

UMIE: Unified Multimodal Information Extraction with Instruction Tuning

Lin Sun^{1*}, Kai Zhang^{2*}, Qingyuan Li^{3*}, Renze Lou⁴

¹Department of Computer Science, Hangzhou City University, China

²Department of Computer Science and Engineering, Ohio State University, USA

³College of Computer Science and Technology, Zhejiang University, China

⁴Department of Computer Science and Engineering, Pennsylvania State University, USA
sunl@hzcw.edu.cn, liqingyuan@zju.edu.cn

Abstract

Multimodal information extraction (MIE) gains significant attention as the popularity of multimedia content increases. However, current MIE methods often resort to using task-specific model structures, which results in limited generalizability across tasks and underutilizes shared knowledge across MIE tasks. To address these issues, we propose UMIE, a unified multimodal information extractor to unify three MIE tasks as a generation problem using instruction tuning, being able to effectively extract both textual and visual mentions. Extensive experiments show that our single UMIE outperforms various state-of-the-art (SoTA) methods across six MIE datasets on three tasks. Furthermore, in-depth analysis demonstrates UMIE's strong generalization in the zero-shot setting, robustness to instruction variants, and interpretability. Our research serves as an initial step towards a unified MIE model and initiates the exploration into both instruction tuning and large language models within the MIE domain. Our code, data, and model are available at <https://github.com/ZUCC-AI/UMIE>.

Introduction

Information extraction (IE), a task aiming to derive structured information from unstructured texts, plays a crucial role in the domain of natural language processing. As the popularity of multimedia content continues to increase (Zhu et al. 2022), multimodal information extraction (MIE) has drawn significant attention from the research community (Zhang et al. 2018; Chen et al. 2022; Li et al. 2022). MIE aims to deliver structured information of interest from multiple media sources such as textual, visual, and potentially more. It is considered a challenging task due to the inherent complexity of media formats and the necessity to bridge cross-modal gaps, where traditional text-based IE methods often struggle.

MIE includes multimodal named entity recognition (MNER) (Moon, Neves, and Carvalho 2018; Zhang et al. 2018; Sun et al. 2021), multimodal relation extraction (MRE) (Zheng et al. 2021b; Wang et al. 2022a), and multimodal event extraction (MEE) (Li et al. 2020, 2022). Current methods for usually focus on a specific task mentioned

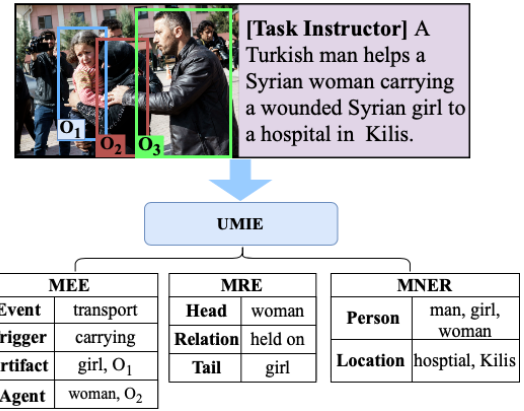


Figure 1: Unifying three key MIE tasks in a single multimodal model. Given a task instructor, UMIE performs the corresponding task by extracting textual and visual mentions (MNER and MEE) or inferring the relationship between two given mentions (MRE). O_1 , O_2 , and O_3 are visual objects.

above, which mostly uses a task-specific model structure with dataset-specific tuning for the task at hand. Such a paradigm leads to a few limitations: Firstly, it results in a lack of generalizability as they often overfit the patterns of a specific task or even dataset. Secondly, the need to design, train, and maintain a separate model for each task is time-consuming, impeding the progress of deploying practical multimodal systems at scale. Lastly, due to their independent training nature, these methods fail to effectively leverage the shared knowledge across different MIE tasks, undermining the performance of each task.

To address these challenges, in this work, we propose a unified multimodal information extractor (UMIE), a single model that unifies different MIE tasks as generation problems with instruction tuning (Ouyang et al. 2022; Wang et al. 2022d). As shown in Figure 1, given the same text and image, UMIE can perform various MIE tasks following different task instructors and generate corresponding structured outputs. In particular, UMIE enables both extraction of textual span and visual objects (e.g., MEE), which is rarely considered by previous MNER and MRE works.

Specifically, we repurpose all MIE datasets and train a

*These three authors contributed equally to this work.

UMIE model to perform each task in a generation fashion by following corresponding instructions. In addition, we design a visual encoder and a gated attention module to dynamically integrate visual clues from images for robust cross-modal feature extraction. Consequently, our singular UMIE model outperforms various SoTA methods on each of the six MIE datasets across all three tasks, demonstrating the effectiveness of our framework. Furthermore, in-depth analysis showcases strong generalization ability and robustness to instructions. Our contributions are summarized as follows:

- We present UMIE, an end-to-end model that unifies MIE tasks into a generation problem. To the best of our knowledge, this study is the first step towards unified MIE and initiates exploration into both instruction tuning and large language models in the MIE field.
- We propose a gate attention module to dynamically utilize visual features and conduct in-depth quantitative analysis on each MIE task, shading light into the role of visual features and validating our gate mechanism’s effectiveness and importance in our model.
- Extensive experiments show the effectiveness across six datasets of three MIE tasks, generalization ability in the zero-shot setting, and instruction robustness of our UMIE model. We will release all MIE datasets with standard format and models trained on them, as a benchmark and starting point for future studies in this area of unified multimodal information extraction.

