Rethinking Multimodal Entity and Relation Extraction from a Translation Point of View

Changmeng Zheng[1], **Junhao Feng**[3], **Yi Cai**[3], **Xiao-Yong Wei**[2,1]*, **Qing Li**[1]
[1]Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR, China
[2]Department of Computer Science, Sichuan University, China
[3]School of Software Engineering, South China University of Technology, China
csczheng@comp.polyu.edu.hk, cswei@scu.edu.cn

Abstract

We revisit the multimodal entity and relation extraction from a translation point of view. Special attention is paid on the misalignment issue in text-image datasets which may mislead the learning. We are motivated by the fact that the cross-modal misalignment is a similar problem of cross-lingual divergence issue in machine translation. The problem can then be transformed and existing solutions can be borrowed by treating a text and its paired image as the translation to each other. We implement a multimodal back-translation using diffusion-based generative models for pseudo-paralleled pairs and a divergence estimator by constructing a high-resource corpora as a bridge for low-resource learners. Fine-grained confidence scores are generated to indicate both types and degrees of alignments with which better representations are obtained. The method has been validated in the experiments by outperforming 14 state-of-the-art methods in both entity and relation extraction tasks. The source code is available at https://github.com/thecharm/TMR.

1 Introduction

Multimodal language understanding has received intensive attention recently for its advantage of mining semantics by collaborating the cross-modal inference (Yang et al., 2019a). Examples include methods for multimodal name entity recognition (MNER) (Zhang et al., 2018) and multimodal relation extraction (MRE) (Zheng et al., 2021a). Both benefit from the collaborative reasoning based on the alignment of textual and visual content. However, statistics on commonly adopted text-image relation benchmarks (e.g., TRC (Vempala and Preoţiuc-Pietro, 2019) and Twitter100k (Hu et al., 2017)) shows that the misalignment rate between images and texts is as high as 60%. Noise introduced by the misalignment can mislead the learning and degrade the performance of resulting models.

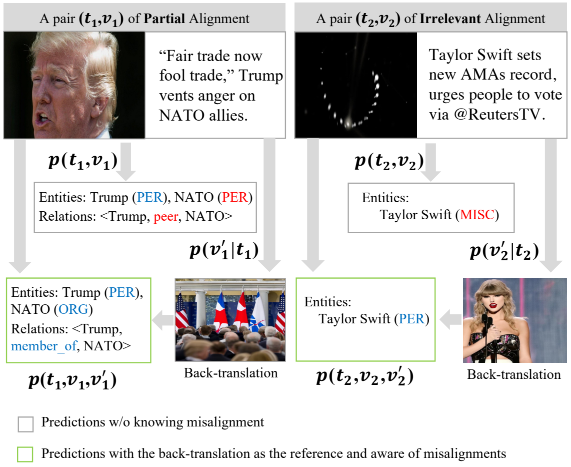

A pair (t_1,v_1) of **Partial** Alignment
A pair (t_2,v_2) of **Irrelevant** Alignment
"Fair trade now fool trade," Trump vents anger on NATO allies.
Taylor Swift sets new AMAs record, urges people to vote via @ReutersTV.
$p(t_1,v_1)$
$p(t_2,v_2)$
Entities: Trump (PER), NATO (PER)
Relations: <Trump, peer, NATO>
Entities: Taylor Swift (MISC)
$p(v'_1|t_1)$
$p(v'_2|t_2)$
Entities: Trump (PER), NATO (ORG)
Relations: <Trump, member_of, NATO>
Back-translation
Entities: Taylor Swift (PER)
Back-translation
$p(t_1,v_1,v'_1)$
$p(t_2,v_2,v'_2)$
□ Predictions w/o knowing misalignment
🟩 Predictions with the back-translation as the reference and aware of misalignments

Figure 1: Partial (left) and irrelevant (right) alignments in text-image pairs and the results of using generative back-translation to help the inference in multimodal entity and relation extraction tasks.

As shown in Fig. 1, the misalignment can be categorized into *partial* and *irrelevant* alignment. In case of incomplete alignment, textual entities (e.g., NATO) might be mismatched to the visual evidence (e.g., person) which results in incorrect labels (e.g., PER). This further leads to underline relations between entities (e.g., <Trump, president of, USA), <USA, member of, NATO>) missing from the extractions. In case of irrelevant alignment, the textual entities might be randomly matched to visual evidence (e.g., MISC) resulting in dirty data for inference. While the misalignment with the ambiguity/distraction it brings to the learning has long been noticed, it has been rarely studied and addressed (Sun et al., 2021). The challenge is that it is nearly impossible to know the degree of misalignment prior to the inference. Otherwise, the inference may has already been done.

In this paper, we conduct a pilot study to address this problem. The motivation is that the misalignment of cross-modal pairs is a similar problem to the divergence of cross-lingual machine translations (Carpuat et al., 2017). The problem can thus be transformed by treating the text-image pairs in

[*] Corresponding author

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6810–6824
July 9-14, 2023 ©2023 Association for Computational Linguistics

MNER/MRE as translations to each other. The divergence problem is more widely studied and existing solutions such as back-translation (Edunov et al., 2018) can be borrowed.

While this sounds appealing, it introduces new challenges as follows.

Modality Gap: The cross-lingual divergence is defined in a monomodal setting. The divergence can be measured explicitly by using features such as difference of sentence lengths, ratio of aligned words, and number of unaligned contiguous sequences (Carpuat et al., 2017). However, those features are not available in a cross-modal setting. We address it in an implicitly way in which disalignment of cross-lingual words (e.g., textual words and visual patches) is indicated by the divergence of their representations in the embedded space.

Parallelism: The detection/assessment of cross-lingual divergence relies on large-scale parallel corpora, in which the sentences are aligned into word-level. The alignment is symmetric which makes high quality back-translation possible. However, in the cross-modal setting, MNER/MRE benchmark datasets are with a small scale due to the high cost of name entities labeling. The datasets are not well paralleled and there is no word-level alignment. We address those problems by taking advantage of the latest development of diffusion-based generative models (Saharia et al., 2022). Those models are trained on large-scale and better paralleled datasets, with which the back-translation can be conducted in a generate-to-translate way, in a sense that, for each text sentence, we can generate an image as its visual language "translation". Visual grounding (Yang et al., 2019b) can then be employed to make the alignment into word-level. More details will be given in Section 3.3.

