Exploring the Trade-Off within Visual Information for MultiModal Sentence Summarization

Minghuan Yuan
Institute of Information Engineering, Chinese Academy of Sciences
School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
yuanminghuan@iie.ac.cn

Shiyao Cui*
Institute of Information Engineering, Chinese Academy of Sciences
School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
cuishiyao@iie.ac.cn

Xinghua Zhang
Institute of Information Engineering, Chinese Academy of Sciences
School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
zhangxinghua@iie.ac.cn

Shicheng Wang
Institute of Information Engineering, Chinese Academy of Sciences
School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
wangshicheng@iie.ac.cn

Hongbo Xu
Institute of Information Engineering, Chinese Academy of Sciences
School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
hbxu@iie.ac.cn

Tingwen Liu*
Institute of Information Engineering, Chinese Academy of Sciences
School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
liutingwen@iie.ac.cn

ABSTRACT

MultiModal Sentence Summarization (MMSS) aims to generate a brief summary based on the given source sentence and its associated image. Previous studies on MMSS have achieved success by either selecting the **task-relevant** visual information or filtering out the **task-irrelevant** visual information to help the textual modality to generate the summary. However, enhancing from a single perspective usually introduces **over-preservation** or **over-compression** problems. To tackle these issues, we resort to Information Bottleneck (IB), which seeks to find a maximally compressed mapping of the input information that preserves as much information about the target as possible. Specifically, we propose a novel method, $\mathbf{T}^3$, which adopts IB to balance the **T**rade-off between **T**ask-relevant and **T**ask-irrelevant visual information through the variational inference framework. In this way, the task-irrelevant visual information is compressed to the utmost while the task-relevant visual information is maximally retained. With the holistic perspective, the generated summary could maintain as many key elements as possible while discarding the unnecessary ones as far as possible. Extensive experiments on the representative MMSS dataset demonstrate the superiority of our proposed method. Our code is available at https://github.com/YuanMinghuan/T3.

CCS CONCEPTS

• **Information systems** → **Summarization**; **Multimedia information systems**; • **Mathematics of computing** → **Information theory**.

*Corresponding author

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07.
https://doi.org/10.1145/3626772.3657753

KEYWORDS

Summarization, Multimodality, Information Bottleneck

ACM Reference Format:
Minghuan Yuan, Shiyao Cui, Xinghua Zhang, Shicheng Wang, Hongbo Xu, and Tingwen Liu. 2024. Exploring the Trade-Off within Visual Information for MultiModal Sentence Summarization. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3626772.3657753

1 INTRODUCTION

With the shift from unimodal towards multimodal in both academia and industry in recent years [18, 36], more research on MultiModal Summarization (MMS) has emerged [15, 20]. As an extension of Textual Summarization, MMS expands the breadth of summarization by incorporating multiple modalities, such as images, leading to a wider range of applications [20]. As the foundation of MMS, MultiModal Sentence Summarization (MMSS) aims to generate a brief summary based on the given source sentence and its associated image [31], which has drawn growing research attention.

As the saying goes, *an image is worth a thousand words*, which emphasizes the significant value of visual information to enhance the textual semantics. However, only a small part of the 'thousand words' about the image holds great relevance to the target summary in MMSS, while the rest is irrelevant. Therefore, the visual information in the image could be divided into two categories: **task-relevant** and **task-irrelevant**. Specifically, the task-relevant information refers to the visual elements mentioned in the target summary, thus such information should be included in the generated summary. Meanwhile, the task-irrelevant information implies the visual background or attributes that should be excluded from the target summary optionally.

Unluckily, previous studies on MMSS either devote to selecting the task-relevant visual information [33, 40] or focus on filtering out the task-irrelevant visual information [31, 59]. Despite their success, these methods are exposed to the issues of **over-preservation** and

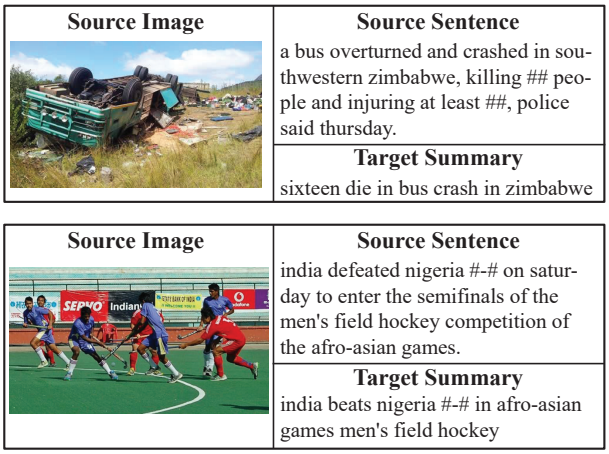


Figure 1: Examples of MultiModal Sentence Summarization.

over-compression since they are merely designed from a single perspective. Specifically, over-preservation refers to the excessive retention of some unnecessary details, leading to the presence of redundant textual terms in the generated summary. For example, in the top half of Fig. 1, the visual attribute 'overturned' depicting the bus may result in the redundant term 'overturn' in the generated summary. Over-compression may occur when the decisive visual elements are compressed with the task-irrelevant visual information at the same time, resulting in the lack of significant details in the generated summary. Taking the bottom half of Fig. 1 as an instance, the visual elements 'male player' and 'hockey stick' may be compressed with the visual background simultaneously.

To tackle the above issues, Information Bottleneck (IB) is leveraged to find a maximally compressed mapping of the input image that preserves sufficient information about the target summary. Accordingly, we propose a novel method termed as $\mathbf{T}^3$, which adopts IB to balance the **T**rade-off between **T**ask-relevant and **T**ask-irrelevant visual information from a holistic perspective. With the two-fold goal above, two mutual information terms in IB cooperate with each other correspondingly. Specifically, one term aims to compress the task-irrelevant visual information, and the other intends to preserve the task-relevant visual information. These two terms are optimized collaboratively as a whole to avoid the potential problems of over-preservation and over-compression. In this way, we obtain expressive visual representations with minimal sufficient predictive power to perform MMSS.

Overall, our main contributions are as follows:

- To the best of our knowledge, we take the lead in considering the task-relevant and task-irrelevant visual information from a holistic perspective in MMSS.
- We propose a novel method, $\mathbf{T}^3$, which explores the trade-off between the task-relevant and task-irrelevant visual information under Information Bottleneck.
- Extensive experiments on the representative MMSS dataset show that our proposed method significantly outperforms the competitive baseline methods.

2 RELATED WORK

2.1 MultiModal Summarization

With increasing research on multimodality [18, 36], MultiModal Summarization (MMS) has attracted great attention [15, 20]. MMS takes multimodal content, typically text and images, as input. According to the output modality, MMS can be divided into two taxonomies. One is only textual output [10, 30–32, 52]. The other is multimodal output, usually text and images [3, 4, 9, 35, 49, 57, 60, 61, 65, 71], sometimes video is added as well [19, 21]. As the foundation of MMS, MultiModal Sentence Summarization (MMSS) aims to generate a brief summary based on the source sentence and its associated image [31], which has drawn increasing attention from researchers.

Existing MMSS studies could be roughly divided into two series with respect to how the source image helps to enhance the textual modality. One series of methods focus on compressing the task-irrelevant visual information for summary generation. Specifically, Li et al. [31] introduced an image filtering mechanism with two types of image filters to strain the task-irrelevant visual information out. Xiao et al. [59] proposed a Coarse-to-Fine contribution network to eliminate the interference of useless visual information. The other series of researches concentrate on preserving the task-relevant visual information. Li et al. [33] proposed a multimodal selective gate network that considers reciprocal relationships between textual and multilevel visual features. Lin et al. [40] proposed a prompt-guided image encoding module and an explicit source critical token learning module to capture the critical to-be-summarized information. In addition to the above two series of methods, Jing et al. [22] introduced an additional pre-training stage, and Yong et al. [62] utilized a larger language model as the backbone to further boost the performance of MMSS.

Different from the two major series of methods, we take a holistic perspective to balance the trade-off between compressing the task-irrelevant visual information and preserving the task-relevant visual information collaboratively, thereby alleviating the potential impacts of over-preservation and over-compression.

