SMSMO: Learning to Generate Multimodal Summary for Scientific Papers

Xinyi Zhong[a,*], Zusheng Tan[a,*], Shen Gao[b], Jing Li[c], Jiaxing Shen[a], Jingyu Ji[a], Jeff Tang[d], Billy Chiu[a,**]

[a] *School of Data Science, Lingnan University, Hong Kong, China*
[b] *University of Electronic Science and Technology of China, China*
[c] *Harbin Institute of Technology (Shenzhen), China*
[d] *The Hong Kong Polytechnic University, China*

Abstract

Nowadays, publishers like Elsevier increasingly use graphical abstracts (i.e., a pictorial paper summary) along with textual abstracts to facilitate scientific paper readings. In such a case, automatically identifying a representative image and generating a suitable textual summary for individual papers can help editors and readers save time, facilitating them in reading and understanding papers. To tackle the case, we introduce the dataset for Scientific Multimodal Summarization with Multimodal Output (SMSMO). Unlike other multimodal tasks which performed on generic, medium-size contents (e.g., news), SMSMO needs to tackle longer multimodal contents in papers, with finer-grained multimodality interactions and semantic alignments between images and text. For this, we propose a cross-modality, multi-task

*Equal Contribution.
**Corresponding author.
Email addresses: xinyizhong@ln.edu.hk (Xinyi Zhong), allentan@ln.hk (Zusheng Tan), shengao@pku.edu.cn (Shen Gao), li.jing@hit.edu.cn (Jing Li), jiaxingshen@ln.edu.hk (Jiaxing Shen), jingyuji@ln.edu.hk (Jingyu Ji), billychiu@ln.edu.hk (Billy Chiu)

learning summarizer (CMT-Sum). It captures the intra- and inter-modality interactions between images and text through a cross-fusion module; and models the finer-grained image-text semantic alignment by jointly generating the text summary, selecting the key image and matching the text and image. Extensive experiments conducted on two newly introduced datasets on the SMSMO task showcase our model's effectiveness.
Keywords: Multi-task, Multimodal Scientific Summarization,
Cross-modality Fusion

1. Introduction

As scientific publications continue to increase (especially fuelled after global challenges like COVID-19 and breakthrough technologies like ChatGPT), they have become an important knowledge source for data science and artificial intelligence (AI) research[1]. To help scientists/scholars to stay well-versed in the deluge of information, it is essential to advance natural language processing (NLP) technologies for scientific document summarization.

Scientific literature is deemed to be visually-rich documents, conveying not only text, but also images (e.g., charts, tables and figures). Images help readers to gain a visualized understanding of the paper while the text provides more details related to it. As illustrated in Figure 1, the theme of the paper is a model which can jointly perform "Chinese named entity recognition" (NER) and "Chinese word segmentation" (CWS), with "shared information"

[1]According to the latest Stanford AI Index Report [1], there is a 1.4-time growth in publication (from about 350k to 500k) in the last 5 years, especially in topics like multimodal language models, generative AI and healthcare AI.

and "self-attention" (see Text Abstract A1 to A3). Here, the Graphical Abstract (a schematic diagram[2]) represent the relationships between different model elements (e.g., self-attention, NER, CWS) and features (e.g., shared information) through visual components such as colour and lines. However, the diagram alone may not be sufficient in clearly expressing specific content. Conversely, the text modality contains detailed descriptions of individual model objects but has limitations in revealing their intrinsic connections. In such a case, it is essential to have a *multimodal summary*, which contains both a textual paper summary (a.k.a., text abstract) and a representative image (a.k.a., graphical abstract) of the given papers. The two sets of information can complement each other and enrich summarization, thereby helping readers save time and read the papers more effectively.

Scientific document summarization has been a long-standing research topic in NLP [4, 5, 6]. The output of existing scientific summarization systems are usually text-only [7, 8, 9, 10, 11, 12, 13, 14]. Recently, Multimodal Summarization with Multimodal Output (MSMO) has been explored in several areas, including news headline generation, legal fact-checking and social media post summarization [15, 16, 17]. MSMO models aim at generating both image and text summaries using a joint model. Compared to the text-only methods, which only produce an unimodal summary, MSMO provides a better user experience with an easier and faster way to get useful informa-

[2]From the survey work conducted by Yoon et al., (2017) [2] and Yang et al., (2019) [3], paper authors commonly used schematic diagrams and photos as graphical abstract to enhance the illustration of their models and backgrounds in the paper, equally popular are charts and table for better-presenting result. We describe more details in Sec. 2.2.