Related Work

Multimodal Named Entity Recognition (MNER). The MNER task aims to recognize mentions and classify them into predefined categories from texts, using additional visual clues provided in images. Pioneering work (Moon, Neves, and Carvalho 2018; Lu et al. 2018; Zhang et al. 2018) focuses on utilizing visual features for improved representation learning. Several works (Arshad et al. 2019; Yu et al. 2020) point out that failure recognitions are due to unrelated images that corrupt visual attention and mislead entity recognition. As a remedy, Yu et al. (2020) propose the Unified Multimodal Transformer (UMT), in which a visual gate dynamically utilizes visual information for final representation. Additionally, RIVA (Sun et al. 2020) and RpBERT (Sun et al. 2021) explicitly predict the relevance of the given image and text, using text-image relationship binary classification as an auxiliary task, thus addressing the misleading issue introduced by irrelevant visual content.

Recently, ITA (Wang et al. 2022b) capitalizes on the objects in the given image and inputs with all text, thus leading to a unified representation and a better attention mechanism over texts. MNER-QG (Jia et al. 2023) frames MNER as a machine reading comprehension task, querying the models for entity recognitions. MoRe (Wang et al. 2022a) enhances the MNER model with related texts obtained via information retrieval techniques (Zhang et al. 2023; Shen et al. 2023).

Multimodal Relation Extraction (MRE). MRE (Zheng et al. 2021b,a) aims to identify the semantic relationships between two entities based on the given text image pair. HVP-

NeT (Chen et al. 2022) fuses image information as a prefix for better text representation. MoRe (Wang et al. 2022a) retrieves related texts from the entire Wikipedia dumps for boosting both MNER and MRE performance. Follow-up work (Hu et al. 2023) further retrieves relevant images to the object, text, and image for better retrieval augmentation (Xie et al. 2023; Yue et al. 2023). However, they involve time-consuming retrieval over a large-scale collection and require an external knowledge base. Therefore, these two retrieval-augmented methods cannot be directly compared with our work, which focuses on model development itself.

Multimodal Event Extraction (MEE). MEE aims to extract events (i.e., **Event Detection**) and arguments for the event (i.e., **Event Argument Extraction**) from multiple modalities. VAD (Zhang et al. 2017) leverages additional images to enhance event extraction by alleviating ambiguity in the text modality. Tong et al. (2020) propose DRMM, which recurrently uses related images from a constructed supplementary image set for augmenting text-only event detection, thereby improving the disambiguation of trigger words. Li et al. (2020) propose M²E² data and obtain weak supervision from text-only, image-only, and image-caption data to encode visual and textual data into a joint representation space for extraction. Follow-up work (Li et al. 2022) pre-trains a vision and language model in an event-level alignment with a contrastive learning fashion over a large event-rich dataset and generalize to the M²E² dataset. Recently, Unicl (Liu, Chen, and Xu 2022) proposes a unified contrastive learning framework to bridge the modality gaps.

Notably, MRE and MNER focus on leveraging visual information to enhance extraction from texts, rather than extraction over images while MEE may extract visual objects as arguments in an event. To the best of our knowledge, our proposed UMIE unifies the extraction from text and image in the multimodal IE domain for the first time.

Unified Information Extraction. Despite the diversity and heterogeneity of information extraction (IE) tasks, several works unify these IE tasks as text-to-structure generation (Lu et al. 2022b) or semantic matching (Lou et al. 2023). However, these models focus on text-only IE tasks. Our UMIE model solves multimodal IE tasks universally.

Meanwhile, instruction tuning (Lou, Zhang, and Yin 2023) fine-tunes models to follow instructions, showing unprecedented zero-shot generalization abilities when models perform unseen tasks given new instructions (Sanh et al. 2022; Ouyang et al. 2022; Chung et al. 2022). However, as pointed out by recent works (Zhang, Gutiérrez, and Su 2023), such instruction-tuned LLMs fail to achieve decent results in IE tasks due to the low incidence of these tasks during instruction tuning. Additionally, Chen et al. (Chen and Feng 2023) provide evidence that LLMs like ChatGPT¹ and GPT-4 (OpenAI 2023) yield poor results on both MNER and MRE tasks. Given high long-term costs and potential risks of using test data for training, such LLMs are not ideal choices for information extraction tasks. Therefore, open-sourced IE-specific models are more transparent, effective, and cost-efficient in practice, which is the focus of our work.

¹<https://openai.com/blog/chatgpt>

Unified Multimodal Information Extractor

Model Overview

As shown in Figure 2, the Unified Multimodal Information Extractor (UMIE) consists of four major modules: 1) Text encoder for instruction-following and text comprehension; 2) Visual encoder for visual representation; 3) Gated attention for cross-modal representation; 4) Text decoder for information extraction. The UMIE model utilizes a transformer-based encoder-decoder architecture to perform MIE and generate structured outputs in an auto-regressive fashion. For the textual input prefixed with a task instructor, we use a text encoder to generate text representations. For the image, we equip it with visual understanding abilities via our proposed visual encoder and gated attention mechanism for dynamic visual clue integration, the details of which will be introduced in the following sections. Finally, we employ a text decoder to generate the structural results for MIE tasks. In this work, we utilize FLAN-T5 (Chung et al. 2022), a Transformer-based encoder-decoder large language model (LLM), to initialize the structure and parameters of the encoder and decoder components of the UMIE model.

