Abstract

MultiModal Summarization (MMS) aims to generate a concise summary based on multimodal data like texts and images and has wide application in multimodal fields. Previous works mainly focus on the coarse-level textual and visual features in which the overall features of the image interact with the whole sentence. However, the entities of the input text and the objects of the image may be underutilized, limiting the performance of current MMS models. In this paper, we propose a novel Visual Enhanced Entity-Level Interaction Network (VE-ELIN) to address the problem of underutilization of multimodal inputs at a fine-grained level in two ways. We first design a cross-modal entity interaction module to better fuse the entity information in text and the object information in vision. Then, we design an object-guided visual enhancement module to fully extract the visual features and enhance the focus of the image on the object area. We evaluate VE-ELIN on two MMS datasets and propose new metrics to measure the factual consistency of entities in the output. Finally, experimental results demonstrate that VE-ELIN is effective and outperforms previous methods under both traditional metrics and ours.

1 Introduction

MultiModal Summarization (MMS) takes multimodal data like texts and images as input and aims to generate a concise summarization as output. This task has attracted much attention in the research community (Li et al., 2019, 2018b; Zhu et al., 2018) because it can be widely used in various real-world applications, such as social media (Zhang et al., 2022a), meeting (Zhong et al., 2021), and e-commerce products (Li et al., 2020a).

Recent studies primarily concentrate on the cross-modal interaction and filtering of visual features, which have achieved promising performances. For instance, Yu et al. (2021) explores various ways of image-text fusion to utilize multimodal information based on the application of generative Pre-trained Language Models (PLMs) to the task. Zhang et al. (2022b) adopts knowledge distillation from the vision-language pre-trained model to improve image selection. Liang et al. (2023) designs a target-oriented contrastive objective to discard needless visual information. Despite their effectiveness, current methods mainly focus on the coarse-level rather than fine-grained visual and textual features, which conduct interactions between the global image and sentence semantics. This might lead to an insufficient utilization of crucial local information. As shown in Figure 1, there are three fine-grained entities "Nicole Cooke", "Gold", and "Beijing Olympics" in the input text, and three object regions in the image corresponding to them while previous methods are not able to extract the fine-grained information adequately.

Thus, we consider utilizing the inherent entity information in the text and object information in the image so that the output summary maintains key entities with high coherence. In this paper, we propose a novel Visual Enhanced Entity-Level Interaction Network (VE-ELIN) for Multimodal Summarization. The proposed VE-ELIN addresses...
the problem of incomplete generation of entity information in two ways. Firstly, we design the cross-modal Entity Interaction (EI) module which can better fuse the entity information in text and the object information in vision and provide richer multimodal representation. In particular, the EI module includes three levels of features, namely sentence, entity, and object level. We encode the input text using a textual encoder to obtain sentence-level features and use a pre-trained Named Entity Recognition model (Yan et al., 2021) to get entity-level features. Moreover, we use the image object detection model (Carion et al., 2020) to capture the objects in the image and encode them to obtain the object-level features. Secondly, to further distill features from vision information, we apply CLIP (Radford et al., 2021) and integrate it into our object-guided Visual Enhancement (VE) module. The VE module can fully extract the visual features and enhance the focus of the image on the object area to better inject visual information into the multimodal decoder.

In addition to conventional evaluation methods, we introduce novel metrics to measure the factual consistency of entities in the output summarization. Specifically, we count the number of entities in the output and compare it with the entities in the target summary. Then, we compute the proportion of entities named EntityScore and the similarity between entities named SimilarScore.

We evaluate VE-ELIN on two MMS datasets, which have different text lengths and input image numbers. The experimental results demonstrate that VE-ELIN is effective and outperforms previous methods under both traditional metrics and ours.

In summary, our contributions are as follows:

- To the best of our knowledge, we are the first to identify the significance of fine-grained entity information for the multimodal summarization task.
- We propose a unified Visual Enhanced Entity-Level Interaction Network (VE-ELIN) to generate high-quality summaries while capturing key entity information in the original text.
- We propose two new metrics EntityScore and SimilarScore to further assess the factual consistency of entities in the output. The experimental results demonstrate the effectiveness of our proposed VE-ELIN.

