Shap-CA: Shapley Value-based Contrastive Alignment for Multimodal Information Extraction

Anonymous ACL submission

Abstract

Recently, Multimodal Information Extraction (MIE) has attracted a lot of attention. Most of the existing methods focus on direct Image-Text interactions and face significant challenges due to the semantic and modality gaps between images and text. In this paper, we introduce a new paradigm of Image-Context-Text interaction, leveraging large multimodal models (LMMs) to generate descriptive textual context as a bridge to address these gaps. Following this paradigm, we propose a novel method, Shapley Value-based Contrastive Alignment (Shap-CA), which aligns both context-text and context-image pairs. Shap-CA first applies the Shapley value to measure the individual contribution of each element in context-text/context-image pairs to the overall semantic/modality overlaps, and then employs a contrastive learning strategy to maximize the contributions from relevant pairs and minimize those from irrelevant ones. Furthermore, we incorporate an adaptive fusion module for selective cross-modal fusion. Extensive experiments across four MIE datasets demonstrate that our method significantly outperforms existing state-of-the-art methods. Code will be released upon acceptance.

1 Introduction

The exponential growth of social media platforms has initiated a new phase of communication, characterized by the sharing of multimodal data, primarily texts and images. This diverse landscape necessitates advanced techniques for multimodal information extraction (MIE) (Wang et al., 2022a; Chen et al., 2022a; Yuan et al., 2023), which primarily aims to utilize auxiliary image inputs to enhance the performance of identifying entities/relations within the unstructured text.

To the best of our knowledge, the majority of previous methods on MIE mainly concentrate on the direct interaction between images and text. These approaches either (1) encode images directly and then employ efficient attention mechanisms to facilitate image-text interactions (Lu et al., 2018; Moon et al., 2018; Arshad et al., 2019; Yu et al., 2020; Sun et al., 2021; Wang et al., 2022c; Xu et al., 2022), or (2) recognize finer-grained visual objects from images and use Graph Neural Networks or attention mechanisms to foster object-text interactions (Wu et al., 2020; Zheng et al., 2020; Zhang et al., 2021a; Zheng et al., 2021a; Chen et al., 2022b,a; Jia et al., 2022, 2023; Yuan et al., 2023; Chen et al., 2023).

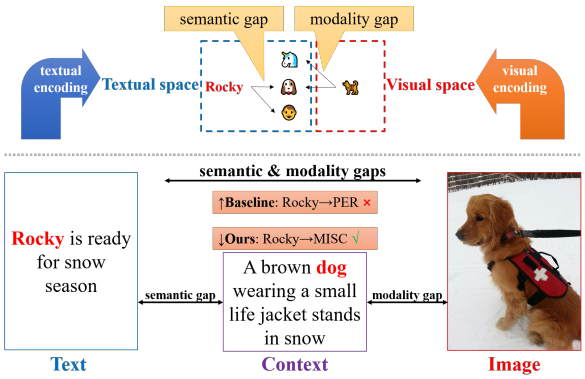

Figure 1: The semantic and modality gaps.

Despite the remarkable advancements made by these methods, this direct Image-Text interaction paradigm suffers from the simultaneous presence of semantic and modality gaps. As shown in the upper half of Figure 1, "Rocky" in the post refers to the name of a dog. The semantic gap refers to the meaning gap between the literal expression and what the image refers to, as "Rocky" can also be a cat or a person. The modality gap results from the under-explored problem of how to map text and images into a unified feature space. The coexistence of these two gaps weaken the link between images and text, potentially leading to erroneous predictions of entities or relations.

While the field of generative large multimodal

models (LMMs) has experienced rapid development, partially bridging the two kinds of gaps mentioned, they still exhibit restricted performance on MIE (Sun et al., 2023) due to their open-ended generative nature similar to large language models (LLMs). However, LMMs demonstrate impressive results on instruction-following and visual comprehension (Liu et al., 2023b,a), making them excel at generating descriptive textual context for images. This inspires us to utilize the generated descriptive context as the intermediate form to fill in the gaps. In this way, we can make use of the powerful generative ability of LMMs and reduce the burden of aligning image and text directly. As shown in Figure 1, the textual context only exhibits a semantic gap with the text and a modality gap with the image, forming the Image-Context-Text interaction paradigm and beneficial for cross-modal interpretation and understanding.

Following the new interaction paradigm, we propose a novel Shapley Value-based Contrastive Alignment (Shap-CA) method, which leverages the context to construct more coherent and compatible representations for MIE. The Shapley value, derived from cooperative game theory (Dubey, 1975), offers an equitable measure of each player's value based on their individual contribution to the overall cooperation. Applying this principle, we adapt the Shapley value to the alignment of both context-text and context-image pairs. In a set of context-text or context-image pairs, each element is treated as a player. Shap-CA first uses the Shapley value to evaluate the contribution of every player to the overall semantic or modality overlap. Then, a contrastive learning strategy is employed to maximize the contribution from relevant pairs, while minimizing the interactive contributions from irrelevant pairs. This alignment method effectively strengthens the links between contexts and their corresponding texts/images, thereby bridging semantic and modality gaps. Next, we further design an adaptive fusion module to obtain the informative fused features across modalities. This module assesses the relevance of each modal feature to the bridging context, strategically weighting their importance to achieve a finer-grained selective cross-modal fusion. Finally, a linear-chain CRF (Lafferty et al., 2001) or a word-pair contrastive layer is employed for prediction.

The main contributions of this paper can be summarized as follows:

- We are the first to introduce the Image-Context-Text interaction paradigm and leverage LMMs to generate descriptive context as a bridge to mitigate semantic and modality gaps for MIE.
- We propose a novel Shapley value-based contrastive alignment method, capturing semantic and modality relationships within and across image-text pairs for coherent and effective multimodal representations.
- Extensive experiments demonstrate that our method substantially outperforms existing state-of-the-art methods on four MIE datasets.

