MPBoCo: Multimodal Prompt-based Boundary-enhanced Continual Framework for Joint Entity and Relation Extraction

Guanglu Sun[1], **Xinyu Liu**[1], **Lili Liang**[1], **Yu Yang**[1], **Fei Lang**[1], **Suxia Zhu**[1], **Ming Liu**[2]
[1]Harbin University of Science and Technology, Harbin, China
[2]Harbin Institute of Technology, Harbin, China
sunguanglu@hrbust.edu.cn, 2010400001@stu.hrbust.edu.cn
lianglili@hrbust.edu.cn, 2210403074@stu.hrbust.edu.cn
langfei@hrbust.edu.cn, zhusuxia@hrbust.edu.cn, mliu@ir.hit.edu.cn

Abstract

In real-world scenarios, multimodal information continuously evolves, with new entity and relation types emerging, necessitating timely updates to multimodal knowledge graphs for supporting downstream tasks. However, existing methods struggle to balance real-time adaptability and computational efficiency in continual learning scenarios. To this end, this paper proposes the Continual Multimodal Entity and Relation Joint Extraction (CMERJE) task and a Multimodal Prompt-based Boundary-enhanced Continual (MPBoCo) framework. Specifically, MPBoCo incrementally stores task-specific knowledge via learnable multimodal prompts, dynamically matches relevant prompts for each instance, and fuses them into a frozen backbone model for task-specific reasoning. Subsequently, the boundary-enhanced dual-branch module leverages the auxiliary branch to preserve local syntactic continuity and provide boundary guidance. Experimental results demonstrate that MPBoCo achieves superior performance in real-world scenarios, significantly outperforming baseline methods by 5.5% and 7.2% in 10-task and 5-task settings, respectively.

1 Introduction

Multimodal Knowledge Graphs (MMKGs) (Liang et al., 2024) provide foundational support for downstream tasks, such as multimodal reasoning (Lee et al., 2024) and question answering (Hu et al., 2025). In real-world scenarios, multimodal information is continuously generated and evolves, with new entity and relation types frequently emerging, which necessitates timely updates of MMKGs to maintain their relevance and completeness. In this setting, visual modalities provide complementary information that alleviates the insufficiency and ambiguity of textual semantics.

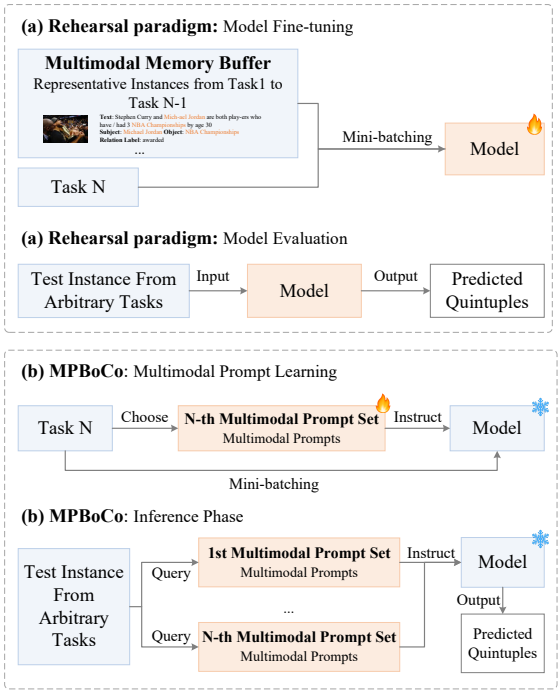

Figure 1: Comparison of rehearsal-based paradigm and MPBoCo during training and testing phases.

However, existing methods struggle to address the aforementioned challenges. Multimodal extraction methods (Yu et al., 2020; Yuan et al., 2023) are developed under a static data assumption. When faced with newly emerging types, they typically require retraining on both historical and newly collected data, which is impractical in continual scenarios due to privacy and computational overhead. In addition, existing continual extraction methods (Chen et al., 2024; Liu et al., 2025) often rely on the pipeline-based strategy, which exacerbates error propagation and limits efficiency. To this end, this paper introduces the Continual Multimodal Entity and Relation Joint Extraction (CMERJE) task for the first time, aiming to enable models to perform joint extraction of entities and relations from multimodal information in continual scenarios.

* Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 26514–26524

The widely used rehearsal-based continual learning paradigm (Wang et al., 2019) stores previous instances in the memory buffer, as shown in Figure 1(a). However, storage limitations and privacy concerns (Shokri and Shmatikov, 2015) motivate the exploration of alternative paradigms. Recently, the prompt-based paradigm (Wang et al., 2022) has been introduced, which injects a small set of learnable parameters (i.e., prompts) during training, aiming to guide the frozen model in performing task-specific reasoning. Despite their efficiency, prompt-based paradigms exhibit limitations in sequence labeling tasks, which require token-level fine-grained semantics. In contrast, prompts typically encode global-level task-specific knowledge, potentially weakening boundary semantics.

To address these challenges, this paper proposes MPBoCo, a Multimodal Prompt-based Boundary-enhanced Continual framework. MPBoCo stores task-specific knowledge using learnable multimodal key-prompt pairs, dynamically matches the most relevant pairs for each instance, and then uses them to guide the frozen backbone model in task-specific reasoning. Prompts for completed tasks are frozen to prevent forgetting and interference. To enhance fine-grained boundary modeling, a boundary-enhanced dual-branch module preserves local syntactic continuity via an auxiliary branch, providing reliable supervision for boundary predictions. The main contributions can be summarized:

- This paper constructs CMERJE , the first dedicated dataset for the proposed Continual Multimodal Entity and Relation Joint Extraction (CMERJE) task, providing a benchmark for future research.
- This paper proposes a multimodal prompt-based boundary-enhanced continual framework, employing instance-level prompt matching and hierarchical prompt fusion to guide a frozen backbone for task-specific reasoning. This is the first application of prompt learning to continual multimodal sequence labeling.
- The designed boundary-enhanced dual-branch fusion module preserves local syntactic and continuity features, providing structured guidance and boundary supervision signals.
- Experiments show that MPBoCo significantly outperforms baselines, effectively balancing knowledge retention and new type acquisition.

