Enhancing Multimodal Entity and Relation Extraction With Variational Information Bottleneck

Shiyao Cui, Jiangxia Cao, Xin Cong, Jiawei Sheng, Quangang Li, Tingwen Liu, and Jinqiao Shi

***Abstract*—This article studies the multimodal named entity recognition (MNER) and multimodal relation extraction (MRE), which are important for content analysis and various applications. The core of MNER and MRE lies in incorporating evident visual information to enhance textual semantics, where two issues inherently demand investigations. The first issue is modality-noise, where the task-irrelevant information in each modality may be noises misleading the task prediction. The second issue is modality-gap, where representations from different modalities are inconsistent, preventing from building the semantic alignment between the text and image. To address these issues, we propose a novel method for MNER and MRE by MultiModal representation learning with Information Bottleneck (MMIB). For the first issue, a refinement-regularizer probes the information-bottleneck principle to balance the predictive evidence and noisy information, yielding expressive representations for prediction. For the second issue, an alignment-regularizer is proposed, where a mutual information-based item works in a contrastive manner to regularize the consistent text-image representations. To our best knowledge, we are the first to explore variational IB estimation for MNER and MRE. Experiments show that MMIB achieves the state-of-the-art performances on three public benchmarks.**

***Index Terms*—Multimodal named entity recognition, multimodal relation extraction, information bottleneck.**

I. Introduction

NAMED entity recognition (NER) [1] and relation extraction (RE) [2] are two fundamental tasks for information extraction, benefiting various applications like knowledge graph construction [3] and so forth. Specifically, NER aims to extract entities of interest and RE seeks to decide the semantic relations between entities from unstructured texts, respectively. Previous NER and RE efforts have been mainly devoted in newswire domain, while social media platforms (e.g. Twitter) are drawing increasing research attention due to their unprecedent development. However, texts in social media are usually short and accompanied with slangs [4], [5], where the inadequate and ambiguous semantics raise the difficulty to recognize the entities and relations between them.

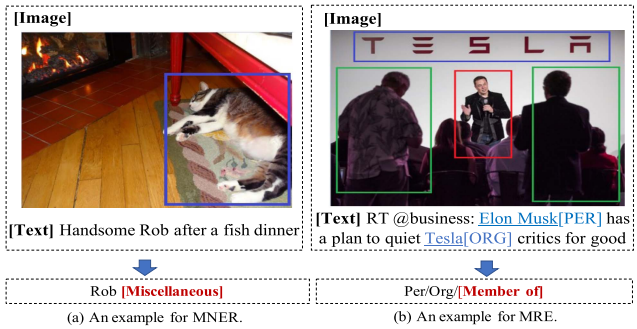

Fig. 1. Examples from Twitter datasets for MNER and MRE tasks.

With the prevalence of multimodal posts in social media, in recent years, images in social posts are leveraged to facilitate NER and RE towards social media, namely **Multimodal NER (MNER)** [6] and **Multimodal RE (MRE)** [2]. As fundamental tasks for language understanding towards multi-modal social media, MNER and MRE have been widely studied together [7], [8], [9]. Specifically, MNER and MRE both achieve their goals by incorporating the evident visual information to enhance the textual semantics. Early researchers directly explore the whole image as global-level visual clues. These methods either encode the raw image as one feature vector [10] for the semantic enhancement, or roughly segment the image into multiple grids for cross-modal interaction [11], [12]. For the more fine-grained visual clues, the following researchers extract the salient objects in the image and encode them as local-level visual hints for task predictions [4], [5], [7], [13], [14], [15], [16], [17]. Despite their remarkable success, two issues are still inadequately addressed.

The first issue is **modality-noise**, where both modalities contain task-irrelevant information, which can not contribute to the final tasks and even be noises misleading the prediction. Taking the MNER in Fig. 1(a) as an example. For the text-modality, the span "Rob" is likely to be predicted as a person-type (PER) entity, suffering from the noises from the words ("handsome" and "dinner") which pose characteristics of a person to "Rob". Fortunately, from the overview of the text and image, the "Cat" object (e.g. the region in the blue box) in the image is evident to guide the correct prediction. Meanwhile, for the image-modality, the noises could lie in two-folds: 1) in the global-level, most regions in the image are not informative to the target entity recognition; 2) in the local-level, the evident region expresses a

Manuscript received 14 April 2023; revised 21 September 2023; accepted 3 December 2023. Date of publication 24 January 2024; date of current version 7 February 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFB3100600 and in part by the Youth Innovation Promotion Association of CAS under Grant 2021153. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. Suma Bhat. *(Corresponding author: Tingwen Liu.)*

Shiyao Cui, Jiangxia Cao, Xin Cong, Jiawei Sheng, Quangang Li, and Tingwen Liu are with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100193, China, and also with the University of Chinese Academy of Sciences, Beijing 100193, China (e-mail: cuishiyao@iie.ac.cn; caojiangxia@iie.ac.cn; congxin@iie.ac.cn; shengjiawei@iie.ac.cn; liquangang@iie.ac.cn; liutingwen@iie.ac.cn).

Jinqiao Shi is with the Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: shijinqiao@bupt.edu.cn).

Digital Object Identifier 10.1109/TASLP.2023.3345146

more complicated visual semantics ("A cat lies on the carpet") than what we need exactly (just a "Cat"). Here, the redundant visual elements may disturb the attention assignments towards the useful visual regions, thus impede the final task predictions. Similarly, in Fig. 1(b), the redundant noisy regions in green boxes may drive the relation prediction between persons, rather than between an organization and a person. Despite previous efforts towards modality-noise [7], [18], these methods only attempt to alleviate the visual noises but ignore the noise-filtering in both modalities.

Even though the redundant noises are erased, **modality-gap** still exists, where the representations of the input text and image are inconsistent. Specifically, since the textual and visual representations are respectively obtained from different encoders, they maintain different feature spaces and distributions. Such a modality-gap makes it confusing for the text and image to "understand" each other and struggling to grasp the cross-modality correlations. For MNER in Fig. 1(a), ideally, the textual entity "Rob" should hold stronger semantic relevance with the "Cat" in the blue box than other regions. However, the disparity between textual and visual representations prevents from building such an alignment, hindering the exploration to predictive visual hints. For MRE in Fig. 1(b), the disagreement between representations makes it hard to align the annotated textual entities with the visual person/organization objects for relation decision. Though prior researches [5], [19] convert visual objects into the textual labels for the consistent expression, they inevitably suffer from the error sensitivity of textual labels produced by external off-the-shelf tools.

In this article, we propose a novel approach for MNER and MRE by MultiModal Representation learning with variational Information Bottleneck, which is termed as **MMIB**. Our method tackles the issues above by regularizing the data distributions based on the variational auto-encoder [20] framework. Specifically, for **modality-noise**, we design a *Refinement-Regularizer (RR)*, which explores the Information-Bottlenack (IB) principle for text-image representation learning. IB principle aims at deriving representations in terms of *a trade-off between having a concise representation with its own information and one with general predictive power* [21]. In our tasks, IB refines text/image representations which are diminished from the noisy information but predictive for final task predictions. Our devised RR consists of two IB-terms respectively upon the representation learning of the input text and image, yielding robust representations to each modality for prediction. For **modality-gap**, an *Alignment-Regularizer (AR)* is proposed. Since mutual information (MI) could measure the association between two variables, AR drives the consistent text-image representations by maximizing the MI between the representations of the paired text-image while minimizing that between the unpaired ones. Based on the regularizers above, expressive text-image representations are derive to facilitate the incorporation to evident visual information, thus the performances of MNER and MNRE could be both improved.

Overall, we summarize our contributions as follows:

- We explore a fresh perspective to solve MNER and MRE via representation learning with variational information bottleneck principle.
- We propose a novel method, MMIB, which employs a refinement-regularizer and an alignment-regularizer respectively for modality-noise and modality-gap.
- Experiments show that MMIB achieves new state-of-the-art performances on three MNER and MRE public benchmarks, and extensive analyses verify the effectiveness of our proposed method.

II. Preliminary

A. Task Formulation

1) MNER: Given a sentence and its associated image, MNER aims to identify the textual named entities and classify the identified entities into the predefined entity types. We formulate the task into a sequence labeling paradigm. Formally, let $(x_1^t, x_2^t, \ldots, x_n^t)$ denote the input sentence containing n tokens, we aim to predict the corresponding label sequence $Y = (y_1, y_2, \ldots, y_n)$ where y_i is a predefined label which follows BIO-tagging schema [12] so that the entities of interests could be derived.

2) MRE: Given a sentence and its associated image, the goal of MRE is to detect the semantic relations between two annotated entities. Formally, let $(x_1^t, x_2^t, \ldots, x_n^t)$ denote the sentence containing two entities $E_1 = (x_i^t, \ldots, x_{i+|E_1|-1}^t)$ and $E_2 = (x_j^t, \ldots, x_{j+|E_2|-1}^t)$, the task is formulated as a classification problem to decide the relation types Y between E_1 and E_2 from the predefined relation types.