2 Related Work

Multimodal Information Extraction MIE is a field of study that mainly strives to augment the identification of entities and relations by utilizing additional image inputs, which primarily consists of three tasks: multimodal named entity recognition (MNER), multimodal relation extraction (MRE) and multimodal joint entity-relation extraction (MJERE). Specifically, **MNER** (Wang et al., 2022a; Sun et al., 2021) and **MRE** (Zheng et al., 2021a; Chen et al., 2022a) are concerned with identifying entities and relations separately, while **MJERE** (Yuan et al., 2023) aims to extract entities and their associated relations jointly. Most of existing methods (Lu et al., 2018; Moon et al., 2018; Yu et al., 2020; Zheng et al., 2020; Wu et al., 2020; Zheng et al., 2021a; Sun et al., 2021; Xu et al., 2022; Yuan et al., 2023; Chen et al., 2023) have been dedicated to the paradigm of direct Image-Text interaction. For instance, Sun et al. (2021) developed a pretrained multimodal BERT model to control the images' effect on text, and Yuan et al. (2023) designed an edge-enhanced graph network to facilitate the alignment between visual objects and the text. Despite these significant strides, these approaches overlook the potential dual gaps present between images and text. There are some other studies (Wang et al., 2022c,a) that leverage external knowledge bases to facilitate model reasoning. From a knowledge perspective, we regard LMM as a knowledge base (Petroni et al., 2019) to generate informative bridging context to enhance model performance.

Large Multimodal Models Large multimodal models have recently gained substantial traction in the research community (Alayrac et al., 2022; Li

et al., 2023a). Similar to the trend observed with large language models, several studies have indicated that scaling up the training data (Bai et al., 2023; Zhao et al., 2023; Dai et al., 2023) or model size (Bai et al., 2023; Lu et al., 2023) can significantly enhance these large multimodal models' capabilities. Moreover, visual instruction tuning (Liu et al., 2023b) can equip large multimodal models with excellent instruction-following, visual understanding, and natural language generation abilities. This advancement empowers these models to excel in interpreting images according to instructions and generating informative textual contexts. However, their open-ended generative characteristics have resulted in less than satisfactory performance when directly applied to information extraction tasks (Sun et al., 2023).

3 Methodology

3.1 Task Definition

Given an input text $\boldsymbol{t} = \{t_1, \ldots, t_{n_t}\}$ with n_t tokens and its attached image $\boldsymbol{I}$, our method aims to predict the output entity/relation labels $\boldsymbol{y}$. The format of the labels $\boldsymbol{y}$ is task-dependent. Specifically, for MNER, they are sequential labels. For MRE and MJERE, they are word-pair labels (Yuan et al., 2023).

3.2 Overview

The architecture of Shap-CA is shown in Figure 2. Initially, we utilize a LMM, e.g., LLaVA-1.5 (Liu et al., 2023a), to extract the textual bridging context from the image. Subsequently, we employ a pretrained textual transformer to extract features from the text and the context. In parallel, an image encoder is used to derive visual features from the image. Following this, we apply a Shapley value-based contrastive alignment to construct more coherent representations. The adaptive fusion module is then employed to obtain the informative features across modalities. Finally, these comprehensive representations are fed into a CRF or a word-pair contrastive layer for prediction.

3.3 Encoding Module

Context Generation Given an image $\boldsymbol{I}$, we utilize a pretrained LMM to generate a textual context as the bridging context $\boldsymbol{c} = \{c_1, \ldots, c_{n_c}\}$.

Textual and Visual Encoding In order to acquire contextualized textual features, we concatenate the input text $\boldsymbol{t}$ with the bridging context $\boldsymbol{c}$. Following this, we employ a pretrained textual transformer to extract both sentence-level and token-level features:

$$\boldsymbol{x_t}, \boldsymbol{H_t}, \boldsymbol{x_c}, \boldsymbol{H_c} = \text{Transformer}([\boldsymbol{t}; \boldsymbol{c}]) \quad (1)$$

where $\boldsymbol{x_t}, \boldsymbol{x_c} \in \mathbb{R}^d$ represent the sentence-level features, while $\boldsymbol{H_t} \in \mathbb{R}^{n_t \times d}$, $\boldsymbol{H_c} \in \mathbb{R}^{n_c \times d}$ denote the token-level features of the input text and context, respectively. Simultaneously, we employ an image encoder to extract the visual features of the image $\boldsymbol{I}$:

$$\boldsymbol{x_v}, \boldsymbol{H_v} = \text{ImageEncoder}(\boldsymbol{I}) \quad (2)$$

where $\boldsymbol{x_v}, \boldsymbol{H_v}$ denotes the global visual feature and the regional visual features of the image, respectively.

3.4 Shapley Value-based Contrastive Alignment

To construct more coherent representations, we leverage the Shapley value to perform both context-text alignment and context-image alignment. The Shapley value, originating from cooperative game theory (Dubey, 1975), offers a solution for the equitable allocation of total benefits derived from cooperation among players, based on their individual marginal contributions. The Shapley value's desirable mathematical properties have led to its widespread application in diverse fields, from economics (Gul, 1989) to machine learning (Ghorbani and Zou, 2019; Jia et al., 2019b). This work is the first to utilize the Shapley value for multimodal alignment through contrastive learning (Chen et al., 2020).

3.4.1 Preliminary

In the context of a cooperative game, suppose we have k players, represented by $\boldsymbol{K} = \{1, \ldots, k\}$, and a utility function $\boldsymbol{u} : 2^k \rightarrow \mathbb{R}$ that assigns a reward to each coalition (subset) of players. The Shapley value of player i is then defined as (Jia et al., 2019a):

$$\phi_i(\boldsymbol{u}) = \frac{1}{k} \sum_{\boldsymbol{S} \subseteq \boldsymbol{K} \setminus \{i\}} \frac{1}{\binom{k-1}{|\boldsymbol{S}|}} [\boldsymbol{u}(\boldsymbol{S} \cup \{i\}) - \boldsymbol{u}(\boldsymbol{S})] \quad (3)$$

The Shapley value essentially quantifies the average marginal contribution of a player to all potential coalitions (subsets). We detail the context-text alignment as an example as follows.