2 Related Work

This paper is the first to propose CMERJE, bridging the gap between multimodal joint extraction and continual multimodal extraction, and adopts a prompt-based paradigm to avoid the storage and privacy limitations of rehearsal-based methods.

2.1 Multimodal Joint Information Extraction

Early multimodal entity-relation extraction methods mainly followed a pipeline strategy, performing entity recognition (Yu et al., 2020) and relation prediction (Zheng et al., 2021) sequentially. To alleviate semantic ambiguity and insufficient contextual cues, recent studies adopted object detection (Wu et al., 2020; Zheng et al., 2020; Liu et al., 2024) to identify key visual regions, followed by visual encoders (e.g., VGGNet (Zhang et al., 2018), ResNet (Sun et al., 2021; Liu et al., 2022)) to extract global- or region-level representations, and finally leveraged graph-based modeling, including graph neural networks (Zhang et al., 2021) and scene graphs (Wang et al., 2023a; Cheng et al., 2023), to capture structured relationships and facilitate fine-grained cross-modal interactions. Building on the observation of bidirectional dependency between entities and relations, Yuan et al. (2023) proposed the Joint Multimodal Entity-Relation Extraction (JMERE) task and studies (Yuan et al., 2025; Huang et al., 2025; Liu et al., 2026a) explicitly modeled structured interactions across modalities and subtasks. However, existing JMERE methods generally assume a static label space and offline training, which makes them unsuitable for continual scenarios.

2.2 Continual Multimodal Information Extraction

In real-world scenarios such as social media, multimodal posts evolve continuously, requiring models to incrementally learn new knowledge without retraining from scratch. Existing methods can be broadly categorized into regularization-based (Kirkpatrick et al., 2017; Farajtabar et al., 2020), architecture-based (Mallya and Lazebnik, 2018; Qin et al., 2021), rehearsal-based (Han et al., 2020; Wang et al., 2023b; Zhang et al., 2024), and prompt-based paradigms (Wang et al., 2022). These methods are predominantly designed for unimodal scenarios. Yu et al. (2024) pointed out that multimodal continual learning is more challenging than its unimodal counterpart, as models must continually handle modality heterogeneity introduced by new tasks

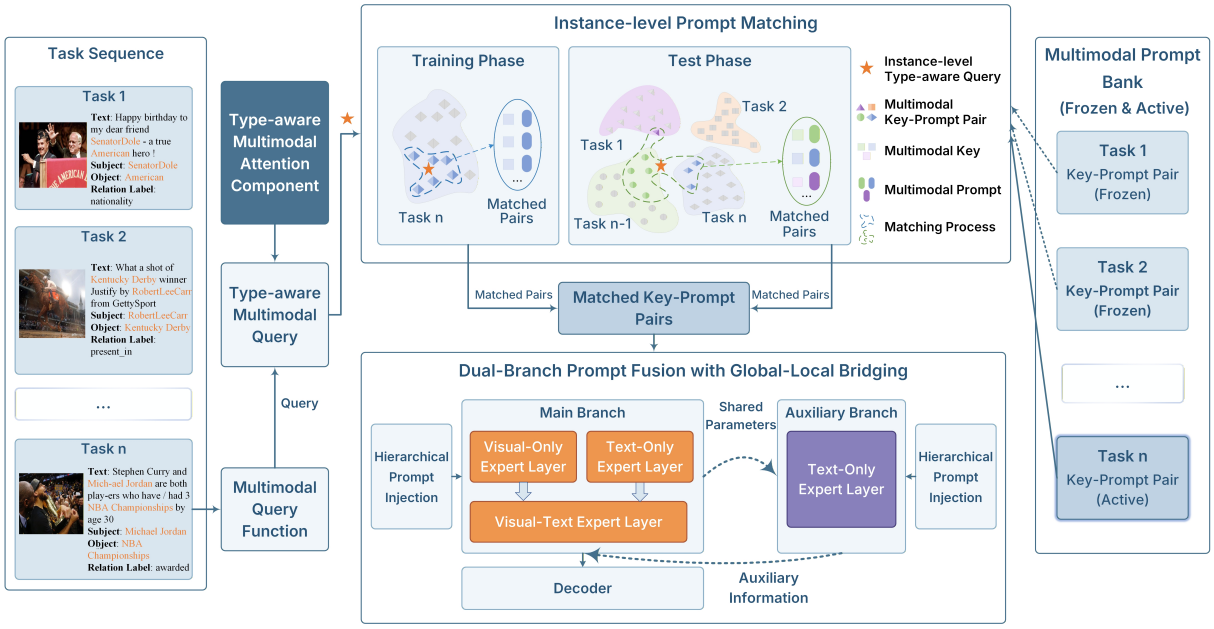

Figure 2: Overall architecture of the proposed MPBoCo framework.

while preserving cross-modal semantic alignment. To extract structured knowledge from evolving multimodal data for downstream applications, recent studies (Chen et al., 2024; Liu et al., 2025) have explored continual multimodal named entity recognition and relation extraction tasks. However, such pipeline strategies exacerbate error propagation and computational inefficiency in continual scenarios.

2.3 Prompt-based Continual Learning

Prompt learning (Li and Liang, 2021; Zhou et al., 2022; Khattak et al., 2023) adapts pre-trained models to downstream tasks by injecting a small number of task-specific prompts, such as manual templates (Cui et al., 2021) and learnable prompts (Wang et al., 2022; Smith et al., 2023; Wang et al., 2024). These methods are mainly designed for unimodal tasks, particularly image classification, and cannot be directly applied to sequence labeling tasks. To this end, this paper proposes a dual-branch prompt fusion, bridging the gap between the boundary-sensitive requirements of sequence labeling and the global-level semantic encoding of prompt-based paradigms.

3 Preliminaries

To address continual learning in CMERJE, this paper adopts a prompt-based paradigm that freezes the pre-trained backbone and encodes task-specific knowledge via learnable prompts, optimizing only the prompt parameters and the decoder.

Vision-language pre-trained models (VLMs) can learn generalizable multimodal representations from large-scale image-text data. Existing VLMs are typically categorized into dual-tower and single-tower architectures. Compared to a dual-tower architecture that requires separate encoding, a single-tower model can achieve deeper cross-modal interactions, making it more suitable for multimodal understanding under a frozen backbone. Thus, this paper adopts VLMO (vision-language pre-trained model) as the backbone, supporting both dual- and single-tower encoding for different tasks. It projects tokens and visual patches into a shared semantic space and encodes modality-specific representations via a multi-expert Transformer.