Specifically, the UMIE computes the hidden vector representation $h^e = \{h_1, \dots, h_n\} \in \mathbb{R}^{n \times d_t}$ of the input text $\{w_1, \dots, w_n\}$ where d_t is the dimension of token embeddings:

$$h^e = \text{Text-Encoder}(w_1, \dots, w_n). \quad (1)$$

In the gated attention module, we use the textual feature h^e as the query and the visual feature h^v as the key and value for the cross-attention computation. For dynamic incorporation of the text-aware visual representation, we design a gated signal g to control the final output of the cross-modal representation c (refer to Section ‘‘Gated Attention Module’’):

$$c = \text{Gated-Attention}(h^e, h^v). \quad (2)$$

Using the cross-modal representation c , the text decoder generates the output structure in an autoregressive manner, starting with the input of the start token $\langle s \rangle$ as the initial step. Such a generation process ends with the end token $\langle /s \rangle$. At step i , the text decoder represents the state H_i^d conditioned on the cross-modal state c and previous states $[H_1^d, \dots, H_{i-1}^d]$. Formally,

$$H_i^d = \text{Text-Decoder}([c; H_1^d, \dots, H_{i-1}^d]), \quad (3)$$

The text decoder consists of N -layer Transformers, additionally inserting a third sub-layer that performs multi-head attention over the output c of the gated attention module, a similar approach as described in (Vaswani et al. 2017). Based on the states $[H_1^d, \dots, H_i^d]$, we can decode a text sequence via a linear projection and softmax function.

Visual Encoder

In MIE tasks, the associated image typically offers a valuable visual clue, directing the models towards the information of interest. Therefore, to incorporate semantic knowledge, we encode the local objects. Furthermore, we consider

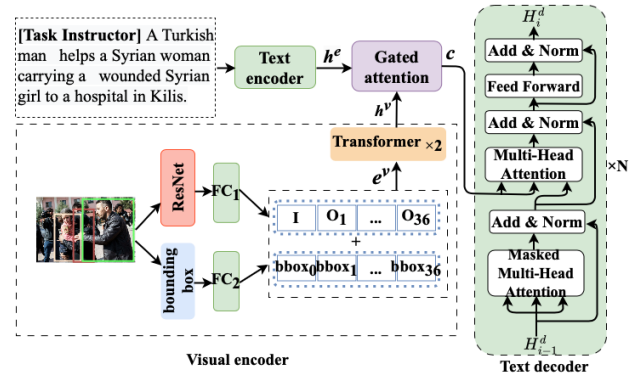


Figure 2: Illustration of the UMIE model. The visual encoder encodes an image and objects into features that are dynamically integrated with textual features in the gated attention module and the text decoder generates information extraction results autoregressively.

the image’s global feature for additional context information. By collecting both global and regional images, our visual encoder has the potential to extract more visual clues, leading to improved information utilization and more accurate extraction.

Specifically, we utilize semantic objects detected by off-the-shelf visual grounding toolkit (Tan and Bansal 2019). This toolkit provides up to 36 local region-of-interest (RoI) features as well as bounding box (bbox) coordinates which are x and y coordinates of the top left corner and the bottom right corner of the object rectangle. If the number of objects are less than 36, the remaining object features are padded with zeros. To unify the RoI inputs, we rescale the image and a series of visual objects to a size of 224×224 pixels.

Given an image I and its objects $\{O_1, \dots, O_{36}\}$, the backbone ResNet-101 (He et al. 2016) extracts visual features as $f = \{f_I, f_{O_1}, \dots, f_{O_{36}}\} \in \mathbb{R}^{37 \times d_v}$. Then two fully connected (FC) layers are applied to visual features f_i and their corresponding bounding box coordinates $bbbox_i$ to obtain visual embeddings e_i^v :

$$e_i^v = \text{FC}_1(f_i) + \text{FC}_2(bbbox_i), \quad (4)$$

where $\text{FC}_1 \in \mathbb{R}^{d_v \times d_t}$ and $\text{FC}_2 \in \mathbb{R}^{4 \times d_t}$. The visual embeddings e^v are further processed by a 2-layer Transformer (Vaswani et al. 2017) to obtain the final visual features $h^v \in \mathbb{R}^{37 \times d_t}$:

$$h^v = \text{Transformer}_{\times 2}(e^v). \quad (5)$$

The features h^v can represent the integration of global image and RoI local visual information.

Gated Attention Module

Previous work illustrated the potential irrelevance of images (Sun et al. 2021) or inductive bias (Yu et al. 2020) introduced by the images. To alleviate such an issue, we design a gate attention module shown in Figure 3 to dynamically integrate visual and textual features and the gate module controls the contribution of visual features for the text decoder.

Task		Task Instructor
MNER		Please extract the following entity type: person, location, miscellaneous, organization.
MRE		Please extract the following relation between [head] and [tail]: part of, contain, present in, none, held on, member of, peer, lace of residence, locate at, alternate names, neighbor, subsidiary, awarded, couple parent, nationality, place of birth, charges, siblings, religion, race.
MEE	MED	Please extract the following event type: arrest jail, meet, attack, transport, demonstrate, phone write, die, transfer.
	MEAE	Given an [TYPE] event, please extract the corresponding argument type: [Argument Roles]. (e.g., <i>given an arrest jail event, please extract the corresponding argument type: person, place, agent...</i>)

Table 1: The description of task instructors. As shown in the last row of the table, the specific instruction of MEAE will be determined by the event type detected by MED.

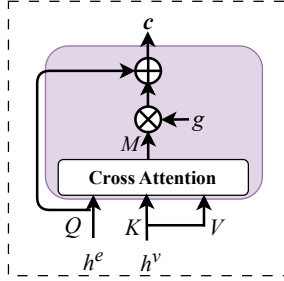


Figure 3: Illustration of the gated attention module.

