Continual Multimodal Knowledge Graph Construction

Xiang Chen[1,3] , **Jingtian Zhang**[2,3] , **Xiaohan Wang**[2,3] , **Ningyu Zhang**[2,3*] , **Tongtong Wu**[4] , **Yuxiang Wang**[5] , **Yongheng Wang**[6] and **Huajun Chen**[1,3*]
[1]College of Computer Science and Technology, Zhejiang University
[2]School of Software Technology, Zhejiang University
[3]Zhejiang University - Ant Group Joint Research Center for Knowledge Graphs
[4]Monash University,
[5]Hangzhou Dianzi University,
[6]Zhejiang Lab
{xiang_chen, zhangjintian, wangxh07, zhangningyu, huajunsir}@zju.edu.cn
tongtong.wu@monash.edu, lsswyx@hdu.edu.cn, wangyh@zhejianglab.com

Abstract

Current Multimodal Knowledge Graph Construction (MKGC) models struggle with the real-world dynamism of continuously emerging entities and relations, often succumbing to catastrophic forgetting—loss of previously acquired knowledge. This study introduces benchmarks aimed at fostering the development of the continual MKGC domain. We further introduce **MSPT** framework, designed to surmount the shortcomings of existing MKGC approaches during multimedia data processing. MSPT harmonizes the retention of learned knowledge (stability) and the integration of new data (plasticity), outperforming current continual learning and multimodal methods. Our results confirm MSPT's superior performance in evolving knowledge environments, showcasing its capacity to navigate balance between stability and plasticity.

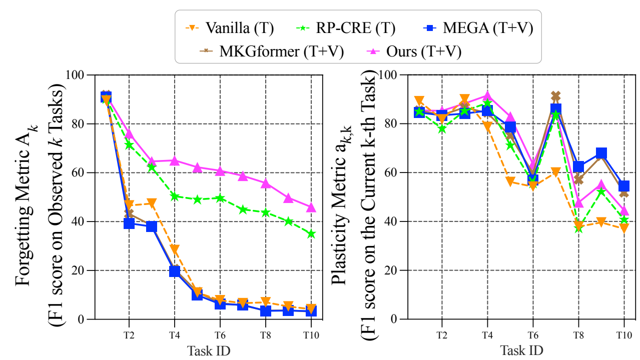


Figure 1: Results on incremental MRE (IMRE) benchmark. We benchmark MSPT against the Vanilla Training approach, multimodal KGC models such as MEGA and MKGformer, as well as the continual RE method RP-CRE.

1 Introduction

The rise of multimodal data on social media platforms has sparked significant interest among knowledge graph and multimedia researchers in the domain of multimodal knowledge graphs [Liu and et al, 2019; Zhu *et al.*, 2022; Zheng *et al.*, 2023; Hu *et al.*, 2023; Liang *et al.*, 2024]. To address the limitations of relying on human-curated multimodal data and to systematically extract insights from vast multimedia repositories, the concept of Multimodal Knowledge Graph Construction (MKGC) has been proposed [Zhang *et al.*, 2023; Liu *et al.*, 2023]. MKGC leverages multimodal data as an additional information source to disambiguate polysemous terms and perform tasks like Multimodal Named Entity Recognition (MNER)[Lu *et al.*, 2022] and Multimodal Relation Extraction (MRE)[Zheng *et al.*, 2021a]. However, existing MKGC architectures [Zheng *et al.*, 2021a; Chen *et al.*, 2022b] primarily focus on "static" knowledge graphs, where entity categories and relations remain fixed throughout the learning process. These models lack adaptability, especially when confronted with new entity categories and relations.

Addressing the dynamic nature of streaming data, replete with emerging entity categories and relations, the research community has developed the continuous Knowledge Graph Completion (KGC) methods [Monaikul *et al.*, 2021; Wang *et al.*, 2022a; Xia *et al.*, 2022], seeking to balance the integration of new entity categories and relations (plasticity) with the preservation of established knowledge (stability). While current continual KGC strategies are largely text-centric, neglecting the demands of MKGC, the latter's capacity to handle multimodal data can provide richer insights than text-only models. Historical evaluations [Chen and et al, 2022; Hu *et al.*, 2023] have demonstrated the superiority of MKGC in static KG settings. However, the preliminary experimental results in Figure 1 reveal significant hurdles when directly transferring MKGC models to a continual learning environment. Notably, MKGC models not only fall short of unimodal counterparts in previous tasks but also show limited effectiveness on current task test sets during continuous task training.

* Corresponding author.

We posit that this observation may decline in MKGC models during continual learning may stem from disparate convergence rates among different modalities [Wang *et al.*, 2020], leading to two primary challenges: *Challenge 1: How to alleviate the imbalanced learning dynamics across modalities to enhance the plasticity?* The differential learning dynamics in multimodal settings can hinder the adaptability of MKGC models to new entity categories and relations, especially when employing replay strategies. This imbalance may result in inferior representations, undermining the addition of new knowledge. *Challenge 2: How to reduce the forgetting in the process of multimodal interaction?* Continual MKGC models face the unique problem of varying forgetting rates across modalities, unlike their unimodal counterparts. These disparities can disproportionately affect secondary modalities, increasing the risk of forgetting and jeopardizing the performance of prior tasks. Addressing these challenges necessitates the development of continual MKGC models that ensure uniform multimodal forgetting and robust modality integration to manage the retention and acquisition of knowledge.

To overcome the highlighted challenges in continual MKGC, we introduce the Multimodal Stability-Plasticity Transformer (**MSPT**), a novel framework that advances the stability-plasticity trade-off through strategic multimodal optimization. Our method is distinguished by two pivotal modules: *(1) Gradient Modulation for Balanced Learning:* We propose a gradient modulation technique to address the imbalanced learning dynamics across modalities, thereby preserving the model's ability to learn new information. By adaptively tuning gradients according to each modality's optimization contribution, our approach ensures nuanced representation development for both modalities, enhancing plasticity. *(2) Hand-in-Hand Multimodal Interaction with Attention Distillation:* Deviating from traditional cross-attention multimodal interaction, MSPT calculates inter-modal self-query affinities against an external learnable key. This decoupling of fusion parameters allows for a more deliberate modulation of forgetting rates, promoting consistent knowledge retention. And the attention distillation is utilized to refine this process, leveraging the multimodal interaction outputs to preserve crucial attention patterns. The results of our thorough experiments demonstrate that **MSPT** outperforms both traditional MKGC and continual unimodal KGC models in various class-incremental settings, showcasing its potential in the field of continual MKGC[1].

