Collaborative Multi-LoRA Experts with Achievement-based Multi-Tasks Loss for Unified Multimodal Information Extraction

Li Yuan[1,2], **Yi Cai**[2,1*], **Xudong Shen**[1], **Qing Li**[3], **Qingbao Huang**[4], **Zikun Deng**[1,2], **Tao Wang**[5]
[1]School of Software Engineering, South China University of Technology, Guangzhou, China
[2]Key Laboratory of Big Data and Intelligent Robot (SCUT), MOE of China
[3]Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[4]School of Electrical Engineering, Guangxi University, Nanning, China
[5]Department of Biostatistics & Health Informatics, King's College London, London, United Kingdom
{seyuanli, se3048xdshen}@mail.scut.edu.cn, {ycai, zkdeng}@scut.edu.cn, csqli@comp.polyu.edu.hk, qbhuang@gxu.edu.cn, tao.wang@kcl.ac.uk

Abstract

Multimodal Information Extraction (MIE) has gained attention for extracting structured information from multimedia sources. Traditional methods tackle MIE tasks separately, missing opportunities to share knowledge across tasks. Recent approaches unify these tasks into a generation problem using instruction-based T5 models with visual adaptors, optimized through full-parameter fine-tuning. However, this method is computationally intensive, and multi-task fine-tuning often faces gradient conflicts, limiting performance. To address these challenges, we propose collaborative multi-LoRA experts with achievement-based multi-task loss (C-LoRAE) for MIE tasks. C-LoRAE extends the low-rank adaptation (LoRA) method by incorporating a universal expert to learn shared multimodal knowledge from cross-MIE tasks and task-specific experts to learn specialized instructional task features. This configuration enhances the model's generalization ability across multiple tasks while maintaining the independence of various instruction tasks and mitigating gradient conflicts. Additionally, we propose an achievement-based multi-task loss to balance training progress across tasks, addressing the imbalance caused by varying numbers of training samples in MIE tasks. Experimental results on seven benchmark datasets across three key MIE tasks demonstrate that C-LoRAE achieves superior overall performance compared to traditional fine-tuning methods and LoRA methods while utilizing a comparable number of training parameters to LoRA.

1 Introduction

Recently, growing attention has been paid to multimodal information extraction (MIE) [Zhang *et al.*, 2018; Li *et al.*, 2020; Zheng *et al.*, 2021b], which focuses on extracting structured information from multi-media sources, including text and images. Compared with text information extraction [Zhou *et al.*, 2022], MIE is generally a more challenging task as it requires bridging cross-modal information. MIE commonly consists of three key sub-tasks: multimodal named entity recognition (MNER) [Wang *et al.*, 2022b; Jia *et al.*, 2023], multimodal relation extraction (MRE) [Zheng *et al.*, 2021b; Yue *et al.*, 2023], and multimodal event extraction (MEE) [Li *et al.*, 2020; Tong *et al.*, 2020; Liu *et al.*, 2022], as shown in Figure 1. Traditional methods only concentrate on individual tasks and employ task-specific models trained with supervised learning [Zheng *et al.*, 2021b; Li *et al.*, 2020]. However, designing, training, and maintaining individual models for each task is both time-consuming and resource-intensive. Besides, this approach fails to efficiently leverage shared knowledge across different MIE tasks, hindering their performance and impeding the large-scale deployment of applications.

Figure 1: Three Key MIE Tasks, where O_1,O_2, and O_3 are visual objects.

To address the various MIE tasks in a unified way to share Information across different MIE tasks, [Sun *et al.*, 2024] proposed an approach called the Unified Multimodal Information Extractor (UMIE). This approach converts the training objectives of different MIE tasks into generation problems and utilizes a T5 model with a visual adaptor to generate outputs for various MIE tasks by full-parameter fine-tuning (FPFT) with multi-task instruction learning. However, this method has two key issues that limit its practical applica-

* Corresponding author.

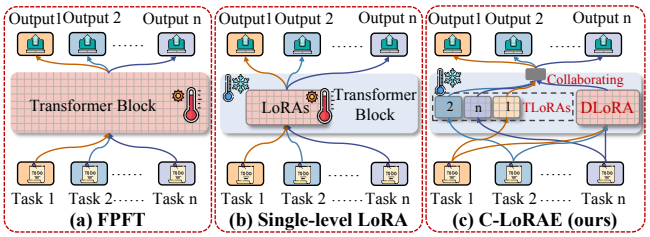

Figure 2: Different fine-tuning methods for MIE.

tion and performance. (1) FPFT necessitates significant training time and hardware resources (e.g., GPUs) [Li and Liang, 2021], as illustrated in Figure 2(a). This approach challenges adapting to larger language models with limited hardware. (2) Multi-task instruction learning often aggregates diverse task data indiscriminately, leading to suboptimal performance [Dai *et al.*, 2024; Chen *et al.*, 2024b], known as *negative transfer* [2021; 2023]. Variations in instruction formats and knowledge domains across tasks can cause gradient conflicts during simultaneous training.

An effective approach to address the first problem caused by full-parameter fine-tuning involves utilizing parameter-efficient fine-tuning methods (PEFT) [Li and Liang, 2021], such as LoRA [2021]. By freezing the pre-trained model weights, LoRA reduces the number of trainable parameters in the Transformer. Specifically, it involves training only a pair of low-rank decomposition weight matrices for each linear layer. This strategy has been widely adopted in fine-tuning large language models (LLMs) and large vision-language models (LVLMs) [Liu *et al.*, 2023a; Zhang *et al.*, 2025b]. Recent studies [Wu *et al.*, 2024; Dou *et al.*, 2024] have integrated the Mixture of Experts (MoE) [2012] concept into LoRA, using multiple LoRA modules as distinct experts to improve the model's capacity to apply world knowledge for downstream tasks, as shown in Figure 2(b). However, the limitation of *negative transfer* persists, as single-level LoRA-based approaches typically mix all instruction tasks without addressing potential instruction conflicts, thereby limiting the model's performance.