Low-Resource Benchmarks: The assessment of the divergence needs datasets on large-scale. This is not the case in MNER/MRE scenario. We borrow the idea of using high-resource corpora as a bridge to address the low-resource learning issue (Haddow et al., 2022; Gu et al., 2020). In this papaer, a new multimodal dataset is constructed for multimodal divergence estimation. An estimator is built which generates fine-grained confidence scores over 3 alignment categories of *strengthen, weaken, and complement*. It enables better argumentation for MNER/MRE than the simple similarity-based filtering schemes adopted previously. It also preserves the text-image pairs that are not well-aligned but with complementary evidence. More details will be given in Section 3.4.

2 Related Work

2.1 Multimodal Entity and Relation Extraction

As the core components of knowledge graph construction, named entity recognition (NER) and relation extraction (RE) have received much more attention in the past few years. Previous studies (Zhang et al., 2018; Zheng et al., 2021b) revealed that incorporating visual information into text-based methods (Lample et al., 2016; Soares et al., 2019) can help improve the NER and RE performance, especially when sentences are short and ambiguous. These methods can be roughly divided into three categories: (1) encoding the features of the whole image and design effective attention mechanisms to capture the visual information related to texts (Lu et al., 2018; Moon et al., 2018). (2) incorporating object or region level visual features segmented from input image into textual-based methods with graph structure or transformers (Wu et al., 2020; Zheng et al., 2020; Zhang et al., 2021a; Zheng et al., 2021a). (3) hybrid fusion of multi-level visual features with textual expressions (Chen et al., 2022b,a). Despite the consistent improvement achieved by these attention-based methods, one major issue is that the texts and images are not always aligned well with each other. Recently, Sun et al. (2021) proposed RpBERT to address the above issue by learning a text-image similarity score to filter out the irrelevant visual representations. Zhao et al. (2022) explored inter-modal and intra-modal image-text relations by utilizing external matching from the dataset. However, some pairs not well aligned but with complementary will be neglected.

2.2 Vision-Language Pretraining Models

Large-scale pretrained models (PTMs) such as BERT (Kenton and Toutanova, 2019) and ViT (Dosovitskiy et al., 2020) have shown their strong abilities in representative learning and become a milestone in machine learning. Due to the success of PTMs in computer vision and natural language processing, many works are trying to adopt PTMs in multimodal domain (Han et al., 2021). Indeed, multimodal PTMs (Zhang et al., 2021b; Kim et al., 2021; Radford et al., 2021) can learn universal cross-modal representations and signifi-

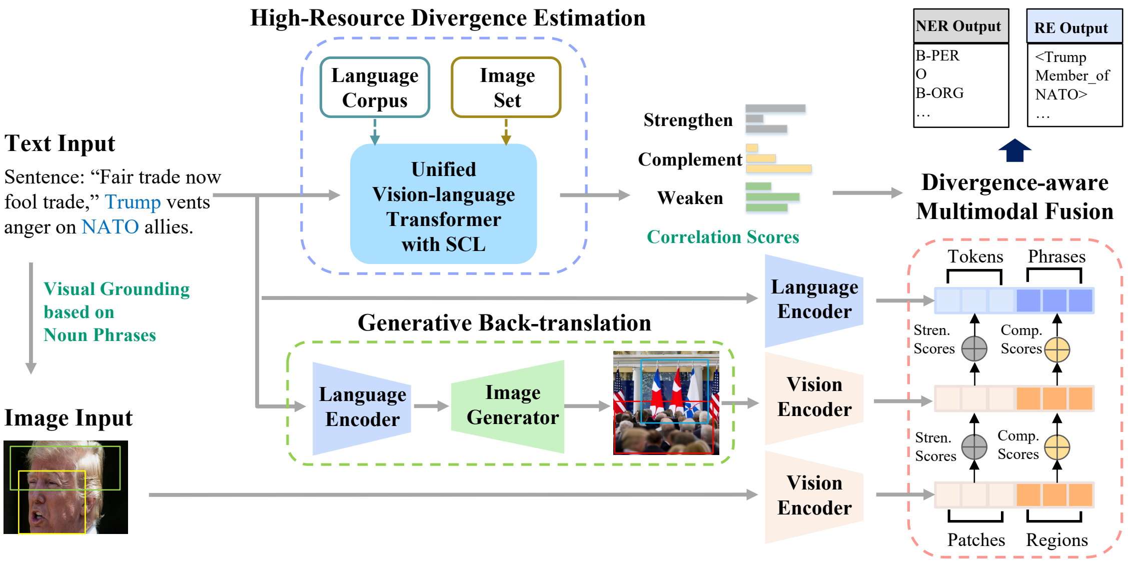


Figure 2: The framework of the proposed **T**ranslation motivated **M**ultimodal **R**epresentation learning (TMR), which generates divergence-aware cross-modal representations by introducing two additional streams of Generative Back-translation and High-Resource Divergence Estimation.

cantly boost the downstream multimodal tasks (Antol et al., 2015; Datta et al., 2008). These methods differ mainly in the architecture for multimodal interactions. However, text-image misalignment has rarely been studied, although it is critical in real-world applications.

2.3 Generative Diffusion Models

Diffusion Models (Ho et al., 2020; Song et al., 2020) have emerged as a powerful toolkit in the family of generative models, with record-breaking results on many downstream applications such as image synthesis (Dhariwal and Nichol, 2021), video generation (Ho et al., 2022), and molecular generation (Jin et al., 2018). Recently, Imagen (Saharia et al., 2022) proposed a text-to-image diffusion model and a comprehensive benchmark for performance evaluation. Rombach et al. (2022) presented stable diffusion, a relatively lightweight text-to-image generator trained with large-scale web-crawled data and can synthesis photo-realistic images in few seconds. In this paper, we propose to transfer knowledge in generative diffusion models as back-translation for multimodal NER and RE tasks.

3 Translation Motivated Multimodal Representation Learning

3.1 Problem Formulation

Give a pair of a sentence t and an image v, our interest is the joint probability $p(t, v)$, on the basis which the "translation" using either modality as the source "language" can be obtained/evaluated (e.g., using $p(t \mid v)$ or $p(v \mid t)$) (Carpuat et al., 2017). However, in the multimodal information extraction scenario, the translation is not a goal. We use it as a conceptual solution-seeking mindset. Specifically, our target is to build a function $g(t, v)$ which learns the representations of $p(t, v)$. We propose to make the learner aware of the modality misalignment (divergence) using

- Back-Translation: a generative diffusion model is employed as a predictor for $p(v' \mid t)$ which generates the back-translation of v. The divergence can be embedded by integrating the representations of v and v';
- High-Resource Divergence Estimation: we learn a function $d(t, v)$ to estimate the cross-modal divergence. The function is learned on a high-resource corpora independently and can be used to adjust $p(t, v)$.