2 Related Works

2.1 Advancements in MKGC

Multimodal Named Entity Recognition. Advancements in MNER have shifted from text-only approaches to also harnessing visual cues. Studies such as those by [Zhang *et al.*, 2018; Lu and et al, 2018; Moon *et al.*, 2018; Arshad *et al.*, 2019] have introduced interactions between CNN-driven visual and RNN-based textual features. Others, UMT [Yu *et al.*, 2020] and UMGF [Zhang *et al.*, 2021], have suggested utilizing fine-grained semantic correspondences with a combination of transformer and visual backbones, taking into account regional image features to represent objects. The ITA [Wang *et al.*, 2022b] model exploits self-attention to enrich text embeddings with image spatial context, showing superiority over text-centric models.

Multimodal Relation Extraction. Researchers have started exploring techniques to link entities mentioned in the textual content with corresponding objects depicted in associated images. Some examples include work done by [Zheng *et al.*, 2021b], who presents an MRE dataset that can associate the textual entities and visual objects for enhancing relation extraction. Then [Zheng *et al.*, 2021a] revises the MRE dataset based on [Zheng *et al.*, 2021b] and utilizes scene graphs to align textual and visual representations. [Wan and et al, 2021] also collects and labels four MRE datasets based on four famous works in China to address the scarcity of resources for multimodal social relations.

2.2 Continual Knowledge Graph Construction

Continual learning addresses catastrophic forgetting in the following strategies: consolidation-based methods [Zenke *et al.*, 2017; Liang *et al.*, 2023] that adjust parameter updates through regularization, dynamic architectures [Rusu *et al.*, 2016] that evolve with data, and rehearsal-based methods [Sprechmann *et al.*, 2018; Chaudhry *et al.*, 2019] using memory banks to preserve knowledge. The latter has exhibited superior performance in continual KGC [Monaikul *et al.*, 2021]. To address the challenge of continual RE, memory interaction methods [Cui *et al.*, 2021] have been proposed to effectively utilize representative samples. Additionally, prototype methods [Han *et al.*, 2020; Cui *et al.*, 2021] are increasingly employed to abstract relation information and mitigate overfitting. In the context of continual NER, the ExtendNER method [Monaikul *et al.*, 2021] tackles class-incremental learning by creating a unified NER classifier that encompasses all encountered classes over time. Moreover, approaches [Xia *et al.*, 2022; Wang *et al.*, 2022a] prevent forgetting of previous NER tasks by utilizing stored or generated data from earlier tasks during training. However, previous studies have focused on continual KGC and have not been readily applicable to MKGC due to the inherent challenges posed by multimodal data.

3 Preliminaries

3.1 Delineation of MKGC Tasks

Definition 1. *MNER. This subtask emphasizes the extraction of named entities from textual content and its associated images. Given a token sequence, denoted as* $x^t = [w_1, \ldots, w_m]$*, and its affiliated image patch sequence* x^v*, the principal goal of continual MNER is to consistently model the sequence tags' distribution, expressed as* $p(y|(x^t, x^v))$*. Within this context, for task* T_k*, the label sequence* y *is defined as* $y = [y_1, \ldots, y_m]$ *and integrates emergent entity types from the entity category set* $\mathcal{E}_k$*.*

Definition 2. *MRE. This subtask focuses on extracting relationships between designated entity pairs from token se-*

[1]Our data and code are available at https://github.com/zjunlp/ContinueMKGC

quences. For a given task T_k, and provided with a token sequence x^t and its corresponding image patch sequence x^v, the goal is to infer the relationship of a specific entity pair, (e_h, e_t), derived from x^t. A key challenge lies in computing the probability distribution over possible relations r from the set $\mathcal{R}_k$, expressed as $p(r|(x^t, x^v, e_h, e_t))$. This is made more complex by the potential addition of novel relations to $\mathcal{R}_k$.

3.2 Class-Incremental Continual Learning

We define a class-incremental continual learning scenario as a series of K separate tasks, each with its schema classes and MKGC corpus. Formally, the tasks are denoted as:

$$\mathcal{T} = [(\mathcal{S}^1, C^1), (\mathcal{S}^2, C^2), ..., (\mathcal{S}^K, C^K)]. \tag{1}$$

The k-th task T_k includes a distinct set of entity types $\mathcal{E}_k$ and relations $\mathcal{R}_k$, along with an MKGC corpus C^k which is divided into training, validation, and testing subsets $\mathcal{D}_k$, $\mathcal{V}_k$, and $\mathcal{Q}_k$, respectively. Each training instance in $\mathcal{D}_k$ consists of a textual input x^t, a sequence of image patches x^v—utilizing ViT encoding—and a corresponding label y, which is either an entity from $\mathcal{E}_k$ or a relation from $\mathcal{R}_k$. Learners are restricted to use only the data from $\mathcal{D}_k$ during the training phase of task T_k, and to ensure non-overlapping classes between tasks, we enforce $\mathcal{E}_i \cap \mathcal{E}_j = \emptyset$ and $\mathcal{R}_i \cap \mathcal{R}_j = \emptyset$ for $i \neq j$. This setup follows the convention of several benchmark methodologies [Masana *et al.*, 2023]. In our class-incremental MKGC setting, after training completes on $\mathcal{D}k$, the model undergoes evaluation across an aggregated test set $\cup i = 1^k \mathcal{Q}_i$, which includes all class categories up to the current task. This differs from task-incremental learning, where evaluation is confined to the specific task $\mathcal{S}^k$. The evaluation metrics are introduced as follows:

DEFINITION 3. *Forgetting Metric ($\mathbf{A}_k$): ** Measures the $F1$ score on aggregate test sets $\bigcup_{i=1}^k \mathcal{Q}_i$ for tasks $\{T_i\}_{i=1}^k$ post-training on T_k. It indicates the model's ability to prevent catastrophic forgetting, especially in sequential data with new entity categories and relations.*

DEFINITION 4. *Plasticity Metric ($\mathbf{U}_k$): ** Defined by the $F1$ score on the current task T_k, showcasing the model's capacity to learn new tasks while retaining existing knowledge, a critical aspect of continual learning.*

4 Methodology

4.1 Framework Overview

As illustrated in Figure 2, our continual KGC framework adopts a dual-stream Transformer structure with the task-specific paradigm, including:

(1) Structure. We incorporate a Visual Transformer (ViT) [Dosovitskiy *et al.*, 2021] for visual data and BERT for textual data. Building on prior research [Clark *et al.*, 2019; Chen *et al.*, 2022a], which indicates that manipulating the upper layers of language models (LMs) more effectively leverages knowledge for downstream tasks, our framework engages in multimodal interactions and attention distillation within the top three layers of the Transformers.

