Neural Computing and Applications (2024) 36:15653–15666
https://doi.org/10.1007/s00521-024-09908-3

ORIGINAL ARTICLE

Align vision-language semantics by multi-task learning for multi-modal summarization

Chenhao Cui[1] · Xinnian Liang[2] · Shuangzhi Wu[3] · Zhoujun Li[2]

Received: 18 July 2023 / Accepted: 23 April 2024 / Published online: 17 May 2024

Abstract

Most current multi-modal summarization methods follow a cascaded manner, where an off-the-shelf object detector is first used to extract visual features. After that, these visual features are fused with language representations for the decoder to generate the text summary. However, the cascaded way employs separate encoders for different modalities, which makes it hard to learn the joint vision and language representation. In addition, they also ignore the semantics alignment between paragraphs and images for multi-modal summarization tasks, which are crucial to a precise summary. To tackle these issues, in this paper, we propose ViL-Sum to jointly model paragraph-level *Vi*sion-*L*anguage Semantic Alignment and Multi-Modal *Sum*marization. Our ViL-Sum contains two components for better learning multi-modal semantics and aims to align them. The first one is a joint multi-modal encoder. The other one is two well-designed tasks for multi-task learning, including image reordering and image selection. Specifically, the joint multi-modal encoder converts images into visual embeddings and attaches them with text embedding as the input of the encoder. The reordering task guides the model to learn paragraph-level semantic alignment, and the selection task guides the model to select summary-related images in the final summary. Experimental results show that our proposed ViL-Sum outperforms current state-of-the-art methods on most automatic and manual evaluation metrics. In further analysis, we find that two well-designed tasks and a joint multi-modal encoder can effectively guide the model to learn reasonable paragraph-image and summary-image relations.

Keywords Multi-modal summarization · Semantic alignment · Multi-task learning

1 Introduction

The dramatic increase in multi-modal data (including text, image, audio, and video) on the Internet makes research on multi-modal summarization necessary. Multi-modal summarization aims to generate a condensed summary, which can cover salient information from one or more modalities inputs [1, 2]. Different from traditional pure text summary; Zhu et al. [3] points out that generated summaries with both text and images can effectively improve the quality of generated summary and increase the satisfaction of users. The information of different modalities is complementary and verifiable to each other. Utilizing multi-modal information helps the model better locate key content and generate better summaries. Intuitively, people can grasp key information easier from multiple modalities than only from the text. This task is defined as multi-modal summarization with multi-modal outputs (MSMO). Figure 1 shows an example of this task, which gets text and images

Chenhao Cui and Xinnian Liang contributed equally to this work.

✉ Zhoujun Li
lizj@buaa.edu.cn

Chenhao Cui
cuich@buaa.edu.cn

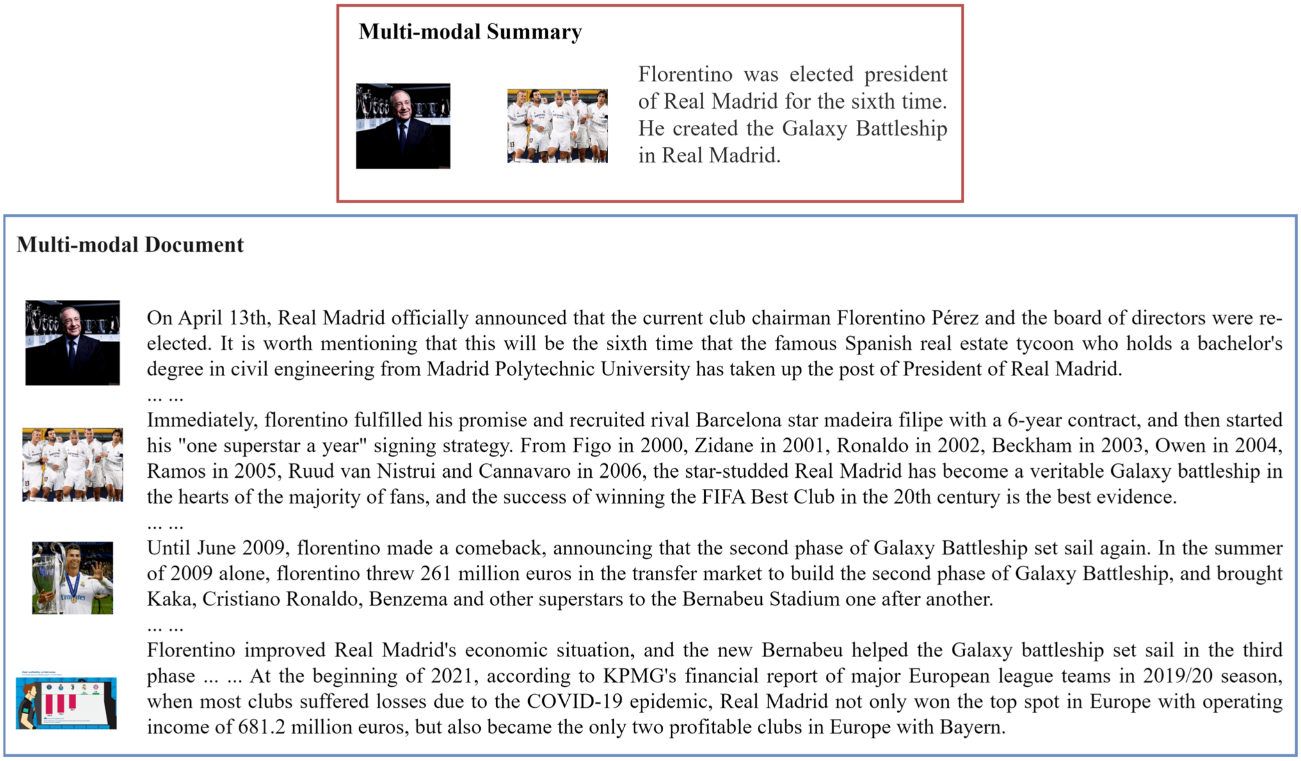
Xinnian Liang
xnliang@buaa.edu.cn

Shuangzhi Wu
wushuangzhi2010@163.com

[1] School of Cyber Science and Technology, Beihang University, Beijing 100191, China

[2] School of Computer Science and Engineering, Beihang University, Beijing 100191, China

[3] Cloud Xiaowei, Tencent, Beijing 100089, China



Multi-modal Summary

Florentino was elected president of Real Madrid for the sixth time. He created the Galaxy Battleship in Real Madrid.

Multi-modal Document

On April 13th, Real Madrid officially announced that the current club chairman Florentino Pérez and the board of directors were re-elected. It is worth mentioning that this will be the sixth time that the famous Spanish real estate tycoon who holds a bachelor's degree in civil engineering from Madrid Polytechnic University has taken up the post of President of Real Madrid.

... ...

Immediately, florentino fulfilled his promise and recruited rival Barcelona star madeira filipe with a 6-year contract, and then started his "one superstar a year" signing strategy. From Figo in 2000, Zidane in 2001, Ronaldo in 2002, Beckham in 2003, Owen in 2004, Ramos in 2005, Ruud van Nistrui and Cannavaro in 2006, the star-studded Real Madrid has become a veritable Galaxy battleship in the hearts of the majority of fans, and the success of winning the FIFA Best Club in the 20th century is the best evidence.

... ...

Until June 2009, florentino made a comeback, announcing that the second phase of Galaxy Battleship set sail again. In the summer of 2009 alone, florentino threw 261 million euros in the transfer market to build the second phase of Galaxy Battleship, and brought Kaka, Cristiano Ronaldo, Benzema and other superstars to the Bernabeu Stadium one after another.

... ...

Florentino improved Real Madrid's economic situation, and the new Bernabeu helped the Galaxy battleship set sail in the third phase ... ... At the beginning of 2021, according to KPMG's financial report of major European league teams in 2019/20 season, when most clubs suffered losses due to the COVID-19 epidemic, Real Madrid not only won the top spot in Europe with operating income of 681.2 million euros, but also became the only two profitable clubs in Europe with Bayern.