2 Related Work

2.1 Multimodal Interaction

Object detection aims to predict a set of bounding boxes and corresponding category labels for the targeted objects in an image, which is a fundamental task in computer vision. Named Entity Recognition aims to identify the named entities in the text and can be widely used in information retrieval (Brandsen et al., 2022), and knowledge graphs (Zamini et al., 2022). Due to the rapid development of social media platforms such as Twitter, Multimodal Named Entity Recognition (MNER) (Zhao et al., 2022) has attracted increasing attention. Given image-text pairs, MNER aims to recognize the named entities in the text and classify the corresponding types. In the study of MNER, aligning the instance information in images with entities in text is an intuitive idea. However, in the field of multimodal summarization, there has been limited research on fine-grained interaction between visual and textual modalities.

2.2 Multimodal Summarization

Text summarization aims to extract important information from text and generate a concise summary. With the increasing of multimodal data on the internet, researchers have shown a growing interest in multimodal summarization. Different from traditional text summarization, multimodal summarization aims to generate summaries based on data from various modalities, e.g., video, image, audio, and text.

Existing multimodal summarization tasks contain sports summarization (Tjondronegoro et al., 2011), movies summarization (Evangelopoulos et al., 2013), video summarization (Sanabria et al., 2018), meeting summarization (Erol et al., 2003; Li et al., 2019), multimodal sentence summarization (Li et al., 2018b), multimodal summarization with multimodal output (Zhu et al., 2018), e-commerce products summarization (Li et al., 2020a) and so on. Previous studies on multimodal summarization tackle the tasks from different aspects. Palaskar et al. (2019) explore the hierarchy attention between the textual article and visual features. Consequent studies utilize fusion forget gate (Liu et al., 2020), visual selective gates (Li et al., 2020b), and contribution network (Xiao et al., 2023), directing the attention of models towards the most salient parts in the visual features for summarization.
3 Methodology

In this section, we introduce the overview of our framework. We first present the brief task formulation and describe the method overview. Then, we detail our proposed module and introduce the training and generation process.

3.1 Task Formulation

In this paper, we focus on the multimodal summarization task, involving a dataset comprising \( n \) triplets \((t_i, v_i, s_i)\), where \( t_i \) represents the \( i \)-th text input, \( v_i \) represents the \( i \)-th image input, and the MMS model is tasked with generating a summary \( s_i \) based on both \( t_i \) and \( v_i \).

3.2 Method Overview

We use VG-GPLM (Yu et al., 2021) as the backbone, which is built upon generative pre-trained language models (e.g., BART), and injects visual features on the encoder side. As shown in Figure 2, the VE-ELIN takes text and image as inputs and generates a summary as output. The multimodal encoder part of VE-ELIN consists of an EI module that can better fuse the entity features in textual and visual information and a VE module that can fully extract the visual features and enhance the focus of the image on the object area. Then, in the multimodal decoder, we fuse the features of different modalities from EI module and VE module and use it as extra input to the decoder.

3.3 Multimodal Encoder

3.3.1 Object-guided Visual Enhancement

Given an image, we first utilize the visual encoder of CLIP (Radford et al., 2021) to extract visual local grid features. CLIP is a dual-stream vision-language pre-trained model that has undergone pre-training with a contrastive loss using 400 million image-text pairs. This model comprises a Transformer (Vaswani et al., 2017) text encoder and an image encoder which could be either Vision Transformer (ViT) (Dosovitskiy et al., 2020) or Residual Convolutional Neural Network (ResNet) (He et al., 2016). In this paper, we apply the ViT image encoder of CLIP and obtain visual features \( V \in \mathbb{R}^{s_v \times d_v} \), where \( s_v \) is the patch numbers and \( d_v \) is the hidden dimension of image features.