2.2 Information Bottleneck

As an important part of information theory, Information Bottleneck (IB) was first proposed by Tishby et al. [55], and was first introduced into deep learning by Tishby and Zaslavsky [56]. On the basis of Tishby and Zaslavsky [56], Shwartz-Ziv and Tishby [54] further analyzed the training dynamics, learning processes, and internal representations in deep learning. Then, Alemi et al. [1] first proposed a variational approximation to IB named Deep Variational Information Bottleneck, which allows us to parameterize the IB using Deep Neural Networks and leverage the reparameterization trick [24] for efficient training. Federici et al. [17] further extended the IB to the unsupervised multi-view setting. Kawaguchi et al. [23] mathematically related IB to generalization errors to provide the first rigorous theory justifying the benefits of IB in deep learning.

IB has shown its sparkles in various areas such as Computer Vision [2, 45], Natural Language Processing [44, 64, 69], and Recommendation Systems [7]. Despite the tremendous achievements above, to the best of our knowledge, IB has not yet been explored in MMSS, which is challenging and deserves the research attention.

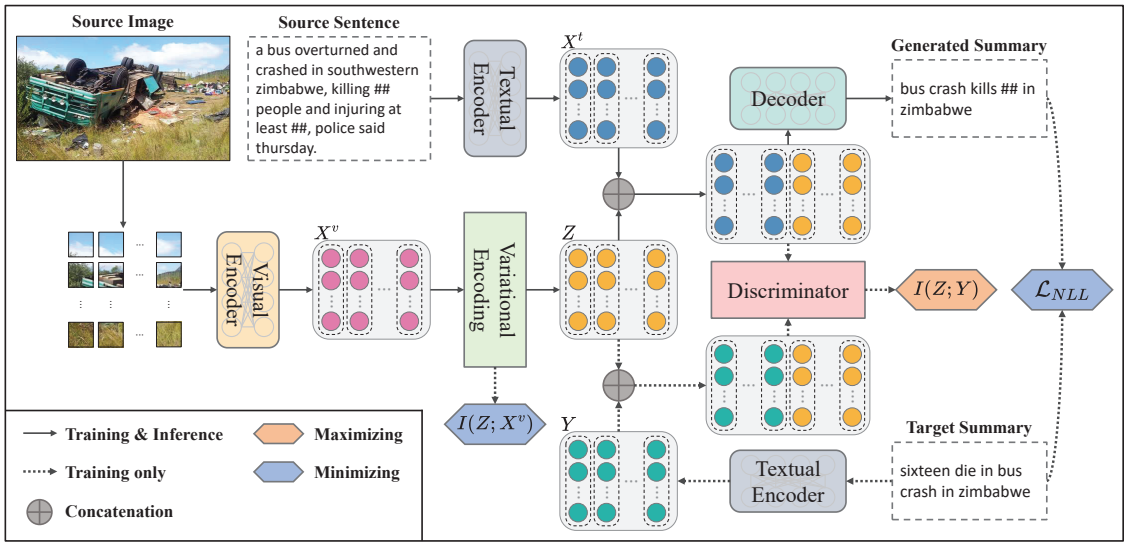


Figure 2: The workflow of our proposed method.

3 PRELIMINARY

3.1 Task Formulation

Based on the source sentence and its associated image, MMSS aims to generate a brief summary. Formally, given the source sentence $X^t = (x_1^t, x_2^t, ..., x_n^t)$ which contains n tokens and its associated image X^v, we aim to learn a generator $\mathcal{G}$ to produce a brief summary $Y = (y_1, y_2, ..., y_l)$ with l tokens:

$$Y = \mathcal{G}_\Theta(X^t, X^v), \tag{1}$$

where Θ denotes the set of learnable parameters of the summary generator $\mathcal{G}$.

3.2 Information Bottleneck Principle

Information Bottleneck (IB) aims to find a maximally compressed mapping of the input random variable X that preserves as much information as possible about the output random variable Y [1, 55, 56]. Formally, IB minimizes the following objective function:

$$\mathcal{L}_{IB} = I(Z;X) - \beta I(Z;Y), \tag{2}$$

where $I(\cdot;\cdot)$ denotes the mutual information of two random variables, Z is the approximated minimal sufficient statistics of X with respect to Y. $\beta > 0$ is the Lagrange multiplier which balances the trade-off between **minimizing** $I(Z;X)$ and **maximizing** $I(Z;Y)$, so that Z could compress the irrelevant information of X while preserving sufficient information towards Y.

4 METHOD

4.1 Overview

In this section, we detail the proposed method, $\mathbf{T}^3$, whose workflow is shown in Fig. 2. Specifically, $\mathbf{T}^3$ consists of five modules: Source Sentence Encoding, Source Image Encoding, Trade-Off within Visual Information, Summary Generation, and Model Training. Note that in the Trade-Off within Visual Information module, the Task-irrelevant Visual Information Compressing submodule and the Task-relevant Visual Information Preserving submodule work collaboratively under Information Bottleneck, driving the visual representations with minimal sufficient predictive ability for MMSS.

4.2 Source Sentence Encoding

Source sentence encoding aims to transform the discrete tokens into continuous representations. We adopt the BART [28] encoder to derive the contextual representations of the source sentence due to its prominent performance in text generation tasks.

Following Lewis et al. [28], two special tokens, <s> and </s>, are appended to the beginning and the end of the source sentence respectively. Formally, let $X^t = (x_0^t, x_1^t, x_2^t, ..., x_n^t, x_{n+1}^t)$ denote the source sentence, where x_0^t and x_{n+1}^t are the two special tokens. We feed X^t into the BART encoder $\mathcal{E}_t$ and obtain the source sentence representation $\boldsymbol{X}^t$ as follows:

$$\boldsymbol{X}^t = \mathcal{E}_t(X^t), \tag{3}$$

where $\boldsymbol{X}^t = (\boldsymbol{x}_0^t, \boldsymbol{x}_1^t, \boldsymbol{x}_2^t, ..., \boldsymbol{x}_n^t, \boldsymbol{x}_{n+1}^t)$, $\boldsymbol{x}_i^t \in \mathbb{R}^{d_t}$ denotes the representation of the i-th token, and d_t is the dimension of the token representation.

4.3 Source Image Encoding

Source image encoding utilizes ViT [14] to transform the numerous pixel values in the source image into few continuous representations, where the following three steps are involved.

First, in order to adapt the input format of ViT, we follow Dosovitskiy et al. [14] to flatten the 2D image $X^v \in \mathbb{R}^{H\times W\times C}$ into a sequence of patches $X_{seq}^v = (x_1^v, x_2^v, ..., x_m^v) \in \mathbb{R}^{m\times(P^2\times C)}$, where H and W are the height and width of the image resolution respectively, C is the number of channels (usually equal to three), P^2 is the resolution of each patch, and $m = H \times W/P^2$ is the resulting number of patches.

In the following, we add positional information to preserve the structural relationship between patches. Specifically, similar to BERT [13], we first pre-append a learnable embedding $\boldsymbol{e}_0^v \in \mathbb{R}^{d_e}$ which represents the whole image. Then, the position embeddings $\boldsymbol{E}_{pos} \in \mathbb{R}^{(m+1)\times d_e}$ are added to acquire the image embeddings $\boldsymbol{E}^v$ as follows:

$$\boldsymbol{E}^v = (\boldsymbol{e}_0^v, \boldsymbol{W}_e \boldsymbol{x}_1^v, \boldsymbol{W}_e \boldsymbol{x}_2^v, ..., \boldsymbol{W}_e \boldsymbol{x}_m^v) + \boldsymbol{E}_{pos}, \tag{4}$$

where $\boldsymbol{W}_e \in \mathbb{R}^{d_e \times (P^2 \times C)}$ denotes the linear transformation matrix for the dimensional alignment, and d_e is the embedding dimension.

Finally, we feed the image embeddings $\boldsymbol{E}^v$ into ViT to obtain the image representations $\boldsymbol{X}^v$ as follows:

$$\boldsymbol{X}^v = \mathcal{E}_v(\boldsymbol{E}^v), \tag{5}$$

where $\boldsymbol{X}^v = (\boldsymbol{x}_0^v, \boldsymbol{x}_1^v, \boldsymbol{x}_2^v, ..., \boldsymbol{x}_m^v)$, $\boldsymbol{x}_j^v \in \mathbb{R}^{d_v}$ denotes the representation of the j-th patch, $\mathcal{E}_v$ denotes ViT, and d_v is the dimension of the patch representation.

