TVOMNER: Textual and Visual Feature Optimization for Multimodal Named Entity Recognition

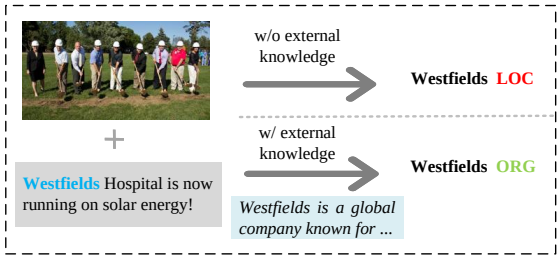
Anonymous EMNLP submission

Abstract

Multimodal Named Entity Recognition (MNER) aims to extract named entities from text by leveraging both textual and visual modalities. Although existing methods focus on enhancing cross-modal interaction or reducing the interference of irrelevant images, two major challenges remain: (1) the textual content is often short and informal, lacking sufficient context to accurately identify ambiguous or low-frequency entities; (2) fine-grained entity information in images that is relevant to the text is rarely utilized. To address these challenges, we propose TVOMNER, a novel framework that focuses on Textual and Visual feature Optimization for MNER. For textual optimization, the model retrieves external knowledge of candidate entities from Wikipedia and incorporates it into the original text to provide richer semantic context. For visual optimization, it integrates (a) heterogeneous text-guided features via a variational autoencoder (VAE), (b) global visual features generated by a visual encoder, and (c) fine-grained entity object-level visual features extracted by large language models (LLMs) and visual grounding (VG) models. These features are adaptively fused and integrated with the textual representation for a subsequent cross-modal attention mechanism and a dynamic gating module. Extensive experiments on the two widely used datasets show that TVOMNER outperforms all baselines and exhibits robust and competitive performance.

1 Introduction

In recent years, Multimodal Named Entity Recognition (MNER) (Lu et al., 2018) has gained significant attention, especially in the context of social media platforms, where textual content is often accompanied by rich visual data. Unlike traditional Named Entity Recognition (NER), which relies solely on text to identify entities such as people, organizations, and locations (Li et al., 2020), MNER incorporates both text and images to enhance entity recognition. It has been widely applied to various downstream natural language processing (NLP) tasks such as relation extraction (Zelenko et al., 2003) and entity linking (Ganea and Hofmann, 2017).



(a) Example of external knowledge

(b) Example of relevant objects

Figure 1: Two examples for the MNER task.

Previous work has primarily focused on two aspects. First, considerable effort has been dedicated to leveraging various attention mechanisms to better align and fuse features from text and image modalities (Yu et al., 2020; Bao et al., 2023). Second, some researchers concentrate on mitigating noise introduced by irrelevant images, which could interfere with the model's ability to extract informative visual features (Sun et al., 2021; Xu et al., 2022; Bai et al., 2025).

Despite the advancements, current methods still

exhibit some limitations. One major issue arises from the fact that text on social media posts is often short and informal, making it difficult for models to correctly classify entities without sufficient context (Ok et al., 2024). In many cases, the model lacks the necessary knowledge to accurately determine the correct entity type, particularly when entities are ambiguous or underrepresented in the training data. As illustrated in Figure 1(a), without access to external knowledge, the model misclassifies "Westfields" as a Location. However, "Westfields" is actually a global company that should be recognized as an Organization. This highlights the need for MNER models that incorporate external knowledge sources to enhance contextual understanding and improve entity recognition in short and noisy social media text.

On the other hand, most existing approaches rely primarily on global visual features extracted by the image encoder (Yu et al., 2020; Wang et al., 2023; Wei et al., 2024), often overlooking fine-grained entity objects present within the image. While some methods attempt to address this by applying simple object detection techniques to extract localized object representations (Wang et al., 2022a; Zheng et al., 2024), they still face challenges. In particular, the presence of irrelevant objects—common in real scenarios—could introduce noise and mislead the model's judgment. As shown in Figure 1(b), with the relevant objects person and the logo of "Harlem Globetrotters", the model could achieve accurate recognition. However, without effective mechanisms to filter or prioritize relevant objects, the model may focus on distracting visual elements, ultimately degrading the accuracy of MNER. Furthermore, when image and text are unrelated, the model should place greater emphasis on the textual information, since the text remains the primary source for entity identification in MNER tasks.

To address the above problems, we propose a novel framework TVOMNER, which focuses on **T**extual and **V**isual features **O**ptimization for the MNER task. Specifically, for textual features, we first retrieve candidate entities in the text from Wikipedia to acquire relevant entity knowledge. This knowledge is then concatenated with the original text and fed into a text encoder to obtain the optimized textual features. For visual features, we process in three complementary ways: (1) we utilize a variational autoencoder (VAE) to mine shared semantic information from heterogeneous text-guided features; (2) we extract global visual features by a standard visual encoder; and (3) we leverage the strong reasoning capabilities of large language models (LLMs) to infer entity-related objects in the image, and extract them via a visual grounding (VG) model. Then we encode these objects to obtain fine-grained, entity object-level visual features. These three types of visual features are adaptively fused to generate optimized visual features. Finally, the optimized textual and visual features are jointly utilized to perform the MNER task. This design enables the model to incorporate external knowledge, pay more attention to fine-grained image entities that are semantically aligned with the text, and maintain robust textual understanding when visual content is irrelevant.

