Visual Enhanced Entity-Level Interaction Network for Multimodal Summarization

Anonymous ACL submission

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

Multi**M**odal **S**ummarization (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.

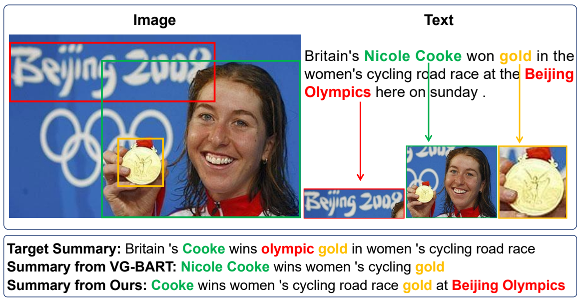

Figure 1: Illustration of multimodal summarization task. The bottom part is the target summary, a summary from the previous method, and ours. The previous method can not adequately leverage fine-grained entity information.

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 $\langle t_i, v_i, s_i \rangle$, 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: $ObjectFeatures$=$V_o \in \mathbb{R}^{n \times 1 \times d_v}$, 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: $AttentionScore$=$A_{i,j}$=$(a_{i,j}) \in \mathbb{R}^{m \times m}$, 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.

$$\hat{A}_{i,j} = \text{Linear}(A_{i,j}) \tag{1}$$

$$V_{filtered} = V * \hat{A}_{i,j} \tag{2}$$

where $V_{filtered} \in \mathbb{R}^{s_v \times d_v}$. 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 $X_t \in \mathbf{R}^{N \times d_t}$, 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 textual dimension.

$$Z_0^{enc} = X_t + E_{pe} \tag{3}$$

As illustrated in Figure 2(a), the encoder is composed of a stack of L encoder layers,

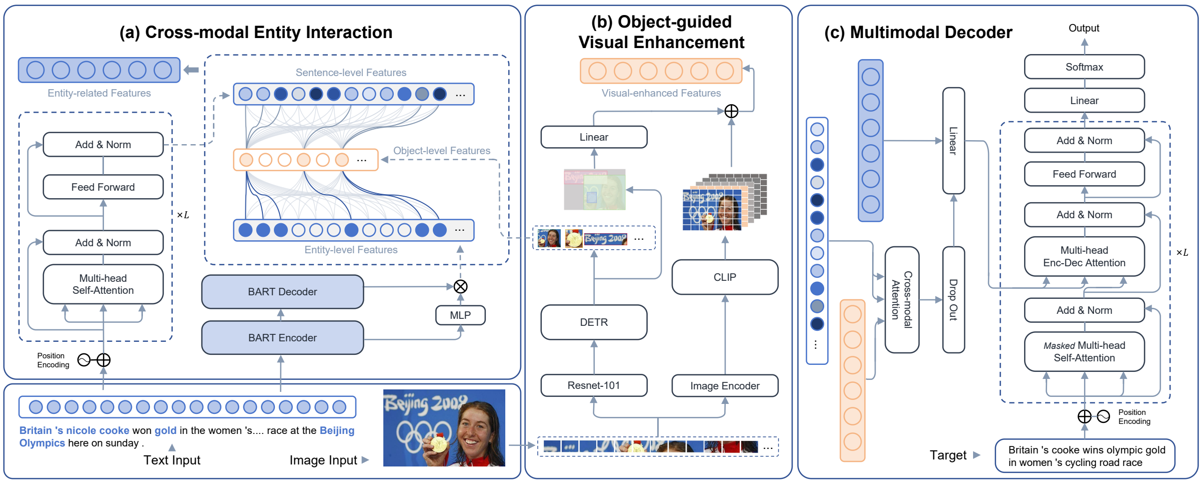

Figure 2: The overview of our model. Given input text and image, our model generates summaries as output through three modules: the cross-modal entity interaction module, object-guided visual enhancement module, and multimodal decoder.