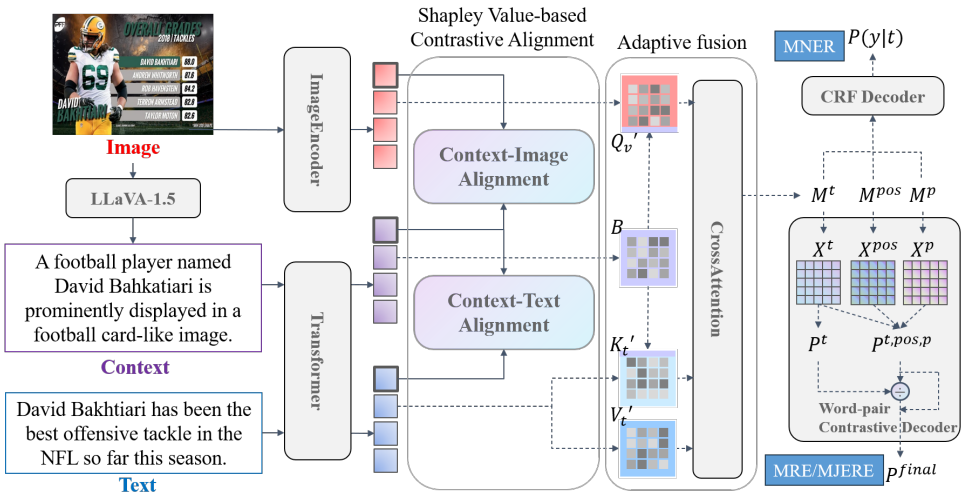

Figure 2: The overall architecture of Shap-CA.

3.4.2 Context-Text Alignment

Shapley Value In the context-text alignment, inputs are a mini-batch of k context-text pairs $\{(\boldsymbol{x_c}^a, \boldsymbol{x_t}^a)\}_{a=1}^k$. Here, we view the k bridging contexts as players, denoted as $\boldsymbol{K} = \{1, \ldots, k\}$ for simplicity. These players, or contexts, collaboratively contribute to the semantic comprehension of a specific text feature. Consider the j-th pooled text feature, $\boldsymbol{x_t}^j$, and a selected subset of context players, denoted as a coalition $\boldsymbol{S} \subseteq \boldsymbol{K}$. The central idea is based on an assumption: if all the contexts within the subset $\boldsymbol{S}$ and the text $\boldsymbol{x_t}^j$ form positive pairs, the utility of $\boldsymbol{S}$ for $\boldsymbol{x_t}^j$ would be represented by the expected semantic overlap between them. This utility captures the collective semantic relationships between the text and the contexts within the coalition, as formalized by:

$$\begin{aligned} \boldsymbol{u}_j(\boldsymbol{S}) &= \sum_{i \in \boldsymbol{S}} p_i \mathrm{sim}(\boldsymbol{x_t}^j, \boldsymbol{x_c}^i) \\ p_i &= \frac{e^{\mathrm{sim}(\boldsymbol{x_t}^j, \boldsymbol{x_c}^i)/\tau}}{\sum_{a \in \boldsymbol{S}} e^{\mathrm{sim}(\boldsymbol{x_t}^j, \boldsymbol{x_c}^a)/\tau}} \end{aligned} \tag{4}$$

Here, $\mathrm{sim}(\boldsymbol{x_t}^j, \boldsymbol{x_c}^i)$ denotes the semantic overlap between the text and each context (i.e., individual semantic contribution), measured by cosine similarity. The weight p_i, computed through a softmax operation with a temperature of τ, models the cooperative behavior among different contexts by normalizing these individual semantic contribution. This approach intuitively suggests that the stronger the semantic overlap a context shares with the text (i.e., the larger the semantic contribution), the more likely it is to form a positive pair with the text in real-world situations. From this perspective, the utility of the coalition can be interpreted as an expectation over the semantic overlaps of each context within $\boldsymbol{S}$ with the text, where the expectation weights are given by the likelihood of each context-text pair being positive. This method naturally prioritizes contexts that have a higher degree of semantic overlap with the text, thereby refining the overall semantic understanding.

However, as indicated by Eqn. 3, the computation of the Shapley value requires an exponentially large number of computations relative to the size of the mini-batch, which poses a challenge during training. To address this, we extend Monte-Carlo approximation methods (Castro et al., 2009; Maleki et al., 2013) to our training setting for estimating the Shapley value. The algorithm is shown in Appendix A.

Contrastive Learning After acquiring the approximated Shapley value $\{\hat{\phi}_1(\boldsymbol{u}_j), \ldots, \hat{\phi}_k(\boldsymbol{u}_j)\}$, we introduce a context-to-text contrastive loss that aims to maximize the average marginal semantic contribution from each context player to the text with which it forms a true positive pair (i.e., the text from the same pair), while simultaneously minimizing the contributions between the true negative pairs (i.e., all other texts not paired with this context).

$$\mathcal{L}_{c2t} = -\frac{1}{k} \sum_{j=1}^{k} \left[\hat{\phi}_j(\boldsymbol{u}_j) - \sum_{i \neq j} \hat{\phi}_i(\boldsymbol{u}_j) \right] \tag{5}$$

In a symmetrical manner, we can treat the k pooled text as players, and derive the text-to-context contrastive loss $\mathcal{L}_{t2c}$. The semantic loss $\mathcal{L}_{semantic}$ is the average of these two contrastive losses:

$$\mathcal{L}_{semantic} = \frac{1}{2}(\mathcal{L}_{c2t} + \mathcal{L}_{t2c}) \tag{6}$$

3.4.3 Context-Image Alignment

Similarly, we can employ Alg. 1 to estimate the Shapley value for a mini-batch of k context-image pairs $\{(\boldsymbol{x_c}^a, \boldsymbol{x_v}^a)\}_{a=1}^k$ and derive the context-to-image loss $\mathcal{L}_{c2v}$ and the image-to-context loss $\mathcal{L}_{v2c}$. The modality loss $\mathcal{L}_{modality}$ is the average of these two contrastive loss:

$$\mathcal{L}_{modality} = \frac{1}{2}(\mathcal{L}_{c2v} + \mathcal{L}_{v2c}) \tag{7}$$