B. Mutual Information

Before going on, we first introduce Mutual Information (MI) since it is one basic concept for information bottleneck. MI is a general metric in information theory [22], which measures the strength of association between random quantities. Formally, MI is defined as follows:

$$I(\boldsymbol{X};\boldsymbol{Y}) = \sum_{y\in\boldsymbol{Y}}\sum_{x\in\boldsymbol{X}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}, \tag{1}$$

where $I(\cdot;\cdot)$ denotes the MI between two random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, $p(\cdot)/p(\cdot,\cdot)$ respectively refer to the marginal/joint **probability** for samples x and y. Meanwhile, MI could also be measured by Kullback–Leibler divergence [23], [24] between two distributions as:

$$I(\boldsymbol{X};\boldsymbol{Y}) = \mathbb{D}_{KL}(\boldsymbol{p}(\boldsymbol{X}|\boldsymbol{Y})||\boldsymbol{p}(\boldsymbol{X})), \tag{2}$$

where $\boldsymbol{p}(\cdot)/\boldsymbol{p}(\cdot|\cdot)$ denote the prior/posterior **distribution** for variables $\boldsymbol{X}$ and $\boldsymbol{Y}$.

C. Information Bottleneck Principle

Information Bottleneck (IB), which was proposed by Tishby et al. [25], aims to derive effective latent features with a tradeoff between having an explicit representation and one with general predictive power [21]. Assuming we have the task input information $\boldsymbol{X}$ and the task-expected information $\boldsymbol{R}$, the IB is formed as:

$$\mathcal{L}_{IB} = \beta I(\boldsymbol{Z};\boldsymbol{X}) - I(\boldsymbol{Z};\boldsymbol{R}), \tag{3}$$

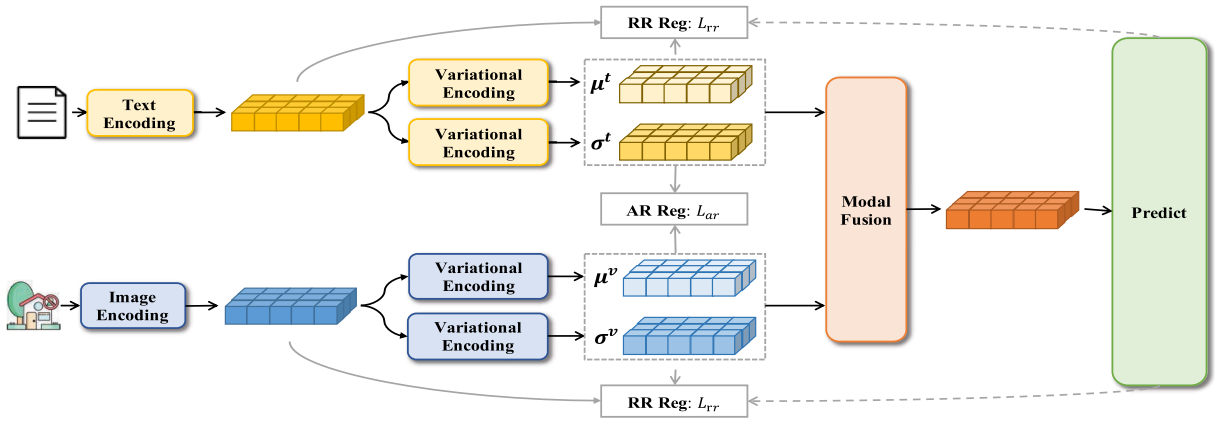

Fig. 2. Toy illustration to the proposed model architecture of multimodal representation learning with variational information bottleneck (MMIB).

Particularly, the goal of (3) is to derive a latent variable $\boldsymbol{Z}$, which filters redundant information provided by the input $\boldsymbol{X}$ while maximally maintain the information for $\boldsymbol{R}$. The objective function could be explained as follows: 1) **minimizing** $I(\boldsymbol{Z};\boldsymbol{X})$ so that $\boldsymbol{Z}$ could discard irrelevant parts for task predictions; 2) **maximizing** $I(\boldsymbol{Z};\boldsymbol{R})$ to enforce $\boldsymbol{Z}$ to retain the predictive information. The $\beta > 0$ is a Lagrangian multiplier to balance the trade-off between the two constraints. In implementation, since it is intractable to directly optimize the MI-based terms, variational approximation [26] is widely adopted for optimization to the corresponding objection functions [24].

III. Method

In this section, we introduce MMIB as Fig. 2 shows. Specifically, MMIB contains three modules: 1) Encoding module converts the texts and images into the real-valued embeddings. 2) Representation learning module conducts variational encoding with our proposed regularizers. 3) Prediction module conducts the modality fusion and performs the final task predictions.

A. Encoding Module

This module conducts the encoding towards the input sentence and image, converting them into the contextualized representations.

1) Text Encoding: To derive the representations for the input sentence, we employ the widely used BERT [27] as the textual encoder. Following Devlin et al. [27], two special tokens are inserted into each sentence, where [CLS] and [SEP] are respectively appended to the beginning and end to the sentence. Formally, let $X^T = (x_0^t, x_1^t, x_2^t, \ldots, x_n^t, x_{n+1}^t)$ denote each processed sentence, where x_0^t and x_{n+1}^t respectively denote the inserted [CLS] and [SEP]. We feed X^T into BERT, and obtain the output representations as $\boldsymbol{X}^T = (\boldsymbol{x}_0^t, \boldsymbol{x}_1^t, \boldsymbol{x}_2^t, \ldots, \boldsymbol{x}_n^t, \boldsymbol{x}_{n+1}^t)$, where $\boldsymbol{x}_i^t \in \mathbb{R}^d$ is the contextualized representation for the i_{th} token.

2) Image Encoding: Given an image X^V, it contains visual information from two aspects: (i) the global-level abstract concepts provided by the whole image; (ii) the local-level semantic units provided by the visual objects. We intend to explore such two-fold visual information, and thus two steps are involved to produce the image representations. **Step1. Object Extraction.** Following Zhang et al. [14], we employ a visual grounding toolkit [28] to extract local objects, which are denoted as $(x_1^v, x_2^v, \ldots, x_m^v)$. **Step2. Image Representation.** Considering the great success of ResNet [29] in computer vision, we utilize it as the visual encoder. Specifically, we first rescale the whole image and its object images into $224 * 224$ pixels, and feed them into ResNet. Since ResNet produces visual representations in the dimension of 2048, a linear transformation matrix $\boldsymbol{W}_v \in \mathbb{R}^{2048\times d}$ works to project the image representations into the same dimension as the textual representations. Finally, we concatenate the image representations together as $\boldsymbol{X}^V = (\boldsymbol{x}_0^v, \boldsymbol{x}_1^v, \ldots, \boldsymbol{x}_m^v)$, where $\boldsymbol{x}_0^v \in \mathbb{R}^{1\times d}$ and $\boldsymbol{x}_i^v \in \mathbb{R}^{1\times d}(i > 0)$ respectively denote the representation of the whole image and i_{th} local object image.

B. Representation Learning Module

This module aims to produce the noise-robust and consistent text-image representations with our proposed regularizers. Since these MI-based regularizers are intractable, variational encoding serves for the representation learning. In the following, we first introduce how the variational encoding works, and then detail the regularizers.

1) Variational Encoding: We devise an encoder in a variational manner for expressive text/image representation learning. Since the latent text/image representation is compressed within each modality, the encoder should explore the intra-modal information propagation. Considering the sufficient ability of Transformer [30] to propagation modeling, we deploy the encoder following the design of a Transformer Layer. Specifically, **we name the encoder as "Attentive Propagation Encoder"**, termed as $\mathrm{APEnc}(\boldsymbol{Q},\boldsymbol{K})$ with a query $\boldsymbol{Q}$ and key-value pair $\boldsymbol{K}$ as input, where a multi-head attention mechanism is first applied towards two variables as:

$$\begin{aligned}\mathrm{MultiHeadAtt}(\boldsymbol{Q},\boldsymbol{K}) &= \boldsymbol{W}'[\mathrm{CA}_1(\boldsymbol{Q},\boldsymbol{K}),\ldots,\mathrm{CA}_h(\boldsymbol{Q},\boldsymbol{K})]^{\mathrm{T}},\\ \mathrm{CA}_j(\boldsymbol{Q},\boldsymbol{K}) &= \mathrm{Softmax}\left(\frac{[\boldsymbol{Q}\boldsymbol{W}_{qj}][\boldsymbol{K}\boldsymbol{W}_{kj}]^{\mathbf{T}}}{\sqrt{d/h}}\right)[\boldsymbol{K}\boldsymbol{W}_{vj}],\end{aligned} \tag{4}$$

where CA_j refers to the j_{th} head of such multi-head attention, h is the number of heads. $\{\boldsymbol{W}_{qj}, \boldsymbol{W}_{kj}, \boldsymbol{W}_{vj}\} \in \mathbb{R}^{d\times(d/h)}$ and

$\mathbf{W}' \in \mathbb{R}^{d\times d}$ are all projections parameter matrices. Then, a fully-connected feed-forward network and a residual layer with layer-normalization are further stacked as follows:

$$\widetilde{\boldsymbol{F}} = \text{LayerNorm}(\boldsymbol{Q} + \text{MultiHeadAtt}(\boldsymbol{Q}, \boldsymbol{K}))$$
$$\boldsymbol{F} = \text{LayerNorm}(\widetilde{\boldsymbol{F}} + \text{FeedForward}(\widetilde{\boldsymbol{F}})), \quad (5)$$

where $\boldsymbol{F}$ is the final output representation of the encoder $\text{APEnc}(\boldsymbol{Q}, \boldsymbol{K})$ containing computation from (4) to (5).

