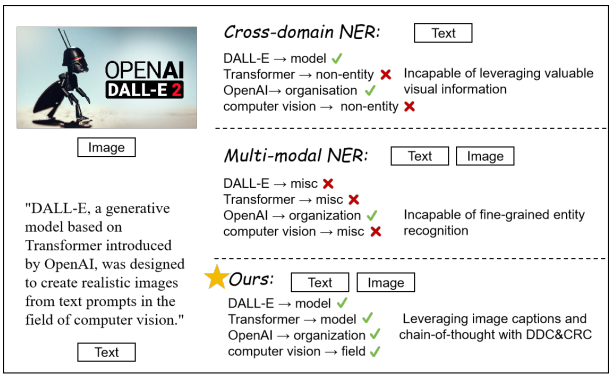


Figure 1: This figure illustrates the strengths and limitations of cross-domain NER and multimodal NER approaches. Our method enhances cross-domain NER through image captions and chain-of-thought.

ities to enhance entity recognition. Approaches like VisualBERT (Li et al., 2019b) improve entity recognition by using image captions as auxiliary information. However, these multimodal methods were not originally designed for cross-domain tasks (Li et al., 2019b)(Wang et al., 2022a)(Wang et al., 2022b). When directly applied to cross-domain NER, they often face significant limitations, such as the use of image captions that do not adapt to domain-specific contexts. Consequently, they fail to fully capture the nuances of domain-specific entities and struggle to generalize to different domains effectively. Figure 1 illustrates the challenges of cross-domain and multimodal NER.

To overcome these challenges, we first generate domain-specific image captions by aligning the semantics of the raw textual context, which are conditioned on textual domain cues. These captions encapsulate both entity information and domain knowledge, enhancing the understanding of potential entities in the original text. This enables a more seamless integration of visual and textual information. We then employ a reasoning chain to progressively process these contextually enriched captions and extract complex entity relationships, facilitating cross-domain and multimodal reasoning. The main contributions of our work are as follows:

- **We propose Domain-specific Dynamic Image Captioning (DDC):** Our approach generates domain-relevant image captions based on the specific contextual information of each domain. This method provides rich additional semantic information through the alignment of text and image semantics conditioned on textual domain cues. By deeply integrating image content with text context, DDC extracts more contextually relevant features from the visual modality, thereby enhancing entity recognition in complex settings.

- **We propose Cross-domain Reasoning Chain (CRC):** In collaboration with domain-specific image captions, CRC enhances the reasoning process by leveraging the contextualized image captions. It ensures a smooth and comprehensive reasoning chain for cross-domain tasks by progressively guiding the exploration of relationships between entities. Through multi-step reasoning, CRC facilitates the deduction of entity relationships, leading to more accurate inference and classification. This significantly improves the model's ability to understand complex cross-domain texts and their interrelated entity relationships.

- Experimental results demonstrate that our method not only significantly outperforms all baseline models in cross-domain NER tasks, but also achieves substantial improvements. Specifically, in the political and scientific domains, our model achieves F1 score increases of 8.22% and 9.58%, respectively, compared to the best baseline. Additionally, our method sets a new state-of-the-art (SOTA) performance in multimodal NER tasks, surpassing the current leading models. Ablation experiments further validate the critical contributions of the DDC and CRC modules in enhancing performance. Furthermore, in few-shot learning scenarios, our method demonstrates exceptional generalization ability in low-resource environments.

2 Related Works

2.1 Named Entity Recognition

NER primarily aims to automatically identify and classify entities in text, such as person, organization, location, etc. (Arora and Park, 2023)(Wang et al., 2023a). In recent years, with the advancement of deep learning, pre-trained language models like BERT have significantly enhanced the performance of NER (Sun et al., 2021). The NER Globalizer system proposed by (Bhowmick et al., 2023) combines local context embeddings and global context information for named entity recognition. In the local part, an attention-based model is used for

entity detection and type classification. The ResuFormer model proposed by (Yao et al., 2023) utilizes a combination of BERT and BiLSTM+CRF structures for named entity recognition, improving model robustness and efficiency through a self-training framework. These deep learning-based methods improve the understanding of complex syntactic structures by leveraging contextual information. .

In addition, recent research has focused on few-shot and zero-shot learning to address the data scarcity issue in low-resource scenarios (Zhu et al., 2024)(Xie et al., 2023). For example, MetaNER uses meta-learning to achieve rapid generalization in low-resource environments (Li et al., 2020). Although these methods improve the model's performance in low-resource scenarios, they are still mainly confined to a single modality of textual data and fail to fully leverage non-textual information (Wang et al., 2022a).

2.2 Cross-domain Named Entity Recognition

Cross-domain NER aims to address the performance degradation encountered when a model trained on one domain is applied to another. Some approaches focus on the data itself, improving cross-domain performance through data augmentation. For instance, (Golde et al., 2024) expanded the entity types and guided the model to learn and understand natural language descriptions of labels. (Yang et al., 2022) proposed semi-factual generation by randomly replacing non-entity words and counterfactual generation by randomly replacing entity words. By combining these two methods to generate augmented instances, the model's generalization ability can be enhanced. In contrast, (Chen et al., 2021) employed cross-domain data augmentation to teach the model patterns across different domains, transforming high-resource domain data into low-resource domain data.