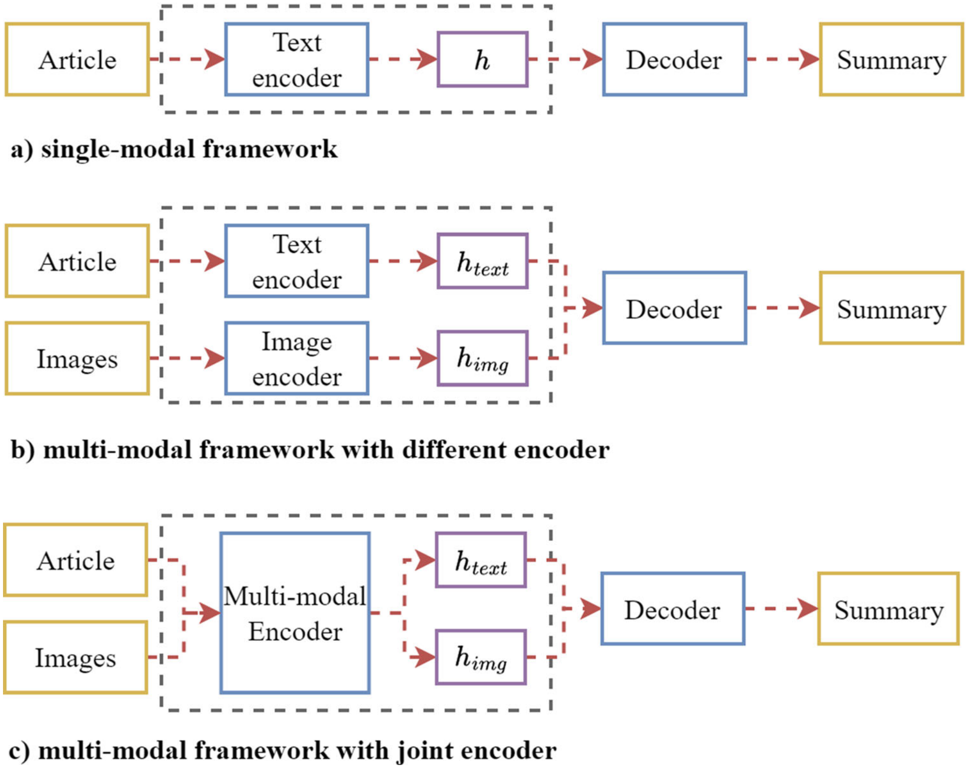
Fig. 1 An example for explaining the semantic alignment between images and paragraphs in the document. "..." means some content is omitted. Each image is aligned to the paragraph at right

as input and generates one summary with two selected images.

Recent single-modal summarization models always employ an encoder–decoder framework with transformers structure [4, 5]. Existing multi-modal models always add separate encoders for different modalities into the single-modal encoder–decoder framework [2, 3, 6–10]. We show the widely used structure of them in Fig. 2a, b. The representation of different modalities is obtained separately from single-modal encoders, which leads to the model cannot effectively capture the interaction between them. Recently, some works have paid attention to how to enhance image text interaction [9, 11] by adding interactive modules or auxiliary tasks.

However, previous works ignore the paragraph-level vision-language semantic alignment, where an example is shown in Fig. 1. The vision-language semantic refers to the meaning conveyed by vision and language. Alignment refers to establishing correspondence between vision and language that share the same meaning. The semantics of each paragraph is highly corresponding to the image on the left. There exists a semantic correspondence between images and paragraphs. If a model can reorder the images based on the semantic meaning of the paragraphs, it indicates that the model comprehends and aligns the semantic features of both the images and the text. Besides, visual-language joint encoding is not well-applied for multi-modal summarization tasks, which has been proven effective on many multi-modal natural language understanding (NLU) tasks (e.g., Visual Question Answering) [12–16].

To improve these deficiencies, in this paper, we propose the Vision-Language Summarization model ViL-Sum with a universal transformer-based encoder–decoder structure. The core of ViL-Sum is a joint multi-modal encoder with two well-designed tasks, image reordering, and image selection, which aims to guide the model to learn better vision-language representations and capture the alignment of paragraph-level vision-language semantics. Specifically, we use a backbone (e.g., ViT [17]) to convert images into visual token embeddings and concatenate them with document token embeddings as the input of the joint multi-modal encoder. The ViL-Sum structure with the joint multi-modal encoder is shown in Fig. 2c. To model paragraph-level vision-language semantic alignment, we propose a simple but effective image reordering task. It forces the model to reorder shuffled input images, which guides the model to learn the corresponding relation between paragraphs and images. To further enhance vision-language representation, we also train ViL-Sum with an image selection task, which selects several summary-related images as part of the multi-modal summary. We follow [9]



a) single-modal framework

b) multi-modal framework with different encoder

c) multi-modal framework with joint encoder

Fig. 2 Multi-modal summarization models with different encoder structures

of using image caption to construct pseudo-labels. Finally, we train ViL-Sum with text summary generation, image selection, and image reordering tasks in a multi-task manner.

Experiments show that our ViL-Sum with multi-task training can outperform baselines by a wide margin. And further analysis demonstrates that the improvement is exactly from the joint modeling and multi-task training. However, the caption of the image is not always available. So, the image selection task is not generalization for all datasets. It is deserved to mention that if we remove the image selection task, our proposed multi-modal encoder and the image reordering task still help the model beat all comparison models.

Our contributions can be summarized as follows:

- We propose a novel vision-language Summarization (ViL-Sum) model, which can jointly encode images and text to capture their interrelation.
- We propose two auxiliary tasks and employ multi-task learning to guide the model to learn the paragraph-level vision and language semantic alignment.
- Our model outperforms all current state-of-the-art methods on most automatic and manual evaluation metrics. And in further analysis, we find that the improvement is exactly from the paragraph-level semantic alignment modeling and multi-task training.

2 Related work

2.1 Single-modal summarization

Recently, text summarization models have achieved remarkable performance with the development of pre-trained language models. Liu and Lapata [18] first apply the pre-trained language model BERT [19] to summarization tasks. They add several transformers as the decoder to the BERT encoder and then train them with different learning rates. Their work outperforms all traditionally trained neural models. Pegasus [4] and BART [5] are two fully pre-trained models for summarization generation with well-designed self-supervised tasks. Their appearance provides powerful base models for summarization and totally changed the research paradigm in the summarization task. After that, more and more summarization works begin to focus on pre-trained language models, including supervised and unsupervised methods [20–26].

2.2 Vision-language representation

Large-scale Transformers-based [27] vision and language representation models [5, 19, 28] have achieve state-of-the-art results on many Natural Language Processing (NLP) tasks. They first pre-trained on a large-scale corpus with self-supervised tasks and then fine-tuned on specific downstream tasks. Most existing vision and language pre-training (VLP) models [29–34] adopt two different encoders to model vision and language separately, which extracts visual features by an object detection model and then, combines the derived object-centric representation of the image and text. Recently, large-scale vision and language representation learning has tried to jointly encode different modalities with the same encoder and achieve promising improvements [12–16, 35–37]. Their success proves the joint modeling of different modalities is practicable. Besides, multi-task learning is effective in vision-Language representation; they perform joint training diverse tasks for better investigating relationships between vision and language [31, 38, 39].

2.3 Multi-modal summarization

Different from single-modal text summarization, multi-modal summarization is a task to generate a condensed summary to cover the primary information from multimedia data. One of the most significant characteristics of this task is it is not only based on text information, but can also employ rich visual information from images, audio, and videos. Multi-modal summarization tasks can be divided into two types with different outputs: single-modal output [1, 6, 7] and multi-modal output [3, 9, 11, 40]. Compared with single-modal output, a multi-modal output summary can increase users-satisfaction [3] and first proposes a large-scale Multi-modal Summarization with Multi-modal Output (MSMO) dataset. To tackle the gap between training and testing in the MSMO task, Zhu et al. [9] propose two methods to obtain pseudo image labels and train the model with multi-modal optimization objectives. Zhang et al. [41] propose to integrate extractive and abstractive summaries and adopt knowledge distillation with a vision and language pre-training model. Zhang et al. [42] propose a location-aware approach to further leverage the image location information. Jiang et al. [43] introduce a cross-modal alignment mechanism by exploiting pseudo image captions to bridge the cross-modal semantic gap. Inspired by MSMO, Li et al. [44] propose the task of video-based Multi-modal Summarization with Multi-modal Output (VMSMO) and a Dual-Interaction-based Multi-modal Summarizer (DIMS) model, including a local conditional self-attention mechanism and a global-attention mechanism to model and summarize multi-modal input.

However, previous works all obtain vision-language representation via separate encoders for different modalities, which has been proved weaker than joint representation in vision-language representation learning research [15, 16]. Besides, they ignored the special paragraph-level semantic alignment between different modalities. In this paper, we proposed a novel vision-language summarization ViL-Sum model with a multi-task learning framework to tackle these issues.