In summary, the main contributions of this paper are as follows:

- We propose TVOMNER, a novel MNER model that focuses on optimizing both textual and visual features. External knowledge of candidate entities is used to optimize textual features while visual features are optimized by fusing heterogeneous text-guided, global, and fine-grained entity object-level visual features.
- Unlike traditional object detection approaches, we leverage the reasoning strength of LLMs combined with a VG model to more accurately extract semantically relevant entities from images, thereby mitigating the impact of unrelated visual noise.
- Extensive experiments on the Twitter-2015 and Twitter-2017 datasets demonstrate the effectiveness of TVOMNER. Ablation and case studies further confirm that each component of the model contributes meaningfully to overall performance.

2 Related Work

MNER extends traditional NER by incorporating both textual and visual information for more accurate entity recognition. Previous methods can be broadly categorized into two types: enhancing **cross-modal interaction** to better align and fuse multimodal features, and mitigating the impact of **irrelevant visual content**.

Cross-modal Interaction. Early approaches typically use simple attention mechanisms to fuse

features between two modalities (Lu et al., 2018; Zhang et al., 2018). With the advancement of the Transformer architecture (Vaswani et al., 2017) , several studies have employed multi-head cross-attention modules to improve cross-modal fusion (Yu et al., 2020; Zhang et al., 2021a; Wang et al., 2022b; Chen et al., 2022; Zhang et al., 2023; Wei et al., 2024). Furthermore, Zeng et al. (2024) utilize the inter-modality connections as a bridge to construct an instruction that fuses multimodal features. Li et al. (2025) implement entity-level language reinforcement in an adaptive multi-scale way. Zhao et al. (2025) design heterogeneous graphs and introduce graph Transformer to enable effective information interaction.

Disambiguation of Irrelevant Images. Another major line of research focuses on addressing the issue of irrelevant or noisy visual information. Chen et al. (2020); Zhang et al. (2021b); Xu et al. (2022) apply contrastive learning to address this issue by constructing positive and negative samples in different ways. Xu et al. (2025) propose an adaptive mix-up image augmentation strategy to harness the complementary benefits of both original and synthesized images. Bai et al. (2025) use CLIP (Radford et al., 2021) prompts to accurately capture visual cues associated with entities. More recently, approaches have begun to incorporate the strong reasoning capabilities of LLMs to further improve MNER performance (Li et al., 2023, 2024). Moreover, some researchers have extended the MNER task from the perspective of visual grounding (Yu et al., 2023; Jia et al., 2023).

However, existing models still struggle with ambiguous entities due to the lack of knowledge and often underutilize fine-grained, object-level visual information. In this work, we tackle these limitations by incorporating external knowledge retrieval for textual optimization and utilizing LLM-based reasoning to extract fine-grained visual entities.

3 Task Definition

Given a sentence $S = \{s_1, s_2, ..., s_n\}$ and its corresponding image I, the MNER task aims to recognize the named entities in S and classify them into specific pre-defined types like *Person* (PER), *Organization* (ORG), *Location* (LOC) and *Miscellaneous* (MISC). Following the BIO tagging schema (Sang and Veenstra, 1999), the output would be $Y = \{y_1, y_2, ..., y_n\}$ where $y_i \in$ {*B-type*, *I-type*, *O*} indicates the label corresponding to s_i, and *type* refers to the above four pre-defined entity types.

4 Methodology

In this section, we introduce the architecture of TVOMNER, which focuses on the optimization of textual and visual features. As shown in Figure 2, TVOMNER consists of three main parts: 1) the *Textual Feature Optimization* (TFO) module; 2) the *Visual Feature Optimization* (VFO) module; 3) the *Cross-modal Feature Fusion* (CFF) module. In the TFO module, we extract the candidate entities in the text and search them in Wikipedia to provide supplementary knowledge to the original text. In the VFO module, we optimize visual features from three perspectives: heterogeneous text-guided features, global visual features, and entity object-level visual features. Finally, in the CFF module, we fuse the optimized features of the two modalities and combine a gating mechanism with the CRF layer for the final sequence labeling.

4.1 Textual Feature Optimization Module

Due to the limited content of text, the model struggles to obtain sufficient contextual information. Therefore, we optimize the textual features in terms of supplementing external knowledge to the framework. This module retrieves the candidate entities in the text from Wikipedia, filters the retrieved knowledge to remove irrelevant noise, and then splits it with the original text into a transformer-based language model.

In the first stage, an entity candidate detection module extracts potential entities from the input text. Inspired by (Ok et al., 2024), we use the RoBERTa (Liu et al., 2019) encoder with BIO tagging to classify each token in the text, determining whether it belongs to an entity span. And this process is guided by cross-entropy loss. After that, we regard the entity candidates as queries to retrieve structured knowledge from Wikipedia. The retrieved entity knowledge introduces global context from Wikipedia, which helps the model understand entities beyond their immediate textual surroundings. At the same time, we perform the calculation of semantic relevance between the retrieved results and the original text to filter low-quality fragments. From Figure 2(a), for entity candidates *Steph Curry* and *NBA*, it would provide the model that *Steph Curry* is a basketball player while *NBA* is a basketball league.