Other methods are based on domain adaptation, aiming to reduce the distributional discrepancies between domains through techniques such as adversarial training and feature alignment. (Wang et al., 2023b) enhanced cross-domain generalization by extracting domain-relevant features and generating corresponding prompts. (Li et al., 2019a) utilized a pointer network to perform entity boundary tagging, integrating adversarial transfer learning to introduce domain-invariant representations into end-to-end sequence labeling models. (Li et al., 2023a) proposed FEWNER, a meta-learning-based cross-domain few-shot NER approach, which effectively adapts to new tasks and reduces overfitting by dividing the network into task-independent and task-specific components, facilitating efficient learning on cross-domain few-shot tasks. (Chen et al., 2023) incorporated logical rules and posterior regularization into deep learning, effectively improving the generalization ability of NER models. With the advent of large language models (LLMs), the underlying reasoning capabilities of LLMs have also been leveraged to help address the challenges posed by cross-domain NER and few-shot learning. (Ashok and Lipton, 2023) exploited the reasoning power of LLMs, guiding the model to predict entities in natural language by adding entity definitions beyond the standard few-shot examples. This allows large language models to generate potential entity lists and corresponding explanations. (Wang et al., 2023a) proposed a method that transforms the NER task into a text generation problem, enhancing performance in low-resource NER scenarios through labeling and self-verification strategies. (Xie et al., 2023) employed a decomposition strategy, converting the NER task into a series of sub-tasks and proposed a two-stage majority voting strategy to improve zero-shot NER performance. Similarly, (Arora and Park, 2023) utilized a decomposition approach, splitting the task into span detection and span classification steps. Additionally, some researchers have proposed prompt templates to further enhance cross-domain performance. For example, (Zhu et al., 2024) introduced an innovative prompt template and label injection instructions, enabling large models to output entities and thereby improving few-shot NER performance.

3 Method

As shown in Fig. 2, we propose a novel cross-domain NER method that introduces two key innovations: Domain-specific Dynamic Image Captioning (DDC) and Cross-domain Reasoning Chain (CRC). Firstly, our DDC generates domain-relevant image captions that align with the textual context. Unlike traditional methods that rely on predefined, static descriptions, DDC generates captions for each image based on the current domain context, effectively utilizing visual and textual information. This approach goes beyond simply concatenating images as supplementary input, instead converting visual content into semantically rich support, tightly aligned with the textual context. As a result,

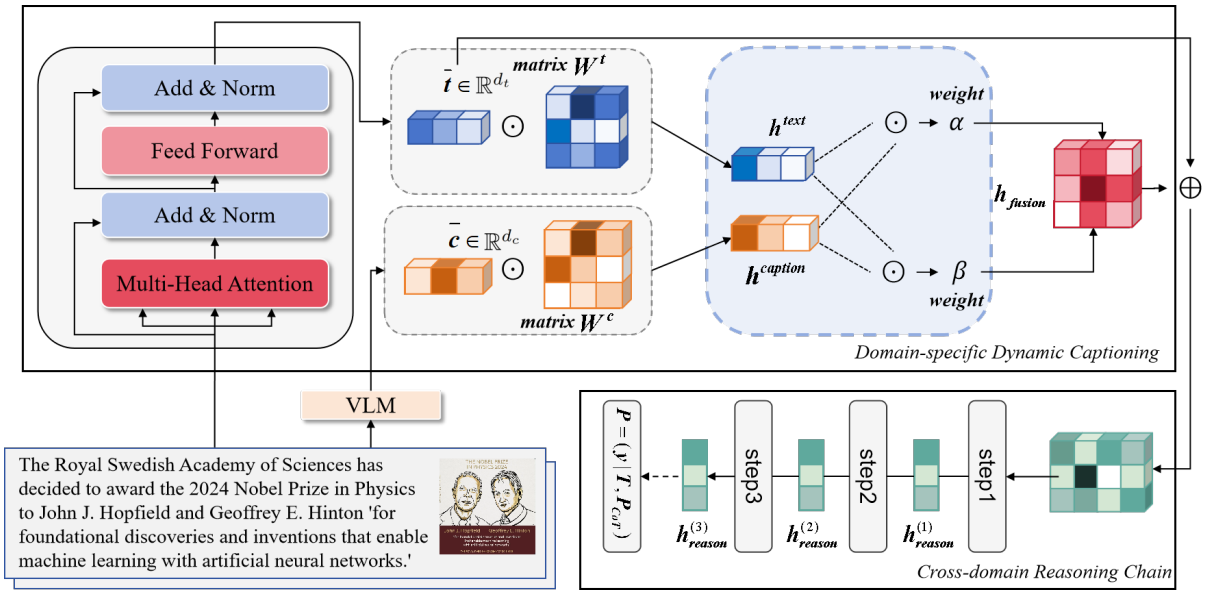


Figure 2: The overall architecture of the proposed method.