3 Methodology

We show the main architecture of our ViL-Sum model in Fig. 3. Firstly, we employ a backbone network as the image tokenizer to convert images into visual token embeddings in Fig. 3a. Then, text embeddings and visual token embeddings are concatenated as the input of the main encoder–decoder framework in Fig. 3b. Finally, we train the ViL-Sum model in a multi-task manner. In the following sections, we will first introduce vision-language joint representation. Then, we will describe the details of multi-task learning.

3.1 Vision-language joint representation

First of all, we formalize the input and output of our ViL-Sum as (D, I) and (S, I_S), where $D = \{t_1, t_2, \ldots, t_T\}$ refers to the sequence of tokens from the input document, $I = \{\text{img}_0, \text{img}_1, \ldots, \text{img}_M\}$ refers to the sequence of input images from the input document, $S = \{t_1, t_2, \ldots\}$ refers to the sequence of tokens from gold text summary, and $I_S = \{\text{img}_1, \text{img}_2, \ldots, \text{img}_K\}$ refers to K selected images for the multi-modal summary.

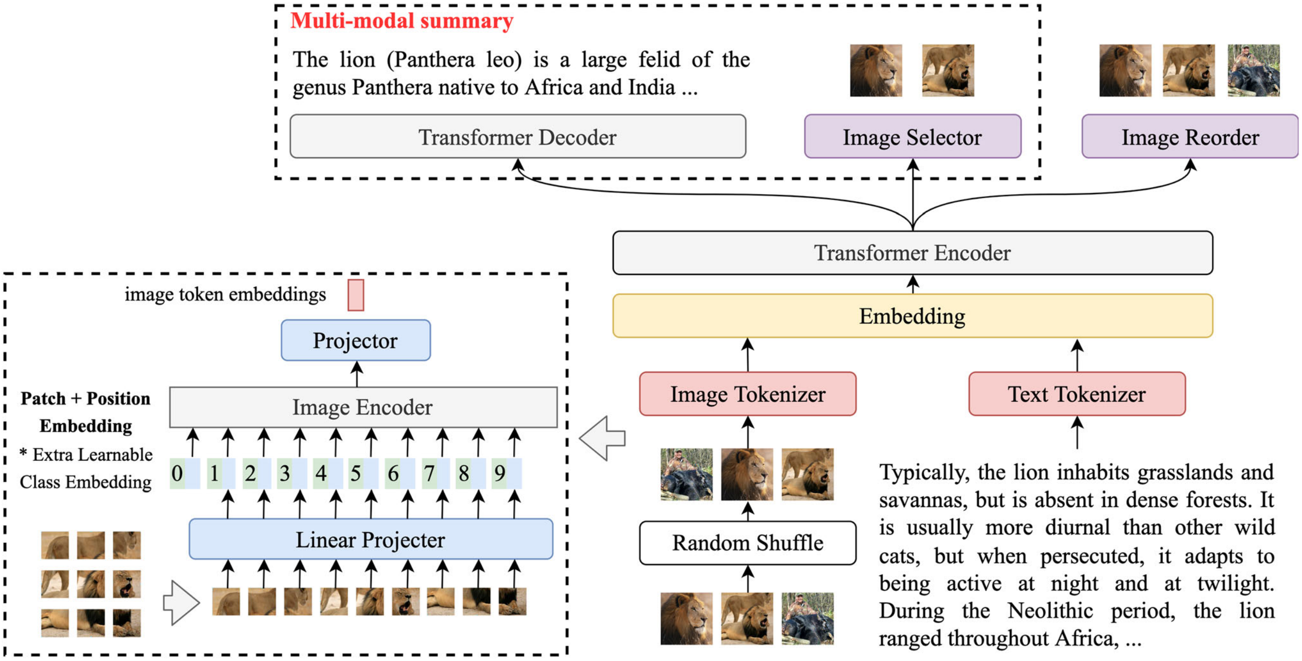
3.1.1 Document embeddings

Each document is firstly converted into the sequence of tokens $\{t_1, t_2, \ldots, t_T\}$, and then, two special tokens "$\langle s \rangle$"and "$\langle \backslash s \rangle$"are added to represent the start and end of the document. After that, we map each token into vector representation $E_D = \{e_{\text{start}}, e_1, \ldots, e_T, e_{\text{end}}\}$ with text embedding layer.

3.1.2 Image embeddings

Different from previous methods, which extract many image features via existing object detection models. We employ ViT [17] as the backbone, which split each image into several patches and then encode them. The details of the image tokenizer are shown in Fig. 3b.

Firstly, we reshape image $\text{img} \in \mathbb{R}^{H \times W \times C}$ into a sequence of flattened 2D patches $\{\text{img}^p \in \mathbb{R}^{N \times (P^2 \cdot C)}\}_{p=1}^{N}$, where ($H$, W) are the resolution of the original image, C is the number of channels; (P, P) are the resolution of each image patch, and $N = HW/P^2$ is the resulting number of patches. Then, we can obtain a sequence of image patches $\{\text{img}^p\}_{p=1}^{N}$ as the input of the image tokenizer.


Multi-modal summary
The lion (Panthera leo) is a large felid of the genus Panthera native to Africa and India ...
Transformer Decoder
Image Selector
Image Reorder
Transformer Encoder
Embedding
Image Tokenizer
Text Tokenizer
Random Shuffle
Typically, the lion inhabits grasslands and savannas, but is absent in dense forests. It is usually more diurnal than other wild cats, but when persecuted, it adapts to being active at night and at twilight. During the Neolithic period, the lion ranged throughout Africa, ...
image token embeddings
Projector
Patch + Position Embedding
* Extra Learnable Class Embedding
Image Encoder
0 1 2 3 4 5 6 7 8 9
Linear Projecter

a) image tokenizer **b) ViL-Sum model with multi-task learning**

Fig. 3 The overall framework of our proposed ViL-Sum model. **a** Is the detail of the ViT-based image tokenizer. **b** Is the encoder–decoder framework with multi-task learning

Secondly, the patches are linearly projected to patch embeddings $e^p = E \times \mathrm{img}_i^p$, where $E \in \mathbb{R}^{(P^2 \cdot D) \times C}$. We also add a special token "[class]" with learnable embedding e^0. Then, attaching position embeddings and patch embeddings as input Z_0 for the image encoder to retain positional information of images:

$$Z_0 = [e_i^0; e_i^1; \ldots; e_i^N] + E_{\text{pos}} \tag{1}$$

where $Z_0, E_{\text{pos}} \in \mathbb{R}^{(N+1) \times D}$, E_{pos} is position embeddings.

Finally, we employ the pre-trained ViT with L encoder layers as the backbone to encode these patches of each image. This backbone also can be replaced by any other encoders (e.g., linear projection layer).

$$Z_{\ell+1} = \mathtt{EncoderLayer}(Z_\ell), \quad \ell = 1, 2, \ldots, L \tag{2}$$

The global max-pooling of output vectors is obtained as the visual token embedding of image img_i:

$$v_i = \mathtt{MaxPooling}(Z_L) \tag{3}$$

where $v_i \in R^D$. Through the image tokenizer, we can convert the sequence of input images into a sequence of visual token embeddings $E_v = \{v_i\}_{i=1}^M$.

3.1.3 Multi-modal encoder

The input of the multi-modal encoder is the concatenation of visual token embeddings E_v and token embeddings E_D. We can formalize the input as $H_0 = \{E_v; E_D\}$ and then, encode visual and text embeddings with 12 transformer blocks. Finally, we can obtain vision-language representation H_L from the last layer output of this encoder.

$$H_L = \{h_{v_1}, \ldots, h_{v_M}, h_{\text{start}}, h_1, \ldots, h_{\text{end}}\} \tag{4}$$

The vision and language semantics interact with the self-attention mechanism of the transformer structure during the encoding process.

3.2 Visual-enhanced summary generation

The vision-language representations H_L from the previous multi-modal encoder contain multi-modal features of input text and images. After encoding, we feed the representations H_L into the decoder to generate a text summary. The target of the summary generation task is to minimize the negative log-likelihood of the reference y tokens as given input document D and images I via updating model parameters θ. The loss function of the summary generation task is as follows:

$$\mathcal{L}_\theta^{\text{GEN}} = -\sum_{j=1}^{|y|} \log P_\theta(y_j | y_{<j}, D, I) \tag{5}$$

Different from single-modal summarization tasks, this optimization target also depends on the features from input images I, which enhance the final summary generation.

