DIUSum: Dynamic Image Utilization for Multimodal Summarization

Min Xiao[1,2], **Junnan Zhu**[1,2], **Feifei Zhai**[1,3], **Yu Zhou**[1,3*], **Chengqing Zong**[1,2]

[1] State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, CAS, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] Fanyu AI Laboratory, Zhongke Fanyu Technology Co., Ltd, Beijing, China
{ min.xiao, junnan.zhu, yzhou, cqzong }@nlpr.ia.ac.cn, zhaifeifei@zkfy.com

Abstract

Existing multimodal summarization approaches focus on fusing image features in the encoding process, ignoring the individualized needs for images when generating different summaries. However, whether intuitively or empirically, not all images can improve summary quality. Therefore, we propose a novel **D**ynamic **I**mage **U**tilization framework for multimodal **Sum**marization (DIUSum) to select and utilize valuable images for summarization. First, to predict whether an image helps produce a high-quality summary, we propose an image selector to score the usefulness of each image. Second, to dynamically utilize the multimodal information, we incorporate the hard and soft guidance from the image selector. Under the guidance, the image information is plugged into the decoder to generate a summary. Experimental results have shown that DIUSum outperforms multiple strong baselines and achieves SOTA on two public multimodal summarization datasets. Further analysis demonstrates that the image selector can reflect the improved level of summary quality brought by the images.

Introduction

With the development of vision-language pretraining, multimodal summarization has achieved significant progress in recent years. Image information can help enhance or supplement text information to produce high-quality summaries. When fed an image and a text, multimodal summarization methods try to fuse multimodal information and generate a summary. Existing studies mainly concentrate on effectively integrating visual information into the process of feature encoding (Zhu et al. 2018; Li et al. 2018; Xiao et al. 2023; Qiu et al. 2022; Liang et al. 2023).

However, there remains an important issue that has received little attention: *Whether an image helps improve the quality of the summary?* As shown in Figure 1, the green part underlined text represents the content corresponding to the image. It can be observed that the image in Figure 1 (a) is related to a part of the article but does not significantly contribute to generating the summary. In contrast, the image in Figure 1 (b) is associated with both the article and the summary. So that the image can help highlight the core content of the article. The above two cases indicate that different samples require different modality information to obtain accurate summaries. To verify whether the existing model can meet the individual needs of different source modalities, we conduct an empirical study to explore the influence of providing different source modalities. As shown in Table 1, text-only and multimodal input summarization models are trained on the standard benchmarks. In the experiment, we test each sample with the text-only and multimodal models, and then report the higher ROUGE-1 score as the dynamic-input result. Ideally, the dynamic result means the upper bound that can be obtained by dynamically utilizing the source inputs. It also indicates that if we correctly choose the source inputs for each sample, the quality of the summary could be greatly improved.

Existing studies focus on fusing all image information with different methods (Zhang, Zhang, and Pan 2022; Liang et al. 2023), such as attention-based fusion (Li et al. 2018; Libovický and Helcl 2017; Calixto, Liu, and Campbell 2017a), gated-based fusion (Li et al. 2020; Zhou et al. 2017)

(a)

Article: the gulf emirate of dubai is to replicate ancient and modern wonders of the world in a #.#-billion-dollar city to be shaped like a falcon .
Ref Summary: world wonders come to dubai.

(b)

Article: a suicide car bomber killed four security agents in an attack on a us diplomatic convoy in the northern iraqi town of mosul , a us official said tuesday.
Ref Summary: four killed in suicide bomb strike on us diplomatic convoy.

Figure 1: Examples of useful and useless images. The green part indicates that it is associated with the left image.

dataset	text-input (ROUGE-1)	multi-input (ROUGE-1)	dynamic-input (ROUGE-1)
MMS	49.16	49.23	54.64
MSMO	41.92	41.85	45.49

Table 1: The influence of feeding different source modalities on the datasets of MMS (Li et al. 2018) and MSMO (Zhu et al. 2018).

* Corresponding author.

and hierarchical-based fusion (Zhang et al. 2022; Qiu et al. 2022). However, they ignore whether an image is effective for the summarization task. Consequently, multimodal models do not perform better than the unimodal variants on all samples. Some studies try to design more refined architecture to utilize the images. Xiao et al. (2023) propose a filtering module to obtain helpful images for the summary coarsely. Li et al. (2022) apply ReLU to the attention activation function to abandon the unaligned images. However, they only consider incorporating images during the encoding process, making them unable to measure summary-effective image contribution. Specifically, although these methods can obtain more efficient multimodal encoding, there is still a gap between image contribution and summary quality improvement. As analyzed above, we believe evaluating the effectiveness of images directly through their contribution to the summary can further improve the summary quality.