4 Methods

This paper proposes a multimodal prompt-based boundary-enhanced continual framework for the CMERJE task, as illustrated in Figure 2.

4.1 Task Formulation

The CMERJE task is formulated for continual multimodal entity-relation extraction, where the model receives a stream of tasks with multimodal data from distinct label spaces or domains. The goal is to sequentially learn new types from each task and retain knowledge from previous tasks.

Formally, let $\mathcal{T} = \{\mathcal{T}^{(1)}, \ldots, \mathcal{T}^{(N)}\}$ denote the task sequence and $\mathcal{Y} = \{\mathcal{Y}^{(1)}, \ldots, \mathcal{Y}^{(N)}\}$ denote label set, where task $\mathcal{T}^{(n)}$ is associated with dataset

$\mathcal{D}^{(n)}$ and its corresponding label set $\mathcal{Y}^{(n)}$. Notably, CMERJE is formulated under a class-incremental setting, such that the label sets of different tasks are mutually disjoint, i.e. $\forall n \neq n', \mathcal{Y}^{(n)} \cap \mathcal{Y}^{(n')} = \varnothing$.

For the current task $\mathcal{T}^{(n)}$, CMERJE aims to extract a set of quintuples $y = \{(e_\mathrm{s}, t_\mathrm{s}, e_\mathrm{o}, t_\mathrm{o}, r)\}$ from image-text pairs, where each quintuple consists of subject e_s, subject type e_o, object t_s, object type t_o, and their relation r, and $(t_\mathrm{s}, t_\mathrm{o}, r) \in \mathcal{Y}^{(n)}$.

4.2 Type-aware Instance-level Prompt Matching

This module enables knowledge accumulation and reuse by selectively freezing and activating prompts from the multimodal pool, and subsequently performs instance-level prompt matching to guide the frozen backbone for task-specific inference.

Multimodal Prompt Pool: MPBoCo constructs a multimodal prompt pool composed of multiple task-specific learnable key-prompt pairs. Each pair consists of a semantic key that serves as a retrieval cue, and a corresponding prompt that encodes learnable knowledge. Key-prompt pairs learned from different tasks are incrementally stored in a unified multimodal prompt pool $\mathcal{B}$, enabling knowledge accumulation while preventing interference across tasks. Formally, the prompt pool is defined as:

$$\begin{aligned}\mathcal{B} &= \bigcup_{n=1}^{N} \left(\mathbf{K}^{(n)}, \mathcal{P}^{(n)}\right) \\ &= \bigcup_{n=1}^{N} \bigcup_{m=1}^{M} \left(\mathbf{k}^{(n,m)}, \mathbf{P}^{(n,m)}\right)\end{aligned} \tag{1}$$

where $\mathbf{K}^{(n)} = \{\mathbf{k}^{(n,1)}, \ldots, \mathbf{k}^{(n,M)}\} \in \mathbb{R}^{M \times d_k}$ denotes the key set for the n-th task, and $\mathbf{k}^{(n,m)} \in \mathbb{R}^{d_k}$ is the m-th key in $\mathbf{K}^{(n)}$. Similarly, $\mathcal{P}^{(n)} = \{\mathbf{P}^{(n,1)}, \ldots, \mathbf{P}^{(n,M)}\} \in \mathbb{R}^{M \times L_p \times J \times d_p}$ denotes the prompt set for the n-th task, where $\mathbf{P}^{(n,m)} \in \mathbb{R}^{L_p \times J \times d_p}$ is the m-th prompt in $\mathcal{P}^{(n)}$. Here, M is the number of prompt pairs learned per task, L_p is the length of a single prompt, and J is the number of Transformer layers in the frozen pre-trained backbone to which prompts are attached.

To enhance alignment between semantic keys and label semantics, keys are initialized with type descriptions as prior knowledge. Specifically, type descriptions composed of subject, relation, and object types are encoded by the frozen text encoder to obtain the initial key representations. During continual training, only prompt pairs associated with the current task are optimized. Once training is completed, the learned prompt pairs are frozen and retained in the prompt pool to prevent cross-task knowledge interference.

Type-Aware Instance-Level Prompt Matching: To enable adaptive instance-level matching, a multimodal query function projects multimodal inputs into the same representation space as the keys. Following prior researches (Wang et al., 2022; Li et al., 2024), MPBoCo directly adopts the frozen VLMO backbone as the query function to extract the multimodal representation $\mathbf{f} \in \mathbb{R}^{d_k}$.

To further enhance the discriminative power of multimodal representations across different types, the learnable type-aware enhancement components $\mathbf{A} = \bigcup_{n=1}^{N} \mathbf{A}^{(n)}$ are introduced, where $\mathbf{A}^{(n)} = \{\mathbf{a}^{(n,1)}, \ldots, \mathbf{a}^{(n,M)}\} \in \mathbb{R}^{M \times d_k}$. It aims to focus on type-related semantic cues (e.g., facial attributes or architectural appearance) while suppressing irrelevant information (e.g., background noise).

The enhanced multimodal representations are matched with all keys, and the resulting key-prompt pairs guide the frozen backbone for task-specific inference. Following prior work (Wang et al., 2022; Li et al., 2024), cosine similarity is adopted as the scoring function. Given the current task $\mathcal{T}^{(N')}$, the module selects the top-r keys by optimizing the following objective:

$$\mathcal{I}^* = \underset{|\mathcal{I}|=r}{\arg\min} \sum_{(n,m)\in\mathcal{I}} \gamma\left(\mathbf{f} \odot \mathbf{a}^{(n,m)}, \mathbf{k}^{(n,m)}\right) \tag{2}$$

where $\mathcal{I} \subseteq \{(n,m) \,|n \in [1, N'], m \in [1, M]\}$, $\gamma(\cdot)$ is a similarity function, and $\odot$ is the element-wise product. The matched key subset $\mathbf{K}_r$ and their corresponding prompt subset $\mathcal{P}_r$ are defined as:

$$\mathbf{K}_r = \{\mathbf{k}^{(n,m)} \mid (n,m) \in \mathcal{I}^*\} \tag{3}$$

$$\mathcal{P}_r = \{\mathbf{P}^{(n,m)} \mid (n,m) \in \mathcal{I}^*\} \tag{4}$$

where $\mathbf{K}_r \in \mathbb{R}^{r \times d_k}$ and $\mathcal{P}_r \in \mathbb{R}^{r \times L_p \times J \times d_p}$.