4.4 Trade-Off within Visual Information

To avoid the potential over-preservation and over-compression problems, we expect to maximally compress the task-irrelevant visual information while preserving the task-relevant visual information as much as possible. To this end, we resort to Information Bottleneck (IB) to balance the trade-off between the aforementioned two types of visual information. Formally, we aim to minimize the IB-based objection function as follows:

$$\mathcal{L}_{IB} = \alpha I(Z; \boldsymbol{X}^v) - \beta I(Z; Y), \tag{6}$$

where Z denotes the minimal sufficient statistics of the source image $\boldsymbol{X}^v$ with respect to the target summary Y. α, β are the hyperparameters controlling the extent of compressing the task-irrelevant visual information as well as preserving the task-relevant visual information, respectively.

Since the traditional IB cannot perform general gradient back propagation, we utilize variational encoding to apply IB to deep neural networks. Moreover, since accurate mutual information estimation is intractable, we optimize the above objective function by minimizing the upper bound of $I(Z; \boldsymbol{X}^v)$ and maximizing the lower bound of $I(Z; Y)$ via variational approximation, respectively. In the following, we will first introduce the variational encoding of the source image, and then detail the approximation of the two mutual information terms.

4.4.1 Variational Encoding. In order to apply the information bottleneck to deep neural networks, we follow Alemi et al. [1] to obtain the random variable of the source image with variational inference [24]. To be specific, for the j-th patch of the source image, its random variable $\boldsymbol{z}_j$ is obtained as follows:

$$\begin{aligned} \boldsymbol{\mu}_j &= \boldsymbol{W}_2^{\mu}(\varphi(\boldsymbol{W}_1^{\mu}\boldsymbol{x}_j^v + \boldsymbol{b}_1^{\mu})) + \boldsymbol{b}_2^{\mu}, \\ \boldsymbol{\sigma}_j &= \boldsymbol{W}_2^{\sigma}(\varphi(\boldsymbol{W}_1^{\sigma}\boldsymbol{x}_j^v + \boldsymbol{b}_1^{\sigma})) + \boldsymbol{b}_2^{\sigma}, \\ \boldsymbol{z}_j &\sim \mathcal{N}(\boldsymbol{\mu}_j, [diag(\boldsymbol{\sigma}_j)]^2), \end{aligned} \tag{7}$$

where $\boldsymbol{W}_1^{\mu}, \boldsymbol{W}_2^{\mu}, \boldsymbol{W}_1^{\sigma}, \boldsymbol{W}_2^{\sigma}$ and $\boldsymbol{b}_1^{\mu}, \boldsymbol{b}_2^{\mu}, \boldsymbol{b}_1^{\sigma}, \boldsymbol{b}_2^{\sigma}$ are learnable parameters, $\varphi(\cdot)$ denotes the activation function, $\boldsymbol{\mu}_j \in \mathbb{R}^{d_v}$ and $\boldsymbol{\sigma}_j \in \mathbb{R}^{d_v}$ are the mean and variance of the Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_j, [diag(\boldsymbol{\sigma}_j)]^2)$, and $\boldsymbol{z}_j$ is sampled from the above Gaussian distribution.

Since the general sampling process is not differentiable for gradient back propagation, we leverage the reparameterization trick [24] to derive $\boldsymbol{z}_j$ as follows:

$$\boldsymbol{z}_j = \boldsymbol{\mu}_j + \boldsymbol{\sigma}_j \odot \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, diag(\boldsymbol{I})), \tag{8}$$

where $\boldsymbol{\epsilon} \in \mathbb{R}^{d_v}$ is sampled from the standard Gaussian distribution, and $\odot$ denotes the Hadamard product.

Note that since $\boldsymbol{x}_0^v$ serves as the aggregated representation of the whole image, we utilize its sampled representation $\boldsymbol{z}_0$ as the variational representation of the source image $Z \in \mathbb{R}^{1\times d_v}$.

4.4.2 Task-irrelevant Visual Information Compression. As $I(Z; \boldsymbol{X}^v)$ measures the quantity of information contained in Z about $\boldsymbol{X}^v$, we minimize its upper bound to compress the task-irrelevant visual information. Specifically, we estimate the upper bound of $I(Z; \boldsymbol{X}^v)$ through Kullback-Leibler divergence with variational approximation [48] as follows:

$$I(Z; \boldsymbol{X}^v)_{UpperBound} = \mathbb{E}_{\boldsymbol{X}^v \sim p(\boldsymbol{X}^v)} \left[\mathbb{D}_{KL} \left[q(Z|\boldsymbol{X}^v) || r(Z) \right] \right], \tag{9}$$

where $q(Z|\boldsymbol{X}^v)$ denotes the approximated posterior distribution calculated by the variational encoding, and $r(Z)$ is the approximation of the marginal distribution $p(Z)$.

Following Kingma and Welling [24], we assume $r(Z)$ as the standard Gaussian distribution. Consequently, the upper bound of $I(Z; \boldsymbol{X}^v)$ can be reformulated as follows:

$$\begin{aligned} &I(Z; \boldsymbol{X}^v)_{UpperBound} = \\ &\mathbb{E}_{p(\boldsymbol{X}^v)} \left[\mathbb{D}_{KL} \left[\mathcal{N}(\boldsymbol{\mu}, [diag(\boldsymbol{\sigma})]^2) || \mathcal{N}(\boldsymbol{0}, diag(\boldsymbol{I})) \right] \right], \end{aligned} \tag{10}$$

where $\mathcal{N}(\boldsymbol{\mu}, [diag(\boldsymbol{\sigma})]^2)$ is derived from $q(Z|\boldsymbol{X}^v)$ through variational encoding, and $\mathcal{N}(\boldsymbol{0}, diag(\boldsymbol{I}))$ refers to the standard Gaussian distribution.

4.4.3 Task-relevant Visual Information Preservation. As $I(Z; Y)$ measures the quantity of information contained in Z about Y, we maximize its lower bound to preserve the task-relevant visual information. Following Velickovic et al. [58], a discriminator is built to estimate the lower bound of $I(Z; Y)$ as follows:

$$\begin{aligned} I(Z; Y)_{LowerBound} = &\mathbb{E}_{Z \sim q(Z|\boldsymbol{X}^v), Y \sim p(Y)} \log D(Z, Y) \\ &+ \mathbb{E}_{Z \sim q(Z|\boldsymbol{X}^v), \tilde{Y} \sim p(\tilde{Y})} \log(1 - D(Z, \tilde{Y})), \end{aligned} \tag{11}$$

where D denotes the discriminator which measures the consistency between the learned visual information and the target summary. (Z, Y) refers to the representation of the positive sample which is constructed by the pair of (source image, target summary), i.e. (X^v, Y). Meanwhile, $(Z, \tilde{Y})$ refers to the representation of the negative sample, where $\tilde{Y}$ is different from Y.

Here, we detail the construction of the negative samples. Intuitively, negative samples could be constructed via the in-batch random negative sampling. For the i-th positive sample (X_i^v, Y_i), the in-batch random negative sampling replaces Y_i with $Y_j (j \neq i)$, which is the target summary of different instances in the same batch. So the negative sample is built as (X_i^v, Y_j). For each positive sample, the in-batch random negative sampling can be operated multiple times, resulting in multiple negative samples for the positive sample. However, since Y_j expresses completely different semantics from Y_i, it is difficult for the model to learn how to distinguish between the source sentence and the target summary, which is required for the summary generation. Hence, we explore the hard negative

sampling strategy, where the source sentence serves to build the negative sample (X^v, X^t).