To address this problem, we propose a Dynamic Image Utilization framework to select images dynamically for multimodal summarization. First, we propose an image selector to score each image according to two-dimensional features. The image selector predicts whether the image helps produce a higher-quality summary than unimodal input. Specifically, we optimize the image selector with the self-labeling method, which defines image contribution according to whether multimodal input can help produce higher-quality summaries compared to the unimodal one. Then, the decoder dynamically utilizes the multimodal information under the guidance of the image selector. In particular, the image information is plugged into an temporary state with hard and soft guidance from the image selector. The decoder then utilizes the temporary state to generate the summary. Through these steps, the model can acquire more summary-effective image information and provide better multimodal information for summarization.

Our contributions are summarized as follows:

- We propose a Dynamic Image Utilization framework for multimodal summarization to select and dynamically utilize summary-effective images for summarization.
- We innovatively design an image selector to score the usefulness of each image and plug the valuable image with the guidance of the image selector for summarization.
- Experimental results show that our method outperforms strong baselines and achieves SOTA on MMS and MSMO datasets, which demonstrates the conjecture that some multimodal summarization does not require images. Besides, extensive analysis proves that the image selector can reflect the improved level of summary quality brought by the images.

Related Work

Multimodal Summarization Tasks. With the rapid progress of multimedia, many forms of multimodal summary tasks have emerged in the multimodal field, such as multimodal sentence summarization (Li et al. 2018; Jangra et al. 2021), multimodal summarization with multimodal output (Zhu et al. 2018; Liang et al. 2023), video summarization (Sanabria et al. 2018; Yu et al. 2021; Mahasseni, Lam, and Todorovic 2017; Wang et al. 2019), multimodal opinion summarization (Im et al. 2021), topic-aware multimodal summarization (Mukherjee et al. 2022). Although multimodal summarization receives increasing attention, current studies usually focus on injecting the image information into the process of multimodal encoding: Fu, Wang, and Yang (2020) apply bi-hop attention to align different source modalities and bridge the gap between article and video. Li et al. (2020) propose a multimodal selective gate network to select the core text content from visual signals. Zhang et al. (2021) investigate image locations for multimodal summarization via a stack of multimodal fusion blocks, which can formulate the high-order interactions among images and texts. Zhang et al. (2023) enhance the multimodal semantic coverage with multiple visual-aware tasks. However, they neglect that not all images bring positive gains to the summary. Consequently, the redundancy or interference of image information for multimodal models is inevitable.

Effective Vision Encoding. Some studies have noticed the influence of non-textual modalities on the summary. Liang et al. (2023) design two summary-oriented vision modeling tasks to enhance vision representation, which exploits more accurate visual features to generate summaries. Xiao et al. (2023) design a coarse-to-fine network to model different image contributions for summarization. It acquires more explicit image contributions and provides better multimodal encoding for summarization. Li et al. (2022) exploit ReLU-based cross-attention to align text and vision features and abandon unaligned visual features with low-value attention. However, they generally learn the image feature implicitly. Specifically, they only consider the image effectiveness in the encoding process and ignore the individualized needs for images when generating different summaries.

In general, existing studies focus on: 1) Integrating all image information; 2) Implicitly incorporating image information at the encoding layer while ignoring the summary-effectiveness of the image. Inspired by the above studies, we propose DIUSum, which considers image summary effectiveness for better generating the final summary.

Proposed Methods

Overview

In this section, we introduce the details of DIUSum. It is a challenge to simultaneously consider selecting and dynamically utilizing valuable images during the training process. To address this issue, we introduce a mutil-stage training method, which enables the model to accomplish different goals across different stages. As depicted in Figure 2, the model is initiated in the first stage (§). In the second stage, we design an image selector and optimize it with the self-labeling method (§). In the third stage, the decoder dynamically utilizes multimodal information with the guidance of the image selector (§).