Visual-textual Cross Attention. In the attention module, the visual features h^v serve as the key and value, while the textual representations h^e from the text decoder serve as the query. In this way, the textual representations are used to attend to the most relevant visual features, generating the text-aware visual representation $M \in \mathbb{R}^{n \times d_t}$:

$$\begin{aligned} M &= \text{Cross-Attention}(Q = h^e, K = h^v, V = h^v) \\ &= \text{softmax}\left(\frac{QK^T}{\sqrt{d_t}}\right)V, \end{aligned} \quad (6)$$

where Q , K , and V respectively represent query, key, and value. The softmax function is applied on the matrix multiplication of Q and the transpose of K , divided by the square root of the model dimension and used as weights multiplied with V .

Gate Control. To control the text-aware visual feature M , we design a gate signal g to denote the contribution of the visual feature in the text decoding process. The gate signal g is dynamically calculated by the following equation:

$$g = \text{LeakyReLU}(\overline{K} \overline{Q}^T), \quad (7)$$

where \overline{K} is the mean of vectors in K and \overline{Q} is the mean of vectors in Q .

Finally, the gate signal g controls the weight of text-aware visual feature M to deliver the cross-modal information $c \in \mathbb{R}^{n \times d_t}$ by textual features h^e and gated residual visual features $g \cdot M$,

$$\begin{aligned} c &= \text{Gated-Attention}(h^e, h^v) \\ &= h^e + g \cdot M. \end{aligned} \quad (8)$$

Such cross-modal representation is used for structure generation by the text decoder in an auto-regressive fashion, thus dynamically incorporating visual features for multimodal information extraction.

Task Instructor and Text Decoding

With multimodal understanding and generation abilities, UMIE could perform various multimodal information extraction tasks, given different instructions. Tables 1 and 2 show our task instructors and input and output formats used for MNER, MRE, and MEE, where MEE is conducted as multimodal event detection (MED) and multimodal event argument extraction (MEAE) in a cascaded fashion. Such instructors will be used as prefixes during both training and evaluation. For MRE, UMIE simply decodes the triples, including the relationship. For MNER, MED, and MEAE, we introduce `<spot>` to separate the classes. Specifically for event extraction tasks, we detect events by finding and identifying the triggers first by MED, based on which we further extract corresponding arguments by MEAE.

Experiments

We train the UMIE with instruction tuning on various MIE datasets and evaluate the model in both supervised learning and zero-shot settings. In addition, we evaluate the robustness of instruction following of UMIE and showcase the unified extraction abilities of our model.

Experiment Setup

Datasets. We train and evaluate UMIE on several datasets commonly used in MNER, MRE, and MEE tasks: 1) For MNER, we consider Twitter-15 (Zhang et al. 2018), SNAP (Lu et al. 2018), and Twitter-17 (Yu et al. 2020) (a refined version of SNAP), all curated from the social media platform; 2) For MRE, we adopt the MNRE dataset (Zheng et al. 2021b) constructed from the social media domain via crowdsourcing; 3) For MEE, following previous work (Tong et al. 2020), we employ datasets such as ACE2005 (Walker et al. 2006), SWiG (Pratt et al. 2020) for training, and the M²E² dataset for evaluation. ACE2005 features textual event annotations with 33 event types and 36 argument types. We utilize the M²E² dataset from (Li et al. 2020), containing 245 documents labeled with parallel textual and visual




Task	Image	Text	Output
MNER		[MNER Task Instructor] Support the Salvation Arm free kids soccer program by donating a ball or two this summer ! UEFAcom # uclfinal	<i>Person</i> , kids <spot> <i>Organization</i> , Salvation Army
MRE		[MRE Task Instructor] Star Fox will be an exclusive character in Starlink.	Star Fox, <i>member of</i> , Starlink
MEE		[MED Task Instructor] Smoke rises over the Syrian city of Kobani , following a US led coalition airstrike, seen from outside Suruc.	<i>Attack</i> , airstrike
		[MEAE Task Instructor] Smoke rises over the Syrian city of Kobani , following a US led coalition airstrike, seen from outside Suruc.	<i>Attacker</i> , coalition <spot> <i>Target</i> , O ₁

Table 2: The input and output format of three tasks. In particular, the MEE task is be conducted as two cascaded tasks, namely MED and MEAE, with the same input example.

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	MNRE-V1	7,824	975	1,282
	MNRE-V2	12,247	923	832
MEE	M ² E ²	-	-	309

Table 3: The statistics of six MIE datasets.

events. The M²E² event schema aligns with 8 ACE types and 98 SWiG types. The event instances are divided into 1,105 text-only events, 188 image-only events, and 385 multimedia events. Due to partial overlap between the Twitter-17 and SNAP datasets, we have removed 319 Twitter-17 testing samples from the SNAP training set, while retaining 2848 Twitter-17 training samples, ensuring that no data overlaps between the training and test sets. Detailed statistics of data are listed in Table 3. We have reformatted all datasets into a standardized JSON format for training our UMIE model.

Training Configuration. We train our model by employing label smoothing and AdamW, with a learning rate of 5e-5 for FLAN-T5-large and 1e-4 for FLAN-T5-base. The number of training epochs is set to 40. All experiments are conducted on 8 NVIDIA A100 GPUs, each possessing a memory capacity of 40GB. Due to GPU memory limitations, we use different batch sizes: 8 for FLAN-T5-large and 16 for FLAN-T5-base. During the training process, we restrict the output of the text input to a maximum length of 256 and the generated length to 128. We add the task instructor as the prefix in the default setup for FLAN-T5.