3.5 Adaptive Fusion

To facilitate a finer-grained fusion across modalities, we develop an adaptive attention fusion module. This module dynamically weights the importance of different features in two modalities, based on their relevance to the context that connects them. Given the representations $\boldsymbol{H_t}, \boldsymbol{H_c}, \boldsymbol{H_v}$ obtained from Section 3.3, we initially employ linear projection to transform them into a set of matrices: a query matrix $\boldsymbol{Q_v} \in \mathbb{R}^{n_v \times D}$, a key matrix $\boldsymbol{K_t} \in \mathbb{R}^{n_t \times D}$, a value matrix $\boldsymbol{V_t} \in \mathbb{R}^{n_t \times D}$, and a context matrix $\boldsymbol{C} \in \mathbb{R}^{n_c \times D}$. Subsequently, we calculate the bridging term $\boldsymbol{B}$ as follows:

$$\boldsymbol{B} = \text{mean}\left(\frac{\boldsymbol{C}^\top \boldsymbol{C}}{\sqrt{D}}\right) \in \mathbb{R}^D \tag{8}$$

This bridging term is then employed to dynamically modify the content of queries $\boldsymbol{Q_v}$ and keys $\boldsymbol{K_t}$ based on their relevance to the bridging term:

$$\begin{aligned} \boldsymbol{Q_v}' &= \boldsymbol{g}_q \odot \boldsymbol{Q_v} + (1 - \boldsymbol{g}_q) \odot \boldsymbol{B} \\ \boldsymbol{K_t}' &= \boldsymbol{g}_k \odot \boldsymbol{K_t} + (1 - \boldsymbol{g}_k) \odot \boldsymbol{B} \end{aligned} \tag{9}$$

where $\odot$ represents the Hadamard product, $\boldsymbol{g}_q \in \mathbb{R}^{n_v \times D}$ and $\boldsymbol{g}_k \in \mathbb{R}^{n_t \times D}$ are gating vectors that are used to capture the relevance between the text (or image) and the bridging context:

$$\begin{aligned} \boldsymbol{g}_q &= \tanh(\text{Linear}_1(\boldsymbol{Q_v}, \boldsymbol{B})) \\ \boldsymbol{g}_k &= \tanh(\text{Linear}_2(\boldsymbol{K_t}, \boldsymbol{B})) \end{aligned} \tag{10}$$

Following this, we acquire the image-awared text features $\boldsymbol{M}^t \in \mathbb{R}^{n_t \times D}$ using the gated cross-attention:

$$\boldsymbol{M}^t = \text{CrossAttention}(\boldsymbol{Q_v}', \boldsymbol{K_t}', \boldsymbol{V_t}) \tag{11}$$

3.6 Classifier

In addition to the image-awared text features derived from Eqn. 11, we also incorporate each token's part-of-speech embeddings $\boldsymbol{M}^{pos} \in \mathbb{R}^{n_t \times d_1}$, and positional embeddings $\boldsymbol{M}^p \in \mathbb{R}^{n_t \times d_1}$, to enrich the information available for decoding. For MNER, a widely used CRF layer (Akbik et al., 2018; Kenton and Toutanova, 2019) is employed to predict label sequences. For MRE and MJERE, we propose a word-pair contrastive layer to predict word-pair labels (Yuan et al., 2023).

CRF layer Initially, we employ a multi-layer perceptron to integrate the three channels of features:

$$\boldsymbol{M} = \text{MLP}_1(\boldsymbol{M}^t; \boldsymbol{M}^{pos}; \boldsymbol{M}^p) \in \mathbb{R}^{n_t \times d_2} \tag{12}$$

Subsequently, we feed $\boldsymbol{M} = \{\boldsymbol{M}_1, \ldots, \boldsymbol{M}_{n_t}\}$ into a standard CRF layer to derive the final distribution $P(\boldsymbol{y}|\boldsymbol{t})$. The training loss for the input sequence $\boldsymbol{t}$ with gold labels $\boldsymbol{y}^*$ is measured using the negative log-likelihood (NLL):

$$\mathcal{L}_{main} = -\log P(\boldsymbol{y}^*|\boldsymbol{t}) \tag{13}$$

Word-Pair Contrastive layer For each word pair (t_i, t_j), we initially employ three multi-layer perceptrons to separately obtain the three channels of pair-wise features:

$$\boldsymbol{X}_{i,j}^l = \text{MLP}^l(\boldsymbol{M}_i^l; \boldsymbol{M}_j^l; \boldsymbol{M}_i^l - \boldsymbol{M}_j^l) \in \mathbb{R}^{d_2} \tag{14}$$

where $l \in \{t, pos, p\}$ represents the different feature channels. Following this, we use only the text features to generate the initial distribution $P_{i,j}^t$ and incorporate the three channels of features to derive the enhanced distribution $P_{i,j}^{t,pos,p}$ over the label set:

$$P_{i,j}^t = \text{Softmax}(\text{MLP}_2(\boldsymbol{X}_{i,j}^t)) \tag{15}$$

$$P_{i,j}^{t,pos,p} = \text{Softmax}(\text{MLP}_3(\boldsymbol{X}_{i,j}^t; \boldsymbol{X}_{i,j}^{pos}; \boldsymbol{X}_{i,j}^p)) \tag{16}$$

The final distribution is refined by contrasting the two distributions:

$$P_{i,j}^{final} = \text{Softmax}(P_{i,j}^{t,pos,p} + \lambda \log \frac{P_{i,j}^{t,pos,p}}{P_{i,j}^t}) \tag{17}$$

where λ is the refinement scale. We use the cross-entropy loss for the input sequence $\boldsymbol{t}$ and the gold word-pair labels $\boldsymbol{y}^*$:

$$\mathcal{L}_{main} = -\sum_{i=1}^{n_t}\sum_{j=1}^{n_t} \boldsymbol{y}_{i,j}^* \log(P_{ij}^{final}) \tag{18}$$

3.7 Training

In summary, our framework incorporates two self-supervised learning tasks and one supervised learning task, resulting in three loss functions. Considering the semantic loss $\mathcal{L}_{semantic}$, the modality loss $\mathcal{L}_{modality}$ from Sec. 3.4, and the prediction loss $\mathcal{L}_{main}$ from Sec. 3.6, the final loss function is defined as follows:

$$\mathcal{L} = \alpha\mathcal{L}_{semantic} + \beta\mathcal{L}_{modality} + \mathcal{L}_{main} \quad (19)$$

where α, β are hyperparameters.