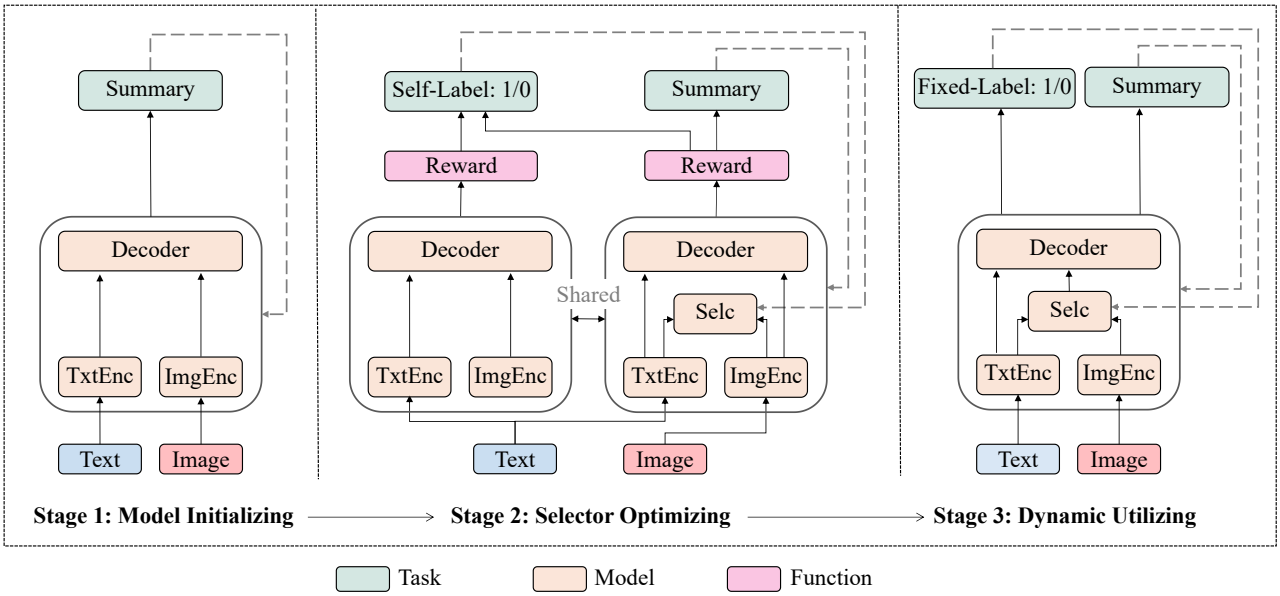

Figure 2: DIUSum framework. "TxtEnc" and "ImgEnc" represent the feature extractor of the text and image, respectively. "Selc" represents the image selector module.

Stage 1: Model Initializing

The purpose of this phase is to obtain a reliable summary model for subsequent phases. Given a dataset D consists of N triples(t, v, s) with text t, image v, summary s. The textual and visual features are extracted separately, which makes the model utilize different modalities dynamically. Then we adopt a straightforward yet effective concatenation fusion to leverage different source modalities for individual requirements.

Textual Feature Extraction. To obtain textual features, we leverage a pre-trained BERT model to extract contextual features h from text t.

$$h = \text{BERT}(t) \tag{1}$$

where $h \in \mathbb{R}^{n \times d}$. Here, d represents the hidden dimension of the BERT model. n denotes the number of tokens in t.

Visual Feature Extraction. In the process of acquiring visual features, we adopt a pre-trained model VGG-19 as the extractor. A 4096-dimensional fully-connected layer is extracted as the global feature, denoted as $v' \in \mathbb{R}^{4096}$. Next, we utilize a fully connected layer to map v' to the d-dimensional feature $g \in \mathbb{R}^d$:

$$g = \text{VGG}(v)W_t + b_t \tag{2}$$

where W_t and b_t are learnable weights in the fully connected layer.

Summary Model Initializing. The objective of the decoder is to generate a target summary $\hat{s} = \{\langle cls \rangle, \cdots, \hat{s}_i, \cdots, \langle sep \rangle\}$ with special tokens $\langle cls \rangle$, $\langle sep \rangle$, respectively. The visual modality g and textual modality h are concatenated together as multimodal inputs m for the decoder as:

$$m = g \oplus h \tag{3}$$

where $\oplus$ indicates the concatenation of two vectors. Afterward, the textual sequence is generated through a transformer-based decoder. The decoder takes the predicted tokens $s_{0;i-1}$ and fusion feature m as inputs and produces the current tokens as outputs:

$$s_i = \text{TransDec}(m, s_{0;i-1}) \tag{4}$$

where the notation $s_{0;i-1}$ represents the sequence of tokens preceding the i^{th} token.

During the summary task training phase, we apply negative log-likelihood for the target word sequence as the generation loss:

$$\mathcal{L}_{gen} = \frac{1}{I}\sum_{i=1}^{I}(-\log P(\hat{s}_i)) \tag{5}$$

where I stands for steps of decoding. And the model is trained with $\mathcal{L}_{gen}$ for T_1 epoch(s).

Stage 2: Image Selector Optimizing

The second phase is dedicated to the optimization of the image selector. Here, we define images that help improve the quality of summaries as summary-effective images. Intuitively, summary-ineffective images should not be used during the summarization process. Following the motivation above, the image selector tries to determine whether the image is summary-effective.

Image Selector Goal. We propose an image selector module to directly measure whether the image is

summary-effective $P(C|t,v)$. The core challenge of modeling $P(C|t,v)$ is defining the goal of optimizing the image selector. To bridge the gap between the role of images for humans and automatic summarization, we propose a self-labeling method to define the target of the selector. As shown in Figure 2, self-labeling obtains feedback from summaries when providing multimodal and unimodal information, respectively. Then it calculates the difference between the above two for the summary. The self-label result of the image selector is formulated as follows:

$$\mathrm{F}(g) = \mathrm{Re}(s, \hat{s}|h, g) - \mathrm{Re}(s, \hat{s}|h) \tag{6}$$

$$\hat{Q} = \begin{cases} 1, & \mathrm{F}\,(g) \leq 0 \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where $\mathrm{Re}(\cdot)$ is the reward of reflecting how similar the generated summary s is to the reference summary $\hat{s}$. Specifically, because $\hat{Q}$ is obtained at training for each batch, we expect $\mathrm{Re}(\cdot)$ to be both computationally efficient and reflective accurate. Li et al. (2018) point out that the accuracy of the first three words can reflect the overall accuracy of the summary. Therefore, $\mathrm{Re}(\cdot)$ is calculated with the edit distance of the first k tokens of the generated summary s and the reference summary $\hat{s}$. Based on comparing the reward of giving different inputs, we could obtain $\hat{Q}$, which stands for the label that determines whether v is summary-effective. Finally, the multimodal information (h, g) and the corresponding self-label $\hat{Q}$ are formed into a new triple $(h, g, \hat{Q})$, which is taken as training data for the image selector.