Due to the intractability to MI-based regularizers, the encoder works in a variational manner. For simplicity, **we illustrate the variational encoding procedure of image representations** $\text{APEnc}^V(\boldsymbol{X}^V, \boldsymbol{X}^V)$ **as an example**. To promote the information propagation between images, each image $\boldsymbol{x}_i^v \in \mathbb{R}^{1\times d}$ is respectively taken as the query to interact with all images representations $\boldsymbol{X}^V \in \mathbb{R}^{(m+1)\times d}$ as key-value pairs. Specifically, the latent gaussian distributional variable $\boldsymbol{z}_i^v$ for each image is produced as follows:

$$\boldsymbol{\mu}_i^v = \text{APEnc}_\mu^V(\boldsymbol{x}_i^v, \boldsymbol{X}^V), \quad (6)$$
$$(\boldsymbol{\sigma}_i^t)^2 = \exp(\text{APEnc}_\sigma^V(\boldsymbol{x}_i^v, \boldsymbol{X}^V)) \quad (7)$$
$$\boldsymbol{z}_i^v \sim \mathcal{N}\left(\boldsymbol{\mu}_i^v, (\boldsymbol{\sigma}_i^v)^2\right), \quad (8)$$

where $\boldsymbol{\mu}_i^v \in \mathbb{R}^{1\times d}$ and $\boldsymbol{\sigma}_i^v \in \mathbb{R}^{1\times d}$ are the mean and variance vector of Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_i^v, (\boldsymbol{\sigma}_i^v)^2)$, and the representation $\boldsymbol{z}_i^v$ of the i_{th} image is sampled from the Gaussian distribution. However, since the sampling process is intractable for back-propagation, we adopt the reparameterization trick [20] as a solution. Mathematically, we first sample $\boldsymbol{\epsilon}$ from the normal Gaussian distribution, and then perform the equivalent sampling to derive $\boldsymbol{z}_i^v$ as follows:

$$\boldsymbol{z}_i^v = \boldsymbol{\mu}_i^v + \boldsymbol{\sigma}_i^v \odot \boldsymbol{\epsilon}, \ \ \boldsymbol{\epsilon} \sim \mathcal{N}(0, \text{diag}(\boldsymbol{I})) \quad (9)$$

where $\odot$ denotes the element-wise product operation. Similarly, we could obtain the representations for all images. Correspondingly, the visual representations $\boldsymbol{Z}^V \in \mathbb{R}^{(m+1)\times d}$ could be derived as follows:

$$\boldsymbol{Z}^V = (\boldsymbol{z}_0^v, \boldsymbol{z}_1^v, \ldots, \boldsymbol{z}_m^v). \quad (10)$$

The above procedure also holds in the case of text encoding via $\text{APEnc}^T(\boldsymbol{X}^T, \boldsymbol{X}^T)$**.** Similarly, we could obtain the textual representations $\boldsymbol{Z}^T \in \mathbb{R}^{(n+1)\times d}$ as follows:

$$\boldsymbol{Z}^T = (\boldsymbol{z}_0^t, \boldsymbol{z}_1^t, \ldots, \boldsymbol{z}_{n+1}^t). \quad (11)$$

2) Refinement-Regularizer (RR): For the modality-noise, the representations of text/images should **be diminished from the task-irrelevant information as much as possible while maximally maintain predictive for final tasks**. Considering that information-bottleneck (IB) could balance information from two variables, we intend to achieve this goal using IB-based regularizers. Specifically, we propose a refinement-regularizer L_{rr}, which consists of two IB-based terms respectively upon the representation learning of text $\boldsymbol{Z}^T$ and images $\boldsymbol{Z}^V$ as follows:

$$L_{rr} = \beta_1 I(\boldsymbol{Z}^T; \boldsymbol{X}^T) - I(\boldsymbol{Z}^T; \boldsymbol{R}) + \beta_2 I(\boldsymbol{Z}^V; \boldsymbol{X}^V) - I(\boldsymbol{Z}^V; \boldsymbol{R}), \quad (12)$$

where $\boldsymbol{X}^T$, $\boldsymbol{X}^V$ respectively denote the original information of input text and image, $\boldsymbol{R}$ refers to the ideal task-expected information for predictions. Considering the independence between $\boldsymbol{Z}^T$ and $\boldsymbol{Z}^V$ [24], the mutual information chain principle could be rewritten as:

$$I(\boldsymbol{Z}^T; \boldsymbol{R}) + I(\boldsymbol{Z}^V; \boldsymbol{R}) = I(\boldsymbol{Z}^T; \boldsymbol{R}|\boldsymbol{Z}^V) + I(\boldsymbol{Z}^V; \boldsymbol{R}) = I(\boldsymbol{Z}^T, \boldsymbol{Z}^V; \boldsymbol{R}). \quad (13)$$

Accordingly, L_{rr} in (12) could be simplified as follows:

$$L_{rr} = \underbrace{\beta_1 I(\boldsymbol{Z}^T; \boldsymbol{X}^T) + \beta_2 I(\boldsymbol{Z}^V; \boldsymbol{X}^V)}_{\text{Minimality}} - \underbrace{I(\boldsymbol{Z}^T, \boldsymbol{Z}^V; \boldsymbol{R})}_{\text{Reconstruction}}, \quad (14)$$

where L_{rr} could be optimized by (i) minimizing the former two terms so that the text/image representations are **compressed from their original modality apart from noises**; (ii) maximizing the reconstruction term to **encourage the predictive representations towards task predictions**. More specifically, L_{rr} refines the text/image representations which maintain evident but limit the disturbing noises. **For the tractable objective function of** L_{rr}**, please refer to Section III-D1 for details.**

3) Alignment-Regularizer (AR): For the modality-gap problem, the representations of text/images should be consistent with each other. We achieve this goal by **enhancing the mutual information (MI) between** $\boldsymbol{Z}^T$ **and** $\boldsymbol{Z}^V$ **for each input text-image pair**, since MI could inherently measure the association between two variables. Specifically, the MI-based alignment-regularizer (AR), L_{ar}, is devised as follows:

$$L_{ar} = -\underbrace{I(\boldsymbol{Z}^T; \boldsymbol{Z}^V)}_{\text{Maximality}}, \quad (15)$$

where L_{ar} is optimized by maximizing $I(\boldsymbol{Z}^T; \boldsymbol{Z}^V)$ for the agreement between representations of paired text/image. **For the tractable objective function of (15), please refer to Section III-D2.**

C. Prediction Module

In this module, we first conduct modality fusion to incorporate information from both modalities, and then perform task predictions for MNER and MRE, respectively.

1) Modality Fusion: To derive the image-aware text representations for the final task predictions, the information interaction between modalities is probed. **The above mentioned attentive propagation encoder (**APEnc**) is again leveraged here, but the key difference lies in that the query and key-value pairs are taken from different modalities**. To exploit the bi-directional interactions between modalities, APEnc works in a coupled way, where two encoders respectively take the image and text as query. Specifically, for $\boldsymbol{Z}^V$ and $\boldsymbol{Z}^T$, the first $\text{APEnc}^{\text{V2T}}$ aggregates information from texts, deriving the text-attended image representations $\boldsymbol{B} \in \mathbb{R}^{(m+1)\times d}$ as :

$$\boldsymbol{B} = \text{APEnc}^{\text{V2T}}(\boldsymbol{Z}^V, \boldsymbol{Z}^T). \quad (16)$$

Then, to enhance the textual representations with visual information, the second $\text{APEnc}^{\text{T2V}}$ derives the image-aware text

representations as follows:

$$\boldsymbol{C} = \text{APEnc}^{\text{T2V}}(\boldsymbol{Z}^T, \boldsymbol{B}), \tag{17}$$

where $\boldsymbol{C} = (\boldsymbol{c}_0, \boldsymbol{c}_1, \ldots, \boldsymbol{c}_n, \boldsymbol{c}_{n+1})$ is the image-aware text representations which are used for final task predictions.

2) MNER Prediction: Since we formulate MNER as a text-level sequence labeling problem, we utilize a standard Conditional Random Field (CRF) to derive the a corresponding label sequence $Y = (y_0, y_1, y_2, \ldots, y_n, y_{n+1})$ for entity recognition. Given the final image-aware text representations $\boldsymbol{C}$ of the sentence, we feed it into CRF to compute the probability to a label sequence Y as follows:

$$p(Y|X^T) = \frac{\exp\left(\sum_{i=0}^{n+1}\left(\boldsymbol{w}_{y_i}^{crf}\boldsymbol{c}_i + \text{Trans}(y_{i-1}, y_i)\right)\right)}{\sum_{Y' \in \boldsymbol{Y}} \exp\left(\sum_{i=0}^{n+1}(\boldsymbol{w}_{y_i}^{crf}\boldsymbol{c}'_i + \text{Trans}(y'_{i-1}, y'_i))\right)}, \tag{18}$$

where $\boldsymbol{w}_{y_i}^{crf}$ is a parameter vector computing the emission score from a token $\boldsymbol{c}_i$ to a label y_i, $\text{Trans}(y_{i-1}, y_i)$ is a learnable transition function from y_{i-1} to y_i, $\boldsymbol{Y}$ is the set of all possible label sequences.