Evaluation Metrics. Among these four MIE tasks (MNER, MRE, MED, and MEAE), we conduct comprehensive comparisons between our UMIE and state-of-the-art multimodal approaches in each task. We use the reported performances from the respective papers for comparison. In terms of the

evaluation metrics, we follow the F1 metric commonly used in each task as described in the respective papers.

Main Results

Table 4 shows the comparison in detail. Our UMIE achieves on-par or significantly better performances compared to the previous baselines.. In particular, the UMIE-XL achieves the SoTA performance on all six datasets, when compared to the best model in each dataset. UMIE marginally outperforms previous SoTA methods on both MRE datasets, with more than a 10% absolute F1 advantage on MNRE-V2. This performance superiority strongly indicates the effectiveness and generality of UMIE in handling various multimodal information extraction tasks, also demonstrating the success of our proposed visual encoder, gated attention module, and training strategies. More importantly, such a unified single checkpoint with capable performance across the board can serve as a reliable starting point for their various MIE tasks. Therefore, MIE practitioners do not need to train and maintain separate models for each specific scenario or dataset.

The M²E² dataset has only a test set available while UMIE can generalize well on this M²E², outperforming previous SoTA results by a large margin. This demonstrates the strong generalization capability of UMIE and the effectiveness of multi-task training. Further in-depth generalization evaluation on MNER and MRE tasks will be discussed in the following experiment.

In addition, the UMIE can achieve even better results by using more powerful backbone models. For instance, transitioning from Base to Large, and then to XLarge size in the FLAN-T5 model family, consistently improves the performance. This indicates the effectiveness and versatility of our model design and training strategies.

Zero-shot Generalization

We evaluate the generalization ability of our UMIE models to unseen MIE tasks in the zero-shot setting. Specifically, we exclude the corresponding training set (e.g., Twitter-17

Method	MNER			MRE		MEE	
	Twitter-15	Twitter-17	SNAP	MNRE-V1	MNRE-V2	M ² E ² MED	M ² E ² MEAE
UMT (Yu et al. 2020)	73.4	73.4	-	-	65.2	-	-
UMGF (Zhang et al. 2021)	74.9	85.5	-	-	-	-	-
MEGA (Zheng et al. 2021a)	72.4	84.4	66.4	-	-	-	-
RpBERT (Sun et al. 2021)	74.4	-	85.7	-	-	-	-
R-GCN (Zhao et al. 2022)	75.0	87.1	-	-	-	-	-
HVPNeT (Chen et al. 2022)	75.3	86.8	-	81.8	-	-	-
ITA (Wang et al. 2022b)	78.0	89.8	90.2	-	66.9	-	-
MoRe (Wang et al. 2022a)	77.3	88.7	89.3	-	65.8	-	-
MNER-QG (Jia et al. 2023)	75.0	87.3	-	-	-	-	-
HamLearning (Liu et al. 2023a)	76.5	87.1	-	-	-	-	-
EviFusion (Liu et al. 2023b)	75.5	87.4	-	-	-	-	-
BGA-MNER (Chen et al. 2023)	76.3	87.7	-	-	-	-	-
WASE (Li et al. 2020)	-	-	-	-	-	50.8	19.2
Unicl (Liu, Chen, and Xu 2022)	-	-	-	-	-	57.6	23.4
UMIE-Base (Ours)	76.1	88.1	87.7	84.3	74.8	60.5	22.5
UMIE-Large (Ours)	77.2	<u>90.7</u>	<u>90.5</u>	<u>85.0</u>	<u>75.5</u>	<u>61.0</u>	<u>23.6</u>
UMIE-XLarge (Ours)	78.2	91.4	91.0	86.4	76.2	62.1	24.5

Table 4: Performance comparison on three multimedia information extraction tasks in F1 score (%). The best performance is marked on bold and the second best result is marked underline.

Method	Twitter-17	MNRE-V2
UMT (Yu et al. 2020)	60.9	-
UMGF (Zhang et al. 2021)	60.9	-
FMIT (Lu et al. 2022a)	64.4	-
CAT-MNER (Wang et al. 2022c)	64.5	-
BGA-MNER (Chen et al. 2023)	64.9	-
ChatGPT	57.5	35.2
GPT-4 (OpenAI 2023)	66.6	42.1
UMIE-Base (Ours)	66.8	67.3
UMIE-Large (Ours)	68.5	68.8
UMIE-XLarge (Ours)	69.9	69.6

Table 5: Performance comparison on zero-shot setting (%). We report the results of each paper and adopt the results of ChatGPT and GPT-4 from Chen and Feng (2023).

or MNRE-V2) and train a checkpoint with the rest of the training data. The baselines are trained with the same rule of the training set exclusion. Therefore, these models are evaluated purely on their ability to generalize from related tasks to unseen MNER (Twitter-17) and MRE (MNRE-V2) datasets.

Table 5 reports final results and we could observe that our UMIE substantially outperforms other zero-shot baselines and such advantages continually enlarge as model size increases. In particular, UMIE outperforms LLMs including ChatGPT and even GPT-4 in both MNER and MRE tasks, strongly indicating the effectiveness of our model. Considering that the generalization is also verified in MEE, these experiments indicate the ability of UMIE to adapt and generalize to new tasks effectively, even without direct training on corresponding task data. The high performance of UMIE in zero-shot setting can be attributed to instruction-following ability and multi-task learning on multiple MIE

MNER	Task Instructor
I0	Given the entity types: person, location, miscellaneous, organization. Please extract the specified entity type.
I1	Identity the following entity type from the given sentence: person, location, miscellaneous, organization.
I2	Please extract entity type in the sentence. Option: person, location, miscellaneous, organization.