In this section, we introduce a general process for learning the representation first (i.e., $g(t, v)$), and then $p(v' \mid t)$ and $d(t, v)$ can be implemented. Once the representation is obtained, multimodal information extraction tasks such as NER and MNRE can be conducted by learning the probability of $p(l \mid g(t, v))$ where l represents the label of name entities or relations depending on the task. The framework is shown in Fig. 2.

3.2 Multi-Grained Representation Learning

To ease the description, let us denote the resulting representation of a text-image pair as $\mathcal{G} = g(t, v)$. It can be implemented using a Transformer model (Kenton and Toutanova, 2019) as long as t and v can be tokenized (e.g., into words or patches) and embedded, so that the joint representation is learned regarding the cross-model correlation (ensured by the multi-head attention). Denote $\boldsymbol{T}$ and $\boldsymbol{V}$ as the tokenized embedding of t and v, respectively, the representation can be learned as

$$\mathcal{G} = \sum softmax\left(\frac{\boldsymbol{W}_d\boldsymbol{V}\boldsymbol{T}^\top}{\sqrt{d}}\right)\boldsymbol{T}, \qquad (1)$$

where d is the dimension of textual embedding $\boldsymbol{T}$ and $\boldsymbol{W}_d$ is a cross-model attention matrix which is learned during the training.

However, granularity is a concern when the representation is cross-modal, because of the aforementioned Modality Gap and Parallelism challenges. We propose to build a multi-grained representation learning scheme, in which a 2-level of granularity is adopted so that a text is tokenized into words and phrases and an image is tokenized into patches and regions. We assume that the cross-modal representation can be generated on a fine scale based on word-patch correlations and the representation is coarse-grained when built on phrase-region correlations (Li et al., 2022).

Let us denote $\boldsymbol{T}^w$ and $\boldsymbol{T}^p$ as the tokenized embedding of the text t at word and phrase level, respectively, in which the phrases is obtained using Stanford Parser following the method in Zhang et al. (2021a). The embedding are encoded using BERT (Kenton and Toutanova, 2019). Similarly, we denote $\boldsymbol{V}^s$ and $\boldsymbol{V}^r$ as the tokenized embedding of the image v at patch and region level, respectively, in which patches are obtained using fixed grid and regions are obtained using the visual grounding method toolkit (Yang et al., 2019b). We set the numbers of patches and regions as 49 and 3, respectively, by following the previous studies (Chen et al., 2022b,a). ResNet50 (He et al., 2016) is then employed to generate the visual embedding. The 2 levels of pairs $(\boldsymbol{T}^w, \boldsymbol{V}^s)$ and $(\boldsymbol{T}^p, \boldsymbol{V}^r)$ are then be substituted into Eq. (1), resulting in the cross-modal representations $\mathcal{G}^f$ and $\mathcal{G}^c$ at fine and coarse level, respectively. A multi-grained representation $\mathcal{G}$ can then be generated as

$$\mathcal{G} = \mathcal{G}^f + \mathcal{G}^c. \qquad (2)$$

3.3 Cross-Modal Back-Translation

We borrow the idea of back-translation from traditional machine translation methods (Edunov et al., 2018), in which the result in the target language is translated back to the source language to verify the quality or divergence. In our case, we treat the text t as a translation from an image v. A back-translation v' can then obtained by using

$$v' = \arg\max\ p(\hat{v} \mid t), \qquad (3)$$

where $\hat{v}$ is an image hypothesis. However, back-translation usually requires parallel corpora to learn the probability of $p(\hat{v} \mid t)$, which is not available in any NER/MNRE settings. We address this problem by taking advantage of recent advance in diffusion-based generative models (Saharia et al., 2022). Those models are trained using large-scale paralleled text-image pairs to learn the ability to generate an image contained on a give text prompt. The objective of those models is thus conceptually similar to Eq. (3). In our case, we use stable diffusion (Rombach et al., 2022), which is trained on a subset of LAION-5B (Schuhmann et al.) dataset. Upon back-translation, we feed the text t as a prompt to stable diffusion. The modal generates a v' which can be used as an approximation of the back-translation from t.

To assess the divergence of translation, we cannot compare v' to v like in text translation, because the cross-modal misalignment is at the semantic level and indicated by the correlation rather than the content. We thus compose a new pair (t, v') and use the process introduced in Section3.2 to generate a back-translated cross-modal representation $\mathcal{G}'$. Since v' is generated directly from t, the alignment between them is better guaranteed than those sampled from user generated content on web or social media. It can be used a pseudo-paralleled pair. Therefore, the original pair (t, v) is better aligned if $\mathcal{G}$ is similar to $\mathcal{G}'$ or otherwise less aligned. There are different ways to use these two representations complementarily. Examples will be given in Section 3.5 under MNER/MRE scenario.

3.4 High-Resource Divergence Estimation

In this subsection, we implement an independent divergence estimator $d(t, v)$. Existing methods address the issue by setting an attention mask on the reasoner trained on low-resource NER/MNRE benchmarks which simply filters out the less attended pairs (Zhang et al., 2018; Wu et al., 2020).

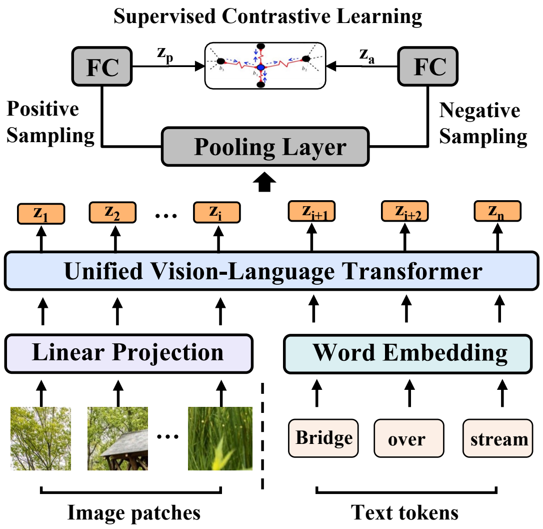


Figure 3: Architecture of our Multimodal Divergence Estimator (MDE), which is trained on high-resource vision-language datasets, and Supervised Contrastive Learning (SCL) is applied to enhance the generalization.