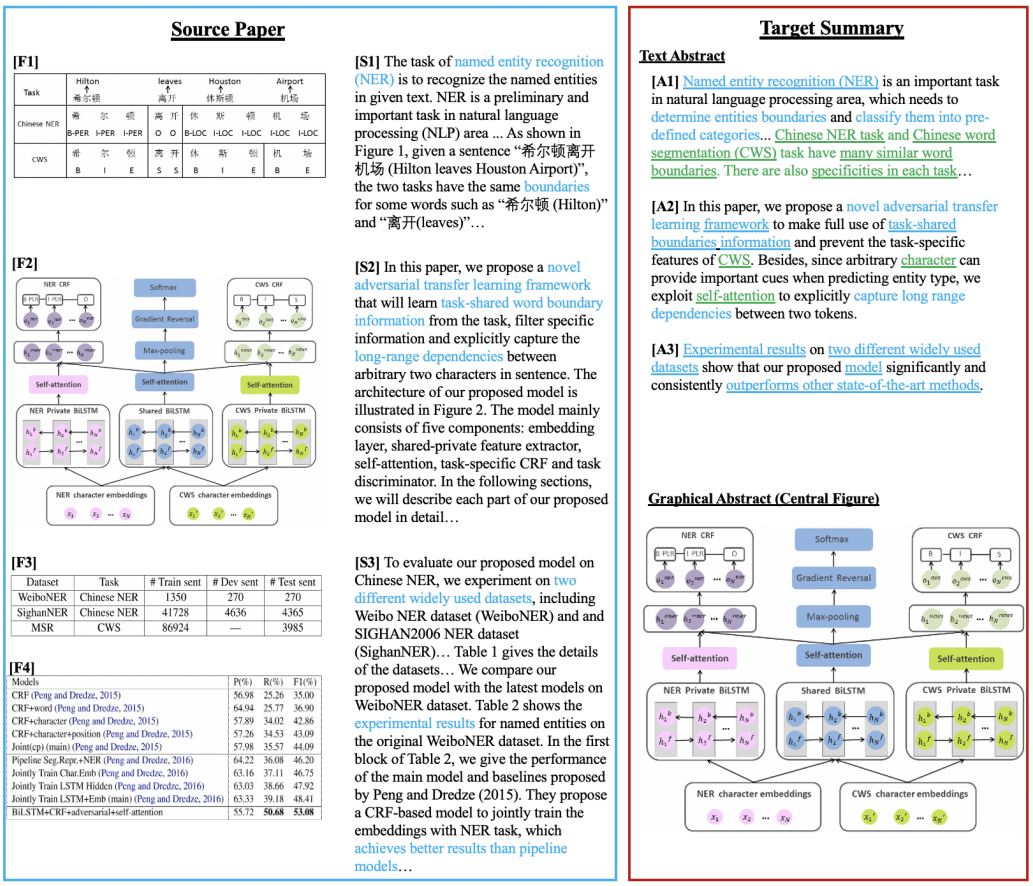


Figure 1: A paper-summary example taken from our $\text{AVIATE}_{\text{SMSMO}}$ dataset. To facilitate understanding, we manually segmented the text abstract into several parts, each corresponding to specific themes or sections within the original paper. The blue words in the text summary represent keywords that exist in the source text, whereas the green words represent concepts presented in the images. Underlined words represent items that presented in both the source text and images.

tion [18]. In this paper, we introduce a novel dataset for *Scientific Multimodal Summarization with Multimodal Output* (SMSMO). The objective of SMSMO is to train models that can generate text summaries while also identifying the key image associated with each individual paper. Its significance lies in the potential to enhance the clarity and accessibility of research findings. Accurate and comprehensive summaries enable better comprehension and quicker assimilation of paper contents, which is critical in the fast-paced research environment. Furthermore, multimodal summaries can facilitate the development of (multimodal) paper retrieval systems, with both text and visual abstracts captured by the search engines. This helps increase the reach of the research as it is no longer restricted to searchability by textual content [19].

In SMSMO, the multimodal information, be it image or text, describes the same paper. These two sets of information complement each other during the summarization process. A direct way to encode the two information sources is to combine them as a global feature vector, using it to generate multimodal summarization [20]. However, images and text generally have distinct feature spaces. Hence, directly combining the two is not an effective approach for capturing the essential information from both modalities. Indeed, this method may introduce noise and hinder the performance of summarization [21]. Different methods have been proposed to fuse the image and text features, ranging from specific task designs to different optimization strategies. For example, Xiao et al. (2024) [22] proposed a multi-stage approach in which they first optimize the text summarizer, then the image selector via self-labelling to preserve the images that are relevant to the generated sum-

mary. Moreover, Zhu et al. (2020) [23] included an image selection task into text summarization, selecting the pseudo key image based on the full source text and the summary generated. Furthermore, Phani et al. (2024) [24] incorporated a selective-gate mechanism for multilingual MSMO tasks, aiming to fuse the text-image features across multilingual news. Typically, these approaches focus on either global or local image-text correspondences, with few effectively addressing both simultaneously. Global-level approaches focus on mapping all images and the entire document into a shared space. For example, Krubinski et al. (2024) [25] used a large language model to unify summary generation and image selection. It can fully extract global-level features across image and text, but there is a large gap between finer-grained feature spaces. While powerful in capturing the high-level theme, they often overlook intricate details, neglecting the fine-grained correspondences across modalities. Conversely, local-level approaches focus on aligning images and text by accumulating similarities of individual patch-phrase pairs [26, 27]. For example, Jin et al. (2024) [28] build a word graph from review text and enrich it by linking detected image objects to their corresponding entities. This approach results in summaries that are rich in specific, detailed information but may lack cohesive structure or fail to convey the document's overarching message. It is important to consider how to effectively learn features from multimodal modalities at different levels to obtain high-quality summaries in SMSMO.

In scientific papers, text and images can convey information at different levels of granularity or with varying degrees of semantic similarity. Par-

ticularly, scientific papers are often organized by *sections* (e.g., IMRaD[3]), in which text and images within the same section exhibit high *intra-* and *inter-modal* correlation. On one hand, there exists a hierarchical semantic relationship within the same modality between visual and textual elements. As showcased in Figure 1 (refer to S1 to S3), textual content at the word level (e.g., "The task of named entity recognition (NER) is to...") contributes to the broader section-level context (i.e., Introduction) within individual papers. On the other hand, individual image semantics typically align with the section's textual content referencing them (e.g., F1 visually illustrates the concept of a NER task), displaying an inter-modal correlation. Furthermore, the theme of Abstract A1 to A3 also aligns with the sectional content of S1 to S3 and F1 to F5. Understanding these correlations enables one modality to compensate for missing information in the other (see the underlined words in A1 to A3). Additionally, by integrating multimodal elements at various levels, the resulting summary can contain a fine-grained description of the textual content as well as the most relevant image, offering a more informative, intuitive, and accessible narrative compared to conventional text-only summaries.

This paper presents CMT-Sum, a cross-modality multi-task learning model for SMSMO. CMT-Sum aims to learn both intra- and inter-modal correlations in paper text and images. Initially, two unimodality encoders are utilized to learn individual image and text features at an intra-modal level. Next, a cross-modality fusion (CFM) module is introduced to capture the

[3]IMRaD (**I**ntroduction, **M**ethods, **R**esults, and **D**iscussion) refers to a common organizational structure in scientific writing.

inter-modal correlation between the image and text. It includes two sub-modules, image-text (section) fusion, and section-word fusion. Therefore, the learned representation comprehensively captures information ranging from global-level semantics to local-level correlations between text and images. The hierarchical and progressive design allows the model to generate sharper intra-modality and inter-modality fusion features effectively. To improve the quality of multimodal output (with consideration of the fine-grained interaction between text and images), we propose a multimodal objective function, in which text summary generation, image selection and image-text relevance matching are jointly optimized. The tasks aim to coalesce various levels of fused semantic features, encompassing word, section and image semantic features, along with their interaction. To evaluate our CMT-Sum, we construct the first dataset for SMSMO in scientific NLP. The experiment conducted on our datasets reveals that CMT-Sum achieves better performance compared to other baseline methods in both automatic evaluations and human assessments.

2. Related Work

2.1. Scientific Document Summarization

Automated summarization of scientific documents is a long-standing research area in NLP ([4, 5, 6]). Significant progress has been made with the development of practical datasets and evaluation tasks. Examples include: *abstract generation* [29], *citation sentence generation* [12], *Related Work section generation* [9] *extreme summarization* (i.e., one-line summary of the entire paper) [11] and *layman text generation* (i.e., generating a simple

text from the source paper that non-experts can understand) [30]. With the advancement in data, different summarization models have also been developed. These include models that exploit citation contexts [31]; and other techniques that exploit the distinctive characteristics in scientific documents such as long length and structure ([10, 32, 33]). The datasets and models denote valuable resources in scientific NLP. They are ***intriguing*** (they help researchers more quickly understand the basic ideas in a piece of research), but ***inadequate*** for scientific summarization. Particularly, the output of these models is usually in a single modality, notably text.