DDC significantly enhances entity recognition performance, particularly in scenarios where context plays a crucial role. Secondly, CRC enables multi-step reasoning that adapts to specific input texts and task requirements. CRC generates reasoning chains that guide entity identification and provide logical steps for entity classification, allowing for a deeper understanding of complex relationships within the text. By leveraging the complementary strengths of DDC and CRC, our approach incorporates both textual and visual information, enhancing entity recognition capabilities in complex, cross-domain, and low-resource environments.

3.1 Domain-specific Dynamic Image Captioning

3.1.1 Formulation

Traditional Named Entity Recognition tasks primarily rely on pure text input. Even in multimodal settings, existing methods often treat images merely as supplementary information, using image captions that do not adapt to task context, leading to a disconnect between image information and textual content. In contrast, our method introduces DDC, which generates image captions based on the specific context of each domain. This approach ensures that image information is fully integrated with text and directly contributes to the entity recognition process. Rather than simply concatenating image captions with the text, DDC treats the generated captions as a key element in the NER task, enhancing semantic understanding and demonstrating strong generalization across domains and in low-resource settings. Specifically, assume we have a text $T = \{t_1, t_2, ..., t_n\}$ and a corresponding image I. The domain-specific image caption is generated through a Visual Language Model (VLM), denoted as $C(T, I)$, and its generation process is defined as follows:

$$C(T, I) = \text{VLM}(T, I; \theta) \tag{1}$$

where θ represents the parameters of the VLM model. The dynamic caption $C(T, I)$ is adjusted based on the domain and context, ensuring that the image caption is not merely an additional piece of information but serves as an effective semantic extension of the text.

3.1.2 Domain-related Caption Generation

In the process of generating $C(T, I)$ within the DDC module, the Visual Language Model (VLM) first projects the text T and the image I into a high-dimensional embedding space to capture semantic features. The text embedding vector $\tilde{t}$ is obtained through the text encoder E_t:

$$\tilde{t} = E_t(T; \theta_t) \tag{2}$$

where θ_t represents the parameters of the text encoder, and $\tilde{t} \in \mathbb{R}^{d_t}$ is the text feature vector. Similarly, the image I is mapped into the feature space, resulting in the embedding $\tilde{i}$:

$$\tilde{i} = E_i(I; \theta_i) \tag{3}$$

where θ_i denotes the parameters of the image encoder, and $\tilde{i} \in \mathbb{R}^{d_i}$ is the image feature vector. The embedding vectors $\tilde{t}$ and $\tilde{i}$ capture the domain-specific semantic information of the text and image.

These features are then combined into a domain-relevant caption using a fusion function f. The generation process of the image caption $C(T, I)$ can be represented as:

$$c = f(\tilde{t}, \tilde{i}; \phi) \tag{4}$$

where ϕ represents the domain-specific parameters, and $c \in \mathbb{R}^{d_c}$ is the fused multimodal feature vector. The fused feature c is then input to the decoder to generate the image caption $C(T, I)$:

$$C(T, I) = D(c; \theta_c) \tag{5}$$

where θ_c denotes the parameters of the decoder.

For example, given an image of a literary award ceremony, the VLM generates a caption that details the award recipient and the award scene. This domain-specific caption provides rich additional semantic information, closely integrating with the original textual information, supporting multimodal understanding.

3.1.3 Deep Text-Image Fusion

We project the text embedding and image caption embedding into a shared feature space, aligning their dimensions using a linear transformation:

$$h_t = \sigma(W_t t + b_t) \tag{6}$$

$$h_c = \sigma(W_c c + b_c) \tag{7}$$

where W_t and W_c are linear transformation matrices, b_t and b_c are bias terms, and σ is an activation function. h_t and h_c are the aligned text and image caption features, respectively. This transformation maps h_t and h_c into a shared feature space to enable further semantic fusion.

Next, the text feature h_t generates a selective weighting coefficient α based on the image caption feature h_c, while the image caption feature h_c generates its selective weighting coefficient β based on the text feature h_t:

$$\alpha = \text{softmax}(h_t \cdot h_c^\top) \tag{8}$$

$$\beta = \text{softmax}(h_c \cdot h_t^\top) \tag{9}$$

where $(\cdot)$ denotes the dot product operation, and α and β represent the selective weighting coefficients for the image caption in the text feature space and for the text in the image caption feature space, respectively.

Finally, we generate the final cross-modal fusion representation h through a bidirectional weighted sum:

$$h = \alpha h_c + \beta h_t \tag{10}$$

This fused feature h captures the bidirectional interaction between the text and image caption at the semantic level, thereby enhancing semantic reasoning capabilities. This fusion approach enables the image caption to supplement implicit information in the text and to help infer potential entities through bidirectional interaction.

3.2 Cross-domain Reasoning Chain

3.2.1 Context-Based Generation

The CRC utilizes multimodal information h and textual context T to construct a multi-step reasoning chain $\{P_i\}_{i=1}^n$, where each step is guided to adaptively select different components of the fusion based on the context. The formula is as follows:

$$P_i(T, h) = f_i\left(h^{(i)}, T^{(i)}\right), \quad i = 1, 2, \ldots, n \tag{11}$$

where f_i denotes the generation function at step i, $h^{(i)}$ represents the fused feature selection at step i, and $T^{(i)}$ represents the semantic information of the text at the given step. This multi-step reasoning chain design enables the model to capture entities embedded within complex textual contexts by adaptively extracting relevant entities. It improves the precision of identifying complex and nested entities.