3) MRE Prediction: MRE is formulated as a classification problem. We derive each annotated entity representation with pooling upon its contained tokens from $\boldsymbol{C}$, and further denote the representations of given entities E_1 and E_2 (explained in Section II-A2) as $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$. Assuming that the gold label Y is the k_{th} relation among all relation types, its corresponding probability is:

$$p(Y|X^T, E_1, E_2) = \text{Softmax}(\boldsymbol{W}([\boldsymbol{E}_1 \oplus \boldsymbol{E}_2]) + \boldsymbol{b})[k], \tag{19}$$

where $\oplus$ is the operation of concatenation, $\boldsymbol{W}$ and $\boldsymbol{b}$ are learnable model parameters.

D. Training Objective

The training objective to the MMIB consists of three components: 1) Objective function to the refinement-regularizer; 2) Objective function to the alignment-regularizer; 3) Objective function to the task predictions. In the following, we will detail them, respectively.

1) Refinement-Regularizer Objective Function: The optimization to L_{rr} in (14) is performed by minimizing the MI-based two terms and maximizing the reconstruction term.

Minimization to the MI-based minimality terms: Considering the challenge of exact computation to MI [31], we estimate these two terms based on a **variational upper bound** through Kullback-Leibler. Since the two terms, $I(\boldsymbol{Z}^T; \boldsymbol{X}^T), I(\boldsymbol{Z}^V; \boldsymbol{X}^V)$, share the similar forms, we take $I(\boldsymbol{Z}^V; \boldsymbol{X}^V)$ as an example to illustrate the process. Specifically, following (2), we first measure $I(\boldsymbol{Z}^V; \boldsymbol{X}^V)$ as follows:

$$I(\boldsymbol{Z}^V; \boldsymbol{X}^V) = \mathbb{D}_{KL}(\boldsymbol{p}_\theta^V(\boldsymbol{Z}^V|\boldsymbol{X}^V)||\boldsymbol{p}(\boldsymbol{Z}^V)). \tag{20}$$

Then, following prior variational works [24], [32], [33], the prior distribution $\boldsymbol{p}(\boldsymbol{Z}^V)$ could be estimated as normal Gaussian distribution $\mathcal{N}(0, \text{diag}(\boldsymbol{I}))$. For the the posterior distribution $\boldsymbol{p}_\theta^V(\boldsymbol{Z}^V|\boldsymbol{X}^V)$, it could be approximated by a variational posterior distribution $\boldsymbol{q}_\phi^V(\boldsymbol{Z}^V|\boldsymbol{X}^V)$. Accordingly, we could obtain the upper bound of $I(\boldsymbol{Z}^V; \boldsymbol{X}^V)$ as:

$$\begin{aligned} I(\boldsymbol{Z}^V; \boldsymbol{X}^V) &\leq \mathbb{D}_{KL}(\boldsymbol{q}_\phi^V(\boldsymbol{Z}^V|\boldsymbol{X}^V)||\boldsymbol{p}(\boldsymbol{Z}^V)) \\ &= \sum_{i=0}^{m+1} \mathbb{D}_{KL}\left(\mathcal{N}\left(\boldsymbol{\mu}_i^v, [\text{diag}(\boldsymbol{\sigma}_i^v)]^2\right)||\mathcal{N}\left(0, \text{diag}(\boldsymbol{I})\right)\right), \end{aligned} \tag{21}$$

where $\boldsymbol{q}_\phi^v$ refers to the parameters of variational encoding in (6). $I(\boldsymbol{Z}^V; \boldsymbol{X}^V)$ is optimized by minimizing its upper bound in (21) as the objective function. Similarly, the object function of $I(\boldsymbol{Z}^T; \boldsymbol{X}^T)$ could also be obtained by following the process from (20) to (21).

Maximization to the reconstruction term: The goal of the reconstruction term is to develop the ideal task-expected information $\boldsymbol{R}$ using the latent variables. We measure the quality of the reconstructed information via the probability to how the learned text/image representations could predict the true label as follows:

$$I(\boldsymbol{Z}^T, \boldsymbol{Z}^V; \boldsymbol{R}) \cong \mathbb{E}_{\boldsymbol{p}_\theta^T(\boldsymbol{Z}^T|\boldsymbol{X}^T)\boldsymbol{p}_\theta^V(\boldsymbol{Z}^V|\boldsymbol{X}^V)}[\log p(Y|\boldsymbol{Z}^T, \boldsymbol{Z}^V)], \tag{22}$$

where Y refers to the targets (i.e. the gold label sequence for MNER and gold relation for MRE). $\boldsymbol{p}_\theta^T(\boldsymbol{Z}^T|\boldsymbol{X}^T), \boldsymbol{p}_\theta^V(\boldsymbol{Z}^V|\boldsymbol{X}^V)$ are estimated via variational encoding $\boldsymbol{q}_\phi^T(\boldsymbol{Z}^T|\boldsymbol{X}^T)\boldsymbol{q}_\phi^V(\boldsymbol{Z}^V|\boldsymbol{X}^V)$ and we could obtain **the variational lower bound** of $I(\boldsymbol{Z}^T, \boldsymbol{Z}^V; \boldsymbol{R})$ as follows:

$$\begin{aligned} I(\boldsymbol{Z}^T, \boldsymbol{Z}^V; \boldsymbol{R}) &\geq \mathbb{E}_{\boldsymbol{q}_\phi^T(\boldsymbol{Z}^T|\boldsymbol{X}^T)\boldsymbol{q}_\phi^V(\boldsymbol{Z}^V|\boldsymbol{X}^V)}[\log p(Y|\boldsymbol{Z}^T, \boldsymbol{Z}^V)], \\ &= \log(p(Y|\boldsymbol{Z}^T, \boldsymbol{Z}^V)) + \sum_{Y' \in \boldsymbol{Y}, Y' \neq Y} \log(1 - p(Y'|\boldsymbol{Z}^T, \boldsymbol{Z}^V)) \end{aligned} \tag{23}$$

where $\boldsymbol{Y}$ refers to the label space of task predictions, Y refers to the true label while Y' as the false labels. To this end, we employ $p(\cdot|\cdot)$ in (18) and (19) respectively as $p(Y|\boldsymbol{Z}^T, \boldsymbol{Z}^V)$ for MNER and MRE. Hence, the reconstruction term could be optimized using the corresponding **task objective function in Section III-D3**.

2) Alignment-Regularizer Objective Function: The core of optimization to L_{ar} lies in the **maximization to the MI-based maximality term** in (15). We optimize L_{ar} following infomax [34], where a neural networks measure MI in a contrastive manner. Correspondingly, we build a discriminator $\mathcal{D}$ to measure the consistence degree between the text-image representations as follows:

$$I(\boldsymbol{Z}^T; \boldsymbol{Z}^V) = \mathbb{E}_{\boldsymbol{p}_\theta^T(\boldsymbol{Z}^T|\boldsymbol{X}^T)\boldsymbol{p}_\theta^V(\boldsymbol{Z}^V|\boldsymbol{X}^V)}[\log \mathcal{D}(\boldsymbol{Z}^T; \boldsymbol{Z}^V)]. \tag{24}$$

Since the direct optimization to MI-based L_{ar} is intractable, we utilize its **lower bound estimated via variational encoding** as the objective function as follows:

$$\begin{aligned} I(\boldsymbol{Z}^T; \boldsymbol{Z}^V) &\geq \mathbb{E}_{\boldsymbol{q}_\phi^T(\boldsymbol{Z}^T|\boldsymbol{X}^T)\boldsymbol{q}_\phi^V(\boldsymbol{Z}^V|\boldsymbol{X}^V)}[\log \mathcal{D}(\boldsymbol{Z}^T; \boldsymbol{Z}^V)] \\ &= \log\left(\mathcal{D}(\boldsymbol{Z}^T, \boldsymbol{Z}^{V+})\right) + \log\left(1 - \mathcal{D}(\boldsymbol{Z}^T, \boldsymbol{Z}^{V-})\right), \end{aligned} \tag{25}$$

where $\boldsymbol{q}_\phi^T$ / $\boldsymbol{q}_\phi^V$ denote the parameters of the variational encoding in Section III-B1, $(\boldsymbol{Z}^T, \boldsymbol{Z}^{V+})$ refers to the sampled text-image representations belonging to the same input pair, otherwise the in-batch negatives $(\boldsymbol{Z}^T, \boldsymbol{Z}^{V-})$ in contrastive mutual information. Since the shape of $\boldsymbol{Z}^T \in \mathbb{R}^{(n+2)\times d}$ and $\boldsymbol{Z}^V \in \mathbb{R}^{(m+1)\times d}$ does not match, we conduct mean-pooling to them for computation, namely:

$$\mathcal{D}(\boldsymbol{Z}^T, \boldsymbol{Z}^V) = \text{Sigmoid}(\text{MLP}(\text{Pooling}(\boldsymbol{Z}^T) \oplus \text{Pooling}(\boldsymbol{Z}^V))), \tag{26}$$

where $\text{Pooling}(\boldsymbol{Z}^T) \in \mathbb{R}^d$, $\text{Pooling}(\boldsymbol{Z}^V) \in \mathbb{R}^d$ and $\oplus$ denotes the concatenation operation between vectors. With (26), the tractable function could be optimized using a standard binary cross-entropy.

