

# Information Screening whilst Exploiting! Multimodal Relation Extraction with Feature Denoising and Multimodal Topic Modeling

Shengqiong Wu<sup>1</sup>, Hao Fei<sup>1\*</sup>, Yixin Cao<sup>2</sup>, Lidong Bing<sup>3</sup>, Tat-Seng Chua<sup>1</sup>

<sup>1</sup> Sea-NExT Joint Lab, School of Computing, National University of Singapore

<sup>2</sup> Singapore Management University, <sup>3</sup> DAMO Academy, Alibaba Group

swu@u.nus.edu haofei37@nus.edu.sg caoyixin2011@gmail.com

l.bing@alibaba-inc.com dcscts@nus.edu.sg

## Abstract

Existing research on multimodal relation extraction (MRE) faces two co-existing challenges, *internal-information over-utilization* and *external-information under-exploitation*. To combat that, we propose a novel framework that simultaneously implements the idea of *internal-information screening* and *external-information exploiting*. First, we represent the fine-grained semantic structures of the input image and text with the visual and textual scene graphs, which are further fused into a unified cross-modal graph (CMG). Based on CMG, we perform structure refinement with the guidance of the graph information bottleneck principle, actively denoising the less-informative features. Next, we perform topic modeling over the input image and text, incorporating latent multimodal topic features to enrich the contexts. On the benchmark MRE dataset, our system outperforms the current best model significantly. With further in-depth analyses, we reveal the great potential of our method for the MRE task. Our codes are open at <https://github.com/ChocoWu/MRE-ISE>.

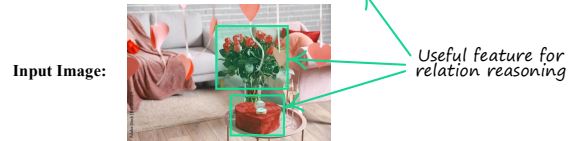
## 1 Introduction

Relation extraction (RE), determining the semantic relation between a pair of subject and object entities in a given text (Yu et al., 2020), has played a vital role in many downstream natural language processing (NLP) applications, e.g., knowledge graph construction (Wang et al., 2019; Mondal et al., 2021), question answering (Cao et al., 2022). But in realistic scenarios (i.e., social media), data is often in various forms and modalities (i.e., texts, images), rather than pure texts. Thus, multimodal relation extraction has been introduced recently (Zheng et al., 2021b), where additional visual sources are added to the textual RE as an enhancement to the relation inference. The essence of a successful MRE lies in the effective utilization of multimodal

\*Corresponding author: Hao Fei

### ➤ Example #1

Input Text: Congratulations to Angela and Mark Salmons, a new life ahead is waiting!



### ➤ Example #2

Input Text: Yessir dropping my first single "Hot summer" with my brothers Migos.



Figure 1: Examples of multimodal relation extraction (MRE). In texts, the subject and object entities are underlined, and linked by the relational tags.

information. Certain efforts have been made in existing MRE work and achieved promising performances, where delicate interaction and fusion mechanisms are designed for encoding the multimodal features (Zheng et al., 2021a; Chen et al., 2022b,a). Nevertheless, current methods still fail to sufficiently harness the feature sources from two information perspectives, which may hinder further task development.

**Internal-information over-utilization.** On the one hand, most existing MRE methods progressively incorporate full-scale textual and visual sources into the learning, under the assumption that all the input information certainly contributes to the task. In fact, prior textual RE research extensively shows that only parts of the texts are useful to the relation inference (Yu et al., 2020), and accordingly propose to prune over input sentences (Zhang et al., 2018). The case is more severe for the visual inputs, as not all and always the visual sources play positive roles, especially on the social media data. As revealed by Vempala and Preoțiu-Pietro (2019), as high as 33.8% of visual information serves no context or even noise in MRE. Xu et al. (2022) thus propose to selectively remove im-

ages from the input image-text pairs. Unfortunately, such coarse-grained instance-level filtering largely hurts the utility of visual features. We argue that a fine-grained feature screening over both the internal image and text features is needed. Taking the example #1 in Fig. 1, the textual expressions ‘*Congratulations to Angela and Mark Salmons*’ and the visual objects of ‘*gift*’ and ‘*roses*’ are valid clues to infer the ‘*couple*’ relation between ‘*Angela*’ and ‘*Mark Salmons*’, while the rest of text and visual information is essentially the task-irrelevant noise.

**External-information under-exploitation.** On the other hand, although compensating the text inputs with visual sources, there can be still information deficiency in MRE, in particular when the visual features serve less (or even negative) utility. This is especially the case for social media data, where the contents are less-informative due to the short text lengths and low-relevant images (Baly et al., 2020). For the example #2 in Fig. 1, due to the lack of necessary contextual information, it is tricky to infer the relation ‘*present in*’ between ‘*Hot summer*’ (an album name) and ‘*Migos*’ (a singer name) based on both the image and text sources. In this regard, more external information should be considered and exploited for MRE. Fortunately, the topic modeling technique offers a promising solution, which has been shown to enrich the semantics of the raw data, and thus facilitate NLP applications broadly (Zeng et al., 2018). For the above same example, if an additional ‘*music*’ topic feature is leveraged into the context, the relation inference can be greatly eased.

Taking into account the above two observations, in this work, we propose a novel framework to improve MRE. As shown in Fig. 4, we first employ the scene graphs (SGs) (Johnson et al., 2015) to represent the input vision and text, where SGs advance in intrinsically depicting the fine-grained semantic structures of texts or images. We fuse both the visual and textual SGs into a cross-modal graph (CMG) as our backbone structure. Next, we reach the first goal of internal-information screening by adjusting the CMG structure via the graph information bottleneck (GIB) principle (Wu et al., 2020), i.e., *GIB-guided feature refinement*, during which the less-informative features are filtered and the task-relevant structures will be highlighted. Then, to realize the second goal of external-information exploiting, we perform *multimodal topic integration*. We devise a latent multimodal topic module

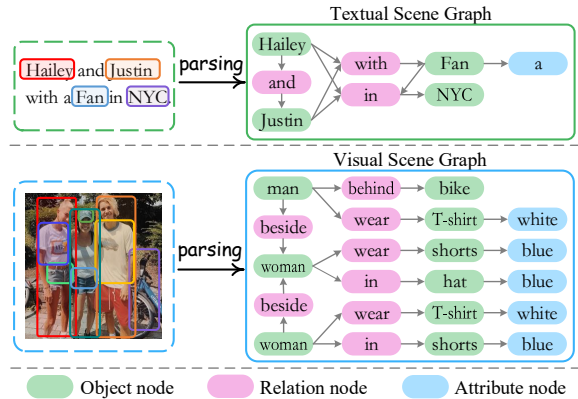


Figure 2: Examples of textual and visual scene graphs.

to produce both the textual and visual topic features based on the multimodal inputs. The multimodal topic keywords are integrated into the CMG to enrich the overall contexts, based on which we conduct the final reasoning of relation for input.

We perform experiments on the benchmark MRE dataset (Zheng et al., 2021a), where the results show that our framework significantly boosts the current state of the art. Further analyses demonstrate that the GIB-guided feature refinement helps in effective input denoising, and the latent multimodal topic module induces rich task-meaningful visual&textual topic features as extended contexts. We finally reveal that the idea of internal-information screening is especially important to the scenario of higher text-vision relevance, while the external-information exploiting particularly works for the lower text-vision relevance case.

To sum up, this work contributes by introducing a novel idea of simultaneous information subtraction and addition for multimodal relation extraction. The internal-information over-utilization and external-information under-exploitation are two common co-existing issues in many multimodal applications, to which our method can be broadly applied without much effort.

## 2 Preliminary

### 2.1 Textual and Visual Scene Graph

There have been the visual scene graph (VSG) (Johnson et al., 2015) and textual scene graph (TSG) (Wang et al., 2018), where both of them include three types of nodes: *object node*, *attribute node*, and *relationship node*. All the nodes come with a specific label text, as illustrated in Fig. 2. In an SG, object and attribute nodes are connected with other objects via pairwise relations. As intrinsically describing the semantic structures of scene

contexts for the given texts or images, SGs are widely utilized as types of external features integrated into downstream applications for enhancements, e.g., image retrieval (Johnson et al., 2015), image generation (Johnson et al., 2018) and image captioning (Yang et al., 2019). We also take advantage of these SG structures for better cross-modal semantic feature learning. Formally, we define a scene graph as  $G=(V, E)$ , where  $V$  is the set of nodes, and  $E$  is the set of edges.

## 2.2 Graph Information Bottleneck Principle

The information bottleneck (IB) principle (Alemi et al., 2017) is designed for information compression. Technically, IB learns a minimal feature  $Z$  to represent the raw input  $X$  that is sufficient to infer the task label  $Y$ . Further, the graph-based IB has been introduced for the graph data modeling (Wu et al., 2020), i.e., by refining a raw graph  $G$  into an informative yet compact one  $G^-$ , by optimizing:

$$\min_{G^-} [-I(G^-, Y) + \beta \cdot I(G^-, G)], \quad (1)$$

where  $I(G^-, G)$  minimizes the mutual information between  $G$  and  $G^-$  such that  $G^-$  learns to be the minimal and compact one of  $G$ .  $I(G^-, Y)$  is the prediction objective, which encourages  $G^-$  to be informative enough to predict the label  $Y$ .  $\beta$  is a Lagrangian coefficient. We will employ the GIB principle for internal-information screening.