3.3 Images reordering

To align the paragraphs and images from the input, in this section, we introduce a simple yet effective task, image reordering, to guide the model to learn semantic alignment. Specifically, we shuffle the order of input images and then, force the ViL-Sum model to predict the original order of input images with a classification head:

$$y_i = P(\text{pos}_i) = \texttt{softmax}(W \cdot h_{v_i} + b) \tag{6}$$

where all input images share one classification head. To train the classification layer, the model computes loss and minimizes the objective function:

$$\mathcal{L}_\theta^{\text{IR}} = \frac{1}{M}\sum_{i=1}^{M}\sum_{c=1}^{C} -\hat{y}_{\text{ic}} \log y_{\text{ic}} \tag{7}$$

where C is the number of categories, depending on the number of input images. We set $C = 10$. If the number of input images is greater than 10, we only keep the first 10 images as input images.

3.4 Images selection

We also train our ViL-Sum with multi-modal output reference following [9]. To build pseudo image selection labels of training data, we employ similarity between image caption and gold summary to select top-K images as labels $\hat{y}$ (K is empirically set as 3). The similarity is the average of ROUGE-1, ROUGE-2, and ROUGE-L scores. The probability to select each image is as follows:

$$y_i = P(\text{img}_i) = \sigma(W \cdot h_{v_i} + b) \tag{8}$$

The loss function of the image selection task is as follows:

$$\mathcal{L}_\theta^{\text{IS}} = \frac{1}{M}\sum_{i=1}^{M} -[\hat{y}_i \log y_i + (1-\hat{y}_i)\log(1-y_i)] \tag{9}$$

3.5 Enhanced by multi-task learning

We train our ViL-Sum with a text summary generation task and two well-designed auxiliary tasks in a multi-task manner, which are used to enhance vision-language representation and paragraph-level semantic alignment. In previous sections, we have introduced the details of them. Finally, ViL-Sum is trained with three tasks: summary generation, image selection, and image reordering, jointly by simultaneously minimizing three loss functions as follows:

$$\mathcal{L}_\theta^{\text{TOTAL}} = L_\theta^{\text{GEN}} + L_\theta^{\text{IS}} + L_\theta^{\text{IR}} \tag{10}$$

It is deserved to mention that the caption of the image is not always available. So, the image selection task is not generalization for all datasets. If we remove the image selection task, we can select images by measuring the similarity between generated summary and vector representations of images. Our proposed multi-modal encoder and the image reordering task still help the model achieve excellent performance.

4 Experimental setup

4.1 Dataset

We employ the MSMO dataset [3] to evaluate the effectiveness of our proposed ViL-Sum. MSMO dataset is a large-scale dataset for the Multi-modal Summarization with Multi-modal Output tasks. Each example in the dataset is a triplet (document, images, summary), which contains more than one image in each example. This dataset contains online news articles (723 tokens on average) paired with multiple image caption pairs (6.58 images on average) and multi-sentence summaries (70 tokens on average). For test data, based on text reference, at most, three images are annotated to produce a multi-modal reference by humans. The detailed statistical information of the MSMO dataset is shown in Table 1.

4.2 Baseline models

We report the existing multi-modal summarization methods (ATG, ATL, HAN, GR) [3], MMR [45] and MOF$_{\text{dec}}^{\text{RR}}$ [9] using multiple metrics. We also report the result of PGC [46], which is a single-modal summarization model.

To prove the effectiveness of our proposed joint representation and multi-task learning, we mainly compare with BART-base [5] model and a reproduced two-stream model BART-cross which has the same structure with MOF$_{\text{dec}}^{\text{RR}}$ and replace GRU and VGG19 [47] with BART and ViT [17], respectively. To be fair, we mainly compare our

Table 1 Statistical information of MSMO dataset

	Train	Valid	Test
#Documents	293,965	10,355	10,261
#AvgTokens (D)	721	766	731
#AvgTokens (S)	70	70	72
#Images	1,928,356	68,520	71,509
#AvgImgs	6.56	6.62	6.97

D refers to the input document. *S* refers to the summary model with BART-base and BART-cross due to previous methods did not employ pre-trained models. The details of these models are as follows:

- *PGC* is the BiGRU-based pointer-generator network that allows both copying words from the input text and generating words from a fixed vocabulary.
- *ATG* is based on the PGC model. It fuses static visual features from VGG19 with text features after the BiGRU encoder. Besides, ATG selects final images by the visual-text attention weight.
- *ATL* replaces the image global features of ATG with local features (multiple pooling features), which select images by measuring the sum of visual attention distribution over the local patch features of each image.
- *HAN* is based on the ATL model and adds a hierarchical attention mechanism. This attention mechanism first attends to the image patches to get the intermediate vectors to represent images and then, attends to these vectors to get the visual context vector.
- *GR* is an extractive method that employs LexRank [48] to rank captions of images and select images based on the rank score. The text summary of it is generated by the PGC model.
- *MMR* is a unified unsupervised graph-based framework for multi-modal summarization that can cover both single-modal output summarization and multi-modal output summarization. According to specific requirements, there are three models: generic multi-modal ranking, modal-dominated multimodal ranking, and non-redundant text-image multi-modal ranking. MMR* is the corresponding MMR model that truncates the input text to 10 sentences.
- *MOF* $^{\mathrm{RR}}_{\mathrm{dec}}$ is based on ATG model. This model first constructs pseudo-labels of image selection for the final summary. Specifically, it employs the ROUGE score to measure the relevance of image caption and summary text.
- *UniMS* is a unified multi-modal summarization framework that integrates extractive and abstractive summaries and adopts knowledge distillation to improve image selection.
- *LAMS* investigates image locations for multi-modal summarization via a stack of multi-modal fusion block and formulates the high-order interactions among images and texts.
- *SITA* proposes a novel coarse-to-fine image text alignment mechanism to identify the most relevant sentence of each image and applies a cross-modal retrieval model to retrieve reference caption for an image from the golden summary.
- *BART-base* is a pre-trained seq2seq generation model, which achieved promising results in many generations of NLP tasks, especially on text summarization. We employ this model to confirm visual features' contribution to a summary generation.
- *BART-cross* is a BART-based model with the same model structure as previous ATG, ATL, HAN, GR, and MOF$^{\mathrm{RR}}_{\mathrm{dec}}$. It first encodes images with ViT and then, fuses text representation from the BART encoder output. The fusion of image and text representations employs cross-attention like the ATG model. This is the main comparison model.

For a fair comparison, we construct this BART-cross model to prove the effectiveness of joint multi-modal encoder and multi-task training in our ViL-Sum. Because our ViL-Sum without multi-task training only changes the encoding mechanism from separate encoders to the joint multi-modal encoder.

4.3 Implementation details

We train our model for 10 epochs on 8xV100 GPUs using Adam [49] with $\beta_1 = 0.9$, $\beta_2 = 0.99$, a batch size of 64. We also use a linear learning rate warm-up with 1000 steps. The weight-decay is set as 10^{-4}. The model is initialized with ViT-B/16 and BART-base parameters. The max length of input images and tokens is 10 and 512, respectively. For the image tokenizer, we employ the same setting with ViT-b/16 in [17]. During testing, we generate the summary with a beam size of 3, and the minimum and maximum decoding lengths are set as 15 and 150 separately.

4.4 Evaluation metrics

We evaluate the pictorial summary with the MMAE metric [3].[1]

MMAE consists of three sub-metrics: ROUGE score (ROUGE-L), Image Precision (IP), and Image Text Relevance ($\mathrm{MAX}_{\mathrm{sim}}$). ROUGE [50] score can measure the salience of text in generated summary, which is widely used for measuring summarization systems. The image precision can measure the salience of selected images and is computed as Eq. (11).

$$\mathbf{IP} = \frac{|\mathrm{ref}_{\mathrm{img}} \cap \mathrm{rec}_{\mathrm{img}}|}{|\mathrm{rec}_{\mathrm{img}}|} \tag{11}$$

where $\mathrm{ref}_{\mathrm{img}}$ and $\mathrm{rec}_{\mathrm{img}}$ denote reference images and recommended images by MSMO systems, respectively. $\mathrm{MAX}_{\mathrm{sim}}$ can measure the relevance between selected

[1] Comment: [9] also proposes a MMAE+ to better evaluate MSMO task. However, the author did not release their MR model, which is the core component of their MMAE+. We find that the performance of MMAE and MMAE+ is very closer and consistent.