To facilitate knowledge sharing among MIE tasks while minimizing conflicts arising from integrating diverse instructional tasks within a lower training budget, this study proposes Collaborative Multi-LoRA Experts (C-LORAE) with an achievement-based multi-task loss for MIE tasks. As illustrated in Figure 3, we extend the vanilla LoRA in three ways. First, inspired by the MOE framework [Jacobs *et al.*, 1991], we introduce a universal LoRA expert along with a set of task-specific experts. This approach assimilates shared multi-model information across various instruction tasks while preserving task-specific knowledge. This configuration enhances the model's generalization ability across multiple tasks while reducing multi-task gradient conflicts. Second, we utilize mutual information maximization to facilitate the effective exchange of information between task-specific experts and the universal expert. Additionally, to enable autonomous and efficient collaboration between the task-specific experts and the universal expert, we propose an expert-motivated gate router that determines whether the features of each token should be output by the universal expert or a task-specific expert. This mechanism allows the model to leverage both generalization and specialization simultaneously.

Moreover, since the number of training samples for individual MIE tasks varies significantly, a multi-task model can easily become biased toward the task with the larger number of training samples [Yun and Cho, 2023], leading to suboptimal overall performance. Therefore, we use achievement-based multi-task loss, which is derived from the current performance across different tasks, to dynamically balance the training progress across instruction tasks and task-specific experts. Our main contributions can be summarized as follows:

- We propose a C-LoRAE method to address MIE tasks, consisting of a universal LoRA expert and a set of task-specific LoRA experts. This approach facilitates knowledge sharing across MIE tasks while resolving gradient conflicts that arise when instruction fine-tuning a large language-vision model on diverse MIE datasets.
- We propose an achievement-based multi-task loss to address the training imbalance caused by varying numbers of training samples in MIE tasks. This loss function balances the training progress of our proposed C-LoRAE model by considering the current performance across different MIE instruction tasks.
- We conducted experiments across seven datasets from MIE tasks. The results demonstrate that our proposed model, C-LoRAE, which utilizes significantly fewer training parameters compared to full fine-tuning, achieves superior overall performance across most tasks and establishes state-of-the-art performance on four datasets. Moreover, C-LoRAE consistently outperforms fine-tuning with vanilla LoRA across all datasets.

2 Method

In this section, we first introduce the collaborative multi-LoRA experts to replace the vanilla linear layers in LVLMs, as depicted in Figure 3. Then, we propose an achievement-based multi-task loss to address the training imbalance caused by varying numbers of training samples in MIE tasks.

2.1 Problem Formulation

As illustrated in Figure 3, a Transformer-based LVLM, such as UMIER [Sun *et al.*, 2024] and LLaVA [Liu *et al.*, 2023a], can be structured to address MIE in a unified way as follows:

$$T_k^{ans} = f_{\text{LVLM}}(f_{vis}(I_k), f_{emb}(T_k^{in})) \quad (1)$$

where f_{emb} denotes the transformation of input text T^{in} into a word embedding matrix, and f_{vis} represents the visual encoder with a visual adapter to convert an image into a sequence of visual embeddings. The triplets $(T_k^{in}, I_k, T_k^{ans})_{k=1}^{K}$ represent the training data for the k-th training sample among all K samples.

2.2 Preliminary

Low-rank Adaptation (LoRA) is an efficient method for fine-tuning LLMs and LVLMs. Given that our proposed approach builds on the foundational principles of LoRA, it is crucial to provide an overview of this method. Formally, a

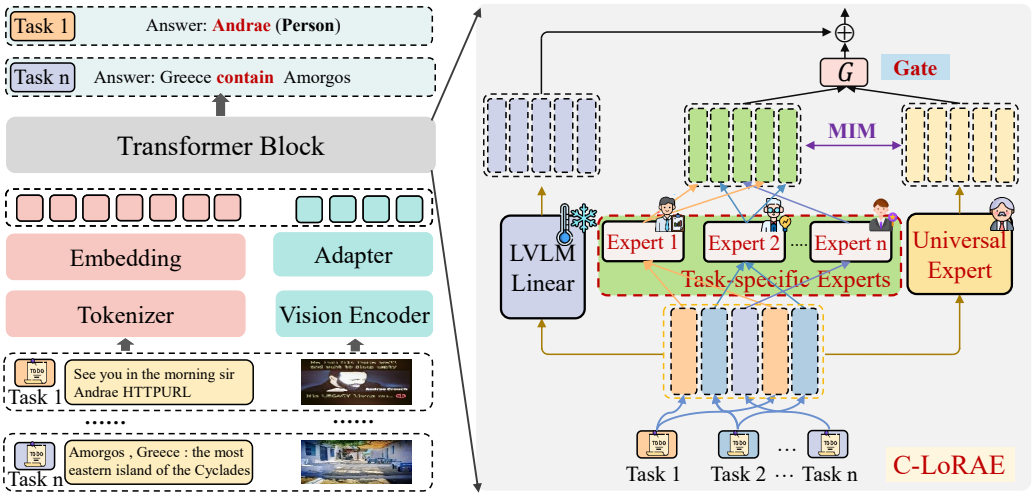

Figure 3: Overall framework of our Collaborative Multi-LoRA Experts. Our model is based on the General Large Visual-Language Model. The input text is tokenized and embedded and then concatenated with the visual input to feed into the LLMs.

linear layer $h = W_0 x$ with input $x \in \mathbb{R}^{d_{in}}$ and weight matrix $W_0 \in \mathbb{R}^{d_{out} \times d_{in}}$. In this setup, the LoRA module comprises an original linear layer from LVLMs with a LoRA layer decomposition, which is defined as follows:

$$h = LoRA(x) = W_0 x + BAx \tag{2}$$

where $A \in \mathbb{R}^{r \times d_{in}}$ and $B \in \mathbb{R}^{d_{out} \times r}$ are the low rank matrices. Since $r \ll \min(d_{out}, d_{in})$, the number of trainable parameters in A and B is significantly smaller than that in W_0. During the training of a LoRA module, only the matrices A and B are updated.