3.2.2 Collaborative Reasoning with Multimodal Information

The CRC works in conjunction with the DDC to enhance the reasoning capabilities through the multimodal fused representation h generated by DDC. The image captions complement the textual entity information and provide CRC with richer contextual support. In the reasoning chain of CRC, the image caption acts as part of the reasoning process, helping to reveal implicit relationships between images and text. For example, when describing a scientific experiment, the image caption generated

by DDC of experimental equipment can assist CRC in deducing possible research methods.

The collaborative reasoning process in CRC with multimodal information is expressed as follows:

$$P_i(T, h) = f_i(g(h^{(i)}, c^{(i)}), T^{(i)}) \quad (12)$$

where $c^{(i)}$ represents the selective feature of the image caption generated by DDC at step i. This formula demonstrates the multimodal collaborative reasoning process, where at each reasoning step P_i, a key feature in the image caption is selected from the fused representation h to help identify implicit relationships within the text. We then concatenates the original text $T = \{t_1, t_2, ..., t_n\}$ with the obtained P_{CoT} as $Z = [T; P_{CoT}]$. The transformer-based encoder integrates information from the CRC P_i into the token representations $Z = \{z_1, \cdots, z_n\}$ by leveraging its attention mechanism. This allows each token representation to encode contextually relevant signals from both the input sentence T and the auxiliary information. In our research, the sequence $Z = \{z_1, \cdots, z_n\}$ is passed through a CRF layer to model the dependency structure of the label sequence y. The conditional probability of y given T and P_{CoT} is expressed as:

$$P(y|T, P_{CoT}) = \frac{\prod_{i=1}^{n} \psi(y_{i-1}, y_i, z_i)}{\sum_{y' \in Y} \prod_{i=1}^{n} \psi(y'_{i-1}, y'_i, z_i)} \quad (13)$$

Here, $\psi(y_{i-1}, y_i, z_i)$ and $\psi(y'_{i-1}, y'_i, z_i)$ denote the potential functions capturing the relationships between labels and token representations. The model's parameters are optimized by minimizing the negative log-likelihood of the predicted label sequence with respect to the ground-truth labels y^*, formulated as:

$$\mathcal{L}_{\text{NLL}}(\theta) = -\log P_\theta(y^*|T, P_{CoT}) \quad (14)$$

4 Experiments and Results

4.1 Experiment Settings

4.1.1 Dataset

To evaluate our method, we selected four datasets: **CoNLL2003** (Tjong Kim Sang and De Meulder, 2003), **CrossNER** (Liu et al., 2021), **Twitter 2015** (Zhang et al., 2018), and **Twitter 2017** (Lu et al., 2018), with detailed dataset statistics shown in AppendixA. We first conducted pre-training on the CoNLL2003 dataset to enable the model to capture basic entity recognition capabilities. Subsequently, we performed experiments on the CrossNER, Twitter 2015, and Twitter 2017 datasets.

4.1.2 Implementation Details

We conducted our experiments on an NVIDIA 3090 GPU using the Pytorch framework for training and evaluation. The backbone of our model is bert-large-cased. We used the Adam optimizer with a linear warmup learning rate schedule, where 10% of the training steps were allocated for warmup. The learning rate was set to 2e-05 during the pre-training phase and 1e-05 during the fine-tuning phase. To prevent overfitting, we applied a weight decay of 0.01 for regularization, and the maximum gradient norm was set to 1.0 to avoid gradient explosion. The model was trained for 200 epochs, with a batch size of 2 due to hardware constraints. Model performance was evaluated using the F1 score, and we monitored the model by evaluating on the validation set every 10 epochs.

4.1.3 Baselines

To verify the effectiveness of our method, we compared it with several competitive models. First, we selected several multimodal NER models for comparison, including: a. **VEC-MNER**(Wei et al., 2024): enhances text representations with visual features, adopting a fusion strategy between visual scene graphs and text features. b. **VisualPT-MoE**(Zhu et al., 2024): leverages a mixture of experts (MoE) structure to integrate multiple image representations. c. **DPE-MNER**(Zheng et al., 2024): fuses visual and textual information at different granularities through incremental multimodal representation. d. **UniNER-7B**(Zhou et al., 2024): distills a large language model to produce a compact cross-domain NER model. e. **LST-NER**(Zheng et al., 2022): uses a graph matching algorithm to transfer label information between source and target domains. f. **PromptNER**(Shen et al., 2023): unifies entity localization and typing through a dual-slot prompt template, treating them as a single prompt-learning task.

4.2 Results and Discussions

4.2.1 Main Results

The results are presented in Table 1. In our experiments, we assessed the entity recognition ability of the model in different domains and compared it with several baseline models. The metric used

Table 1: F1 scores of different models on CrossNER dataset across five domains and on Twitter 2015 and Twitter 2017 datasets.