4.3 Boundary-Enhanced Dual-Branch Prompt Fusion

This module bridges global prompt representations and local boundary semantics through hierarchical prompt fusion and a dual-branch decoder.

Hierarchical Prompt Fusion Strategy: This module adopts a hierarchical prompt fusion strategy, where prompts are injected layer-wise into the Transformer backbone by augmenting the key and value components of the attention mechanism, instead of appending prompts to the input tokens.

For a Transformer layer with prompt fusion, the h-head self-attention is computed as:

$$\mathrm{MHSA}(\mathcal{P}_r) = [\mathrm{SA}_1(\mathcal{P}_r); \ldots; \mathrm{SA}_h(\mathcal{P}_r)]\, \mathbf{W}' \quad (5)$$

$$\mathrm{SA}_\xi(\mathcal{P}_r) = \mathrm{softmax}\left(\frac{\left(\mathbf{Q}\mathbf{W}_\xi^{\mathrm{q}}\right)\left[\mathcal{P}_r^{\mathrm{k}}; \mathbf{K}\mathbf{W}_\xi^{\mathrm{k}}\right]^\top}{\sqrt{d_h}} \right) \cdot \left[\mathcal{P}_r^{\mathrm{v}};\, \mathbf{V}\mathbf{W}_\xi^{\mathrm{v}}\right] \quad (6)$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ denote the query, key, and value components, and $\mathrm{SA}_\xi(\cdot)$ represents the ξ-th attention head. The scaling factor is defined as $d_h = d_p/h$. $\mathbf{W}_M$, $\mathbf{W}_z^Q$, $\mathbf{W}_z^K$, and $\mathbf{W}_z^V$ are learnable projection matrices. The matched prompt set $\mathcal{P}_r$ is split for concatenation with keys and values, denoted as $\mathcal{P}_r^{\mathrm{k}}$ and $\mathcal{P}_r^{\mathrm{v}} \in \mathbb{R}^{r\times(L_p/2)\times J\times d_p}$.

Auxiliary Branch: The upper Transformer layers of VLMO perform cross-modal interaction and fusion via VL-FFN, generating multimodal representation $\mathbf{H}_{\mathrm{VL}}$. MPBoCo preserves the VL-FFN as the main branch, which exploits visual context to alleviate textual ambiguity in social media scenarios. However, noisy or weakly aligned visual regions may introduce boundary shifts, which may weaken semantics such as lexical morphology and syntactic continuity after fusion. To this end, this paper retains a parallel auxiliary branch (T-FFN) that generates the textual representation $\mathbf{H}_{\mathrm{T}}$. It provides fine-grained boundary cues to guide the main branch, improving entity boundary modeling while retaining strong multimodal type discrimination.

Dual-Branch Fusion: The label space of the auxiliary task exhibits a structured dependency on that of the main task, as entity span detection is closely related to joint entity-relation extraction.

To explicitly encode this inter-task correspondence, this paper introduces a task transformation matrix $\mathbf{W}_{\mathrm{T2VL}} \in \mathbb{R}^{|\mathcal{Z}|\times|\mathcal{Y}|}$, where $|\mathcal{Z}|$ and $|\mathcal{Y}|$ denote the size of label spaces for auxiliary and main branches, respectively. Each entry $\mathbf{W}_{\mathrm{T2VL}}[j,k]$ represents the transformation weight from an auxiliary label $z_j \in \mathcal{Z}$ to a main-branch label $y_k \in \mathcal{Y}$.

The initialization of this matrix is based on prior knowledge between labels. For each auxiliary label z_j, a corresponding valid label subset $\mathcal{C}_j \subseteq \mathcal{Y}$ is defined. For valid transitions ($y_k \in \mathcal{C}_j$), the transformation weight is set to $\mathbf{W}_{\mathrm{T2VL}}[j,k] = |\mathcal{C}_j|^{-1}$, while invalid transitions ($y_k \notin \mathcal{C}_j$) are assigned zero. For example, under the IOB2 tagging scheme, the auxiliary label `B` (entity beginning) can only be mapped to type-specific beginning labels such as `B-PER` or `B-ORG`, but not to the non-entity label `O`.

After obtaining the transition matrix, auxiliary span predictions are incorporated into the emission scores of the main-branch decoder, allowing explicit boundary guidance for joint prediction. In CMERJE, entities may play different semantic roles across relations, making role awareness crucial for modeling relation directionality. To alleviate this ambiguity, this paper explicitly projects entity representations into role-specific semantic spaces. The subject- and object-specific representations $\mathbf{H}_{\mathrm{VL}}^{\mathrm{s}}, \mathbf{H}_{\mathrm{VL}}^{\mathrm{o}}$ can be computed:

$$\mathbf{H}_{\mathrm{VL}}^{\mathrm{s}} = \gamma_{\mathrm{VL}}^{\mathrm{s}}\left(\mathbf{H}_{\mathrm{VL}} + \lambda_{\mathrm{T}}\left(\gamma_{\mathrm{T}}^{\mathrm{s}}\left(\mathbf{H}_{\mathrm{T}}\right)\mathbf{W}_{\mathrm{T2VL}}\right)\right) \quad (7)$$

$$\mathbf{H}_{\mathrm{VL}}^{\mathrm{o}} = \gamma_{\mathrm{VL}}^{\mathrm{o}}\left(\mathbf{H}_{\mathrm{VL}} + \lambda_{\mathrm{T}}\left(\gamma_{\mathrm{T}}^{\mathrm{o}}\left(\mathbf{H}_{\mathrm{T}}\right)\mathbf{W}_{\mathrm{T2VL}}\right)\right) \quad (8)$$

where λ_{T} controls the contribution of auxiliary-branch, and $\gamma_{\mathrm{VL}}^{\mathrm{s/o}}$ and $\gamma_{\mathrm{T}}^{\mathrm{s/o}}$ are role-specific projection functions for the main and auxiliary branches. Finally, the role-specific representations are separately decoded for subject and object prediction:

$$\tilde{\mathbf{H}}_{\mathrm{VL}} = \{\mathbf{H}_{\mathrm{VL}}^{\mathrm{s}}, \mathbf{H}_{\mathrm{VL}}^{\mathrm{o}}\} \quad (9)$$

$$\tilde{\mathbf{H}}_{\mathrm{T}} = \{\gamma_{\mathrm{T}}^{\mathrm{s}}(\mathbf{H}_{\mathrm{T}}), \gamma_{\mathrm{T}}^{\mathrm{o}}(\mathbf{H}_{\mathrm{T}})\} \quad (10)$$