By feeding different source modalities to generate the summary and compare their feedback, the self-label can have a global view of the image contribution and determine whether it is helpful for generating a better summary.

Image Selector Module. Given the multimodal information (h, g), the goal of the image selector module is to predict $\hat{Q}$. We believe that the usefulness of images is related to the following aspects: 1) Whether the text-only information is sufficient to provide a summary; 2) Whether the generation of the final summary relies on multimodal information. Therefore, the image selector judge with two-dimensional information: text-only feature h_0 and multimodal feature (h_0, g). First, we project bi-modal features (h_0, g) to vectors of the same dimension with h_0.

$$x^M = \mathrm{Linear}(h_0 \oplus g) \tag{8}$$

Then we employ a two-layer MLP to classify the image:

$$q = \mathrm{MLP}(h_0 \oplus x^M) \tag{9}$$

where q is the prediction value of the image selector for whether the image v is summary-effective or not.

During the phase of optimizing the image selector (shown in Figure 2-Stage 2), we add a binary cross-entropy loss:

$$\mathcal{L}_{selc}(q, \hat{Q}) = \mathrm{BCE}(q, \hat{Q}) \tag{10}$$

By minimizing the cross-entropy loss, the image selector is able to have a global and prior view of the benefits the image brings for summary generation.

Training Objectives. To maintain the ability of generation, the summary task and the image selector are optimized together:

$$\mathcal{L} = \mathcal{L}_{gen} + \alpha \mathcal{L}_{selc}(q, \hat{Q}) \tag{11}$$

where α is the weight for image selector optimization. The optimization of the image selector includes the acquisition and fitting of the self-label $\hat{Q}$. As a result, the predictions of the image selector in this phase are not utilized for dynamic guiding. In addition, to ensure the image selector optimization process is reliable and stable, the training epoch T_2 in this stage is set to $T_2 > 1$. Accordingly, the self-label $\hat{Q}$ does not rely only on the last trained model from the first stage.

Stage 3: Image Dynamically Utilizing

The primary target of the third stage is to facilitate the model with the dynamic utilization of the image. After optimizing the image selector, it can provide guidance to utilize the image information dynamically.

Image Plugging. With prediction results from the image selector, the fixed model framework is unable to decide whether to send images to the decoder automatically. Therefore, we plug the image information into a temporary state with hard and soft guidance from the image selector.

First, the hard guidance assists in incorporating an image selectively based on the prediction category. The formula is mathematically represented as:

$$q^B = \begin{cases} 1, & q \geq 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

$$temp\text{-}s = q^B \cdot g + (1 - q^B) \cdot h_0 \tag{13}$$

where q^B indicates the binary classification prediction of the image selector: 1 denotes the image being summary-effective, and 0 indicates it is summary-ineffective. When the image is summary-ineffective, the text starting feature h_0 is plugged into the temporary state to guarantee the meaningfulness of $temp\text{-}s$.

Furthermore, the score value q also indicates the probability assigned by the selector to the usefulness of the image, which is the soft guidance. In other words, q serves as a quantitative measure of the image's utility for summarization. To maximize the potential of this signal, we simultaneously inform the decoder about the extent of image utility, building on the presumption that the image is indeed summary-effective. Equation 13 is replaced with:

$$temp\text{-}s = q \cdot q^B \cdot g + (1 - q^B) \cdot h_0 \tag{14}$$

By employing the aforementioned approach, we not only prevent summary-ineffective images from being fed to the decoder but also navigate the model in assigning appropriate weights to images during summary generation.

The temporary state $temp\text{-}s$ and textual modality h are concatenated together as multimodal inputs m' for the decoder. Thus for decoding, the equation 3 is replaced with:

$$m' = temp\text{-}s \oplus h \tag{15}$$

Through hard and soft guidance, the summary-effective images are dynamically fed into the decoder.