## 2.3 Latent Multimodal Topic Modeling

We introduce a latent multimodal topic (LAMO) model. Technically, we first represent the input text  $T$  with a bag-of-words (BoW) feature  $\mathbf{b}^T$ , and represent image  $I$  with a visual BoW (VBoW)<sup>1</sup>  $\mathbf{b}^I$ . The topic generative process is described as follows:

- Draw a topic distribution  $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ .
- For each word token  $w_i^T$  and visual token  $w_j^I$ :
  - Draw  $w_i^T \sim \text{Multinomial}(\boldsymbol{\chi}, \boldsymbol{\theta})$ ,
  - Draw  $w_j^I \sim \text{Multinomial}(\boldsymbol{\psi}, \boldsymbol{\theta})$ .

where  $\boldsymbol{\mu}$  and  $\boldsymbol{\sigma}$  are the mean and variance vector for the posterior probability  $p(\boldsymbol{\theta}|T, I)$ .  $\boldsymbol{\chi} \in \mathbb{R}^{K \times U^T}$  and  $\boldsymbol{\psi} \in \mathbb{R}^{K \times U^I}$  are the probability matrices of *textual topic-word* and *visual topic-word*, respectively.  $K$  is the pre-defined topic numbers, and  $U^T$  and  $U^I$  are textual and visual vocabulary size.

As depicted in Fig. 3,  $\boldsymbol{\mu}$  and  $\boldsymbol{\sigma}$  are produced from a cross-modal feature encoder upon  $T$  and  $I$ . The topic distribution is yielded via  $\boldsymbol{\theta} = \text{Softmax}(\boldsymbol{\mu} + \boldsymbol{\sigma} \cdot \boldsymbol{\varepsilon})$ , where  $\boldsymbol{\varepsilon} \in \mathcal{N}(0, \mathbf{I})$ . Then, we autoregressively reconstruct the input  $\mathbf{b}^T$  and

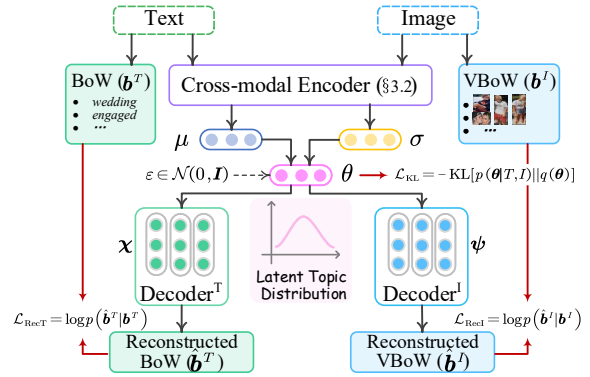


Figure 3: The schematic of our latent multimodal topic (LAMO) model.

$\mathbf{b}^I$  based on  $\boldsymbol{\theta}$ :

$$p(\mathbf{b}_i^T | \boldsymbol{\chi}, \boldsymbol{\theta}) = \text{Softmax}(\boldsymbol{\theta} \cdot \boldsymbol{\chi} | \mathbf{b}_{<i}^T), \quad (2)$$

$$p(\mathbf{b}_i^I | \boldsymbol{\psi}, \boldsymbol{\theta}) = \text{Softmax}(\boldsymbol{\theta} \cdot \boldsymbol{\psi} | \mathbf{b}_{<i}^I). \quad (3)$$

Then, with the activated  $k$ -th topic (via  $\text{argmax}$  over  $\boldsymbol{\theta}$ ), we obtain the distributions of the textual and visual topic words by slicing the  $\boldsymbol{\chi}[k, :] \in \mathbb{R}^{U^T}$  and  $\boldsymbol{\psi}[k, :] \in \mathbb{R}^{U^I}$ .

As shown in Fig. 3, the objective of topic modeling is derived as follows:

$$\mathcal{L}_{\text{LAMO}} = \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{RecT}} + \mathcal{L}_{\text{RecI}}. \quad (4)$$

Appendix §A.5 extends the description of LAMO.

## 3 MRE Framework

As shown in Fig. 4, our overall framework consists of five tiers. First, the model takes as input an image  $I$  and text  $T$ , as well as the subject  $v_s$  and object entity  $v_o$ . We represent  $I$  and  $T$  with the corresponding VSG and TSG. Then, the VSG and TSG are assembled as a cross-modal graph, which is further modeled via a graph encoder. Next, we perform GIB-guided feature refinement over the CMG for internal-information screening, which results in a structurally compact backbone graph. Afterwards, the multimodal topic features induced from the latent multimodal topic model are integrated into the previously obtained feature representation for external-information exploitation. Finally, the decoder predicts the relation label  $Y$  based on the enriched features.

### 3.1 Scene Graph Generation

We employ the off-the-shelf parsers to generate the VSG (i.e.,  $G^I=(V^I, E^I)$ ) and TSG (i.e.,  $G^T=(V^T, E^T)$ ), respectively. We denote the representations of VSG nodes as  $\mathbf{X}^I=\{\mathbf{x}_1^I, \dots, \mathbf{x}_n^I\}$ , where each node embedding  $\mathbf{x}_i^I$  is the concatenation of the object region representations and the corresponding node label embeddings. We directly

<sup>1</sup>Note that visual topic words are visual objects.

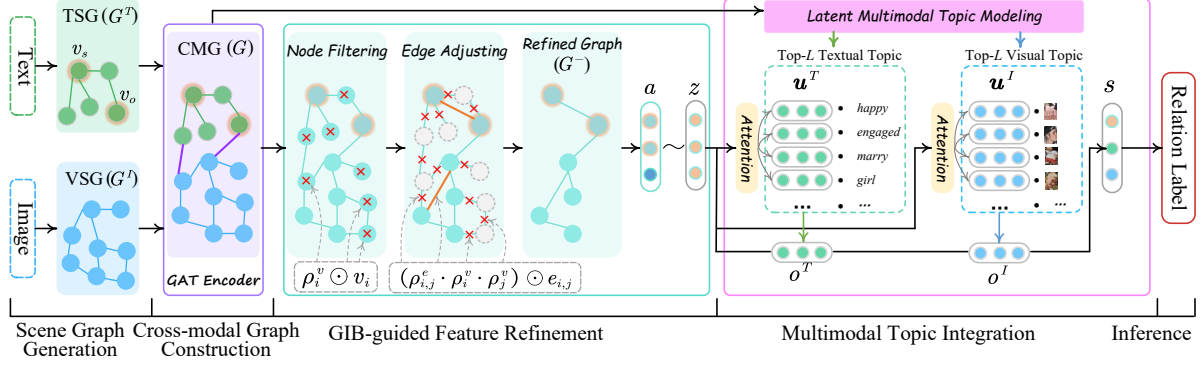


Figure 4: Overview of our proposed framework.

represent the TSG nodes as  $\mathbf{X}^T = \{\mathbf{x}_1^T, \dots, \mathbf{x}_m^T\}$ , where each  $\mathbf{x}_j^T$  is the contextualized word embedding. Note that both visual objects and text token representations are obtained from the CLIP (Radford et al., 2021) encoder, which ensures an identical embedding space across the two modalities. More details are provided in the appendix §A.1&§A.2.

### 3.2 Cross-modal Graph Construction

Next, we consider merging the VSG and TSG into one unified backbone cross-modal graph (CMG). Let’s denote CMG as  $G=(V, E)$ , where  $V(=V^T \cup V^I)$  is the union of  $V^I$  and  $V^T$ .  $E(=E^T \cup E^I \cup E^\times)$  is the set of edges, including the *intra-modal* edges ( $E^I$  and  $E^T$ ), and *inter-modal* hyper-edges  $E^\times$ . To build the cross-modal hyper-edges between each pair of VSG node  $v_i^I$  and TSG node  $v_j^T$ , we measure the relevance score  $s$  in between:

$$s_{v_i^I, v_j^T} = \cos(\mathbf{x}_i^I, \mathbf{x}_j^T). \quad (5)$$

A hyper-edges  $e_{i,j}^\times$  is created if  $s_{v_i^I, v_j^T}$  is larger than a pre-defined threshold  $\lambda$ . Node representations from VSG and TSG are copied as the CMG’s node representations, i.e.,  $\mathbf{X}=\mathbf{X}^T \cup \mathbf{X}^I$ . We denote each edge  $e_{i,j}(\in E)=1$  if there is an edge between nodes, and  $e_{i,j}=0$  and vice versa. Next, a graph attention model (GAT; Velickovic et al., 2018) is used to fully propagate the CMG:

$$\mathbf{H} = \{\mathbf{h}_1, \dots, \mathbf{h}_{m+n}\} = \text{GAT}(G, \mathbf{X}). \quad (6)$$

### 3.3 GIB-guided Feature Refinement

In this step, we propose a GIB-guided feature refinement (GENE) module to optimize the initial CMG structure such that we fine-grainedly prune the input image and text features. Specifically, with the GIB guidance, we 1) filter out those task-irrelevant nodes, and 2) adjust the edges based on their relatedness to the task inference.