4.4 Joint Training Objective

This module jointly optimizes entity-relation prediction, boundary modeling, prompt consistency, and key regularization in an end-to-end objective.

Dual-Branch Decoding: Following prior work (Yu et al., 2020; Liu et al., 2026a), this paper adopts Conditional Random Fields (CRF) as the decoder for sequence labeling, where the conditional probability can be defined:

$$p\left(y \mid \mathbf{H}\right) = \frac{\prod_{i=0}^{N-1} \psi\left(y_i, y_{i+1}; \mathbf{H}\right)}{\sum_{y'\in Y}\prod_{i=0}^{N-1}\psi\left(y_i, y_{i+1}; \mathbf{H}\right)} \quad (11)$$

where $\psi(\cdot)$ denotes a potential function, and Y is a set of all possible labels. Under the dual-branch architecture, losses for both branches are optimized using the CRF negative log-likelihood:

$$\mathcal{L}_{\mathrm{D}} = -\left(\log p\left(y \mid \tilde{\mathbf{H}}_{\mathrm{VL}}\right) + \log p\left(y_{\mathrm{BIO}} \mid \tilde{\mathbf{H}}_{\mathrm{T}}\right)\right) \quad (12)$$

where $y \in \mathcal{Y}^{(n)}$, and the auxiliary-branch optimizes entity boundary prediction using labels y_{BIO}.

Prompt Consistency Loss: During the instance-level prompt matching, prompt selection is a discrete operation, which prevents direct updates to

type-aware enhancement components and multimodal key-prompt pairs. To enable the end-to-end optimization, this paper introduces a prompt consistency loss $\mathcal{L}_{\text{Top}}$, which encourages the similarity scores between the enhanced query representations and candidate prompts to match the one-hot encoding of the ground-truth prompt index $\hat{y}$.

Multimodal Key Regularization: Since the semantic keys $\boldsymbol{K}$ are learnable parameters, they are susceptible to catastrophic semantic drift during continual training, potentially deviating from their initial semantics. Therefore, this paper regularizes the learned keys by minimizing the mean squared error between updated keys and their initial values:

$$\mathcal{L}_{\text{Key}} = \text{MSE}(\mathbf{K}^{(N')}, \mathbf{K}'^{(N')}) \tag{13}$$

where $\mathbf{K}^{(N')}$ and $\mathbf{K}'^{(N')}$ denote keys before and after training on task $T^{(N')}$, respectively.

In summary, the final training objective consists of three loss components, defined as follows:

$$\mathcal{L} = \mathcal{L}_{\text{D}} + \mathcal{L}_{\text{Top}} + \mathcal{L}_{\text{Key}} \tag{14}$$

During inference, the model predicts subject and object label sequences by maximizing the conditional probability. The final relational quintuple outputs are then constructed using the DualSeqTag decoding strategy introduced in (Liu et al., 2026a).

5 Experiments

5.1 Experimental Setup

Datasets: To support research on continual multimodal joint entity and relation extraction, this paper constructs a new dataset, named **CMERJE**. It is built upon existing multimodal named entity recognition (MNER) (Zhang et al., 2018) and multimodal relation extraction (MRE) (Zheng et al., 2021) datasets, with supplementary annotations inspired by MNER-MI (Huang et al., 2024) to improve label balance. The detailed construction process is described in Appendix A.

Implementation Details: This paper adopts a frozen pre-trained VLMO as our backbone and updates only the decoder and prompt-related parameters. Prompt hyperparameters are tuned on the validation set. A total prompt length of 50 achieves the best performance (25 for keys and values), with one type-specific prompt selected per task. Given that VLMO consists of 10 unimodal and 2 multimodal expert layers, prompts are injected into 8 unimodal layers and all multimodal ones. MPBoCo is trained using Adam with a cosine learning rate schedule and warm restarts.

Evaluation Metric: Following prior work (Chen et al., 2024), we adopt average F1 (AF1) as the evaluation metric. AF1 provides a comprehensive evaluation under class imbalance by averaging the F1-score across all learned tasks. Let F1_n denote the F1-score on the test set after training up to task n. The AF1 score is defined as:

$$\text{AF1}_N = \frac{1}{N}\sum_{n=1}^{N}\text{F1}_n \tag{15}$$

5.2 Baselines

To demonstrate the effectiveness of MPBoCo, we compare it with several baselines. Non-continual baselines include Vanilla BERT, which fine-tunes BERT on each task and serves as a lower bound for catastrophic forgetting, as well as three multimodal extraction models, including UMT (Yu et al., 2020), UMGF (Zhang et al., 2021), and HiTIMI (Liu et al., 2026b). In addition, UMT, UMGF, and HiTIMI are extended with random replay (RR) and influence replay (IR) strategies for multimodal continual learning, where RR replays randomly selected samples and IR selectively replays high-impact samples for better knowledge retention.

5.3 Experimental Results

For the non-continual baselines, Vanilla BERT suffers from severe catastrophic forgetting. Although UMT, UMGF, and HiTIMI achieve strong initial performance through cross-modal fusion, their performance degrades rapidly, highlighting the necessity of continual learning strategies.

For the continual baselines, EWC and RP-CRE remain inferior to multimodal continual methods, indicating the importance of visual information for semantic disambiguation. Replay-based multimodal methods further improve knowledge retention, with influence replay (IR) consistently outperforming random replay (RR) by replaying high-impact samples, thereby exhibiting greater stability across variations in modality distributions. Unlike replay-based methods, MPBoCo matches semantically relevant prompts for task-specific reasoning, achieving superior knowledge retention.