We argue that the training is easy to be biased by replying low-resource benchmarks which are neither sufficient on scale nor designed for divergence assessment purpose. More importantly, the filtering scheme also ignores pairs that are less aligned but with complementary evidence (e.g., Fig. 1). We construct a high-resource corpora which serves as a bridge to train the estimator independently. Furthermore, the estimator generates for each pair 3 confidence scores ($\alpha_s, \alpha_c, \alpha_w$) over the category set {*strengthen, complement, weaken*} for a more detailed divergence estimation. It can then be utilized as an augmenter (instead of a filter) for better representations of $\mathcal{G}$ and $\mathcal{G}'$ as

$$\begin{bmatrix} \mathcal{G}^* \\ \mathcal{G}'^* \\ 0 \end{bmatrix}^\top = \begin{bmatrix} \alpha_s \\ \alpha_c \\ \alpha_w \end{bmatrix}^\top \begin{bmatrix} \mathcal{G}^f & \mathcal{G}'^f \\ \mathcal{G}^c & \mathcal{G}'^c \\ 0 & 0 \end{bmatrix}, \tag{4}$$
$$w.r.t. \quad \alpha_s + \alpha_c + \alpha_w = 1.$$

High-Resource Corpora Construction Different from Sun et al. (2021) using limited data crawled from social media (e.g., Twitter), we collect data from large-scale public image-text datasets to enhance the generalization of our estimator. We randomly select 100k data from MSCOCO (Lin et al., 2014) as the "Strengthen" samples, since the dataset contains fine-grained aligned image-text pairs designed for tasks like Visual grounding and Scene graph generation. LAION-400M (Schuhmann et al., 2021) is chosen as the "Complement" dataset since it is built on web paired data and no strict rules are applied for the alignment between image contents and text tokens. Similar to MSCOCO, we select 100k image-text pairs from LAION-400M as training samples. We generate negative samples as the "Weaken" (unaligned) data by substituting the images in the "Strengthen" and "Complement" data with a different image randomly sampled from the two datasets. Finally, we accumulate 400k training samples, with 100k, 100k, 200k for "Strengthen", "Complement" and "Weaken", respectively. To verify the effectiveness and generalization, we further construct a in-domain test set of 10k data sampled from the two datasets and a out-of-domain test set of 1k data from the SBU dataset which contains both fine-grained and coarse-grained aligned text-image pairs. More supportive evidences and the generalization experiments are provided in Appendix B.3.

Model Design We adopt the same structure as ViLT (Kim et al., 2021) that leverages a unified transformer to encode visual and textual contents. To be more specific, the input image v (or its back-translation v') is sliced into patches and flattened. Then a linear projection is applied to transfer the visual features to the same dimensions of token embeddings. The text and image embeddings are concatenated into a sequence Z and iteratively updated through D-dimensional Transformers. We get the pooled representations of the multimodal input sequence M as final output z. Details can be found in Figure 3 and Section 4.4.3.

Supervised Contrastive Learning Conventional supervised methods use Cross-entropy Loss to distinguish samples with different classes. However, since our pretraining data are constructed on different datasets, simply applying cross-entropy loss will lead the model to learn a short-cut by utilizing the domain difference other than the semantic alignment. This results in poor generalization performance. To tackle this problem, we propose to use the supervised contrastive learning (Khosla et al., 2020) instead to push away the distance between anchors and negative samples generated from the positive classes "Strengthen" and "Complement".

A self-supervised learning loss can be written

$$L_{self} = -\sum_{i \in I} log \frac{exp(z_i \cdot z_{j(i)}/\tau)}{\sum\limits_{a \in A(i)} exp(z_i \cdot z_a/\tau)} \tag{5}$$

where z is the output of our estimator model, τ is a scalar temperature parameter. i, j, a denote the anchor point, positive and negative samples, respectively. We can simply generalize the Eq. (5)to incorporate supervision as:

$$L_{sup} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} log \frac{exp(z_i \cdot z_p/\tau)}{\sum\limits_{a \in A(i)} exp(z_i \cdot z_a/\tau)} \tag{6}$$

where $P(i)$ is the set of indices of positives and $|P(i)|$ denotes its cardinality.

3.5 Multimodal Information Extraction

We use the augmented representations $\boldsymbol{\mathcal{G}}^*$ and $\boldsymbol{\mathcal{G}}'^*$ for two tasks of NER and MNRE.

Named Entity Recognition Following (Chen et al., 2022b; YU et al.), we adopt the CRF decoder to perform the NER task. We fuse the $\boldsymbol{\mathcal{G}}^*$ with its back-translation $\boldsymbol{\mathcal{G}}'^*$ using using multi-head extension (Kenton and Toutanova, 2019) and denoted the final representation for a pair (t, v) as

$$\bar{\boldsymbol{\mathcal{G}}} = Multihead(\boldsymbol{\mathcal{G}}^*, \boldsymbol{\mathcal{G}}'^*) \in \mathbb{R}^{n \times d} \tag{7}$$

which consists of the representation of n words from the text t. NER is then a task to predict probabilities of those words over a set of predefined entity labels (e.g., PER, ORG). Let us denote this label set as $\mathcal{L} = \{l\}$. The probabilities are then denoted as $Y = [y] \in \mathbb{R}^{n \times |\mathcal{L}|}$ and calculated as

$$p(y \mid \bar{\boldsymbol{\mathcal{G}}}) = \frac{\prod_{i=1}^{n} F_i(y_{i-1}, y_i, \bar{\boldsymbol{\mathcal{G}}})}{\sum_{l_j \in \mathcal{L}} \prod_{i=1}^{n} F_i(y_{i-1,j}, y_{i,j}, \bar{\boldsymbol{\mathcal{G}}})}, \tag{8}$$

where $y_{i,j}$ denotes the probability of the i^{th} word over the j^{th} label, and F represents potential functions in CRF. We use the maximum conditional likelihood estimation as the loss function

$$L_{ner} = -\sum_{i=1}^{n} log\Big(p(y|\bar{\boldsymbol{\mathcal{G}}})\Big). \tag{9}$$