**Node Filtering** We assign a 0 or 1 value  $\rho_i^v$  to a node  $v_i$  indicating whether to prune or keep  $v_i$ , i.e., via  $\rho_i^v \odot v_i$ . We sample the value from the *Bernoulli* distribution, i.e.,  $\rho_i^v \in \{0, 1\} \sim \text{Bernoulli}(\pi_i^v)$ , where  $\pi_i^v \in (0, 1)$  is a parameter. While the sampling is a discrete process, we make it differentiable via the concrete relaxation method (Jang et al., 2017):

$$\rho_i^v = \text{Sigmoid}\left(\frac{1}{\tau} \left(\log \frac{\pi_i^v}{1 - \pi_i^v} + \log \frac{\epsilon}{1 - \epsilon}\right)\right), \quad (7)$$

where  $\tau$  is the temperature,  $\epsilon \sim \text{Uniform}(0, 1)$ . We estimate  $\pi_i^v$  by considering both the  $v_i$ ’s *l-order context* and the influence of target entity pair:

$$\begin{aligned} \mathbf{r}_i^v &= \text{Att}(v_i, \varphi(v_i); \mathbf{H}), \\ \pi_i^v &= \text{Sigmoid}(\text{FFN}([\mathbf{r}_i^v; \mathbf{h}_s; \mathbf{h}_o])), \end{aligned} \quad (8)$$

where  $\text{Att}(\cdot)$  is an attention operation,  $\varphi(v_i)$  is the *l-order* neighbor nodes of  $v_i$ ,  $\mathbf{h}_s$  and  $\mathbf{h}_o$  are the representations of the subject and object entity.

**Edge Adjusting** Similarly, we take the same sampling operation (Eq. 7) to generate a signal  $\rho_{i,j}^e$  for any edge  $e_{i,j}$ , during which we also consider the *l-order context* features and the target entity pair:

$$\begin{aligned} \mathbf{r}_{i,j}^e &= \text{Att}(v_i, \varphi(v_i), v_j, \varphi(v_j); \mathbf{H}), \\ \pi_{i,j}^e &= \text{Sigmoid}(\text{FFN}([\mathbf{r}_{i,j}^e; \mathbf{h}_s; \mathbf{h}_o])), \end{aligned} \quad (9)$$

where  $\varphi(v_i)$  and  $\varphi(v_j)$  are the *l-order* neighbor nodes of  $v_i$  and  $v_j$ . Instead of directly determining the existence of  $e_{i,j}$  with  $\rho_{i,j}^e$ , we also need to take into account the existences of  $v_i$  and  $v_j$ , i.e.,  $(\rho_{i,j}^e \cdot \rho_i^v \cdot \rho_j^v) \odot e_{i,j}$ . Because even if  $\rho_{i,j}^e=1$ , an edge is non-existent when its affiliated nodes are deleted.

Thereafter, we obtain an adjusted CMG, i.e.,  $G^-$ , which is further updated via the GAT encoder, resulting in new node representations  $\mathbf{H}^-$ . We apply pooling operation on  $\mathbf{H}^-$  to obtain the overall graph presentation  $\mathbf{g}$ , which is concatenated with two entity representations as the context feature  $\mathbf{a}$ :

$$\mathbf{a} = [\mathbf{g}; \mathbf{h}_s; \mathbf{h}_o]. \quad (10)$$

**GIB Optimization** To ensure that the above-adjusted graph  $G^-$  is sufficiently informative (i.e.,

not wrongly pruned), we consider a GIB-guided optimization. We denote  $\mathbf{z}$  as the compact information of the resulting  $G^-$ , which is sampled from a Gaussian distribution parameterized by  $\mathbf{a}$ . Then, we rephrase the raw GIB objective (Eq. 1) as:

$$\mathcal{L}_{\text{GIB}} = \min_{\mathbf{z}} [-I(\mathbf{z}, Y) + \beta \cdot I(\mathbf{z}, G)]. \quad (11)$$

The first term  $-I(\mathbf{z}, Y)$  can be expanded as:

$$\begin{aligned} -I(\mathbf{z}, Y) &\leq -\int p(Y, \mathbf{z}) \log q(Y|\mathbf{z}) dY d\mathbf{z} + H(Y) \\ &:= \mathcal{L}_{\text{CE}}(q(Y|\mathbf{z}), Y), \end{aligned} \quad (12)$$

where  $q(Y|\mathbf{z})$  is a variational approximation of the true posterior  $p(Y, \mathbf{z})$ . For the second term  $I(\mathbf{z}, G)$ , we estimate its upper bound via reparameterization trick (Kingma and Welling, 2014):

$$\begin{aligned} I(\mathbf{z}, G) &\leq \int p(\mathbf{z}|G) \log \frac{p(\mathbf{z}|G)}{r(\mathbf{z})} d\mathbf{z} dG \\ &:= \text{KL}(p(\mathbf{z}|G)||r(\mathbf{z})). \end{aligned} \quad (13)$$

We run GENE several iterations for sufficient refinement. In Appendix §A.4 we detail all the technical processes of GIB-guided feature refinement.

### 3.4 Multimodal Topic Integration

We further enrich the compressed CMG features with more semantic contexts, i.e., the multimodal topic features. As depicted in Sec. §2.3, our LAMO module takes as input the backbone CMG representation  $\mathbf{H}$  and induces both the visual and textual topic keywords that are semantically relevant to the input content. Note that we only retrieve the associated top- $L$  textual and visual keywords, separately. Technically, we devise an attention operation to integrate the embeddings of the multimodal topic words ( $\mathbf{u}^T$  and  $\mathbf{u}^I$ , from CLIP encoder) into the resulting feature representation  $\mathbf{z}$  of GENE:

$$\begin{aligned} \alpha_i^{T/I} &= \frac{\exp(\text{FFN}([\mathbf{u}_i^{T/I}; \mathbf{z}]))}{\sum_i^L \exp(\text{FFN}([\mathbf{u}_i^{T/I}; \mathbf{z}]))}, \\ \mathbf{o}^{T/I} &= \sum_i^L \alpha_i^{T/I} \mathbf{u}_i^{T/I}. \end{aligned} \quad (14)$$

We finally summarize these three representations as the final feature:

$$\mathbf{s} = [\mathbf{z}; \mathbf{o}^T; \mathbf{o}^I]. \quad (15)$$

### 3.5 Inference and Learning

Based on  $\mathbf{s}$ , a softmax function predicts the relation label  $\hat{Y}$  for the entity pair  $v_s \& v_o$ . The training of our overall framework is based on a warm-start strategy. First, GENE is trained via  $\mathcal{L}_{\text{GIB}}$  (Eq. 11) for learning the sufficient multimodal fused representation in CMG, and refined features from compacted CMG. Then LAMO module is unsupervisedly pre-trained separately via  $\mathcal{L}_{\text{LAMO}}$  (Eq. 4) on the well-learned multimodal fused representations

	Train	Develop	Test	Total
#Sentence	7,356	931	914	9,201
#Instance	12,247	1,624	1,614	15,485
#Entity	16,863	2,174	2,143	21,180
#Relation	12,247	1,624	1,614	15,485
#Image	7,356	931	914	9,201

Table 1: Summary of the dataset and splits in our experiments. ‘#’ denotes the number.

so as to efficiently capture the task-related topic. Once the two modules have converged, we train our overall framework with the final cross-entropy task loss  $\mathcal{L}_{\text{CE}}(\hat{Y}, Y)$ , together with the above two learning loss:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \eta_1 \mathcal{L}_{\text{GIB}} + \eta_2 \mathcal{L}_{\text{LAMO}}. \quad (16)$$

## 4 Experiment

### 4.1 Setting

We experiment with the MRE dataset<sup>2</sup>, which contains 9,201 text-image pairs and 15,485 entity pairs with 23 relation categories. The statistical information of the MRE dataset is listed in Table 1. Note that a sentence may contain several entity pairs, and thus a text-image pair can be divided into several instances, each with only one entity pair. We follow the same split of training, development, and testing, as set in Zheng et al. (2021a). We compare our method with baselines in two categories: **1) Text-based RE methods** that traditionally leverage merely the texts of MRE data, including, *BERT* (Devlin et al., 2019), *PCNN* (Zeng et al., 2015), *MTB* (Soares et al., 2019), and *DP-GCN* (Yu et al., 2020). **2) Multimodal RE methods** as in this work, including, *BERT+SG* (Zheng et al., 2021a), *MEGA* (Zheng et al., 2021a), *VisualBERT* (Li et al., 2019), *ViLBERT* (Lu et al., 2019), *RDS* (Xu et al., 2022), *MKGformer* (Chen et al., 2022a), and *HVP-Net* (Chen et al., 2022b).