Table 6: Three variants of MNER instructor.

tasks which allow our model to transfer knowledge from one task to another. This proves that our model is not only capable of learning from different sources of multimodal data but also successfully applying learned knowledge to unseen task data, underscoring its flexibility and versatility.

Robustness to Instruction-Following

An ideal model with instruction-following capability would understand and execute instructions correctly, regardless of their phrasing. In such context, users can get decent results without carefully crafting the instructions, which requires time-consuming and tricky prompt engineering. Therefore, to evaluate the robustness of the instruction-following ability of UMIE models, as shown in Table 6, we provide UMIE models with three kinds of task instructors for the MNER task. The results in Table 7 demonstrate that UMIE models have decent robustness to various instructions, maintaining consistently high performances with all model sizes.

	Method	Twitter-15	Twitter-17	SNAP
I0	UMIE-Base	75.8	88.0	87.8
	UMIE-Large	76.8	90.7	90.1
	UMIE-XLarge	78.0	92.1	91.7
I1	UMIE-Base	75.9	88.1	87.6
	UMIE-Large	77.2	90.4	90.0
	UMIE-XLarge	77.9	91.0	90.8
I2	UMIE-Base	76.0	88.2	87.5
	UMIE-Large	77.0	90.5	90.4
	UMIE-XLarge	78.2	91.2	91.0

Table 7: Performance of UMIE with three MNER instructors in F1 score (%).

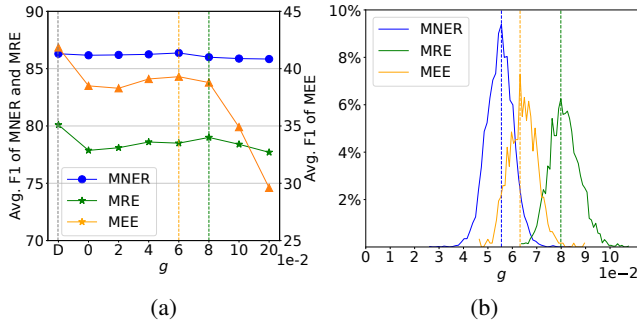


Figure 4: (a) Performances of UMIE-Base averaged over three MIE tasks *w.r.t* fixed gate value g where D denotes the dynamic gate value of the UMIE; (b) g value distribution of our dynamic gate module in three MIE tasks of UMIE-Base.

Gate Control Ablation

To disentangle the effects of the gated fusion mechanism and visual information in our UMIE model, we conduct an ablation study. Concretely, we fix gate value g in Eq. (7) for using visual information and observe the impact on our model’s performance in Figure 4. Based on Figure 4(b), it is evident that in the UMIE model, the effective range of the gate value is limited to 0-0.1. When the fixed gate value exceeds 0.1, a significant drop in performance is observed across all MIE tasks. Therefore, our primary focus is to present and analyze the performance specifically within the range of $g \in [0, 0.1]$.

In comparison to our dynamic gate value (D), any fixed gate value results in a decline in performance, particularly in MRE and MEE tasks. Specifically, when compared to $g = 0$, representing only text without visual information, the dynamic gate value shows an average improvement of 0.1% F1 score in MNER, 2.3% in MRE, and 3.3% in MEE. This indicates that MRE and MEE are more heavily influenced by visual features, whereas MNER is hardly sensitive to visual aids. This observation can be attributed to the fact that the MRE task involves inferring relationships between entities based on visual information, and the MEE task utilizes visual mentions as arguments, both of which require a high level of understanding of images. Further evidence supporting this can be seen in the distribution of the g values shown in Figure 4(b). Specifically, the peak values of the histogram

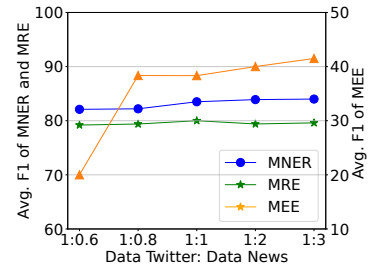


Figure 5: Averaged performance of three MIE tasks *w.r.t* training sampling ratios of Twitter and News corpora.

for MEE ($\approx 6.3e-2$) and MRE ($\approx 8e-2$) are higher compared to that of MNER ($\approx 5.5e-2$), implying a stronger emphasis on visual features in the former two tasks.

Figure 4(b) demonstrates the effectiveness of our gate module in assigning appropriate weights to each task. The peak values of the histogram for MRE ($\approx 6.3e-2$) and MEE ($\approx 8e-2$) coincide with the highest performances observed in the fixed gate value experiments, represented by the yellow dotted line for MEE and the green dotted line for MRE in Figure 4(a). By dynamically adjusting the weights for individual input instances based on their specific characteristics, our model achieves optimal results. This finding serves as strong evidence for the effectiveness and necessity of our dynamic gated fusion mechanism.

Training Materials

We also delve deep into the effects of training corpus, namely Twitter including Twitter-15, Twitter-17, SNAP, MNRE-V1, and MNRE-V2, and News including ACE2005 and M²E², on the performance of the UMIE model. In Figure 5, using UMIE-Base as an example, we observe that the Twitter corpus primarily contributes to MNER and MRE tasks, while the News corpus helps UMIE adapt to MEE. Importantly, increasing training examples from the News domain does not negatively affect the tasks from the Twitter domain. This indicates that UMIE series models exhibit strong compatibility when trained on cross-domain corpora and have the potential to benefit even more from larger-scale training materials.