Relation Extraction We merge the representations of textual entities, fine-grained and coarse-grained image-text pairs, as well as noun phrases to predict final relations. For a given pair of entities (e_i, e_j) corresponding to the i^{th} and j^{th} words from t, we generate its representation as

$$\ddot{\boldsymbol{\mathcal{G}}}_{i,j} = \boldsymbol{T}_i \oplus \boldsymbol{T}_j \oplus \mathbf{p} \oplus \mathbf{h} \tag{10}$$

where $\boldsymbol{T}_i$ and $\boldsymbol{T}_j$ denote the embeddings of the two entities, respectively, $\oplus$ indicates the concatenation operation, $\mathbf{p}$ denote the summed features of noun phrases in the text t, and $\mathbf{h}$ denotes the summed representation of the text-image pair and its back-translation (i.e., $\mathbf{h} = \boldsymbol{\mathcal{G}}^* + \boldsymbol{\mathcal{G}}'^*$). We can then aggregate the likelihoods of this representation over a set of relation labels $\mathcal{R} = \{r\}$ as $p(r \mid \ddot{\boldsymbol{\mathcal{G}}}_{i,j}) = softmax(\ddot{\boldsymbol{\mathcal{G}}}_{i,j})$. Finally, we can calculate the RE loss with cross-entropy loss function

$$L_{re} = -\sum_{i=1}^{n} log\Big(p(r \mid \ddot{\boldsymbol{\mathcal{G}}}_{i,j})\Big). \tag{11}$$

4 Experiment

4.1 Experimental Settings

Datasets and Metrics We adopt three publicly available datasets for evaluating our proposed method on MNER and MRE, including: 1) **Twitter15** (Lu et al., 2018) and **Twitter17** (Zhang et al., 2018) are two datasets for MNER, which include user posts on Twitter during 2014-2015 and 2016-2017, respectively. 2) **MNRE** (Zheng et al., 2021a) is a manually-annotated dataset for MRE task, where the texts and images are crawled from Twitter and a subset of Twitter15 and Twitter17. Statistics and experimental details are provided in Appendix. We use precision, recall and F1 value as the default evaluation metric and compare such results in the following sections.

Baselines We compare our method with two groups of state-of-the-art (SOTA) methods as follows.

Text-based Methods: *CNN-BLSTM-CRF* (Ma and Hovy, 2016), *HBiLSTM-CRF* (Lample et al., 2016), and *BERT-CRF* (Kenton and Toutanova, 2019) are classical sequence-labeling methods which show excellent prediction results on NER in newswire domain. *PCNN* (Zeng et al., 2015) is a distantly-supervised method for relation extraction, leveraging the knowledge from external knowledge base. *MTB* (Soares et al., 2019) is a SOTA method for many text-based RE tasks.

Previous SOTA Multimodal Approaches: *AdapCoAtt* (Zhang et al., 2018) is the pioneer work that extracts named entities with co-attention mechanism. *RpBERT* (Sun et al., 2021) explicitly calculates image-text similarities by learning a classifier on Twitter data. *OCSGA* (Wu et al., 2020), *UMT* (YU et al.), *UMGF* (Zhang et al., 2021a), and *MEGA* (Zheng et al., 2021a) are the NER/RE methods that align fine-grained object features with textual representations with Transformers or Graph Neural Networks. *VisualBERT* (Li et al., 2019) is a vision-language pretraining model that can be

Modality	Methods	*Twitter-2015*			*Twitter-2017*			*MNRE*		
		Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Text	CNN-BLSTM-CRF	66.24	68.09	67.15	80.00	78.76	79.37	-	-	-
	HBiLSTM-CRF	70.32	68.05	69.17	82.69	78.16	80.37	-	-	-
	BERT-CRF	69.22	74.59	71.81	83.32	83.57	83.44	-	-	-
	PCNN	-	-	-	-	-	-	62.85	49.69	55.49
	MTB	-	-	-	-	-	-	64.46	57.81	60.86
Text+Image	AdapCoAtt	69.87	74.59	72.15	85.13	83.20	84.10	-	-	-
	OCSGA	74.71	71.21	72.92	-	-	-	-	-	-
	RpBERT	71.15	74.30	72.69	-	-	-	-	-	-
	UMT	71.67	75.23	73.41	85.28	85.34	85.31	62.93	63.88	63.46
	UMGF	74.49	75.21	74.85	86.54	84.50	85.51	64.38	66.23	65.29
	VisualBERT	68.84	71.39	70.09	84.06	85.39	84.72	57.15	59.48	58.30
	MEGA	70.35	74.58	72.35	84.03	84.75	84.39	64.51	68.44	66.41
	HVPNeT	73.87	**76.82**	75.32	85.84	87.93	86.87	83.64	80.78	81.85
	MKGFormer	-	-	-	86.98	88.01	87.49	82.67	81.25	81.95
	TMR w/o BT.	74.99	75.18	75.08	84.89	88.16	86.49	88.13	84.69	86.37
	TMR w/o MDE.	74.70	76.05	75.37	85.53	87.93	86.72	89.45	86.09	87.73
	TMR (our method)	**75.26**	76.49	**75.87**	**88.12**	**88.38**	**88.25**	**90.48**	**87.66**	**89.05**

Table 1: The Overall Performance of TMR compared to several baselines on three benchmark datasets for MNER and MRE. We show the prediction results of TMR variants (without Back Translation (BP) or Multimodal Divergence Estimation (MDE)) in the bottom rows.

applied for MNER and MRE tasks. *HVPNet* (Chen et al., 2022b) and *MKGFormer* (Chen et al., 2022a), the latest SOTA for MNER and MRE, which develops a hierarchical structure to learn visual prefix from multiple views.

4.2 Comparison to SOTA

The results are shown in Table 1. It is easy to see our method outperforms other SOTA methods on on all datasets.

When compared to models relying on pure textual information, visual features contribute to the performance gain by 5% on MNER and 20% on MRE. Due to the short and ambiguous characteristics of texts in social media, it is difficult to identify entities and their relations in limited context.

Incorporating multi-grained visual and textual information performs better than relying on object or image level information solely. The SOTA method HVPNeT and our MTR gain better results (88.35% and 86.87% in Twitter-2017 dataset) than UMGF (85.51%) and UMT (85.31%) which align image and text in fine-grained object-level.