To obtain the representations of the positive and negative samples, we use the variational encoding described in Sec. 4.4.1 to encode the image part and the same encoder as in Eq. (3) to encode Y and $\tilde{Y}$. Since the shape of $\boldsymbol{Y} \in \mathbb{R}^{(l+1)\times d_t}$ varies with l, we employ the pooling operation to simplify the calculation of the discriminator as follows:

$$D(\boldsymbol{Z}, \boldsymbol{Y}) = \varphi(\boldsymbol{W}(\boldsymbol{Z} \oplus \text{Pooling}(\boldsymbol{Y})) + \boldsymbol{b}), \tag{12}$$

where $\boldsymbol{W}, \boldsymbol{b}$ denote the learnable parameters, $\varphi(\cdot)$ and $\oplus$ denote the activation function and the concatenation operation, respectively.

Therefore, the lower bound of $I(\boldsymbol{Z}; \boldsymbol{Y})$ could be maximized using the binary cross entropy.

4.5 Summary Generation

To fuse the semantics of textual and visual modalities, we concatenate representations of the source sentence $\boldsymbol{X}^t$ and the learned visual information $\boldsymbol{Z}$ as the final input representation $\boldsymbol{X}$ as follows:

$$\boldsymbol{X} = [\boldsymbol{X}^t \oplus \boldsymbol{Z}], \tag{13}$$

where $\oplus$ denotes the concatenation operation.

Since the summary is generated in an auto-regressive manner, we feed $\boldsymbol{X}$ obtained in Eq. (13) and all previous tokens $\hat{y}_1, \hat{y}_2, ..., \hat{y}_{k-1}$ into the BART [28] decoder $\mathcal{D}$ to generate the k-th token as follows:

$$\hat{\boldsymbol{p}}_k = \mathcal{D}(\boldsymbol{X}, \hat{y}_1, \hat{y}_2, ..., \hat{y}_{k-1}), \tag{14}$$

$$\hat{y}_k = V[\text{argmax}(\hat{\boldsymbol{p}}_k)], \tag{15}$$

where $\hat{\boldsymbol{p}}_k \in \mathbb{R}^{|V|}$ denotes the token distribution on the vocabulary V of the k-th token in the generated summary, and $\hat{y}_i$ denotes the i-th token in the generated summary.

4.6 Model Training

We adopt the negative log likelihood to supervise the training of the summary generation, which could be formulated as follows:

$$\mathcal{L}_{NLL} = -\frac{1}{l}\sum_{k=1}^{l} \log(\hat{\boldsymbol{p}}_k(y_k)), \tag{16}$$

where l is the length of the generated summary, $\hat{\boldsymbol{p}}_k$ is calculated in Eq. (14), and y_k denotes the k-th token in the target summary.

Finally, we combine the objective function of summary generation and the trade-off within visual information for the final objective function $\mathcal{L}$ as follows:

$$\mathcal{L} = \mathcal{L}_{NLL} + \mathcal{L}_{IB}. \tag{17}$$

5 EXPERIMENT

In this section, we conduct experiments to validate the effectiveness of our proposed method. Besides, a series of analyses and discussions are performed to show how our proposed method works.

5.1 Settings

5.1.1 Dataset. We conduct experiments on the representative MMSS dataset [31], which contains 62,000/2,000/2,000 samples for the train/validation/test set respectively. Each sample in the dataset is a triplet, i.e. the source sentence, the source image, and the target summary. Tab. 1 shows the length statistics of the textual modality.

Table 1: Length statistics. The format of Avg (Min, Max) displays the average, minimum, and maximum number of words in the source sentence and target summary.

Dataset	Source Sentence	Target Summary
Train	21.68 (11, 63)	7.72 (2, 25)
Validation	24.36 (11, 47)	7.68 (3, 17)
Test	22.97 (11, 51)	7.67 (3, 24)

5.1.2 Evaluation Metrics. Following previous studies [31, 33, 40, 59], we adopt Rouge-1, Rouge-2, and Rouge-L [39] as primary metrics. Since there is currently only one dataset available, following Xiao et al. [59], we additionally employ BLEU [47], BertScore [66], and MoverScore [67] as supplementary metrics for a comprehensive evaluation. Specifically, for the generated summary and the target summary, Rouge is a recall-oriented metric that measures the overlapping of content between them; BLEU is a precision focused metric that measures the overlapping of n-gram between them; BertScore is a model based metric that measures the similarity between them by calculating the maximal cosine similarity greedily on token embeddings from RoBERTa [42]; MoverScore is also a model based metric that uses Word Mover's Distance [27] acting on n-gram embeddings from DistilBERT [53] to measure the semantic distance between them. Note that we adopt Rouge-2 as the metric to determine the optimal model during the training phase.

5.1.3 Implementation Details. We use BART-base[1] to initialize both the textual encoder and the decoder. ViT-base[2] is utilized as the initialization of the visual encoder, where the image resolution is resized to 224×224 in advance, and the resolution of each patch P^2 is set to 32^2, resulting in the number of patches being 49. The dimension of the token representation d_t and the patch representation d_v are both 768. We adopt the Grid Search strategy to determine the optimal values of the hyperparameters α and β in $\mathcal{L}_{IB}$. Specifically, we search for α among the values of {0.05, 0.075, 0.1, 0.125} and for β among the values of {1.0, 1.5, 2.0, 2.5}. Ultimately, the optimal values of α, β are 0.075, 2.0 respectively.

During the training phase, we set the batch size to 16 and utilize AdamW optimizer [43] with the learning rate of 5e-6 for 50 epochs of training. During the testing phase, we utilize the Beam Search strategy with the beam size of 10 to generate the summary.

5.2 Baselines

To validate the effectiveness of our proposed method, we choose four series of baselines for comparison.

Text-only baselines perform MMSS using the textual modality solely without the source image. **Lead** simply uses the first eight words of the source sentence as the generated summary. **Compress** [12] uses integer linear programming to compress the source sentence based on the syntactic structure. **ABS** [51] derives the summary using an attentive CNN-based encoder and a neural language model based decoder. **SEASS** [70] summarizes the source sentence via textual selective encoding.

[1]https://huggingface.co/facebook/bart-base
[2]https://huggingface.co/google/vit-base-patch32-224-in21k

Table 2: Main results of the primary metrics. The results marked with † and ‡ are provided by Li et al. [31] and Lin et al. [40], respectively.

Method	Rouge-1	Rouge-2	Rouge-L
Lead†	33.64	13.40	31.84
Compress†	31.56	11.02	28.87
ABS†	35.95	18.21	31.89
SEASS†	44.86	23.03	41.92
Multi-Source†	39.67	19.11	38.03
Doubly-Attentive†	41.11	21.75	39.92
VG-BART‡	51.02	27.80	48.13
MAtt	47.28	24.85	44.48
CFSum	47.86	25.64	44.64
TGSMR	48.19	25.64	45.27
BART-MMSS	<u>52.15</u>	<u>29.22</u>	<u>49.24</u>
$\mathbf{T}^3$ (ours)	**53.71**	**30.96**	**50.62**
Δ	2.99%↑	5.95%↑	2.80%↑

Vanilla baselines directly inject visual information to enhance the semantics without taking the task-relevance and task-irrelevance into account. **Multi-Source** [37] is a generative method with hierarchical attention mechanisms. **Doubly-Attentive** [6] utilizes a doubly-attentive mechanism to model the visual information. **VG-BART** [63] uses BART [28] as the backbone with attention-based add-on layers to incorporate the visual information.

Compressing task-irrelevant visual information baselines solely focus on compressing the task-irrelevant visual information, disregarding the preservation of the task-relevant visual information. **MAtt** [31] proposes a hierarchical attention mechanism for visual information filtering. **CFSum** [59] filters the interferential visual information according to the consistency between textual and visual modalities.

Preserving task-relevant visual information baselines only preserve the task-relevant visual information, while neglecting the compression to the task-irrelevant visual information. **TGSMR** [33] designs the visual selective gates to capture semantic clues from three levels. **BART-MMSS** [40] uses the prompt-guided encoding to capture the critical information from the source image.

5.3 Main Results

Tab. 2 and Tab. 3 present the overall results on the primary and supplementary metrics respectively. In both tables, the best and suboptimal results are displayed in **bold** and <u>underline</u>, and the bottom row (Δ) shows the relative improvement of our proposed method over the strongest baseline BART-MMSS. The p-value of the significant test between our proposed method and the strongest baseline BART-MMSS is less than 0.05.