Previous studies have indicated that different regions of visual features contribute unequally to summary generation (Li et al., 2020b; Liu et al., 2020; Xiao et al., 2023). For instance, given the input sentence and image, the target summary is “Britain’s Cooke wins Olympic gold in women’s cycling road race.”, as shown in Figure 1. In the image, the People, Gold Medal, and Olympic Logo components are more relevant to the target summary, while the features corresponding to the rest of the sections are less important. Thus we design a simple feature filter to enhance the focus on the image objects and better utilization of input visual features. In practice, we follow Carion et al. (2020) to detect the objects in the image using ResNet-101 as a backbone. As shown in Figure 2(b), two features are obtained after going through DETR, one is the visual features of each object marked with the bounding box: \( \text{ObjectFeatures} = V_o \in \mathbb{R}^{n \times 1 \times d_v} \), where \( n \) is the object numbers. For instance, there are three objects in the image, then \( n=3 \). In addition, we set the maximum number of objects to 64.

The other is the attention score matrix of the whole image: \( \text{AttentionScore} = A_{i,j} \in \mathbb{R}^{m \times m} \), where \( a_{i,j} \in [0,1], i,j \in [0,m] \) and are the indexes of the matrix, the closer the value is to the object area the closer it is to 1. We design a simple features filter through the attention score matrix, in practice, we transform \( A_{i,j} \) through a linear layer to the same dimension as the image features, and then fuse it with the image features.

\[
\hat{A}_{i,j} = \text{Linear}(A_{i,j}) \quad (1)
\]

\[
V_{\text{filtered}} = V \ast \hat{A}_{i,j} \quad (2)
\]

where \( V_{\text{filtered}} \in \mathbb{R}^{s_v \times d_v} \). The filtered visual features are represented in Figure 2 as visual-enhanced features.

3.3.2 Cross-modal Entity Interaction

We design this module to capture entity-related textual and visual information through three features: sentence-level features, entity-level features, and object-level features. Finally, get the entity-related feature as output and add it to the text-vision fusion in Section 3.4.

Sentence-level Features. At the entry of the framework, the input text is first tokenized and converted to a sequence of token embeddings \( X_t \in \mathbb{R}^{N \times d_t} \), and the positional encodings \( E_{pe} \in \mathbb{R}^{N \times d_t} \) are added to it, in which \( N \) is the sequence length and \( d_t \) is the textural dimension.

\[
Z_0^{nc} = X_t + E_{pe} \quad (3)
\]

As illustrated in Figure 2(a), the encoder is composed of a stack of \( L \) encoder layers,
each containing two sub-layers: Multi-head Self-Attention (MSA) and Feed-Forward Network (FFN). After each sub-layer, there is a residual connection (Wang et al., 2019) followed by a layer normalization (LN). We obtain the sentence-level features $T_s$ through the encoder.

$$Z'_l = LN(MSA(Z'_{l-1}^{enc}) + Z'_{l-1}^{enc})$$ (4)
$$T_s = LN(FFN(Z'_l) + Z'_l)$$ (5)

where $T_s \in \mathbb{R}^{N \times d_t}$.

Entity-level features. Following Yan et al. (2021), we use the Seq2Seq model with the pointer mechanism to generate the entity index sequences, which are then mapped to sentence-level features to obtain entity-level features. This part includes two components.

1. BART Encoder encodes the input sentence $X = t_i$ into vectors $H^e$

$$H^e = Encoder(X)$$ (6)

where $H^e \in \mathbb{R}^{N \times d_t}$, and $d_t$ is the hidden dimension.