5.4 Ablation Study

To evaluate the effectiveness of each component in MPBoCo, we conduct ablation studies in Table 2.

Method	CMERJE										
	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	Avg
Vanilla BERT	81.6	32.9	24.1	14.4	7.3	6.2	4.4	5.4	3.8	3.6	18.4
UMT	84.2	35.5	26.4	15.7	8.8	6.4	5.5	6.7	5.3	4.3	19.9
UMGF	83.3	34.6	25.6	15.0	6.8	5.1	5.3	4.8	2.2	3.5	18.6
HiTIMI	**85.0**	37.5	29.4	16.2	7.4	7.0	6.5	6.8	4.9	3.1	20.4
UMT-RR	84.2	36.3	27.4	22.9	6.9	8.1	7.0	7.3	6.2	6.4	21.3
UMGF-RR	83.3	36.1	26.2	21.2	6.3	4.7	5.2	5.7	4.6	3.7	19.7
HiTIMI-RR	85.0	37.6	32.4	26.1	9.2	7.9	6.4	7.7	8.8	6.3	22.7
UMT-IR	84.2	35.3	27.0	24.1	8.7	8.1	8.1	7.7	6.5	6.9	21.6
UMGF-IR	83.3	35.4	27.3	21.2	6.4	5.1	5.8	5.6	5.3	3.9	19.9
HiTIMI-IR	85.0	38.2	32.8	26.3	8.6	7.2	7.3	8.8	8.9	6.9	23.0
MPBoCo (Ours)	82.7	**48.7**	**33.7**	**28.5**	**20.4**	**17.6**	**17.4**	**13.4**	**11.7**	**10.7**	**28.5**

Table 1: Experimental results on the CMERJE dataset under the 10-task continual setting.

Ablation Variants	CMERJE										
	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	Avg
w/o KReg	80.0	45.4	32.1	27.1	**21.3**	17.5	16.6	13.3	11.5	10.3	27.5
w/o TAE	77.0	34.5	21.4	13.4	12.3	7.1	4.8	3.2	3.0	2.2	17.9
w/o TFFN	77.3	44.2	31.1	28.2	18.4	16.1	12.9	10.1	9.3	8.0	25.6
w/o TM	78.6	46.9	33.0	28.5	18.1	16.7	15.1	12.4	6.4	4.6	26.0
w/o SO	75.3	45.6	25.8	21.6	12.1	8.4	6.1	5.7	1.3	0.8	20.3
MPBoCo	**82.7**	**48.7**	**33.7**	**28.5**	20.4	**17.6**	**17.4**	**13.4**	**11.7**	**10.7**	**28.5**

Table 2: Ablation results of important components in the MPBoCo framework.

Method	CMERJE					
	T1	T2	T3	T4	T5	Avg
Vanilla BERT	78.4	29.8	24.3	14.8	8.9	31.3
UMT	**81.4**	31.0	25.2	15.9	8.6	32.4
UMGF	79.2	29.7	24.4	15.0	8.7	31.4
HiTIMI	81.2	30.1	26.4	15.6	9.1	32.5
UMT-RR	81.4	35.9	27.8	16.2	12.2	34.7
UMGF-RR	79.2	33.6	26.7	19.3	10.4	33.8
HiTIMI-RR	81.2	35.0	26.2	19.6	13.6	35.1
UMT-IR	81.4	33.1	29.2	19.0	13.2	35.2
UMGF-IR	79.2	34.7	27.0	19.3	9.8	34.0
HiTIMI-IR	81.2	33.4	28.9	20.2	12.7	35.3
MPBoCo (Ours)	75.1	**46.3**	**36.7**	**29.0**	**25.7**	**42.5**

Table 3: Experimental results on the CMERJE dataset under the 5-task continual setting.

Instance-level prompt matching. *w/o KReg* removes the key regularization process, allowing keys to deviate from their initial semantics, thereby reducing matching accuracy and cross-task generalization. *w/o TAE* removes the type-aware enhancement components, weakening the model's ability to focus on type-relevant semantics.

Boundary-enhanced dual-branch fusion. *w/o TFFN* removes the auxiliary text branch, degrading entity boundary predictions and highlighting its role in preserving local textual continuity. *w/o TM* replaces the task-specific transition matrix with a fully connected one, removing explicit structural guidance and leading to performance degradation. *w/o SO* removes subject/object-specific representations, weakening directional relation modeling and compromising the stability of joint decoding.

5.5 Robustness Analysis

The granularity of task partitioning is crucial for evaluating the performance of continual learning, as it affects the task sequence length (increasing inter-task interference) and the number of relation types within each task (increasing intra-task learning complexity). To comprehensively evaluate the generalization ability of MPBoCo under different task granularities, we construct additional continual learning settings on the CMERJE dataset.

As shown in Table 3, MPBoCo consistently outperforms baselines in the 5-task setting, demonstrating strong robustness to task granularity. This gain mainly stems from task-specific prompts and the multimodal type-aware prompt matching strategy, which effectively matches relevant prompts

Prompt Layers (Uni-Multi)	CMERJE										
	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	Avg
4-1	74.3	45.8	32.4	25.4	18.0	15.2	15.4	11.3	10.5	10.0	25.8
6-1	77.8	46.7	34.0	27.3	18.9	16.0	15.6	12.8	7.5	5.0	26.2
8-1	80.5	47.3	31.5	27.7	18.4	15.7	14.7	12.3	10.0	8.7	26.7
10-1	76.6	44.9	20.7	14.3	12.9	15.0	13.0	9.8	8.0	5.1	22.0
4-2	76.2	46.3	30.3	27.0	16.7	13.1	12.5	9.4	6.9	4.7	24.3
6-2	78.6	47.2	32.2	26.2	19.9	16.7	16.9	12.8	10.2	8.9	27.0
8-2	82.7	48.7	33.7	28.5	20.4	17.6	17.4	13.4	11.7	10.7	28.5
10-2	77.2	46.8	33.5	25.2	17.4	14.2	13.5	10.0	7.0	5.0	25.0

Table 4: Performance sensitivity of prompt attachment layers under the 10-task continual setting.