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.

$$Z_l' = \mathrm{LN}(\mathrm{MSA}(Z_{l-1}^{enc}) + Z_{l-1}^{enc}) \tag{4}$$

$$T_s = \mathrm{LN}(\mathrm{FFN}(Z_l') + Z_l') \tag{5}$$

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 $\mathbf{H}^e$

$$\mathbf{H}^e = \mathrm{Encoder}(X) \tag{6}$$

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

(2) BART Decoder is to get the index probability distribution for each step $P_t = P\left(y_t \mid X, Y_{<t}\right)$. 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:

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

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

$$\mathbf{h}_t^{d_t} = \mathrm{Decoder}(\mathbf{H}^e; \hat{Y}_{<t}) \tag{8}$$

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

$$\mathbf{E}^e = \mathrm{TokenEmbed}(X) \tag{9}$$

$$\hat{\mathbf{H}}^e = \mathrm{MLP}(\mathbf{H}^e) \tag{10}$$

$$\bar{\mathbf{H}}^e = \alpha \times \hat{\mathbf{H}}^e + (1 - \alpha) \times \mathbf{E}^e \tag{11}$$

$$\mathbf{G}^{d_t} = \mathrm{TokenEmbed}(G) \tag{12}$$

$$P_t = \mathrm{Softmax}([\bar{\mathbf{H}}^e \otimes \mathbf{h}_t^{d_t}; \mathbf{G}^{d_t} \otimes \mathbf{h}_t^{d_t}]) \tag{13}$$

where TokenEmbed is the embeddings shared between the Encoder and Decoder; $\mathbf{E}^e$, $\hat{\mathbf{H}}^e$, $\bar{\mathbf{H}}^e \in \mathbb{R}^{n \times d_t}$; $\alpha \in [0, 1]$ is a hyper-parameter; $\mathbf{G}^{d_t} \in \mathbb{R}^{l \times d_t}$; $[\cdot; \cdot]$ 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.

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

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.

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

$$\alpha_e = \text{Sigmoid}(W_{e1} R_e + W_{e2} H_{T_e}) \quad (17)$$

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

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.

$$\alpha_s = \text{Sigmoid}(W_{s1} M_e + W_{s2} H_{T_s}) \quad (19)$$

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

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

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

where y_i is the ith generated token on the decoder side. For the text-vision fusion process above, the training loss is the commonly used cross-entropy loss function $\mathcal{L}_{ce}$.

Dataset	Size	S.Len (M/A/M)	T.Len (M/A/M)	I.Num (M/A/M)
MMSS				
train	62,000	11/21.68/63	2/7.72/25	1/1/1
dev	2,000	11/24.35/47	3/7.68/17	1/1/1
test	2,000	11/22.97/51	3/7.67/24	1/1/1
average	-	23.00	7.69	1
MM-Sum-En				
train	303,828	7/461.82/39,282	1/22.12/172	0/2.35/118
dev	11,437	55/440.59/1,686	8/21.15/41	0/2.24/30
test	11,460	61/438.11/1,667	7/21.23/42	0/2.09/26
average	-	446.84	21.50	2.23

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. "I.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 $\langle sentence, image, summary \rangle$. 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 $\langle article, images, summary \rangle$. 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 7×7 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 $threshold = 0.95$. 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

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore	MoverScore
MMSS						
$Lead^{\star\top}$	33.64	13.40	31.84	-	-	-
$Compress^{\star\top}$	31.56	11.02	28.87	-	-	-
$ABS^{\star\top}$	35.95	18.21	31.89	-	-	-
$SEASS^{\star\top}$	44.86	23.03	41.92	-	-	-
$Multi\text{-}Source^{\star}$	39.67	19.11	38.03	-	-	-
$Doubly\text{-}Attention^{\star}$	41.11	21.75	39.92	-	-	-
$MAtt^{\star}$	47.28	24.85	44.48	-	-	-
$MSE^{\star}$	45.63	23.68	42.97	-	-	-
$CFSum^{\star}$	47.86	25.64	44.64	48.83	86.98	32.36
$VG\text{-}BART$	52.02	29.67	49.45	57.94	91.86	47.36
$Ours\ (VE\text{-}ELIN)$	**54.20**	**31.24**	**51.47**	**60.16**	**92.22**	**49.15**
MM-Sum-En						
$mT5^{\wedge\top}$	36.99	15.18	29.64	-	-	-
$VG\text{-}mT5^{\wedge}$	37.17	14.88	29.41	-	-	-
$SOV\text{-}MAS^{\wedge}$	37.26	15.02	29.61	-	-	-
$VG\text{-}BART$	37.39	15.99	30.35	40.81	90.11	27.37
$Ours\ (VE\text{-}ELIN)$	**39.97**	**18.09**	**32.47**	**45.44**	**90.61**	**30.85**

Table 2: Experimental results on test set of multimodal sentence summarization (MMSS) dataset and test set of Multilingual Multimodal abstractive Summarization for English (MM-Sum-En) dataset. "$\star$" marks the experimental results reported by Xiao et al. (2023) and "$\wedge$" indicates that they were reported by Liang et al. (2022). "$\top$" denotes this method only leverages text modality data.