We use the pre-trained language-vision model CLIP (vit-base-patch32) to encode the visual and textual inputs. We set the learning rate as 2e-5 for pre-trained parameters, and 2e-4 for the other parameters. The threshold value  $\lambda$  is set to 0.25; the temperature  $\tau$  is 0.1; and  $\beta$  is set to 0.01. All the dimensions of node representations and GAT hidden sizes are set as 768-d. We utilize the 2-order (i.e.,  $l = 2$ ) context of each node to refine the nodes and edges of CMG. For the latent topic modeling, we pre-define the number of topics as 10, and then we choose the Top-10 textual and visual keywords to enhance the semantic contexts

<sup>2</sup><https://github.com/thecharm/Mega>

	Acc.	Pre.	Rec.	F1
<b>• Text-based Methods</b>				
BERT <sup>†</sup>	-	63.85	55.79	59.55
PCNN <sup>†</sup>	72.67	62.85	49.69	55.49
MTB <sup>†</sup>	72.73	64.46	57.81	60.86
DP-GCN <sup>‡</sup>	74.60	64.04	58.44	61.11
<b>• Multimodal Methods</b>				
BERT(Text+Image) <sup>‡</sup>	74.59	63.07	59.53	61.25
BERT+SG <sup>†</sup>	74.09	62.95	62.65	62.80
MEGA <sup>†</sup>	76.15	64.51	68.44	66.41
VisualBERT <sup>†</sup> <sub>base</sub>	-	57.15	59.48	58.30
ViLBERT <sup>†</sup> <sub>base</sub>	-	64.50	61.86	63.16
RDS <sup>†</sup>	-	66.83	65.47	66.14
HVPNeT <sup>†</sup>	-	83.64	80.78	81.85
MKGformer <sup>†</sup>	<u>92.31</u>	<u>82.67</u>	<u>81.25</u>	<u>81.95</u>
<b>Ours</b>	<b>94.06</b>	<b>84.69</b>	<b>83.38</b>	<b>84.03</b>
w/o GENE (Eq. 11)	92.42	82.41	81.83	82.12
w/o $I(z, G)$ (Eq. 13)	93.64	83.61	82.34	82.97
w/o LAMO (Eq. 4)	92.86	82.97	81.22	82.09
w/o $\sigma^T$	93.05	83.95	82.53	83.23
w/o $\sigma^I$	93.63	84.03	83.18	83.60
w/o VSG&TSG	93.12	83.51	82.67	83.09
w/o CMG	93.97	84.38	83.20	83.78

Table 2: Main results. ‘w/o  $I(z, G)$ ’ means GENE adjustment without GIB guidance. ‘w/o CMG’ means VSG and TSG are not connected with hyper-edge  $E^X$ . ‘w/o VSG&TSG’ means our method uses the embeddings of visual and text inputs without structural SG modeling. Baselines with the superscript ‘†’ are copied from their raw papers (Chen et al., 2022a); with ‘‡’ are from our re-implementation.

of compressed CMG. All models are trained and evaluated using the NVIDIA A100 Tensor Core GPU. Following existing MRE work, we adopt accuracy (Acc.), precision (Pre.), recall (Rec.), and F1 as the major evaluation metrics.

## 4.2 Main Results

Table 2 shows the overall results. First, compared to the traditional text-based RE, multimodal methods, by leveraging the additional visual features, exhibit higher performances consistently. But without carefully navigating the visual information into the task, most MRE baselines merely obtain incremental improvements over text-based ones. By designing delicate text-vision interactions, HVPNeT and MKGformer achieve the current state-of-the-art (SoTA) results. Most importantly, our model boosts the SoTA with a very significant margin, i.e., with improvements of 1.75%(=94.06-92.31) in accuracy and 2.08%(=84.03-81.95) in F1. This validates the efficacy of our method.

**Model Ablation** In the lower part of Table 2, we also study the efficacy of each part of our designs. First of all, we see that both the GENE and

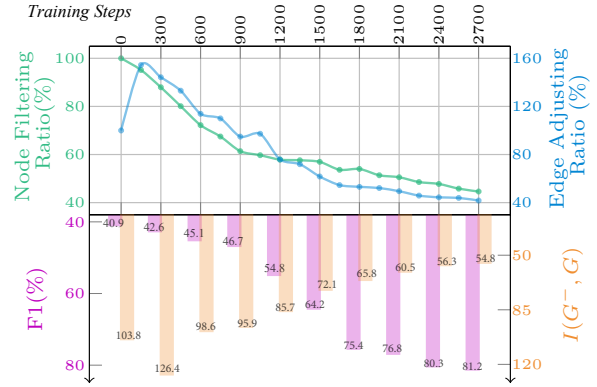


Figure 5: The trends of changing ratio of nodes and edges, along with the task performance and the mutual information between  $G$  and  $G^-$ . The model is without LAMO. This is done on the developing set.

LAMO modules show big impacts on the results, i.e., exhibiting a drop in F1 by 1.91% and 1.94% F1, respectively. This confirms their fundamental contributions to the whole system. More specifically, the GIB guidance is key to the information refinement in GENE, while both the textual and visual topic features are key to LAMO. Also, it is critical to employ the SG for the structural modeling of the multimodal inputs. And the proposal of the cross-modal graph is also helpful to task modeling.

## 4.3 Analysis and Discussion

To gain a deeper understanding of how our proposed methods succeed, we conduct further analyses to answer the following questions.

**►RQ1:** Does GENE helps by really denoising the input features?

**A:** We first investigate the working mechanism of GENE on internal-information screening. We plot the trajectories of the node filtering and the edge adjusting, during which we show the changing trends of overall performances and the mutual information  $I(G^-, G)$  between the raw CMG ( $G$ ) and the pruned one ( $G^-$ , i.e.,  $z$ ). As shown in Fig. 5, along with the training process both the number of nodes and edges decrease gradually, while the task performance climbs steadily as  $I(G^-, G)$  declines. These clearly show the efficacy of the task-specific information denoising by GENE.

**►RQ2:** Are LAMO induced task-relevant topic features beneficial to the end task?

**A:** Now, we consider visualizing the learned contextual features in our system, including the  $z$  without integrating the topic features, and the  $s$  with rich topic information injected. We separately

Topic	Textual keywords	Visual keywords (ID)
#Politic	trump, president, world, new, china, leader, summit, meet, korean, senate	#1388, #1068
#Music	tour, concert, video, live, billboard, album, styles, singer, taylor, dj	#1446, #1891
#Love	wife, wedding, engaged, ring, son, baby, girl, love, rose, annie	#434, #1091
#Leisure	photo, best, beach, lake, island, bridge, view, florida, photograph, great	#679, #895
#Idol	metgala, hailey, justin, taylor, rihanna, hit, show, annual, pope, shawn	#1021, #352
#Scene	contain, near, comes, american, in, spotted, travel, to, from, residents	#535, #167
#Sports	team, man, world, cup, nike, nba, football, join, play, chelsea	#1700, #109
#Social	google, retweet, twitter, youtube, netflix, acebook, flight, butler, series, art	#1043, #1178
#Show	show, presents, dress, interview, shot, speech, performing, attend, portray, appear	#477, #930
#Life	good, life, please, family, dog, female, people, boy, soon, daily	#613, #83

Topic	Visual Keywords(ID)	Topic	Visual Keywords(ID)
#Politic	 #1388, #1068	#Music	 #1466, #1891
#Love	 #434, #1091	#Leisure	 #679, #895
#Idol	 #1021, #352	#Scene	 #535, #167
#Sports	 #1700, #109	#Social	 #1043, #1178
#Show	 #477, #930	#Life	 #613, #83

Table 3: Top 10 key textual topic keywords and top 2 visual topic keywords discovered by LAMO.

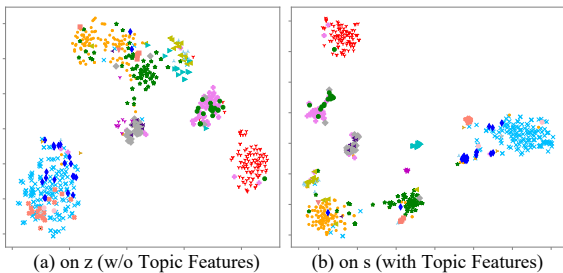


Figure 6: T-SNE visualization of the contextualized features with or without topic features. Different colors indicate different ground-truth relation labels.

project  $z$  and  $s$  into the ground-truth relational labels of the MRE task, as shown in Fig. 6. We see that both  $z$  and  $s$  have divided the feature space into several clusters clearly, thanks to the GIB-guided information screening. However, there are still some wrongly-placed or entangled instances in  $z$ , largely due to the input feature deficiency. By supplementing more contexts with topic features, the patterns in  $s$  become much clearer, and the errors reduce. This indicates that LAMO induces topic

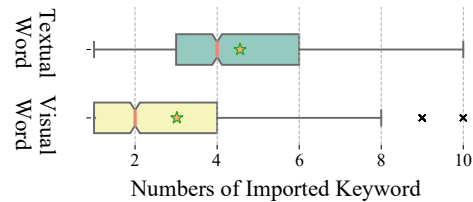


Figure 7: Distribution of numbers of textual and visual topic keywords imported for MRE.

information beneficial to the task.

Meanwhile, we demonstrate what latent topics LAMO can induce. In Table 3 we show the top 10 latent topics with both the textual and visual keywords, where we notice that the latent topic information is precisely captured and modeled by LAMO. Further, we study the variance of the latent topics in two modalities, exploring the different contributions of each type. Technically, we analyze the numbers of the imported topic keywords of textual and visual ones respectively, by observing the attention weights  $\alpha_i^{T/I}$  (Eq. 14). In Fig. 7 we plot the distributions. It can be found that the

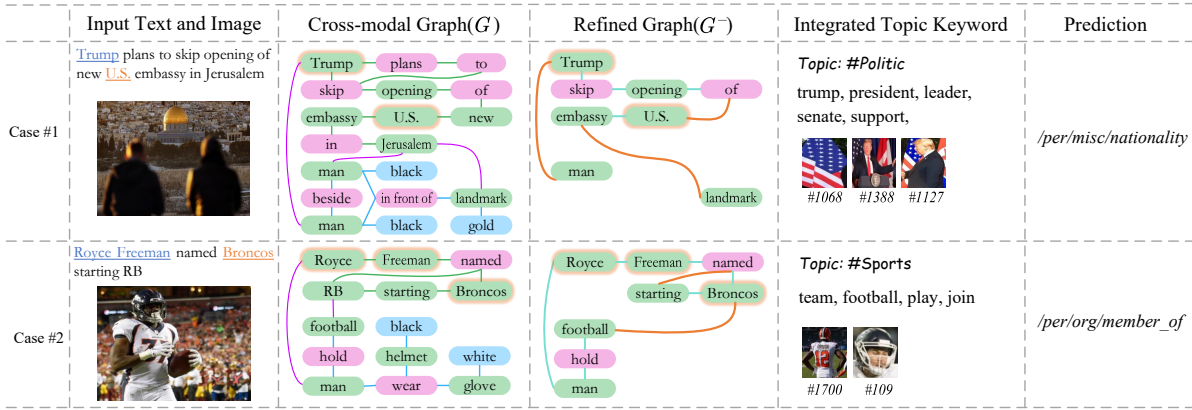


Figure 8: Quantitative results of two testing examples, where our model made correct predictions.