4 Experiments

We conduct our experiments on four MIE datasets and compare our approaches with a number of previous approaches.

4.1 Datasets and Implementation Details

Experiments are conducted on 2 MNER datasets, 1 MRE dataset and 1 MJERE dataset: Twitter-15 (Zhang et al., 2018) and Twitter-17 (Yu et al., 2020) for MNER[1], MNRE[2] (Zheng et al., 2021b) for MRE, and MJERE[3] (Yuan et al., 2023) for MJERE, noting that the last dataset and the task share the same name. Statistical details of these datasets are provided in Appendix B. Implementation details are shown in Appendix C.

4.2 Baselines

We compare Shap-CA with previous state-of-the-art methods, which primarily fall into two categories. The first group of methods only consider text input, including BERT-CRF (Devlin et al., 2018), MTB (Baldini Soares et al., 2019), ChatGPT and GPT-4[4]. Secondly, we consider several latest multimodal methods, including OCSGA (Wu et al., 2020), UMGF (Zhang et al., 2021a), MEGA (Zheng et al., 2021a), MAF (Xu et al., 2022), ITA (Wang et al., 2022b), CAT-MNER (Wang et al., 2022c), MNER-QG (Jia et al., 2023), EEGA (Yuan et al., 2023), and MQA (Sun et al., 2023).

4.3 Main Results

Table 1 presents a comprehensive comparison between our proposed method and the baselines, including both text-based methods and previous state-of-the-art multimodal methods.

The first observation from Table 1 is the significant superiority of our method, Shap-CA, across all datasets. In the entity/relation extraction tasks, represented by Twitter-15, Twitter-17, MJERE_{e} and MNRE, Shap-CA outperforms the closest competitor by 1.33%, 1.07%, 0.7% and 1.05% F1 scores, respectively, underscoring the effectiveness of our proposed Image-Context-Text interactions. Furthermore, in the most challenging task—MJERE_{j}, which demands the simultaneous extraction of entities and their associated relations—Shap-CA excels, achieving the highest F1 score of 55.95%. This result underlines Shap-CA's potential in more complex MIE scenarios.

Next, it's evident that incorporating images generally enhances the performance of MIE when comparing the state-of-the-art multimodal methods to their text-only counterparts.

Finally, we compare methods based on LLMs (e.g., ChatGPT, GPT-4) and LMMs (e.g., MQA) with other approaches. Despite the common belief that larger models possess better generalization abilities, our experimental results indicate their underperformance compared to our method and other previous multimodal approaches. This suggests that existing large models, despite their extensive capabilities, might not yet be fully optimized for information extraction tasks, potentially due to their open-ended generative nature and pre-training scenarios that are misaligned with the specifics of MIE.

5 Analysis

5.1 Ablation Study

We conduct a comprehensive ablation study to further analyze the effectiveness of our method. The results of these model variants are also presented in Table 1.

Importance of Bridging Context Our method fundamentally relies on the bridging context to enhance cross-modal comprehension and interpretation. To evaluate its effectiveness, we introduce two variants: (1) Shap-CA w/o Context in Alignment, which directly aligns image-text inputs based on the Shapley value, and (2) Shap-CA w/o Context in Fusion, which simply applies vanilla cross-attention fusion without the context involved. Our results show that Shap-CA consistently outperforms these variations across all datasets. For instance, the F1

[1]The datasets are available at https://github.com/jefferyYu/UMT, https://github.com/Multimodal-NER/RpBERT

[2]https://github.com/thecharm/MNRE

[3]https://github.com/YuanLi95/EEGA-for-JMERE

[4]Please refer to Appendix D for detailed settings about LLMs.

Modality	Method	MNER		MJERE		MRE
		Twitter-15	Twitter-17	MJERE_j	MJERE_e	MNRE
Text	BERT-CRF	71.81	83.44	-	-	-
	MTB	-	-	-	-	60.86
	ChatGPT	50.21	57.50	13.37	51.74	35.20
	GPT4	57.98	66.61	21.51	56.63	42.11
Text + Image	OCSGA	72.92	-	49.64	76.07	-
	UMGF	74.85	85.51	51.45	76.75	65.29
	MEGA	72.35	84.39	53.18	77.22	66.41
	MAF	73.42	86.25	53.62	76.81	-
	ITA	75.60	85.72	-	-	66.89
	CAT-MNER	75.41	85.99	-	-	-
	MNER-QG	74.94	87.25	-	-	-
	EEGA	-	-	55.29	78.59	-
	MQA	50.6	62.6	-	-	61.6
	Shap-CA w/o Context in Alignment	74.87	86.19	54.07	77.74	66.29
	Shap-CA w/o Context in Fusion	75.24	87.31	53.35	77.49	66.41
	Shap-CA w/o Shapley Value	75.38	86.58	54.14	77.48	65.72
	Shap-CA w/o word-pair Contrastive	-	-	54.73	78.02	67.12
	Shap-CA	**76.93**	**88.32**	**55.95**	**79.29**	**67.94**

Table 1: Performance comparison (F1 score) of our approach and state-of-the-art approaches on four MIE datasets. Following Yuan et al. (2023), on the MJERE dataset, we demonstrate both joint entity-relation extraction and named entity recognition results, denoted as MJERE_j, MJERE_e respectively. The results of large language models on the MJERE dataset are reproduced by us.

score decreases from 76.93% to 74.87% on the Twitter-15 dataset when the context in alignment is removed. This highlights that the textual context can effectively bridge the semantic and modality gaps between images and text, demonstrating its value in MIE.

Role of Shapley Value To understand the impact of the Shapley value-based contrastive alignment, we replace it with InfoNCE (Oord et al., 2018), a widely used approach in the field of contrastive learning (Baevski et al., 2020; Ma et al., 2020; Morgado et al., 2021). As shown in Table 1, Shap-CA apparently outperforms the variant without the Shapley Value on all datasets, which validates our proposed alignment method's ability to construct more coherent multimodal representations for MIE.