Table 2 The main results of all comparison models on different metrics

	Model	ROUGE-1	ROUGE-2	ROUGE-L	MAX$_{\text{sim}}$	IP	MMAE
1	PGC [46]	41.11	18.31	37.74	–	–	–
	ATG [3]	40.63	18.12	37.53	25.82	59.28	3.35
	ATL [3]	40.86	18.27	37.75	13.26	62.44	3.26
	HAN [3]	40.82	18.30	37.70	12.22	61.83	3.25
	GR [3]	37.13	15.03	30.21	26.60	61.70	3.20
	MMR [45]	39.22	15.46	–	–	–	–
	MMR* [45]	41.72	17.33	–	–	–	–
	MOF$^{\text{RR}}_{\text{dec}}$ [9]	41.20	18.33	37.80	26.38	65.45	**3.37**
	UniMS [41]	42.94	20.50	40.96	29.72	69.38	–
	LAMS [42]	43.07	20.28	39.34	–	–	–
	SITA [43]	**43.64**	**20.53**	**41.03**	**33.47**	**76.41**	-
2	BART-base	43.75	20.70	40.66	–	–	–
	BART-cross	43.67	20.65	40.65	30.25	65.98	3.45
	ViL-Sum	**44.29***	**20.96***	**41.34**	32.17	66.27	3.48
	ViL-Sum+SEL	44.20	20.90	41.22	34.47	68.18	3.51
	ViL-Sum+REO	44.21	20.98	41.20	34.35	69.03	3.52
	ViL-Sum+SEL,REO	44.16	20.88	41.21	**34.52***	**71.73**	**3.55***

Bold values indicate the best results of existing methods and our method

Significant improvements are marked with *(*t*-test, $p < 0.05$)

Table 3 Comparison of text summary results on ROUGE scores between multi-modal models and their single-modal models

Model	ROUGE-1	ROUGE-2	ROUGE-L
PGC	41.11	18.31	37.74
ATL	40.86	18.27	37.75
MOF$^{\text{RR}}_{\text{dec}}$	41.20	18.33	37.80
BART	41.83	19.83	39.74
UniMS	42.94	20.50	40.96
BERTSum	41.51	19.43	38.85
SITA	**43.64**	**20.53**	**41.03**
BART-base	43.75	20.70	40.66
ViL-Sum	**44.29**	**20.96**	**41.34**

Bold values indicate the best results of existing methods and our method

images and generated text summary, which trains an image text retrieval [51] model with max-margin loss to evaluate Image Text relevance. Finally, Zhu et al. [3] choose the linear regression results of 3 metrics as MMAE with human judgments and the weight for ROUGE-L, MAX$_{\text{sim}}$, and IP is 1.641, 0.854, 0.806, respectively; the intercept is 1.978.

We report the results of ROUGE-1/2/L, MAX$_{\text{sim}}$, IP, and MMAE of each model to comprehensively measure their performance. The results of our model are all the averages of three different checkpoints.

Table 4 Results evaluated by human annotators

Systems	Human score
BART-base	3.29
BART-cross	3.46
ViL-Sum (best)	**3.78**
Reference	**4.02**

Bold values indicate the best results of existing methods and our method

Table 5 Results of ViL-Sum under different numbers *K* of images

K	ROUGE-L	MAX$_{\text{sim}}$	IP	MMAE
1	40.97	34.63	70.94	3.54
2	41.12	34.33	70.40	3.53
3	**41.21**	**34.52**	**71.73**	**3.55**
4	41.08	34.49	70.61	3.53

Bold values indicate the best results of existing methods and our method

5 Results

5.1 Overall performance

The main results of all models are shown in Table 2. Previous baseline models in block 1 are based on the pointer network with BiGRU. Models in block 2 are our implementation based on the BART-base model. SEL means selection task and REO means reordering task. All reported results of ours are the average of 3 different checkpoints.

We can see that compared with the baselines, our ViL-Sum gains significant improvement on most metrics, and ViL-Sum+selection reordering achieves the best comprehensive performance except IP. SITA gains a notable improvement on IP metric. It trains a cross-modal retrieval model to retrieve reference caption for image and provides supervision signal, which is very beneficial for image selection and the image text alignment. Nevertheless, our ViL-Sum still outperforms SITA on other metrics. Compared with BART-cross, we can see that the joint representation and multi-task training both bring satisfactory improvement, which proved the effectiveness of our proposed methods. Interestingly, the introduction of image features hurts the performance of all single-modal summarization models, especially BiGRU-based models.

To further demonstrate that visual information indeed benefits text summary generation, we compare four groups of multi-modal models with the single-modal models on which they are base, as shown in Table 3. The single-modal models are above the dashed line, and the multi-modal models based on them are below the dashed line. In the first group, the multi-modal model (ATL and $\mathrm{MOF}^{\mathrm{RR}}_{\mathrm{dec}}$) are not superior to the single-modal model (PGC). These works concluded that long documents already contain enough information and that too many images would introduce noise for summary generation. On the contrary, the other three groups of results show that the multi-modal models have achieved performance improvements compared with the single-modal models. The results indicate that introducing and exploiting image information effectively can improve text summarization generation.

5.2 Performance of joint representation

Firstly, we can see that the performance of ATG, ATL, HAN, and GR all hurt ROUGE scores by simply introducing images as independent visual features. Through the multi-modal objective optimization, $\mathrm{MOF}^{\mathrm{RR}}_{\mathrm{dec}}$ has a significant improvement on IP and does not decrease the quality of generated text summary. This situation proves that modeling vision and language information independently did not bring in the revenue for text summary generation. The results of BART-cross, which also introduces images as independent features, also have lower ROUGE scores than BART-base. This situation proves again the previous conclusion.

Different from previous performance on ROUGE score, our ViL-Sum with joint vision-language representation obtains better ROUGE scores, and the Image Precision (IP) and $\mathrm{MAX}_{\mathrm{sim}}$ both have a significant improvement. This demonstrates that using the joint multi-modal encoder to obtain vision-language representation is better than using separate encoders with cross-attention to fuse multi-modal features.

Table 6 Results of ViL-Sum with different image tokenizers

	ROUGE-L	$\mathrm{MAX}_{\mathrm{sim}}$	IP	MMAE
ViT	**41.21**	**34.52**	**71.73**	**3.55**
Linear	40.18	33.89	70.44	3.51
Vision	41.10	34.28	71.04	3.54

Bold values indicate the best results of existing methods and our method

5.3 Performance of multi-task learning

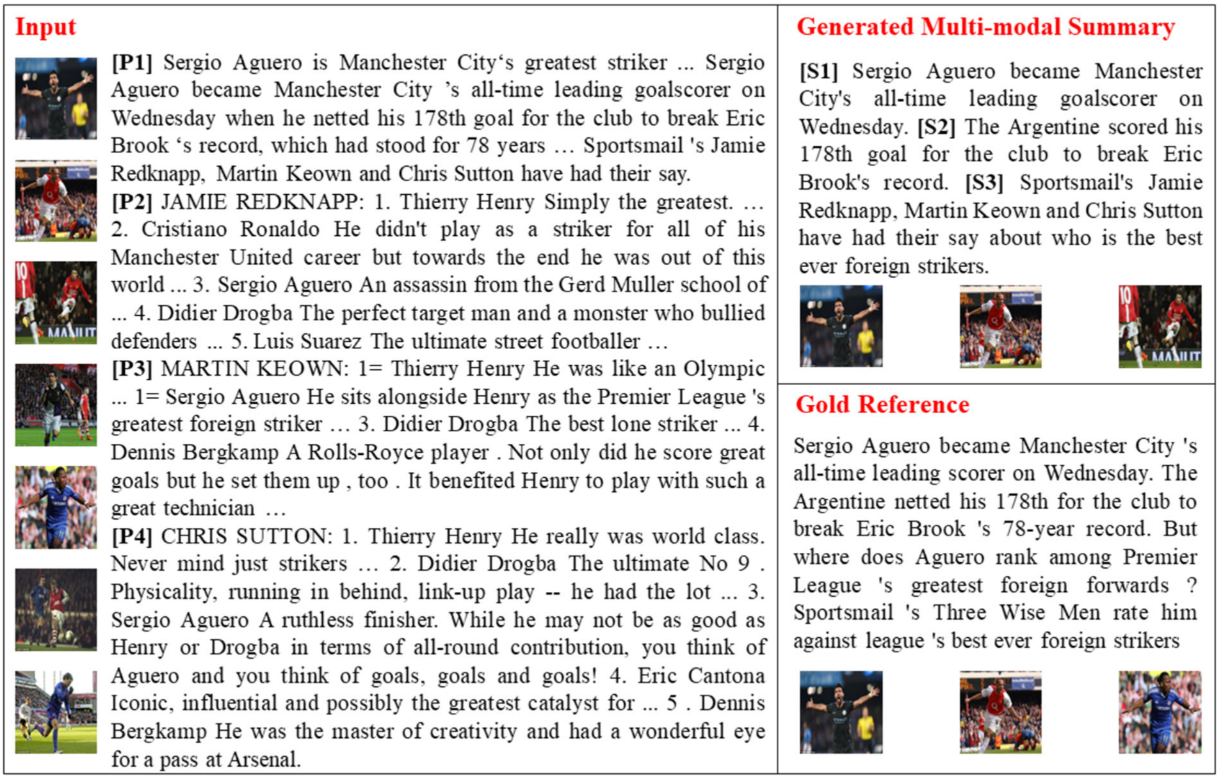
The result of ViL-Sum without multi-task learning has achieved good performance and is better than BART-cross. In this section, we will analyze the influence of our proposed multi-task learning. From the results, we can see that the introduction of image selection and reordering bring a slight decrease in ROUGE scores. Meanwhile, the IP and $\mathrm{MAX}_{\mathrm{sim}}$ scores increase significantly, which makes the overall score MMAE better than ViL-Sum without multi-task training.