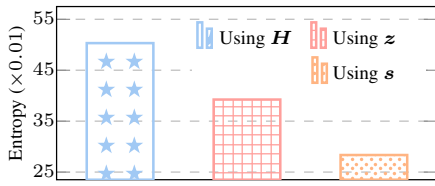


Figure 9: The entropy under various model settings.

model tends to make use of more textual contexts, compared with the visual ones.

►RQ3: How do GENE and LAMO collaborate to solve the end task?

A: As demonstrated previously, the GENE is able to relieve the issue of noisy information, and LAMO can produce latent topics to offer additional clues for relation inference. Now we study how these two modules cooperate together to reach the best results. First, we use the learned feature  $c^*$  to calculate task entropy  $-\sum p(Y|c^*) \log p(Y|c^*)$ , where lower entropy means more confidence of the correct predictions. We compute the entropy using  $H$  (initial context feature), using  $z$  (with denoised context feature) and using  $s$  (with feature denoising and topic enriched context), respectively, which represents the three stages of our system, as shown in Fig. 9. As seen, after the information denoising and enriching by GENE and LAMO respectively, the task entropy drops step by step, indicating an effective learning process with the two modules.

We further empirically perform a case study to gain an intuitive understanding of how the two modules come to play. In Fig. 8 we illustrate the two testing instances, where we visualize the constructed cross-model graph structures, the refined graphs ( $G^-$ ) and then the imported multimodal topic features. We see that GENE has fine-grainedly removed those noisy and redundant nodes, and adjusted the node connections that are more knowledgeable for the relation prediction.

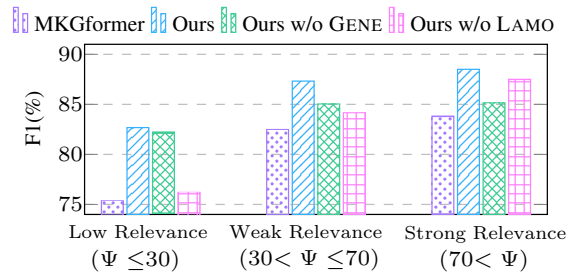


Figure 10: Results under varying text-image relevance.

For example, in the refined graph, the task-noisy visual nodes, ‘man’ and textual nodes, ‘in’, ‘plans’ are removed, and the newly-generated edges (e.g., ‘Trump’ → ‘US’, and ‘Broncos’ → ‘football’) allow more efficient information propagation. Also, the model correctly paid attention to the topic words retrieved from LAMO that are useful to infer the relation, such as ‘president’, ‘leader’ in case #1, and ‘team’, and ‘football’ in case #2.

►RQ4: Under what circumstances do the internal-information screening and external-information exploiting help?

A: In realistic scenarios, a wide range of multimodal tasks is likely to face the issues of internal-information over-utilization and external-information under-exploitation (or simultaneously). Especially for the data collected from the online web, the vision and text pairs are not well correlated. Finally, we take one step further, exploring when our idea of internal-information screening and external-information exploiting aids the tasks in such cases. Technically, we first measure the vision-language relevance  $\Psi$  of each image-text pair by matching the correspondence of the VSG and TSG structures. And then, we group the instances by their relevance scores, and finally make predictions for different groups. From Fig. 10, it is observed that for the inputs with higher text-



vision relevance, the GENE plays a greater role than LAMO, while under the case with less cross-modal feature relevance, LAMO contributes more significantly than GENE. This is reasonable because most of the high cross-modal relevance input features come with rich yet even redundant information, where the internal-information screening is needed for denoising. When the input text-vision sources are irrelevant, the exploitation of external features (i.e., latent topic) can be particularly useful to bridge the gaps between the two modalities. On the contrary, **MKGformer** performs quite badly especially when facing with data in low vision-language relevance. Integrating both the LAMO and GENE, our system can perform consistently well under any case.

## 5 Related Works

As one of the key subtasks of the information extraction track, relation extraction (RE) has attracted much research attention (Yu et al., 2020; Chen et al., 2022c; Tan et al., 2022; Guo et al., 2023). The recent trend of RE has shifted from the traditional textual RE to the recent multimodal RE, where the latter additionally adds the image inputs in the former one for better performances, under the intuition that the visual information can offer complementary features to the purely textual input from other modalities. Zheng et al. (2021b) pioneers the MRE task with a benchmark dataset, which is collected from the social media posts that come with rich vision-language sources. Later, more delicate and sophisticated methods are proposed to enhance the interactions between the input texts and images, and achieve promising results Zheng et al. (2021a); Chen et al. (2022b,a).

On the other hand, increasing attention has been paid to exploring the role of different information in the RE task. As extensively revealed in prior RE studies, only a few parts of the input sentence can provide real clues for the relation inference (Xu et al., 2015; Yu et al., 2020), which inspires the proposal of textual feature pruning methods (Zhang et al., 2018; Jin et al., 2022). More recently, Vempala and Preoȃiuc-Pietro (2019); Li et al. (2022) have shown that not always the visual inputs serve positive contributions in existing MRE models, as the social media data contains many noises. Xu et al. (2022) thus introduce an instance-level filtering approach to directly drop out those images less-informative to the task. However, such coarse-grained aggressive data deletion

will inevitably abandon certain useful visual features. In this work we propose screening the noisy information from both the visual and textual input features, in a fine-grained and more controllable manner, i.e., structure denoising via graph information bottleneck technique (Wu et al., 2020). Also, we adopt the scene graph structures to model both the vision and language features, which partially inherits the success from Zheng et al. (2021a) that uses visual scene graphs to represent input images.

Due to the sparse and noisy characteristics of social media data, as well as the cross-modal information detachment, MRE also suffers from feature deficiency problems. We thus propose modeling the latent topic information as additional context features to enrich the inputs. Multimodal topic modeling has received considerable explorations (Chu et al., 2016; Chen et al., 2021), which extends the triumph of the textual latent topic models as in NLP applications (Zhu et al., 2021; Fu et al., 2020; Xie et al., 2022). We however note that existing state-of-the-art latent multimodal models (An et al., 2020; Zosa and Pivovarova, 2022) fail to navigate the text and image into a unified feature space, which leads to irrelevant vision-text topic induction. We thus propose an effective latent multimodal model that learns coherent topics across two modalities. To our knowledge, we are the first to attempt to integrate the multimodal topic features for MRE.

## 6 Conclusion

In this paper, we solve the internal-information over-utilization issue and the external-information under-exploitation issue in multimodal relation extraction. We first represent the input images and texts with the visual and textual scene graph structures, and fuse them into the cross-modal graphs. We then perform structure refinement with the guidance of the graph information bottleneck principle. Next, we induce latent multimodal topic features to enrich the feature contexts. Our overall system achieves huge improvement over the existing best model on the benchmark data. Further in-depth analyses offer a deep understanding of how our method advances the task.

## Acknowledgments

The work is substantially supported by Alibaba Group through the Alibaba Innovative Research (AIR) Program, and is also partially supported by the Sea-NExT Joint Lab at the National University of Singapore.

## Limitiations

The main limitations of our work lie in the following two aspects: First, we take sufficient advantage of the scene graph (SG) structures, which are obtained by external SG parsers. Therefore, the overall performance of our system is subject to the quality of the SG parser to some extent. However, our system, by equipping with the refinement mechanism, is capable of resisting the quality degradation of SG parsers to a certain extent. Second, the performance of the latent multimodal topic model largely relies on the availability of large-scale text-image pairs. However, the size of the dataset of MRE is limited, which may limit the topic model in achieving the best effect.