The retrieved entity knowledge is then concate-

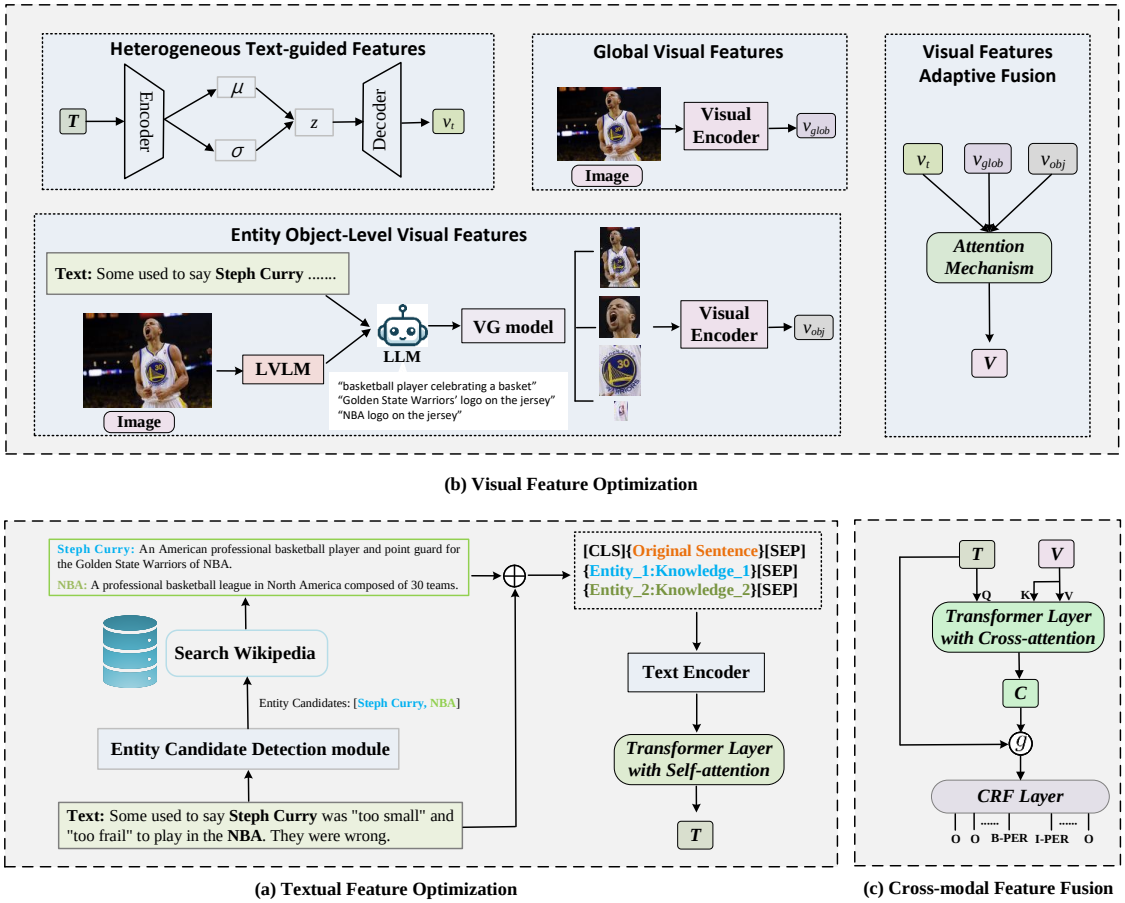


Figure 2: The framework of TVOMNER. It consists of three main modules: (a) Textual Feature Optimization; (b) Visual Feature Optimization; (c) Cross-modal Feature Fusion.

nated with the original text, forming an augmented input representation. We build a template and add [CLS] and [SEP] tokens that conform to the input format of the text encoder. The template is represented as follows:

"[CLS] {*Original Sentence*} [SEP] {*Entity_1*: *Knowledge_1*} [SEP] {*Entity_2*: *Knowledge_2*} [SEP] ... {*Entity_n*: *Knowledge_n*} [SEP] "

where $Entity\_i$ and $Knowledge\_i$ denote the entity candidates and their corresponding retrieved knowledge, respectively.

Subsequently, we also utilize RoBERTa to encode the retrieval-augmented text S_{rag} to get the token-level representation $T_{rag} = \{t_{[CLS]}, t_1, ..., t_n, t_{[SEP]}, t_{n+1}..., t_{n+m}, t_{[SEP]}\}$, where n and m represent the length of the original text and retrieved knowledge, respectively. Then we feed T_{rag} into a Transformer-based self-attention module to obtain $\hat{T}$:

$$\hat{T} = \text{LN}(\text{MP}(\text{softmax}(\frac{QK^T}{\sqrt{d_t}})V)) \quad (1)$$

where $Q = K = V = \hat{T}$ as query, key, and value. LN is layer normalization (Ba et al., 2016) and MP is mean pooling. It leverages the multi-head self-attention architecture to dynamically capture long-range dependencies and contextual relationships within the text. Finally, we obtain optimized textual features T by taking the first n tokens from $\hat{T}$ as an equal-length sequence with the original sentence S for subsequent processing.

4.2 Visual Feature Optimization Module

In the MNER task, the text, as the subject, contains adequate shared semantic information of textual and visual modalities. Especially in irrelevant text-image pairs, text plays a more important role in the inference of models. Similarly, the global context of relevant images, in conjunction with the entities present in them, furnishes the model with a substantial amount of entity-related information. The integration of such information is of paramount importance. Consequently, we optimize visual fea-

tures by incorporating heterogeneous text-guided features, global visual features, and entity object-level visual features.

Heterogeneous Text-guided Features. Following (Sui et al., 2024), we choose to utilize the cross-reconstruction of VAE (Yi et al., 2023) to extract the semantic information shared with the visual modality from heterogeneous textual features.