2.3 Collaborative Multi-LoRA Experts

Our method aims to facilitate knowledge sharing among MIE tasks while reducing gradient conflicts arising from mixing diverse instruction tasks. To achieve this, we propose collaborative multi-LoRA experts: a universal expert dedicated to learning shared knowledge information from all tasks, alongside a set of task-specific LoRA experts designed for learning task-specific features within individual tasks. Moreover, we employ mutual information maximization to facilitate the exchange of information between task-specific experts and the universal expert. Finally, we propose an experts-motivated gate router to obtain final token representations, which are customized for each token from different tasks. This router also helps to determine whether to utilize the universal LoRA expert or the task-specific LoRA for each token.

Universal LoRA Expert

To maintain knowledge exchange among MIE tasks and enhance the model's generalization ability by training on extensive instructions [Wei *et al.*, 2022], we propose a universal LoRA expert (ULoRA). ULoRA adopts the vanilla LoRA framework and is capable of learning universal multimodal representations across various tasks. Formally, given the *i*-th training sample in *n* instruction task x_i^n, the output of the universal expert U_i^n is defined as:

$$U_i^n = ULoRA(x_i^n) \tag{3}$$

where the trainable matrices from the universal expert are denoted as $A^U \in \mathbb{R}^{r \times d_{in}}$ and $B^U \in \mathbb{R}^{d_{out} \times r}$. Each element $x_{i,j}^n$ in x_i^n represents the *j*-th token of the sample.

Task-specific LoRA Experts

To address the issue of *negative transfer* caused by gradient conflicts across MIE tasks, we introduce a set of task-specific experts that are optimized solely by their respective task datasets. Specifically, for each instruction task, we designate a task-specific expert, resulting in *N*=3 (MNER, MRE, and MEE) task-specific experts in the MIE task framework. Assuming that the *i*-th training sample belongs to the instruction task *n*, the task-specific LoRA expert $T^n LoRA$ outputs D_i^n of x_i^n are as follows:

$$D_i^n = T^n LoRA(x_i^n) \tag{4}$$

Considering that each task-specific LoRA expert fine-tunes solely on its dataset and thus has less data for learning compared to the universal LoRA expert, we define the trainable matrices for the *n*-th task-specific LoRA expert as $A_n^D \in \mathbb{R}^{\frac{r}{N} \times d_{in}}$ and $B_n^D \in \mathbb{R}^{d_{out} \times \frac{r}{N}}$. This approach aims to minimize the number of training parameters while ensuring that the model retains the capability to learn task-specific features.

Collaborating of Experts

(1) Mutual Information Maximization To harmonize task-specific experts and the universal expert, we regard the universal expert as a teacher who imparts essential universal knowledge to students in various fields, enhancing their ability to model the bridge between text and image [Ahn *et al.*, 2019]. We employed the variational information maximization method to maximize the mutual information (MIM) between universal expert representations U_i^n and task-specific expert representations D_i^n, which is defined as:

$$\mathcal{L}_{\text{MIM}} = \sum_{n \in N} \sum_{i \in |n|} \text{MIM}(D_i^n, U_i^n) \tag{5}$$

where $|n|$ denotes the number of samples in task n, and $\mathcal{L}_{\text{MIM}}$ the loss function for optimizing mutual information. MIM($\cdot$)

Task	Dataset	Train	Dev	Test
MNER	Twitter-15	4,000	1,000	3,257
	Twitter-17	2,848	723	723
	SNAP	3,971	1,432	1,459
MRE	MRE-V1	7,824	975	1,282
	MRE-V2	12,247	923	832
MEE	ACE2005	4424	224	-
	SWiG	16451	3350	-
	S2E2	-	-	309

Table 1: The statistics of six MIE datasets.

refers to the specific MIM method used in the optimization.

(2) Experts-motivated Gate Router To tailor the contributions of the task-specific expert representation D_i^n and the universal expert representation U_i^n across various tasks, we introduce an expert-motivated gate router. For each input token $x_{i,j}^n$, the router $\mathcal{G}$ learns to activate the most suitable weights from the universal expert and the *n*-th task expert. The calculation of the gate router is defined as follows:

$$\mathcal{G}_{i,j}^n = \text{softmax}(W_g(x_{i,j}^n)) \tag{6}$$

where $x_{i,j}^n$ represent the *j*-th token of the input sequence x_i^n and $G_{i,j}^n = \{g_{i,j}^{n,1}; g_{i,j}^{n,2}\}$ denotes the gate weights for the universal expert and *n*-th task-specific expert. Additionally, $W_g \in \mathbb{R}^{2\times d_{out}}$ is shared among all the experts, and represents the linear gate weights. Thus, the output of the collaborative multi-LoRA experts $h_{i,j}^n$ is defined as follows:

$$h_{i,j}^n = g_{i,j}^{n,1} * U_{i,j}^n + g_{i,j}^{n,2} * D_{i,j}^n \tag{7}$$

We use a weighted sum of the collaborative multi-LoRA experts, $h_{i,j}^n$, and the original linear layer on input $x_{i,j}^n$ to compute final output $\widetilde{y}_{i,j}^n$, which can be expressed as follows:

$$\widetilde{y}_{i,j}^n = W_0(x_{i,j}^n) + \frac{\alpha}{r} h_{i,j}^n \tag{8}$$

where α is a hyper-parameter to facilitate tuning trainable low-rank matrices *r*.

2.4 Achievement-based Multi-task Loss

Since the numbers of training samples for individual MIE tasks can significantly differ, the multi-task models can be easily biased toward the dominant task, which has larger numbers of training samples [Yun and Cho, 2023]. Thus, to balance training progress across instruction tasks in C-LoRAE, we propose an achievement-based multi-task loss. This approach, originally inspired by focal loss, was introduced to address the class imbalance in object detection [Lin *et al.*, 2017]. Formally, we define the achievement p_c of each task as the ratio of the current and single-task metric F_1 score of the state-of-the-art model. Thus, the achievement-based weight can be defined as follows:

$$w_t^m = (1 - \frac{s_t^m}{\partial \cdot p_m})^\gamma \tag{9}$$

where w_t^m and s_t^m represent the weight and F_1 score of the m-th dataset at the t-th training epoch, respectively. The P_m denotes the SoTA results of the m-th dataset. However, task weights can unintentionally increase if the F_1 score of the multi-task model surpasses that of single-task models. To prevent this unintended increase and encourage the model to perform above the SoTA, we introduce a slight margin $\partial > 1$. Reducing task weights is analogous to theoretically decreasing the learning rate, which could lead to underfitting in the corresponding tasks. To mitigate the risk of underfitting, we normalize task weights using $softmax$.