## References

- Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. 2017. Deep variational information bottleneck. In *Proceedings of the ICLR*.
- Minghui An, Jingjing Wang, Shoushan Li, and Guodong Zhou. 2020. Multimodal topic-enriched auxiliary learning for depression detection. In *Proceedings of the COLING*, pages 1078–1089.
- Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020. What was written vs. who read it: News media profiling using text analysis and social media context. In *Proceedings of the ACL*, pages 3364–3374.
- Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021. Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the EACL*, pages 1676–1683.
- Shulin Cao, Jiaxin Shi, Zijun Yao, Xin Lv, Jifan Yu, Lei Hou, Juanzi Li, Zhiyuan Liu, and Jinghui Xiao. 2022. Program transfer for answering complex questions over knowledge bases. In *Proceedings of the ACL*, pages 8128–8140.
- Jiaxin Chen, Zekai Wu, Zhenguo Yang, Haoran Xie, Fu Lee Wang, and Wenyin Liu. 2021. Multimodal fusion network with latent topic memory for rumor detection. In *Proceedings of the ICME*, pages 1–6.
- Xiang Chen, Ningyu Zhang, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, and Huajun Chen. 2022a. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *Proceedings of the SIGIR*, pages 904–915.
- Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022b. Good visual guidance make A better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. In *Proceedings of the NAACL Findings*, pages 1607–1618.
- Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022c. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the WWW*, pages 2778–2788.
- Lingyang Chu, Yanyan Zhang, Guorong Li, Shuhui Wang, Weigang Zhang, and Qingming Huang. 2016. Effective multimodality fusion framework for cross-media topic detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(3):556–569.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the NAACL*, pages 4171–4186.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the ICLR*.
- Zihao Fu, Lidong Bing, Wai Lam, and Shoaib Jameel. 2020. Dynamic topic tracker for kb-to-text generation. In *Proceedings of COLING*, pages 2369–2380.
- Jiuxiang Gu, Shafiq R. Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. 2019. Unpaired image captioning via scene graph alignments. In *Proceedings of the ICCV*, pages 10322–10331.
- Jia Guo, Stanley Kok, and Lidong Bing. 2023. Towards integration of discriminability and robustness for document-level relation extraction. In *Proceedings of the EACL*, pages 2598–2609.
- John R. Hershey and Peder A. Olsen. 2007. Approximating the kullback leibler divergence between gaussian mixture models. In *Proceedings of the ICASSP*, pages 317–320.
- Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *Proceedings of the ICLR*.
- Yifan Jin, Jiangmeng Li, Zheng Lian, Chengbo Jiao, and Xiaohui Hu. 2022. Supporting medical relation extraction via causality-pruned semantic dependency forest. In *Proceedings of the COLING*, pages 2450–2460.
- Justin Johnson, Agrim Gupta, and Li Fei-Fei. 2018. Image generation from scene graphs. In *Proceedings of the CVPR*, pages 1219–1228.

- Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *Proceedings of the CVPR*, pages 3668–3678.
- Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *Proceedings of the ICLR*.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.
- Lei Li, Xiang Chen, Shuofei Qiao, Feiyu Xiong, Huajun Chen, and Ningyu Zhang. 2022. On analyzing the role of image for visual-enhanced relation extraction. *CoRR*, abs/2211.07504.
- Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *CoRR*.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the NIPS*, pages 13–23.
- Ishani Mondal, Yufang Hou, and Charles Jochim. 2021. End-to-end construction of NLP knowledge graph. In *Proceedings of the ACL-IJCNLP Findings*, pages 1885–1895.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the ICML*, pages 8748–8763.
- Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Proceedings of the NIPS*, pages 91–99.
- Sebastian Schuster, Ranjay Krishna, Angel X. Chang, Li Fei-Fei, and Christopher D. Manning. 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the EMNLP*, pages 70–80.
- Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the ACL*, pages 2895–2905.
- Qingyun Sun, Jianxin Li, Hao Peng, Jia Wu, Xingcheng Fu, Cheng Ji, and Philip S. Yu. 2022. Graph structure learning with variational information bottleneck. In *Proceedings of the AAAI*, pages 4165–4174.
- Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the ACL*, pages 1672–1681.
- Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning? In *Proceedings of the NIPS*.
- Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *Proceedings of the ICLR*.
- Alakananda Vempala and Daniel Preotjuc-Pietro. 2019. Categorizing and inferring the relationship between the text and image of Twitter posts. In *Proceedings of the ACL*, pages 2830–2840.
- Yu-Siang Wang, Chenxi Liu, Xiaohui Zeng, and Alan Yuille. 2018. Scene graph parsing as dependency parsing. In *Proceedings of the NAACL*, pages 397–407.
- Zihao Wang, Kwun Ping Lai, Piji Li, Lidong Bing, and Wai Lam. 2019. Tackling long-tailed relations and uncommon entities in knowledge graph completion. In *Proceedings of the EMNLP*, pages 250–260.
- Tailin Wu, Hongyu Ren, Pan Li, and Jure Leskovec. 2020. Graph information bottleneck. In *Proceedings of the NIPS*, pages 20437–20448.
- Qianqian Xie, Jimin Huang, Tulika Saha, and Sophia Ananiadou. 2022. GRETEL: graph contrastive topic enhanced language model for long document extractive summarization. In *Proceedings of the COLING*, pages 6259–6269.
- Bo Xu, Shizhou Huang, Ming Du, Hongya Wang, Hui Song, Chaofeng Sha, and Yanghua Xiao. 2022. Different data, different modalities! reinforced data splitting for effective multimodal information extraction from social media posts. In *Proceedings of the COLING*, pages 1855–1864.
- Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. 2015. Classifying relations via long short term memory networks along shortest dependency paths. In *Proceedings of the EMNLP*, pages 1785–1794.
- Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-encoding scene graphs for image captioning. In *Proceedings of the CVPR*, pages 10685–10694.
- Bowen Yu, Mengge Xue, Zhenyu Zhang, Tingwen Liu, Yubin Wang, and Bin Wang. 2020. Learning to prune dependency trees with rethinking for neural relation extraction. In *Proceedings of the COLING*, pages 3842–3852.

- Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural motifs: Scene graph parsing with global context. In *Proceedings of the CVPR*, pages 5831–5840.
- Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the EMNLP*, pages 1753–1762.
- Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R. Lyu, and Irwin King. 2018. Topic memory networks for short text classification. In *Proceedings of the EMNLP*, pages 3120–3131.
- Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the EMNLP*, pages 2205–2215.
- Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li, and Tao Wang. 2021a. Multimodal relation extraction with efficient graph alignment. In *Proceedings of the MM*, pages 5298–5306.
- Changmeng Zheng, Zhiwei Wu, Junhao Feng, Ze Fu, and Yi Cai. 2021b. MNRE: A challenge multimodal dataset for neural relation extraction with visual evidence in social media posts. In *Proceedings of the ICME*, pages 1–6.
- Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the ACL/IJCNLP*, pages 1571–1582.
- Elaine Zosa and Lidia Pivovarova. 2022. Multilingual and multimodal topic modelling with pretrained embeddings. In *Proceedings of the COLING*, pages 4037–4048.

## A Extended Method Specification

### A.1 Scene Graph Generating

We mainly follow the prior practice of SG applications (Yang et al., 2019; Gu et al., 2019) to acquire the visual scene graph (VSG) and textual scene graph (TSG). A VSG or TSG contains three types of nodes, including the object, attribute, and relation nodes.

For VSG, we employ the FasterRCNN (Ren et al., 2015) as an object detector to obtain all the object nodes, and use MOTIFS (Zellers et al., 2018) as a relation classifier to obtain the relation labels (nodes) as well as the relational edges, which is trained using the Visual Genome (VG) dataset (Krishna et al., 2017). We then use an attribute classifier to obtain attribute nodes. For TSG generation, we first convert the sentences into dependency trees with a dependency parser, which is then transformed into the scene graph based on the rules defined at Schuster et al. (2015). Note that the object nodes in VSG are image regions, while the object nodes in TSG are textual tokens.

### A.2 Node Embedding

In Section 3.1, we directly give the representations of nodes in VSG and TSG. Here, we provide the encoding process in detail.

**Visual Node Embedding** In VSG, the visual feature vector of an object node is extracted from its corresponding image region; the feature of the attribute node is the same as its connected object, while the visual feature vector of a relationship node is extracted from the union image region of the two related object nodes. Specifically, for each visual node, we first rescale it to  $224\text{-d} \times 224\text{-d}$ . Subsequently, following Dosovitskiy et al. (2021), each visual node is split into a sequence of fixed-size non-overlapping patches  $\{p_k \in \mathbb{R}^{P \times P}\}$ , where  $P \times P$  is the patch size. Then, we map all patches of  $i$ -th visual node to a  $d$ -dimensional vector  $\mathbf{X}_i^{PC}$  with a trainable linear projection. For each sequence of image patches, a [CLS] token embedding  $\mathbf{x}_{CLS} \in \mathbb{R}^{d_1}$  is appended for the sequence of embedded patches, and an absolute position embeddings  $X_i^{POS}$  also added to retain positional information. The visual region of  $i$ -th node is represented as:

$$\mathbf{Z}_i = [\mathbf{x}_{CLS}; \mathbf{X}_i^{PC}] + X_i^{POS}, \quad (17)$$

where  $[\cdot]$  denotes a concatenation. Then, we feed the input matrix  $\mathbf{Z}_i$  into the CLIP vision encoder to

acquire the representation  $\hat{\mathbf{x}}_i^I$ . Note that the [CLS] token is utilized to serve as a representation of an entire image region:

$$\hat{\mathbf{x}}_i^I = \text{CLIP}(\mathbf{Z}_i)_{[\text{CLS}]}. \quad (18)$$

where  $\hat{\mathbf{x}}_i^I \in \mathbb{R}^{d_1}$ . Since the category label of each node can provide the auxiliary semantic information, a label embedding layer is built to embed the word label of each node into a feature vector. Given the one-hot vectors of the category label of each node, we first map it into an embedded feature vector  $\bar{\mathbf{x}}_i^I$  by an embedding matrix  $\mathbf{W}^{\text{label}} \in \mathbb{R}^{d_2 \times C_{\text{label}}}$ , where is initialized by Glove embedding (i.e.,  $d_2 = 300$ ),  $C_{\text{label}}$  is the number of categories. And then, the embedding features of the category label corresponding to the node are fused to the visual features to obtain the final visual node embedding:

$$\mathbf{x}_i^I = \text{Tanh}(\mathbf{W}_1[\hat{\mathbf{x}}_i^I; \bar{\mathbf{x}}_i^I]). \quad (19)$$

where  $\mathbf{W}_1 \in \mathbb{R}^{d_1 \times (d_1 + d_2)}$ .

