

# MNER-MI: A Multi-image Dataset for Multimodal Named Entity Recognition in Social Media

Shizhou Huang<sup>1</sup>, Bo Xu<sup>2</sup>, Changqun Li<sup>1</sup>, Jiabo Ye<sup>1</sup>, and Xin Lin<sup>1,3,\*</sup>

<sup>1</sup>School of Computer Science and Technology, East China Normal University, Shanghai, China

<sup>2</sup>School of Computer Science and Technology, Donghua University, Shanghai, China

<sup>3</sup>Shanghai Key Laboratory of Multidimensional Information Processing

huangshizhou@ica.stc.sh.cn, xubo@dhu.edu.cn, 52215901009@stu.ecnu.edu.cn,

jiabo.ye@stu.ecnu.edu.cn, xlin@cs.ecnu.edu.cn

## Abstract

Recently, multimodal named entity recognition (MNER) has emerged as a vital research area within named entity recognition. However, current MNER datasets and methods are predominantly based on text and a single accompanying image, leaving a significant research gap in MNER scenarios involving multiple images. To address the critical research gap and enhance the scope of MNER for real-world applications, we propose a novel human-annotated MNER dataset with multiple images called MNER-MI. Additionally, we construct a dataset named MNER-MI-Plus, derived from MNER-MI, to ensure its generality and applicability. Based on these datasets, we establish a comprehensive set of strong and representative baselines and we further propose a simple temporal prompt model with multiple images to address the new challenges in multi-image scenarios. We have conducted extensive experiments to demonstrate that considering multiple images provides a significant improvement over a single image and can offer substantial benefits for MNER. Furthermore, our proposed method achieves state-of-the-art results on both MNER-MI and MNER-MI-Plus, demonstrating its effectiveness. The datasets and source code can be found at <https://github.com/JinFish/MNER-MI>.

**Keywords:** multimodal named entity recognition, multiple images, social media

## 1. Introduction

Recently, multimodal named entity recognition (MNER) has emerged as a vital research area within NER, as it can improve text-based NER by incorporating accompanying images as additional contextual information (Xu et al., 2022b). This fusion of text and images has shown promising potential to enhance the accuracy and scope of entity recognition in various real-world scenarios. Current MNER approaches focus on obtaining better text and image representations (Yu et al., 2020; Wang et al., 2022d), establishing better text-image interaction (Zhang et al., 2021; Chen et al., 2022), and reducing the hindrance caused by image noise (Sun et al., 2021; Xu et al., 2022b).

With the proliferation of user-generated content in social media, posts containing both text content and multiple images are becoming increasingly common, with over 42% of tweets containing more than one image according to (Zhang et al., 2018). However, current MNER datasets (Lu et al., 2018; Zhang et al., 2018; Wang et al., 2022c) and methods are predominantly based on text and a single accompanying image, leaving a significant research gap in real MNER scenarios involving multiple images. This limitation highlights the urgent need for novel datasets and approaches that can better address multi-image scenarios, enabling

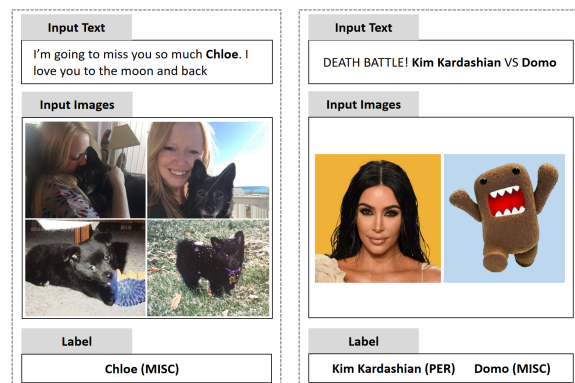


Figure 1: Two examples of multimodal named entity recognition with multiple images. Considering only one image may incorrectly determine the type of entity. The first example is shown on the left and the second on the right.

more accurate and robust MNER across diverse social media content.

Moreover, the current works underestimate the importance of multiple images and overlook the necessity of considering multiple images to understand multi-image posts in real-world applications fully. For instance, during the annotation process of WikiDiverse (Wang et al., 2022c), only the first image is retained in instances containing multiple images, disregarding the valuable information pro-

\* Corresponding author.

vided by the other images. To illustrate the advantages of considering multiple images for understanding multi-image posts, we present the following examples: (1) Posts with multiple images can help alleviate the ambiguity present in posts with only one image. As shown in the first example in Figure 1, it is challenging to determine the type of `Chloe` with only the text and the first image, as it is unclear whether `Chloe` corresponds to a person or a dog in the first image. However, with the assistance of multiple images, the third and fourth images reveal that its type is MISC, resolving the ambiguity. (2) Posts with multiple images provide abundant information that can be utilized to identify more entities in the text. As illustrated in the second example shown in Figure 1, with only the text and the first image, we can only determine that the type of `Kim Kardashian` is Person, and the type of `Domo` remains undetermined. Nevertheless, by incorporating multiple images, we obtain additional evidence to classify the type of `Domo` is Miscellaneous.

To bridge the critical research gap in multiple images at MNER and enhance the scope of MNER for real-world applications, we introduce a novel human-annotated dataset, named **MNER-MI** (Multimodal Named Entity Recognition with Multiple Images). The MNER-MI dataset comprises 8,576 instances collected from Twitter and each instance contains at least 2 images and at most 4 images (the maximum number of images on Twitter). Both text and image information are considered in the annotation process, and the annotation is only performed for the text. To further enhance the generality and applicability of our dataset, we extend MNER-MI with the TWITTER-2017 dataset (Yu et al., 2020), which consists of instances containing only one image. The extended dataset is called **MNER-MI-Plus**, which provides a collection of both single and multiple image instances.