We report the ablation study results of two auxiliary tasks in the second block of Table 2. From the results, we can see that image selection and reordering both can bring improvement in IP and $\mathrm{MAX}_{\mathrm{sim}}$ scores. The combination of two tasks can push the overall score MMAE higher. The comparison of these models demonstrates that the introduction of multi-task learning exactly improved the vision-language representation and semantic alignment, which is reflected in the improvement of the multi-modal metrics: IP, $\mathrm{MAX}_{\mathrm{sim}}$ and MMAE.

6 Discussion

6.1 Human evaluation

We randomly sample 100 examples from the test set to conduct the human evaluation. The multi-modal summary of golden reference, BART-base, BART-cross, and our ViL-Sum (best) is evaluated by three human annotators. Each annotator will score each example with a rating scale from 1 (worst) to 5 (best). Table 4 shows the average scores from three annotators (t-test, $p < 0.05$). We can see that annotators tend to give the multi-modal summary from BART-cross and our ViL-Sum higher scores. In addition, our ViL-Sum outperforms two strong baselines by a wide margin and is close to the references. It is noteworthy that



Input

[P1] Sergio Aguero is Manchester City‘s greatest striker ... Sergio Aguero became Manchester City 's all-time leading goalscorer on Wednesday when he netted his 178th goal for the club to break Eric Brook ‘s record, which had stood for 78 years … Sportsmail 's Jamie Redknapp, Martin Keown and Chris Sutton have had their say.
[P2] JAMIE REDKNAPP: 1. Thierry Henry Simply the greatest. … 2. Cristiano Ronaldo He didn't play as a striker for all of his Manchester United career but towards the end he was out of this world ... 3. Sergio Aguero An assassin from the Gerd Muller school of ... 4. Didier Drogba The perfect target man and a monster who bullied defenders ... 5. Luis Suarez The ultimate street footballer …
[P3] MARTIN KEOWN: 1= Thierry Henry He was like an Olympic ... 1= Sergio Aguero He sits alongside Henry as the Premier League 's greatest foreign striker … 3. Didier Drogba The best lone striker ... 4. Dennis Bergkamp A Rolls-Royce player . Not only did he score great goals but he set them up , too . It benefited Henry to play with such a great technician …
[P4] CHRIS SUTTON: 1. Thierry Henry He really was world class. Never mind just strikers … 2. Didier Drogba The ultimate No 9 . Physicality, running in behind, link-up play -- he had the lot ... 3. Sergio Aguero A ruthless finisher. While he may not be as good as Henry or Drogba in terms of all-round contribution, you think of Aguero and you think of goals, goals and goals! 4. Eric Cantona Iconic, influential and possibly the greatest catalyst for ... 5 . Dennis Bergkamp He was the master of creativity and had a wonderful eye for a pass at Arsenal.

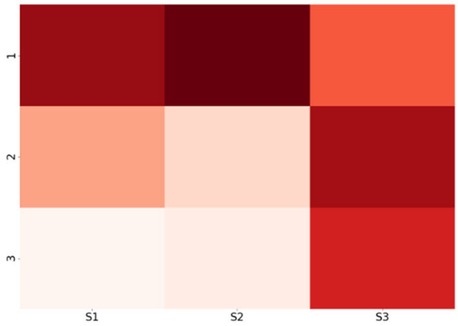
Generated Multi-modal Summary

[S1] Sergio Aguero became Manchester City's all-time leading goalscorer on Wednesday. **[S2]** The Argentine scored his 178th goal for the club to break Eric Brook's record. **[S3]** Sportsmail's Jamie Redknapp, Martin Keown and Chris Sutton have had their say about who is the best ever foreign strikers.

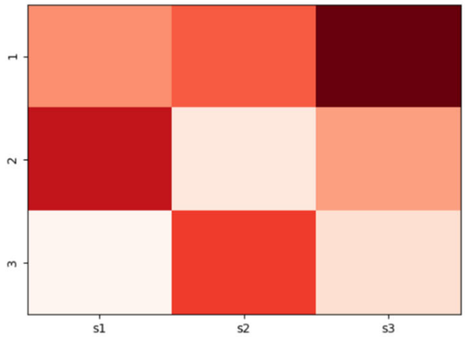
Gold Reference

Sergio Aguero became Manchester City 's all-time leading scorer on Wednesday. The Argentine netted his 178th for the club to break Eric Brook 's 78-year record. But where does Aguero rank among Premier League 's greatest foreign forwards ? Sportsmail 's Three Wise Men rate him against league 's best ever foreign strikers

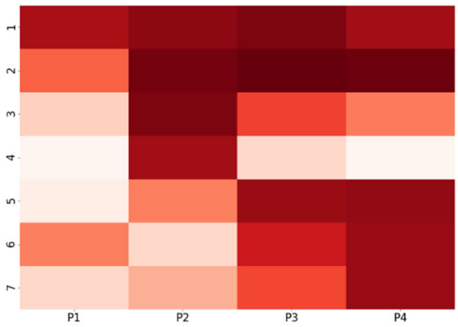
a) Example from test set and generated multi-modal summary



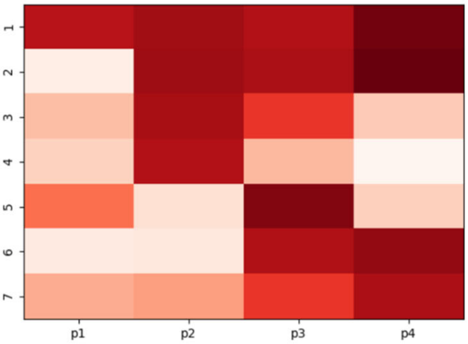
b) Heatmap of ViL-Sum for summary sencences and selected images



d) Heatmap of BART-cross for summary sencences and selected images



c) Heatmap of ViL-Sum for selected paragraphs and images



e) Heatmap of BART-cross for selected paragraphs and images

◀ **Fig. 4** Example from the test set with the generated multi-modal summary. **a** Is the full example. **b, d** are heatmaps that show the relevance of the summary and selected images. **c, e** are heatmaps that show the relevance of selected paragraphs and images. Each color block means cosine similarity between the image and text object. The darker color refers to higher similarity (colour figure online)

our work shows a higher improvement in human evaluation compared with the improvement in metrics. This reflects the gap between human evaluation and metrics.

6.2 Impact of different numbers of images

Table 5 depicts the experimental results of our model performance varying with different K (the number of selected summary-related images at the final summary). Since the golden reference in the test set contains three images, the consistency between training and test makes the model perform best when K is 3. Overall, our model is not very sensitive with K. With different K, our ViL-Sum all achieve excellent performance, which proves our method can identify the real importance images from multi-modal inputs. Besides, we guess the image selection of the MSMO dataset is simple due to the data from the news.