Model	Politics	Science	Music	Literature	AI	Avg.	Twitter 2015	Twitter 2017
UMT	-	-	-	-	-	-	73.41	85.31
VisualPT-MoE	-	-	-	-	-	-	75.63	87.42
VEC-MNER	-	-	-	-	-	-	74.89	84.51
DPE-MNER	-	-	-	-	-	-	77.56	87.90
PromptNER	73.61	71.23	64.61	60.09	57.79	66.47	-	-
UniNER-7B	66.90	70.80	70.60	64.90	62.90	67.40	-	-
LST-NER	68.51	66.48	72.04	66.73	60.69	67.07	-	-
Ours	**8.22** ↑ **76.73**	**9.58** ↑ **76.06**	**3.2** ↑ **75.24**	**5.72** ↑ **72.45**	**4.65** ↑ **65.34**	**6.23** ↑ **73.63**	**2.4** ↑ **79.96**	**3.64** ↑ **91.54**

Table 2: Ablation study results on the impact of DDC and CRC modules.

Model	Politics	Science	Music	Literature	AI	Avg.	Twitter 2015	Twitter 2017
w/o DDC+CRC	73.61	71.23	64.61	60.09	57.79	66.47	76.52	88.19
w/o DDC	76.02	75.41	72.95	64.64	63.35	71.24	77.43	88.94
w/o CRC	74.47	73.09	67.34	64.08	60.52	68.73	76.47	88.57

in the table is the F1 score, which measures the model's performance in cross-domain NER tasks supported by image semantics.

The results show that our method achieves an overall F1 score of 73.63 across all domains, outperforming our baseline models. This improvement highlights the effectiveness of our DDC and CRC in enhancing text-image fusion and contextual reasoning. Notably, in the politics and science domains, our method outperforms baseline models with improvements of 8.22% and 9.58%, respectively. The image captions generated by DDC enrich the textual context, allowing the model to better distinguish between complex entities in multimodal settings. And the CRC module significantly improves the model's ability to handle implicit relationships in complex domain-specific contexts. However, our method faces some challenges in the AI domain, where image captions provide limited contextual support for abstract entities. In this domain, textual reasoning is more prominent for entity recognition, which might explain the slightly lower performance (F1 score of 65.34). When compared with multimodal baselines, our method achieves state-of-the-art performance. On the Twitter 2017 dataset, our model attains an F1 score of 91.54, surpassing the best baseline model, DPE-MNER, by 3.64%. Similarly, on the Twitter 2015 dataset, our model achieves an F1 score of 79.96, outperforming other multimodal models and setting a new SOTA in multimodal NER tasks.

4.2.2 Ablation Study

To validate the effectiveness of the DDC and CRC modules, we conducted an ablation study. In this study, we progressively removed the DDC and CRC modules and evaluated their impact on the model's performance. The results are shown in Table 2:

Impact of Removing Both DDC and CRC. When both DDC and CRC modules are removed, the model's average F1 score drops to 66.47, indicating that the synergy of these two modules is crucial to the model's overall performance. In particular, in the science (71.23) and music (64.61) domains, the model's performance declines significantly without the image caption and reasoning chain, suggesting that these domains have a strong dependency on multimodal information.

Impact of Removing DDC. When the DDC module is removed, the model's average F1 score decreases to 71.24. Specifically, in the music (72.95) and science (75.41) domains, the absence of image captions leads to a decline in performance. This demonstrates that the dynamic image captions generated by DDC are essential for enriching textual context and enhancing entity recognition capabilities.

Table 3: Performance comparison across domains with different K values.

Samples	$K = 20$					$K = 50$				
Domain	**Pol.**	**Sci.**	**Mus.**	**Lit.**	**AI**	**Pol.**	**Sci.**	**Mus.**	**Lit.**	**AI**
BiLSTM-CRF(Lample et al., 2016)	41.75	42.54	37.96	35.78	37.59	53.46	43.65	41.54	44.73	56.13
Coach(Liu et al., 2020)	46.15	48.71	43.37	41.64	41.55	60.97	51.56	48.73	51.15	56.09
Multi-Cell LSTM(Jia and Zhang, 2020)	59.58	60.55	67.12	63.92	55.39	68.21	70.47	66.85	58.67	58.48
BERT-tagger(Devlin et al., 2019)	61.01	60.34	64.73	61.79	53.78	66.13	68.41	63.44	58.93	58.16
NNShot(Yang and Katiyar, 2020)	60.93	60.67	64.21	61.64	54.27	66.33	67.94	63.19	59.17	57.34
StructShot(Yang and Katiyar, 2020)	63.31	62.95	67.27	63.48	55.16	67.16	70.21	65.33	59.73	58.74
templateNER(Cui et al., 2021)	63.39	62.64	62.00	61.84	56.34	58.39	65.23	64.57	64.49	56.58
LST-NER(Zheng et al., 2022)	64.06	64.03	68.83	64.94	**57.78**	68.51	72.04	66.73	60.69	61.25
Ours	**67.26**	**70.68**	**68.85**	**65.77**	57.67	**75.75**	**73.57**	**74.82**	**67.08**	**62.36**

Impact of Removing CRC. When only the CRC module is removed, the F1 scores in the politics (74.47) and literature (64.08) domains drop considerably, indicating that the CRC module plays a crucial role in handling complex textual relationships and multi-entity associations in these domains. However, in other domains, such as AI (60.52), the performance remains relatively stable, suggesting that the contribution of CRC is more significant for reasoning tasks involving complex textual information.