2.2. Graphical Abstract

The long and complex structure of scientific text poses a challenge in identifying the key semantic components and converting them into a structured format. Hence, journal publishers have been exploring concise summaries in other modalities like images (a.k.a. *graphical abstract*). A graphical abstract (GA) provides a concise image summary of a paper's theme and contribution. Regarding this, Yoon et al., (2017) [2] reported a 350% increase of GAs used in social science from 2011 to 2015. In computer science, Yang et al., (2019) [3] examined the papers accepted in top conferences like ICCV (International Conference on Computer Vision) and CVPR (Conference on Computer Vision and Pattern Recognition). They observed that more than half of the authors (68% in ICCV and 65% in CVPR) incorporated the "teaser figures" (a form of GA) in their paper submissions. Among these figures, almost half of them are diagrams and pictures, which are used to provide an overview of the proposed methods, models and backgrounds. Another typical use of GA is for better presenting research findings, using charts, tables

or plots. It is essential to have both text and visual modalities. Particularly, the image modality (e.g., charts) represents the relationships between elements/concepts and data features through visual components such as colour and lines, while the text modality contains more detailed descriptions of individual elements and conveys deep insights [3]. Hence, these two sets of information can complement each other and enrich summarization. Nowadays, leading publishers of scientific articles (e.g., Elsevier) also suggest authors provide *multimedia summaries* (i.e., a textual abstract supplied with GA) to facilitate the searching process [34].

2.3. Multimodal Summarization

In general NLP, multimodal summarization (MMS) is rapidly expanding, with various applications such as review summarization [35] and discussion summarization [36]. Different from the traditional Single summarization with Single Output (**SSO**), MMS aims to extract salient information from various input modalities, including text and images, to produce a concise summary encapsulating the core multi-modal semantics. MMS methodologies are broadly classified into two categories: Multi-modal Summarization with Single-modal Output (MSSO), characterized by a unimodal summary (e.g., text), and Multi-modal Summarization with Multi-modal Output (MSMO), which generates both textual and visual summaries for comprehensive representation.

In MSSO, researchers often concerned about how to improve the quality of text summarization through multimodal data sources. For example, Li et al. (2018) [37] presented an extractive approach aimed at summarizing sentences from a collection of articles, audio clips, images, and videos. To

address the noise present in multimodal sources, Lu et al. (2024) [38] use cascade gates to balance the contribution of each modality. Besides, Argadea et al. (2024) [39] introduce a two-level attention mechanism, which involves a first-level pairwise computation of the attention weights between text and other modalities, followed by a second-level attention that focuses on the pairwise attention feature. A different approach was taken by Jin et al. (2024) [28], who employed a bi-hop graph to achieve alignment between different modalities. Their method first aligns the word with its corresponding sentence in the document and then aligns the sentence with the image caption, thereby establishing a connection between the image and the text. Other studies also explored attention activation [40], selective gating [41], and self-labeling [22] techniques to guide the selection and filtering of multimodal noise, thereby improving summarization performance. Apart from reducing data noise, some studies explore modality-specific features. For example, Zhang et al. (2021) [42] leveraged image location information via multimodal fusion blocks to capture high-order text-image interactions. In the e-commerce domain, Li et al. (2020) [43] used both product images and textual descriptions of product aspects to enhance their multimodal summarization model. Other than that, some studies explore using pre-training models/strategies. For example, Jing et al. (2023) [44] used contrastive pre-training to connect text and image attributes semantically. They aim to align the text and image representations of images and text by enhancing their similarity. Also, Liu et al. (2023) [45] employed knowledge distillation techniques to extract relevant information from pre-trained vision-language models, improving their multimodal headline generation model. Another

type of approach aims to model the intricate relationship between semantic elements like words/phrases and image segments. For example, Jiang et al. (2023) [46] and Li et al. (2020) [47] partition the image into patches and model the similarity between these patches and word representations. Subsequently, they identify the patches that exhibit high similarity with the text and utilize them as the image gate to guide the text encoding procedure. Along similar lines, Xiao et al. (2023)[27] present two visual complement modules at the word and phrase levels. By leveraging images to enhance semantic understanding at these levels, they facilitate comprehensive multi-modal alignment. Some approaches focus on aligning semantic details across modalities at the attention layer, bridging the semantic divide between text and image models. For instance, Yu et al. (2021) [48] enhance pre-trained text embeddings (BART) by integrating visual cues through a newly introduced cross-attention mechanism in each encoder layer. Suman et al. (2021) [49] and Overbay et al. (2023) [17] embed cross-attention layers into the Transformer architecture [50], allowing simultaneous observation of input texts and images. There are also other alignment techniques, depending on optimal transport [51] and video-text time correspondences [52]. This enables nuanced cross-modal learning, leading to superior text summarization quality.

Unlike MSSO, MSMO enhances the interaction between multi-modal features by incorporating auxiliary tasks (e.g., key image selection), resulting in better text summarization performance. For instance, Zhu et al. (2018) [18] proposed the first MSMO model, where they use a cross-attention mechanism to fuse the text-image features for better text generation, and the coverage

mechanism is used to help select representative images. Later on, Zhu et al. (2020) [23] improved the MSMO model by replacing the coverage mechanism with pseudo image labels. These labels were obtained by comparing the image caption with the target summary and the order in which the images appear. Overbay et al. (2023) [17] and Liu et al. (2024) [53] utilize hierarchical attention to merge textual and visual features for generating a summary, enclosing also a key frame from associated videos to enrich the summary. Zhang, Meng et al. (2022) [54] introduced a joint model, which simultaneously outputs abstractive and extractive text summaries and a representative image. Variant MSMO tasks have also emerged recently. For example, Krubiński et al. (2023) [15] proposed a dataset and explored the use of a hierarchical attention mechanism for MMSO in Czech news. Subsequently, the authors explored the application of large language models like BART and T5 for generating news headlines from images and videos [25]. Additionally, Phani et al. (2024) [24] propose a selective gate to align the text-image semantics in multilingual news. In the legal domain, Yao et al. (2023) [16] introduced an MSMO dataset for explanation generation and legal fact-checking. They further explore using a shared encoder with multitask training to predict the veracity (Supported or Refuted) based on textual and visual evidence while also generating relevant explanations for the predictions.