As in Figure 2 (b) depicted, the VAE encoder takes the optimized T as input and obtains the mean vector μ and the standard deviation vector σ of a latent distribution: $\mu = TW_\mu$, $\sigma = TW_\sigma$, where W_μ and $W_\sigma \in \mathbb{R}^{d_T \times d_z}$ are trainable matrices. Using the reparameterization strategy (Kingma et al., 2013), we sample a latent variable $z = \mu + \sigma \odot \epsilon$ from this distribution, where $\epsilon \in (0, \mathbf{I})$ and $\mathbf{I}$ is an identity matrix. Then z is decoded to generate text-guided visual features $v_t = zW_z$, where $W_z \in \mathbb{R}^{d_z \times d_v}$ and d_v is the dimension of the visual features.

Finally, the training loss of VAE is calculated as follows:

$$L_{VAE} = \|v_{glob} - v_t\|^2 + \mathrm{KL}(q(z|T)\|p(z)) \quad (2)$$

where v_{glob} is the global feature of images, which will be introduced in the next section. $\mathrm{KL}(\cdot)$ represents the Kullback-Leibler divergence between two distributions. $q(z|T) = N(\mu, \sigma^2)$ is the distribution of z and $p(z) = N(0, \mathbf{I})$ is a standard normal distribution.

Global Visual Features. As one of the two modalities, the information of images is undeniably crucial. To capture the global features of images, we use Vision Transformer (ViT) (Dosovitskiy et al., 2020) as the basic image encoder.

Given the input image I, it is divided into N patches, each represented as a vector $y_i \in \mathbb{R}^{d_v}$. A learnable position embedding $P \in \mathbb{R}^{N \times d_v}$ is added to the patch embeddings to incorporate spatial information. The Transformer encoder processes the sequence of embeddings through multiple layers, producing the final output $H \in \mathbb{R}^{N \times d_v}$. The [CLS] token is appended to the sequence, and its final hidden state is used as the global visual feature representation $v_{glob} \in \mathbb{R}^{d_v}$.

Entity Object-Level Visual Features. Entities within images serve as primary carriers of semantic information. Extracting features of entities in images that are relevant to the text can significantly enhance the model's comprehension of named entities. To eliminate the interference of irrelevant information in images, we leverage the advanced reasoning capabilities of LLMs to accurately identify visual entities relevant to the text.

Inspired by Jian et al. (2024), we utilize a large visual-language model (LVLM) to convert images into corresponding detailed textual captions. Then we feed the original text and image captions into LLMs to extract entities in images along with their attributes potentially relevant to the text. The prompt templates are presented in Appendix A. For example, in Figure 2(b), given the text and image captions, the outputs of LLM would be ["*basketball player celebrating a basket*", "*Golden State Warriors' logo on the jersey*", "*NBA logo on the jersey*"], which are considered relevant to the text.

After obtaining the relevant entities and their attributes, we employ an existing referring expression comprehension VG model (Liu et al., 2024b) to localize the regions of interest (RoIs) in images associated with these entities. Specifically, we use the outputs of LLM as referring expressions, which are fed into the model for subsequent visual grounding. Similar to the process of global visual features, we also utilize ViT as visual encoder to capture entity object-level visual features. For the i-th RoI of visual entities, the embedding of it would be $v_i \in \mathbb{R}^{d_v}$. Then we concatenate all of the vectors to get $v_{sum} \in \mathbb{R}^{N \times d_v}$, where N is the number of RoIs. Finally, we apply the average pooling to v_{sum} to obtain the final vector $v_{obj} \in \mathbb{R}^{d_v}$ that encapsulates rich visual entity information.

Visual Features Adaptive Fusion. Now we have obtained text-guided visual features v_t, global visual features v_{glob} and entity object-level visual features v_o, we employ an attention mechanism to adaptively fuse them so as to generate the optimized visual features for subsequent processing of TVOMNER. Inspired by (Sui et al., 2024), we feed them into different feedforward neural networks (FFNNs) and utilize the sigmoid function to get the fusion weights ω_t, ω_g and ω_o. Then we use the following equation to calculate the final optimized visual features V:

$$V = \omega_t \cdot v_t + \omega_g \cdot v_{glob} + \omega_o \cdot v_{obj} \quad (3)$$

where we set $\omega_t + \omega_g + \omega_o = 1$.

4.3 Cross-modal Feature Fusion module

In order to integrate the optimized visual and textual features, this module employs a multi-head cross-attention mechanism to obtain cross-modal fusion features C. The textual features T serve as query, while the visual features V serve as key and value. Layer normalization and FFNN are subsequently employed to obtain the final output feature. After the L-th layer, we acquire the cross-modal fusion features $C = \{c_{cls}, c_1, \ldots, c_n, c_{sep}\}$, which have abundant multimodal information, thereby laying a solid foundation for TVOMNER to explore the relationship between the two modalities.

After that, we feed C and T into a gating mechanism to dynamically determine the fusion weights of them. The gate g is calculated as follows:

$$g = \text{sigmoid}(\ (\text{LN}(c_{[cls]} + t_{[cls]})W_g^1)\, W_g^2\ ) \quad (4)$$

where $W_g^1 \in \mathbb{R}^{dt \times dt}$ and $W_g^2 \in \mathbb{R}^{dt \times 1}$. Then we get the final token-level features R:

$$R = g \odot T + (1 - g) \odot C \quad (5)$$

R is subsequently input into the Bi-directional Long Short-Term Memory (BiLSTM) network and Conditional Random Field (CRF, Huang et al. (2015)) decoder for final entity recognition:

$$e_{hidden} = \{e_1, e_2, \cdots, e_n\} = \text{BiLSTM}(R) \quad (6)$$

$$p(y|e_{hidden}) = \frac{\exp(\sum_i E^{y_i} e_i + \text{Tr}(y_{i-1}, y_i, e))}{\sum_{y'} \exp(\sum_i E^{y'_i} e_i + \text{Tr}(y'_{i-1}, y'_i, e))} \quad (7)$$

where y_i is the predicted label possibility of the i-th token. And the training loss function of this is:

$$L_{NER} = -\frac{1}{N} \sum_{i=1}^{N} \log p(y|e_{hidden}) \quad (8)$$

The CRF layer can leverage global information to ensure that these words are correctly labeled as the same entity, rather than labeling each word in isolation.