From Tab. 2, we can observe that: (1) The latter two series of baselines generally outperform the vanilla baselines. This phenomenon indicates that **compressing the task-irrelevant visual information and preserving the task-relevant visual information are both efficacious for performance improvement in MMSS**. (2) $\mathbf{T}^3$ leads the baselines that either compress the task-irrelevant visual information or preserve the task-relevant visual information by miles. This phenomenon highlights **the importance of modeling the two aspects of visual information from a holistic perspective to achieve a balance between them**. (3) $\mathbf{T}^3$ **achieves the state-of-the-art performance on primary metrics over all baseline methods**. Particularly, $\mathbf{T}^3$ is far ahead of the strongest baseline BART-MMSS with improvements of 2.99%, 5.95%, and 2.80% on Rouge-1, Rouge-2, and Rouge-L, respectively. This verifies the effectiveness of our proposed method once again.

Table 3: Main results of the supplementary metrics.

Method	BLEU	BertScore	MoverScore
CFSum	48.83	86.98	32.36
BART-MMSS	<u>55.52</u>	<u>90.83</u>	<u>61.09</u>
$\mathbf{T}^3$ (ours)	**59.68**	**91.99**	**63.96**
Δ	7.49%↑	1.28%↑	4.70%↑

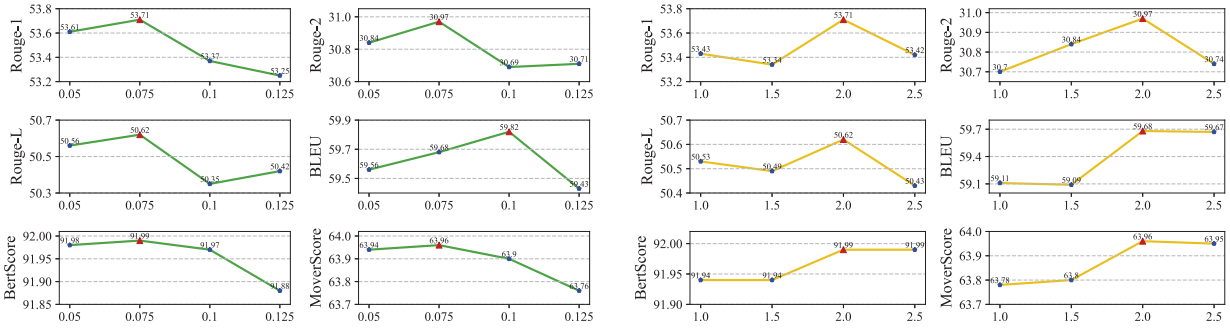
Following Xiao et al. [59], we further assess our proposed method on supplementary metrics. We employ CFSum as a baseline since it is the only one providing the results of supplementary metrics. Besides, to compare $\mathbf{T}^3$ with the strongest baseline BART-MMSS, we reproduce BART-MMSS using its official implementation and calculate the corresponding metrics. We present the results in Tab. 3, where $\mathbf{T}^3$ surpasses BART-MMSS with relative improvements of 7.49%, 1.28%, and 4.70% on BLEU, BertScore, and MoverScore, respectively. This phenomenon is consistent with what we observed in Tab. 2 and reveals the superiority of our proposed method.

5.4 Ablation Study

Following previous studies [40, 59], we conduct the ablation study on the test set to evaluate the effectiveness of each component of our proposed method. Tab. 4 shows the results and we have the following analyses: (1) **w/o Com.**: We remove the task-irrelevant visual information compression module to prove its effectiveness. The decrease in performance highlights the significance of filtering out the task-irrelevant visual information. (2) **w/o Pre.**: To confirm how the preservation of the task-relevant visual information influences the generated summary, we remove the corresponding module. The decline in performance demonstrates the importance of the task-relevant visual information preservation. (3) **w/o Com. & Pre.**: We remove the trade-off within visual information where the image information is injected directly. We notice that the performance drops more severely than removing the module of compression or preservation individually. This phenomenon emphasizes the significance of considering both categories of visual information simultaneously in MMSS. (4) **w/o image**: To verify the importance of visual information, we black out the source image to achieve the goal of removing visual modality without altering the model architecture. The performance drop indicates that visual information could enhance the semantics in MMSS. One might wonder why the performance decline of w/o image is relatively moderate. We attribute it to the fact that with the merit of Information Bottleneck, the blacked-out image is learned as the prompt which helps to refine the crucial semantics for MMSS. (5) **w/o hard negative**: To explore the effect of our proposed hard negative sampling strategy, we replace it with the in-batch random negative sampling. The

Table 4: Ablation study results, where Com. and Pre. are short for the Compression of the task-irrelevant visual information and the Preservation of the task-relevant visual information, respectively.

Method	Rouge-1	Rouge-2	Rouge-L	BLEU	BertScore	MoverScore
T^3	**53.71**	**30.96**	**50.62**	**59.68**	**91.99**	**63.96**
w/o Com.	52.70	29.54	49.57	59.28	91.83	63.64
w/o Pre.	53.28	30.44	50.19	59.28	91.92	63.86
w/o Com. & Pre.	52.26	29.27	49.12	58.42	91.76	63.40
w/o image	53.46	30.49	50.30	59.52	91.94	63.88
w/o hard negative	53.27	30.25	50.32	59.02	91.93	63.73



(a) The control of compressing task-irrelevant visual information. **(b) The control of preserving task-relevant visual information.**

Figure 3: Visual Information Trade-Off Analysis. The red triangle marks the highest point in each line chat.

decrease in performance reflects the effectiveness of our proposed negative sampling strategy to boost performance.

Above experiments verify the effectiveness of each component in **T**^{3}, revealing the design rationality of our proposed method[3].

5.5 Visual Information Trade-Off Analysis

We analyze the trade-off within visual information by controlling the two coefficients, α and β, in $\mathcal{L}_{IB}$. Since the optimal performance of the model is achieved with $\alpha = 0.075$ and $\beta = 2.0$ (see Sec. 5.1.3 for details), we vary the values of either α or β to explore the performance variance. Specifically, to examine the compression of the task-irrelevant visual information, we vary α from 0.05 to 0.125 in steps of 0.025 while keeping β at 2.0. Similarly, we vary β from 1.0 to 2.5 in steps of 0.5, with α fixed at 0.075, to investigate the preservation of the task-relevant visual information. We show the results in Fig. 3 and have the following analyses: (1) Despite the performance fluctuations with different values of α and β, our proposed method, **T**^{3}, produces performance superior to both of them. This phenomenon reflects the importance to compress the task-irrelevant and preserve the task-relevant visual information. (2) The performance could vary with the values of α and β obviously. This demonstrates the significance of finding a decent balance between the compression of the task-irrelevant and the preservation of the task-relevant visual information.

[3]Due to the unique characteristics of different evaluation metrics, the absolute values of the differences on different evaluation metrics in the ablation study vary in order of magnitude, which is consistent with Xiao et al. [59].

5.6 Negative Sampling Analysis

To further explore the effectiveness of our proposed negative sampling strategy, we replace it with the in-batch random negative sampling. Specifically, with the source image and its corresponding target summary as the positive sample, the in-batch random negative sampling constructs the negative samples using the source image and target summaries of other in-batch instances (see Sec. 4.4.3 for details). Since the batch size is 16, the maximum multiple of negative samples to positive samples is set to 8. The comparison between the hard negative sampling and the in-batch random negative sampling with 1, 2, 4, and 8 multiples of negative samples is shown in Fig. 4. We can find that: (1) Using the in-batch random negative sampling, the performance follows a trend of decrease-then-rise with the increasing multiples of negative samples. We attribute this to the limitation of using the in-batch random negative samples to model the distinction between the source and the target. Specifically, since target summaries of other instances express completely different semantics with the target summary, it is hard for the model to learn how to retain the essential need for the summary generation. For this reason, it needs to increase the number of negative samples to provide sufficient semantics to drive the model optimization. (2) Hard negative sampling excels the in-batch random negative sampling consistently. We believe this is due to the use of the source sentence to construct negative samples, which may help to refine the vital information required for summary generation. This phenomenon not only verifies the effectiveness of the hard negative sampling, but also shows the efficiency of our proposed negative sampling strategy.