2. BART Decoder is to get the index probability distribution for each step $P_t = P(\hat{y}_t \mid X, Y_{<t})$. However, since $Y_{<t}$ contains the pointer and tag index, it cannot be directly inputted to the Decoder. We use the Index2Token conversion to convert indexes into tokens:

$$\hat{y}_t = \begin{cases} X_{y_t}, & \text{if } y_t \leq n, \\ G_{y_{t-n}}, & \text{if } y_t > n \end{cases}$$ (7)

After converting each $\hat{y}_t$ this way, we can get the last hidden state $\hat{h}^d_{t} \in \mathbb{R}^{d_t}$ with $\hat{Y}_{<t} = [\hat{y}_1, \ldots, \hat{y}_{t-1}]$ as follows

$$\hat{h}^d_{t} = Decoder(H^e; \hat{Y}_{<t})$$ (8)

Then, we can use the following equations to achieve the index probability distribution $P_t$

$$E^e = TokenEmbed(X)$$ (9)
$$\hat{H}^e = MLP(H^e)$$ (10)
$$\hat{H}^e = \alpha \times \hat{H}^e + (1 - \alpha) \times E^e$$ (11)
$$G^{dl} = TokenEmbed(G)$$ (12)
$$P_t = \text{Softmax}([\hat{H}^e \otimes \hat{h}^d_{t}; G^{dl} \otimes \hat{h}^d_{t}])$$ (13)

where TokenEmbed is the embeddings shared between the Encoder and Decoder; $E^e, \hat{H}^e, G^{dl} \in \mathbb{R}^{l \times d_t}$; $\alpha \in [0, 1]$ is a hyper-parameter; $\otimes$ means concatenation in the first dimension; $\otimes$ means the dot product. Finally, we map the index $P_t$ to the sentence-level features Eq.(5) to get entity-level features.

$$T_e = \text{Map}(P, T_s)$$ (14)

During the training phase, we use the negative log-likelihood loss and the teacher forcing method. During the inference, we use an autoregressive manner to generate the target sequence. In the overall framework of our model, the NER part is pre-trained in advance, and in the overall model training, it is used for inference.
Cross-modal Entity Interaction. Firstly, we employ multi-head self-attention on the interaction features to exploit contexts of the same modality.

\[ D_m = \text{MultiHeadAttn}(H_m, H_m, H_m) \quad (15) \]

\( H_m \) is the interaction features, where \( m \in \{ T_e, V_o, T_s \} \). Then, we interact entity features with object features via a gated cross-attention module.

\[ R_e = \text{MultiHeadAttn}(H_{T_e}, D_{V_o}, D_{V_o}) \quad (16) \]

\[ \alpha_e = \text{Sigmoid}(W_c1 R_e + W_c2 H_{T_e}) \quad (17) \]

\[ M_e = \alpha_e \cdot R_e + (1 - \alpha_e) \cdot H_{T_e} \quad (18) \]

where \( M_e \) is object-aware entity representations. Similarly, we obtain entity-aware object representations \( M_o \). After that, we fuse visual information from \( M_e \) to the sentence-level features \( T_s \).

\[ \alpha_s = \text{Sigmoid}(W_{s1} M_e + W_{s2} H_{T_e}) \quad (19) \]

\[ M_s = \alpha_s \cdot R_s + (1 - \alpha_s) \cdot H_{T_s} \quad (20) \]

Finally, we add \( M_s \) and \( M_o \) to get the output entity-related features \( Z_{er} \) of the cross-modal entity interaction module.

\[ Z_{er} = M_s + M_o \quad (21) \]

3.4 Multimodal Decoder

We inject visual information through the vision-guided multi-head attention mechanism. The query \( Q \) is from the obtained filtered visual features \( V_{filtered} \) in Section 3.3.1, and the key \( K \) and value \( V \) are from the obtained sentence-level features \( T_s \) in Section 3.3.2. Then, we apply a cross-modal multi-head attention (CMA) to get the text queried visual features \( Z_v \). Finally, we add the entity-related features \( Z_{er} \) and \( Z_v \) to get the text-visual fusion features \( Z_k \).

\[ Z'_v = \text{CMA}(V_{filtered}; T_s, T_s) \quad (22) \]

\[ Z_v = \text{Dropout}(\text{concat}(T_s, Z'_v)) \quad (23) \]

\[ Z_k = \text{Linear}(Z_{er} + Z_v) \quad (24) \]

The text-visual fusion features will be input into the decoder of BART to generate the corresponding summary.