(2) Task-specific paradigm. For the MRE task, we employ a task-specific approach by fusing the [CLS] token representations from both ViT and BERT models. This integrated representation enables us to derive the probability distribution over the relation set $\mathcal{R}$ for the given task.

$$p(r|(x^t, x^v, e_h, e_t)) = \texttt{Softmax}(W \cdot [\boldsymbol{h}^t_{cls}; \boldsymbol{h}^v_{cls}]), \tag{2}$$

where $\boldsymbol{h}^t \in \mathbb{R}^{m_t \times d_t}$ and $\boldsymbol{h}^v \in \mathbb{R}^{m_v \times d_v}$ represent the output sequence embeddings from BERT and ViT, respectively. In the context of MNER, for fair benchmarking against prior work, we employ a CRF function akin to that in the MSPT framework. For the entity tag sequence $y = [y_1, \ldots, y_m]$, we enhance the BERT embeddings with $\boldsymbol{h}^v_{cls}$ and positional embeddings $\mathbb{E}^t_{\text{pos}}$ to capture visual information. The probability of a tag sequence y within the predefined label set Y is computed using the BIO tagging scheme that follows in [Lample *et al.*, 2016] as:

$$p(y_i|(x^t, x^v)) = \texttt{Softmax}(W \cdot [\boldsymbol{h}^t_i; (\boldsymbol{h}^v_{cls} + \mathbb{E}^t_{\text{pos}_i})]). \tag{3}$$

4.2 Balanced Multimodal Learning Dynamics

Modulating Optimization with Gradient. Diverse convergence rates across modality-specific parameters can lead to imbalanced learning dynamics during continual learning, potentially hampering current task performance. To address this, we propose a gradient modulation strategy to fine-tune the optimization of visual and textual encoders, depicted in Figure 2(b). Building upon concepts, we adapt these to the k-th task using the *Stochastic Gradient Descent* (SGD) algorithm:

$$\begin{aligned} \theta^{v(k)}_{n+1} &= \theta^{v(k)}_n - \eta \nabla_{\theta^v} \mathcal{L}_{CE}(\theta^{v(k)}_n) \\ &= \theta^{v(k)}_n - \eta \varphi(\theta^{v(k)}_n), \end{aligned} \tag{4}$$

where $\varphi(\theta^{v(k)}_n) = \frac{1}{N} \sum_{x \in B_n \nabla_{\theta^v}} \ell(x; \theta^{v(k)}_n)$ is an unbiased estimation of the full gradient, B_n represent a random mini-batch with N samples at step n of optimization, and $\nabla_{\theta^v} \ell(x; \theta^{v(k)}_n)$ denotes the gradient w.r.t. batch B_n.

Drawing from [Peng and et al, 2022], to counteract imbalanced multimodal learning dynamics, we introduce an adaptive gradient modulation mechanism for visual and textual modalities. This is based on quantifying their respective contributions to the learning goal via the contribution ratio γ_n:

$$\begin{aligned} s^v_i &= \sum_{y=1}^{M} 1_{y=y_i} \cdot \text{softmax}(W^v_n \cdot f^v_n(\theta^v, x^v_i))_y, \\ s^t_i &= \sum_{y=1}^{M} 1_{y=y_i} \cdot \text{softmax}(W^t_n \cdot f^t_n(\theta^t, x^t_i))_y, \end{aligned} \tag{5}$$

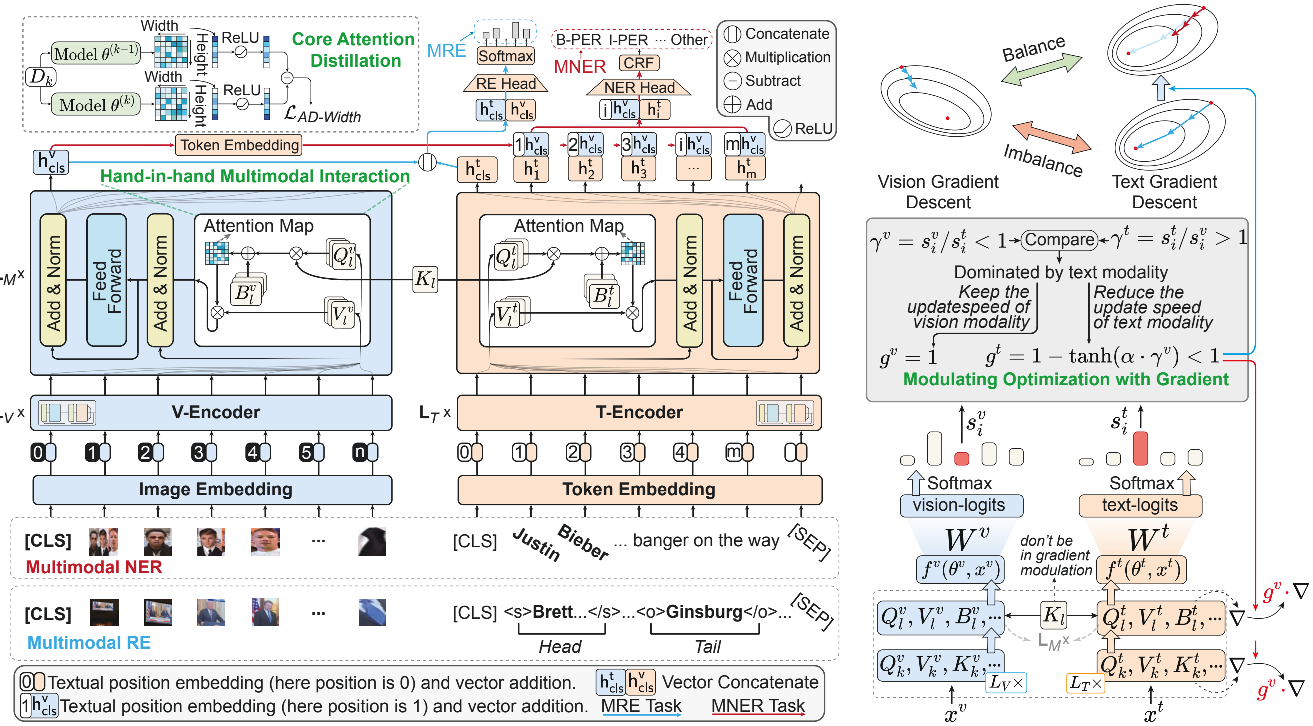
$$\gamma^t_n = \frac{\sum_{i \in B_n} s^t_i}{\sum_{i \in B_n} s^v_i}. \tag{6}$$