Finally, the loss function of our entire framework is calculated as:

$$L = L_{NER} + \lambda L_{VAE} \quad (9)$$

where λ is a hyperparameter.

5 Experiments

5.1 Settings

Datasets We conducted our experiments on two publicly available and widely used English MNER datasets: Twitter-2015 (Zhang et al., 2018) and Twitter-2017 (Lu et al., 2018). There are four predefined entity types: *Person* (PER), *Organization* (ORG), *Location* (LOC) and *Miscellaneous* (MISC). The detailed data distribution is presented in Appendix B Table 3.

Implementations All experiments are conducted on a NVIDIA 4090 GPU with Pytorch 2.1.0. We use RoBERTa-large and ViT as basic encoders, and employ Minigpt-v2 (Chen et al., 2023) to generate image captions and gpt-4o-2024-11-20 (OpenAI) to output referring expressions. All images are reshaped into 224×224 resolution, and the patch size N is 32. We set the learning rate and dropout rate to 3e-5 and 0.3. And we utilize AdamW (Loshchilov and Hutter, 2017) as the optimizer with a batch size of 64. The hyperparameter λ is set to 1e-3.

Evaluation Consistent with the majority of MNER tasks, we employ precision (Pre.), recall (Rec.), and the F1 score (F1) to evaluate the performance of the proposed model and use the F1 score for each entity type.

Baselines We compare our method with the following representative baselines, which contain text-only NER and MNER models:

- Text-only: BiLSTM-CRF (Huang et al., 2015); CNN-BiLSTM-CRF (Ma and Hovy, 2016); BERT-BiLSTM-CRF (Souza et al., 2020).
- Multimodal: UMT (Yu et al., 2020); MAF (Xu et al., 2022); M3S (Wang et al., 2023); GPT-NER (Li et al., 2023); MGCMT (Liu et al., 2024a); ICKA (Zeng et al., 2024); VEC-MNER (Wei et al., 2024); AMLR (Li et al., 2025), VCRMNER (Bai et al., 2025).

5.2 Main Results

Table 1 illustrates the experimental results between our proposed method and other baseline models on the two Twitter datasets. We have the following observations:

Firstly, our model significantly outperforms all baseline methods, which verifies the effectiveness

Modality	Model	Twitter-2015							Twitter-2017						
		PER	**LOC**	**ORG**	**MISC**	**Pre.**	**Rec.**	**F1**	**PER**	**LOC**	**ORG**	**MISC**	**Pre.**	**Rec.**	**F1**
Text	BiLSTM-CRF	76.77	72.56	41.33	26.80	68.14	61.09	64.42	85.12	72.68	72.50	52.56	79.42	73.43	76.31
	CNN-BiLSTM-CRF	80.86	75.39	47.77	32.61	66.24	68.09	67.15	87.99	77.44	74.02	60.82	80.00	78.76	79.37
	BERT-BiLSTM-CRF	84.74	80.51	60.27	37.29	69.22	74.59	71.81	90.25	83.05	81.13	62.21	83.32	83.57	83.44
Text+Vision	UMT	85.24	81.58	63.03	39.45	71.67	75.23	73.41	91.56	84.73	82.24	70.10	85.28	85.34	85.31
	MAF	85.67	81.69	61.82	40.42	72.02	75.25	73.60	90.91	84.51	84.30	70.51	86.15	85.64	85.89
	M3S	86.05	81.32	62.97	41.36	74.92	75.14	75.03	92.73	84.81	82.49	69.53	86.93	85.21	86.06
	GPT-NER	–	–	–	–	42.96	75.37	54.73	–	–	–	–	52.19	75.03	61.56
	MGCMT	–	–	–	–	73.57	75.59	74.57	–	–	–	–	86.03	76.16	86.09
	ICKA	87.01	83.85	65.87	48.28	72.36	78.75	75.42	93.99	87.24	86.24	75.76	85.13	89.19	87.12
	VEC-MNER	86.11	81.03	62.86	40.60	74.56	75.23	74.89	93.88	81.27	85.49	73.40	87.42	87.61	87.51
	AMLR	85.90	82.19	65.95	40.20	75.45	75.20	75.31	93.22	86.13	84.46	68.42	86.96	86.90	86.93
	VCRMNER	–	–	–	–	75.48	**78.23**	76.83	–	–	–	–	87.76	**89.79**	88.76
	TVOMNER	**87.53**	**84.14**	**67.44**	**50.11**	**77.54**	77.01	**77.27**	**95.47**	**88.50**	**87.08**	**76.33**	**89.76**	88.92	**89.34**

Table 1: Performance comparison on Twitter-2015 and Twitter-2017 datasets.

of our dual-modality feature optimization strategy. TVOMNER surpasses the strongest baseline VCRMNER, with F1-score improvements of 0.44 and 0.58 on the two datasets, respectively. Unlike the methods that decompose MNER into multiple stages same as us (VCRMNER, AMLR, MAF, M3S *.etc*), our method incorporates external knowledge from Wikipedia and leverages the strong reasoning capabilities of LLMs to identify relevant entity objects in images. While several utilize the knowledge of pre-trained models such as CLIP (ICKA) or purely employ ChatGPT to perform MNER (GPT-NER), our TVOMNER integrates multimodal knowledge from an LVLM Minigpt-v2, substantially improving the reasoning accuracy of LLMs in cross-modal scenarios. Furthermore, when the image is irrelevant to the text, TVOMNER could pay more attention to textual information, ensuring superior performance.