In contrast to the growing work of MMS in different domains, there is inadequate work in scientific MMS. Some emerging works include Yang et al. (2019) [3] and Atri et al. (2021; 2023) [7, 55], who incorporated paper images and presentation videos for paper summarization. But still, their out-

puts are represented in a single modality, either text or images (not both). Other than that, most existing studies on text and vision alignment concentrate on mapping images with texts at either a broad, global level (covering entire documents and all images) or a more specific, localized correspondence (between image patches and individual words/phrases). This tends to overlook the intricate hierarchical semantic relationships within multimodal scientific text, spanning from words to sections. Furthermore, existing MSMO datasets in general domains often lack labelled images in the training set, which somewhat restricts supervised training for image selection [27]. In contrast to these approaches, our model is designed to grasp the hierarchical semantic structure from words to sections and the nuanced correlations between images and text. Moreover, we will develop a dataset for *Scientific Multimodal Summarization with Multimodal Output* (SMSMO). This dataset aims to facilitate multimodal learning with supervised information in scientific contexts.

3. Problem Definition

In SMSMO, a summarizer takes a paper along with its corresponding images as the input, and generates a *multimodal summary*. This summary encompasses both textual abstract (i.e., a text summary) and graphical abstract (i.e., a representative image for the paper). Formally, each paper input consists of a sequence of word tokens $X_t = w_1, ..., w_m$, and a sequence of paper images $X_i = img_1, ..., img_n$. The output text is a word sequence $Y_t = y_1, y_2, ..., y_t$, while the output image is the representative image $Y_i = img_n$. The summarization model can be viewed as an optimization

problem of its set of trainable parameters (θ):

$$\arg\max_{\theta} MODEL(Y_t, Y_i|X_t, X_i; \theta) \quad (1)$$

4. Our Model

Our SMSMO incorporates multimodal information into scientific summarization, aiming to improve the (summarization) performance and the diversity of generated summaries. On the one hand, the multimodal information, be it image or text, describes the same paper. Hence, these two sets of information can complement each other and enrich summarization. On the other hand, a multimodal summary helps readers save time and read the papers more effectively, with the graphical abstracts help readers to gain a brief, visualized understanding of the paper while the text abstracts provide more details related to it. Currently, cutting-edge scientific summarizers typically consider summaries of a single modality, either text or images (not both) (e.g., [3, 7, 55]). Here, we introduce a cross-modality, multi-task learning model (CMT-Sum). It captures not only the intra-modal features within individual paper text and images, but also their inter-modality correlation.

As shown in Figure 2 (left), CMT-Sum comprises three modules: the *Feature Encoder* encodes the intra-modal features of images/text in individual papers; the *Cross Fusion Module* (**CFM**) learns cross-modality correlation and fuses the intra- and inter-modal features; the *Multimodal Objective Generator* (**MOG**) utilizes the fused features to output the text abstract and chooses the key image as the graphical abstract for individual papers. Additionally, it computes a fine-grained alignment score (Image-Text Matching loss) between images and text.

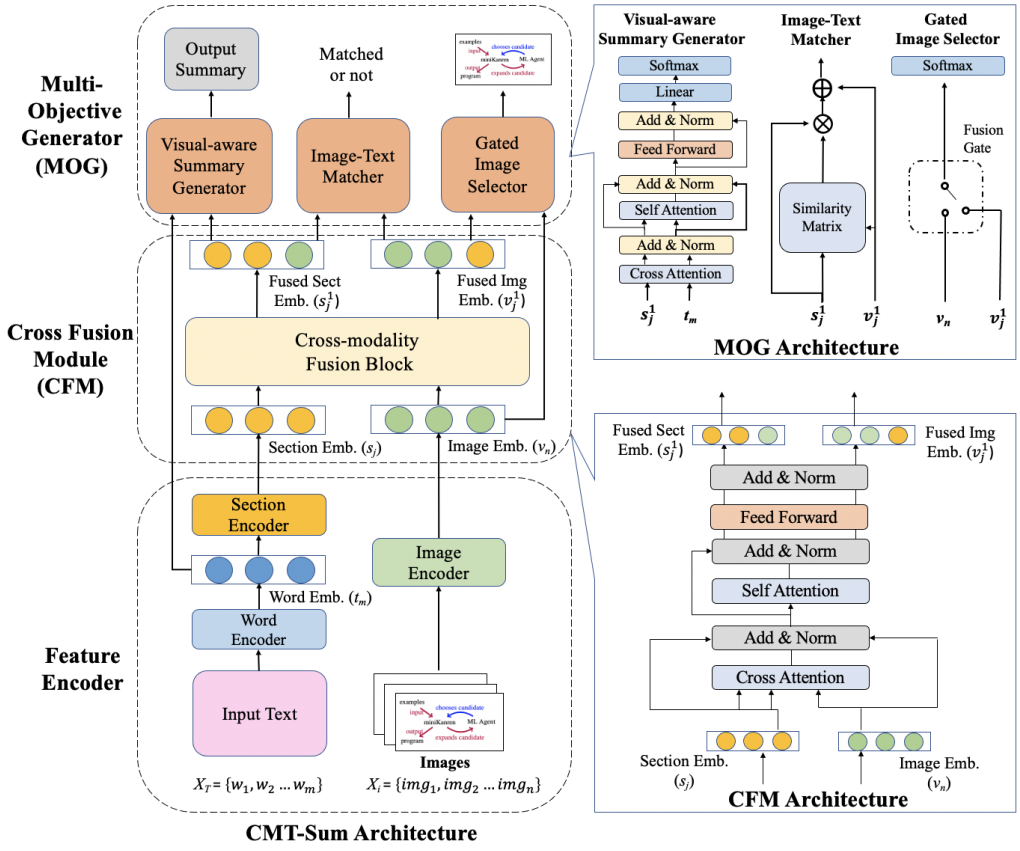


Figure 2: The overview of our CMT-Sum.

4.1. Feature Encoder

To encode the intra-modality features in images and text, we deploy a Feature Encoder which contains an *Image Encoder* and a *Text Encoder*.

4.1.1. Image Encoder

Given a set of paper images $X_i = \{img_1, img_2, ..., img_n\}$, we utilize the ResNet-101 model [56] to encode image features. These features are then fed into a Transformer encoder [50] to learn the intra-modal information among individual images. The visual embeddings of the n^{th} image (v_n) is learned as follows:

$$v_n = ResNet\left(img_n\right) \tag{2}$$

4.1.2. Text Encoder

Paper texts are usually long, consisting e.g., 100-200 sentences. Generally, a paper is divided into multiple sections, each describing certain themes. Here, we hierarchically encode paper text. Particularly, a local *word* encoder will encode individual word contents, followed by a global *section* encoder to obtain a sequentially contextualized embedding for each paper, using all the surrounding sections as global context (see Figure 2 bottom left). Intuitively, our hierarchical encoder first absorbs the local word context on each section level, which is then transferred to a global, section-level paper context.