To dynamically assess the contribution ratio γ^t_n between textual and visual modalities, we introduce a modulation coefficient g^t_n that adaptively regulates the gradient, defined as:

$$g^t_n = \begin{cases} 1 - \tanh\left(\alpha \cdot \gamma^t_n\right) & \gamma^t_n > 1 \\ 1 & \text{otherwise} \end{cases}, \tag{7}$$

$$G^{t(k)} = \text{Avg}(g^{t(k)}_n), \tag{8}$$

where α is a hyper-parameter that adjusts the influence of modulation. , $G^{t(k)}$ is the averaged modulation coefficient



(a) Hand-in-hand Multimodal Interaction with Attention Distillation. (b) Balanced Multimodal Learning Dynamics.

Figure 2: Overview of our MSPT framework.

of the model trained after the task k. We further propose to balance the multimodal learning rhythm by integrating the coefficient g_n^t into the SGD optimization process of task k in iteration n as follows:

$$\theta_{n+1}^{v(k)} = \begin{cases} \theta_n^{v(k)} - \eta g_n^{t(k)} \varphi(\theta_n^{v(k)}) & k = 1 \\ \theta_n^{v(k)} - \eta G^{t(k-1)} \varphi(\theta_n^{v(k)}) & k > 1 \end{cases} . \tag{9}$$

Remark 1. *Through gradient modulation in task 1 and using the average coefficient from the preceding $k-1$ tasks to influence training on the current k-th task, we ensure a smoother transition in balanced multimodal learning across tasks.*

4.3 Hand-in-hand Multimodal Interaction via Attention Distillation

Inspired by the concept of collaborative progress with the saying goes *"hand in hand, no one is left behind"*, our method establishes a coherent framework for continual learning by integrating a dual-stream Transformer with attention distillation. As depicted in Figure 2(a), the model promotes uniform learning across various modalities, reducing unequal forgetting and enhancing multimodal resilience.

Hand-in-hand Multimodal Interaction

The self-attention mechanism (SAM) [Vaswani and et al, 2017], central to Transformer-based architectures, derives attention maps through self-key and self-query similarity calculations. Our proposed multimodal interaction approach introduces a unique attention generation process using shared learnable keys (K_W) and corresponding self-queries to enhance knowledge consolidation and retention. This method aims to counteract catastrophic forgetting by embedding previous task knowledge into the attention framework. It also promotes a tighter integration between visual and textual encoders, minimizing fusion bias and inconsistency associated with forgetting. Additionally, by regulating updates to K_W, our strategy preserves knowledge from earlier tasks, safeguarding against information degradation during new task.

Applying linear transformations to the input tensors X_v and X_t, we obtain the visual self-query $Q^{X_v} = W^{q_v} X_v$ and self-value $V^{X_v} = W^{v_v} X_v$, alongside the textual self-query $Q^{X_t} = W^{q_t} X_t$ and self-value $V^{X_t} = W^{v_t} X_t$ using the visual and textual encoders' parameters $W^{q_v}, W^{v_v}, W^{q_t}, W^{v_t}$, respectively. We introduce a shared external key K^s that supersedes the original self-key, generating updated attention maps for both modalities. For the k-th task, utilizing a ViT and BERT model, we denote the prescaled attention matrix at the l-th layer as $A_l^{(k)}$ and the resulting SAM output as $Z_l^{(k)}$, prior to softmax activation. Note that details on multi-head attention and normalization are omitted for conciseness.

$$\mathrm{A}_l^{v(k)} = \frac{Q_l^{X_v} (K_l^s)^\top + B_l^v}{\sqrt{d_v/H}}, \mathrm{A}_l^{t(k)} = \frac{Q_l^{X_t} (K_l^s)^\top + B_l^t}{\sqrt{d_t/H}}, \tag{10}$$

$$\begin{aligned} Z_l^{v(k)} &= \mathtt{Softmax}\left(\mathrm{A}_l^{v(k)}\right) V_l^{X_v}, \\ Z_l^{t(k)} &= \mathtt{Softmax}\left(\mathrm{A}_l^{t(k)}\right) V_l^{X_t}, \quad l = L-2, \dots, L, \end{aligned} \tag{11}$$

where L denotes the encoder's total number of layers, and B_l^v

IMRE Benchmark											
Model	Resource	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Vanilla	Text	89.6	46.6	47.4	28.4	11.0	7.8	6.6	7.0	5.3	4.2
EWC [Kirkpatrick *et al.*, 2016]	Text	89.2	38.4	37.8	37.0	10.4	7.1	5.8	2.2	2.9	3.0
EMR [Wang *et al.*, 2019]	Text	84.8	56.3	42.4	30.6	19.3	19.8	13.0	9.8	9.6	10.9
EMAR-BERT [Han *et al.*, 2020]	Text	90.3	57.0	61.1	49.1	48.4	45.5	37.6	31.0	30.1	25.8
RP-CRE [Cui *et al.*, 2021]	Text	90.3	71.5	62.2	50.4	49.1	49.7	45.0	43.8	40.2	35.1
UMT [Yu *et al.*, 2020]	Text+Image	90.2	38.2	37.2	18.3	10.3	6.5	5.5	3.2	3.3	2.8
UMGF [Zhang *et al.*, 2021]	Text+Image	90.7	41.5	38.3	18.8	10.7	6.0	5.2	3.3	3.5	3.0
MEGA [Zheng *et al.*, 2021a]	Text+Image	91.1	39.3	38.0	19.7	9.9	6.3	5.9	3.5	3.6	3.3
MKGformer [Chen *et al.*, 2022b]	Text+Image	**91.8**	43.2	38.5	20.5	11.2	6.7	5.7	3.6	3.7	3.2
M-[EWC]	Text+Image	91.8	46.3	40.0	31.8	11.8	7.4	6.8	4.5	5.8	3.9
M-[R-Replay]	Text+Image	91.8	60.2	61.4	50.5	43.1	40.0	34.5	31.7	28.4	30.8
M-[R-Replay & EWC]	Text+Image	91.8	62.3	57.6	49.5	44.5	38.7	31.9	34.3	31.3	28.0
Joint Training	Text+Image	91.8	82.0	82.6	79.2	77.2	75.8	71.4	72.6	67.8	67.0
MSPT	Text+Image	91.7	**76.2**	**64.7**	**65.1**	**62.3**	**60.9**	**58.8**	**55.8**	**49.8**	**46.0**

Table 1: Forgetting Metric $\mathbf{A}_k$ (F1 score (%)) on IMRE benchmark. The best performance results other than the upper bound model (Joint Training) are bolded. We explain uniformly that "R-Replay" refers to replay with random sampling. "M-[]" denotes the continual learning strategies applied to the multimodal method (here is MKGformer).

and B_l^t serve as bias terms for the vision and text attention maps, respectively. Nnote that the external key K_l^s is not constrained by the current feature input, allowing for end-to-end optimization and the integration of prior knowledge.