To comprehensively evaluate the performance of baselines and our methods on MNER-MI and MNER-MI-Plus, we establish a diverse and representative set of baselines, including text-based NER methods, MNER methods, and large language models. In addition, we compare the performance of the single-image MNER methods on our proposed datasets with that of the single-image MNER datasets. We find that the current single-image MNER methods perform well in single-image scenarios but poorly in multi-image scenarios, which highlights the fact that the existing single-image MNER methods are not directly applicable to multi-image scenarios, as well as the complexity and challenges in our proposed dataset.

Although utilizing multiple images provides a wealth of contextual information, it also brings new challenges. One of the main challenges is effi-

ciently representing these multiple images. To address this challenge, we further propose a simple yet powerful model called the Temporal Prompt Model with Multiple Images (TPM-MI). We treat multiple images as frames in a video, which allows us to exploit temporal information to establish relationships between the images and understand the interplay between them. Additionally, we couple the multiple images as prompts with the text, enabling effective interaction between the images and the text. Note that we propose the model for demonstrating a reasonable level of performance, and future works can be improved and explored based on it.

Our main contributions are summarized as follows:

- We introduce a novel and challenging human-annotated dataset, MNER-MI, to bridge the research gap in MNER and enhance the scope of MNER for real-world applications. To ensure the generality and applicability of our dataset, we extend it with the TWITTER-2017, resulting in MNER-MI-Plus. To the best of our knowledge, we are the first to propose the limitations of MNER in multi-image scenarios and introduce a multi-image MNER dataset.
- We establish a comprehensive set of strong and representative baselines on MNER-MI and MNER-MI-Plus. Experimental results demonstrate that utilizing multiple images significantly enhances model performance in multi-image scenarios compared to using a single image alone, which demonstrates the potential of multiple images in facilitating a better understanding of multimodal content. Additionally, we observe that the current single-image MNER methods perform poorly on our datasets, which highlights the challenges and difficulty of our proposed datasets.
- To address the challenges in multiple images, we propose a temporal prompt model with multiple images (TPM-MI), which models multiple images as frames in a video and couples the multiple images as prompts with the text for interaction between the images and the text. Experimental results demonstrate that the proposed method and its variants achieve state-of-the-art results on both MNER-MI and MNER-MI-Plus, which demonstrates the effectiveness of our method.

## 2. Related Work

### 2.1. Multimodal Named Entity Recognition

Multimodal named entity recognition (MNER) introduces additional modalities as extra information on top of text, including visual contents (Moon et al., 2018), acoustic contents (Sui et al., 2021), which can effectively improve the performance of text-based named entity recognition. In this paper, we focus on MNER for images and text. The current MNER methods focus on the representation of text and images, the interaction of text and images, and the reduction of the effect of image noise.

Regarding the representation of text and images, most of the early works (Moon et al., 2018; Lu et al., 2018; Zhang et al., 2018) directly use text encoder and image encoder to obtain respective representations. Subsequently, (Yu et al., 2020; Wang et al., 2022d) use a more advanced text encoder for obtaining a better text representation, and (Wu et al., 2020) propose to use the objects in the image as the image representation. In addition, (Wang et al., 2022b) propose to use the image objects, image caption and text in the image as the image representation.

In terms of the interaction of text and images, current methods are mainly based on the attention mechanism. Specifically, (Zhang et al., 2021) use a graph-based method to achieve the interaction between the image and the text, and (Yu et al., 2020) use the mechanism of attention to establish a bi-directional relationship between text and image. Recently, (Chen et al., 2022), (Wang et al., 2022d) and (Xu et al., 2023) project image representation as the prompts to allow the image representation to interact with each layer of the text encoder.

For the reduction of the effect of image noise, the core idea of the current approaches is to train a text-image matching classifier for determining whether an image can help text for named entity recognition. Specifically, (Sun et al., 2021), (Xu et al., 2022b) and (Xu et al., 2022a) propose the use of exogenous supervised datasets, a self-supervised approach and a reinforcement learning approach to learn the text-image matching classifier, respectively.

In addition, (Wang et al., 2022a) extract knowledge of text and images to help MNER, and (Jia et al., 2022) propose a machine reading comprehension framework to locate regions in images more accurately.

### 2.2. Datasets for Multimodal Named Entity Recognition

To the best of our knowledge, there are currently four public MNER datasets: SNAP (Lu et al., 2018),

TWITTER-2015 (Zhang et al., 2018), TWITTER-2017 (Yu et al., 2020), and WikiDiverse (Wang et al., 2022c). Specifically, TWITTER-2015 and TWITTER-2017 are two widely used datasets in the social media domain. WikiDiverse is a multimodal entity linking dataset constructed on the basis of Wikinews, which contains entity spans and entity labels that can be used for MNER.

However, the current datasets are predominantly based on text and a single accompanying image, leaving a significant research gap in scenarios involving multiple images. To address this issue and fully exploit multi-image posts in social media, we propose the multi-image MNER dataset MNER-MI and MNER-MI-Plus.

## 3. Datasets

In this paper, we introduce two novel MNER datasets, namely MNER-MI and MNER-MI-Plus.