Our model outperforms HVPNet and MKGFormer which leverage hierarchical visual representations or powerful vision-language pretraining embeddings, in a relatively large margin (from 82% to 89%) on the MRE task. We observe a more obvious performance improvement on MRE datasets compared to that on MNRE. The difference comes from the different distributions of MRE and MNRE datasets. Our statistics show that the proportion of complementary cases is significantly higher in MRE (51.5%) than in MNRE (15.7%). As mentioned in the paper, the proposed back-translation helps the two tasks by providing additional contextual information for inference. This benefits the complementary cases the most because it makes the identification of indirect relationships possible (otherwise, those cases will be considered as misalignments or used incorrectly like in the similarity-based methods).

4.3 Ablation Study

In this section, we conduct extensive experiments with the variants of our model to analyze the effectiveness of each component.

Back-translation: We ablate the procedure of generating back-translation images and the results in Table 1 show the component can boost model performance by 1-3% in MNER and MRE. Still, our ablated model gains comparable or superior performance against baselines which demonstrates the effectiveness of back-translation.

Multimodal Divergence Estimation: Compared with similarity-score based method RpBERT, our model shows stronger extraction and generalization performance with 3.18% improvement on Twitter-2015 dataset. Also, our model achieves significant improvements (3% to 7%) over attention-based methods, revealing that TMR can improve conventional NER/RE methods by decomposing the divergence into fine-grained level.

4.4 Other Essentials of the Model

4.4.1 Low-resource Performance

We conduct experiments in low-resource scenarios following the setting of Chen et al. (2022b), by randomly sampling 5% to 50% from original training set. From the results in Figure 4, we can observe: 1) The methods utilizing multi-grained features (HVPNet and TMR) consistently outperform object-level models in MNER (UMGF) and MRE (MEGA). Multi-grained features can provide global and local views and help models infer entities and relations efficiently. 2) Moreover, our proposed TMR model performs better than HVPNet with external knowledge from generative diffusion models, which addresses the information lack problem in low-resource scenarios.

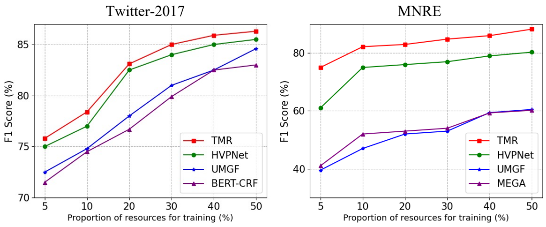


Figure 4: Performances in low-resource setting on MNER and MRE tasks.

4.4.2 Improvements on Complementary Cases

To demonstrate the effectiveness of correlation decomposition, we further compare our method with SOTA method HVPNeT on complementary cases of MNRE test set. We argue that previous similarity-based methods ignore the cross-modal divergence, especially when texts and images are complementary. We export 832 cases with "complement score" higher than 0.5 from 1614 test samples. Our model achieves significant improvements against HVPNeT, especially on some categories (e.g., Present in, Locate at and Residence) that rely on deeper understanding of visual scenarios.

Table 2: Our results on complementary cases compared to HVPNeT (Chen et al., 2022b) on the MNRE test set. Six main categories are selected for comparison.

Category	Count	TMR	HVPNeT
Peer	98	**91.00**	89.30
Member_of	46	**97.87**	82.11
Contain	33	**98.46**	95.65
Present_in	44	**91.95**	79.01
Locate_at	18	**97.14**	75.68
Residence	13	**83.87**	66.67
Overall	832	**87.37**	77.93

4.4.3 Generalization Performance of Multimodal Divergence Estimator

We extend conventional similarity score into fine-grained level and weight the importance of incorporated visual information based on the pretrained divergence estimator. To verify the generalizations to data in other domain, we first construct test set collected with in-domain data (i.e., by sampling on MSCOCO and LAION400M). Then, We first request 2 annotators to label 1k test samples on out-of-domain data and then ask other 2 to review and rectify the test set. As shown in Table 3, we compare the estimator trained with different loss function. The results indicate that the model with cross-entropy loss suffers the generalization problem when transferred into out of domain data. The possible reason is that the model may learn a shortcut from the difference of image/text style on the data from the two datasets, other than taking the image-text correlation into consideration. We improve it by introducing negative sampling on in-domain data to reduce the style bias and the F1 value on out-of-domain data increases from 61.8 to 80.01. We further apply the supervised contrastive learning to pull together the positive samples and push apart negative ones, resulting in better generalization performance.

Model Setting	In Domain	Out of Domain
Cross-entropy	98.56	61.80
Negative Sampling	92.57	80.01
Supervised Contrastive	93.26	**86.21**

Table 3: The generalization experiment of the Multimodal Divergence Estimator (MDE). Origin. is the dataset with 10k data sampling from pretraining data, while SBU is the 1k dataset for human evaluation. F1 value is used for evaluation metric.

4.4.4 Case Study

To validate the effectiveness and robustness of our method, we conduct case analysis for multimodal divergence estimation. Previous works simply calculate the image-text similarity with attention mechanism (HVPNeT) or pretrained classifier (RpBERT). As a result, visual information with low similarity score will be filtered out. We notice that our model and RpBERT can identify entities correctly when images are well-aligned with sentence in S1. However, RpBERT fails to extract the ORG entity "Foran" since it outputs a much lower similarity score. Our model successfully captures

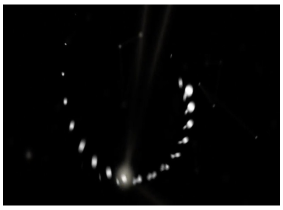
Strengthen	Complement	Weaken
S1: A beautiful Timber Frame bridge over a stream in Auburn (LOC), PA (LOC). Relational Triplet: (Auburn, contain, PA)	S2: Cross country: Foran (ORG)'s Mia Williams (PER), Kevin Preneta take firsts. Relational Triplet: (Mia Williams,member_of, Foran)	S3: Taylor Swift (PER) sets new AMAs (MISC) record, urges people to vote via @ReutersTV.  Relational Triplet: (Taylor Swift, awarded, AMAs)
Similarity Score: 0.76 MDE Score - Strengthen: 0.954 Complement: 0.045 Weaken: 0.001	Similarity Score: 0.24 MDE Score - Strengthen: 0.000 Complement: 0.927 Weaken: 0.072	Similarity Score: 0.14 MDE Score - Strengthen: 0.000 Complement: 0.073 Weaken: 0.926
RpBERT: Auburn (LOC), PA (LOC) Ours: Auburn (LOC), PA (LOC)	RpBERT: Foran (PER), Mia Williams (PER) Ours: Foran (ORG), Mia Williams (PER)	HVPNeT: (/per/misc/present_in) Ours: (/per/misc/awarded)