Training Objectives. To maintain the ability of the image selector, the summary task and the image selector are also optimized together. The self-labels obtained at the end of the second stage are employed as fixed labels $\hat{Q}^F$ of the image selector. The training loss is calculated as follows:

$$\mathcal{L} = \mathcal{L}_{gen} + \beta \mathcal{L}_{selc}(q, \hat{Q}^F) \tag{16}$$

where β is the weight for image selector optimization. It is worth mentioning that β is not a hyperparameter. In particular, compared to the generation task, the image selector task is relatively simpler, making the model prone to over-train the image selector to reduce the overall loss rapidly. Therefore, to avoid the model excessively optimizing the image selector while preserving its capabilities, we set β to decrease linearly with the current stage training epochs ep:

$$\beta = \alpha - \gamma * ep \tag{17}$$

where γ represents the linear decline rate. The training epoch in this stage is set to T_3.

By employing the multi-stage training strategy, on the one hand, we guarantee both the reliability and stability of image selector training. On the other hand, it allows the model to progressively acquire the ability to dynamically utilize source modalities for generating high-quality summaries. In the inference phase, the model propagates forward according to this stage. Therefore, the multi-stage training strategy also ensures the consistency of the model training and inference phases.

Experiment

Settings

We experiment with the MMS (Li et al. 2018) and MSMO (Zhu et al. 2018) datasets. MMS and MSMO contain 66,000 and 314,575 examples, respectively. Each sample in MMS is a triple of sentence, image, and summary, while the sample in MSMO is a triple of sentence, several images, and summary. Our task involves the input of a text and an image to generate a summary. As a result, our method retain a random image for each sample in the MSMO dataset and discard other images. We call this dataset MSMO (Single Img). Some statistical information is shown in Table 2.

We set both the text embedding dimension and the hidden dimension as 768. The batch size is set to 8. For texts in MMS dataset, we use the max text encoding length of 60, and the max text decoding length is 20. While for texts in MSMO dataset, we use the max text encoding length of 300, and the max text decoding length is 120. We use the BertAdam (Kingma and Ba 2014) optimizer and set the learning rate as $1e{-}4$, with the warmup portion as 0.1. When calculating the edit distance of the first k tokens of the generated and the reference summary, k is set to 5 and 8 for MMS and MSMO, respectively. The MSMO and MMS datasets have inconsistent scales and summary lengths, which leads to distinct parameter settings for each dataset during the multi-stage training process. For MMS dataset, the training epoch for three stages is $T_1 = 15, T_2 = 5, T_3 = 10$, and the learning weights of the image selector are $\alpha = 1, rt = 0.1$. For MSMO dataset, the training epoch for three stages is $T_1 = 8, T_2 = 2, T_3 = 10$, and the learning weights of the image selector are $\alpha = 0.5, rt = 0.05$. The base model trains 30 and 20 epochs for MMS and MSMO, respectively. In the test phase, we employ beam search and set the beam size as 4 to generate the summary.

Dataset	Subset	Size	Avg.Len (S)	Avg.Len (R)
MMS	train	62,000	21.68	7.72
	dev	2,000	24.35	7.68
	test	2,000	22.97	7.67
MSMO	train	293,964	720.87	70.12
	dev	10,355	766.08	70.02
	test	10,256	730.80	72.16

Table 2: Statistical information about datasets MMS and MSMO. "Avg.Len (S)" and "Avg.Len (R)" denote the average number of words in the source text (S) and reference summary (R), respectively.

Comparative Methods

Lead: Exploiting the first eight words as the summary.
SEASS (Zhou et al. 2017): It constructs a second-level sentence representation with a sentence encoder and a selective gate for summarization.
Doubly-Attn (Calixto, Liu, and Campbell 2017b): It employs dual attention mechanisms to narrow the gap between the image and the translation.
MAtt (Li et al. 2018): It proposes modality attention and image filtering for multimodal summarization.
CFSum (Xiao et al. 2023): It focuses on modeling different contributions of images for summarization and effectively enhances the multimodal representation for summarization.
PGN (See, Liu, and Manning 2017): It generates the current summary word by copying words from the source text or producing new words from the generator.
MSMG (Zhu et al. 2020): It proposes a multimodal objective function to close the distance between model output and multimodal reference.
MMRank (Zhu et al. 2021): It presents an unsupervised graph-based summarization model covering both single-modal and multimodal output summarization.
BertAbs: It is our base model with a text encoder BERT, an image encoder VGG and a transformer decoder.
BertAbs-txt: BertAbs is fed only with textual modality.

Automatic Evaluation Results

Our methods are reported with seven automatic metrics, including ROUGE-1, ROUGE-2, ROUGE-L (Lin and Hovy 2002), BLEU (Papineni et al. 2002), BERTScore (Zhang* et al. 2020), MoverScore (Zhao et al. 2019), and edit distance of the first 5 tokens denoted as ED-top5. More details of evaluation scripts are given in the appendix.