3) Task Objective Function: We introduce the task-specific objection function for MNER and MRE, respectively.

MNER: For MNER, given the sentence X^T and its golden sequence labels Y, we could obtain the probability of Y as $p(Y|X^T)$ from (18). Correspondingly, we could compute the negative log-likelyhood loss L_{ner} as follows:

$$L_{task}^{ner} = -\log(p(Y|X^T)). \tag{27}$$

MRE: For MRE, given the sentence X^T and the annotated head/tail entity E_1/E_2, we could obtain the probability $p(Y|X^T, E_1, E_2)$ from (19). The objection function for optimization could be derived as follows:

$$L_{task}^{re} = -\log(p(Y|X^T, E_1, E_2)). \tag{28}$$

4) Overall Objective Function: With the proposed regularizers and task objectives, the final objection function L could be written as follows:

$$\begin{aligned} L &= L_{rr} + L_{ar} + L_{task} \\ &= \beta_1 I(\boldsymbol{Z}^T; \boldsymbol{X}^T) + \beta_2 I(\boldsymbol{Z}^V; \boldsymbol{X}^V) - I(\boldsymbol{Z}^T, \boldsymbol{Z}^V; \boldsymbol{R}) \\ &\quad - I(\boldsymbol{Z}^T; \boldsymbol{Z}^V) \\ &\quad + L_{task}, \end{aligned} \tag{29}$$

where L could be optimized with the tractable functions (21) (L_{rr}), (26) (L_{ar}) and (27), (28) (L_{task}). β_1, β_2 are coefficients in IB (12) which penalize the learned representations from the noises.

IV. Experiments

A. Experimental Setup

1) Dataset: We conduct experiments on *Twitter-2015* [1] and *Twitter-2017* [6] for MNER, and *MNRE* [13] dataset for MRE. These datasets are collected from multimodal posts on Twitter, and each twitter post consists of one piece of text and a image. *Twitter-2015* and *Twitter-2017* both contain four types of entities: Person (**PER**), Location (**LOC**), Organization (**ORG**) and Miscellaneous (**MISC**). Table I shows the number of entities for each type and the division of multimodal tweets in the training, development, and test sets of the two dataset. The *MNRE* dataset contains 9,201 sentence-image pairs with 23 relation categories. Table II shows the detailed statistics and we compare it with the widely used relation extraction dataset SemEval-2010 Task 8 [35], reflecting the effectiveness of MNRE for task evaluation.

TABLE I
BASIC STATISTICS OF TWITTER-2015 AND TWITTER-2017

Entity Type	Twitter15			Twitter17		
	Train	**Dev**	**Test**	**Train**	**Dev**	**Test**
PER	2217	552	1816	2943	626	621
LOC	2091	522	1697	731	173	178
ORG	928	247	939	1674	375	395
MISC	940	225	726	701	150	157
Total	7176	1546	5078	6049	1324	1351
Tweets	4000	1000	3257	3373	723	723

TABLE II
STATISTICS OF MNRE COMPARED TO SEMEVAL-2010 DATASET

Dataset	Sentence	Entity	Relation	Image
MNRE	9,201	30,970	23	9,201
SemEval-2010	10,717	21,434	9	0

2) Metric: We adopt the same evaluation metric as our baselines. For **MNER**, an entity is correctly recognized when its span and entity type both match the gold answer. For **MRE**, the relation between a pair of entities is correctly extracted when the predicted relation type meets the gold answer. We utilize the evaluation code released by Chen et al. [7], where the Precision (**P**), Recall (**R**) and F1-score (**F1**) are used for performance evaluation.

3) Hyperparameters: We adopt BERT_{base} and ResNet50 as the textual and visual encoders for a fair comparision with previous study [7]. Accordingly, the hidden size of the all latent variables and representations are 768. Parameter optimization are performed using AdamW [36], where the decay is 0.01, the learning rate is 3e-5 and the batch size is 8. We manually tune the Lagrangian multiplier β_1, β_2 and achieve the best results with $\beta_1 = 0.1$ and $\beta_2 = 0.1$. The maximum length of the input sentence is 128 for MNER and 80 for MRE by cutting the longer ones and padding the shorter ones. The model implementation code has been released for reproductivity check.

B. Baselines

For a comprehensive comparison, we compare our method with four groups of baselines.

Text-based baselines: To verify the effectiveness of introducing visual images, we choose the baselines which conduct the both tasks without incorporating the visual information. The **NER** baselines contain 1) **CNNBiLSTM-CRF** [37] utilizes word- and character-level representations via BiLSTM and CNN for NER; 2) **HBiLSTMCRF** [38] replaces the CNN in CNNBiLSTM-CRF with an LSTM; 3) **T-NER** [39] is a Twitter-specific NER system with various features to boost performance, and we refer to its results on our used datasets from Xu et al. [12]. The **RE** baselines involve 1) **PCNN** [40] uses convolutional networks with piecewise pooling; 2) **MTB** [41] is a RE-oriented pre-training model based on BERT.

Vanilla multimodal baselines: To confirm the necessity of solving modality-noise and modality-gap, we choose vanilla multimodal baselines which ignore these two issues. Specifically, the **MNER** baselines contain 1) **GVATT** [6] utilizes attention mechanism to combine the image- and text-level information; 2) **AdapCoAtt** [1] designs an adaptive co-attention network to explore visual information; 3) **UMT** [11] proposes a multimodal interaction module to obtain image-aware word representations; 4) **FMIT** [16] designs a flat multi-modal interaction transformer (FMIT) layer where a unified lattice structure is used for cross-modality interaction. We cite the performance of FMIT with one FMIT layer for a fair comparison with MMIB. For **MRE**, it contains 1) **VisualBERT** [42], where Chen et al. [7] utilize the pre-trained multimodal model for text-image encoding and fusion. 2) **BERT-SG** [13] directly incorporates visual features extracted using a Scene Graph (SG) Tool [43].

Multimodal baselines considering modality-gap: We choose baselines which derive the consistent text-image representations. For **MNER**, the baselines include 1) **UMGF** [14] builds a multimodal graph for semantic alignment. 2) **MAF** [12] employs contrastive learning method for consistent representations. 3) **DebiasCL** [44] derives the shared text-image semantic space via de-biased contrastive learning. For **MRE**, **MEGA** [13] develops a dual graph for semantic agreement.

Multimodal baselines considering modality-noise: To validate the superiority of MMIB for tackling modality-noise, we choose baselines which attempt to allevite the issue. For **MNER**, the baselines involve 1) **M3S** [18] cascades the MNER tasks with Named Entity Segmentation and Named Entity Categorization to alleviate the visual bias. 2) **HVPNet** [7] alleviates the noises of irrelevant visual objects by exploring the hierarchical visual features as pluggable visual prefix. For **MRE**, **MKGformer** [9] utilizes a correlation-aware fusion module to alleviate the noisy information.

Note that the original results of UMT, UMGF and MEGA only involve one task, and we refer to their results from Chen et al. [7].

C. Main Results

Tables III and IV respectively show the final model performances upon MNER and MRE, with significant difference ($p < 0.05$) between MMIB and all baselines. Reading from the resutls, we have observations as follows.

1) *Visual information indeed helps to boost performances:* As multimodal methods generally outperform the text-based ones upon two tasks, we could see the necessarity of incorporating visual information for semantics enhancement. However, the performance improvements are still limitted, which demands further exploration to multimodal methods.
2) *Bridging the modality-gap could improve the model performances:* Multimodal baselines considering modality-gap advantage the vanilla ones generally, revealing the importance of the consistent text-image representations.