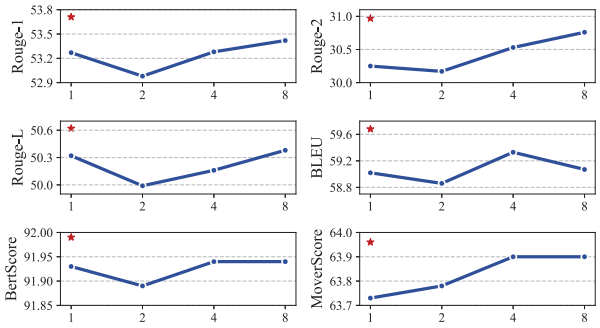


Figure 4: Comparison of different negative sampling strategies. The values on the horizontal axis denote the multiples of negative samples relative to positive samples. The red star marks the result obtained with the hard negative sampling, while the blue line shows the results derived with the in-batch random negative sampling.

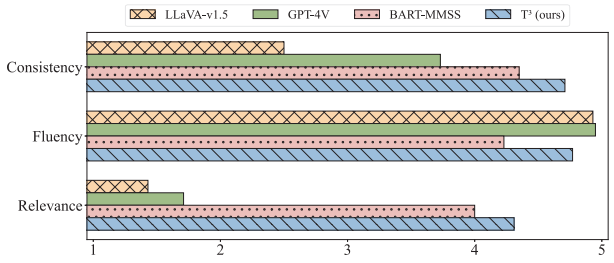


Figure 5: Human evaluation results.

5.7 Comparison with VLP-based and MLLM-based Methods

Due to the success of Vision-Language Pre-training (VLP) [8] and Multimodal Large Language Model (MLLM) [29] in various multimodal tasks, we additionally compare our proposed method with VLP-based and MLLM-based methods.

For VLP-based methods, we explore two methods based on typical VLP models. **UniG** [59] is the variant of CFSum. It directly utilizes UNITER [11], a VLP model with strong multimodal understanding ability, as its encoder. **BLIP-MMSS** finetunes BLIP [34], a VLP model with powerful generation capability for multimodal tasks such as image caption, for MMSS. As for the MLLM-based method, we employ **LLaVA-v1.5** [41], which is the top-ranked open source MLLM in various evaluation benchmarks, for the comparison. To maintain the comparability in terms of parameters with other methods as much as possible, we choose LLaVA-v1.5-7B as it is the smallest size of LLaVA-v1.5. As the performance of few-shot In-Context Learning [5] is highly sensitive to the selection of demonstrations [68], and even the order of demonstrations [50, 68], we leave the few-shot scenario for future work, and prompt LLaVA-v1.5 for MMSS in a zero-shot manner with the following prompt:

> *<Source Image>*
> ***Source Sentence**: <Source Sentence>*
> *Please provide a summary based on the above source image and the above source sentence. Please note that the summary should reflect the main idea of the source sentence with reference to the source image. The length of the summary should be shorter than the length of the source sentence.*
> ***Summary**:*

Reading from Tab. 5, we have the following findings: (1) VLP-based methods lead MLLM-based method by a large margin on all metrics. We believe this is because VLP-based methods benefit from the finetuning. This indicates that there is still a long way to go before MLLM can be directly adopted for MMSS. (2) BLIP-MMSS is comprehensively superior to UniG on all metrics, suggesting that VLP models that value generation ability are more appropriate for MMSS than those pay attention to understanding capability. (3) Our proposed method obviously excels both VLP-based and MLLM-based methods. We attribute this to the fact that both VLP and MLLM mainly focus on the visual modality, leading to relatively poor performance.

5.8 Human Evaluation

In order to evaluate the generated summary with accordance to human preference, human evaluation is performed based on the following three dimensions [16, 26]:

- **Consistency** assesses the factivity between the generated summary and the source sentence and image. The score will be penalized when the generated summary contains hallucinated facts that are not entailed by the source sentence and image.
- **Fluency** measures whether the generated summary follows the rules of the language. When the generated summary contains fewer grammar errors, it receives a higher fluency score.
- **Relevance** refers to the relevance between the generated summary and the target summary. When the generated summary contains less irrelevant contents while missing fewer relevant contents, it gets a higher relevance score.

Specifically, we first randomly sample 50 samples from the test set. Then, three experts are asked to perform the human evaluation. Each dimension is scored using the five-point Likert scale [38] with integers ranging from 1 to 5, where the better the quality of the generated summary, the higher the score. Note that the generated summary will receive a full score when it is exactly the same as the target summary. Lastly, we average the scores of the three experts to get the final score. Following Fabbri et al. [16], we compute the Krippendorff's alpha [25] and achieve 0.52, 0.49, and 0.54 on consistency, fluency, and relevance, respectively, reflecting a reasonable inter-annotator agreement of the human evaluation.

We compare our proposed method with the strongest baseline **BART-MMSS** and two MLLM-based methods: **LLaVA-v1.5** and **GPT-4V(ision)** [46]. We prompt GPT-4V using the same settings as LLaVA-v1.5 (see Sec. 5.7 for details) and conduct the experiment via the official ChatGPT webpage.

Fig. 5 presents the results and we have the following analyses: (1) $\mathbf{T}^3$ outperforms the strongest baseline BART-MMSS in all three dimensions, which once again confirms the effectiveness of balancing the trade-off within visual information from a holistic perspective. (2) $\mathbf{T}^3$ surpasses both two MLLM-based methods on consistency and

Table 5: Comparison with VLP-based and MLLM-based methods.

Method	Rouge-1	Rouge-2	Rouge-L	BLEU	BertScore	MoverScore
UniG	46.22	24.28	43.47	46.85	86.57	30.95
BLIP-MMSS	48.43	26.76	45.85	52.86	91.05	62.32
LLaVA-v1.5	10.43	1.93	9.63	15.03	84.31	52.17
T^3	**53.71**	**30.96**	**50.62**	**59.68**	**91.99**	**63.96**

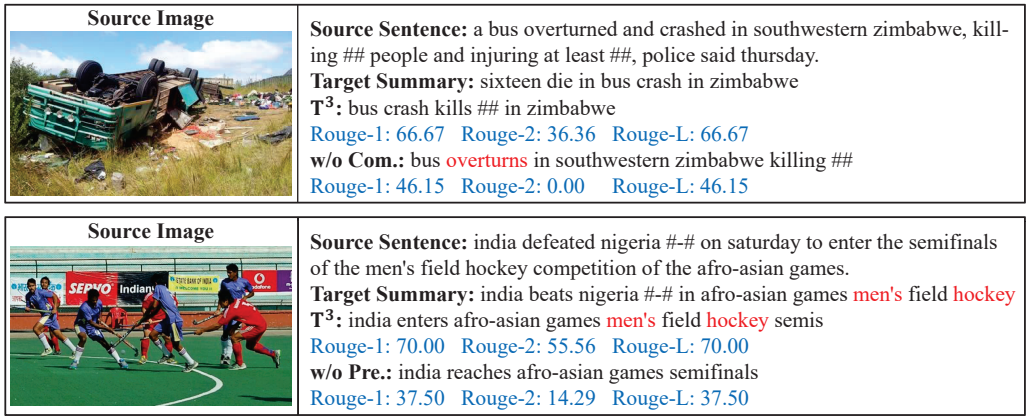


Figure 6: Case study, where Com. and Pre. are short for the Compression of the task-irrelevant visual information and the Preservation of the task-relevant visual information, respectively.

relevance, but is slightly inferior on fluency. We attribute this phenomenon to two aspects. One is that with the merit of large-scale pre-training, MLLMs maintain strong natural language generation ability, resulting in superb performance on fluency. The other is that MLLMs pay excessive attention to the image. Specifically, the summaries generated by MLLM-based methods tend to describe the image rather than sketch the source sentence, thereby losing the semantics of the source sentence. As a result, MLLM-based methods fall short in the consistency and relevance.

Overall, the human evaluation demonstrates the superiority of our proposed method.

5.9 Case Study

To provide an intuitive understanding of how our proposed method balances the trade-off between the task-relevant and task-irrelevant visual information, we show some cases between our proposed method and its variants in Fig. 6.