Conclusion

This work proposes a unified framework for all MIE tasks with instruction tuning and multi-task learning based on encoder-decoder LLMs. The gated attention module can effectively yield cross-modal information for decoding the MIE results. Extensive experiments over three MIE tasks and six datasets show that our single UMIE model outperforms various prior task-specific SoTA methods across the board by a large margin, i.e., on average + 0.9% for MNER, + 7.0% for MRE, and + 2.8% for MEE. Our framework has strong generalization ability and robustness to instruction. These desired properties make UMIE a foundation model in the MIE domain, demonstrating its broad applicability and promising potential for future work.

Acknowledgements

This work was supported in part by Zhejiang Provincial Natural Science Foundation of China under Grant No. LGN22F020002 and Key Research and Development Program of Zhejiang Province under Grant No. 2022C03037.

References

- Arshad, O.; Gallo, I.; Nawaz, S.; and Calefati, A. 2019. Aiding Intra-Text Representations with Visual Context for Multimodal Named Entity Recognition. In *Proceedings of IC-DAR*.
- Chen, F.; and Feng, Y. 2023. Chain-of-Thought Prompt Distillation for Multimodal Named Entity and Multimodal Relation Extraction. *arXiv:2306.14122*.
- Chen, F.; Liu, J.; Ji, K.; Ren, W.; Wang, J.; and Wang, J. 2023. Learning Implicit Entity-object Relations by Bidirectional Generative Alignment for Multimodal NER. *arXiv:2308.02570*.
- Chen, X.; Zhang, N.; Li, L.; Yao, Y.; Deng, S.; Tan, C.; Huang, F.; Si, L.; and Chen, H. 2022. Good Visual Guidance Make A Better Extractor: Hierarchical Visual Prefix for Multimodal Entity and Relation Extraction. In *Findings of NAACL*, 1607–1618.
- Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S.; Webson, A.; Gu, S. S.; Dai, Z.; Suzgun, M.; Chen, X.; Chowdhery, A.; Narang, S.; Mishra, G.; Yu, A.; Zhao, V. Y.; Huang, Y.; Dai, A. M.; Yu, H.; Petrov, S.; Chi, E. H.; Dean, J.; Devlin, J.; Roberts, A.; Zhou, D.; Le, Q. V.; and Wei, J. 2022. Scaling Instruction-Finetuned Language Models. *CoRR*, abs/2210.11416.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of CVPR*.
- Hu, X.; Guo, Z.; Teng, Z.; King, I.; and Yu, P. S. 2023. Multimodal Relation Extraction with Cross-Modal Retrieval and Synthesis. *arXiv:2305.16166*.
- Jia, M.; Shen, L.; Shen, X.; Liao, L.; Chen, M.; He, X.; Chen, Z.; and Li, J. 2023. MNER-QG: an end-to-end MRC framework for multimodal named entity recognition with query grounding. In *Proceedings of AAAI*.
- Li, M.; Xu, R.; Wang, S.; Zhou, L.; Lin, X.; Zhu, C.; Zeng, M.; Ji, H.; and Chang, S.-F. 2022. CLIP-Event: Connecting Text and Images with Event Structures. In *Proceedings of CVPR*.
- Li, M.; Zareian, A.; Zeng, Q.; Whitehead, S.; Lu, D.; Ji, H.; and Chang, S.-F. 2020. Cross-media Structured Common Space for Multimedia Event Extraction. In *Proceedings of ACL*.
- Liu, J.; Chen, Y.; and Xu, J. 2022. Multimedia Event Extraction From News With a Unified Contrastive Learning Framework. In *Proceedings of ACM Multimedia*.
- Liu, P.; Li, H.; Ren, Y.; Liu, J.; Si, S.; Zhu, H.; and Sun, L. 2023a. A Novel Framework for Multimodal Named Entity Recognition with Multi-level Alignments. *arXiv preprint arXiv:2305.08372*.
- Liu, W.; Zhong, X.; Hou, J.; Li, S.; Huang, H.; and Fang, Y. 2023b. Integrating Large Pre-trained Models into Multimodal Named Entity Recognition with Evidential Fusion. *arXiv preprint arXiv:2306.16991*.
- Lou, J.; Lu, Y.; Dai, D.; Jia, W.; Lin, H.; Han, X.; Sun, L.; and Wu, H. 2023. Universal Information Extraction as Unified Semantic Matching. In *Proceedings of AAAI*.
- Lou, R.; Zhang, K.; and Yin, W. 2023. Is Prompt All You Need? No. A Comprehensive and Broader View of Instruction Learning. *arXiv preprint arXiv:2303.10475*.
- Lu, D.; Neves, L.; Carvalho, V.; Zhang, N.; and Ji, H. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of ACL*.
- Lu, J.; Zhang, D.; Zhang, J.; and Zhang, P. 2022a. Flat Multimodal Interaction Transformer for Named Entity Recognition. In *Proceedings of COLING*.
- Lu, Y.; Liu, Q.; Dai, D.; Xiao, X.; Lin, H.; Han, X.; Sun, L.; and Wu, H. 2022b. Unified Structure Generation for Universal Information Extraction. In *Proceedings of ACL*.
- Moon, S.; Neves, L.; and Carvalho, V. 2018. Multimodal Named Entity Recognition for Short Social Media Posts. In *Proceedings of NAACL*.
- OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155.
- Pratt, S.; Yatskar, M.; Weihs, L.; Farhadi, A.; and Kembhavi, A. 2020. Grounded situation recognition. In *Proceedings of ECCV*.
- Sanh, V.; Webson, A.; Raffel, C.; Bach, S.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Raja, A.; Dey, M.; Bari, M. S.; Xu, C.; Thakker, U.; Sharma, S. S.; Szczechla, E.; Kim, T.; Chhablani, G.; Nayak, N. V.; Datta, D.; Chang, J.; Jiang, M. T.; Wang, H.; Manica, M.; Shen, S.; Yong, Z. X.; Pandey, H.; Bawden, R.; Wang, T.; Neeraj, T.; Rozen, J.; Sharma, A.; Santilli, A.; Févry, T.; Fries, J. A.; Teehan, R.; Scao, T. L.; Biderman, S.; Gao, L.; Wolf, T.; and Rush, A. M. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *Proceedings of ICLR*.
- Shen, T.; Geng, X.; Tao, C.; Xu, C.; Long, G.; Zhang, K.; and Jiang, D. 2023. Unifier: A Unified Retriever for Large-Scale Retrieval. *arXiv:2205.11194*.
- Sun, L.; Wang, J.; Su, Y.; Weng, F.; Sun, Y.; Zheng, Z.; and Chen, Y. 2020. RIVA: A Pre-trained Tweet Multimodal Model Based on Text-image Relation for Multimodal NER. In *Proceedings of COLING*.
- Sun, L.; Wang, J.; Zhang, K.; Su, Y.; and Weng, F. 2021. RpBERT: a text-image relation propagation-based BERT model for multimodal NER. In *Proceedings of AAAI*.
- Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of EMNLP-IJCNLP*.