Core Attention Distillation

Accordingly, we introduce a method for refining the attention matrices within the dual-stream Transformer's interaction module. This involves stabilizing attention maps through a distillation function that leverages learnable shared keys K^s to mitigate forgetting. Considering the attention maps on the*visual side* across consecutive steps k and $(k-1)$, we quantify the distillation loss in the width dimension as:

$$\mathcal{L}^v_{\text{AD-width}}\left(\mathrm{A}_l^{v(k-1)}, \mathrm{A}_l^{v(k)}\right) = \sum_{h=1}^{H} \delta_W\left(\mathrm{A}_l^{v(k-1)}, \mathrm{A}_l^{v(k)}\right), \quad (12)$$

where H and W represent the height and weight of the attention maps. The total distance between attention maps a and b along the h or w dimension is represented by $\delta(a, b)$.

Our proposed attention distillation framework incorporates a crucial asymmetric distance function δ, diverging from typical continual learning approaches that use symmetric Euclidean distance for model outputs comparison across tasks k and $(k-1)$[Wang *et al.*, 2022a; Douillard *et al.*, 2020]. Symmetric distances tend to equally penalize shifts in attention from both new and old tasks, potentially impeding learning by increasing the loss when attention to previous tasks is maintained. Although preserving past tasks' knowledge mitigates forgetting, over-penalization can inadvertently suppress newly acquired insights, creating a tension between preserving past knowledge and embracing new information. To address this, we suggest δ, an asymmetric distance measure that conserves prior knowledge while sustaining the model's adaptability, aligning with findings in computer vision[Pelosin *et al.*, 2022]. The modified function, δ_W, is specifically crafted to balance the trade-off between plasticity and forgetting, illustrated as follows:

$$\delta_W\left(\mathbf{A}_l^{v(k-1)}, \mathbf{A}_l^{v(k)}\right) = \left\|\mathcal{F}_{\text{asym}}\left(\sum_{h=1}^{H} \mathbf{A}_{l_{w,h}}^{v(k-1)} - \sum_{h=1}^{H} \mathbf{A}_{l_{w,h}}^{v(k)}\right)\right\|. \quad (13)$$

We employ $\mathcal{F}_{\text{asym}}$, an asymmetric distance function, with ReLU [Nair and Hinton, 2010] integrated as $\mathcal{F}_{asym}$ in subsequent experiments. The attention distillation loss is:

$$\mathcal{L}_{\text{AD}} = \mathcal{L}^v_{\text{AD-width}} + \mathcal{L}^t_{\text{AD-width}}. \quad (14)$$

Remark 2. *This setup permits the development of new attention patterns during the k-th task without penalties, while attention absent in the current but present in the $(k-1)$-th task is penalized, promoting targeted knowledge retention.*

4.4 Training Objective

Our model leverages a cross-entropy loss ($\mathcal{L}$CE) to effectively recognize entities and relations, while an attention distillation loss ($\mathcal{L}$AD) mitigates the issue of catastrophic forgetting. We formulate the combined loss function as:

$$\mathcal{L}_{\text{all}} = \lambda\mathcal{L}_{\text{AD}} + \mathcal{L}_{\text{CE}}. \quad (15)$$

Here, λ serves as the weighting factor for the attention distillation loss. Additionally, we adopt the rehearsal strategy from PR-CRE to retain a concise memory set—merely six examples per task—for continual learning alignment, and optimizing memory footprint.

5 Experiments

5.1 Incremental MKGC Benchmarks

IMNER Benchmark. We utilize the established Twitter-2017 MNER dataset, which consists of multimodal tweets from 2016-2017, containing examples with multiple entity categories. To simulate more realistic learning conditions and reduce labeling ambiguity, we transition to a class-incremental framework, modifying the dataset such that each entity category is exclusive to a single task.

IMRE Benchmark. For our IMRE benchmark, we partition the dataset into 10 subsets for 10 distinct tasks. The original benchmark imposes two constraints that are at odds with the principles of lifelong learning: (1) clustering semantically related relations, and (2) excluding the "N/A" (not applicable) class. To rectify this, every task incorporates the "N/A" class, and relations are randomly sampled without bias, enhancing the benchmark's diversity and adherence to real-world lifelong learning conditions.

Model	Resource	IMNER Benchmark PER →	ORG →	LOC →	MISC	PER →	LOC →	ORG →	MISC
Vanilla	Text	74.2	38.9	20.8	12.6	74.2	26.9	35.0	12.8
EWC [Kirkpatrick *et al.*, 2016]	Text	76.2	40.5	20.7	14.4	76.3	28.9	36.9	13.4
EMR [Wang *et al.*, 2019]	Text	72.5	45.8	35.3	20.5	72.5	43.2	48.7	22.5
EMAR-BERT [Han *et al.*, 2020]	Text	73.2	48.5	42.5	28.7	73.2	45.5	50.3	30.8
ExtendNER [Monaikul *et al.*, 2021]	Text	50.7	53.3	47.8	41.4	55.6	52.9	57.9	47.9
UMT [Yu *et al.*, 2020]	Text+Image	71.2	37.3	19.1	11.6	71.2	24.9	33.6	11.6
UMGF [Zhang *et al.*, 2021]	Text+Image	73.0	39.1	18.6	10.9	73.1	26.9	34.3	12.6
MEGA [Zheng *et al.*, 2021a]	Text+Image	72.5	38.7	18.9	11.2	72.3	24.6	33.8	11.5
MKGformer [Chen *et al.*, 2022b]	Text+Image	77.6	38.4	19.0	11.8	74.5	25.2	34.0	11.3
M-[EWC]	Text+Image	74.5	39.9	20.9	13.9	74.5	28.3	35.1	13.4
M-[R-Replay]	Text+Image	74.5	37.6	22.3	15.7	74.5	28.6	42.1	19.5
M-[R-Replay & EWC]	Text+Image	74.5	41.5	24.6	21.0	74.5	28.6	40.7	17.0
Joint Training	Text+Image	77.6	71.2	73.7	69.1	79.8	76.2	71.7	69.1
MSPT	Text+Image	**79.8**	**66.4**	**63.2**	**48.2**	**77.8**	**62.6**	**64.8**	**62.3**

Table 2: Forgetting Metric $\mathbf{A}_k$ (F1 score (%)) on IMNER benchmark with two different order perturbations.