**Dataset Collection.** Since the links to tweets provided by TWITTER-2015 are often mixed with other links in the tweets, and TWITTER-2017 and SNAP do not provide direct links to the original posts, it is difficult for us to extend the existing MNER datasets on social media. We follow (Lu et al., 2018) to collect tweets from Twitter<sup>1</sup>, and different from the two widely used MNER datasets on social media (TWITTER-2015 and TWITTER-2017), we do not pick certain topics and do not only take data from a fixed number of months in one given year. We collect tweets from each month in the years 2019, 2020, 2021, and 2022 to provide a more diverse and unbiased dataset, which also makes it more challenging. Firstly, we filter out non-English tweets, repeated tweets, tweets with a text length of less than 3, and tweets with less than 2 images. Then, we save the original links corresponding to each Tweet data, allowing future works to easily extend our dataset and help with MNER through information other than tweet text and images (e.g., meta-information: author bios, comments, etc.). Finally, we get the 10K+ tweets for annotation, where each tweet contains up to 4 images (the maximum number of images on Twitter).

**Human Annotation.** We employ three graduate students with backgrounds in named entity recognition to annotate the tweets. After ensuring that all annotators understand the annotation requirements, the annotators use an annotation tool called doccano<sup>2</sup> to annotate the tweets. Each annotator can see the text and all the images in the tweet during the annotation process and uses both the text and all the images to identify the entity as well

<sup>1</sup><https://archive.org/details/twitterstream>

<sup>2</sup><https://github.com/doccano/doccano>

as determine the category of the entity. The annotators follow the BIO2 (Sang and Veenstra, 1999) annotation standard to annotate the tweets and encompass the same four types of named entities as the SNAP, TWITTER-2015, and TWITTER-2017: Person, Location, Organization, and Miscellaneous. We aggregate the annotations using majority voting. In addition, if any of the annotators think that the tweet reveals personal information, or that there is sensitive or harmful information, then the tweet is discarded. We adopt Fleiss Kappa (Fleiss, 1971) to measure the annotation agreement, and the Fleiss score between the three annotators is  $\mathcal{K} = 0.87$ , indicating a substantial annotation agreement. Finally, we get 8,576 tweets as our dataset called MNER-MI.

**Extension of MNER-MI.** Considering that MNER-MI only contains multi-image tweets, we propose to extend it with a dataset containing single-image tweets to obtain an MNER dataset containing both single-image tweets and multiple-image tweets, allowing for the evaluation of the performance of a model in both single-image scenarios and multi-image scenarios. There are two widely used public single-image MNER datasets on social media: TWITTER-2015 (Zhang et al., 2018) and TWITTER-2017 (Yu et al., 2020). Since TWITTER-2015 contains more noisy data (e.g., labeling inconsistency), we only chose TWITTER-2017 to combine with MNER-MI to obtain the dataset, named MNER-MI-Plus.

An alternative approach for evaluating model performance in single-image and multi-image scenarios is to train separate models on the single-image MNER dataset and the multi-image MNER dataset. However, this would lead to two different models, one of which is trained for single-image scenarios and the other for multi-image scenarios, and this approach does not allow us to determine the robustness of a model, i.e., we cannot determine whether a single model can handle both single-image and multi-image scenarios. To avoid this problem, we choose to merge a single-image MNER dataset, which allows us to evaluate the model in both single-image scenarios and multi-image scenarios.

**Dataset Analysis.** As shown in Table 1, MNER-MI comprises 8,576 tweets and 11,862 named entities, divided into training, development, and test sets containing 6,856, 860, and 860 tweets, respectively. On average, each tweet in MNER-MI contains around 3 images. MNER-MI-Plus merges the training set, development set, and test set of MNER-MI with the training set, development set, and test set of TWITTER-2017, respectively. It contains a total of 13,395 tweets and 20,586 named entities. The training, development, and test sets of MNER-MI-Plus contain 10,229, 1,583, and 1,583 tweets, respectively, with each tweet having an av-

erage of approximately 2 images due to the integration of single-image tweets. The named entity type statistics are also shown in Table 1.

Type	MNER-MI			MNER-MI-Plus		
	Train	Dev	Test	Train	Dev	Test
Person	4,529	573	439	7,472	1,199	1,060
Location	1,878	210	156	2,609	383	334
Organization	1,273	165	92	2,947	540	487
Miscellaneous	2,054	260	233	2,755	410	390
Total	9,734	1,208	920	15,783	2,532	2,271
# One Image	0	0	0	3,373	723	723
# Two Images	3,711	446	455	3,711	446	455
# Three Images	814	110	135	814	110	135
# Four Images	2,331	304	270	2,331	304	270
# Images per Tweet	2.799	2.835	2.785	2.206	1.997	1.970
# Tweets	6,856	860	860	10,229	1,583	1,583

Table 1: Statistics of MNER-MI and MNER-MI-Plus.

We compare MNER-MI and MNER-MI-Plus with four existing MNER datasets in Table 2. Compared with existing datasets, MNER-MI contains more annotated samples, and more images, and is the first MNER dataset for multi-image scenarios. MNER-MI-Plus further extends MNER-MI for both single-image scenarios and multi-image scenarios, enabling the evaluation of the performance of a model in both single-image scenarios and multi-image scenarios, and becoming the largest MNER dataset that we know of so far. In summary, MNER-MI and MNER-MI-Plus offer valuable resources for the research community in MNER, and these datasets address the limitations of the existing datasets and better align with real-world multimodal content and aim to foster advancements in MNER, especially in the context of contemporary social media platforms with diverse multimedia information.