In the upper half of Fig. 6, we present the case of **w/o Com.**, which removes the compression of the task-irrelevant visual information in $\mathbf{T}^3$. We can see that w/o Com. produces the summary with a redundant term 'overturns', while $\mathbf{T}^3$ generates the summary as concise as the target summary. This intuitively reflects the necessity of compressing the task-irrelevant visual information.

In the lower half of Fig. 6, we show the case of **w/o Pre.**, which removes the preservation of the task-relevant visual information in $\mathbf{T}^3$. It can be observed that w/o Pre. eliminates the visual elements 'male player' and 'hockey stick', resulting in the absence of the significant terms 'men' and 'hockey' in the generated summary. As for $\mathbf{T}^3$, these important details are retained. This phenomenon intuitively demonstrates the importance of preserving the task-relevant visual information.

To sum up, with the trade-off within visual information, the summary generated by $\mathbf{T}^3$ encompasses the necessary factors while eliminating the redundancy. The above cases highlight the importance of the trade-off within visual information and verify the effectiveness of our proposed method again.

6 CONCLUSION

In this paper, we propose an elegant method, $\mathbf{T}^3$, for MultiModal Sentence Summarization (MMSS). Considering the issues of over-preservation and over-compression of visual information in MMSS, we resort to Information Bottleneck (IB) for an effective solution. Specifically, with the two mutual information terms in IB collaborating with each other, our proposed method could acquire the visual representation that maximally compresses the task-irrelevant visual information while preserving the task-relevant visual information as most. Experiments on the representative MMSS dataset show that our proposed method is far ahead of all competitive baseline methods, and extensive analyses are performed to confirm the effectiveness of our proposed method.

ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China (Grant No.2021YFB3100600) and the Youth Innovation Promotion Association of CAS (Grant No.2021153).

REFERENCES

[1] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. 2017. Deep Variational Information Bottleneck. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=HyxQzBceg

[2] Anton Bardera, Jaume Rigau, Imma Boada, Miquel Feixas, and Mateu Sbert. 2009. Image Segmentation Using Information Bottleneck Method. *IEEE Trans. Image Process.* 18, 7 (2009), 1601–1612. https://doi.org/10.1109/TIP.2009.2017823

[3] Jingwen Bian, Yang Yang, and Tat-Seng Chua. 2013. Multimedia summarization for trending topics in microblogs. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, Qi He, Arun Iyengar, Wolfgang Nejdl, Jian Pei, and Rajeev Rastogi (Eds.). ACM, 1807–1812. https://doi.org/10.1145/2505515.2505652

[4] Jingwen Bian, Yang Yang, Hanwang Zhang, and Tat-Seng Chua. 2015. Multimedia Summarization for Social Events in Microblog Stream. *IEEE Trans. Multim.* 17, 2 (2015), 216–228. https://doi.org/10.1109/TMM.2014.2384912

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[6] Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-Attentive Decoder for Multi-modal Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, 1913–1924. https://doi.org/10.18653/v1/P17-1175

[7] Jiangxia Cao, Jiawei Sheng, Xin Cong, Tingwen Liu, and Bin Wang. 2022. Cross-Domain Recommendation to Cold-Start Users via Variational Information Bottleneck. In *38th IEEE International Conference on Data Engineering, ICDE 2022, Kuala Lumpur, Malaysia, May 9-12, 2022*. IEEE, 2209–2223. https://doi.org/10.1109/ICDE53745.2022.00211

[8] Feilong Chen, Duzhen Zhang, Minglun Han, Xiuyi Chen, Jing Shi, Shuang Xu, and Bo Xu. 2023. VLP: A Survey on Vision-language Pre-training. *Int. J. Autom. Comput.* 20, 1 (2023), 38–56. https://doi.org/10.1007/s11633-022-1369-5

[9] Jingqiang Chen and Hai Zhuge. 2018. Abstractive Text-Image Summarization Using Multi-Modal Attentional Hierarchical RNN. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 4046–4056. https://doi.org/10.18653/V1/D18-1438

[10] Jingqiang Chen and Hai Zhuge. 2018. Extractive Text-Image Summarization Using Multi-Modal RNN. In *14th International Conference on Semantics, Knowledge and Grids, SKG 2018, Guangzhou, China, September 12-14, 2018*. IEEE, 245–248. https://doi.org/10.1109/SKG.2018.00033

[11] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX (Lecture Notes in Computer Science, Vol. 12375)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 104–120. https://doi.org/10.1007/978-3-030-58577-8_7

[12] James Clarke and Mirella Lapata. 2008. Global Inference for Sentence Compression: An Integer Linear Programming Approach. *J. Artif. Intell. Res.* 31 (2008), 399–429. https://doi.org/10.1613/jair.2433

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=YicbFdNTTy

[15] Erkut Erdem, Menekse Kuyu, Semih Yagcioglu, Anette Frank, Letitia Parcalabescu, Barbara Plank, Andrii Babii, Oleksii Turuta, Aykut Erdem, Iacer Calixto, Elena Lloret, Elena Simona Apostol, Ciprian-Octavian Truica, Branislava Sandrih, Sanda Martincic-Ipsic, Gábor Berend, Albert Gatt, and Grazina Korvel. 2022. Neural Natural Language Generation: A Survey on Multilinguality, Multimodality, Controllability and Learning. *J. Artif. Intell. Res.* 73 (2022), 1131–1207. https://doi.org/10.1613/jair.1.12918

[16] Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Trans. Assoc. Comput. Linguistics* 9 (2021), 391–409. https://doi.org/10.1162/TACL_A_00373

[17] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. 2020. Learning Robust Representations via Multi-View Information Bottleneck. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=B1xwcyHFDr

[18] Muskan Garg, Seema Wazarkar, Muskaan Singh, and Ondrej Bojar. 2022. Multimodality for NLP-Centered Applications: Resources, Advances and Frontiers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, 6837–6847. https://aclanthology.org/2022.lrec-1.738

[19] Anubhav Jangra, Adam Jatowt, Mohammed Hasanuzzaman, and Sriparna Saha. 2020. Text-Image-Video Summary Generation Using Joint Integer Linear Programming. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12036)*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer, 190–198. https://doi.org/10.1007/978-3-030-45442-5_24

[20] Anubhav Jangra, Sourajit Mukherjee, Adam Jatowt, Sriparna Saha, and Mohammad Hasanuzzaman. 2023. A Survey on Multi-modal Summarization. *ACM Comput. Surv.* 55, 13s (2023), 296:1–296:36. https://doi.org/10.1145/3584700

[21] Anubhav Jangra, Sriparna Saha, Adam Jatowt, and Mohammed Hasanuzzaman. 2021. Multi-Modal Supplementary-Complementary Summarization using Multi-Objective Optimization. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 818–828. https://doi.org/10.1145/3404835.3462877

[22] Liqiang Jing, Yiren Li, Junhao Xu, Yongcan Yu, Pei Shen, and Xuemeng Song. 2023. Vision Enhanced Generative Pre-trained Language Model for Multimodal Sentence Summarization. *Mach. Intell. Res.* 20, 2 (2023), 289–298. https://doi.org/10.1007/S11633-022-1372-X

[23] Kenji Kawaguchi, Zhun Deng, Xu Ji, and Jiaoyang Huang. 2023. How Does Information Bottleneck Help Deep Learning?. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 16049–16096. https://proceedings.mlr.press/v202/kawaguchi23a.html

[24] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1312.6114

[25] Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. (2011).