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., 2020). This approach is a multilingual language model pre-trained on a large dataset of 101 languages that is a text-only baseline. 2) **VG-mT5** (Liang et al., 2022). This approach implements the vision-guided multi-head attention fusion method to inject visual features into the mT5 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 CFSum (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 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.

Model	R-1	R-2	R-L
MMSS			
Ours(VE-ELIN)	**54.20**	**31.24**	**51.47**
- w/o $\mathcal{M}_{VE}$&$\mathcal{M}_{EI}$&V_f	52.02	29.67	49.45
- w/o $\mathcal{M}_{VE}$&$\mathcal{M}_{EI}$	53.60	31.10	50.80
- w/o $\mathcal{M}_{VE}$	53.42	31.03	51.02
- w/o $\mathcal{M}_{EI}$	53.30	30.97	50.85
MM-Sum-En			
Ours(VE-ELIN)	**39.97**	**18.09**	**32.47**
- w/o $\mathcal{M}_{VE}$&$\mathcal{M}_{EI}$&V_f	37.39	15.99	30.35
- w/o $\mathcal{M}_{VE}$&$\mathcal{M}_{EI}$	39.30	17.60	31.90
- w/o $\mathcal{M}_{VE}$	39.74	17.96	32.28
- w/o $\mathcal{M}_{EI}$	39.51	17.84	32.04

Table 3: Ablation study on two datasets, the top row of each model shows the experimental results from the MMS dataset and the bottom row shows the results from the MM-Sum dataset. R-1/2/L denotes ROUGE-1/2/L, "$\mathcal{M}_{VE}$" denotes visual enhancement module, "$\mathcal{M}_{EI}$" denotes entity interaction module, and "V_f" denotes visual features.

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

[1] https://www.bbc.com/

Dataset	Source	Target			VG-BART			Ours (VE-ELIN)		
	E.Num	E.Num	E.Score	S.Score	E.Num	E.Score	S.Score	E.Num	E.Score	S.Score
MMSS										
dev	3,013	1,422	100	100	616	48.80	91.53	703	**61.87**	**93.74**
test	3,117	1,429	100	100	620	58.47	93.35	641	**59.60**	**93.47**
average	3,065	1,425.5	100	100	618	53.64	92.44	672	**60.74**	**93.61**
MM-Sum-En										
dev	72,412	19,300	100	100	6,461	37.96	90.27	7,293	**43.28**	**91.31**
test	72,403	19,200	100	100	6,310	37.02	90.14	7,272	**43.01**	**91.20**
average	72,407.5	19,250	100	100	6,385.5	37.49	90.21	7,282.5	**43.15**	**91.26**

Table 4: The Entity evaluation metrics in the output summarization. "Source" refers to the input text of the datasets, and "Target" refers to the reference summary. "E.Num" denotes the number of entities in the text, "E.Score" refers to the EntityScore, which is the proposed evaluation metric, and the "S.Score" means SimlarScore metric, which is obtained by doing similarity calculations between the entities in the summaries generated by Target/VG-BART/Ours and the entities in the "Target" respectively.

that the image features filter does help to improve the quality of the output summaries. The results show that the visual enhancement module further improves the model performance, indicating that the objects in the images are beneficial to the visual modality information.

The model without entity interaction module makes relative contributions to the MMS model. We can see a certain growth of three ROUGE metrics compared with others in Section 4.4, showing that focusing on the object visual features of the image is effective. The results indicate that our entity interaction module improves the quality of the output summaries and has a large improvement on the model performance.

5.2 Entity Consistency

As shown in Table 4, we formulate some new metrics to assess the quality of output summarization. Specifically, we utilize the NER model trained with BART with an accuracy of 93.8% to count the number of entities in the output summarization generated by the proposed method and the baseline, which is represented in Table 4 by "E.Num". In the process of counting, if an entity in the generated summary is also among the entities in the corresponding target summary, the entity is recorded as a valid entity. Then, the ratio of the number of valid entities to the number of entities in the target summary is calculated and named EntityScore, which is expressed as "E.Score" in Table 4.

$$\text{EntityScore} = \frac{N_{generated}}{N_{target}} \tag{26}$$

where $N_{generated}$ and N_{target} is the entity numbers in generated summary and target summary. 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 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

Alex Brandsen, Suzan Verberne, Karsten Lambers, and Milco Wansleeben. 2022. Can bert dig it? named entity recognition for information retrieval in the archaeology domain. *Journal on Computing and Cultural Heritage (JOCCH)*, 15(3):1–18.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924, Vancouver, Canada. Association for Computational Linguistics.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

James Clarke and Mirella Lapata. 2008. Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research*, 31:399–429.