To measure the ground truth distribution and the predicted tagging distribution of the m-th task at the t-th training epoch, we employ the cross-entropy error $\mathcal{L}_m^t$. Then, we use $\mathcal{L}_m^t$ with their corresponding weight w_t^m to compute the multi-task loss $\mathcal{L}_{MT}$. Finally, we jointly optimize the multi-task loss and the loss of MIM. Therefore, the final jointly optimized objective is computed as follows:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{MT} + \beta \cdot \mathcal{L}_{\text{MIM}} \\ \mathcal{L}_{MT} &= \sum_{t=1}^{T} \sum_{m=1}^{M} w_t^m \cdot \mathcal{L}_m^t \\ \mathcal{L}_m^t &= \sum_{(x^m, y^m) \in D^m} CE(p(\widetilde{y}^m | x^m, y^m; \theta), y^m) \end{aligned} \tag{10}$$

where β is a hyperparameter that regulates the influence on optimizing the MIM term, while θ represents the trainable parameters. D^m denotes the complete sample set of the m-th task dataset.

3 Experiments

We conducted extensive experiments on various MIE datasets to evaluate the effectiveness and efficiency of the proposed methods for MIE. Additional details and supplementary experimental results are provided in the Appendix.

3.1 Implementation Details

All models utilized the Adam optimizer with a learning rate of 1e-4 and a decay factor of 0.5. During training, experiments were conducted on four NVIDIA RTX 3090 GPUs or GPUs of equivalent performance. The maximum text input length was set to 256, and training lasted for 20 epochs. To accommodate GPU memory constraints, different batch sizes were used: 10 for FLAN-T5-base, 4 for FLAN-T5-large, and 2 for T5-XLarge on each GPU. We configured the number of task-specific LoRA experts to 3 to match the MIE instruction task number, and the influence on optimizing the MIM term β is 0.01. Detailed discussions on the low-rank dimensionality and dropout rate in our proposed C-LoRAE will be provided in subsequent sections. The implementation is publicly available at https://github.com/YuanLi95/C-LORAE.

3.2 Dataset Details

Following the experimental setup in [Sun *et al.*, 2024], we train and evaluate the proposed C-LoRAE method on MNER, MRE, and MEE tasks,

MNER: Following previous studies, we used the Twitter-15 dataset [Zhang *et al.*, 2018], the SNAP dataset [Lu *et al.*, 2018], and the Twitter-17 dataset [Yu *et al.*, 2020]. These datasets are sourced from the social media platform and are commonly employed in prior research.**MRE:** We adopt the

Model	MNER			MRE		MEE		All
	Twitter-15	Twitter-17	SNAP	MRE-V1	MRE-V2	MED	MEAE	
UMT(Yu *et al.*)	73.4	73.4	-	-	65.2	-	-	-
UMGF(Zhang *et al.*)	74.9	85.5	-	-	-	-	-	-
MEGA(Zheng *et al.*)	72.4	84.4	66.4	-	-	-	-	-
RpBERT(Sun *et al.*)	74.4	-	85.7	-	-	-	-	-
R-GCN(Zhao *et al.*)	75.0	87.1	-	-	-	-	-	-
ITA(Wang *et al.*)	**78.9**	89.8	90.2	-	66.9	-	-	-
MoRe(Wang *et al.*)	77.3	88.7	89.3	-	65.8	-	-	-
HVPNeT(Chen *et al.*)	75.3	86.8	-	81.8	81.9	-	-	-
MKGFormer(Chen *et al.*)	-	87.5	-	-	**82.0**	-	-	
MNER-QG(Jia *et al.*)	75.0	87.3	-	-	-	-	-	-
HamLearning(Liu *et al.*)	76.5	87.1	-	-	-	-	-	-
EviFusion(Liu *et al.*)	75.5	87.4	-	-	-	-	-	-
BGA-MNER(Chen and Feng)	76.3	87.7	-	-	-	-	-	-
WASE(Li *et al.*)	-	-	-	-	-	50.8	19.2	-
Unicl(Liu *et al.*)	-	-	-	-	-	57.6	23.4	-
UMIER$_{Base}$(Sun *et al.*) †	76.1	88.1	87.7	84.3	74.8	60.5	22.5	494.0
UMIER$_{Large}$	77.2	90.7	90.5	85.0	75.5	61.0	23.6	503.5
UMIER$_{XLarge}$	78.2	**91.4**	**91.0**	**86.4**	76.2	**62.1**	**24.5**	509.8
C-LoRAE$_{Base}$	74.4↓(1.9)	86.3↓(1.8)	85.8↓(1.9)	87.3↑(3.0)	81.4↑(6.6)	60.7↑(0.2)	28.4↑(5.9)	503.9↑(9.9)
C-LoRAE$_{Large}$	75.9↓(1.3)	88.9↓(1.8)	89.3↓(1.2)	89.2↑(4.3)	82.8↑(7.3)	**64.5**↑(3.5)	31.2↑(7.6)	521.3↑(17.8)
C-LoRAE$_{XLarge}$	77.8↓(0.4)	89.8↓(1.6)	90.5↓(0.5)	**89.6**↑(3.2)	**83.2**↑(6.0)	63.5↑(1.4)	**32.6**↑(8.1)	527.0↑(17.2)

Table 2: The comparative test results for three multimedia information extraction tasks are based on the F_1 score. All values represent the total score across all tasks. The best performance is highlighted in bold, while the second-best is underlined. For the baseline, we used the results from the [Sun *et al.*, 2024] report. The results that were not reported are marked with "-".