Dataset	Size	# Images	Scenarios
SNAP	6.8K	6,882	SI
WikiDiverse	7.9K	7,969	SI
TWITTER-2015	8.3K	8,357	SI
TWITTER-2017	4.8K	4,819	SI
<b>MNER-MI</b>	<b>8.5K</b>	<b>24,201</b>	<b>MI</b>
<b>MNER-MI-Plus</b>	<b>13.3K</b>	<b>29,020</b>	<b>SI + MI</b>

Table 2: A comparison with other MNER datasets. SI and MI represent that this dataset is used for single-image scenarios and multi-image scenarios, respectively.

## 4. Method

In this section, we first formulate our problem, multimodal named entity recognition with multiple images, and then introduce our proposed method: TPM-MI, and finally describe the main components in the proposed model: (1) Multi-Image Representation, (2) Projection, (3) Text Representation.



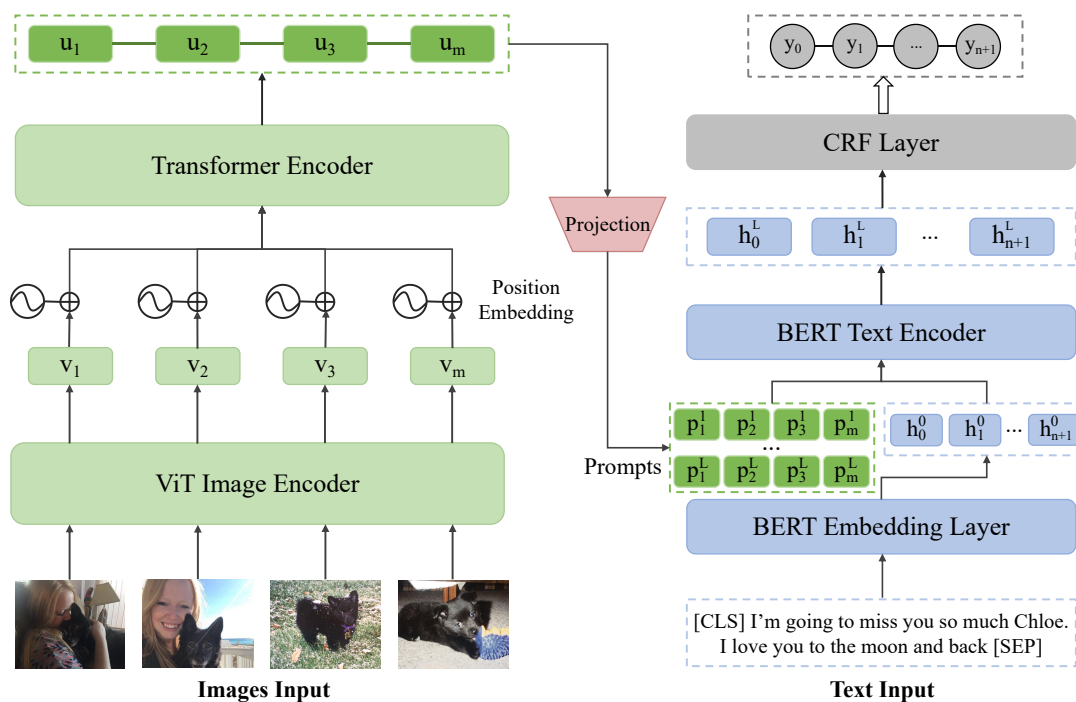


Figure 2: Overall framework of TPM-MI.

#### 4.1. Problem Formulation

Given a text  $S$  and its associated images  $\{I_1, I_2, \dots, I_m\}$  as input, where  $m$  is the maximum number of images ( $m = 4$  in this paper). The objective of MNER with multiple images is to extract the named entities from  $S$  and assign them to one of the pre-defined types. In line with previous existing work on MNER, we approach the task as a sequence labeling problem. Let  $S = (s_1, s_2, \dots, s_n)$  denote a sequence of input words, and  $y = (y_1, y_2, \dots, y_n)$  represent the corresponding label sequence, where  $y_i \in \mathcal{Y}$  and  $\mathcal{Y}$  is the pre-defined label set with the BIOES tagging schema.

#### 4.2. Overall Framework

Our overall framework of TPM-MI is shown in Figure 2, and the overall process is as follows.

For multi-image representation, we first input multiple images into ViT (Dosovitskiy et al., 2020) to obtain a representation of each image separately, then we add a learnable positional embedding for each image to indicate the temporal order, and next input all images representation into a Transformer Encoder (Vaswani et al., 2017) to obtain the overall representation of multiple images. Then, we project the representation of the multiple images as prompts through projection for subsequent interaction with the text. For text representation, we first input the text into the Embedding layer of BERT to obtain a representation of each token, then we input the prompts with the representation of the token into the BERT to obtain the final text representation.

Finally, we input the final text representation into a conditional random field layer (Lafferty et al., 2001) to obtain prediction results.

#### 4.3. Multi-Image Representation

We use ViT (Dosovitskiy et al., 2020) as the image encoder for obtaining the representation of each image in  $m$  input images and use the Transformer Encoder to establish relationships between the images. The entire process is described as follows.