In addition, we observe that under unimodal settings, the BERT-based model demonstrates significant advantages over other approaches. This indicates the critical importance of pre-trained language models for MNER tasks. Furthermore, multimodal approaches consistently outperform unimodal methods, as visual information provides complementary context that enhances the model's understanding of textual entities.

5.3 Ablation Study

To further validate the contribution of each components, we conduct a series of ablation studies, with experimental results summarized in Table 2. We systematically designed experiments for the TFO and VFO modules - the two most influential components of TVOMNER - as follows:

Method	Twitter-2015			Twitter-2017		
	Pre.	**Rec.**	**F1**	**Pre.**	**Rec.**	**F1**
TVOMNER	**77.54**	**77.01**	**77.27**	**89.76**	**88.92**	**89.34**
w/o O-T	76.79	76.42	76.60	88.57	88.03	88.30
w/o Wiki	77.12	76.73	76.92	88.93	88.56	88.74
w/o LLMs	77.34	76.37	76.85	89.03	88.21	88.62
w/o VAE	76.90	76.25	76.57	88.79	87.94	88.36
w/o O-V	73.64	76.02	74.81	87.12	86.45	86.78
w/o O-T-V	72.43	75.88	74.11	86.75	85.30	86.02

Table 2: Results of ablation experiments.

'**w/o** O-T' removes the entity candidates; '**w/o** Wiki' removes knowledge from Wikipedia; '**w/o** LLMs' removes the entities and their attributes output from LLMs; '**w/o** VAE' removes the heterogeneous text-guided features generated from VAE; **w/o** O-V replaces the optimized visual features with global features. '**w/o** O-T-V' replaces both the optimized textual and visual features. The detailed explanation of them is listed in Appendix C.

Effects of TFO module. As shown in Table 2, '**w/o** O-T' drops 0.67 and 1.04 F1 scores, and '**w/o** VAE' drops 0.35 and 0.60 F1 scores on the two datasets, respectively. We can conclude that emphasizing candidate entities in the text can guide the model's attention more effectively toward them, thereby improving recognition performance. When relevant external knowledge about these entities is further incorporated, the model gains a deeper understanding of their semantic meaning, which facilitates more accurate results.

Effects of VFO module. According to Table 2, we can also see that '**w/o** LLMs' drops 0.42 and

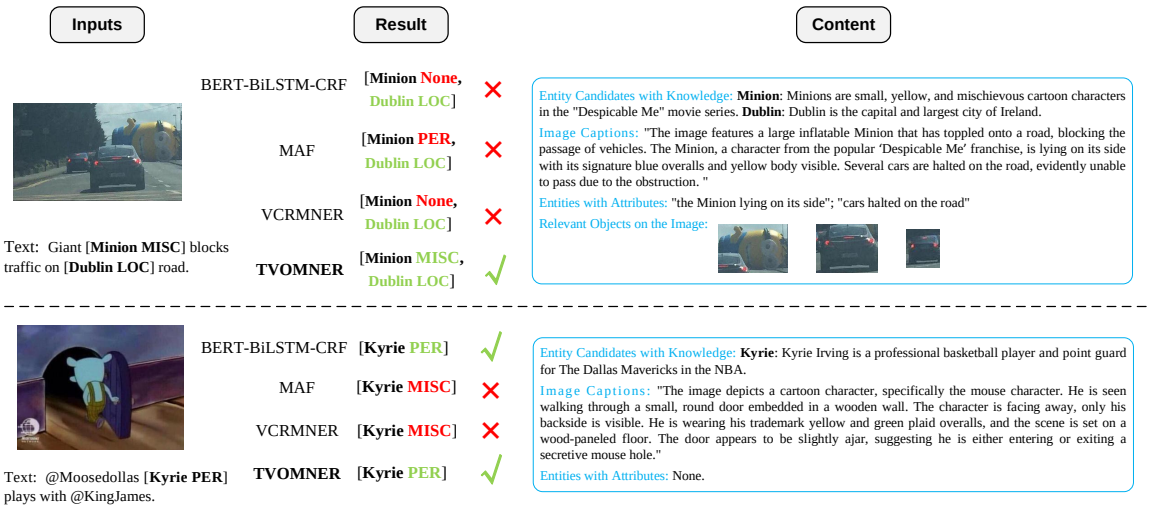


Figure 3: Case study. Each case is accompanied by the recognition results of four models. And the right part is generated by our model TVOMNER.

0.72 F1 scores and '**w/o** VAE' drops 0.70 and 0.98 F1 scores, respectively. When the relevant entity objects extracted by the LLMs are replaced with random objects, the model performance declined, indicating the critical role of semantically relevant visual entities. Furthermore, without heterogeneous text-guided features, the model can be distracted by irrelevant visual content, leading to the omission of implicit visual semantics embedded in the text. Finally, when optimized visual features are replaced by global features, it drops the F1 scores by 2.46 and 2.56, demonstrating the critical impact of the VFO module.