\[ \log p_\theta(y) = \sum_{i=1}^{n} \log p_\theta(y_i | Z_k, y_1, \ldots, y_{i-1}) \quad (25) \]

where \( y_i \) is the \( i \)th generated token on the decoder side. For the text-visual fusion process above, the training loss is the commonly used cross-entropy loss function \( \mathcal{L}_{ce} \).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size</th>
<th>S.Len (M/A/M)</th>
<th>T.Len (M/A/M)</th>
<th>L.Num (M/A/M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMSS</td>
<td>train</td>
<td>62,000</td>
<td>11/21/69/63</td>
<td>27/72/25/5</td>
</tr>
<tr>
<td></td>
<td>dev</td>
<td>2,000</td>
<td>11/24/35/47</td>
<td>37/68/17</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>2,000</td>
<td>11/22/97/51</td>
<td>37/67/24/4</td>
</tr>
<tr>
<td>average</td>
<td></td>
<td>-</td>
<td>23.00</td>
<td>7.69</td>
</tr>
<tr>
<td>MM-Sum-En</td>
<td>train</td>
<td>305,828</td>
<td>7/461.32/19/282</td>
<td>1/22/12/172</td>
</tr>
<tr>
<td></td>
<td>dev</td>
<td>11,437</td>
<td>55/440.586/1,686</td>
<td>82/15/41</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>11,460</td>
<td>61/483/114/667</td>
<td>7/31/23/42</td>
</tr>
<tr>
<td>average</td>
<td></td>
<td>-</td>
<td>446.84</td>
<td>21.50</td>
</tr>
</tbody>
</table>

Table 1: The statistics of MMSS and MM-Sum-En datasets. "S.Len" and "T.Len" refer to the number of words in the source text and the target summary. "L.Num" denotes the number of images corresponding to each text. "M/A/M" means Minimum/Average/Maximum.

4 Experiments

4.1 Dataset

We evaluate our method on the MultiModal Sentence summarization (MMSS) (Li et al., 2018a) and Multilingual Multimodal abstractive Summarization for English (MM-Sum-En) dataset on mid-high-resource scenario (Liang et al., 2022). The MMSS dataset contains 62,000 samples in the training set, 2,000 in the validation set, and 2,000 in the test set, and each sample is a triplet of \( \langle \text{sentence}, \text{image}, \text{summary} \rangle \). The MM-Sum dataset for English contains 326,725 samples and 867,817 images in total which crawled from the BBC News, where each sample is constructed of a news article and some images and presented as \( \langle \text{article}, \text{images}, \text{summary} \rangle \). We count some basic information about the dataset, which is shown in Table 1.

4.2 Experimental Settings

For image processing, we utilize the vision encoder of the "ViT-B/32" version of CLIP (Radford et al., 2021), the image patches are \( 7 \times 7 \) and the dimension of output visual features is 768. We apply the "Resnet-101" version of DETR (Carion et al., 2020) for object detection with \( \text{threshold} = 0.95 \). For textual generative pre-trained language models, we adopt BART-base (Lewis et al., 2020) as our textual encoder and decoder, where the textual dimension is also 768. We train the Named Entity Recognition (NER) model proposed by Yan et al. (2021) as a tool for extracting text entities. During training, for MMSS, we set the dropout to 0.1, the batch size is 120, the maximum training epochs is 50, and the beam size is 5. The learning rate is \( 2e-5 \) and the 5
times the learning rate for vision-related modules of the MMS model and the loss function is cross entropy. We leverage AdamW (Loshchilov and Hutter, 2018) as optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a weight decay of 1e-2. Additionally, we apply a scheduler to decay the learning rate to 95% of the current one after every 10 epochs. The maximum input length is 64 and the maximum output length is 32. For the MM-Sum-En dataset, the parameters are the same as in MMSS except that the maximum input length is 1024, the maximum output length is 256, the batch size is 10, and the maximum training epochs is 20. We save our best model checkpoint according to the best ROUGE-2 score on the validation set. All models are trained and tested on a single NVIDIA 3090Ti GPU.