Firstly, we follow (Dosovitskiy et al., 2020) and resize every image to  $224 \times 224$  pixels, then feed them into the ViT, which splits every image into a sequence of  $14 \times 14 = 196$  non-overlapping patches with a pixel size of  $16 \times 16$ , which are then linearly embedded to get each path representation  $(z_1, z_2, \dots, z_{196})$ . Then, a learnable special token [CLS] with the same dimensions as these patches is added at the beginning of them to get  $([CLS], z_1, z_2, \dots, z_{196})$ . Next, we use the representation of the activation at [CLS] token in the last layer of ViT to obtain the representation of the  $i$ -th image  $v_i \in \mathbb{R}^{d_v}$  and get the representation of  $m$  images  $\mathbf{V} = (v_1, v_2, \dots, v_m) \in \mathbb{R}^{d_v \times m}$ , where  $d_v$  is the dimension of the image representation. If the number of images is less than  $m$ , we use the zero vectors to fill it up to  $m$  images.

Then, to indicate the positional and temporal information of multi-images (e.g. the first image usually contains more information and words indicating the location of an image may appear in the text), we further add learnable temporal positional encod-

ing  $\mathbf{T} = (t_1, t_2, \dots, t_m)$  onto  $\mathbf{V}$ :  $\mathbf{C} = \mathbf{V} + \mathbf{T}$ , where  $t_i \in \mathbb{R}^{d_v}$  is the positional embedding of the  $i$ -th image and  $\mathbf{C} \in \mathbb{R}^{d_v \times m}$  is the multi-images representation with position information. This temporal positional encoding is not just an additional feature in our model, it can provide prior knowledge to help our model to dynamically capture these patterns (e.g., which image is the first image).

Finally, to establish relationships between the multi-images for a more global image representation, we model the multi-images as frames in a video, and feed  $\mathbf{C}$  into a Transformer Encoder (Vaswani et al., 2017) following (Ju et al., 2022) to obtain the final multi-images representation  $\mathbf{U} = (u_1, u_2, \dots, u_m) \in \mathbb{R}^{d_v \times m}$  using its own self-attention mechanism.

#### 4.4. Projection

Inspired by multimodal prompt learning (Liang et al., 2022; Li et al., 2023; Khattak et al., 2023) and recent MNER methods (Wang et al., 2022d; Chen et al., 2022; Xu et al., 2023), projecting image representations as prompts can better guide text representations during subsequent interactions with the text. We project  $\mathbf{U}$  as prompts for subsequent interaction with each Transformer layer of the text encoder:

$$\mathbf{P}^l = \mathbf{W}_p^l \mathbf{U}, 1 \leq l \leq L \quad (1)$$

where  $L$  is the number of layers of Transformer in the text encoder,  $\mathbf{P}^l \in \mathbb{R}^{d_t \times m}$  is the prompts corresponding to the  $l$ -th Transformer layer,  $\mathbf{W}_p^l \in \mathbb{R}^{d_t \times d_v}$  is the weight metric corresponding to the  $l$ -th Transformer layer,  $d_t$  is the dimension of the text representation. Projecting the image as a different prompts for each layer in the text encoder can have a better guiding effect on the text.

#### 4.5. Text Representation

We use BERT (Devlin et al., 2019) as the text encoder and feed prompts and text into BERT for interaction. The entire process is described as follows.

Firstly, we follow (Devlin et al., 2019) and add a [CLS] token and a [SEP] token at the beginning and end of the text input as  $S' = (s_0, s_1, \dots, s_n, s_{n+1})$ , where  $s_1$  to  $s_n$  is the original input text. Then we feed the  $S'$  to the Embedding Layer of BERT to get the text representation of the 0-th Transformer layer  $\mathbf{H}^0 = (h_0^0, h_1^0, \dots, h_n^0, h_{n+1}^0) \in \mathbb{R}^{d_t \times (n+2)}$ , where  $d_t$  is the dimension of the text representation.

Then, we input the text representation of the  $(l-1)$ -th layer  $\mathbf{H}^{l-1}$  with  $\mathbf{P}^l$  into the  $l$ -th Transformer layer in BERT to obtain the representation of  $l$ -th layer  $\mathbf{H}^l$ . Specifically, we first project the  $\mathbf{H}^{l-1}$  as the 'queries'  $\mathbf{Q}^l$ , 'keys'  $\mathbf{K}^l$  and 'values'  $\mathbf{V}^l$  of the

$l$ -th layer:

$$\mathbf{Q}^l = \mathbf{W}_Q^l \mathbf{H}^{l-1}; \mathbf{K}^l = \mathbf{W}_K^l \mathbf{H}^{l-1}; \mathbf{V}^l = \mathbf{W}_V^l \mathbf{H}^{l-1} \quad (2)$$

where  $\{\mathbf{W}_Q^l, \mathbf{W}_K^l, \mathbf{W}_V^l\} \in \mathbb{R}^{d_t \times d_t}$  are the weight matrices. Next, we follow (Chen et al., 2022) and project  $\mathbf{P}^l$  as additional 'keys'  $\mathbf{K}_p^l$  and 'values'  $\mathbf{V}_p^l$  for interaction with  $(l-1)$ -th text representation:

$$\mathbf{K}_p^l = \phi_k^l \mathbf{P}^l; \mathbf{V}_p^l = \phi_v^l \mathbf{P}^l \quad (3)$$

$$\mathbf{H}^l = \text{Softmax}\left(\frac{(\mathbf{Q}^l)^T [\mathbf{K}_p^l; \mathbf{K}^l]}{\sqrt{d_t}}\right) [\mathbf{V}_p^l; \mathbf{V}^l]^T \quad (4)$$

where  $\{\phi_k^l, \phi_v^l\} \in \mathbb{R}^{d_t \times d_t}$  are the weight matrices,  $\mathbf{H}^l \in \mathbb{R}^{(n+2) \times d_t}$ , and after  $L$  layers of Transformer, we obtain the final text representation  $\mathbf{H}^L \in \mathbb{R}^{(n+2) \times d_t}$ .