Effects of Both. When we replace the optimized textual and visual features, the F1 scores drop by 3.12 and 3.33, indicating that the TFO and VFO modules work synergistically and play a vital role in maintaining high model performance.

5.4 Case Study

To further illustrate the effectiveness of TVOMNER, we present two examples selected from the datasets in Figure 3. We choose three classic baselines to compare with TVOMNER and the right part Content is generated by our model.

In the first example, our model correctly identifies "Minion" as a MISC entity, whereas the other three models fail to do so. As shown on the right, TVOMNER retrieves the candidate entity "Minion" from Wikipedia and obtains knowledge that it is a well-known cartoon character. Using this external knowledge in conjunction with the visual content, TVOMNER is able to accurately recognize it as a MISC entity.

In the second example, both our model and BERT-BiLSTM-CRF correctly identify "Kyrie" as a person, while the other models fail. In this case, the accompanying image is irrelevant to the text. As a result, the text-only method BERT-BiLSTM-CRF makes the correct prediction without being affected by visual noise. In contrast, the two multimodal models, MAF and VCRMNER, are misled by irrelevant visual content. Notably, the LLM determines that there are no entities in the image related to the text and therefore outputs nothing. This enables TVOMNER to focus more on the heterogeneous textual information, resulting in the correct classification of "Kyrie" as a person.

6 Conclusion

In this paper, we propose TVOMNER, a novel framework for the MNER task that focuses on textual and visual feature optimization. By retrieving candidate entities from Wikipedia, TVOMNER enhances its understanding of ambiguous or rare entities that may not be well represented during training. Through the adaptive fusion of heterogeneous text-guided features, global visual features, and fine-grained object-level visual cues, the model effectively attends to relevant visual entities. Moreover, when the image and text are irrelevant, TVOMNER is able to rely more heavily on textual information. Experimental results demonstrate the superior performance of our proposed TVOMNER framework.

7 Limitations

Although our TVOMNER has demonstrated its effectiveness on the MNER task, there are still some limitations to be addressed in the future: 1) About the heterogeneous text-guided features, we employ VAE to generate the features. However, there would be more advanced generation models that can be utilized. 2) The application potential of LLMs has yet to be fully exploited. In the future, we will explore more methods to utilize the powerful capabilities of LLMs, thereby improving the performance of the MNER models.

8 Ethics Statement

All models and datasets utilized in this study are publicly available and distributed under permissible licenses. The training data has been fully desensitized.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Yu Bai, Lianji Wang, Xiang Liu, Haifeng Chi, and Guiping Zhang. 2025. Vcrmner: Visual cue refinement in multimodal ner using clip prompts. *Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning@ COLING 2025*, page 61.

Xigang Bao, Mengyuan Tian, Zhiyuan Zha, and Biao Qin. 2023. Mpmrc-mner: A unified mrc framework for multimodal named entity recognition based multimodal prompt. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 47–56.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

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

Xiang Chen, Ningyu Zhang, Lei Li, 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, Seattle, United States. Association for Computational Linguistics.

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

Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep joint entity disambiguation with local neural attention. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2619–2629.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging.

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

Pu Jian, Donglei Yu, and Jiajun Zhang. 2024. Large language models know what is key visual entity: An llm-assisted multimodal retrieval for vqa. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10939–10956.

Diederik P Kingma, Max Welling, et al. 2013. Auto-encoding variational bayes.

Enping Li, Tianrui Li, Huaishao Luo, Jielei Chu, Lixin Duan, and Fengmao Lv. 2025. Adaptive multi-scale language reinforcement for multimodal named entity recognition. *IEEE Transactions on Multimedia*.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering*, 34(1):50–70.

Jinyuan Li, Han Li, Zhuo Pan, Di Sun, Jiahao Wang, Wenkun Zhang, and Gang Pan. 2023. Prompting chatgpt in mner: Enhanced multimodal named entity recognition with auxiliary refined knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2787–2802.

Jinyuan Li, Han Li, Di Sun, Jiahao Wang, Wenkun Zhang, Zan Wang, and Gang Pan. 2024. LLMs as bridges: Reformulating grounded multimodal named entity recognition. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1302–1318, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Peipei Liu, Gaosheng Wang, Hong Li, Jie Liu, Yimo Ren, Hongsong Zhu, and Limin Sun. 2024a. Multi-granularity cross-modal representation learning for named entity recognition on social media. *Information Processing & Management*, 61(1):103546.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2024b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

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

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Hyunjong Ok, Taeho Kil, Sukmin Seo, and Jaeho Lee. 2024. Scanner: Knowledge-enhanced approach for robust multi-modal named entity recognition of unseen entities. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7718–7730.

OpenAI. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/.

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

Erik F Sang and Jorn Veenstra. 1999. Representing text chunks. *arXiv preprint cs/9907006*.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer.

Xuhui Sui, Ying Zhang, Yu Zhao, Kehui Song, Baohang Zhou, and Xiaojie Yuan. 2024. Melov: Multimodal entity linking with optimized visual features in latent space. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 816–826.

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.

Jie Wang, Yan Yang, Keyu Liu, Zhiping Zhu, and Xiaorong Liu. 2023. M3s: Scene graph driven multi-granularity multi-task learning for multi-modal ner. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:111–120.