Module MI	AD	GM	MM	IMNER PER →	LOC →	ORG →	MISC	AVG
	✓	✓	✓	76.9	**71.1**	54.6	41.9	61.1
✓		✓	✓	77.8	42.0	64.7	41.2	56.4
✓	✓		✓	69.1	35.8	54.8	18.6	44.6
✓	✓	✓		77.8	27.5	36.1	11.9	38.3
✓	✓	✓	✓	**79.5**	62.6	**64.8**	**62.3**	**66.8**

Table 3: Ablation Study. "MI": Multimodal Interaction; "AD": Attention Distillation; "GM": Gradient Modulation; "MM": Memory.

5.2 Compared Baselines

We benchmark our MSPT against SOTA multimodal baselines to demonstrate its effectiveness: *1)* UMT [Yu *et al.*, 2020]; *2)* UMGF [Zhang *et al.*, 2021]; *3)* MEGA [Zheng *et al.*, 2021a]; *4)* MKGformer. Apart from previous multimodal approaches, we also compare MSPT with typical continual learning methods for a fair comparison as follows: *1)* **Vanilla** fine-tunes a BERT model on new task data without memory, acting as a **lower bound** for catastrophic forgetting. *2)* **Joint Training** retains all data in memory, retraining the MKGformer for each task, establishing an upper-performance limit. *3)* **EWC** constrains critical parameter shifts to preserve performance on prior tasks. *4)* **EMR** combines new task data with a memory of key past samples for incremental learning. *5)* **EMAR-BERT** employs reconsolidation and activation techniques to address catastrophic forgetting. *6)* **RP-CRE** represents the forefront in continual relation extraction, using stored relation samples to refine prototypes. *7)* **ExtendNER** applies KD, leveraging an existing NER model to guide the learning of a subsequent model.

5.3 Performance on IMRE Benchmark

Experiments on the IMRE benchmark (Table 1) yield several insights: *(1)* Fine-tuning unimodal BERT (Vanilla approach) with new examples leads to performance degradation due to overfitting and catastrophic forgetting. Surprisingly, multimodal models, expected to outperform Vanilla, delivered inferior results, emphasizing the need for research in continual multimodal learning. *(2)* Our method MSPT outperforms all existing MKGC models. While other continual learning approaches, utilize memory modules and sampling strategies to reduce forgetting, they are outstripped by MSPT in the 10-split IMRE benchmark, highlighting our method's effective use of multimodal interactions.

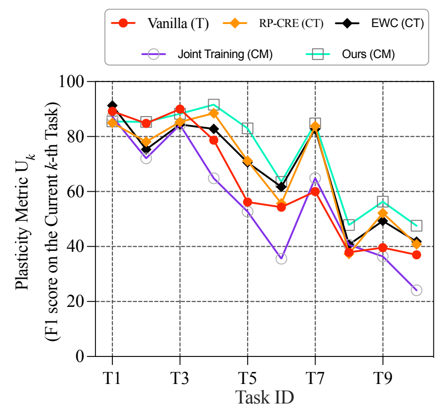


Figure 3: Performance in plasticity on the IMRE Benchmark.

5.4 Performance on IMNER Benchmark

In this section, we thoroughly compare MSPT with baseline methods across two task orders, detailed in Table 2. The insights are as follows: **(1) Overall performance:** Despite variations in MKGC model performance, these models generally lag behind unimodal BERT in continual MNER tasks, highlighting unresolved challenges in multimodal continual learning. Yet, MSPT significantly outshines all competing methods on the IMNER benchmark, demonstrating its robustness and ability to overcome the limitations of previous MKGC approaches in continual settings. **(2) Task order robustness:** To test MSPT's robustness and order independence, we evaluate it on two entity-type permutations: "PER → ORG → LOC → MISC" and "PER → LOC → ORG → MISC". MSPT consistently tops baselines across permutations, indicating it is not bound to a particular order and can generalize effectively. This across-the-board superiority on the IMNER benchmark confirms the method's effectiveness. **(3)** The "M-[]" series methods surpass both RP-CRE and our MSPT but do not reach the performance levels of SOTA unimodal continual RE methods, suggesting that simple transfer-based strategies are inadequate for optimal performance.

5.5 Ablation Study and Analysis

Effect of Each Component. Table 3 reveals that each component generally enhances model performance. Specifically,

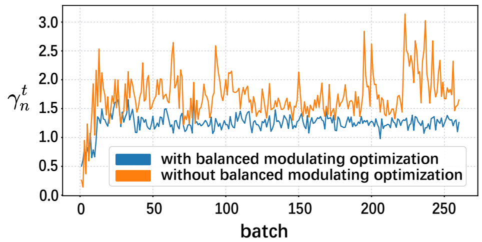


Figure 4: Change of contribution ratio γ_n^t during training.

the "MM" strategy boosts the average forgetting metric by **28.5%**, resonating with evidence that rehearsal is effective for continual KGC. "GM" leads to a **22.2%** increase in F1 score, suggesting the necessity of balancing learning rates across modalities to reduce forgetting. "AD" yields a **10.4%** F1 score improvement, indicating that preserving attention patterns aids in retaining prior knowledge. "MI" shows a **5.7%** F1 score gain, confirming its crucial role in consistent learning. Notably, omitting "MI" resulted in a temporary performance spike on the second task, potentially due to the self-attention mechanism's efficacy in short-term learning, enhanced by attention distillation. However, our findings suggest this approach is less suitable for longer task sequences. The performance declines across all tasks when the other components are removed, further validating the effectiveness of each proposed element.

Plasticity Assessment. Our evaluation of model plasticity, depicted in Figure 3, indicates that MSPT surpasses other models employing continual learning strategies like RP-CRE and EWC. We found that Joint Training exhibits the lowest plasticity due to its reliance on replaying all previous tasks' data, which hampers the model's ability to adapt to new tasks. The results highlight the superior plasticity of MSPT, which outperforms other continual learning approaches and competes with leading multimodal methods. Through attention distillation, MSPT strikes a balance between maintaining past knowledge and adapting to new information, thereby mitigating catastrophic forgetting effectively.