TABLE III
OVERALL MNER RESULTS

Method	Twitter15			Twitter17		
	P	**R**	**F1**	**P**	**R**	**F1**
CNN-BiLSTM-CRF	66.24	68.09	67.15	80.00	78.76	79.37
HBiLSTM-CRF	70.32	68.05	69.17	82.69	78.16	80.37
T-NER	69.54	68.65	69.09	-	-	-
GVATT	73.96	67.90	70.80	83.41	80.38	81.87
AdapCoAtt	72.75	68.74	70.69	84.16	80.24	82.15
UMT	71.67	75.23	73.41	85.28	85.34	85.31
FMIT	74.18	75.03	74.60	85.55	85.29	85.42
UMGF	**74.49**	75.21	74.85	86.54	84.50	85.51
MAF	71.86	75.10	73.42	86.13	86.38	86.25
DebiasCL	74.45	76.13	75.28	87.59	86.11	86.84
M3S	74.92	75.14	75.03	86.93	85.21	86.06
HVPNeT	73.87	76.82	75.32	85.84	87.93	86.87
MMIB	74.44	**77.68**	**76.02**	**87.34**	87.86	**87.60**

TABLE IV
OVERALL MRE PERFORMANCES

Method	MRE		
	P	**R**	**F1**
PCNN	62.85	49.69	55.49
MTB	64.46	57.81	60.86
AdapCoAtt	64.67	57.98	61.14
VisualBERT	57.15	59.48	58.30
BERT+SG	62.95	62.65	62.80
UMT	62.93	63.88	63.46
UMGF	64.38	66.23	65.29
MEGA	64.51	68.44	66.41
HVPNeT	**83.64**	80.78	81.85
MKGformer	82.67	81.25	81.95
MMIB	83.49	**82.97**	**83.23**

3) *Alleviating the modality-noise could bring performance gain to multi-modal methods:* Multimodal baselines considering modality-noise outperform the vanilla multimodal ones, reflecting the importance of noise mitigation. Despite of this, the performances upon two tasks are still barely satisfactory, which calls for further investigation to this issue.
4) *Our proposed MMIB achieves the best results upon two tasks:* As the two tables show, MMIB outperforms the prior multimodal baselines considering the modality-gap and modality-noise problems. This demonstrates the effectiveness of our method to handle the two issues. We attribute such performance advantage to that the refinement-regularizer and alignment-regularizer could effectively derive the "good" text-image representations for the final task predictions.

V. ANALYSIS AND DISCUSSION

A. Ablation Study

To probe how each regularizer contributes to the final performances, we conduct an ablation study upon two tasks in three

TABLE V
ABLATION STUDY

Ablation	Dataset	Metrics			
		P	R	F1	Δ F1
MMIB	Twitter15	74.44	77.68	76.02	-
	Twitter17	87.34	87.86	87.60	-
	MRE	83.49	82.97	83.23	-
$-L_{rr}$	Twitter15	73.36	76.61	74.95	-1.07
	Twitter17	86.67	87.56	87.11	-0.49
	MRE	82.10	81.71	81.91	-1.32
$-L_{ar}$	Twitter15	74.05	75.75	74.89	-1.13
	Twitter17	88.07	85.79	86.91	-0.69
	MRE	82.48	80.93	81.70	-1.53
$-L_{rr}$ & L_{ar}	Twitter15	74.02	74.81	74.42	-1.60
	Twitter17	87.15	85.86	86.50	-1.10
	MRE	80.62	80.00	80.31	-2.92

benchmarks. Following prior studies [7], [12], [14], we report the ablation performances in the test set in Table V. Specifically, we have analysis as follows.

1) *Refinement-regularizer (L_{rr}):* Removing the refinement-regularizers degrades the final performances upon three benchmarks, revealing the importance of limiting task-relevant noisy information. Without L_{rr}, the noisy information in both modalities may disturb the incorporation to evident visual information, hurting the performances.
2) *Alignment-regularizer (L_{ar}):* When the alignment-regularizer is removed, the model performances decline upon all benchmarks. This phenomenon indicates the necessarity of bridging the modality-gap and verifies the effectiveness of L_{ar} to drive the consistent text-image representations.
3) L_{rr} & L_{ar}: The final performances severely drop when we simultaneously remove all the regularizers. This reveals that the both regularizers are functional and work with each other collaboratively to boost the final performances.

B. Analysis for Modality-Noise

To confirm that MMIB derives text/image representations which dismisses from redundant noises and maintain task-predictive, we present an visualization of two cases. To quantify how much information the original representations $\mathbf{X}^T/\mathbf{X}^V$ attend to the learned representations $\mathbf{Z}^T/\mathbf{Z}^V$, we define a contribution-score to measure it. We illustrate the computation to contribution-score using the visual representations as an example. Specifically, we first derive a matrix $\mathbf{A}^V \in \mathbb{R}^{(m+1)\times(m+1)}$ by dot-product between $\mathbf{X}^V/\mathbf{Z}^V$. Then, we respectively sum each row of $\mathbf{A}^V$ obtaining $\hat{\mathbf{A}}^V \in \mathbb{R}^{m+1}$, where $\hat{\mathbf{A}}^V_i$ is a scalar indicating how the i_{th} image contributes to final visual representations $\mathbf{Z}^V$. Following this process, we compute $\hat{\mathbf{A}}^V$, $\hat{\mathbf{A}}^T$ and visualize them.

The visualization to MMIB and MMIB w/o L_{rr} are respectively deployed in the color map of red and green. As Fig. 3(a) shows for MRE, the task-irrelevant information ("RT @Del: Tune in to") are weakened in MMIB. For the visual modality, in MMIB w/o L_{rr}, each image evenly attends to the final

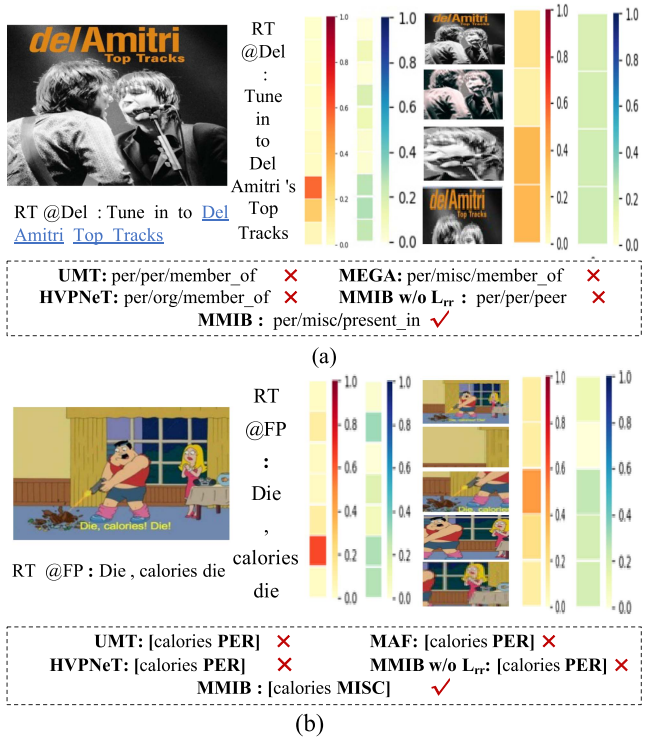

Fig. 3. Case study for modality-noise: (a) A case for MRE task; (b) a case for MNER task.

visual representations, which makes the attention mechanism in modality fusion wrongly incorporate the visual information of two singers, resulting in the prediction to `per/per/peer`. Meanwhile, such task-irrelevant visual information are dismissed in MMIB, facilitating the task prediction. For MNER in Fig. 3(b), we observe that the task-irrelevant information (i.e. the textual "RT @FP" and visual background in the image) attends little to the final image representations, where the redundant noises in each modality are dimissed. With merit of L_{rr}, MMIB makes the correct prediction for these two cases.

Besides, we compare MMIB with some baselines for performance analysis. Specifically, UMT, one of vanilla multimodal baselines, predicts the relation as `peer` between persons in MNRE and assigns the entity type `PER` to "calories" in MNER. Since UMT conducts the cross-modality interaction with all the grids in the image via the attention mechanism, the noisy grids could disturb the attention assignments and thus lead to the wrong predictions. For MEGA, the noisy information could distract the semantic correlations between the textual entities and visual objects. MAF considers the relevance between the image and text globally, but the lack of fine-grained noise filter still produces the wrong prediction. Though HVPNet attempts to alleviate the visual noises by visual prefixes, it ignores the textual noisy semantics. For example in Fig. 3(b), the token "die" internally maintain semantics about a person, thus the visual prefix is misled by the salient person objects. As a result, HVPNet fails to perform the correct predictions.

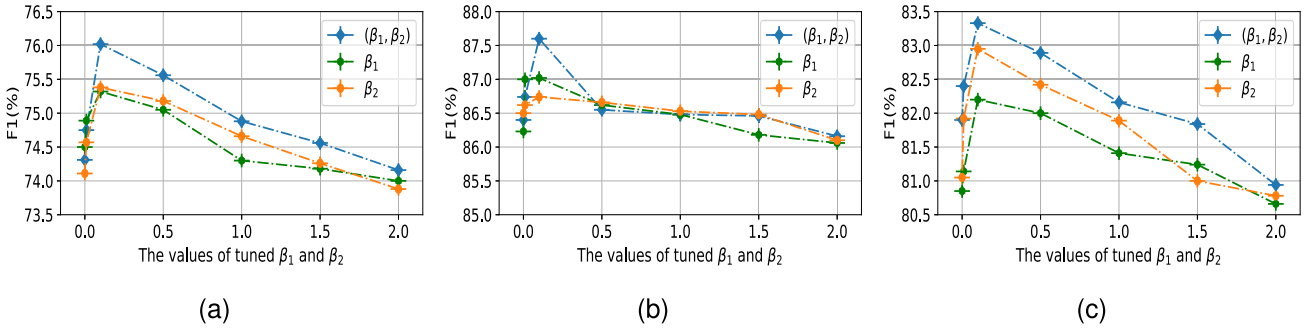

Fig. 4. Change F1 for our model with different values of β_1 and β_2 under different datasets: (a) The change of F1 on Twitter15 dataset; (b) the change of F1 on Twitter17; and (c) the change of F1 on MRE dataset.