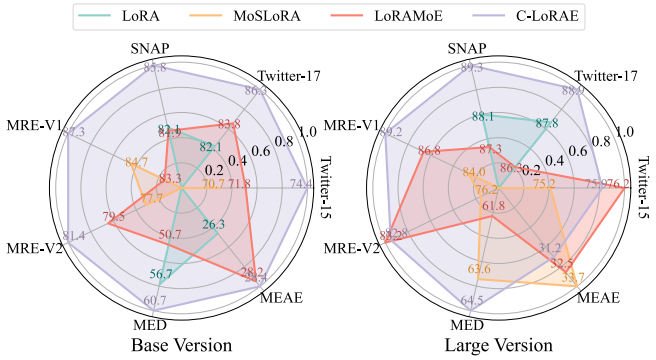

Figure 4: Normalized performance across seven tasks for LoRA, MoSLoRA, LoRAMoE, and our proposed C-LoRAE.

MRE data set [Zheng *et al.*, 2021b] constructed from social media, which is currently the only available MRE data set. This data set is divided into versions V1 and V2, with V2 being the refined version of V1. **MEE:** Follow the previous task setting [Tong *et al.*, 2020], we train our proposed method using datasets such as ACE2005 and SWiG, and subsequently evaluate its performance using the M^2E^2 dataset. We reformat all datasets into a standardized JSON format to ensure uniform representation across all tasks. Detailed statistics for each dataset are presented in Table 1.

3.3 Main Results

As depicted in Table 2, we observe that UMIER shows improvement over most previous independent task models, particularly as the scale of Flan-T5 increases, resulting in consistent performance enhancements. Although UMIER exhibits capabilities close to the SoTA method (ITA) in the MNER task, there is a significant performance gap in the MRE V2 task compared to the MRE SoTA (MKFGformer). Even with the largest model, UMIER$_{XLarge}$, there is a decrease of 5.6% in terms of F_1 score in the MRE-V2 dataset. One possible explanation is the *negative transfer* caused by fine-tuning the full-parameter model on various instruction tasks, leading the model to fail to achieve optimal performance on MRE tasks.

We replaced the vanilla linear layer of the Transformer in the UMIER model with the proposed C-LoRAE structure, denoted as C-LoRAE. Despite having a limited number of trainable parameters, our model achieves comparable, and in many cases superior, performance relative to the original UMIER model. Specifically, for the "All" metric, which aggregates F_1 scores across all datasets to measure overall performance, C-LoRAE consistently and significantly outperforms the original UMIER model. Notably, in the *Large* and *XLarge* versions, C-LoRAE demonstrates improvements of 17.8 and 17.2 points, respectively. These results suggest that the proposed method facilitates knowledge sharing across MIE tasks while reducing conflicts arising from the integration of diverse instruction data types. Besides, Our model shows substantial improvements in the MRE and MEE tasks compared to UMIER. The *XLarge* version sets a new SOTA across four datasets, indicating that the MRE and MEE tasks benefit significantly from our collaborative multi-LoRA experts approach, especially in comparison to the MNER task.

We further compared the performance of the proposed method with current mainstream efficient fine-tuning approaches, as shown in Figure 4. The results demonstrate that our method outperforms the single-stage LoRA approach across most tasks, with a particularly pronounced advantage

Model	MNER			MRE		MEE		All	TP
	Twitter-15	Twitter-17	SNAP	MRE-V1	MRE-V2	MED	MEAE		
C-LoRAE$_{Large}$	75.9	88.9	89.3	89.2	82.8	64.0	31.2	521.3	39M
UMIER$_{Large}$(*Only* ULoRA)	74.5	87.8	88.1	82.1	74.9	61.0	25.3	493.7	19M
UMIER$_{Large}$(*Only* TLoRA)	72.9	85.2	82.3	85.8	81.2	62.4	29.5	499.3	19M
C-LoRAE$_{Large}$(*w/o* EGR)	74.9	87.8	88.4	86.4	80.8	61.8	27.8	507.9	38M
C-LoRAE$_{Large}$(*w/o* AML)	74.3	87.2	86.8	87.6	81.7	63.3	29.8	510.7	39M
C-LoRAE$_{Large}$(*w/o* MIM)	75.4	88.2	88.1	87.1	82.0	63.8	30.2	514.8	39M

Table 3: An ablation study was conducted on the test F_1 scores using the large version of the model. TP denotes the number of trainable parameters during optimization, $Only$ indicates the exclusive use of a specific module, and w/o signifies the exclusion of a module.

Method	Twitter-17	MRE-V2
UMT(Yu *et al.*)	60.9	-
UMGF(Zhang *et al.*)	60.9	-
FMIT(Lu *et al.*)	64.4	-
CAT-MNER(Wang *et al.*)	64.5	-
BGA-MNERChen *et al.*	64.9	-
ChatGPT ‡	57.5	35.2
GPT-4 ‡	66.6	42.1
UMIER$_{Base}$	66.8	67.3
UMIER$_{Large}$	68.5	68.8
C-LoRAE$_{Base}$	66.2	76.5
C-LoRAE$_{Large}$	67.8	78.2

Table 4: Comparison of the generalization ability. The ChatGPT‡ and GPT-4‡ result in in-context learning setting come from [Chen and Feng, 2023] .

Methods		TP	fwd FLOPs (G)	fwd+bwd FLOPs (G)
UMIE (Base)	Full Fine-tuning	287.0	21.3	63.8
	Vanilla LoRA	15.5	22.0	65.9
	MoSLoRA	15.8	26.3	78.9
	LoRAMoE	30.5	22.7	68.2
	C-LoRA(*ours*)	30.4	23.2	69.5
UMIE (Large)	Full Fine-tuning	843.0	73.2	219.0
	Vanilla LoRA	20.3	74.1	223.4
	MoSLoRA	20.4	89.1	267.3
	LoRAMoE	40.0	75.2	225.5
	C-LoRAE(*ours*)	39.9	75.7	227.2

Table 5: The sensitivity analysis assesses the different models. TP represents the number of trainable parameters during optimization, while *fwd* and *bwd* indicate a single forward and backward process, respectively. G represents the GFLOPs.

on the T5-Base backbone. This can be attributed to the dual-stage LoRA structure in our framework, which effectively learns general knowledge while significantly reducing conflicts between tasks.