Effectiveness of the Word-Pair Contrastive Layer We further delve into the impact of our proposed word-pair contrastive layer by evaluating the performance of Shap-CA w/o word-pair Contrastive. As expected, it can enhance the model's predictive power by contrasting the distributions before and after the integration of the three channels.

5.2 How Alignments Affect Performance

In this section, we further discuss the influence of alignments on our model's performance on the MJERE dataset. Specifically, we examine two key coefficients, α and β in Eqn. 19, which represent context-text alignment and context-image alignment, respectively. As shown in Figure 3, the model's performance exhibits a consistent pattern of variation in response to changes in α and β. We will discuss the trend using α as an example. $\alpha = 0$ means w/o context-text alignment, resulting in the poorest performance due to the inability to effectively bridge the semantic gap . When $\alpha = 0.2$, the model reaches its optimal performance. However, when α exceeds 0.2, the alignment task may interfere with the main task, leading to declining performance.

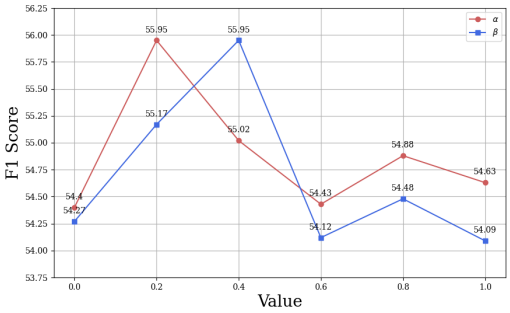

Figure 3: The model performance with different α or β

Generator	$MJERE_j$			$MJERE_e$		
	P	R	F1	P	R	F1
BU	54.11	56.92	55.48	76.57	80.92	78.69
VinVL	53.64	57.86	55.67	77.02	80.73	78.83
BLIP-2	**54.50**	57.08	55.76	**77.33**	80.83	79.03
LLaVA-1.5	54.03	**58.02**	**55.95**	77.19	**81.50**	**79.29**

Table 2: Performance effect from the different context generators.

	Twitter-17→Twitter-15			Twitter-15→Twitter-17		
	P	R	F1	P	R	F1
UMT	64.67	63.59	64.13	67.80	55.23	60.87
UMGF	67.00	62.18	66.21	69.88	56.92	62.74
FMIT	66.72	69.73	68.19	70.65	59.22	64.43
CAT-MNER	74.86	63.01	68.43	70.69	59.44	64.58
Shap-CA	72.79	65.83	**69.14**	71.27	59.95	**65.12**

Table 3: Comparison of the generalization ability. Results of other methods are from Li et al. (2020); Wang et al. (2022c)

5.3 Impact of Different Context Generators

We conduct experiments to explore the impact when using weaker context generators. In particular, we examine the results of four different context generators on the MJERE dataset. These include BU (Anderson et al., 2018), a pretrained image captioning model [5] proposed by (Luo et al., 2018) with the Bottom-Up features, VinVL (Zhang et al., 2021b), a captioning model finetuned on MS-COCO[6] (Lin et al., 2014), BLIP-2 (Li et al., 2023b), a computation-friendly pretrained Vision-Language model[7] and LLaVA-1.5 (Liu et al., 2023a), a newly proposed visual instruction-tuned, large multimodal model[8]. As shown in Table 2, the performance disparity across these context generators is minimal. This observation implies that our method is robust to the variance in context generators, maintaining consistent performance regardless of the quality of the generated context.

5.4 Generalization Analysis

Table 3 presents a comparison of our method's generalization ability against several previous state-of-the-art approaches. The experimental setup involves training on one dataset and testing on another. For instance, Twitter-17→Twitter-15 denotes that the model is trained on the Twitter-17 dataset and then tested on the Twitter-15 dataset. When examining the F1 score, our method significantly outperforms all other methods in both tasks. This exceptional performance highlights our method's strong generalization ability.

5.5 Case Study

In Appendix E, we present two cases that highlight the effectiveness of our proposed method in bridging both semantic and modality gaps, challenging for EEGA (Yuan et al., 2023) which follows a direct Image-Text interaction paradigm. The first case discusses Stephen Curry's achievements in the NBA. Although EEGA recognizes the player, it incorrectly extracts "NBA" as an entity, leading to an inaccurate "present_in" relation between "Stephen Curry" and "NBA". In contrast, our method, utilizing the context featuring a "trophy" and "celebration", successfully bridges the semantic and modality gaps, leading to an accurate prediction. The second case is more nuanced, involving a scene with Elizabeth Taylor and George Stevens. Though EEGA can identify both persons, it fails to infer the intimate relationship implied by the image and text, predicting a mere "peer" relation. However, our method, leveraging the context that emphasizes "intimacy and connection", effectively bridges the gaps and accurately predicts the relation "couple" between "Elizabeth Taylor" and "George Stevens".

6 Conclusion

In this paper, we introduce a novel paradigm of Image-Context-Text interaction to address the challenges posed by the semantic and modality gaps in traditional MIE approaches. Following this paradigm, we propose a novel method, Shapley Value-Based Contrastive Alignment (Shap-CA), which performs both context-text and context-image alignments. Shap-CA first employs the Shapley value to measure the individual contributions of context, text, and image to the overall semantic or modality overlaps, and then applies a contrastive learning strategy to perform the alignment by optimizing those contributions. Furthermore, Shap-CA incorporates an adaptive fusion module for selective cross-modal fusion. Our experiments demonstrate that Shap-CA significantly outperforms previous state-of-the-art approaches across four MIE datasets. In our analysis, we show Shap-CA's robust performance in the face of varying context quality and its strong generalization ability.

[5] https://github.com/ruotianluo/self-critical.pytorch

[6] https://github.com/microsoft/Oscar

[7] https://github.com/salesforce/LAVIS/tree/main/projects/blip2

[8] https://github.com/haotian-liu/LLaVA

Limitations

In this paper, we use the context generator trained on general-purpose datasets (e.g., ShareGPT[9]). Although effective, this approach may not fully leverage domain-specific nuances. The performance and efficiency of our method can be further enhanced by utilizing a context generator specifically trained on data from the relevant domain (e.g., Twitter domain). Besides, the MJERE dataset assumes a relation between each pair of entities, which ignores entities that do not have relations with others, leading to a potential discrepancy from the actual entity-relation distributions in real-world scenarios. We leave constructing a more realistic dataset and further verify the effectiveness of our method on it as future work.