C. Analysis to the Information Bottleneck

To analyze how the information bottleneck principle works, we explore how the Lagrangian multipliers, β_1, β_2 in L_{rr}, influence the final performances. We employ $\{0, 0, 0.01, 0.1, 0.5, 1.0, 1.5, 2.0\}$ as the candidate values of β_1, β_2 and perform three series of experiments: 1) We fix $\beta_1 = 1$ and tune β_2 with the candidate values; 2) We fix $\beta_2 = 1$ and tune β_1 with the candidate values; 3) We tune β_1, β_2 simultaneously and keep their values as the same value. The the fluctuation of $F1$ performances towards the experiments above are respectively presented with the orange, green and blue curves in Fig. 4.

Reading from Fig. 4, we could see that the best performances are achieved when both β_1 and β_2 are set as 0.1. Specifically, we have observations and analysis as follows. 1) Tuning β_1, β_2 simultaneously could generally achieve better performances than tuing them individually, which demonstrates that these two IB-terms in L_{rr} are collaborative with each other. 2) The performances degrade severely when β_1 or β_2 is set as 0.0. Such a phenomenon manifests the importance of discarding noisy information from the original textual and visual information. 3) The worst performances are achieved when β_1 or β_2 is set as 2.0. The reason might be that the model compress the original textual and visual information too much to reconstruct the task-expected information. 4) The best performances are achieved with $\beta_1 = 0.1, \beta_2 = 0.1$, which reveals that the noises in both modalities do not overweight the valid information. Even so, it is important to discard the noises and control the trade-off between the noisy and valid information for predictions.

D. Visualization for Modality-Gap

To confirm that MMIB produces the consistent text-image representations, we perform a text-image representation visualization. Specifically, we first randomly choose 150 pairs of entity-objects regarding PER, LOC and ORG types from the input text-image pairs, and gather their representations produced by trained MMIB with/without the alignment regularizer. Then, t-SNE [45] serves to visualize their representations, where it is trained 500 iterations with perplexity of 10. We deploy the results in Twitter17 dataset as an example and present the results in Fig. 5, where the three rows are respectively visualization to entity/objects of PER, LOC and ORG type. Without the

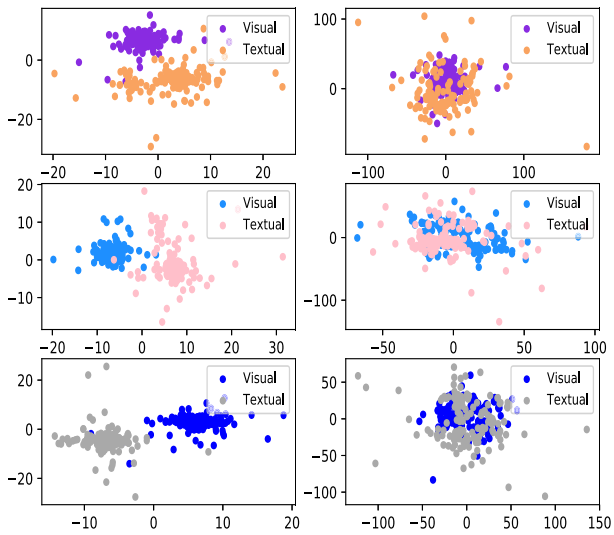

Fig. 5. Representation visualization to the issue of modality-gap.

alignment-regularizer, the entity/object representations respectively scatter in their own spaces as the left columns in Fig. 5 shows. On the contrary, the right column illustrates representation visualization obtained from the complete MMIB. We could see that the alignment-regularizer could achieve the semantic agreements between textual and visual representations. This phenomenon manifests the effectiveness of the alignment-regularizer for consistent text-image representation learning.

E. Error Analysis

To analyze the potential weakness of MMIB, we conduct an error analysis. Specifically, we respectively select 100 random mistakenly predicted instances from three datasets, and observe that MMIB struggles when the content of text and image are obviously irrelevant. Fig. 6 shows some typical instances. For MNER in Fig. 6(a), the entity of interest "Celtic" refers to a foolball team of ORG-type. However, the salient object in the image is a billboard without obvious clues about an organization-type football team. With the visual clues of the billboard, MMIB recognizes "Celtic" as MISC-type. The similar situation also happens in Fig. 6(b) for MRE. Specifically, the

Fig. 6. Error Analysis for the proposed MMIB model: (a) An instance for MNER task; (b) an instance for MRE task.

sentence depicts the "tour" in "New York City", while the image poses objects about a person and a house. Here, no obvious semantic relevance is presented between the content of text and image. Hence, guided by such visual information, MMIB predicts `per/loc/place_of_residence` as the relation between given entities.

We attempt to infer the reasons for such phenomenon. Despite that MMIB could discard the noisy information, some noisy information would be inevitably introduced when the content of text-image are obviously irrelevant. We think that this points out an interesting direction and could inspire more future works to improve MRE and MNER.

VI. Related Work

A. Multimodal Entity and Relation Extraction

Multimodal named entity recognition (**MNER**) and multimodal relation extraction (**MRE**) have raised great research interests due to the increasing amount of the user-generated multimodal content. Existing studies could be grouped into two lines. The **first** line of studies directly incorporate the whole image as global visual clues to enhance the textual semantics. The image is encoded as either one feature vector [1], [6], [10] or multiple vectors corresponding to grids [2], [8], [11], [12], [46], [47]. Correspondingly, attention-based mechanisms are designed to promote the text-image interaction, deriving expressive textual representations for the final tasks. However, these methods struggle to build the fine-grained mapping between visual objects and named entities. The **second** line of studies [4], [5] explore the fine-grained object-level visual information to boost the model performances. Specifically, these studies employ the pretrained object detection model to extract the salient objects in the image, where the visual information are acquired from both the global image and local object image. Apart from attention mechanisms [4], [5], various strategies are proposed to sufficiently explore the text-image interaction. For example, Zheng et al. [13] and Zhang et al. [14] utilized multi-modal graph structure to capture the semantic correlation between texts and visual objects. Lu et al. [16] designed a unified lattice structure for cross-modal interaction. Wang et al. [18] develop scene graphs as a structured representation of the visual contents for semantic interaction. To provide prior information about entity types and image regions, Jia et al. [17] and Jia et al. [15] proposed machine reading comprehension (MRC) based methods to promote the cross-modality interaction. Though the studies above have achieved great success, modality-noise and modality-gap are two inadequately addressed issues. While most previous works ignore the modality-noise problem, Chen et al. [7] attempt to prevent visual information from irrelevant objects via hierarchical visual features. However, we focus on noises in both modalities and aim to derive representations which are not only noise-robust but also predictive enough for downstream tasks. For the issue of modality-gap, though it could be bridged by converting the image or objects to their textual descriptions [5], [13], [19], such operation suffers from the error-propagation from external tools. Different from prior works, we attempt to solve these two issues simultaneously via representation learning from the information-theoretic perspective, which is more generalizable and adaptive to other tasks.

B. Information Bottleneck

Information bottleneck (IB) [25] is an important concept in information theory and has attached great research attention. IB is adapted into deep neural networks with variational inference, and thus is called variational information bottleneck (VIB) [21]. VIB could serve as a regularization technique for representation learning and has shown its sparkles in various fields of computer vision [33], [48], natural language processing [23], [49], [50] and recommendation systems [24]. Despite of the great achievements above, VIB has not been explored in MNER and MRE yet. In this article, we adopt VIB to derive the task-expected and noise-robust representations for MNER and MRE, and achieve great model performances upon three public benchmarks.

VII. Conclusion

In this article, we propose a novel approach, MMIB, for multimodal named entity recognition (MNER) and multimodal relation extraction (MRE). MMIB aims to tackles the issues of modality-noise and modality-gap in these two tasks via representation learning with information theory. Specifically, for the first issue, a refinement-regularizer, which explores the information-bottleneck principle, is devised to draw text-image representations which are diminished from the redundant noises and retain predictive for final tasks. For the second issue, an alignment-regularizer, which contains an MI-term, is proposed to derive the consistent cross-modality representations. Experiments on three benchmarks show that MMIB significantly outperforms state-of-the-art baselines. In the future, we will adapt MMIB for other multimodal tasks for multimedia analysis.

References

[1] Q. Zhang, J. Fu, X. Liu, and X. Huang, "Adaptive co-attention network for named entity recognition in tweets," in *Proc. 32nd AAAI Conf.*, 2018, pp. 5674–5681.

[2] C. Zheng, Z. Wu, J. Feng, Z. Fu, and Y. Cai, "MNRE: A challenge multimodal dataset for neural relation extraction with visual evidence in social media posts," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2021, pp. 1–6.