Imbalance Modulation Analysis. Our evaluation investigates our method's capability to mitigate training imbalances by monitoring the discrepancy ratio γn^t, which reflects inter-modality disparity. Figure 4 demonstrates that our method successfully minimizes γn^t, indicating its effectiveness in rectifying the common issue of modality imbalance in datasets. Through nuanced modulation, our approach ensures equitable learning across modalities, promoting a balanced contribution to the learning process.

Model Dependence on Rehearsal Size. The performance of rehearsal-based continual MKGC models is inherently linked to the rehearsal size, which governs the volume of training samples preserved. We assessed our model's robustness by evaluating its performance under varying rehearsal sizes. Our MSPT model consistently outperforms competing methods on the IMRE benchmark, regardless of the allocated rehearsal size, as depicted in Figure 5. This steadfastness highlights our method's capability to maintain performance even when faced with constraints on rehearsal size. Remarkably, the superiority of our model becomes more apparent with smaller rehearsal sizes, showcasing its effective utilization of limited memory resources.

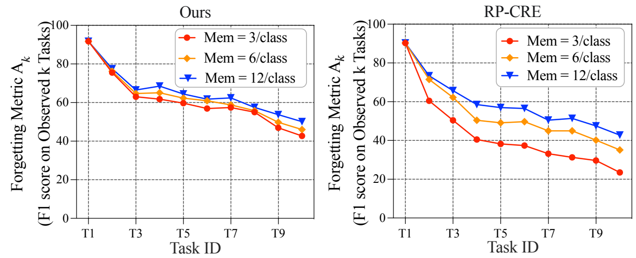


Figure 5: Analysis on rehearsal size.

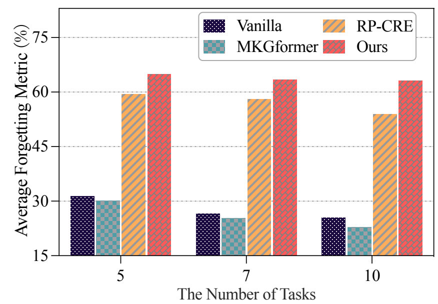


Figure 6: Sensitive Analysis on Task Numbers.

Sensitive Analysis on Task Numbers. We assess how the number of tasks impacts the MSPT model's performance, using the IMRE benchmark with 5, 7, and 10 tasks. All methods, including RP-CRE, MKGformer, and Vanilla, were tested under uniform experimental conditions: identical random seeds, hyperparameters, and task sequences. MSPT demonstrates superiority over RP-CRE and other baselines for all task quantities, showcasing consistent performance regardless of the number of tasks. This consistency confirms the robustness and adaptability of MSPT for continual MRE.

6 Conclusion and Future Work

Our study introduces the novel concept of continual MKGC, addressing the critical and practical challenge of continuously recognizing new entity categories and relations within a knowledge graph. We present a benchmark for MKGC and propose a unique approach named MSPT, which adeptly combats the dual challenges of catastrophic forgetting and plasticity, central issues in continual learning. MSPT employs a harmonized multimodal training approach to improve the detection of novel patterns, alongside a synergistic multimodal interaction with attention distillation to effectively retain previous knowledge. Comprehensive experiments and analysis demonstrate the superiority of MSPT over existing techniques in the context of continual learning. Future work will aim to expand our approach to a broader range of MKGC and investigate rehearsal-free strategies for continual MKGC.

Acknowledgments

We would like to express gratitude to the anonymous reviewers for kind comments. This work was supported by the National Natural Science Foundation of China (No. 62206246), the Fundamental Research Funds for the Central Universities (226-2023-00138), Zhejiang Provincial Natural Science Foundation of China (No. LGG22F030011), Yongjiang Talent Introduction Programme (2021A-156-G), CCF-Baidu Open Fund, and Information Technology Center and State Key Lab of CAD&CG, Zhejiang University.

References

[Arshad *et al.*, 2019] Omer Arshad, Ignazio Gallo, Shah Nawaz, and Alessandro Calefati. Aiding intra-text representations with visual context for multimodal named entity recognition. In *ICDAR 2019*, pages 337–342. IEEE, 2019.

[Chaudhry *et al.*, 2019] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with A-GEM. In *ICLR 2019*. OpenReview.net, 2019.

[Chen and et al, 2022] Xiang Chen and et al. Good visual guidance make a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. In *Findings of NAACL 2022*, pages 1607–1618, July 2022.

[Chen *et al.*, 2022a] Xiang Chen, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, Huajun Chen, and Ningyu Zhang. Lightner: A lightweight tuning paradigm for low-resource NER via pluggable prompting. In *COLING 2022*, pages 2374–2387. International Committee on Computational Linguistics, 2022.

[Chen *et al.*, 2022b] Xiang Chen, Ningyu Zhang, and et al. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *SIGIR*, SIGIR '22, page 904–915, New York, NY, USA, 2022.

[Clark *et al.*, 2019] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of bert's attention. In Tal Linzen, Grzegorz Chrupala, Yonatan Belinkov, and Dieuwke Hupkes, editors, *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 276–286. Association for Computational Linguistics, 2019.

[Cui *et al.*, 2021] Li Cui, Deqing Yang, Jiaxin Yu, Chengwei Hu, Jiayang Cheng, Jingjie Yi, and Yanghua Xiao. Refining sample embeddings with relation prototypes to enhance continual relation extraction. In *ACL*, pages 232–243, Online, August 2021.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, and Alexander Kolesnikov. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR 2021*. OpenReview.net, 2021.

[Douillard *et al.*, 2020] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *ECCV 2020*, volume 12365, pages 86–102. Springer, 2020.

[Han *et al.*, 2020] Xu Han, Yi Dai, Tianyu Gao, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Continual relation learning via episodic memory activation and reconsolidation. In *ACL*, pages 6429–6440, Online, July 2020.

[Hu *et al.*, 2023] Xuming Hu, Junzhe Chen, Aiwei Liu, Shiao Meng, Lijie Wen, and Philip S. Yu. Prompt me up: Unleashing the power of alignments for multimodal entity and relation extraction. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 5185–5194. ACM, 2023.

[Kirkpatrick *et al.*, 2016] James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796, 2016.

[Lample *et al.*, 2016] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 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, 2016. Association for Computational Linguistics.