Figure 5: The first line shows the three correlation categories, and the second row indicates representative samples with their ground-truth entity and relation types. The third line presents the comparison between our decomposed multimodal divergence estimation (MDE) score and conventional similarity score, and the bottom is the prediction results of our model and corresponding baselines.

the semantics of "team competition" and it can be used to complement the missing semantics, which helps extract "Foran" as a name of organization and the relation "member_of" between the two entities. Another case is that when the image is irrelevant to textual contents in S3, HVPNeT gives the wrong prediction due to the misleading of the image. Our method can address this problem by generating a back-translation image of "Taylor Swift" and the "awarding scene", as shown in Figure 1.

5 Conclusion

We have revisited the misalignment issue in multimodal benchmarks. By borrowing the ideas from translation methods, we have implemented multimodal versions of back-translation and high-resource bridging, which provide a multi-view to the misalignment between modalities. The method has been validated in the experiments and outperforms 14 SOTA methods.

Acknowledgments

This work has been supported by the Hong Kong Research Grants Council under the Collaborative Research Fund (Project No. C1031-18G), the National Natural Science Foundation of China (62076100), Fundamental Research Funds for the Central Universities, SCUT (x2rjD2220050), the Science and Technology Planning Project of Guangdong Province (2020B0101100002), CAAI-Huawei MindSpore Open Fund, CCF-Zhipu AI Large Model Fund.

Limitations

The study is in its initial form. The efficiency is a major concern. This mainly results from the use of generative diffusion models, which are under heavy development. We believe this will be addressed soon in the near future. Further, the proposed framework is not end-to-end. It may introduce extra effort for training. We will deal with this issue in the future study.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Marine Carpuat, Yogarshi Vyas, and Xing Niu. 2017. Detecting cross-lingual semantic divergence for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 69–79.

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. *arXiv preprint arXiv:2205.02357*.

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*.

Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (Csur)*, 40(2):1–60.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500.

Jiatao Gu, Yong Wang, Yun Chen, Kyunghyun Cho, and Victor OK Li. 2020. Meta-learning for low-resource neural machine translation. In *2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 3622–3631. Association for Computational Linguistics.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.

Jonathan Ho, Tim Salimans, Alexey A Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022. Video diffusion models. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*.

Yuting Hu, Liang Zheng, Yi Yang, and Yongfeng Huang. 2017. Twitter100k: A real-world dataset for weakly supervised cross-media retrieval. *IEEE Transactions on Multimedia*, 20(4):927–938.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. 2018. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pages 2323–2332. 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.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270.

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. *arXiv preprint arXiv:1908.03557*.

Zejun Li, Zhihao Fan, Huaixiao Tou, Jingjing Chen, Zhongyu Wei, and Xuanjing Huang. 2022. Mvptr: Multi-level semantic alignment for vision-language pre-training via multi-stage learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4395–4405.

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 *European conference on computer vision*, pages 740–755. Springer.

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.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1064–1074.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Christoph Schuhmann, Robert Kaczmarczyk, Aran Komatsuzaki, Aarush Katta, Richard Vencu, Romain Beaumont, Jenia Jitsev, Theo Coombes, and Clayton Mullis. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *NeurIPS Workshop Datacentric AI*, FZJ-2022-00923. Jülich Supercomputing Center.

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.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.

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.

Alakananda Vempala and Daniel Preoţiuc-Pietro. 2019. Categorizing and inferring the relationship between the text and image of twitter posts. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, pages 2830–2840.

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.

Sibei Yang, Guanbin Li, and Yizhou Yu. 2019a. Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4145–4154.

Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. 2019b. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4683–4693.

Jianfei YU, Jing JIANG, Li YANG, and Rui XIA. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer.(2020). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3342–3352.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1753–1762.

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021a. Multimodal 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 *Thirty-Second AAAI Conference on Artificial Intelligence*.

Fei Zhao, Chunhui Li, Zhen Wu, Shangyu Xing, and Xinyu Dai. 2022. Learning from different text-image pairs: A relation-enhanced graph convolutional network for multimodal ner. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3983–3992.

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 Detailed Statistics of Dataset

Table 4: Statistics of the Twitter-2015 Dataset.

Category	Train	Dev	Test	Total
Person	2217	552	1816	4583
Location	2091	522	1697	4308
Organization	928	247	839	2012
Misc	940	225	726	1881
Total Entity	6176	1546	5078	12784

Table 5: Statistics of the Twitter-2017 Dataset.

Category	Train	Dev	Test	Total
Person	2943	626	621	4190
Location	731	173	178	1082
Organization	1674	375	395	2444
Misc	701	150	157	1008
Total Entity	6049	1324	1351	8724

Table 6: The Statistics of MNRE Dataset Compared to SemEval-2010 Task 8 Dataset.

Statistics	SemEval-2010	MNRE
Word	205k	258k
Sentence	10,717	9,201
instance	8,853	15,485
Entity	21,434	30,970
Relation	9	23
Image	-	9,201

B Experimental Details

B.1 Multimodal Named Entity Recognition

This section details the training procedures and hyperparameters for named entity recognition. We use the BERT-base-uncased model from hugging face library . We follow UMGF (Zhang et al., 2021a) to revise some wrong annotations in the Twitter-2015 dataset. We utilize Pytorch to conduct experiments with 1 Nvidia 3090 GPUs. All optimizations are performed with the AdamW optimizer with a linear warmup of learning rate 3e-5 over the first 10% of gradient updates to a maximum value, then linear decay over the remainder of the training. And weight decay on all non-bias parameters is set to 0.01. We set the number of grounding regions and image patches to 3 and 49, respectively. Max length of noun phrases is set to 4 and max length for sentence is set to 80.

B.2 Multimodal Relation Extraction

This section details the training procedures and hyperparameters for relation extraction. Similar to NER, we use the BERT-base-uncased model from hugging face library. We set the number of grounding regions and image patches to 3 and 49, respectively. Max length of noun phrases is set to 6 and max length for sentence is set to 128. We set the initialized learning rate to 1e-2.