3.4 Ablation Study

To investigate the effectiveness of different components in the C-LoRAE structure, we conduct ablation studies on the universal LoRA expert (ULoRA), task-specific LoRA experts (TLoRA), mutual information maximization (MIM) module, and the experts-motivated gate router (EGR). Additionally, we evaluate the effect of the proposed achievement-based multi-task loss (AML).

The results are reported in Table 3. First, we observe that employing *only* ULoRA, which utilizes the standard LoRA for fine-tuning UMIER, results in a performance decrease of 27.6 in the ALL metric. Notably, the MRE-V1, MRE-V2, and MEE datasets exhibit significant declines compared to others. This indicates that using a single LoRA for fine-tuning all instructional tasks does not consistently improve performance, particularly when gradient conflicts among these tasks are present. Subsequently, we implemented task-specific LoRA for each instruction task (*only* TLoRA). This approach led to a notable performance decline in the MNER task compared to *only* ULoRA, primarily due to the MNER task having less available training data than the others. Additionally, ULoRA struggles to effectively leverage shared knowledge from other tasks, which limits its performance enhancement capabilities.

Thirdly, excluding EGR (*w/o* EGR) and replacing it with element-wise addition results in significant performance degradation across most sub-tasks. This suggests that the importance of expert collaboration varies among different tasks, and simple addition cannot effectively manage diverse instructional tasks. Finally, the AML strategy successfully balances performance across various instructional tasks, resulting in overall performance improvements. Moreover, while the model remains functional without the MIM module, its absence reduces overall performance.

3.5 Generalization Analysis

To investigate the generalization capability of the C-LoRAE structure, we followed the settings of previous work [Wang *et al.*, 2022c; Sun *et al.*, 2024] and evaluated its performance on the Twitter-17 and MRE-V2 datasets. Specifically, we excluded these datasets and only used the remaining training data in model training.

Table 4 presents the results. We observe that the proposed C-LoRAE maintains strong generalization ability, despite utilizing fewer trainable parameters. Considering that generalization has also been verified in MEE, these experiments suggest that the proposed structure can effectively adapt to and generalize across new datasets, even without direct training on those specific datasets. The results demonstrate that C-LoRAE not only mitigates conflicts arising from mixing diverse instruction data types through task-specific LoRA but also benefits from the Universal LoRA, which enables knowledge sharing across all datasets. Furthermore, C-LoRAE effectively learns from multimodal data sources and successfully transfers knowledge across various tasks.

3.6 Sensitivity Analysis

In this section, we analyze the sensitivity of the proposed C-LoRAE under the *Base* (rank = 128) and *Large* (rank = 64) versions. For comparison, we evaluate the vanilla UMIER, MoSLoRA, and LoRAMoE. We examined the floating-point operations (FLOP) with a batch size set to 1, which serves as a measure of computational efficiency, indicating the number of floating-point computations required for a single forward (fwd) and backward (bwd) process. We also used trainable parameters (TP) as a metric for comparison.

It can be observed that our method is similar to LoRAMoE in terms of TP and FLOPs. Although TP is slightly higher than that of MoSLoRA, the FLOPs are lower than this method. This demonstrates that the proposed method improves performance without significantly increasing training resources compared to other approaches, and it is considerably more efficient than full fine-tuning.

4 Related Work

4.1 Multimodal Information Extraction

Multimodal Information Extraction encompasses MNER, MRE, and MEE. Previous studies have treated these tasks in isolation, overlooking potential correlations between them.

MNER addresses the limitations of textual information by incorporating visual data. Early studies enhanced textual representations with visual features [Zhang *et al.*, 2018; Zhang *et al.*, 2021; Zhang *et al.*, 2025a]. To mitigate the issue of inaccurate image recognition, [Yu *et al.*, 2020] proposed a unified multimodal transformer that dynamically merges visual and textual representations through a gating mechanism. Similarly, [Sun *et al.*, 2021] introduced an auxiliary binary classification task to filter irrelevant visual content. Recent approaches, such as ITA [2022b] and MNER-QG [2023], enhance unified representations and text-focused attention mechanisms by incorporating image information (e.g., regional object tags) into inputs.

MRE focuses on extracting relationships between entities and is a newer field compared to MNER [Zheng *et al.*, 2021a; Zheng *et al.*, 2021b; Chen *et al.*, 2022b; Yuan *et al.*, 2023; Yuan *et al.*, 2024b]. Early methods employed multimodal fusion layers to align textual and visual features, thereby enhancing text representation [Chen *et al.*, 2022a; Zhao *et al.*, 2023]. However, they often lacked sufficient background information. Recent approaches address this limitation by utilizing retrieval-augmented generation to incorporate relevant images or knowledge from sources such as the Web or Wikipedia, thereby improving relationship identification [Yue *et al.*, 2023; Yuan *et al.*, 2024a; Yuan *et al.*, 2025]

MEE aims to extract events and event parameters from multiple modalities. Unlike MNER and MRE, which solely utilize visual information to enhance text extraction, MEE also extracts visual objects as parameters within events [Tong *et al.*, 2020; Li *et al.*, 2020]. To better bridge the gap between modalities, [Liu *et al.*, 2022] proposed a cross-modal contrastive learning framework to improve the similarity between their representations.

UMIE emerged more recently than unified information extraction (UIE) in text modality. [Sun *et al.*, 2024] was the first to propose a unified multimodal information extractor (UMIER), which utilizes the T5 model with a visual adapter to extract both textual and visual information effectively. However, this method utilizes full-parameter fine-tuning combined with instruction tuning, necessitating significant training time and hardware resources. This requirement poses challenges for LLMs and environments with limited hardware resources.

4.2 Parameter-Efficient Fine-Tuning

With the rise of LLMs and LVLMs [Li *et al.*, 2023; Liu *et al.*, 2023a; Chen *et al.*, 2023b; Chen *et al.*, 2024a], traditional fine-tuning becomes impractical due to the high computational demands of models with billions of parameters [Houlsby *et al.*, 2019]. Parameter-efficient fine-tuning (PEFT) methods, such as LoRA [2021], offer a solution by enabling task-specific customization with lower computational cost. LoRA, which applies low-rank updates, maintains the model's core capabilities, making it widely used for fine-tuning LLMs and LVLMs. Recent studies, such as LoRAMoE [2024] and MoSLoRA [2024], have incorporated the Mixture of Experts approach [Yuksel *et al.*, 2012] into LoRA, leveraging multiple LoRA modules as distinct experts to enhance the model's ability to utilize world knowledge for solving downstream tasks.