Table 4: Few-shot performance of our model on the CrossNER dataset across different domains.

Domain	**5-shot**	**10-shot**	**20-shot**	**50-shot**
Politics	49.00	59.81	67.26	75.75
Science	57.56	66.44	70.68	73.57
Music	50.28	62.46	68.85	74.82
Literature	46.55	56.74	65.77	67.08
AI	41.72	45.08	57.67	62.36

4.2.3 Few-shot Study

To evaluate our method's performance in low-resource scenarios, we conducted experiments with 20-shot and 50-shot settings across five domains: politics, science, music, literature, and AI. The experimental results are shown in Table 3. In the 20-shot setting, our method achieves higher F1 scores in most domains, especially in politics (67.26) and science (70.68), where it outperforms the second-best method by significant margins. However, it performs slightly lower in the AI (57.67) domains. The lower performance in AI is likely due to the abstract nature of its entities, which makes it harder for the model to generalize with limited data. In the 50-shot setting, our method dominates across most domains, with significant improvements over other methods, especially in politics(75.75) and music (74.82), demonstrating its robustness in low-resource settings. The results show that as the number of training samples increases, the F1 scores improve significantly, approaching stable levels in each domain. We also tested additional k-values (5-shot, 10-shot, 20-shot, 50-shot). The results are shown in Table 4. The performance improves with more data, especially in the 50-shot setting, where the model stabilizes. Even with 5-shot and 10-shot settings, our method maintains a reasonable recognition ability, demonstrating adaptability in data-scarce situations.

5 Conclusions

We propose a cross-domain NER method that synergizes Domain-specific Dynamic Image Captioning (DDC) with Cross-domain Reasoning Chain (CRC), achieving significant performance improvements across diverse domains. By employing DDC to generate context-aware visual semantics through text-image alignment and constructing CRC for progressive deduction entity relationships via multi-step contextualized reasoning, our method effectively addresses the challenges of both the scarcity of high-quality annotated data in cross-domain settings and the limitations of incorporating multimodal information, particularly demonstrating strong generalization capabilities in low-resource scenarios. These advancements establish new state-of-the-art performance while preserving interpretability through explicit reasoning pathways.

6 Limitations

Our method has limitations in certain scenarios. First, while DDC enhances context comprehension through text-image alignment, its performance may be limited in domains where visual information has little relevance, leading to a reduced impact on tasks where textual reasoning is dominant. Additionally, although the CRC facilitates entity relationship reasoning, complex relationships may still be missed due to the inherent challenges of progressive deduction in dynamic, evolving data streams. In future work, we aim to improve these areas by exploring enhanced image-text synergy in domain-specific contexts and refining the multi-step reasoning process to handle more complex entity interactions.

7 Risks

The datasets utilized in our research are all publicly available, and no personal data or sensitive information is collected or processed. The prompts used in our method are designed to extract entities and their relationships from these datasets, ensuring no private or confidential information is involved. Additionally, the method avoids the inclusion of any harmful, discriminatory, or unethical content, respecting the rights of individuals and groups. Our approach adheres to the terms of use and licensing agreements associated with publicly accessible large language models and datasets.

References

Chinatsu Aone, Mary Ellen Okurowski, James Gorlinsky, and Bjornar Larsen. 1999. *A trainable summarizer with knowledge acquired from robust NLP techniques*. Advances in Automatic Text Summarization.

Eiji Aramaki, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Mashuichi, and Kazuhiko Ohe. 2009. TEXT2TABLE: Medical text summarization system based on named entity recognition and modality identification. In *Proceedings of the BioNLP 2009 Workshop*, pages 185–192, Boulder, Colorado. Association for Computational Linguistics.

Jatin Arora and Youngja Park. 2023. Split-NER: Named entity recognition via two question-answering-based classifications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 416–426, Toronto, Canada. Association for Computational Linguistics.

Dhananjay Ashok and Zachary C. Lipton. 2023. Promptner: Prompting for named entity recognition. *Preprint*, arXiv:2305.15444.

Satadisha Saha Bhowmick, Eduard C. Dragut, and Weiyi Meng. 2023. Globally aware contextual embeddings for named entity recognition in social media streams. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 1544–1557.

H. Chen, W. Chung, J.J. Xu, G. Wang, Y. Qin, and M. Chau. 2004. Crime data mining: a general framework and some examples. *Computer*, 37(4):50–56.

Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. 2021. Data augmentation for cross-domain named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5346–5356, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhijun Chen, Hailong Sun, Haoqian He, and Pengpeng Chen. 2023. Learning from noisy crowd labels with logics. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 41–52.

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Naji Esmaail, Nazlia Omar, Masnizah Mohd, Fariza Fauzi, and Zainab Mansur. 2024. Named entity recognition in user-generated text: A systematic literature review. *IEEE Access*, 12:136330–136353.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open information extraction from the web. *Commun. ACM*, 51(12):68–74.

Jonas Golde, Felix Hamborg, and Alan Akbik. 2024. Large-scale label interpretation learning for few-shot named entity recognition. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2915–2930, St. Julian's, Malta. Association for Computational Linguistics.