#### 4.6. CRF Decoder

Following (Chen et al., 2022), we adopt the conditional random field (CRF) (Lafferty et al., 2001) decoder to perform the NER task, and we feed the final text representation  $\mathbf{H}^L$  into a standard CRF layer, which predicts the probability of a sequence of predictions  $y$  through the  $\mathbf{H}^L$  as follows:

$$p(y|\mathbf{H}^L) = \frac{\exp(\sum_{i=1}^n E_{h_i, y_i} + \sum_{i=0}^n T_{h_i, y_{i+1}})}{Z(\mathbf{H}^L)} \quad (5)$$

$$Z(\mathbf{H}^L) = \sum_{y \in \mathcal{Y}} \exp\left(\sum_{i=1}^n E_{h_i, y_i} + \sum_{i=0}^n T_{h_i, y_{i+1}}\right) \quad (6)$$

where  $E_{h_i, y_i}$  is the emission score of label  $y_i$  for the  $i$ -th token,  $T_{h_i, y_{i+1}}$  is the transition score from label  $y_i$  to label  $y_{i+1}$ ,  $\mathcal{Y}$  represents the pre-defined label set with the BIO tagging schema. To train the module, we use the log-likelihood loss as our loss function, which is defined as follows:

$$L_{ner} = -\frac{1}{|D_{ner}|} \sum_{j=1}^N \log(p(y|\mathbf{H}^L)) \quad (7)$$

where  $D_{ner}$  is the batch of training examples and  $N$  is the batch size.

## 5. Experiments

In this section, we conduct various experiments to comprehensively evaluate the performance of our proposed datasets MNER-MI and MNER-MI-Plus. Following many recent works (Chen et al., 2022; Xu et al., 2023), we use the precision ( $\mathbf{P}$ ), recall ( $\mathbf{R}$ ) and F1 score ( $\mathbf{F1}$ ) as evaluation metrics.

## 5.1. Baselines

**Text-based models:** For text modality, we explore several well-known models commonly used in named entity recognition tasks. Specifically, we consider BiLSTM-CRF (Huang et al., 2015), which utilizes a bidirectional LSTM with a CRF layer. Building upon BiLSTM-CRF, we also investigate CNN-BiLSTM-CRF (Ma and Hovy, 2016) and HBiLSTM-CRF (Lample et al., 2016), which incorporate additional character-level word representations using CNN and LSTM, respectively. Additionally, we include BERT (Devlin et al., 2019), a powerful transformer-based text encoder, and BERT-CRF, which combines BERT with a CRF-based decoder.

**Multimodal named entity recognition models:** For our multimodal experiments involving both text and image modalities, we use the current representative MNER models as baselines. Specifically, GVATT-HBiLSTM-CRF (Lu et al., 2018) and AdaCAN-CNN-BiLSTM-CRF (Zhang et al., 2018) use the attention mechanism to combine text and images based on HBiLSTM-CRF and CNN-BiLSTM-CRF, respectively. UMT (Yu et al., 2020) proposes a multimodal interaction module for establishing bi-directional relationships between text and images. OCSGA (Wu et al., 2020) uses an object detector to extract the objects in the image and use the text labels of these objects as the image representation. UMGF (Zhang et al., 2021) proposes an approach based on a graph model to establish the relationship between text and images. MAF (Xu et al., 2022b) proposes a general matching and alignment framework to align text and image representations as well as to alleviate the impact of image noise. ITA (Wang et al., 2022b) extracts the objects, caption, and text in the image as the image representation. promptMNER (Wang et al., 2022d), HVPNeT (Chen et al., 2022) and VisualPT-MoE (Xu et al., 2023) all project image representations as prompts to achieve interaction with each layer of the text encoder. All of the above methods are single-image MNER methods that use information from the first image only.

For a fair comparison, we stitch the images into a single image, which allows single-image methods can use the information of multiple images. We apply this approach for UMT, UMGF, and VisualPT-MoE, resulting in UMT-MI, UMGF-MI, and VisualPT-MoE-MI, respectively. TPM-MI is the approach we propose in this paper.

**Large language models:** Considering the advancements in large language models that can perform various tasks in a zero-shot manner. We use a large language model GPT4 and a multimodal large language model MiniGPT-4 (Zhu et al., 2023) as a baseline for studying the performance of large language models on this task.

For the prompts used in ChatGPT, we fol-

lows (Qin et al., 2023) and improve upon it: *‘Please identify Person, Organization, Location and Miscellaneous Entity from the given text, and respond to the result in JSON format that contains the following keys: Person, Organization, Location and Miscellaneous. Text: [Text Input]’*. Since MiniGPT-4 impairs the generative capacity of the language model during training, we provide a more detailed prompt for it: *‘Your task is to perform Named Entity Recognition through the given text and an attached image, in which the named entities exist only in the text. There are four types: Person, Organization, Location and Miscellaneous. You should reply with the results in JSON format containing the following keywords: Person, Organization, Location, and Miscellaneous. The text you should perform named entity recognition on is [Text Input]’*.