In contrast to vanilla LoRA, LoRAMoE, and MoSLoRA, which adopt single-level approaches, our method addresses gradient conflicts in multi-task instruction fine-tuning through a two-level LoRA framework. This framework consists of a universal LoRA module for shared knowledge and task-specific modules tailored to individual requirements, thereby minimizing conflicts. Additionally, we introduce an achievement-based multi-task loss function that dynamically adjusts based on task performance, helping to mitigate training imbalances across diverse MIE tasks for more balanced and effective training.

5 Conclusions

This study proposes a collaborative multi-LoRA expert (C-LoRAE) model that consists of a universal LoRA expert and a set of task-specific LoRA experts to model various MIE tasks jointly. The universal LoRA expert effectively shares knowledge among different instruction tasks, while the task-specific LoRA experts mitigate conflicts arising from diverse instruction data types. Additionally, we designed an achievement-based multi-task loss to address the imbalance caused by the varying numbers of training samples in MIE tasks. This loss function balances the training progress of our proposed C-LoRAE model based on current performance across different MIE instruction tasks. Experiments on seven benchmark datasets across three tasks demonstrate that C-LoRAE not only outperforms traditional full-parameter fine-tuning models and vanilla LoRA fine-tuning methods, achieving state-of-the-art results in four datasets with fewer trainable parameters, highlighting the efficiency and scalability of our method. Future work will explore further optimization of the collaboration among LoRA experts and investigate the application of our approach to other multimodal learning tasks.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (62476097, 62276072, 62402184), the Fundamental Research Funds for the Central Universities, South China University of Technology (x2rjD2240100), Guangdong Provincial Fund for Basic and Applied Basic Research—Regional Joint Fund Project (Key Project) (2023B1515120078), Guangdong Provincial Natural Science Foundation for Outstanding Youth Team Project (2024B1515040010), the Hong Kong Polytechnic University under the Postdoc Matching Fund Scheme (Project No. P0049003). Support from the Guangxi Natural Science Foundation Key Project (No. 2025GXNSFDA069017) is also gratefully acknowledged. We further acknowledge the support from the Postdoc Matching Fund Scheme of The Hong Kong Polytechnic University (Project No. P0049003). TW received funding from the NIHR Maudsley Biomedical Research Centre, Maudsley Charity, King's Together, and MHaPS Early Career Researcher Awards.

References

[Ahn *et al.*, 2019] Sungsoo Ahn, Shell Xu Hu, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the CVPR 2019*, pages 9155–9163, 2019.

[Chen and Feng, 2023] Feng Chen and Yujian Feng. Chain-of-thought prompt distillation for multimodal named entity and multimodal relation extraction, 2023.

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

[Chen *et al.*, 2022b] Xiang Chen, Ningyu Zhang, and Huajun Chen. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *Proceedings of the ACM SIGIR 2022*, pages 904–915, 2022.

[Chen *et al.*, 2023a] Feng Chen, Jiajia Liu, Kaixiang Ji, Wang Ren, Jian Wang, and Jingdong Chen. Learning implicit entity-object relations by bidirectional generative alignment for multimodal ner. In *Proceedings of the ACM MM 2023*, page 4555–4563, 2023.

[Chen *et al.*, 2023b] Jiali Chen, Zhenjun Guo, Jiayuan Xie, Yi Cai, and Qing Li. Deconfounded visual question generation with causal inference. In *Proceedings of the ACM MM 2023*, pages 5132–5142, 2023.

[Chen *et al.*, 2024a] Jiali Chen, Xusen Hei, Yuqi Xue, Yuancheng Wei, Jiayuan Xie, Yi Cai, and Qing Li. Learning to correction: Explainable feedback generation for visual commonsense reasoning distractor. In *Proceedings of ACM MM 2024*, pages 8209–8218, 2024.

[Chen *et al.*, 2024b] Shaoxiang Chen, Zequn Jie, and Lin Ma. Llava-mole: Sparse mixture of lora experts for mitigating data conflicts in instruction finetuning mllms, 2024.

[Dai *et al.*, 2024] Yanqi Dai, Dong Jing, Nanyi Fei, and Zhiwu Lu. CoTBal: Comprehensive Task Balancing for Multi-Task Visual Instruction Tuning, 2024.

[Dou *et al.*, 2024] Shihan Dou, Enyu Zhou, and Xiaoran Fan. Loramoe: Alleviating world knowledge forgetting in large language models via moe-style plugin. In *Proceedings of the ACL 2024*, pages 1932–1945, 2024.

[Houlsby *et al.*, 2019] Neil Houlsby, Andrei Giurgiu, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *Proceedings of the ICML 2019*, pages 2790–2799, 2019.

[Hu *et al.*, 2021] Edward J Hu, Phillip Wallis, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *Proceedings of the ICLR 2021*, 2021.

[Jacobs *et al.*, 1991] Robert A Jacobs, Michael I Jordan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[Jia *et al.*, 2023] Meihuizi Jia, Lei Shen, and Jiaqi Li. Mnerqg: an end-to-end mrc framework for multimodal named entity recognition with query grounding. In *Proceedings of the AAAI 2023*, pages 8032–8040, 2023.

[Li and Liang, 2021] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the ACL 2021*, pages 4582–4597, 2021.

[Li *et al.*, 2020] Manling Li, Alireza Zareian, and Shih-Fu Chang. Cross-media structured common space for multimedia event extraction. In *Proceedings of the ACL 2020*, pages 2557–2568, 2020.

[Li *et al.*, 2023] Junnan Li, Dongxu Li, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the ICML 2023*, pages 19730–19742, 2023.

[Lin *et al.*, 2017] Tsung-Yi Lin, Priya Goyal, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the ICCV 2017*, pages 2980–2988, 2017.