Word Encoder. We utilize the Longformer [57] to encode long paper text with reduced computational costs. Here, the text input is first tokenized and padded to form a fixed-length sequence. The Longformer then captures

contextual word features for each text. The computation of the m^{th} token embeddings within a paper can be expressed as follows:

$$t_m = Longformer(w_m) \tag{3}$$

Section Encoder. We extend the hierarchical encoding scheme in BERTSUM [58] from sentence level to section one. Academic papers often follow the typical IMRaD structure with sections like **I**ntroduction, **M**ethod, **R**esult, and **D**iscussion. This inherent structure can be extracted with off-the-shelf paper parsers like *Grobid* [59]. Then, a `[CLS]` token is added at the start of each section. It collects features for the tokens preceding it. Formally, the token embeddings are mapped into section embeddings as:

$$s_j = \{t_{CLS}, t_1, ..., t_T\}. \tag{4}$$

4.2. Cross Fusion Module (CFM)

A graphical abstract (i.e., visual summary) should cover the main theme of a paper, while the text abstract will also contain the essential information from source articles. Hence, the two sets of information complement each other in the summarization process. Here, we incorporate a cross-fusion module (CFM) to jointly model the visual-textual dependency of the image and text. CFM contains 3 parts: cross-attention, self-attention and feed-forward layers (see Figure 2 bottom right). To fuse the section embeddings $\{s_1, s_2, ..., s_j\}$ and visual embeddings $\{v_1, v_2, ..., v_k\}$, a cross-attention layer is deployed as:

$$\alpha = softmax\ (score(s_j, v_k)) \tag{5}$$

$$CrossAtt_{s \to v} = \alpha\, v_k \tag{6}$$

where s_j is a query section embeddings, v_k is visual image embeddings, and *score* denotes a product function that computes the similarity between individual section and image embeddings. The three layers in CFM are defined as follows:

$$\begin{aligned} s_j^{cross} &= CrossAtt_{s \to v}\left(s_j, \{v_1, v_2, ..., v_k\}\right), \\ s_j^{self} &= SelfAtt_{s \to v}\left(s_j^{cross}, \{s_j^{cross}\}\right), \\ s_j^{out} &= FF\left(s_j^{self}\right) \end{aligned} \tag{7}$$

where s_j^{cross} and s_j^{self} are the results after the cross-attention layer and the self-attention layers (resp.), followed by the feed-forward layers denoted as $FF(\cdot)$.

We learn the inter-modal correlation between text and image using cross-modality attention. The fused representations are denoted as:

$$s_j^1 = CFM\left(s_j, \{v_1, v_2, ..., v_k\}\right) \tag{8}$$

$$v_k^1 = CFM\left(v_k, \{s_1, s_2, ..., s_j\}\right) \tag{9}$$

where s_j^1 is the *image-aware* embeddings of the j^{th} section text after the fusing in CFM, and v_k^1 is the *text-aware* embeddings of the k^{th} images after fusing with text.

4.3. Multimodal Objective Generator (MOG)

Suppose we have the image reference besides the text reference during model training. To utilize the multimodal reference in training, we propose a generator with a multimodal objective function, which considers not only the negative log-likelihood loss of text summary but also a cross-entropy loss for selecting GA and a binary cross-entropy loss on image-text matching. Concretely, we decompose the multimodal summarization into three subtasks: *text summary generation*, *image selection* and *image-text matching* (see Figure 2 top right). The text generator creates the (text) summary; the image selector picks the most relevant figures in the paper as its graphical abstract; and the image-text matcher determines whether an image is related to the text content. Particularly, the text generator employed a hierarchical attention mechanism to enhance its learning of the text features, which combines the local word and the global image-aware section representation (as obtained from the CFM) for decoding the output word at the current state. Concurrently, the key image will be chosen, ensuring that it aligns closely with the semantics of the generated text summary at each stage of the decoding process. Consequently, the target image chosen at the final step aligns closely with the complete semantics of the generated text summary. The matching task captures nuanced text-image expressions within sections, balancing global and local alignment strategies. We apply multi-task learning [60, 61] to train the three subtasks simultaneously. We now describe the task details.

4.3.1. Visual-aware Summary Generation

For summary generation, it needs to incorporate multimodal information. To accomplish this, we designed a hierarchical decoder that initially focuses on the multimodal semantic alignment representation and subsequently directs attention to the existing text summary to extract the relevent context vector for summary generation. Our hierarchical decoder follows the transformer architecture. Specifically, we use the last hidden state of the text representation t_m as the initial state d_0 of the transformer decoder, and the ℓ^{th} generation procedure is:

$$d_\ell = Transformer_{dec}(d_{\ell-1}, y_{\ell-1}, C_{\ell-1}), \tag{10}$$

where d_ℓ denotes the hidden state at ℓ^{th} decoding step, $y_{\ell-1}$ denotes the previous output and $C_{\ell-1}$ is the context vector. Here, we want our context vector to benefit from both the word representation (t) and the image-aware section representation (s^1). Hence, we deploy a hierarchical attention mechanism over the two representations, computing a higher-level context vector. Particularly, we first compute the cross-attention weight β^{sec} between the section content s^1 and the last decoding state $d_{\ell-1}$:

$$\beta^{sec} = softmax\left(score(s^1, d_{\ell-1})\right). \tag{11}$$

where the last decoding state $d_{\ell-1}$ is derived from the decoder input $y_{\ell-1}$, and is combined with the image-aware section representation s^1 to compute the section-relevant score for the input source text. The scores are constrained to a range of 0 to 1 using the *softmax* to obtain section-relevant attention weights β^{sec}. The weight captures the dependency between the decoding state

and individual source sections, which can be an indicator of *section relevancy*. For example, when summarizing research findings, the chart and text in the *Result* section can be more relevant (see the connection between F3, F4, S3 and A3 in Fig. 1). The section-guided attention indicates which section content is relevant when decoding each word. Consequently, we use the section attention to guide the word attention. Formally, the word attention is denoted as:

$$\beta^{word} = softmax\left(\beta^{sec} \cdot score(t, d_{\ell-1})\right) \tag{12}$$

where the last decoding state $d_{\ell-1}$ is combined with the word representation t to compute the word-relevant score, which captures the dependency between the decoding state and individual words. Then, taking the section attention β^{sec} as guide/condition, the attention weight on each word β^{word} can be computed. After that, β^{word} is used to weigh the source word representation t_m to obtain the context vectors:

$$C_\ell = \sum\nolimits_i \beta^{word} t. \tag{13}$$

The context vector (C_ℓ), which contains relevant contents from both the word representation t and the image-aware section representation s^1, are concatenated with the decoder state d_ℓ. A linear layer then uses the concatenated vector to create the probabilities for each word (P_w):

$$d_\ell^{output} = \sigma\left(FF([d_\ell; C_\ell])\right), \tag{14}$$

$$P_w = softmax\left(FF(d_\ell^{output})\right). \tag{15}$$

For the text summary generation task, its loss is computed using negative log-likelihood against the target word y_ℓ:

$$\mathcal{L}_\theta^{GEN} = \sum\nolimits_\ell log P_w(y_\ell). \tag{16}$$

4.3.2. Image Selection

We assume that the importance of an image is related to two aspects: the information conveyed solely in the raw images and the relevancy of the image information that complements/aligns with the text. Hence, the graphical abstract (i.e., the representative image of each paper) is chosen based on two representations, the original image representation (v) and the text-aware image representation v^1. Here, we incorporate a *fusion gate* to weight the two sets of representations. The fusion gate's weight is determined by the last hidden state of the text decoder ($d_{\ell-1}$). That way, the gate uses images as the main guide and text as support to find the salient information. Consequently, the image score is computed as:

$$\gamma = \sigma\left(FF(d_{\ell-1})\right), \tag{17}$$

$$p^{image} = \gamma\, v + (1-\gamma)\, v^1, \tag{18}$$

$$y_i = \sigma\left(FF(p^{image})\right). \tag{19}$$

When generating text summary in eq. 15, the fusion gate is activated to balance the source image representation v and the text-aware image representation v^1. The gate observes the last decoding state $d_{\ell-1}$ during text summary generation to obtain the gating score γ. The score controls whether the image selector focuses more on the original image representation (larger

γ) or the text-aware one (smaller γ). Subsequently, each image's probability p^{image} is calculated, and the highest probable image (y_i) is picked as the graphical abstract. We calculate the loss function for the image selection task as:

$$\mathcal{L}_{\theta}^{IS} = \frac{1}{N} \sum_{i=1}^{N} - \left[\hat{y}_i \, log \, y_i + (1 - \hat{y}_i) \, log(1 - y_i) \right]. \tag{20}$$

The loss function for image selection measures the difference between the predicted image and the ground truth image ($\hat{y}_i$) using cross-entropy. Including image selection with text summary generation seeks to improve the coherence between the text summary and the visual summary, thereby enhancing the accuracy of the ultimate image summary. Consequently, the image selected at the final step of the text summary generation is regarded as the definitive image summary.

4.3.3. Image-Text Matching (ITM)

Unlike generic text (e.g., news), scientific papers are longer and more structured, containing multiple sections in which images and text within the same section often share similar semantic. To capture the section-level semantic alignment between image and text, we proposed the image-text matching (ITM) task to jointly train in our model. ITM helps our model to also consider the sectional image-text alignment information while calculating attention for both image and text. Formally, ITM is defined as follows:

$$y_{ic} = \sigma \left(W_{txt} \, s^1 + W_s \, h_{ic}^{sim} + W_{img} \, v^1 \right) \tag{21}$$

where h_{ic}^{sim} is a similarity matrix whose element denotes the similarity be-

tween individual section text and image representation, W_{txt}, W_s and W_{img} are trainable parameters on the section representation (s^1), similarity matrix (h_{ic}^{sim}) and image representation (v^1) respectively. Intuitively, at each decoding timestep, in addition to the words and images in the papers, our model also attends to the relevant section. The ITM loss is a binary cross-entropy loss that is optimized to predict whether or not individual image-text pair *matches* (i.e., came from the same paper section). We consider an image to belong to the section(s) that has inline-mentioned it (e.g., "Fig. X describes ..."). The loss function for the ITM task is:

$$\mathcal{L}_\theta^{ITM} = -\left[\hat{y}_{ic}\, log\, y_{ic} + (1 - \hat{y}_{ic})\, log\, (1 - y_{ic})\right] \tag{22}$$

where $\hat{y}_{ic}$ is the ground truth label (i.e., 1 for matching pair and 0 otherwise).

4.4. Joint Training

Finally, we jointly trained our CMT-Sum model with the summary generation, image selection and image-text matching. The model simultaneously minimizes the three loss functions:

$$\mathcal{L}_\theta^{TOTAL} = \mathcal{L}_\theta^{GEN} + \mathcal{L}_\theta^{IS} + \mathcal{L}_\theta^{ITM} \tag{23}$$

5. Experimental Settings

5.1. Dataset

Due to the lack of multimodal reference in existing scientific summarization datasets, the gold standard is either pure text or pure images (not both) during the training and validation. Here, we create two new datasets

(namely $\text{AVIATE}_{\text{SMSMO}}$ and $\text{Pubmed}_{\text{SMSMO}}$) from existing scientific summarization datasets to enrich the benchmarks in the SMSMO research area. Table 1 shows our dataset statistics.

	$\text{PubMed}_{\text{SMSMO}}$			$\text{AVIATE}_{\text{SMSMO}}$		
	Train	Valid	Test	Train	Valid	Test
Num. Docs	5167	659	638	1647	205	206
Avg. Num. Words in Articles	4254.77	4225.27	4138.04	4817.20	4858.91	4853.15
Avg. Num. Sections in Articles	15.03	14.41	15.44	13.03	13.17	13.38
Avg. Num. Words in Summary	255.71	250.43	257.55	138.49	137.40	139.82
Avg. Num. Image in Articles	4.50	4.61	4.47	7.07	6.62	6.88

Table 1: Corpus statistics of our dataset.

Yang et al., [3] proposed PubMed, a scientific paper dataset whose figures were annotated for central figure identification. In PubMed, the authors of each paper identified a central figure that represents their papers. We take the central figure as the graphical abstract; we also incorporate the paper text abstract as the ground-truth summary so that the dataset now contains multimodal references, making it suitable for SMSMO task. To train our summarizer with the Image-Text Matching module (see Sec. 4.3.3), we obtained the PDFs of individual papers in PubMed, and extracted their paragraph/section text and images using *Grobid* [59] and *Pdffigures* [62] (resp.). Finally, we obtained 35k paragraph text-image pairs from the dataset. We use the train, valid and test split as provided by Yang et al. [3] (8:1:1). We call this dataset $\text{PubMed}_{\text{SMSMO}}$.