[Liang *et al.*, 2023] Ke Liang, Lingyuan Meng, and et al. Learn from relational correlations and periodic events for temporal knowledge graph reasoning. In *SIGIR*, page 1559–1568, New York, NY, USA, 2023.

[Liang *et al.*, 2024] Ke Liang, Yue Liu, Sihang Zhou, Wenxuan Tu, Yi Wen, Xihong Yang, Xiangjun Dong, and Xinwang Liu. Knowledge graph contrastive learning based on relation-symmetrical structure. *IEEE Trans. Knowl. Data Eng.*, 36(1):226–238, 2024.

[Liu and et al, 2019] Ye Liu and Hui Li et al. MMKG: multimodal knowledge graphs. In *ESWC 2019*, volume 11503, pages 459–474. Springer, 2019.

[Liu *et al.*, 2023] Meng Liu, Ke Liang, Dayu Hu, Hao Yu, Yue Liu, Lingyuan Meng, Wenxuan Tu, Sihang Zhou, and Xinwang Liu. Tmac: Temporal multi-modal graph learning for acoustic event classification. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 3365–3374, New York, NY, USA, 2023.

[Lu and et al, 2018] Di Lu and et al. Visual attention model for name tagging in multimodal social media. In *ACL*, pages 1990–1999, Melbourne, Australia, 2018.

[Lu *et al.*, 2022] Junyu Lu, Dixiang Zhang, Jiaxing Zhang, and Pingjian Zhang. Flat multi-modal interaction transformer for named entity recognition. In *COLING 2022*, pages 2055–2064. International Committee on Computational Linguistics, 2022.

[Masana *et al.*, 2023] Marc Masana, Xialei Liu, Bartlomiej Twardowski, Mikel Menta, Andrew D. Bagdanov, and Joost van de Weijer. Class-incremental learning: Survey and performance evaluation on image classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(5):5513–5533, 2023.

[Monaikul *et al.*, 2021] Natawut Monaikul, Giuseppe Castellucci, Simone Filice, and Oleg Rokhlenko. Continual learning for named entity recognition. In *AAAI 2021*, pages 13570–13577. AAAI Press, 2021.

[Moon *et al.*, 2018] Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. Multimodal named entity recognition for short social media posts. In *NAACL (Long Papers)*, pages 852–860, New Orleans, Louisiana, 2018.

[Nair and Hinton, 2010] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *(ICML-10)*, pages 807–814. Omnipress, 2010.

[Pelosin *et al.*, 2022] Francesco Pelosin, Saurav Jha, Andrea Torsello, Bogdan Raducanu, and Joost van de Weijer. Towards exemplar-free continual learning in vision transformers: An account of attention, functional and weight regularization. In *(CVPR) Workshops*, pages 3820–3829, June 2022.

[Peng and et al, 2022] Xiaokang Peng and et al. Balanced multimodal learning via on-the-fly gradient modulation. In *CVPR*, 2022.

[Rusu *et al.*, 2016] Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *CoRR*, abs/1606.04671, 2016.

[Sprechmann *et al.*, 2018] Pablo Sprechmann, Siddhant M. Jayakumar, Jack W. Rae, Alexander Pritzel, Adrià Puigdomènech Badia, Benigno Uria, Oriol Vinyals, Demis Hassabis, Razvan Pascanu, and Charles Blundell. Memory-based parameter adaptation. In *ICLR 2018*. OpenReview.net, 2018.

[Vaswani and et al, 2017] Ashish Vaswani and et al. Attention is all you need. In *Neurips*, pages 5998–6008, 2017.

[Wan and et al, 2021] Hai Wan and et al. FL-MSRE: A few-shot learning based approach to multimodal social relation extraction. In *AAAI 2021*, 2021.

[Wang *et al.*, 2019] Hong Wang, Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Shiyu Chang, and William Yang Wang. Sentence embedding alignment for lifelong relation extraction. In *NAACL (Long and Short Papers)*, pages 796–806, Minneapolis, Minnesota, June 2019.

[Wang *et al.*, 2020] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR 2020*, 2020.

[Wang *et al.*, 2022a] Rui Wang, Tong Yu, Handong Zhao, Sungchul Kim, Subrata Mitra, Ruiyi Zhang, and Ricardo Henao. Few-shot class-incremental learning for named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 571–582, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[Wang *et al.*, 2022b] Xinyu Wang, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang, and Kewei Tu. ITA: image-text alignments for multimodal named entity recognition. In *NAACL 2022*, pages 3176–3189, 2022.

[Xia *et al.*, 2022] Yu Xia, Quan Wang, Yajuan Lyu, Yong Zhu, Wenhao Wu, Sujian Li, and Dai Dai. Learn and review: Enhancing continual named entity recognition via reviewing synthetic samples. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2291–2300, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[Yu *et al.*, 2020] Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *ACL*, pages 3342–3352, Online, 2020. Association for Computational Linguistics.

[Zenke *et al.*, 2017] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *ICML 2017*, volume 70, pages 3987–3995. PMLR, 2017.

[Zhang *et al.*, 2018] Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. Adaptive co-attention network for named entity recognition in tweets. In *(AAAI-18)*, pages 5674–5681. AAAI Press, 2018.

[Zhang *et al.*, 2021] Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *AAAI 2021,*, pages 14347–14355. AAAI Press, 2021.

[Zhang *et al.*, 2023] Ningyu Zhang, Lei Li, Xiang Chen, Xiaozhuan Liang, Shumin Deng, and Huajun Chen. Multimodal analogical reasoning over knowledge graphs. In *ICLR*, 2023.

[Zheng *et al.*, 2021a] Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li, and Tao Wang. Multimodal relation extraction with efficient graph alignment. In *ACM Multimedia 2021*, pages 5298–5306. ACM, 2021.

[Zheng *et al.*, 2021b] Changmeng Zheng, Zhiwei Wu, Junhao Feng, Ze Fu, and Yi Cai. Mnre: A challenge multimodal dataset for neural relation extraction with visual evidence in social media posts. In *ICME*, pages 1–6, 2021.

[Zheng *et al.*, 2023] Shangfei Zheng, Weiqing Wang, Jianfeng Qu, Hongzhi Yin, Wei Chen, and Lei Zhao. MMKGR: multi-hop multi-modal knowledge graph reasoning. In *ICDE 2023*, pages 96–109. IEEE, 2023.

[Zhu *et al.*, 2022] Rui Zhu, Yiming Zhao, Wei Qu, Zhongyi Liu, and Chenliang Li. Cross-domain product search with knowledge graph. In Mohammad Al Hasan and Li Xiong, editors, *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, 2022.