**Textual Node Embedding** In TSG, we utilize CLIP as the underlying encoder to yield the basic contextualized word representations for each textual node:

$$\{\mathbf{x}_1^T, \dots, \mathbf{x}_m^T\} = \text{CLIP}(\{v_1, \dots, v_m\}), \quad (20)$$

where  $\mathbf{x}_i^T \in \mathbb{R}^{d_1}$ .

### A.3 Graph Encoding

In Section 3.2 and Section 3.3, we introduce a graph attention model (GAT) to encode the cross-modal graph (CMG) and refined graph ( $G^-$ ). Here, we provide a detail. Technically, given a graph  $G = (V, E)$ , where  $V$  is the set of nodes, and  $E$  is the set of edges. And the feature matrix  $\mathbf{X} \in \mathbb{R}^{|V| \times d_1}$  of  $V$  with  $d_1$ -dimensions. The hidden state  $\mathbf{h}_i$  of  $i$ -th node will be updated as follows:

$$\alpha_{i,j} = \frac{\exp(\text{LeakReLU}(\mathbf{W}_2[\mathbf{x}_i; \mathbf{x}_j]))}{\sum_{k \in \mathcal{N}(i)} \exp(\text{LeakReLU}(\mathbf{W}_2[\mathbf{x}_i; \mathbf{x}_k]))}, \quad (21)$$

$$\mathbf{h}_i = \text{ReLU}\left(\sum_j^{m+n} \alpha_{i,j} (\mathbf{W}_3 \mathbf{x}_j)\right), \quad (22)$$

where  $\mathcal{N}(i)$  denotes the neighbors of  $i$ -th node,  $\mathbf{W}_2$  and  $\mathbf{W}_3$  are learnable parameters. In short, we denote the graph encoding as follows:

$$\mathbf{H} = \text{GAT}(G, \mathbf{X}). \quad (23)$$

### A.4 Detailed GIB-guided Feature Refinement

**Introduction to GIB** Here, we provide more background information about the GIB principle. Given the original graph  $G$ , and the target  $Y$ , the goal of representation learning is to obtain the compressed graph  $G^-$  which is maximally informative

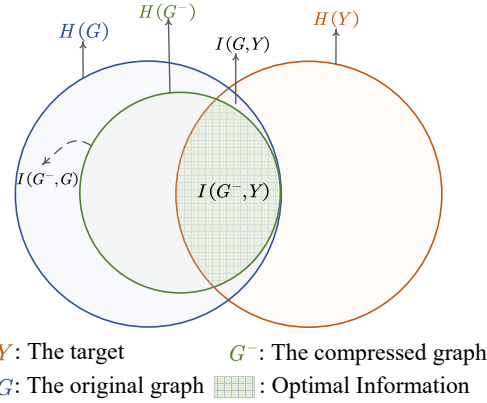


Figure 11: The Venn diagram visualization of GIB.

w.r.t  $Y$  (i.e., sufficiency,  $I(G, Y) = I(G^-, Y)$ ), and without any noisy information (i.e., minimality,  $I(G^-, G) - I(G^-, Y) = 0$ ), as indicated in Fig. 11. To encourage the information compressing process to focus on the target information, GIB was proposed to enforce an upper bound  $I_c$  to the information flow from the original graph to the compressed graph, by maximizing the following objectives:

$$\max_{G^-} I(G^-, Y) \text{ s.t. } I(G^-, G) \leq I_c. \quad (24)$$

Eq. (24) implies that a compressed graph can improve the generalization ability by ignoring irrelevant distractors in the original graph. By using a Lagrangian objective, GIB allows the  $G^-$  to be maximally expressive about  $Y$  while being maximally compressive about  $G$  by:

$$\max_{G^-} I(G^-, Y) - \beta I(G^-, G), \quad (25)$$

where  $\beta$  is the Lagrange multiplier. For the sake of consistency with the main body of the paper, the objective can be rewritten to:

$$\min_{G^-} -I(G^-, Y) + \beta I(G^-, G). \quad (26)$$

However, the GIB objective in Eq. (26) is notoriously hard to optimize due to the intractability of mutual information and the discrete nature of irregular graph data. By assuming that there is no information loss in the encoding process (Tian et al., 2020), the graph representation  $\mathbf{z}$  of  $G^-$  is utilized to optimize the GIB objective in Eq. (1), leading to  $-I(G^-, Y) \sim -I(\mathbf{z}, G)$ ,  $I(G^-, G) \sim I(\mathbf{z}, Y)$ . Therefore, the Eq. (26) can be computed as:

$$-I(G^-, Y) + \beta I(G^-, G) \sim -I(\mathbf{z}, Y) + \beta I(\mathbf{z}, G). \quad (27)$$

**Attention Operation for Node Filtering and Edge Adjusting** In Section 3.3, we utilize the  $l$ -order context to determine whether a node should be filtered or an edge should be adjusted since the nodes and edges in a graph have local dependence,

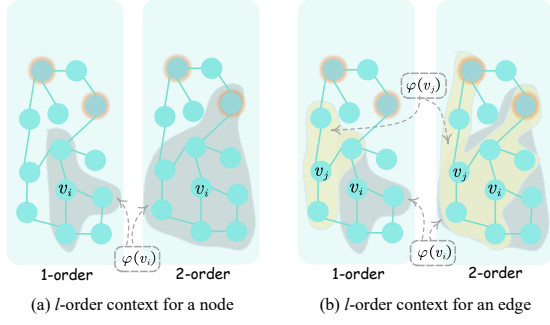


Figure 12: The  $l$ -order context for a node and an edge.

as shown in Fig. 12. Here, we give a detail of the calculation for the  $\text{Att}(\cdot)$  operation in Eq. (8) and Eq. (9). In Eq. (8), the attention operation can be computed as:

$$\begin{aligned} \alpha_{i,k}^v &= \frac{\exp(\mathbf{W}_4[\mathbf{h}_i; \mathbf{h}_k])}{\sum_{c \in \Phi(\{v_i, \varphi(v_i)\})} \exp(\mathbf{W}_4[\mathbf{h}_i; \mathbf{h}_c])} \\ \mathbf{r}_i^v &= \text{Tanh}\left(\sum_{k \in \Phi(\{v_i, \varphi(v_i)\})} \alpha_{i,k}^v (\mathbf{W}_5 \mathbf{h}_k)\right) \end{aligned} \quad (28)$$

where  $\Phi(\{v_i, \varphi(v_i)\})$  is a function to retrieve the index of a node in a set. Similarly, we consider the  $l$ -order context to calculate the  $\mathbf{r}_{i,j}^e$  in Eq.(9):

$$\begin{aligned} \alpha_{i,j,k}^e &= \frac{\exp(\mathbf{W}_6[\mathbf{h}_i; \mathbf{h}_j; \mathbf{h}_k])}{\sum_{c \in \Phi(\{v_i, \varphi(v_i), v_j, \varphi(v_j)\})} \exp(\mathbf{W}_6[\mathbf{h}_i; \mathbf{h}_j; \mathbf{h}_c])} \\ \mathbf{r}_{i,j}^e &= \text{Tanh}\left(\sum_{k \in \Phi(\{v_i, \varphi(v_i), v_j, \varphi(v_j)\})} \alpha_{i,j,k}^e (\mathbf{W}_7 \mathbf{h}_k)\right) \end{aligned} \quad (29)$$

**Detailed GIB Optimization** First, we examine the second term  $I(\mathbf{z}, G)$  in Eq. (11). Same as Sun et al. (2022), we employ variational inference to compute a variational upper bound for  $I(\mathbf{z}, G)$  as follow:

$$I(\mathbf{z}, G) \leq \int p(\mathbf{z}|G) \log \frac{p(\mathbf{z}|G)}{r(\mathbf{z})} d\mathbf{z} dG, \quad (30)$$

where  $r(\mathbf{z})$  is the variational approximation to the prior distribution  $p(\mathbf{z})$  of  $\mathbf{z}$ , which is treated as a fixed  $d_1$ -dimensional spherical Gaussian as in Alemi et al. (2017), i.e.,  $r(\mathbf{z}) = N(\mathbf{z}|0, \mathbf{I})$ . We use reparameterization trick ((Kingma and Welling, 2014)) to sample  $\mathbf{z}$  from the latent distribution according to  $p(\mathbf{z}|G)$ , i.e.,  $p(\mathbf{z}|G) = N(\boldsymbol{\mu}_z, \boldsymbol{\sigma}_z)$ , where  $\boldsymbol{\mu}_z$  and  $\boldsymbol{\sigma}_z$  is the mean vector and the diagonal co-variance matrix of  $\mathbf{z}$ , which can be computed as:

$$\boldsymbol{\mu}_z = \text{FFN}(\mathbf{a}); \boldsymbol{\sigma}_z = \text{Softplus}(\text{FFN}(\mathbf{a})), \quad (31)$$

where  $\mathbf{a}$  is the context feature of  $G^-$  obtained from Eq.(10).  $\mathbf{z}$  is sampled by  $\mathbf{z} = \boldsymbol{\mu}_z + \boldsymbol{\sigma}_z \cdot \varepsilon$ , where  $\varepsilon \in N(0, \mathbf{I})$ . We could reach the following optimization to approximate  $I(\mathbf{z}, G)$ :

$$I(\mathbf{z}, G) = \text{KL}(p(\mathbf{z}|G)||r(\mathbf{z})), \quad (32)$$

where  $\text{KL}(\cdot||\cdot)$  is the Kullback Leibler (KL) divergence (Hershey and Olsen, 2007).