- Tong, M.; Wang, S.; Cao, Y.; Xu, B.; Li, J.; Hou, L.; and Chua, T.-S. 2020. Image enhanced event detection in news articles. In *Proceedings of AAAI*.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of NeurIPS*.
- Walker, C.; Strassel, S.; Medero, J.; and Maeda, K. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57: 45.
- Wang, X.; Cai, J.; Jiang, Y.; Xie, P.; Tu, K.; and Lu, W. 2022a. Named Entity and Relation Extraction with Multi-Modal Retrieval. In *Findings of EMNLP*.
- Wang, X.; Gui, M.; Jiang, Y.; Jia, Z.; Bach, N.; Wang, T.; Huang, Z.; and Tu, K. 2022b. ITA: Image-Text Alignments for Multi-Modal Named Entity Recognition. In *Proceedings of NAACL*.
- Wang, X.; Ye, J.; Li, Z.; Tian, J.; Jiang, Y.; Yan, M.; Zhang, J.; and Xiao, Y. 2022c. CAT-MNER: Multimodal Named Entity Recognition with Knowledge-Refined Cross-Modal Attention. In *Proceedings of ICME*.
- Wang, Y.; Mishra, S.; Alipoormolabashi, P.; Kordi, Y.; Mirzaei, A.; Arunkumar, A.; Ashok, A.; Dhanasekaran, A. S.; Naik, A.; Stap, D.; et al. 2022d. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ Tasks. In *Proceedings of EMNLP*.
- Xie, J.; Zhang, K.; Chen, J.; Lou, R.; and Su, Y. 2023. Adaptive Chameleon or Stubborn Sloth: Unraveling the Behavior of Large Language Models in Knowledge Clashes. *arXiv:2305.13300*.
- Yu, J.; Jiang, J.; Yang, L.; and Xia, R. 2020. Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer. In *Proceedings of ACL*.
- Yue, X.; Wang, B.; Zhang, K.; Chen, Z.; Su, Y.; and Sun, H. 2023. Automatic Evaluation of Attribution by Large Language Models. *arXiv:2305.06311*.
- Zhang, D.; Wei, S.; Li, S.; Wu, H.; Zhu, Q.; and Zhou, G. 2021. Multi-modal Graph Fusion for Named Entity Recognition with Targeted Visual Guidance. In *Proceedings of AAAI*.
- Zhang, K.; Gutiérrez, B. J.; and Su, Y. 2023. Aligning Instruction Tasks Unlocks Large Language Models as Zero-Shot Relation Extractors. In *Findings of ACL*.
- Zhang, K.; Tao, C.; Shen, T.; Xu, C.; Geng, X.; Jiao, B.; and Jiang, D. 2023. LED: Lexicon-Enlightened Dense Retriever for Large-Scale Retrieval. In *Proceedings of WWW*.
- Zhang, Q.; Fu, J.; Liu, X.; and Huang, X. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of AAAI*.
- Zhang, T.; Whitehead, S.; Zhang, H.; Li, H.; Ellis, J. G.; Huang, L.; Liu, W.; Ji, H.; and Chang, S. 2017. Improving Event Extraction via Multimodal Integration. In *Proceedings of ACM Multimedia*.
- Zhao, F.; Li, C.; Wu, Z.; Xing, S.; and Dai, X. 2022. Learning from Different Text-Image Pairs: A Relation-Enhanced Graph Convolutional Network for Multimodal NER. In *Proceedings of ACM Multimedia*.
- Zheng, C.; Feng, J.; Fu, Z.; Cai, Y.; Li, Q.; and Wang, T. 2021a. Multimodal Relation Extraction with Efficient Graph Alignment. In *Proceedings of ACM Multimedia*.
- Zheng, C.; Wu, Z.; Feng, J.; Fu, Z.; and Cai, Y. 2021b. MNRE: A Challenge Multimodal Dataset for Neural Relation Extraction with Visual Evidence in Social Media Posts. In *Proceedings of ICME*.
- Zhu, X.; Li, Z.; Wang, X.; Jiang, X.; Sun, P.; Wang, X.; Xiao, Y.; and Yuan, N. J. 2022. Multi-Modal Knowledge Graph Construction and Application: A Survey. *IEEE Transactions on Knowledge and Data Engineering*.