[Liu *et al.*, 2022] Jian Liu, Yufeng Chen, and Jinan Xu. Multimedia event extraction from news with a unified contrastive learning framework. In *Proceedings of the ACM MM 2022*, page 1945–1953, 2022.

[Liu *et al.*, 2023a] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the NIPS 2023*, 2023.

[Liu *et al.*, 2023b] Weide Liu, Xiaoyang Zhong, and Yuming Fang. Integrating large pre-trained models into multimodal named entity recognition with evidential fusion, 2023.

[Liu *et al.*, 2024] Peipei Liu, Hong Li, and Limin Sun. Hierarchical aligned multimodal learning for ner on tweet posts. In *Proceedings of the AAAI 2024*, pages 18680–18688, 2024.

[Lu *et al.*, 2018] Di Lu, Leonardo Neves, and Heng Ji. Visual attention model for name tagging in multimodal social media. In *Proceedings of the ACL 2018*, pages 1990–1999, 2018.

[Lu *et al.*, 2022] Junyu Lu, Dixiang Zhang, Jiaxing Zhang, and Pingjian Zhang. Flat multi-modal interaction transformer for named entity recognition. In *Proceedings of the COLING 2022*, pages 2055–2064, 2022.

[Sun *et al.*, 2021] Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. Rpbert: a text-image relation propagation-based bert model for multimodal ner. In *Proceedings of the AAAI 2021*, pages 13860–13868, 2021.

[Sun *et al.*, 2024] Lin Sun, Kai Zhang, and Renze Lou. UMIE : Unified Multimodal Information Extraction with Instruction Tuning. In *Proceedings of the AAAI 2024*, pages 19062–19070, 2024.

[Tong *et al.*, 2020] Meihan Tong, Shuai Wang, and Tat-Seng Chua. Image enhanced event detection in news articles. In *Proceedings of the AAAI 2020*, pages 9040–9047, 2020.

[Wang *et al.*, 2022a] Xinyu Wang, Jiong Cai, and Wei Lu. Named entity and relation extraction with multi-modal retrieval. In *Findings of the EMNLP 2022*, pages 5925–5936, 2022.

[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 multi-modal named entity recognition. In *Proceedings of the NAACL 2022*, pages 3176–3189, 2022.

[Wang *et al.*, 2022c] Xuwu Wang, Jiabo Ye, Zhixu Li, Junfeng Tian, Yong Jiang, Ming Yan, Ji Zhang, and Yanghua Xiao. Cat-mner: Multimodal named entity recognition with knowledge-refined cross-modal attention. In *Proceedings of the ICME 2022*, pages 1–6, 2022.

[Wei *et al.*, 2022] Jason Wei, Maarten Bosma, and Quoc V Le. Finetuned language models are zero-shot learners. In *Proceedings of the ICLR 2022*, 2022.

[Wu *et al.*, 2024] Taiqiang Wu, Jiahao Wang, Zhe Zhao, and Ngai Wong. Mixture-of-subspaces in low-rank adaptation. In *Proceedings of the EMNLP 2024*, pages 7880–7899, 2024.

[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 *Proceedings of the ACL 2020*, pages 3342–3352, 2020.

[Yuan *et al.*, 2023] Li Yuan, Yi Cai, Jin Wang, and Qing Li. Joint multimodal entity-relation extraction based on edge-enhanced graph alignment network and word-pair relation tagging. In *Proceedings of the AAAI 2023*, pages 11051–11059, 2023.

[Yuan *et al.*, 2024a] Li Yuan, Yi Cai, and Junsheng Huang. Few-shot joint multimodal entity-relation extraction via knowledge-enhanced cross-modal prompt model. In *Proceedings of the ACM MM 2024*, pages 8701–8710, 2024.

[Yuan *et al.*, 2024b] Li Yuan, Yi Cai, Jingyu Xu, Qing Li, and Tao Wang. A fine-grained network for joint multimodal entity-relation extraction. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

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

[Yue *et al.*, 2023] Xiang Yue, Boshi Wang, and Huan Sun. Automatic evaluation of attribution by large language models. In *Findings of the EMNLP 2023*, pages 4615–4635, 2023.

[Yuksel *et al.*, 2012] Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.

[Yun and Cho, 2023] Hayoung Yun and Hanjoo Cho. Achievement-based Training Progress Balancing for Multi-Task Learning. In *Proceedings of the ICCV 2023*, pages 16935–16944, 2023.

[Zhang and Yang, 2021] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2021.

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

[Zhang *et al.*, 2021] Dong Zhang, Suzhong Wei, and Guodong Zhou. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *Proceedings of the AAAI 2021*, pages 14347–14355, 2021.

[Zhang *et al.*, 2025a] Jian Zhang, Zhangqi Wang, and Erik Cambria. Mars: A multi-agent framework incorporating socratic guidance for automated prompt optimization. *arXiv preprint arXiv:2503.16874*, 2025.

[Zhang *et al.*, 2025b] Jian Zhang, Zhiyuan Wang, and Jun Liu. Maps: A multi-agent framework based on big seven personality and socratic guidance for multimodal scientific problem solving. *arXiv preprint arXiv:2503.16905*, 2025.

[Zhao *et al.*, 2022] Fei Zhao, Chunhui Li, and Xinyu Dai. Learning from different text-image pairs: A relation-enhanced graph convolutional network for multimodal ner. In *Proceedings of the ACM MM 2022*, page 3983–3992, 2022.

[Zhao *et al.*, 2023] Qihui Zhao, Tianhan Gao, and Nan Guo. Tsvfn: Two-stage visual fusion network for multimodal relation extraction. *Information Processing & Management*, 60(3):103264, 2023.

[Zheng *et al.*, 2021a] Changmeng Zheng, Junhao Feng, and Tao Wang. Multimodal relation extraction with efficient graph alignment. In *Proceedings of the ACM MM 2021*, pages 5298–5306, 2021.

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

[Zhou *et al.*, 2022] Shaowen Zhou, Bowen Yu, and Aixin Sun. A survey on neural open information extraction: Current status and future directions. In *proceedings of the IJCAI 2022*, 2022.