[3] L. Yuan, Y. Cai, J. Wang, and Q. Li, "Joint multimodal entity-relation extraction based on edge-enhanced graph alignment network and word-pair relation tagging," in *Proc. 32nd AAAI Conf.*, 2023, pp. 11051–11059.
[4] Z. Wu, C. Zheng, Y. Cai, J. Chen, H. Leung, and Q. Li, "Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1038–1046.
[5] C. Zheng, Z. Wu, T. Wang, Y. Cai, and Q. Li, "Object-aware multimodal named entity recognition in social media posts with adversarial learning," *IEEE Trans. Multimedia*, vol. 23, pp. 2520–2532, 2021.
[6] D. Lu, L. Neves, V. Carvalho, N. Zhang, and H. Ji, "Visual attention model for name tagging in multimodal social media," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 1990–1999.
[7] X. Chen et al., "Good visual guidance make a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction," in *Proc. Findings Assoc. Comput. Linguistics*, 2022, pp. 1607–1618.
[8] B. Xu et al., "Different data, different modalities! reinforced data splitting for effective multimodal information extraction from social media posts," in *Proc. 29th Int. Conf. Comput. Linguistics*, 2022, pp. 1855–1864.
[9] X. Chen et al., "Hybrid transformer with multi-level fusion for multimodal knowledge graph completion," in *Proc. 45th Int. ACM SIGIR*, 2022, pp. 904–915.
[10] S. Moon, L. Neves, and V. Carvalho, "Multimodal named entity recognition for short social media posts," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics*, 2018, pp. 852–860.
[11] J. Yu, J. Jiang, L. Yang, and R. Xia, "Improving multimodal named entity recognition via entity span detection with unified multimodal transformer," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3342–3352.
[12] B. Xu, S. Huang, C. Sha, and H. Wang, "MAF: A general matching and alignment framework for multimodal named entity recognition," in *Proc. 15th ACM Int. Conf. Web Search Data Mining*, 2022, pp. 1215–1223.
[13] C. Zheng, J. Feng, Z. Fu, Y. Cai, Q. Li, and T. Wang, "Multimodal relation extraction with efficient graph alignment," in *Proc. ACM Multimedia Conf.*, 2021, pp. 5298–5306.
[14] D. Zhang, S. Wei, S. Li, H. Wu, Q. Zhu, and G. Zhou, "Multi-modal graph fusion for named entity recognition with targeted visual guidance," in *Proc. 35th AAAI Conf.*, 2021, pp. 14347–14355.
[15] M. Jia et al., "Query prior matters: A MRC framework for multimodal named entity recognition," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 3549–3558.
[16] J. Lu, D. Zhang, J. Zhang, and P. Zhang, "Flat multi-modal interaction transformer for named entity recognition," in *Proc. 29th Int. Conf. Comput. Linguistics*, 2022, pp. 2055–2064.
[17] M. Jia et al., "MNER-QG: An end-to-end MRC framework for multimodal named entity recognition with query grounding," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 8032–8040.
[18] J. Wang, Y. Yang, K. Liu, Z. Zhu, and X. Liu, "M3S: Scene graph driven multi-granularity multi-task learning for multi-modal NER," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 111–120, 2023.
[19] X. Wang et al., "ITA: Image-text alignments for multi-modal named entity recognition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2022, pp. 3176–3189.
[20] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. 2nd Int. Conf. Learn. Representations*, 2014.
[21] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," in *Proc. 5th Int. Conf. Learn. Representation*, 2017.
[22] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.
[23] C. Zhang, X. Zhou, Y. Wan, X. Zheng, K. Chang, and C. Hsieh, "Improving the adversarial robustness of NLP models by information bottleneck," in *Proc. Findings Assoc. Computat. Linguistics*, 2022, pp. 3588–3598.
[24] J. Cao, J. Sheng, X. Cong, T. Liu, and B. Wang, "Cross-domain recommendation to cold-start users via variational information bottleneck," in *Proc. IEEE 38th Int. Conf. Data Eng.*, 2022, pp. 2209–2223.
[25] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Proc. IEEE Inf. Theory Workshop*, 2015, pp. 1–5.
[26] S. Gershman and N. D. Goodman, "Amortized inference in probabilistic reasoning," in *Proc. 36th Annu. Meeting Cogn. Sci. Soc.*, 2014.
[27] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 4171–4186.
[28] Z. Yang, B. Gong, L. Wang, W. Huang, D. Yu, and J. Luo, "A fast and accurate one-stage approach to visual grounding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4682–4692.
[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[30] A. Vaswani et al., "Attention is all you need," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[31] P. Colombo, P. Piantanida, and C. Clavel, "A novel estimator of mutual information for learning to disentangle textual representations," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics*, 2021, pp. 6539–6550.
[32] I. Higgins et al., "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. 5th Int. Conf. Learn. Representations*, 2017.
[33] T. Q. Chen, X. Li, R. B. Grosse, and D. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2018, pp. 2615–2625.
[34] P. Velickovic, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, "Deep graph infomax," in *Proc. 7th Int. Conf. Learn. Representations*, 2019.
[35] I. Hendrickx et al., "SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals," 2019, *arXiv:1911.10422*.
[36] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2019.
[37] X. Ma and E. H. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNS-CRF," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2016, pp. 1064–1074, doi: 10.18653/v1/P16-1101.
[38] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural Architectures for Named Entity Recognition," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics*, 2016, pp. 260–270.
[39] A. Ritter, S. Clark, Mausam, and O. Etzioni, "Named entity recognition in tweets: An experimental study," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2011, pp. 1524–1534.
[40] D. Zeng, K. Liu, Y. Chen, and J. Zhao, "Distant supervision for relation extraction via piecewise convolutional neural networks," in *Proc. Empirical Methods Natural Lang. Process.*, 2015, pp. 1753–1762.
[41] L. B. Soares, N. FitzGerald, J. Ling, and T. Kwiatkowski, "Matching the blanks: Distributional similarity for relation learning," in *Proc. Assoc. Comput. Linguistics*, 2019, pp. 2895–2905.
[42] I. Hendrickx et al., "SemEval-2010 Task 8: Multi-Way classification of semantic relations between pairs of nominals," in *Proc. 5th Int. Workshop Semantic Eval.*, Jul. 2010, pp. 33–38. [Online]. Available: https://aclanthology.org/S10-1006
[43] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, "Unbiased scene graph generation from biased training," in *Proc. IEEE/CVF Comput. Vis. Pattern Recognit.*, 2020, pp. 3713–3722.
[44] X. Zhang, J. Yuan, L. Li, and J. Liu, "Reducing the bias of visual objects in multimodal named entity recognition," in *Proc. 16th ACM Int. Conf. Web Search Data Mining*, 2023, pp. 958–966. doi: 10.1145/3539597.3570485.
[45] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Proc/ Adv. Neural Inf. Process. Syst.*, 2002, pp. 833–840.
[46] L. Sun, J. Wang, K. Zhang, Y. Su, and F. Weng, "RpBERT: A text-image relation propagation-based BERT model for multimodal NER," in *Proc. 35th AAAI Conf.*, 2021, pp. 13860–13868.
[47] S. Chen, G. Aguilar, L. Neves, and T. Solorio, "Can images help recognize entities? A study of the role of images for multimodal NER," in *Proc. 7th Workshop Noisy User-Generated Text*, 2021, pp. 87–96.
[48] S. Motiian and G. Doretto, "Information bottleneck domain adaptation with privileged information for visual recognition," in *Proc. Comput. Vis. 14th Eur. Conf.*, 2016, pp. 630–647.
[49] J. Zhou, Y. Wu, Q. Chen, X. Huang, and L. He, "Attending via both fine-tuning and compressing," in *Proc. Findings Assoc. Comput. Linguistics*, pp. 2152–2161, 2021.
[50] J. Zhou, Q. Zhang, Q. Chen, L. He, and X. Huang, "A multi-format transfer learning model for event argument extraction via variational information bottleneck," in *Proc. 29th Int. Conf. Comput. Linguistics*, 2022, pp. 1990–2000.

Shiyao Cui received the Ph.D. degree from the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. She is currently a research Assistant with the Institute of Information Engineering, Chinese Academy of Sciences. Her research interests include large language models and knowledge graph.

Jiangxia Cao received the Ph.D. degree from the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. He is currently an algorithm Engineer with Kuaishou Technology, Beijing. His research focuses on large language models for recommendation system.

Xin Cong received the Ph.D. degree from the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. He is currently a Postdoctor with THUNLP Lab, Tsinghua University, Beijing. His research interests include large lanuage models, tool learning, and autonomous agent.

Jiawei Sheng received the Ph.D. degree from the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. He is currently a tenure-track associate Researcher with the Institute of Information Engineering, Chinese Academy of Sciences. His research interests include information extraction, knowledge acquisition, and data mining.

Quangang Li received the Ph.D. degree from the University of Electronic Science and Technology of China, Chengdu, China. He is currently an Associate Professor of engineering with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. He research interests include natural language processing and intelligent systems.

Tingwen Liu received the Ph.D. degree in information security from the Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, China, in 2013. He is currently a Professor with the Institute of Information Engineering, CAS and University of CAS, Beijing. His research interests include knowledge graph, natural language processing, and network content security.

Jinqiao Shi received the Ph.D. degree from the Harbin Institute of Technology, Harbin, China. He is currently a Professor with the School of Cyber Security, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include network measurement, intelligent information processing, and Big Data security analysis.