### 4.3 Compared Methods

Our base model is VG-BART (Yu et al., 2021), which utilizes PLMs as the backbone and injects visual features into the encoder layer through dot production.

We also compare our method with other works with these two datasets. For MMS dataset: 1) **Lead**. The initial eight words are employed as the summary. 2) **Compress** (Clarke and Lapata, 2008). A methodology centered on sentence compression, utilizing syntactic structure as a basis. 3) **ABS** (Rush et al., 2015). An attentive CNN encoder in conjunction with a neural network language model decoder to proficiently summarize sentences. 4) **SEASS** (Zhou et al., 2017). A summarization framework distinguished by its incorporation of textual selective encoding. 5) **Multi-Source** (Libovický and Helcl, 2017). This method integrates multiple source modalities utilizing hierarchical attention mechanisms, addressing challenges in multimodal machine translation. 6) **Doubly-Attention** (Calixto et al., 2017). This approach leverages two distinct attention mechanisms to incorporate visual features, narrowing the gap between image and translation. 7) **MAtt** (Li et al., 2018b). This approach proposes modality attention and image-filtering techniques tailored for multimodal summarization. 8) **MSE** (Li et al., 2020a). This approach advocates for the application of visual selective gates in multimodal summarization. 9) **CFSum** (Xiao et al., 2023). This approach proposes a contribution network that selects more important parts of images for multimodal summarization, which is a strong baseline.

For MM-Sum-En dataset: 1) **mT5** (Xue et al., 2020b). A methodology centered on sentence compression, utilizing syntactic structure as a basis. 2) **Doubly-Attention** (Calixto et al., 2017). This approach leverages two distinct attention mechanisms to incorporate visual features, narrowing the gap between image and translation. 3) **MAtt** (Li et al., 2018b). This approach proposes modality attention and image-filtering techniques tailored for multimodal summarization. 4) **MSE** (Li et al., 2020a). This approach advocates for the application of visual selective gates in multimodal summarization. 5) **CFSum** (Xiao et al., 2023). This approach proposes a contribution network that selects more important parts of images for multimodal summarization, which is a strong baseline.
This approach is a multilingual language model pre-trained on a large dataset of 101 languages that is a text-only baseline. 2) VG-mTS (Liang et al., 2022). This approach implements the vision-guided multi-head attention fusion method to inject visual features into the mTS model. 3) SOV-MAS (Liang et al., 2022). This approach applies two summary-oriented visual modeling tasks to enhance the MMS model based on the pre-trained language models (e.g., BART).

For all the above models trained on MM-Sum-En, we follow the same monolingual experimental settings in the mid-high-resource scenario, as employed by Liang et al. (2022).

### 4.4 Main Results

Following Xiao et al. (2023) and Liang et al. (2022), we report our experiment results with 6 automatic metrics: ROUGE-1, ROUGE-2, ROUGE-L (Lin, 2005), BLEU (Papineni et al., 2002), MOVER (Zhao et al., 2019) and BERTScore (Zhang et al., 2019).

Overall, compared with previous works on MMSS as shown in Table 2, our proposed method demonstrates significant improvements across all 6 reported evaluation metrics. Compared with the strong baseline CFSum (Xiao et al., 2023), our method achieves 6.64 higher points on ROUGE-1 than it, demonstrating the effectiveness of our proposed method. Comparing VG-BART with those that design gate-based pre-filters or other networks based on the vision-language pre-trained encoder (e.g., MSE (Li et al., 2020b) and CF-Sum (Xiao et al., 2023)), we find that our base model, which straightforwardly employs a PLM and integrates visual features, proves to be more effective in enhancing model performance. Furthermore, VE-ELIN outperforms the base model VG-BART, showing that the image processing and visual enhancement we use in the model and the added entity-level features complement each other and significantly improve the quality of the output summarization. The experimental effects of each module are specified in the ablation study 5.1. In the MM-Sum-En dataset, we observe the same results as in MMSS dataset, the performance of our proposed method is improved compared to others.