Ethics Statement

In this paper, we use the publicly available datasets for our experiments. The context generator we use is also an open-source large multimodal model trained on publicly available datasets. Therefore, we believe that all data we use does not invade the privacy of any user.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Omer Arshad, Ignazio Gallo, Shah Nawaz, and Alessandro Calefati. 2019. Aiding intra-text representations with visual context for multimodal named entity recognition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 337–342. IEEE.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Javier Castro, Daniel Gómez, and Juan Tejada. 2009. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730.

Feng Chen and Yujian Feng. 2023. Chain-of-thought prompt distillation for multimodal named entity and multimodal relation extraction. *arXiv preprint arXiv:2306.14122*.

Feng Chen, Jiajia Liu, Kaixiang Ji, Wang Ren, Jian Wang, and Jingdong Wang. 2023. Learning implicit entity-object relations by bidirectional generative alignment for multimodal ner. *arXiv preprint arXiv:2308.02570*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

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 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 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 makes a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. *arXiv preprint arXiv:2205.03521*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[9] https://sharegpt.com/

Pradeep Dubey. 1975. On the uniqueness of the shapley value. *International Journal of Game Theory*, 4:131–139.

Amirata Ghorbani and James Zou. 2019. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pages 2242–2251. PMLR.

Faruk Gul. 1989. Bargaining foundations of shapley value. *Econometrica: Journal of the Econometric Society*, pages 81–95.

Meihuizi Jia, Lei Shen, Xin Shen, Lejian Liao, Meng Chen, Xiaodong He, Zhendong Chen, and Jiaqi Li. 2023. Mner-qg: An end-to-end mrc framework for multimodal named entity recognition with query grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8032–8040.

Meihuizi Jia, Xin Shen, Lei Shen, Jinhui Pang, Lejian Liao, Yang Song, Meng Chen, and Xiaodong He. 2022. Query prior matters: a mrc framework for multimodal named entity recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3549–3558.

Ruoxi Jia, David Dao, Boxin Wang, Frances A Hubis, Nezihe M Gürel, Bo Li, Ce Zhang, Costas J Spanos, and Dawn Song. 2019a. Efficient task-specific data valuation for nearest neighbor algorithms. *Proceedings of the VLDB Endowment*, 12(11):1610–1623.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J Spanos. 2019b. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1167–1176. PMLR.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

John D Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. 2023a. Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020*, 1(2).

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. Flat: Chinese ner using flat-lattice transformer. *arXiv preprint arXiv:2004.11795*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1990–1999, Melbourne, Australia. Association for Computational Linguistics.

Yadong Lu, Chunyuan Li, Haotian Liu, Jianwei Yang, Jianfeng Gao, and Yelong Shen. 2023. An empirical study of scaling instruct-tuned large multimodal models. *arXiv preprint arXiv:2309.09958*.

Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6964–6974.

Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. 2020. Active contrastive learning of audio-visual video representations. In *International Conference on Learning Representations*.

Sasan Maleki, Long Tran-Thanh, Greg Hines, Talal Rahwan, and Alex Rogers. 2013. Bounding the estimation error of sampling-based shapley value approximation. *arXiv preprint arXiv:1306.4265*.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity recognition for short social media posts. *arXiv preprint arXiv:1802.07862*.

Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. 2021. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12486.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rpbert: a text-image relation propagation-based bert model for multimodal ner. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13860–13868.

Yuxuan Sun, Kai Zhang, and Yu Su. 2023. Multimodal question answering for unified information extraction. *arXiv preprint arXiv:2310.03017*.

Xinyu Wang, Jiong Cai, Yong Jiang, Pengjun Xie, Kewei Tu, and Wei Lu. 2022a. Named entity and relation extraction with multi-modal retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5925–5936, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xinyu Wang, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang, and Kewei Tu. 2022b. ITA: Image-text alignments for multi-modal named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3176–3189, Seattle, United States. Association for Computational Linguistics.

Xuwu Wang, Jiabo Ye, Zhixu Li, Junfeng Tian, Yong Jiang, Ming Yan, Ji Zhang, and Yanghua Xiao. 2022c. Cat-mner: multimodal named entity recognition with knowledge-refined cross-modal attention. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. 2020. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1038–1046.

Bo Xu, Shizhou Huang, Chaofeng Sha, and Hongya Wang. 2022. Maf: a general matching and alignment framework for multimodal named entity recognition. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 1215–1223.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3342–3352, Online. Association for Computational Linguistics.

Li Yuan, Yi Cai, Jin Wang, and Qing Li. 2023. Joint multimodal entity-relation extraction based on edge-enhanced graph alignment network and word-pair relation tagging. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11051–11059.

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021a. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14347–14355.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021b. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Bo Zhao, Boya Wu, and Tiejun Huang. 2023. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*.

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 29th ACM International Conference on Multimedia*, 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 *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Changmeng Zheng, Zhiwei Wu, Tao Wang, Yi Cai, and Qing Li. 2020. Object-aware multimodal named entity recognition in social media posts with adversarial learning. *IEEE Transactions on Multimedia*, 23:2520–2532.

A Shapley Value Approximation Algorithm

We present the detailed algorithm in Alg. 1. It begins with a random permutation of k context players and an initial stride of s, which is set as $\text{batch\_size}/2$. At each iteration, the algorithm scans each player in the current permutation and calculates the marginal contribution when the player is added to the coalition formed by the preceding players. This marginal contribution serves as an unbiased estimate of the Shapley value for that player, improving with each iteration. Subsequently, the algorithm updates the permutation

through a cyclic shift of the current stride and halve the stride for the next iteration. This gradual reduction in stride results in more stable marginal contributions over time, as the size of the subsets for each player changes less dramatically on average. The process continues until the stride falls to zero.