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 267–274, New York, NY, USA. Association for Computing Machinery.

Xiaodong He and David Golub. 2016. Character-level question answering with attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1598–1607, Austin, Texas. Association for Computational Linguistics.

Chen Jia and Yue Zhang. 2020. Multi-cell compositional LSTM for NER domain adaptation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5906–5917, Online. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Jing Li, Billy Chiu, Shanshan Feng, and Hao Wang. 2023a. Few-shot named entity recognition via meta-learning (extended abstract). In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 3805–3806.

Jing Li, Shuo Shang, and Ling Shao. 2020. Metaner: Named entity recognition with meta-learning. In *Proceedings of The Web Conference 2020*, WWW '20, page 429–440, New York, NY, USA. Association for Computing Machinery.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2023b. A survey on deep learning for named entity recognition : Extended abstract. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 3817–3818.

Jing Li, Deheng Ye, and Shuo Shang. 2019a. Adversarial transfer for named entity boundary detection with pointer networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5053–5059. International Joint Conferences on Artificial Intelligence Organization.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019b. Visualbert: A simple and performant baseline for vision and language. *Preprint*, arXiv:1908.03557.

Zihan Liu, Genta Indra Winata, Peng Xu, and Pascale Fung. 2020. Coach: A coarse-to-fine approach for cross-domain slot filling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 19–25, Online. Association for Computational Linguistics.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13452–13460.

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, Melbourne, Australia. Association for Computational Linguistics.

Diego Mollá, Menno van Zaanen, and Daniel Smith. 2006. Named entity recognition for question answering. In *Proceedings of the Australasian Language Technology Workshop 2006*, pages 51–58, Sydney, Australia.

Desislava Petkova and W. Bruce Croft. 2007. Proximity-based document representation for named entity retrieval. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, page 731–740, New York, NY, USA. Association for Computing Machinery.

Yongliang Shen, Zeqi Tan, Shuhui Wu, Wenqi Zhang, Rongsheng Zhang, Yadong Xi, Weiming Lu, and Yueting Zhuang. 2023. PromptNER: Prompt locating and typing for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12492–12507, Toronto, Canada. Association for Computational Linguistics.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *Preprint*, arXiv:1409.1556.

Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rpbert: A text-image relation propagation-based bert model for multimodal ner. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13860–13868.

Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. 2020. Multi-modal knowledge graphs for recommender systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 1405–1414, New York, NY, USA. Association for Computing Machinery.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023a. Gpt-ner: Named entity recognition via large language models. *Preprint*, arXiv:2304.10428.

Xinyu Wang, Jiong Cai, Yong Jiang, Pengjun Xie, Kewei Tu, and Wei Lu. 2022a. Named entity and relation extraction with multi-modal retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5925–5936, Abu Dhabi, United

Arab Emirates. Association for Computational Linguistics.

Xuwu Wang, Jiabo Ye, Zhixu Li, Junfeng Tian, Yong Jiang, Ming Yan, Ji Zhang, and Yanghua Xiao. 2022b. 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.

Yuanyi Wang, Haifeng Sun, Jiabo Wang, Jingyu Wang, Wei Tang, Qi Qi, Shaoling Sun, and Jianxin Liao. 2024. Towards semantic consistency: Dirichlet energy driven robust multi-modal entity alignment. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 3559–3572.

Zihan Wang, Ziqi Zhao, Zhumin Chen, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. 2023b. Generalizing few-shot named entity recognizers to unseen domains with type-related features. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2228–2240, Singapore. Association for Computational Linguistics.

Pengfei Wei, Hongjun Ouyang, Qintai Hu, Bi Zeng, Guang Feng, and Qingpeng Wen. 2024. Vecmner: Hybrid transformer with visual-enhanced cross-modal multi-level interaction for multimodal ner. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, ICMR '24, page 469–477, New York, NY, USA. Association for Computing Machinery.

Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. Empirical study of zero-shot NER with ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7935–7956, Singapore. Association for Computational Linguistics.

Linyi Yang, Lifan Yuan, Leyang Cui, Wenyang Gao, and Yue Zhang. 2022. FactMix: Using a few labeled in-domain examples to generalize to cross-domain named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5360–5371, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yi Yang and Arzoo Katiyar. 2020. Simple and effective few-shot named entity recognition with structured nearest neighbor learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6365–6375, Online. Association for Computational Linguistics.

Kaichun Yao, Jingshuai Zhang, Chuan Qin, Xin Song, Peng Wang, Hengshu Zhu, and Hui Xiong. 2023. Resuformer: Semantic structure understanding for resumes via multi-modal pre-training. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 3154–3167.

Yawen Zeng, Qin Jin, Tengfei Bao, and Wenfeng Li. 2023. Multi-modal knowledge hypergraph for diverse image retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3):3376–3383.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Xin Zhang, Jingling Yuan, Lin Li, and Jianquan Liu. 2023. Reducing the bias of visual objects in multimodal named entity recognition. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, page 958–966, New York, NY, USA. Association for Computing Machinery.