B.3 Multimodal Divergence Estimation

We adapt the main structure of ViLT (Kim et al., 2021) to decompose the image-text correlation, as shown in Figure 5. For all experiments, we use AdamW optimizer with base learning rate of 1e-4 and weight decay of 1e-2. The learning rate was warmed up for 10% of the total training steps and was decayed linearly to zero for the rest of the training. We resize the shorter edge of input images to 384 and limit the longer edge to under 640 while preserving the aspect ratio. Patch projection of the model yields 12 × 20 patches for an image with a resolution of 384 × 640. We use the BERT-based-uncased tokenizer to tokenize text inputs. We pretrain the model for 100K steps on 8 NVIDIA V100 GPUs with a batch size of 32.

B.4 More Cases of Generative Back-translation

We provide more examples in Figure 6 to illustrate the power of generative back-translation. Compared to extract entities and their relations with only original images, the generated images provide a different view and help to align the image and text from a translation perspective.

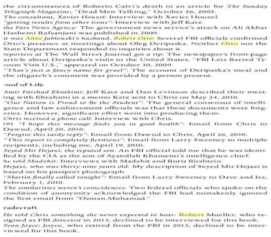
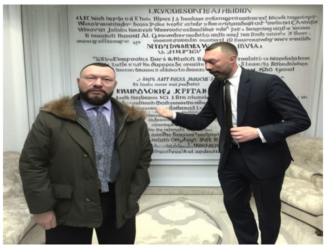
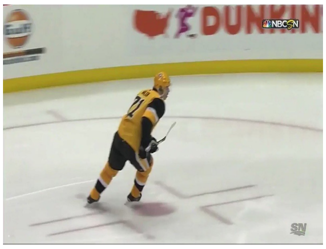
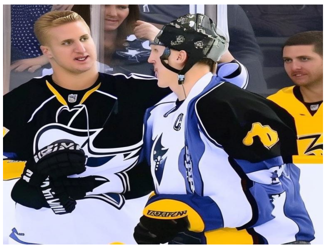
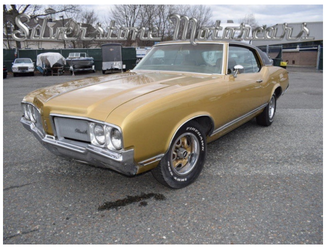
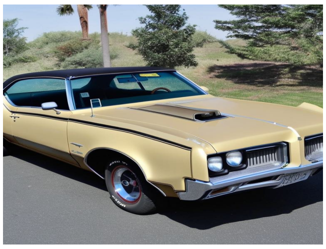
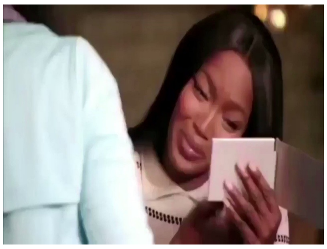
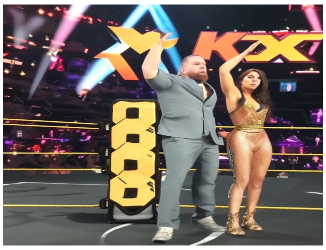
Sentence	Facts	Original Image	Generated Image
Wow! # Bob0tto met with Deripaska as well, wonder why Mueller refused to be interviewed.	per/per/peer	 the circumstances of Roberto Calvi's death in an article for *The Sunday Telegraph Magazine*, "Dead Men Talking," October 26, 2003. *The consultant, Xavier Houzel*: Interview with Xavier Houzel. *"getting results from other irons"*: Interview with Jeff Katz. *the Fars News Agency*: The government news service's attack on Ali Akbar Hashemi Rafsanjani was published in 2009. *it was Anne Jablonski's husband, Robert Otto*: Several FBI officials confirmed Otto's presence at meetings about Oleg Deripaska. Neither Otto nor the State Department responded to inquiries about it. *reporters at* The Wall Street Journal *got a tip*: The newspaper's front-page article about Deripaska's visits to the United States, "FBI Lets Barred Tycoon Visit U.S.," appeared on October 30, 2009. *"That's just a fancy name for gruel"*: The account of Deripaska's meal and the oligarch's comment was provided by a person present. **roof of Life** *Amir Farshad Ebrahimi*: Jeff Katz and Dan Levinson described their meeting with Ebrahimi in a memo Katz sent to Chris on May 24, 2010. *"Our Nation is Proud to Be the Student"*: The general consensus of intelligence and law enforcement officials was that these documents were forgeries. However, significant effort went into producing them. *Chris received a phone call*: Interview with Chris. 191 *"I hope this message finds you in good health"*: Email from Chris to Dawud, April 20, 2010. *"Forgive this tardy reply"*: Email from Dawud to Chris, April 26, 2010. *"This report is completely fictitious"*: Email from Larry Sweeney to multiple recipients, including me, April 19, 2010. *Seyed Mir Hejazi, the reputed son*: An FBI official told me that he was identified by the CIA as the son of Ayatollah Khameini's intelligence chief. *he told Madzhit*: Interviews with Madzhit and Boris Birshtein. *Hejazi, who was thirty-nine years old*: My description of Seyed Mir Hejazi is based on his passport photograph. *"Martin finally called tonight"*: Email from Larry Sweeney to Dave and Ira, February 1, 2010. *The similarities weren't coincidences*: Two federal officials who spoke on the condition of anonymity acknowledged the FBI had mistakenly ignored the first email from "Osman.Muhamad." **radecraft** *He told Chris something she never expected to hear*: Robert Mueller, who resigned as FBI director in 2013, declined to be interviewed for this book. *Sean Joyce*: Joyce, who retired from the FBI in 2013, declined to be interviewed for this book.	
Stamkos and Malkin dropping is what makes hockey so great. Two superstars just chuckin knucks.	per/per/peer		
eBay: Oldsmobile Cutlass 1970 Gold Original Daily Driver Classic Car.	org/misc/other		
They let Mike and Maria skip NXT to do nothing	per/per/couple		

Figure 6: Examples of the back-translation images generated by diffusion models. The images form a different view and help to extract entities and relations precisely.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 6

☑ A2. Did you discuss any potential risks of your work?
Section 7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
No response.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

□ 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)?
No response.

□ 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?
No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

□ 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.
No response.

C ☒ Did you run computational experiments?

Left blank.

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

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?
No response.

□ 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?
No response.

□ 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.)?
No response.

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.