Algorithm 1: Cyclic Shapley Value Approximation

INPUT: k context players, a pooled text $\boldsymbol{x_t}^j$, an initial stride $s = \text{batch\_size}/2$
OUTPUT: Approximated Shapley value of k players $\{\hat{\phi}_1(\boldsymbol{u}_j), \ldots, \hat{\phi}_k(\boldsymbol{u}_j)\}$
$P \leftarrow$ Random permutation of the players;
while $s > 0$ **do**
  for $i \in \{1, \ldots, k\}$ **do**
    Compute the marginal contribution $m_{P[i]}$ when adding player $P[i]$ to the preceding coalition $P[1 \sim i-1]$;
    Update $\hat{\phi}_{P[i]}(\boldsymbol{u}_j)$ with $m_{P[i]}$;
  end
  Apply a cyclic shift to P by s steps;
  $s = s/2$;
end

B Detailed Statistics of Datasets

The detailed statistics of the four MIE datasets are shown in Table 4.

C Implementation Details

C.1 Model Configuration

In order to fairly compare with the previous approaches(Devlin et al., 2018; Wu et al., 2020; Zhang et al., 2021a; Zheng et al., 2021a; Xu et al., 2022; Wang et al., 2022b,c; Jia et al., 2023; Yuan et al., 2023), we use BERT-based model with the dimension of 768 as the textual encoder. For visual encoder, we experiment with the ViTB/32 in CLIP[10] and ResNet152 models as potential alternatives. We find ResNet152 with the dimension of 2048 to provide the most effective and consistent results in our settings. To generate the texual descriptive contexts of images, we use LLaVA-1.5 (Liu et al., 2023a), a newly proposed visual instruction-tuned, large multimodal model[11]. The prompt we use is shown in Table 5 and the sampling temperature is set to 1.0. We apply the spaCy[12] library of the en_core_web_sm version to parse the given sentence and obtain each word's part of speech. The dimensions of positional embeddings and part-of-speech embeddings are 100, while other features' hidden dimensions are all set to 512.

Dataset	Train	Val	Test
Twitter-15	4,000	1,000	3,257
Twitter-17	3,373	723	723
MNRE	12,247	1,624	1,614
MJERE	3,617	495	474

Table 4: Size of the datasets in numbers of tweets.

C.2 Training Configuration

Shap-CA is trained by Pytorch on single NVIDIA RTX 3090 GPU. During training, the model is fine-tuned by AdamW (Loshchilov and Hutter, 2017) optimizer with a warmup cosine scheduler of ratio 0.05 and a batch size of 28. We use the grid search to find the learning rate over $[1 \times 10^{-5}, 5 \times 10^{-4}]$. The learning rate of encoders is 5×10^{-5}. And the learning rate of other modules is set to 3×10^{-4}, 1×10^{-4}, 1×10^{-4} and 1×10^{-4} for Twitter-15, Twitter-17, MJERE and MNRE, respectively. The refinement scale λ in Eqn. 17 is set to 1.0. An early stopping strategy is applied to determine the number of training epochs with a patience of 7. We choose the model performing the best on the validation set and evaluate it on the test set. We report the average results from 3 runs with different random seeds.

D Large Language Model Experiments

For large language models, the MNER and MRE results are from Chen and Feng (2023), and the results on the MJERE dataset are implemented by us via in-context learning with 10 in-context examples. These large language models are provided with the textual context and the text as input. As shown in Table 6, our prompt consists of an instruction, followed by 10 in-context examples, and finally a question. Each in-context example consists of the textual context of the image, the text and the corresponding entity-relation answer. The question has the same format as in-context examples except there is no answer.

[10] https://github.com/openai/CLIP
[11] https://github.com/haotian-liu/LLaVA
[12] https://github.com/explosion/spaCy

Please generate a clear and concise textual description for this image sourced from a Twitter post. Describe this image in English as detailed as possible and avoid repetition in your explanation.

Table 5: Prompts employed for LLaVA-1.5.

Please perform Named Entity Recognition and Relation Extraction in social media content. Given a tweet and accompanying textual descriptions of its related image, extract entities from the text and identify their associated relations. Entity types include [Entity category], while relation types encompass [Relation category].

Image description: The tweet is about a basketball game. There is a basketball player, holding up a trophy and yelling in celebration.
Text: Stephen Curry and Michael Jordan are both players who have/had 3 NBA Championships by age 30.
Answer:
Stephen Curry (per), Michael Jordan (per), peer.
Stephen Curry (per), NBA Championships (misc), awarded.
Michael Jordan (per), NBA Championships (misc), awarded.

...
<More In-Context Examples>
...

Image description: A black and white photo of a man and a woman. They are standing close to each other, with the man having a tie on. The man is looking at the camera while holding the woman, which gives an air of intimacy and connection between the two.
Text: Elizabeth Taylor and Montgomery Clift in A Place In the Sun, 1951 (dir. George Stevens).
Answer:

Table 6: Prompts employed for ChatGPT/GPT4 on the MJERE dataset.

E Case Study

Cases mentioned in Section 5.5 are shown in Figure 4.

Text: Stephen Curry and Michael Jordan are both players who have/had 3 NBA Championships by age 30.

Context: The tweet is about a basketball game. There is a basketball player, holding up a **trophy** and yelling in **celebration**.

Prediction:

Gold: Stephen Curry$_{per}$, NBA Championships$_{misc}$, *awarded*
EEGA: Stephen Curry$_{per}$, NBA$_{misc}$, *present_in* ✗
Ours: Stephen Curry$_{per}$, NBA Championships$_{misc}$, *awarded* ✓

Text: Elizabeth Taylor and Montgomery Clift in A Place In the Sun, 1951 (dir. George Stevens).

Context: A black and white photo of a man and a woman. They are standing close to each other, with the man having a tie on. The man is looking at the camera while holding the woman, which gives an air of **intimacy and connection** between the two.

Prediction:

Gold: Elizabeth Taylor$_{per}$, George Stevens$_{per}$, *couple*
EEGA: Elizabeth Taylor$_{per}$, George Stevens$_{per}$, *peer* ✗
Ours: Elizabeth Taylor$_{per}$, George Stevens$_{per}$, *couple* ✓

Figure 4: Two cases of the predictions by EEGA and Shap-CA (ours).