Fei Zhao, Chunhui Li, Zhen Wu, Shangyu Xing, and Xinyu Dai. 2022. Learning from different text-image pairs: A relation-enhanced graph convolutional network for multimodal ner. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 3983–3992, New York, NY, USA. Association for Computing Machinery.

Junhao Zheng, Haibin Chen, and Qianli Ma. 2022. Cross-domain named entity recognition via graph matching. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2670–2680, Dublin, Ireland. Association for Computational Linguistics.

Zihao Zheng, Zihan Zhang, Zexin Wang, Ruiji Fu, Ming Liu, Zhongyuan Wang, and Bing Qin. 2024. Decompose, prioritize, and eliminate: Dynamically integrating diverse representations for multimodal named entity recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4498–4508, Torino, Italia. ELRA and ICCL.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024. Universalner: Targeted distillation from large language models for open named entity recognition. *Preprint*, arXiv:2308.03279.

Xingyu Zhu, Feifei Dai, Xiaoyan Gu, Bo Li, Meiou Zhang, and Weiping Wang. 2024. Gl-ner: Generation-aware large language models for few-shot named entity recognition. In *Artificial Neural Networks and Machine Learning – ICANN 2024*, pages 433–448, Cham. Springer Nature Switzerland.

A Dataset Details

Table 5 shows the details of the datasets used in our experiments.

B Case Study

To illustrate the effectiveness of our proposed approach, we examine specific cases as shown in

Table 5: The statistics of the dataset.

Dataset	Type Num	Sentence Num		
		Train	Dev	Test
CoNLL2003	4	14987	3466	3684
Politics	9	200	541	651
Science	17	200	450	543
Music	13	100	380	456
Literature	12	100	400	416
AI	14	100	350	431
Twitter2015	4	3999	999	3256
Twitter2017	4	3373	723	723

Fig.3. Baseline models rely exclusively on textual inputs and often fail to perform well in scenarios requiring multimodal or contextual understanding. Competing methods such as PromptNER and UniNER employ static prompts or generic templates, which restrict their ability to adapt to varying domain-specific contexts. Similarly, LST-NER, while effective in low-resource cross-domain tasks through label transfer mechanisms, lacks the capacity to fully leverage multimodal or generated contextual information. In contrast, our proposed framework addresses these limitations by introducing DDC, which adaptively generate visual captions aligned with textual context, and CRC that performs multi-step reasoning for fine-grained entity classification. By integrating dynamic visual and contextual information, our approach demonstrates superior adaptability and accuracy in complex multimodal and cross-domain NER tasks.

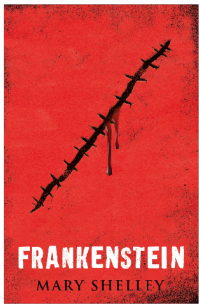
Text: Daniel Bolya presented his research on neural networks at the International Conference on Machine Learning (ICML) in New York in 2023.

Reasoning Chain:

Step 1. Identified entities: "Daniel Bolya," "neural networks," "International Conference on Machine Learning (ICML)," "New York,".
Step 2. Classification:
- "Daniel Bolya" is a Researcher (refers to a person conducting research).
- "neural networks" is an Algorithm (refers to a computational method).
- "International Conference on Machine Learning (ICML)" is a Conference (refers to a scientific event).
- "New York" is a Location (refers to a geographical place).

Step 3. Explanation: "Daniel Bolya" is a Researcher because the context suggests he is presenting research. "Neural networks" is an Algorithm based on its usage in machine learning. "International Conference on Machine Learning" is a Conference, as it refers to a known academic event. "New York" is a Location due to its geographical nature, and "2023" is a Metric because it refers to a specific time.

Prediction: Ours —— researcher | algorithm | conference | location ✓ PromptNER —— person | algorithm | award | country ×
UniNER —— person | algorithm | award | loacation × LST-NER —— person | misc | conference | location ×



Text: Mary Shelley wrote 'Frankenstein' during the summer of 1816, known as the Year Without a Summer, while staying near Lake Geneva, a pivotal moment in Gothic literature.

Reasoning Chain:

Step 1: Identified entities: "Mary Shelley,""Frankenstein,""Year Without a Summer,""Lake Geneva,""Percy Shelley,""Lord Byron,""Gothic literature."
Step 2: Classification: - "Mary Shelley" is a Writer. - "Frankenstein" is a Book. - "Year Without a Summer" is a Event.
- "Lake Geneva" is a Location. - "Gothic literature" is a Literary genre.
Step 3: Explanation: "Mary Shelley" is classified as Writers because she is historically recognized an author contributing to English literature."Frankenstein" is classified as a Book because it refers to a specific novel written by Mary Shelley. "Year Without a Summer" is classified as an Event, referring to a notable climatic anomaly affecting that period."Lake Geneva" is classified as a Location, as it is a geographical reference tied to the setting where the writers stayed."Gothic literature" is classified as a Literary genre, as it describes the style and thematic context of the novel.

Prediction: Ours —— writer | book | event | location | literarygenre ✓ PromptNER —— person | person | O | location | O ×
UniNER —— person | person | O | loacation | misc × LST-NER —— writer | person | misc | location | O ×

Figure 3: This is the figure of case study.