Then, we examine the first term in Eq. (11), which encourages  $\mathbf{z}$  to be informative to  $Y$ . We expand  $I(\mathbf{z}, Y)$  as:

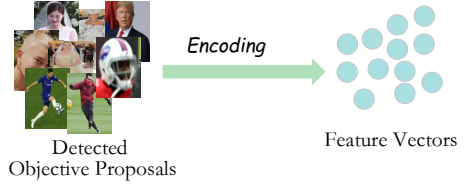
$$\begin{aligned} -I(\mathbf{z}, Y) &\leq -\int p(Y, \mathbf{z}) \log q(Y|\mathbf{z}) dY d\mathbf{z} + H(Y) \\ &:= \mathcal{L}_{\text{CE}}(q(Y|\mathbf{z}), Y), \end{aligned} \quad (33)$$

where  $q(Y|\mathbf{z})$  is the variational approximation of the true posterior  $p(Y, \mathbf{z})$ . Eq. (33) indicates that minimizing  $-I(\mathbf{z}, Y)$  is achieved by minimization of the classification loss between  $Y$  and  $\mathbf{z}$ , we model it as an MLP classifier with parameters. The MLP classifier takes  $\mathbf{z}$  as input and outputs the predicted label.

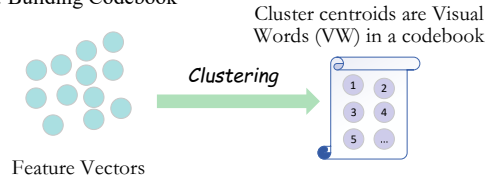
Step 1: Detecting Objective Proposal



Step 2: Featuring Objective Proposal



Step 3: Building Codebook



Step 4: Representing Images

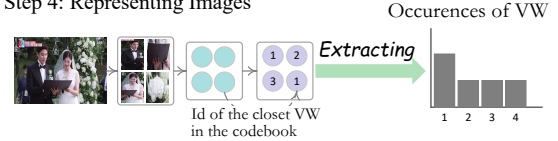


Figure 13: The extraction process of visual BoW features.

## A.5 Detailed Latent Multimodal Topic Modeling

**Visual BoW Feature Extraction** As mentioned in Section 2.3, we represent image  $I$  with visual BoW (VBoW) features. Here, we introduce how to extract VBoW features from an image. We compute the objective-level visual words in the following four steps, as shown in Fig. 13:

- **Step 1: Detecting Objective Proposal.** We first employ a Faster-RCNN (Ren et al., 2015) as an objective detector to extract all the objective proposals in the training dataset.
- **Step 2: Featuring Objective Proposal:** We use a pre-trained vision language model to obtain the feature descriptors (vectors) of each objective proposal.
- **Step 3: Building the Codebook:** After obtaining the feature vectors, these feature vectors are clustered by a kmeans algorithm, where the number of clusters is set to 2,000. Cluster centroids are taken as visual words.
- **Step 4: Representing Images:** Similar to the extraction of Bag-of-word (BoW) features for text representation, we build the Visual Bag-of-Word (VBoW) features for images. Specifically, using this codebook, each feature vector of the objective proposal in an image is replaced with the id of the nearest learned visual word.

### Detailed Latent Topic Modelling Optimization

In Section 2.3, we directly provide the optimal objective. In the following, we introduce how to optimize LAMO concretely. First of all, the prior parameters of  $\theta$ ,  $\mu$  and  $\sigma$  are estimated from the input data and defined as:

$$\mu = f_{\mu}(f(\mathbf{H})), \log\sigma = f_{\sigma}(f(\mathbf{H})), \quad (34)$$

where  $\mathbf{H}$  is the contextualized representation obtained from CMG,  $f(\cdot)$  is an aggregation function, and  $f_*(\cdot)$  is a neural perceptron that linearly transforms inputs, activated by a non-linear transformation. Note that we can generate the latent topic variable  $\varpi$  from  $p(\theta|T, I)$  by sampling, i.e.,  $\varpi = \mu + \sigma \cdot \varepsilon$ , where  $\varepsilon \in \mathcal{N}(0, \mathbf{I})$ . Then we employ Gaussian softmax to draw topic distribution  $\theta$ :

$$\theta = \text{Softmax}(\text{FFN}(\varpi)) \quad (35)$$

Similar to previous neural topic models only for handling text (Bianchi et al., 2021), we consider autoregressively reconstructing the textual and visual BoW features of input by learned topic distribution  $\theta$ :

$$p(\mathbf{b}_i^T | \chi, \theta) = \text{Softmax}(\theta \cdot \chi | \mathbf{b}_{<i}^T), \quad (36)$$

$$p(\mathbf{b}_i^T | \psi, \theta) = \text{Softmax}(\theta \cdot \psi | \mathbf{b}_{<i}^T). \quad (37)$$

The objective function of latent multimodal topic modeling is to maximize the evidence lower bound

(ELBO), as derived as follows:

$$\begin{aligned} \mathcal{L}_{LAMO} = & KL(q(\theta) || p(\theta|T, I)) \\ & - \mathbb{E}_{q(\theta)} [p(\mathbf{b}^T | \theta, \chi)] \\ & - \mathbb{E}_{q(\theta)} [p(\mathbf{b}^I | \theta, \psi)] \\ = & \mathcal{L}_{KL} + \mathcal{L}_{RecT} + \mathcal{L}_{RecI}, \end{aligned} \quad (38)$$

where  $q(\theta)$  is the prior probability of  $\theta$ , set as a standard Normal prior  $\mathcal{N}(0, \mathbf{I})$ .

## B Extended Experiments Setting

### B.1 Baselines

We compare our model with two categories of baseline systems.

**Text-based Methods,** which only leverage the texts of MRE data.

- **BERT** (Devlin et al., 2019) is only fine-tuned on the dataset by Zheng et al. (2021a).
- **PCNN** (Zeng et al., 2015) leverages external knowledge graphs to extract relations in a distantly supervised manner, which is employed in MRE dataset by Zheng et al. (2021a).
- **MTB** (Soares et al., 2019) is a RE-oriented pretraining model based on BERT, which is applied in MRE dataset by Zheng et al. (2021a).
- **DP-GCN** (Yu et al., 2020) propose dynamical pruning GCN for relation extraction, we re-implemented the framework and apply it to the MRE dataset.

**Multimodal Methods,** which utilize the additional visual information to enhance the textual RE.

- **BERT+SG** (Zheng et al., 2021a) simply concatenate the textual representation with visual features extracted.
- **MEGA** (Zheng et al., 2021a) leverage the alignment between textual and visual graphs to learn better semantic representation for MRE.
- **VisualBERT** (Li et al., 2019) is a single-stream structure via self-attention to discover implicit alignments between language and vision, which is then fine-tuned on the MRE dataset by Chen et al. (2022a).

- **ViLBERT** (Lu et al., 2019) consider employing two parallel streams for visual and language processing, which is then fine-tuned on the MRE dataset by Chen et al. (2022a).
- **HVPNet** (Chen et al., 2022b) propose to incorporate visual features into each self-attention layer of BERT.
- **MKGformer** (Chen et al., 2022a) introduce a hybrid transformer architecture, in which the underlying two encoders are utilized to capture basic textual and visual features, and the upper encoder to model the interaction features between image and text.
- **RDS** (Xu et al., 2022) design a data discriminator via reinforcement learning to determine whether data should utilize additional visual information for the relation inference.

## B.2 Calculating Text-image Relevance

In Fig. 10 we measure the relevance of input text-image pairs. Technically, we adopt the CLIP model to yield a vision-language matching score. Instead of directly feeding the whole picture and sentence into CLIP, we take a finer-grained method. Because in the MRE data, the picture and sentence pair collected from social media sources comes with low correlations, and if directly measuring their relevance at the instance level, our preliminary experiment shows that the highest text-image relevance score by CLIP is only 45%. Thus, we measure the picture and sentence pair by matching their correspondence of the VSG and TSG structures. We take their object nodes and the attribute nodes at the treatment targets, and calculate the vision-language pairs with CLIP at the node level:

$$\Psi(I, T) = \frac{1}{Z} \sum_{i,j} \text{CLIP}(\mathbf{x}_i^I, \mathbf{x}_j^T | G^I, G^T), \quad (39)$$

where  $Z$  is the normalization term.



## ACL 2023 Responsible NLP Checklist

---

### A For every submission:

- A1. Did you describe the limitations of your work?  
8
- A2. Did you discuss any potential risks of your work?  
8
- A3. Do the abstract and introduction summarize the paper’s main claims?  
1
- A4. Have you used AI writing assistants when working on this paper?  
*Left blank.*

### B Did you use or create scientific artifacts?

4

- B1. Did you cite the creators of artifacts you used?  
4
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?  
4
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?  
4
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?  
4
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?  
4
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.  
4

### C Did you run computational experiments?

4

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?  
4.1

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

4.1

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

4.2

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

4.1

**D  Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Left blank.*

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

*No response.*

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

*No response.*

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

*No response.*

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

*No response.*

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

*No response.*