6.3 Impact of different image tokenizer

To further evaluate the effectiveness of joint modeling and multi-task learning, we replace the backbone of the image tokenizer to observe the performance of ViL-Sum. We replace the ViT backbone with Linear Layer and an image tokenizer from Vision Transformer [52]. Both of them have much smaller parameters than the ViT backbone. Specifically, linear is the simple version of ViT which replaces the transformer image encoder with a simple linear layer to map the images into visual token embeddings. Vision is an image tokenizer from Vision Transformer [52], which can convert one image into several visual token embeddings. Table 6 reports the results of them. We can see that the ViT exactly provides better visual features than the other two backbones. However, the performance does not drop sharply with the replacement of the image tokenizer. This proves that Our proposed two strategies are robust and the ViL-Sum is flexible with different image tokenizers.

6.4 Case study and relevance visualization

We select one typical example from the test set and visualize the relevance of (1) summary sentences and selected images; (2) selected paragraphs and images; (3) all tokens and images in Figs. 4 and 5. Each color block means a cosine similarity between the image and text object. The darker color refers to a higher similarity in the heatmap. With our proposed methods, the generated summary contains high-quality summary with three related images as shown in Fig. 4a. From different relevant visualizations, we can see that our ViL-Sum can effectively align the semantic representation of summary sentences and selected images as shown in Fig. 4b. The input images can be aligned with paragraphs by training with image reordering

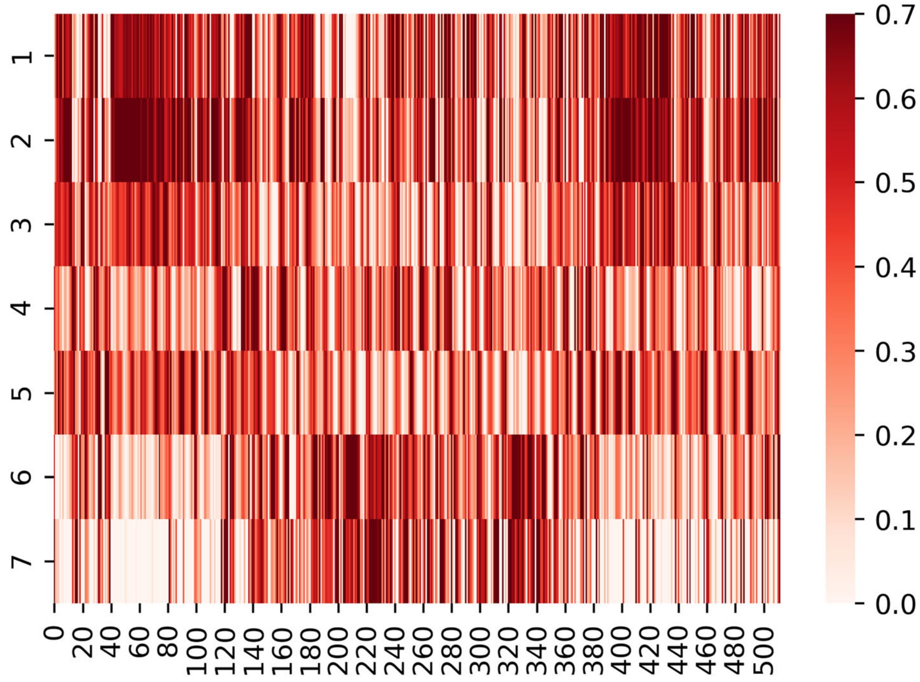


Fig. 5 The heatmap shows the relevance of all input tokens and images. The darker color refers to higher similarity (colour figure online)

as shown in Fig. 4c. By comparing Fig. 4b, d, as well as Fig. 4c, e, we can observe that ViL-Sum surpasses BART-cross in paragraph-level vision-language semantic alignment, primarily attributed to the incorporation of multi-tasking learning.

We also report the heatmap of all input tokens and images in Fig. 5, which is consistent with Fig. 4b, c. This case proves that the multi-task training really helps ViL-Sum learn reasonable relations between images and input paragraphs.

7 Conclusion

In this paper, we propose a novel Vision-Language Summarization (ViL-Sum) model, which can enhance the vision-language representation and the paragraph-level semantics alignment through multi-task training and joint modeling. A multi-modal encoder jointly encodes images and text to capture their interrelation. The reordering task and image selection task guide the model to learn paragraph-level vision and language semantic alignment. Our ViL-Sum achieves new state-of-the-art results on most automatic and manual evaluation metrics. Further analysis demonstrates that the improvement is from the joint multi-modal encoder and multi-task training.

In human evaluation, we have observed a gap between human evaluation and metrics. Introducing more appropriate evaluation metrics contribute to the development of multi-modal summarization. Besides, we only use the MSMO dataset due to the lack of other datasets. Our proposed image reordering task is straightforward yet effective, we will extend our method to more scenarios (e.g., vision-language pre-training models) and modalities (e.g., audio and video) in the future. Furthermore, we plan to generalize our method to other multi-modal tasks, such as multi-modal question answering.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62276017, U1636211, 61672081), the 2022 Tencent Big Travel Rhino-Bird Special Research Program, and the Fund of the State Key Laboratory of Software Development Environment (Grant No. SKLSDE-2021ZX-18).

Data availability All data generated or analyzed during this study are included in this published article.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Evangelopoulos G, Zlatintsi A, Potamianos A et al (2013) Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. IEEE Trans Multimed 15(7):1553–1568. https://doi.org/10.1109/TMM.2013.2267205
2. Li H, Zhu J, Ma C et al (2017) Multi-modal summarization for asynchronous collection of text, image, audio and video. In: EMNLP 2017. Association for Computational Linguistics, Copenhagen, pp 1092–1102. https://doi.org/10.18653/v1/D17-1114
3. Zhu J, Li H, Liu T et al (2018) MSMO: Multimodal summarization with multimodal output. In: Proceedings of the 2018 conference on empirical methods in natural language processing. Association for Computational Linguistics, Brussels, pp 4154–4164. https://doi.org/10.18653/v1/D18-1448
4. Zhang J, Zhao Y, Saleh M et al (2019) Pegasus: pre-training with extracted gap-sentences for abstractive summarization. arXiv:1912.08777
5. Lewis M, Liu Y, Goyal N et al (2020) BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: ACL 2020. Association for Computational Linguistics, pp 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703
6. Li H, Zhu J, Liu T et al (2018) Multi-modal sentence summarization with modality attention and image filtering. In: IJCAI-18. International joint conferences on artificial intelligence organization, pp 4152–4158. https://doi.org/10.24963/ijcai.2018/577
7. Chen J, Zhuge H (2018) Abstractive text-image summarization using multi-modal attentional hierarchical RNN. In: EMNLP. Association for Computational Linguistics, Brussels, pp 4046–4056. https://doi.org/10.18653/v1/D18-1438
8. Khullar A, Arora U (2020) MAST: Multimodal abstractive summarization with trimodal hierarchical attention. In: Proceedings of the first international workshop on natural language processing beyond text. Association for Computational Linguistics, pp 60–69. https://doi.org/10.18653/v1/2020.nlpbt-1.7
9. Zhu J, Zhou Y, Zhang J et al (2020) Multimodal summarization with guidance of multimodal reference. Proc AAAI Conf Artif Intell 34(05):9749–9756. https://doi.org/10.1609/aaai.v34i05.6525
10. Im J, Kim M, Lee H et al (2021) Self-supervised multimodal opinion summarization. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers). Association for Computational Linguistics, pp 388–403. https://doi.org/10.18653/v1/2021.acl-long.33
11. Zhang L, Zhang X, Pan J (2022) Hierarchical cross-modality semantic correlation learning model for multimodal summarization. Proc AAAI Conf Artif Intell 36(10):11676–11684. https://doi.org/10.1609/aaai.v36i10.21422
12. Li G, Duan N, Fang Y et al (2020) Unicoder-vl: a universal encoder for vision and language by cross-modal pre-training. Proc AAAI Conf Artif Intell 34(07):11336–11344. https://doi.org/10.1609/aaai.v34i07.6795
13. Li X, Yin X, Li C et al (2020) Oscar: object-semantics aligned pre-training for vision-language tasks. In: ECCV 2020
14. Zhang P, Li X, Hu X et al (2021) Vinvl: making visual representations matter in vision-language models. In: CVPR 2021
15. Xu H, Yan M, Li C et al (2021) E2E-VLP: End-to-end vision-language pre-training enhanced by visual learning. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers). Association