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*.

Berna Erol, D-S Lee, and Jonathan Hull. 2003. Multimodal summarization of meeting recordings. In *2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698)*, volume 3, pages III–25. IEEE.

Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. 2013. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568.

David Graff and C Cieri. 2003. English gigaword, linguistic data consortium.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020a. Aspect-aware multimodal summarization for chinese e-commerce products. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8188–8195.

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, and Chengqing Zong. 2018a. Multi-modal sentence summarization with modality attention and image filtering. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4152–4158. International Joint Conferences on Artificial Intelligence Organization.

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, Chengqing Zong, et al. 2018b. Multi-modal sentence summarization with modality attention and image filtering. In *IJCAI*, pages 4152–4158.

Haoran Li, Junnan Zhu, Jiajun Zhang, Xiaodong He, and Chengqing Zong. 2020b. Multimodal sentence summarization via multimodal selective encoding. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5655–5667.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196, Florence, Italy. Association for Computational Linguistics.

Yunlong Liang, Fandong Meng, Jiaan Wang, Jinan Xu, Yufeng Chen, and Jie Zhou. 2023. D2 tv: Dual knowledge distillation and target-oriented vision modeling for many-to-many multimodal summarization. *arXiv preprint arXiv:2305.12767*.

Yunlong Liang, Fandong Meng, Jinan Xu, Jiaan Wang, Yufeng Chen, and Jie Zhou. 2022. Summary-oriented vision modeling for multimodal abstractive summarization. *arXiv preprint arXiv:2212.07672*.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting* *of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Vancouver, Canada. Association for Computational Linguistics.

C Lin. 2005. Recall-oriented understudy for gisting evaluation (rouge). *Retrieved August*, 20:2005.

Nayu Liu, Xian Sun, Hongfeng Yu, Wenkai Zhang, and Guangluan Xu. 2020. Multistage fusion with forget gate for multimodal summarization in open-domain videos. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1834–1845, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated Gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, pages 95–100, Montréal, Canada. Association for Computational Linguistics.

Shruti Palaskar, Jindřich Libovický, Spandana Gella, and Florian Metze. 2019. Multimodal abstractive summarization for how2 videos. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6587–6596, Florence, Italy. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A large-scale dataset for multimodal language understanding. In *NeurIPS*.

Dian Tjondronegoro, Xiaohui Tao, Johannes Sasongko, and Cher Han Lau. 2011. Multi-modal summarization of key events and top players in sports tournament videos. In *2011 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 471–478. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. 2019. Learning deep transformer models for machine translation. *arXiv preprint arXiv:1906.01787*.

Min Xiao, Junnan Zhu, Haitao Lin, Yu Zhou, and Chengqing Zong. 2023. CFSum coarse-to-fine contribution network for multimodal summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8538–8553, Toronto, Canada. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A unified generative framework for various ner subtasks. *arXiv preprint arXiv:2106.01223*.

Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021. Vision guided generative pre-trained language models for multimodal abstractive summarization. *arXiv preprint arXiv:2109.02401*.

Mohamad Zamini, Hassan Reza, and Minou Rabiei. 2022. A review of knowledge graph completion. *Information*, 13(8):396.

Huakui Zhang, Cai Yi, Bingshan Zhu, Haopeng Ren, and Qing Li. 2022a. Multimodal topic modeling by exploring characteristics of short text social media. *IEEE Transactions on Multimedia*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zhengkun Zhang, Xiaojun Meng, Yasheng Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. 2022b. Unims: A unified framework for multimodal summarization with knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11757–11764.

Gang Zhao, Guanting Dong, Yidong Shi, Haolong Yan, Weiran Xu, and Si Li. 2022. Entity-level interaction via heterogeneous graph for multimodal named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6345–6350.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. *arXiv preprint arXiv:2104.05938*.

Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. 2017. Selective encoding for abstractive sentence summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1095–1104, Vancouver, Canada. Association for Computational Linguistics.

Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. Msmo: Multimodal summarization with multimodal output. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4154–4164.