## 5.2. Experimental Settings

All experiments are conducted on NVIDIA GeForce RTX 3090 GPUs with PyTorch 1.7.1, and the parameters settings of our model and baselines are as follows:

- We use BERT-base<sup>3</sup> and ViT-base-patch16<sup>4</sup> as the text encoder (except BiLSTM-based methods) and image encoder for all methods, respectively.
- We use the AdamW (Loshchilov and Hutter, 2018) as the optimizer and use the grid search in the development set to find the learning rate within  $[1e^{-5}, 7e^{-5}]$ , the batch size within  $[8, 32]$ . The framework uses mini-batch backpropagation for training. We select the model that performs best on the development set and evaluate it on the test set.

## 5.3. Performance Comparison

As shown in Table 3, we first compare all text-based methods. We find that the BERT-based model significantly outperforms the BiLSTM-based model on both datasets, demonstrating the advantage of pre-trained language models. In addition, we find that GPT4 fails to achieve satisfactory results compared to fine-tuning methods, and this phenomenon is also found by (Qin et al., 2023), which indicates that the current large language model still faces challenges in NER.

Next, we compare multimodal methods of using the first image with the text-based methods. We find that almost all multimodal models significantly outperform their corresponding text-based models on both datasets, such as GVATT-HBiLSTM-CRF

<sup>3</sup><https://huggingface.co/bert-base-uncased>

<sup>4</sup><https://huggingface.co/google/vit-base-patch16-224>

Modality	Model	MNER-MI			MNER-MI-Plus		
		P	R	F1	P	R	F1
Text Only	BiLSTM-CRF	64.03	65.91	64.96	73.65	70.74	72.17
	CNN-BiLSTM-CRF	64.89	66.89	65.87	73.71	71.97	72.83
	GPT4	64.28	67.91	66.05	63.76	69.12	66.33
	HBiLSTM-CRF	64.51	68.55	66.47	72.19	74.34	73.25
	BERT	69.04	73.54	71.22	77.35	79.19	78.26
	BERT-CRF	70.78	75.05	72.85	80.15	78.52	79.33
Text + Single Image	MiniGPT4	59.87	62.37	61.09	62.22	64.27	63.23
	GVATT-HBiLSTM-CRF	67.83	67.19	67.51	76.31	73.11	74.68
	AdaCAN-CNN-BiLSTM-CRF	67.89	68.24	68.06	75.67	73.85	74.75
	OCSGA	75.75	72.04	73.85	81.44	79.13	80.27
	UMT	74.23	74.03	74.13	81.71	79.50	80.59
	MAF	74.91	73.60	74.25	80.17	81.29	80.73
	UMGF	73.74	75.30	74.51	82.31	79.65	80.96
	ITA	74.95	74.21	74.58	79.64	81.46	80.54
	promptMNER	75.80	73.46	74.61	81.13	81.39	81.26
	VisualPT-MoE	74.77	75.01	74.89	82.72	80.64	81.67
HVPNeT	74.93	75.28	75.10	81.88	80.94	81.41	
Text + Multiple Images	UMT-MI	76.56	75.90	76.23	82.26	82.96	82.61
	UMGF-MI	75.88	77.14	76.50	82.55	82.25	82.40
	VisualPT-MoE-MI	76.87	76.38	76.62	82.61	82.79	82.70
	<b>TPM-MI</b>	<b>77.45</b>	<b>77.19</b>	<b>77.32<sup>†</sup></b>	<b>83.66</b>	<b>83.18</b>	<b>83.42<sup>†</sup></b>

Table 3: Performance comparison on MNER-MI and MNER-MI-Plus. The marker † refers to significant test p-value < 0.05 when compared with VisualPT-MoE-MI.

vs. HBiLSTM-CRF, AdaCAN-CNN-BiLSTM-CRF vs. CNN-BiLSTM-CRF, and other MNER methods (except MiniGPT4 vs. BERT-CRF). This indicates that image information in social media posts contributes to named entity recognition. In addition, we find that MiniGPT-4 does not perform better than GPT4, probably because the multimodal large language models are weaker at understanding instructions than text-based large language models.

Finally, we compare multimodal methods of using multiple images. We find that methods using multiple images always perform better than their corresponding single-image methods, indicating that the use of multiple images in multi-image scenarios provides a better understanding of multimodal content and aids MNER. In addition, we find that our proposed methods TPM-MI significantly outperform the other methods, which demonstrates that the image position information and the relationship between images can help obtain a better multi-image representation.

#### 5.4. Performance Comparison on Different Datasets

In Figure 3, we show the performance of models under the multi-image MNER dataset MNER-MI and the single-image MNER dataset TWITTER-2017. We find the models perform well on TWITTER-2017 (each model performs above 85.0), but poorly on MNER-MI. The maximum performance degrada-

tion of the model is more than **10 points** despite the use of multiple images, indicating that current single-image MNER methods are not directly suitable for multi-image scenarios, and that simply transferring existing methods to multi-image scenarios does not achieve satisfactory results, as well as highlighting that our dataset exists with different challenges than the single-image MNER dataset.

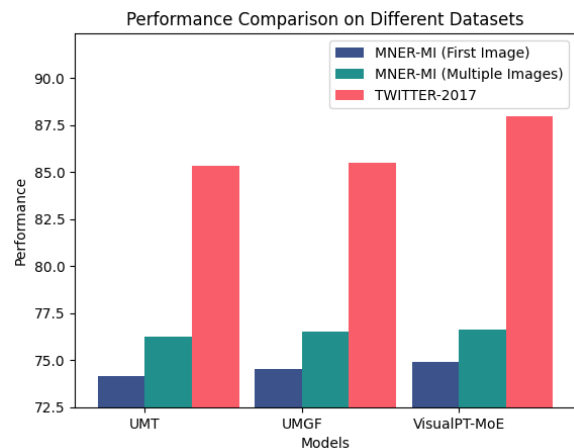


Figure 3: Performance comparison on different datasets. First Image means that the model uses only the representation of the first image, and Multiple Images means that the model uses the representation of all images.