Xinyu Wang, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang, and Kewei Tu. 2022a. 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, Jiabo Ye, Zhixu Li, Junfeng Tian, Yong Jiang, Ming Yan, Ji Zhang, and Yanghua Xiao. 2022b. Cat-mner: multimodal named entity recognition with knowledge-refined cross-modal attention. In *2022 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE.

Pengfei Wei, Hongjun Ouyang, Qintai Hu, Bi Zeng, Guang Feng, and Qingpeng Wen. 2024. Vecmner: Hybrid transformer with visual-enhanced cross-modal multi-level interaction for multimodal ner. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 469–477.

Bo Xu, Shizhou Huang, Chaofeng Sha, and Hongya Wang. 2022. Maf: A general matching and alignment framework for multimodal named entity recognition. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 1215–1223, New York, NY, USA. Association for Computing Machinery.

Bo Xu, Haiqi Jiang, Jie Wei, Hongyu Jing, Ming Du, Hui Song, Hongya Wang, and Yanghua Xiao. 2025. Enhancing multimodal named entity recognition through adaptive mixup image augmentation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1802–1812.

Jing Yi, Yaochen Zhu, Jiayi Xie, and Zhenzhong Chen. 2023. Cross-modal variational auto-encoder for content-based micro-video background music recommendation. *IEEE Transactions on Multimedia*, 25:515–528.

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

Jianfei Yu, Ziyan Li, Jieming Wang, and Rui Xia. 2023. Grounded multimodal named entity recognition on

social media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9141–9154.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *Journal of machine learning research*, 3(Feb):1083–1106.

Qingyang Zeng, Minghui Yuan, Jing Wan, Kunfeng Wang, Nannan Shi, Qianzi Che, and Bin Liu. 2024. Icka: an instruction construction and knowledge alignment framework for multimodal named entity recognition. *Expert Systems with Applications*, 255:124867.

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

Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021b. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–842.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. 32.

Xin Zhang, Jingling Yuan, Lin Li, and Jianquan Liu. 2023. Reducing the bias of visual objects in multimodal named entity recognition. In *Proceedings of the Sixteenth ACM international conference on web search and data mining*, pages 958–966.

Jiachen Zhao, Shizhou Huang, and Xin Lin. 2025. A graph interaction framework on relevance for multimodal named entity recognition with multiple images. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1237–1246.

Zihao Zheng, Zihan Zhang, Zexin Wang, Ruiji Fu, Ming Liu, Zhongyuan Wang, and Bing Qin. 2024. Decompose, prioritize, and eliminate: Dynamically integrating diverse representations for multimodal named entity recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4498–4508.

Appendix

A Prompts

Prompts for Image Captioning:

> You are an AI assistant that excels at image captioning. <Img> #I </Img> describe this picture in detail.

Prompts for generating relevant entity objects with their attributes:

> Given a sentence and the description of its corresponding image. Text content and image content may be relevant or irrelevant. Output the entities along with their attribute on the picture that may be relevant to entities appearing in the text. Do not output entities in the image description that are not relevant to the text.
>
> Note: Output common objects and group them into general categories that are not duplicated, merging essentially similar entities. Avoid extracting abstract or non-specific entities.
>
> Examples:
> **Text**: Some used to say Steph Curry was "too small" and "too frail" to play in the NBA. They were wrong.
> **Caption**: "The image shows a basketball player wearing a white Golden State Warriors jersey with blue and yellow accents. The jersey prominently displays the number 30 and the team logo, a bridge enclosed in a circle. The player is clenching their fists in front of their chest, showcasing muscular arms. The background is a blurred arena filled with fans and ambient lights, hinting at a game or event in progress. The jersey also features an NBA logo near the left shoulder. The player's confident stance suggests they may have just made a significant play".
> **Outputs**: ["basketball player celebrating a basket", "'Golden State Warriors' logo on the jersey", "NBA logo on the jersey"]
>
> **Text**: #T
> **Caption**: #C
> **Outputs**: #O

In our prompts template, #I is the visual embeddings of the given image; #T is the original text; #C is the caption of the image; #O is the outputs of LLMs.

B Distribution of Two Twitter Datasets

The detailed data distribution of two twitter datasets is as follows:

Entity Types	Twitter-2015			Twitter-2017		
	Train	Dev	Test	Train	Dev	Test
PER	2217	552	1816	2943	626	621
LOC	2091	522	1697	731	173	178
ORG	928	247	839	1674	375	395
MISC	940	225	726	701	150	157
Total entities	6176	1546	5078	6049	1324	1351
Total tweets	4000	1000	3257	3373	723	723

Table 3: Distribution of Twitter-2015 and Twitter-2017 datasets.

C Ablation Study Details

Our ablation study systematically evaluates key components through the following experimental configurations:

- '**w/o** O-T' removes the entity candidates, feeding only the original text into the text encoder to get T_{origin} for subsequent processing.
- '**w/o** Wiki' removes retrieved knowledge from Wikipedia, feeding the original text with the pure entity candidates (without knowledge) into the encoder.
- '**w/o** LLMs' removes the entities and their attributes output from LLMs. As an alternative, we employ an object detection module to randomly sample five objects on images for encoding as entity object-level visual features.
- '**w/o** VAE' removes the heterogeneous text-guided features generated from VAE. As a result, just v_{glob} and v_{obj} are fed into the attention mechanism to fuse.
- **w/o** O-V replaces optimized visual features V with the global visual features v_{glob}, which will be fed into the CFF module.
- '**w/o** O-T-V' removes both the optimized textual and visual features, feeding only T_{origin} and v_{glob} into the CFF module.