[26] Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural Text Summarization: A Critical Evaluation. 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 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 540–551. https://doi.org/10.18653/V1/D19-1051

[27] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From Word Embeddings To Document Distances. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015 (JMLR Workshop and Conference Proceedings, Vol. 37)*, Francis R. Bach and David M. Blei (Eds.). JMLR.org, 957–966. http://proceedings.mlr.press/v37/kusnerb15.html

[28] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

[29] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. 2023. Multimodal Foundation Models: From Specialists to General-Purpose Assistants. *CoRR* abs/2309.10020 (2023). https://doi.org/10.48550/ARXIV.2309.10020 arXiv:2309.10020

[30] Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. Aspect-Aware Multimodal Summarization for Chinese E-Commerce Products. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 8188–8195. https://doi.org/10.1609/AAAI.V34I05.6332

[31] Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, and Chengqing Zong. 2018. Multi-modal Sentence Summarization with Modality Attention and Image Filtering. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, Jérôme Lang (Ed.). ijcai.org, 4152–4158. https://doi.org/10.24963/ijcai.2018/577

[32] Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. Multi-modal Summarization for Asynchronous Collection of Text, Image, Audio and Video. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, 1092–1102. https://doi.org/10.18653/V1/D17-1114

[33] Haoran Li, Junnan Zhu, Jiajun Zhang, Xiaodong He, and Chengqing Zong. 2020. Multimodal Sentence Summarization via Multimodal Selective Encoding. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, Donia Scott, Núria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, 5655–5667. https://doi.org/10.18653/v1/2020.coling-main.496

[34] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (Eds.). PMLR, 12888–12900. https://proceedings.mlr.press/v162/li22n.html

[35] Mingzhe Li, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. 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 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 9360–9369. https://doi.org/10.18653/V1/2020.EMNLP-MAIN.752

[36] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2022. Foundations and Recent Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. *CoRR* abs/2209.03430 (2022). https://doi.org/10.48550/arXiv.2209.03430 arXiv:2209.03430

[37] Jindrich Libovický and Jindrich Helcl. 2017. Attention Strategies for Multi-Source Sequence-to-Sequence Learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, 196–202. https://doi.org/10.18653/v1/P17-2031

[38] R Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology* (1932).

[39] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[40] Dengtian Lin, Liqiang Jing, Xuemeng Song, Meng Liu, Teng Sun, and Liqiang Nie. 2023. Adapting Generative Pretrained Language Model for Open-domain Multimodal Sentence Summarization. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (Eds.). ACM, 195–204. https://doi.org/10.1145/3539618.3591633

[41] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved Baselines with Visual Instruction Tuning. *CoRR* abs/2310.03744 (2023). https://doi.org/10.48550/ARXIV.2310.03744 arXiv:2310.03744

[42] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[43] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=Bkg6RiCqY7

[44] Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2021. Variational Information Bottleneck for Effective Low-Resource Fine-Tuning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=kvhzKz-_DMF

[45] Saeid Motiian and Gianfranco Doretto. 2016. Information Bottleneck Domain Adaptation with Privileged Information for Visual Recognition. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII (Lecture Notes in Computer Science, Vol. 9911)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer, 630–647. https://doi.org/10.1007/978-3-319-46478-7_39

[46] OpenAI. 2024. GPT-4V(ision) system card. https://openai.com/research/gpt-4v-system-card

[47] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. ACL, 311–318. https://doi.org/10.3115/1073083.1073135

[48] Ben Poole, Sherjil Ozair, Aäron van den Oord, Alexander A. Alemi, and George Tucker. 2019. On Variational Bounds of Mutual Information. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 5171–5180. http://proceedings.mlr.press/v97/poole19a.html

[49] Xueming Qian, Mingdi Li, Yayun Ren, and Shuhui Jiang. 2019. Social media based event summarization by user-text-image co-clustering. *Knowl. Based Syst.* 164 (2019), 107–121. https://doi.org/10.1016/J.KNOSYS.2018.10.028

[50] Mathieu Ravaut, Shafiq Joty, Aixin Sun, and Nancy F Chen. 2023. On Context Utilization in Summarization with Large Language Models. *arXiv e-prints* (2023), arXiv–2310.

[51] Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton (Eds.). The Association for Computational Linguistics, 379–389. https://doi.org/10.18653/v1/d15-1044

[52] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A Large-scale Dataset for Multimodal Language Understanding. *CoRR* abs/1811.00347 (2018). arXiv:1811.00347 http://arxiv.org/abs/1811.00347

[53] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR* abs/1910.01108 (2019). arXiv:1910.01108 http://arxiv.org/abs/1910.01108

[54] Ravid Shwartz-Ziv and Naftali Tishby. 2017. Opening the Black Box of Deep Neural Networks via Information. *CoRR* abs/1703.00810 (2017). arXiv:1703.00810 http://arxiv.org/abs/1703.00810

[55] Naftali Tishby, Fernando C. N. Pereira, and William Bialek. 2000. The information bottleneck method. *CoRR* physics/0004057 (2000). http://arxiv.org/abs/physics/0004057

[56] Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop, ITW 2015, Jerusalem, Israel, April 26 - May 1, 2015*. IEEE, 1–5. https://doi.org/10.1109/ITW.2015.7133169

[57] Akanksha Tiwari, Christian von der Weth, and Mohan S. Kankanhalli. 2018. Multimodal Multiplatform Social Media Event Summarization. *ACM Trans. Multim. Comput. Commun. Appl.* 14, 2s (2018), 38:1–38:23. https://doi.org/10.1145/3115433

[58] Petar Velickovic, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R. Devon Hjelm. 2019. Deep Graph Infomax. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=rklz9iAcKQ

[59] Min Xiao, Junnan Zhu, Haitao Lin, Yu Zhou, and Chengqing Zong. 2023. CFSum Coarse-to-Fine Contribution Network for Multimodal Summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 8538–8553. https://aclanthology.org/2023.acl-long.476

[60] Shize Xu, Liang Kong, and Yan Zhang. 2013. A cross-media evolutionary timeline generation framework based on iterative recommendation. In *International Conference on Multimedia Retrieval, ICMR'13, Dallas, TX, USA, April 16-19, 2013*, Ramesh C. Jain, Balakrishnan Prabhakaran, Marcel Worring, John R. Smith, and Tat-Seng Chua (Eds.). ACM, 73–80. https://doi.org/10.1145/2461466.2461480

[61] Rui Yan, Xiaojun Wan, Mirella Lapata, Wayne Xin Zhao, Pu-Jen Cheng, and Xiaoming Li. 2012. Visualizing timelines: evolutionary summarization via iterative reinforcement between text and image streams. In *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, Xue-wen Chen, Guy Lebanon, Haixun Wang, and Mohammed J. Zaki (Eds.). ACM, 275–284. https://doi.org/10.1145/2396761.2396799

[62] Qian Yong, Jueqi Wei, YiRen Zhang, XiLun Zhang, Chao Wei, Simiao Chen, Yunhe Li, Cheng Ye, Bing Huang, and Hao Wang. 2023. CGSMP: Controllable Generative Summarization via Multimodal Prompt. In *Proceedings of the 1st Workshop on Large Generative Models Meet Multimodal Applications, LGM3A 2023, Ottawa ON, Canada, 2 November 2023*, Zheng Wang, Cheng Long, Shihao Xu, Bingzheng Gan, Wei Shi, Zhao Cao, and Tat-Seng Chua (Eds.). ACM, 45–50. https://doi.org/10.1145/3607827.3616841

[63] Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 3995–4007. https://doi.org/10.18653/v1/2021.emnlp-main.326

[64] Cenyuan Zhang, Xiang Zhou, Yixin Wan, Xiaoqing Zheng, Kai-Wei Chang, and Cho-Jui Hsieh. 2022. Improving the Adversarial Robustness of NLP Models by Information Bottleneck. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 3588–3598. https://doi.org/10.18653/v1/2022.findings-acl.284

[65] Litian Zhang, Xiaoming Zhang, and Junshu Pan. 2022. Hierarchical Cross-Modality Semantic Correlation Learning Model for Multimodal Summarization. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 11676–11684. https://doi.org/10.1609/AAAI.V36I10.21422

[66] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=SkeHuCVFDr

[67] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 563–578. https://doi.org/10.18653/V1/D19-1053

[68] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-shot Performance of Language Models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 12697–12706. http://proceedings.mlr.press/v139/zhao21c.html

[69] Jie Zhou, Yuanbin Wu, Qin Chen, Xuanjing Huang, and Liang He. 2021. Attending via both Fine-tuning and Compressing. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021 (Findings of ACL, Vol. ACL/IJCNLP 2021)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 2152–2161. https://doi.org/10.18653/v1/2021.findings-acl.189

[70] Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. 2017. Selective Encoding for Abstractive Sentence Summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, 1095–1104. https://doi.org/10.18653/v1/P17-1101

[71] Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. MSMO: Multimodal Summarization with Multimodal Output. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 4154–4164. https://doi.org/10.18653/V1/D18-1448