for Computational Linguistics, pp 503–513. https://doi.org/10.18653/v1/2021.acl-long.42
16. Zhou L, Palangi H, Zhang L et al (2020) Unified vision-language pre-training for image captioning and VQA. Proce AAAI Conf Artif Intell 34(07):13041–13049. https://doi.org/10.1609/aaai.v34i07.7005
17. Dosovitskiy A, Beyer L, Kolesnikov A et al (2021) An image is worth 16x16 words: transformers for image recognition at scale. In: International conference on learning representations. https://openreview.net/forum?id=YicbFdNTTy
18. Liu Y, Lapata M (2019) Text summarization with pretrained encoders. In: EMNLP-IJCNLP 2019. Association for Computational Linguistics, Hong Kong, China, pp 3730–3740. https://doi.org/10.18653/v1/D19-1387
19. Devlin J, Chang MW, Lee K et al (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). Association for Computational Linguistics, Minneapolis, pp 4171–4186. https://doi.org/10.18653/v1/N19-1423
20. Zhong M, Liu P, Chen Y et al (2020) Extractive summarization as text matching. In: ACL 2020. Association for Computational Linguistics, pp 6197–6208. https://doi.org/10.18653/v1/2020.acl-main.552
21. Liang X, Wu S, Li M et al (2021) Improving unsupervised extractive summarization with facet-aware modeling. In: Findings of the association for computational linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, pp 1685–1697. https://doi.org/10.18653/v1/2021.findings-acl.147
22. Liang X, Li J, Wu S et al (2021) Improving unsupervised extractive summarization by jointly modeling facet and redundancy. In: IEEE/ACM Transactions on audio, speech, and language processing, pp 1–1. https://doi.org/10.1109/TASLP.2021.3138673
23. Liang X, Li J, Wu S et al (2022) An efficient coarse-to-fine facet-aware unsupervised summarization framework based on semantic blocks. In: Proceedings of the 29th international conference on computational linguistics. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, pp 6415–6425. https://aclanthology.org/2022.coling-1.558
24. Cattan A, Eden L, Kantor Y et al (2023) From key points to key point hierarchy: structured and expressive opinion summarization. In: Rogers A, Boyd-Graber J, Okazaki N (eds) Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers). Association for Computational Linguistics, Toronto, pp 912–928. https://doi.org/10.18653/v1/2023.acl-long.52
25. Adams G, Fabbri A, Ladhak F et al (2023) Generating EDU extracts for plan-guided summary re-ranking. In: Rogers A, Boyd-Graber J, Okazaki N (eds) Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers). Association for Computational Linguistics, Toronto, pp 2680–2697. https://doi.org/10.18653/v1/2023.acl-long.151
26. Bao G, Ou Z, Zhang Y (2023) GEMINI: controlling the sentence-level summary style in abstractive text summarization. In: Bouamor H, Pino J, Bali K (eds) Proceedings of the 2023 conference on empirical methods in natural language processing. Association for Computational Linguistics, Singapore, pp 831–842. https://doi.org/10.18653/v1/2023.emnlp-main.53
27. Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. In: Proceedings of the 31st international conference on neural information processing systems. NIPS'17. Curran Associates Inc., Red Hook, pp 6000–6010
28. Radford A, Wu J, Child R et al (2019) Language models are unsupervised multitask learners. OpenAI Blog
29. Tan H, Bansal M (2019) Lxmert: Learning cross-modality encoder representations from transformers. In: Proceedings of the 2019 conference on empirical methods in natural language processing
30. Li C, Yan M, Xu H et al (2021) SemVLP: Vision-language pre-training by aligning semantics at multiple levels. arXiv:2103.07829
31. Lu J, Goswami V, Rohrbach M et al (2020) 12-in-1: Multi-task vision and language representation learning. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR). IEEE Computer Society, Los Alamitos, pp 10434–10443. https://doi.org/10.1109/CVPR42600.2020.01045
32. Yu F, Tang J, Yin W et al (2021) Ernie-vil: knowledge enhanced vision-language representations through scene graphs. Proc AAAI Conf Artif Intell 35(4):3208–3216. https://doi.org/10.1609/aaai.v35i4.16431
33. Jiang C, Xu H, Li C et al (2022) TRIPS: Efficient vision-and-language pre-training with text-relevant image patch selection. In: Goldberg Y, Kozareva Z, Zhang Y (eds) Proceedings of the 2022 conference on empirical methods in natural language processing. Association for Computational Linguistics, Abu Dhabi, pp 4084–4096. https://doi.org/10.18653/v1/2022.emnlp-main.273
34. Jiang C, Ye W, Xu H et al (2023) Vision language pre-training by contrastive learning with cross-modal similarity regulation. In: Rogers A, Boyd-Graber J, Okazaki N (eds) Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers). Association for Computational Linguistics, Toronto, pp 14660–14679. https://doi.org/10.18653/v1/2023.acl-long.819
35. Hu X, Yin X, Lin K et al (2021) Vivo: Visual vocabulary pre-training for novel object captioning. Proc AAAI Conf Artif Intell 35(2):1575–1583. https://doi.org/10.1609/aaai.v35i2.16249
36. Huang Z, Zeng Z, Huang Y et al (2021) Seeing out of the box: end-to-end pre-training for vision-language representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 12976–12985
37. Kim W, Son B, Kim I (2021) ViLT: Vision-and-language transformer without convolution or region supervision. In: Meila M, Zhang T (eds) Proceedings of the 38th international conference on machine learning, proceedings of machine learning research, vol 139. PMLR, pp 5583–5594. https://proceedings.mlr.press/v139/kim21k.html
38. Hu R, Singh A (2021) Unit: multimodal multitask learning with a unified transformer. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 1439–1449
39. Chen J, Zhu D, Shen X et al (2023) Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv:2310.09478
40. Bian J, Yang Y, Zhang H et al (2015) Multimedia summarization for social events in microblog stream. IEEE Trans Multimed 17(2):216–228. https://doi.org/10.1109/TMM.2014.2384912
41. Zhang Z, Meng X, Wang Y et al (2022) Unims: a unified framework for multimodal summarization with knowledge distillation. Proc AAAI Conf Artif Intell 36(10):11757–11764. https://doi.org/10.1609/aaai.v36i10.21431
42. Zhang Z, Wang J, Sun Z et al (2021) Lams: a location-aware approach for multimodal summarization (student abstract). Proc AAAI Conf Artif Intell 35(18):15949–15950. https://doi.org/10.1609/aaai.v35i18.17971
43. Jiang C, Xie R, Ye W et al (2023) Exploiting pseudo image captions for multimodal summarization. In: Rogers A, Boyd-Graber J, Okazaki N (eds) Findings of the association for computational linguistics: ACL 2023. Association for Computational

Linguistics, Toronto, pp 161–175. https://doi.org/10.18653/v1/2023.findings-acl.12

44. Li M, Chen X, Gao S et al (2020) VMSMO: learning to generate multimodal summary for video-based news articles. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, pp 9360–9369. https://doi.org/10.18653/v1/2020.emnlp-main.752
45. Zhu J, Xiang L, Zhou Y et al (2021) Graph-based multimodal ranking models for multimodal summarization. ACM Trans Asian Low-Resour Lang Inf Process. https://doi.org/10.1145/3445794
46. See A, Liu PJ, Manning CD (2017) Get to the point: summarization with pointer-generator networks. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers). Association for Computational Linguistics, Vancouver, pp 1073–1083. https://doi.org/10.18653/v1/P17-1099
47. Liu S, Deng W (2015) Very deep convolutional neural network based image classification using small training sample size. In: 2015 3rd IAPR Asian conference on pattern recognition (ACPR), pp 730–734. https://doi.org/10.1109/ACPR.2015.7486599
48. Erkan G, Radev DR (2004) Lexrank: graph-based lexical centrality as salience in text summarization. J Artif Int Res 22(1):457–479
49. Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. In: ICLR 2015
50. Lin CY (2004) ROUGE: A package for automatic evaluation of summaries. In: Text summarization branches out. Association for Computational Linguistics, Barcelona, pp 74–81. https://aclanthology.org/W04-1013
51. Faghri F, Fleet DJ, Kiros J et al (2018) Vse++: Improving visual-semantic embeddings with hard negatives. In: BMVC
52. Wu B, Xu C, Dai X et al (2020) Visual transformers: token-based image representation and processing for computer vision. arXiv:2006.03677

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.