Results On MMS. We compare our work with our baselines and other work on the MMS dataset. Table 3 shows the results of different models. The results show that **BertAbs**

Method	R-1 ↑	R-2 ↑	R-L ↑	BLEU ↑	BS ↑	MS ↑	ED ↓
Lead	33.30	13.27	31.55	39.37	83.94	20.47	4.20
SEASS △	44.86	23.03	41.92	-	-	-	-
Doubly-Attn △	41.11	21.75	39.92	-	-	-	-
MAtt	44.50	23.37	41.85	43.50	86.42	30.75	3.75
CFSum	47.86	25.64	44.64	48.83	86.98	32.36	3.67
BertAbs-txt	49.14	26.14	46.05	48.99	87.12	34.74	3.62
BertAbs	49.20	26.35	46.21	49.28	87.17	34.76	3.63
DIUSum	**50.39***	**27.34***	**47.33***	**50.68***	**87.36***	**35.28***	**3.59***

Table 3: Experimental results on MMS. "△" marks the results from Li et al. (2018). "BS, MS, ED" represent BERTScore, MoverScore, and edit distance, respectively. The ↑ indicates that the higher value of the indicator is better, while ↓ indicates the opposite. "*" indicates the model performs significantly better than the BertAbs by the 95% confidence interval ($p < 0.05$).

Datasets	Method	R-1 ↑	R-2 ↑	R-L ↑
MSMO (Multi Imgs)	Lead	39.35	17.01	32.54
	PGN	39.76	18.19	36.51
	CFSum	38.62	16.05	36.19
	MSMG	41.16	18.35	37.85
	MMRank	41.72	17.33	-
MSMO (Single Img)	BertAbs-txt	41.92	19.40	38.99
	BertAbs	41.85	19.33	38.94
	DIUSum	**42.23**	**19.83**	**39.34**

Table 4: Experimental results on MSMO.

performs comparably with **BertAbs-txt**, indicating that utilizing the image information uniformly could not help improve the generated summary quality. **DIUSum** builds on **BertAbs** and introduces an image selector to dynamically utilize different source modalities. Generally, our method **DIUSum** outperforms the baselines **BertAbs** and **BertAbs-txt**. And it achieves 1.19 higher points on ROUGE-1 than **BertAbs**, which achieves SOTA on MMS dataset. This proves that image selector plays a significant role in multimodal summarization. In addition, **DIUSum** has the lowest ED-top5 among all methods, which indicates its strength in generating the first few words of the summary.

Results On MSMO. To further validate our method, we experiment on the MSMO dataset shown in Table 4. The results show that **BertAbs-txt** is even better than **BertAbs**, suggesting that most of the images in the MSMO (Single Img) dataset negatively impact the performance of **BertAbs**. In contrast, **DIUSum** obtains 0.38 higher points on ROUGE-1 than **BertAbs**. This proves that **DIUSum** can generalize to the dataset where images are harmful. Furthermore, it is worth mentioning that **DIUSum** exploits only one image in the MSMO dataset, which surpasses other methods using multiple images. This further demonstrates the significance of providing more summary-effective images to the model. However, our method is not significantly improved on MSMO (Single Img) due to some (image, summary) pairs lacking strong correlations, thereby limiting the performance boost of **DIUSum**.

Ablation Study

Ablation Study. To investigate the effectiveness of different components, we further compare DIUSum with several variants in Table 5.

Effectiveness of Image Selector. In the test phase, we test each example with text-only (#1b/#2b) and multimodal inputs (#1a/#2a) to generate summaries, respectively, and select the higher ROUGE-1 score as the dynamic-input results (#1c/#2c). The percentage of #1c/#2c in Table 5 represents the improvement of the dynamic result compared to the multimodal input. Compared with **BertAbs**, it can be observed that 1) the gap between "dynamic-input" and "multi-input" in **DIUSum** is smaller. This indicates that our method correctly exploits source modalities leading to better results. 2) The "text-input" performance of #2b also shows an improvement compared to #1b. This is ascribed to the fact that during the training process in the third stage, we restrict the model from using summary-ineffective images, which maximize the utilization of textual information.

Effectiveness of Hard/Soft Guidance. Removing the soft or hard guidance results in model #3a/#3b. Comparing #2a, #3a, and #3b shows that: 1) Plugging images without soft guidance impairs the model performance. This may ascribe to the disparity in representation with only 0 or 1 hard guidance. 2) Removing hard guidance leads to performance drops, which implies that soft guidance cannot completely eliminate the influence of invaluable images. In summary, combining soft and hard guidance produces the best results.

Necessity of Multi-Stage. Model #4a removes the initialization of the first stage. Model #4b merges Stage 2 and Stage 3, in which the image selector is optimized and utilized simultaneously. The drastic performance drop of model #4a proves that it is vital to optimize DIUSum with model initialization. Model #4b performs worse than **BertAbs**. This is because the training and guiding of the im-

#	Method	R-1
1a	BertAbs (multi-input)	49.20
1b	w/ text-input	49.33
1c	w/ dynamic-input	50.01 (+1.64%)
2a	DIUSum (multi-input)	50.39
2b	w/ text-input	49.98
2c	w/ dynamic-input	50.53 (+0.27%)
3a	w/o soft guidance	49.88
3b	w/o hard guidance	50.09
4a	w/o Stage 1	49.44
4b	merge Stage 2 & 3	48.88

Table 5: Ablation study about DIUSum on MMS.

age selector appear at the same stage, resulting in the image selector optimizing oscillations and failing to converge. In brief, both ablations prove that applying the multi-stage training strategy is necessary.