$\text{AVIATE}_{\text{SMSMO}}$ is a modified version based on the AVIATE dataset [7], which took the first step to study the effect of multimodal signals (i.e., pre-

sentation videos) on paper abstract generation. In the AVIATE dataset, presentation videos from 28 social science and computer science conferences were collected and used to create corresponding paper abstracts (text-only). Here, we utilize the paper sources from AVIATE to build our new dataset. We obtained the open PDFs of individual papers and extracted their paragraph text and images using *Grobid* and *Pdffigures* (like we did in $\text{Pubmed}_{\text{SMSMO}}$). We filter out the data examples which contain no images. Then, we employ a heuristic method to generate the pseudo image selection labels for our data. Specifically, in research articles, images that provide summary information are often captioned with keywords like "*overall*, *framework*, *overview*, etc.". Here, we leverage this property and use a list of summary-related keywords[4] to identify the key images for individual papers. We didn't prioritize the keywords, and we picked the image with the caption that contains most of the keywords (In case there is a tie, we pick the larger image[5]). We compare our keyword lists with the ones generated automatically by Rapid Automatic Keyword Extraction (RAKE) [63] and TextRank [64]. We also compare our methods with Order-ranking and ROUGE-ranking proposed by Zhu et al. [23], which extract GA by considering the image's order appearing in the paper and the ROUGE value between individual image captions and the text abstract. For comparison, we take the manually-labelled key figures in $\text{PubMed}_{\text{SMSMO}}$. We compare this ground truth with the results obtained from ours and other methods, achieving a top-3 accuracy of 62%, no-

[4]The full list of keywords are provided in Appendix A, Table A.8.

[5]Typically, images of greater importance are allocated more space in research articles to accommodate the rich content they need to display [3].

tably higher than the one obtained from the RAKE (53%), TextRank (51%), Order-ranking (48%) and ROUGE-ranking (59%). Consequently, we use our keyword list to obtain the key figures in $\text{AVIATE}_{\text{SMSMO}}$. To ensure the test set is reliable, two volunteers are engaged for post-validation, in which they check if the selected figures can represent the paper given its abstract. The inter-annotator agreement amounts to 0.65 Cohen's kappa, which denotes a fair agreement. Using our methods, we get 2,058 data samples with pseudo image selection labels. The data was split into train, validation, and test sets following the 8:1:1 ratio from Atri et al. [7].

5.2. Implementation Detail

Preprocessing. We tokenized all the characters in the source paper text and target summaries with the Longformer's subwords tokenizer [57].

Model. In the text encoder module of our CMT-Sum model, we initialize our embedding matrix using the word embedding of Longformer [57]. It contains 30,522 vocabularies with an embedding dimension of 4,096. The paper text and summaries share the same vocabulary. The paper image feature is extracted by the ResNet-101 encoder [56], which represents each image by a 2048-dimensional vector. We randomly initialize all trainable parameters using a uniform distribution within $[-0.1, 0.1]$.

Training. During training, we configured the model batch size to 5, the learning rate to 0.0001 and the maximum gradient norm to 1.0. Additionally, we set the dropout ratio to 0.1. We employ an Adam [65] optimizer. The experiments are deployed in Pytorch on an NVIDIA RTX A5000 GPU.

Testing. In the testing phase, we configured the decoding beam size as 5. The minimum and maximum decoding lengths were configured to 100

and 300 (resp.). To avoid repetitive trigrams in the generated summaries, we incorporated trigram blocking [66], and set the length penalty and summary coverage penalty as 0.9 and 5 (resp.) as used by Wu et al., [67].

5.3. Baselines and Evaluation

For evaluation, we compare our model performance against different baselines, covering extractive and abstractive approaches, as well as unimodal and multimodal summarization models.

Unimodal Summarization Models. **Lead3** [68] is a widely used extractive baseline that adopts the first three sentences of individual documents as their summary. TextRank [64] is an extractive baseline, which ranks sentences using graph-based similarity and importance scoring. LexRank [69] is a graph-based extractive baseline. It takes individual sentences as nodes, their (sentence) similarity as edges and extracts the key sentences by their similarity scores. Here, we apply TextRank and LexRank on the paper text and image captions to extract the key text and the representative image (**TextRank (with caption)** and **LexRank (with caption)**). **MemSum** [70] is an extractive method that learns summarization as a multi-step episodic Markov Decision Process (MDP) with awareness of the extraction history. **GoSum** [71] is a graph-based summarization method which encodes the sentence states of the source documents using graph neural networks (GNNs), followed by training an agent's action based on its state in a reinforcement learning environment to evaluate and select sentences and produce an extractive summary. **Lodoss** [72] develop a determinantal regularizer to optimize the segmentation and summarization tasks in parallel, ensuring a set of representative and diverse sentences are selected for the summary.

Seq2Seq (RNN) [73] and **Seq2Seq (Transformer)** [58] are both built upon the standard sequence-to-sequence (seq2seq) architecture. Their difference is that Seq2Seq (RNN) employed the Recurrent Neural Network (RNN) encoder-decoder with a global attention mechanism; while the Seq2Seq (Transformer) use a BERT encoder and a transformer-based decoder to learn summarizing text abstractively. **Pointer Generator (PG)** [74] extends the standard seq2seq architecture to enable word generation from the vocabulary as well as copying words directly from the source document. **LongT5** [75] and **Pegasus-X** [76] are the extension of the T5 and Pegasus encoding methods for handling longer input sequences; **DYLE** [77] is an "extract-and-summarize" method that jointly trains an extractor to select key text snippets and a generator to create a summary from those snippets.

Multimodel Summarization Models. **Mulitmodal Transformer (concatenate)** extend our **Seq2Seq (Transformer)** model. It fuses image and textual features by concatenating their feature vectors, and the vectors to a transformer decoder to generate textual summaries. **Multi-BART** [17] fine-tune Bart model with both the source text and figure caption for better multimodal summarization. **VG-BART** [48] enhances text summarization by integrating visual information using a vision-guided multi-head attention mechanism within a pre-trained BART model; MAST [78] employs a hierarchical trimodal attention technique, first computing pairwise attention weights between text and other modalities, then applying second-level attention to these pairwise features. **CFSum** [27] propose a contribution network that selects more important parts of images for multimodal summarization and effectively enhances the multimodal representation for summarization.

MSMO [18] is the first multimodal summarization model with multimodal output, which applies attention to combine the text-image features for better text generation, and the coverage mechanism is used to help select representative images. **MOF** [23] extended the MSMO model, in which it integrates image precision as an additional training loss. **UNMHG** [25] is a unified model which leverages the large language model to both generate text summary and select GA. **SITransformer** [53] utilizes hierarchical attention for capturing topically-aligned image-text features. **MLASK** [15] develop a Dual-level Interaction Summarizer to generate multimodal summarization. **A2Summ** [52] builds upon the transformer framework and learns inter-modality and intra-modality correlations by contrastive losses.

We employ the widely-used ROUGE [79] to evaluate the generated textual summary. We follow previous works (e.g., [8, 9, 74]) by reporting the F_1 scores of ROUGE-1 (R-1), ROUGE-2 (R-2) and ROUGE-L (R-L). These scores are computed using the *pyrouge* package[6]. Furthermore, we evaluate the quality of the chosen key image using the top-1 and top-3 accuracy metrics introduced by Yang et al. [3]. These metrics determine whether the positive sample is correctly identified within the top-1 or top-3 positions of the predictions.