As shown in Table 1, the average length of input sentences in MMSS is 23, and the average number of input images is 1. In contrast, the length and number of MM-Sum-En are 446.84 and 2.23. Also, MMSS is from the headlines of article pairs

<table>
<thead>
<tr>
<th>Model</th>
<th>R-1</th>
<th>R-2</th>
<th>R-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMSS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours(VE-ELIN)</td>
<td>54.20</td>
<td>31.24</td>
<td>51.47</td>
</tr>
<tr>
<td>- w/o $M_{VE}$ &amp; $M_{EI}$</td>
<td>52.02</td>
<td>29.67</td>
<td>49.45</td>
</tr>
<tr>
<td>- w/o $M_{VE}$ &amp; $M_{EI}$</td>
<td>53.60</td>
<td>31.10</td>
<td>50.80</td>
</tr>
<tr>
<td>- w/o $M_{VE}$</td>
<td>53.42</td>
<td>31.03</td>
<td>51.02</td>
</tr>
<tr>
<td>- w/o $M_{EI}$</td>
<td>53.30</td>
<td>30.97</td>
<td>50.85</td>
</tr>
</tbody>
</table>

Table 3: Ablation study on two datasets, the top row of each model shows the experimental results from the MM dataset and the bottom row shows the results from the MM-Sum dataset. R-1/L denotes ROUGE-1/L, "$M_{VE}$" denotes visual enhancement module, "$M_{EI}$" denotes entity interaction module, and "$V_f$" denotes visual features.

from Gigaword (Graff and Cieri, 2003; Napoles et al., 2012), and MM-Sum-En is sourced from BBC website 1. This indicates that there is a huge difference between the two MMS datasets. Our method still generates high-quality summaries, further demonstrating the robustness and effectiveness of our proposed VE-ELIN.

### 5 Analysis

#### 5.1 Ablation Study

We conduct ablation studies on both MMSS dataset and MM-Sum-En dataset to prove the effectiveness of the different components of our model. The results are shown in Table 3. We have the following conclusions:

The absence of visual features means that it is a text-only model based on pre-trained language models (PLMs) like BART. It shows a decrease in performance across all ROUGE metrics, demonstrating the incorporation of visual information within the MMS model yields noticeable enhancements in performance.

Without the inclusion of the visual enhancement module and entity interaction module, we find a performance degradation of about 1%, this verifies the effectiveness of our proposed modules.

As for the model without the visual enhancement module compared with the previous methods, we find an improvement in the metrics, which shows

1https://www.bbc.com/
statistical results indicate a significant improvement in the number of entities recognized by our approach. Moreover, we concatenate the entities in the model output summary into one sentence \( \hat{X} = \langle x_1, x_2, ..., x_k \rangle \) and the entities in the target summary into another sentence \( \hat{X} = \langle \hat{x}_1, \hat{x}_2, ..., \hat{x}_l \rangle \). Following Zhang et al. (2019), the SimilarScore is then used to calculate the similarity of the two sentences.

\[
\text{SimilarScore} = \text{BERTScore}(X, \hat{X}) \tag{27}
\]

The computational results demonstrate that our proposed method indeed improves the number and quality of entities in the output summarization, thus proving the effectiveness of our model.

### 6 Conclusion

In this paper, we propose a novel framework VE-ELIN for multimodal summarization to alleviate the incomplete generation of entity information in summary. We design a cross-modal entity interaction module to better utilize the entity features in texts and images, and an object-guided visual enhancement module to enhance the focus on the objects while taking full advantage of useful image information. To further evaluate the factual consistency of entities in the output summary, we also propose two new metrics named EntityScore and SimilarScore. Experimental results on two different types of datasets demonstrate that our method is effective and outperforms previous methods under both traditional evaluation metrics and our proposed new metrics.
Limitations

Our approach is limited by the underlying performance of the generative pre-trained language model. In addition, the accuracy of the object detection model DETR and named entity recognition model also limit our performance.

Ethics Statement

We affirm that our work here does not deepen the biases already inherent in the models and the datasets we used are open-sourced. Thus we expect no ethical concerns associated with this research.

References