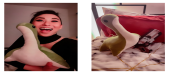
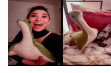


Sentence and Images	Whole Image	Gold	VisualPT-MoE	VisualPT-MoE-MI	TPM-MI
Elon Musk is considering bringing back Vine 		Elon Musk (PER) Vine (ORG)	Elon Musk (PER)	Elon Musk (PER) Vine (ORG)	Elon Musk (PER) Vine (ORG)
After months of being sold out, I finally managed to get myself a Nessie! She favors a wingman. 		Nessie (MISC)	Nessie (PER)	Nessie (MISC)	Nessie (MISC)
Early mornings in the garden with Raffi. 		Raffi (MISC)	Raffi (PER)	Raffi (PER)	Raffi (MISC)

Table 4: A case study for visually showing the effectiveness of multiple images, where the column of the whole image represents the result of stitching multiple images into one whole image.

## 5.5. Case Study

To more visually show the effectiveness of using multiple images in a multi-image scenario, we conduct a case study and compare the different methods as shown in Table 4.

Specifically, we can observe that: the use of multiple images helps to identify the additional entity *Vine* in the first example and removes the ambiguity contained in the single image in the second example, as well as accurately identifying the type of *Nessie*, which demonstrates the multiple images can provide more information and can help models to better understand the multimodal content compared the single image. In addition, we find that VisualPT-MoE-MI incorrectly predicts the type of *Raffi* as PER in the third example, which may be due to the pixel size of the image being reduced and the information of the image being lost in order to stitch the images. TPM-MI can obtain more information about the image by using multiple images directly, which can accurately identify the type of *Raffi* as MISC.

Overall, this case study highlights the benefits of using multiple images for NER, showing how it helps identify additional entities and removes ambiguities from a single image.

## 6. Conclusion

In this paper, in order to address the research gaps in MNER as well as to expand the scope of MNER for real-world applications, we propose a multi-image MNER dataset MNER-MI and extend an MNER dataset MNER-MI-Plus up on it. Based on both datasets, we establish a comprehensive set of representative baseline methods and propose a novel temporal prompt model for the challenges of MNER with multiple images. We have conducted extensive experiments to demonstrate that multiple images can provide more information to better help MNER compared to a single image, and the

effectiveness of our method.

In the future, we plan to further investigate the representation of multiple images. Although we model multiple images as frames in a video in this paper, we recognize the need for more efficient representations to fully capture the unique characteristics of multiple images. In addition, we are aware of the limitations of our approach: we treat each image equally, while in reality, different images have different importance in understanding the post, and we plan to explicitly establish the weight of each image in the future.

## 7. Acknowledgements

We are very grateful to the anonymous reviewers for their hard work and valuable comments. This work is supported by National Science and Technology Major Project (2021ZD0111000/2021ZD0111004), the Science and Technology Commission of Shanghai Municipality Grant (No. 21511100101, 22511105901, 22DZ2229004), the NSF of Shanghai under grant number 22ZR1402000. Xin Lin is the corresponding author.

## 8. Bibliographical References

- Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Good visual guidance make a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1607–1618.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the As-*

- sociation for Computational Linguistics: *Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- 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*.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.
- 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.
- Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. 2022. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, pages 105–124.
- Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122.
- 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.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.
- Yaowei Li, Ruijie Quan, Linchao Zhu, and Yi Yang. 2023. Efficient multimodal fusion via interactive prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2604–2613.
- Sheng Liang, Mengjie Zhao, and Hinrich Schütze. 2022. Modular and parameter-efficient multimodal fusion with prompting. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2976–2985.
- Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.
- 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. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 852–860.
- Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.
- Erik Tjong Kim Sang and Jorn Veenstra. 1999. Representing text chunks. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 173–179.
- Dianbo Sui, Zhengkun Tian, Yubo Chen, Kang Liu, and Jun Zhao. 2021. A large-scale chinese multimodal ner dataset with speech clues. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2807–2818.
- 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.

- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- 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.
- 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.
- Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. 2022c. Wikidiverse: A multimodal entity linking dataset with diversified contextual topics and entity types. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4785–4797.
- Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Jiabo Ye, Ming Yan, and Yanghua Xiao. 2022d. Promptmner: prompt-based entity-related visual clue extraction and integration for multimodal named entity recognition. In *International Conference on Database Systems for Advanced Applications*, pages 297–305. Springer.
- Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. 2020. Multi-modal 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, Ming Du, Hongya Wang, Hui Song, Chaofeng Sha, and Yanghua Xiao. 2022a. Different data, different modalities! reinforced data splitting for effective multimodal information extraction from social media posts. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1855–1864.
- Bo Xu, Shizhou Huang, Ming Du, Hongya Wang, Hui Song, Yanghua Xiao, and Xin Lin. 2023. A unified visual prompt tuning framework with mixture-of-experts for multimodal information extraction. In *International Conference on Database Systems for Advanced Applications*, pages 544–554. Springer.
- Bo Xu, Shizhou Huang, Chaofeng Sha, and Hongya Wang. 2022b. 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.
- Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021. 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.
- 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.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.