Prompt Layers (Uni-Multi)	CMERJE					
	T1	T2	T3	T4	T5	Avg
4-1	72.9	44.5	24.8	12.7	11.5	33.3
6-1	73.9	42.0	30.0	18.9	16.7	36.3
8-1	74.8	44.0	32.6	25.4	20.1	39.4
10-1	73.3	44.9	25.7	14.3	12.9	34.2
4-2	73.3	46.3	21.8	18.7	16.4	35.3
6-2	74.5	44.3	32.8	27.1	21.9	40.1
8-2	75.1	46.3	36.7	29.0	25.7	42.6
10-2	74.2	43.9	29.6	21.9	17.5	37.4

Table 5: Performance sensitivity of prompt attachment layers under the 5-task continual setting.

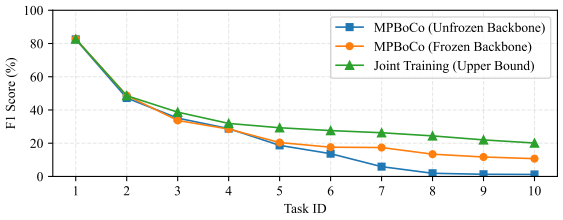

Figure 3: Performance comparison of MPBoCo with frozen backbone, unfrozen backbone, and joint training.

while alleviating inter-task knowledge interference.

The observed performance gains are attributed to task-specific prompts and a multimodal type-aware prompt matching strategy, which enables matching of relevant prompts while substantially alleviating inter-task knowledge interference.

5.6 Further Analysis

Sensitivity Analysis of Prompt Fusion: We analyze the number of prompt layers attached to unimodal and multimodal expert layers, considering 4, 6, 8, or 10 unimodal and 1 or 2 multimodal layers.

Tables 4 and 5 show that few unimodal prompt layers lead to limited task-specific semantics, while too many introduce redundancy. For multimodal experts, two prompt layers consistently yield better performance than a single layer, owing to more effective cross-modal interaction modeling. Overall, the configuration with 8 unimodal and 2 multimodal prompt layers achieves the best performance across all settings, demonstrating a favorable trade-off between knowledge retention and adaptation.

Effect of Backbone Freezing: To investigate the impact of backbone freezing, we train MPBoCo with all backbone parameters unfrozen. This improves adaptation to new tasks but increases forgetting of previous tasks, highlighting the trade-off between plasticity and stability. Figure 3 validates the effectiveness of freezing the backbone in the prompt-based paradigm for achieving a balance between knowledge retention and adaptability.

Effect of Joint Training: To quantify the performance ceiling of MPBoCo in CMERJE, we conduct joint training with all previous and current task data. As illustrated in Figure 3, this establishes an upper bound on achievable performance. MPBoCo closely approaches this bound, demonstrating effective knowledge retention while learning new tasks without access to historical instances.

Conclusion

In this paper, we propose an MPBoCo framework for continuous joint entity-relation extraction in dynamic social media scenarios. MPBoCo leverages task-specific prompts, type-aware prompt matching, and auxiliary predictions to explicitly model cross-modal dependencies and entity boundaries. Extensive experiments on the CMERJE dataset demonstrate that MPBoCo consistently outperforms baselines across continual settings, achieving superior knowledge retention and adaptability.

Limitations

Despite its excellent performance, MPBoCo still has several limitations. First, we utilize pre-trained VLMO expert layers that cover both textual and visual modalities. When handling other modalities, it may be necessary to pre-train expert layers for those modalities in advance. Second, while MPBoCo applies prompt-based paradigms to continual sequence labeling, it remains limited in fine-grained boundary recognition, and improving its adaptability in such tasks is an important direction for future work. This also suggests a promising direction toward unifying parametric knowledge in LLMs with continual learning frameworks.

Acknowledgements

This work was supported by the National Science and Technology Major Project (Grant No. 2025ZD0123702), the National Natural Science Foundation of China (Grant No. 62541604), and the Heilongjiang Provincial Discipline Innovation Project (Grant No. LJGXCG2024-F10).

References

Xiang Chen, Jingtian Zhang, Xiaohan Wang, Ningyu Zhang, Tongtong Wu, Yuxiang Wang, Yongheng Wang, and Huajun Chen. 2024. Continual multimodal knowledge graph construction. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 6225–6233.

Jian Cheng, Kaifang Long, Shuang Zhang, Tian Zhang, Lianbo Ma, Shi Cheng, and Yinan Guo. 2023. Text-image scene graph fusion for multi-modal named entity recognition. *IEEE Transactions on Artificial Intelligence*, 5(6):2828–2839.

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. Association for Computational Linguistics.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. 2020. Orthogonal gradient descent for continual learning. In *Proceedings of the Twenty-Third International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108, pages 3762–3773.

Xu Han, Yi Dai, Tianyu Gao, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2020. Continual relation learning via episodic memory activation and reconsolidation. In *Proceedings of the Fifty-Eighth Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6429–6440.

Zhiqiang Hu, Xinyu Li, Xinyu Pan, Sijie Wen, and Jinsong Bao. 2025. A question answering system for assembly process of wind turbines based on multimodal knowledge graph and large language model. *Journal of engineering design*, 36(7-9):1093–1117.

Shizhou Huang, Bo Xu, Changqun Li, Jiabo Ye, and Xin Lin. 2024. Mner-mi: A multi-image dataset for multimodal named entity recognition in social media. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11452–11462.

Shubin Huang, Yi Cai, Li Yuan, and Jiexin Wang. 2025. A knowledge-enhanced network for joint multimodal entity-relation extraction. *Information Processing & Management*, 62(3):104033.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Maple: Multi-modal prompt learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19113–19122.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences (PNAS)*, 114(13):3521–3526.

Junlin Lee, Yequan Wang, Jing Li, and Min Zhang. 2024. Multimodal reasoning with multimodal knowledge graph. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10767–10782.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597. Association for Computational Linguistics.

Yujie Li, Xin Yang, Hao Wang, Xiangkun Wang, and Tianrui Li. 2024. Learning to prompt knowledge transfer for open-world continual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13700–13708.