Analysis for Image Selector

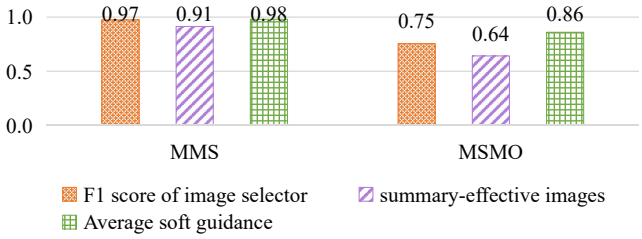

Figure 3: Statistics about the prediction of image selector.

Although the image selector improves the performance of the model in terms of metrics, we still wish to know how the image selector performs on different datasets. As depicted in Figure 3, the orange box represents the F1 score of the image selector, where ground truth labels are acquired by comparing text-input and multi-input results to deduce image effectiveness. The purple and green boxes represent the percentage of summary-effective images, and the average value of the soft guidance for summary-effective images in the test set, respectively. First, the 0.97 and 0.75 F1 scores of the image selector on two datasets confirm that the image selector is capable of predicting summary-effective images. Second, it can be observed that the percentage of summary-effective images in the MMS dataset is greater than that in the MSMO dataset. Third, the image selector assigns lower soft guidance to images in the MSMO dataset. This discrepancy is attributed to the construction process of the datasets: For the MSMO dataset, the relevant images are automatically retrieved with the source text, and many of these images may be irrelevant to the final summary. In contrast, the MMS dataset is created by manually labeling the most matching images. Consequently, a higher percentage of images in the

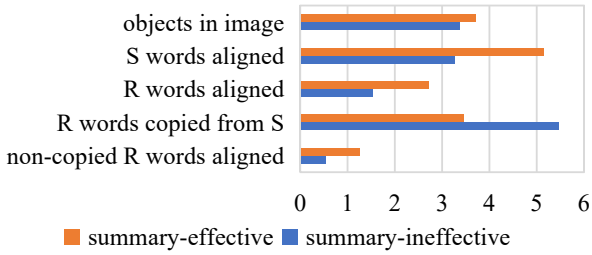

Figure 4: The number of ablated features. "S" and "R" denote the source text and reference summary, respectively.

MSMO dataset are classified as summary-ineffective by the image selector, and lower scores are assigned to the image usage. This explains the limited performance improvement of DIUSum on the MSMO dataset. Additionally, it demonstrates that the image selector can reflect the improved level of summary quality brought by the images.

Furthermore, we aim to understand the characteristics of the summary-effective and summary-ineffective images. Figure 4 plots the average number of ablated features of the two types of images, with 30 randomly selected images from each category. The following conclusions could be safely drawn through the fined-grained analysis: (1) An object detection tool identifies objects in the image with confidence over 0.8. Notably, summary-ineffective images contain fewer objects, offering limited guidance for summarization. (2) We manually align the image with words in the source text (or reference summary). Fewer words align with the summary-ineffective images. It indicates that summary-ineffective images have less correlation with the text. (3) In the summary-ineffective image sample, more words in the summary are copied from the source text, which illustrates that the effectiveness of an image is related to the abstractness of text generation. (4) Fewer non-copied reference words align with images in the summary-ineffective image samples (0.53 on average). This confirms that if the image merely visually represents copied reference words, the text modality often suffices to generate the summary without the help of the images. Across multiple dimensions, summary-effective images show favorable trends, confirming the rationality of our approach.

Conclusion

Based on the observation that existing multimodal summarization models cannot meet the individual needs of different source modalities, this paper focuses on dynamically utilizing image information for summarization. Therefore, we propose a novel framework DIUSum to select and utilize valuable images for summarization. The core module of DIUSum is the image selector, which selects summary-effective images and guides the incorporation of multimodal information for the decoder. Experimental results have shown that DIUSum can improve the quality of the summary. Furthermore, fine-grained analysis demonstrates that the image selector can reflect the improved level of summary quality brought by the images.

Acknowledgments

The research work has been supported by the Natural Science Foundation of China under Grant No. 62106263.

References

Calixto, I.; Liu, Q.; and Campbell, N. 2017a. Doubly-Attentive Decoder for Multi-modal Neural Machine Translation. In *Annual Meeting of the Association for Computational Linguistics*.

Calixto, I.; Liu, Q.; and Campbell, N. 2017b. 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)*, 1913–1924. Vancouver, Canada: Association for Computational Linguistics.

Fu, X.; Wang, J.; and Yang, Z. 2020. Multi-modal summarization for video-containing documents. *arXiv preprint arXiv:2009.08018*.

Im, J.; Kim, M.; Lee, H.; Cho, H.; and Chung, S. 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)*, 388–403. Online: Association for Computational Linguistics.