Wanying Liang, Pasquale De Meo, Yong Tang, and Jia Zhu. 2024. A survey of multi-modal knowledge graphs: Technologies and trends. *ACM Computing Surveys*, 56(11):1–41.

Luping Liu, Meiling Wang, Mozhi Zhang, Linbo Qing, and Xiaohai He. 2022. Uamner: uncertainty-aware multimodal named entity recognition in social media posts. *Applied Intelligence*, 52(4):4109–4125.

Peipei Liu, Gaosheng Wang, Hong Li, Jie Liu, Yimo Ren, Hongsong Zhu, and Limin Sun. 2024. Multi-granularity cross-modal representation learning for

named entity recognition on social media. *Information Processing & Management*, 61(1):103546.

Xinyu Liu, Guanglu Sun, Jing Jin, Fei Lang, and Suxia Zhu. 2026a. Citr: Context-driven implicit triple reasoning for joint multimodal entity-relation extraction. *Information Processing & Management*, 63(2):104388.

Xinyu Liu, Guanglu Sun, Fei Lang, and Suxia Zhu. 2025. Mpcl: Multimodal prompt learning for continual relation extraction with type-aware inter-modality alignment. *Information Fusion*, page 104025.

Xinyu Liu, Guanglu Sun, Lili Liang, Fei Lang, and Suxia Zhu. 2026b. Hitimi: Hierarchical type-driven inter-modality interaction framework for multimodal named entity recognition. *Neurocomputing*, 664:132131.

Arun Mallya and Svetlana Lazebnik. 2018. Packnet: Adding multiple tasks to a single network by iterative pruning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7765–7773.

Qi Qin, Han Peng, Wenpeng Hu, Dongyan Zhao, and Bing Liu. 2021. Bns: building network structures dynamically for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 20608–20620.

Reza Shokri and Vitaly Shmatikov. 2015. Privacy-preserving deep learning. In *Proceedings of the Twenty-Second ACM SIGSAC conference on computer and communications security (ACM SIGSAC)*, pages 1310–1321.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. 2023. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11909–11919.

Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rpbert: a text-image relation propagation-based bert model for multimodal ner. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13860–13868.

Hong Wang, Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Shiyu Chang, and William Yang Wang. 2019. Sentence embedding alignment for lifelong relation extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 796–806.

Jie Wang, Yan Yang, Keyu Liu, Zhiping Zhu, and Xiaorong Liu. 2023a. M3s: Scene graph driven multi-granularity multi-task learning for multi-modal ner. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:111–120.

Xinyi Wang, Zitao Wang, and Wei Hu. 2023b. Serial contrastive knowledge distillation for continual few-shot relation extraction. In *Findings of the Association for Computational Linguistics (ACL Findings)*, pages 12693–12706.

Zhiyuan Wang, Xiaoyang Qu, Jing Xiao, Bokui Chen, and Jianzong Wang. 2024. Incprompt: Task-aware incremental prompting for rehearsal-free class-incremental learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7240–7244.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. 2022. Learning to prompt for continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 139–149.

Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. 2020. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1038–1046. ACM.

Dianzhi Yu, Xinni Zhang, Yankai Chen, Aiwei Liu, Yifei Zhang, Philip S Yu, and Irwin King. 2024. Recent advances of multimodal continual learning: A comprehensive survey. *arXiv*.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3342—-3352. ACL.

Li Yuan, Yi Cai, Jin Wang, and Qing Li. 2023. Joint multimodal entity-relation extraction based on edge-enhanced graph alignment network and word-pair relation tagging. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, pages 11051–11059.

Li Yuan, Yi Cai, Jingyu Xu, Qing Li, and Tao Wang. 2025. A fine-grained network for joint multimodal entity-relation extraction. *IEEE Transactions on Knowledge and Data Engineering*, 37(1):1–14.

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

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, pages 5674–5681.

Zhiheng Zhang, Daojian Zeng, and Xue Bai. 2024. Improving continual few-shot relation extraction

Split	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Train	848	862	634	484	404	235	279	304	225	222
Val	114	110	90	63	56	30	43	49	25	37
Test	118	245	338	418	472	509	538	582	614	647

Table 6: Statistics of CMERJE under the 10-task continual setting.

Split	T1	T2	T3	T4	T5
Train	1152	1087	913	706	639
Val	163	135	133	100	86
Test	162	321	443	556	647

Table 7: Statistics of CMERJE under the 5-task continual setting.

through relational knowledge distillation and prototype augmentation. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 8756–8767.

Changmeng Zheng, Zhiwei Wu, Junhao Feng, Ze Fu, and Yi Cai. 2021. Mnre: A challenge multimodal dataset for neural relation extraction with visual evidence in social media posts. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Changmeng Zheng, Zhiwei Wu, Tao Wang, Yi Cai, and Qing Li. 2020. Object-aware multimodal named entity recognition in social media posts with adversarial learning. *IEEE Transactions on Multimedia*, 23:2520–2532.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 130(9):2337–2348.

A Dataset Construction Details

This section provides detailed descriptions of the construction process and statistics of the CMERJE dataset.

First, to facilitate joint extraction under the continual learning setting, we unify the annotation formats of existing MNER and MRE datasets. Following prior work (Chen et al., 2024), CMERJE is partitioned into 10 sequential tasks based on relation type semantics, with no relation label shared across tasks. This task partitioning ensures clear type-level heterogeneity and enables a controlled evaluation of catastrophic forgetting and knowledge accumulation. In addition to the default 10-task setting, we also construct an alternative 5-task variant by randomly merging different tasks to evaluate the robustness of different methods under varying task granularities. We report detailed dataset statistics for both the 10-task and 5-task settings in Tables 6 and 7, respectively. The statistics include the number of samples, entities, and relations in each task.

Second, to alleviate the label imbalance issue caused by the limited number of samples in certain relation types and the excessively high proportion of the None (N/A) type in the original dataset, we introduce supplementary annotations on the MNER-MI dataset to enhance the representativeness of the continual learning benchmark.

Finally, the original datasets employ a single entity-pair relation annotation paradigm, splitting multiple entity-relation quintuples within the same image-text pair into separate samples. To maintain multimodal semantic integrity, all quintuples within the same image-text pair are merged into a single complete instance, where all possible entity-relation tuples are jointly labeled.