Jangra, A.; Saha, S.; Jatowt, A.; and Hasanuzzaman, M. 2021. Multi-Modal Supplementary-Complementary Summarization Using Multi-Objective Optimization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, 818–828. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380379.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, H.; Zhu, J.; Liu, T.; Zhang, J.; and Zong, C. 2018. Multi-modal Sentence Summarization with Modality Attention and Image Filtering. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 4152–4158. International Joint Conferences on Artificial Intelligence Organization.

Li, H.; Zhu, J.; Zhang, J.; He, X.; and Zong, C. 2020. Multimodal Sentence Summarization via Multimodal Selective Encoding. In *Proceedings of the 28th International Conference on Computational Linguistics*, 5655–5667. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Li, J.; Zhang, Z.; Wang, B.; Zhao, Q.; and Zhang, C. 2022. Inter- and Intra-Modal Contrastive Hybrid Learning Framework for Multimodal Abstractive Summarization. *Entropy*, 24(6).

Liang, Y.; Meng, F.; Xu, J.; Wang, J.; Chen, Y.; and Zhou, J. 2023. Summary-Oriented Vision Modeling for Multimodal Abstractive Summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2934–2951. Toronto, Canada: Association for Computational Linguistics.

Libovický, J.; and Helcl, J. 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)*, 196–202. Vancouver, Canada: Association for Computational Linguistics.

Lin, C.-Y.; and Hovy, E. 2002. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization*, 45–51. Phildadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Mahasseni, B.; Lam, M.; and Todorovic, S. 2017. Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 202–211.

Mukherjee, S.; Jangra, A.; Saha, S.; and Jatowt, A. 2022. Topic-aware Multimodal Summarization. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, 387–398. Online only: Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Qiu, J.; Zhu, J.; Xu, M.; Dernoncourt, F.; Bui, T.; Wang, Z.; Li, B.; Zhao, D.; and Jin, H. 2022. MHMS: Multimodal Hierarchical Multimedia Summarization. arXiv:2204.03734.

Sanabria, R.; Caglayan, O.; Palaskar, S.; Elliott, D.; Barrault, L.; Specia, L.; and Metze, F. 2018. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*.

See, A.; Liu, P. J.; and Manning, C. D. 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)*, 1073–1083. Vancouver, Canada: Association for Computational Linguistics.

Wang, J.; Wang, W.; Wang, Z.; Wang, L.; Feng, D.; and Tan, T. 2019. Stacked memory network for video summarization. In *Proceedings of the 27th ACM international conference on multimedia*, 836–844.

Xiao, M.; Zhu, J.; Lin, H.; Zhou, Y.; and Zong, C. 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)*, 8538–8553. Toronto, Canada: Association for Computational Linguistics.

Yu, T.; Dai, W.; Liu, Z.; and Fung, P. 2021. Vision Guided Generative Pre-trained Language Models for Multimodal Abstractive Summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3995–4007. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Zhang, L.; Zhang, X.; Guo, Z.; and Liu, Z. 2023. CISum: Learning Cross-modality Interaction to Enhance Multimodal Semantic Coverage for Multimodal Summarization.

In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, 370–378. SIAM.

Zhang, L.; Zhang, X.; and Pan, J. 2022. Hierarchical Cross-Modality Semantic Correlation Learning Model for Multimodal Summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10): 11676–11684.

Zhang, L.; Zhang, X.; Pan, J.; and Huang, F. 2022. Hierarchical Cross-Modality Semantic Correlation Learning Model for Multimodal Summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11676–11684.

Zhang*, T.; Kishore*, V.; Wu*, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Zhang, Z.; Wang, J.; Sun, Z.; and Yang, Z. 2021. LAMS: A Location-aware Approach for Multimodal Summarization (Student Abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(18): 15949–15950.

Zhao, W.; Peyrard, M.; Liu, F.; Gao, Y.; Meyer, C. M.; and Eger, S. 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)*, 563–578. Hong Kong, China: Association for Computational Linguistics.

Zhou, Q.; Yang, N.; Wei, F.; and Zhou, M. 2017. Selective Encoding for Abstractive Sentence Summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1095–1104. Vancouver, Canada: Association for Computational Linguistics.

Zhu, J.; Li, H.; Liu, T.; Zhou, Y.; Zhang, J.; and Zong, C. 2018. MSMO: Multimodal Summarization with Multimodal Output. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4154–4164. Brussels, Belgium: Association for Computational Linguistics.

Zhu, J.; Xiang, L.; Zhou, Y.; Zhang, J.; and Zong, C. 2021. Graph-Based Multimodal Ranking Models for Multimodal Summarization. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(4).

Zhu, J.; Zhou, Y.; Zhang, J.; Li, H.; Zong, C.; and Li, C. 2020. Multimodal Summarization with Guidance of Multimodal Reference. In *AAAI Conference on Artificial Intelligence*.