[6]https://github.com/bheinzerling/pyrouge

6. Results

6.1. Main Result

To assess the performance of CMT-Sum, we evaluate it against three types of baselines: Text-only Models (both extractive and abstraction, **SSO**), Multimodal Summarization Models (with text and image as input only, **MSSO**), and Multimodal Summarization with Multimodal Output Models (with text and image as input and output, **MSMO**). Table 2 shows the results. Here, we can obtain several findings. First, we see a better result from the abstractive models, demonstrating the paper summaries in our proposed dataset are generally abstractive in nature; and merely extracting a few sentences from the paper as the summary may not be as effective in capturing the key information.

Second, for the MSSO methods, MAST and CFSum integrate multimodal information through hierarchical attention and word/phrase-image attention (resp.), leading to a more pronounced enhancement in text summarization performance. In contrast, some MSSO methods (e.g., VG-BART, Multimodal Transformer and Multi-BART) perform even worse than the text-only methods when the integration of image semantic information into the text modal is not effectively accomplished. This implies that simply using an attention sub-layer (VG-BART) or concatenating text/image/caption features (Multimodal Transformer and Multi-BART) are not effective in fusing multimodal information, resulting in a decline in performance. In our CMT-Sum, the CFM (Cross Fusion Module) computes multi-modal fusion with self-attention and cross-attention as described in eq. 7, enabling our model to capture both intra-modality (i.e., word and section) and inter-modality

(text and image) correlations within multimodal data. Furthermore, our model benefits from second-level attention computed by eq. 11 and 12, allowing it to selectively attend to relevant word- or section-level semantics for summarization, leading to achieving the best text summary performance.

Finally, for the MSMO methods, we notice that our image selection task in eq. 20 can improve both the visual and textual representation and deepen the degree of multi-modal alignment, resulting in improved accuracy of text summary. Differing from other methodologies that select the key image solely from the text (e.g., UNMHG) or the image hidden state (e.g., MSMO), our image selector chooses the key image by considering the multi-modal semantic alignment representation, computed through the fusion gate in eq. 19. This approach enables our model to pick the key image that captures the essence of both the source text and the source image. Compared to MLASK and MOF, our model incorporates three joint tasks, as computed in eq. 15, 20 and 22, which effectively learns integrated features from the multimodal content. Notably, the selection of the key image considers both the generated text summary and its alignment. Moreover, the quality of the text summary is influenced by the quality of the image selection. This mutual dependency not only enhances the performance of image summaries but also drives improvements in text summaries. Consequently, our model attains superior performance in both tasks.

6.2. Ablation Study

We conducted ablation experiments to assess the impacts of the two fundamental modules in our CMT-Sum: the Cross Fusion Module (CFM) and the Multi-Objective Generator (MOG). Correspondingly, two sets of models

	Model	Pubmed$_{SMSMO}$					AVIATE$_{SMSMO}$				
		R-1	R-2	R-L	Acc@1	Acc@3	R-1	R-2	R-L	Acc@1	Acc@3
Extractive Models (Text only)	Lead3 [80]	16.30	1.01	1.03	-	-	21.71	0.71	11.21	-	-
	LexRank (with caption)	21.21	5.02	12.01	28.46	69.81	25.91	4.42	12.12	13.17	40.48
	TextRank (with caption)	17.88	6.03	11.34	28.77	65.09	16.31	5.82	13.94	17.56	47.80
	Lodoss [72]	19.14	6.33	16.37	-	-	23.18	6.42	19.43	-	-
	MemSum [70]	21.88	7.3	17.76	-	-	28.66	6.08	16.27	-	-
	GoSum [71]	20.45	8.28	18.25	-	-	16.46	6.45	14.93	-	-
Abstractive Models (Text only)	Seq2Seq (RNN)	20.36	3.89	15.12	-	-	28.82	4.78	14.22	-	-
	Seq2Seq (Transformer)	22.12	4.13	17.21	-	-	33.48	5.75	15.7	-	-
	PG [74]	23.67	4.28	16.62	-	-	29.27	4.85	14.72	-	-
	Long-T5 [75]	28.28	7.88	16.23	-	-	30.41	6.95	16.18	-	-
	Pegasus-X [76]	27.46	8.15	16.28	-	-	26.14	6.75	15.26	-	-
	DYLE [77]	30.12	6.85	20.61	-	-	29.05	6.47	25.99	-	-
Multimodal Summarization Models (Text+Image, input only)	Multimodal Transformer (concatenate)	24.85	4.73	20.35	-	-	34.35	6.33	16.24	-	-
	Multi-BART [17]	24.54	6.37	23.04	-	-	24.54	5.83	21.2	-	-
	VG-BART [48]	27.94	7.84	20.83	-	-	24.29	6.55	23.71	-	-
	MAST [78]	28.42	7.91	25.21	-	-	25.12	6.89	24.21	-	-
	CFSum [27]	30.75	8.14	28.67	-	-	28.15	7.06	26.26	-	-
Multimodal Summarization Output Models (Text+Image, input and output)	MSMO [18]	25.32	4.95	21.12	28.62	72.96	32.05	6.17	24.89	21.46	56.09
	MOF [23]	26.75	5.62	23.86	29.12	69.41	32	6.49	17.84	51	67.21
	UNMHG [25]	28.7	6.36	25.88	25.49	51.29	30.74	6.58	25.51	50.24	56.1
	SITransformer [53]	26.5	5.55	23.2	26.9	69.7	24.9	5.57	24.44	57.24	72.5
	A2sum [52]	29.57	7.56	26.71	30.29	73.06	28.59	7.2	25.62	59.51	78.78
	MLASK [15]	31.96	8.72	27.62	27.55	70.07	33.13	6.58	25.83	60	85.37
	CMT-Sum (Ours)	**36.67**	**9.5**	**33.8**	**33.03**	**74.05**	**35.55**	**7.23**	**32.19**	**68.63**	**86.08**

Table 2: The ROUGE and Accuracy scores of all baselines compared on our Pubmed$_{SMSMO}$ and AVIATE$_{SMSMO}$ datasets. The best scores are **bold**.

		Pubmed$_{SMSMO}$					AVIATE$_{SMSMO}$				
Fusion	Tasks	R-1	R-2	R-L	Acc@1	Acc@3	R-1	R-2	R-L	Acc@1	Acc@3
W/o CFM	T	23.05	4.21	21.21	-	-	26.86	4.02	24.09	-	-
	I	-	-	-	26.56	70.6	-	-	-	58.21	78.23
	T+I	31.05	5.98	28.52	27.97	72.64	33.07	5.92	29.72	61.46	80.73
	T+I+M	33.19	6.86	30.83	28.98	73.7	33.71	6.08	30.05	62.93	82.41
With CFM	T	24.37	4.99	22.56	-	-	29.11	4.64	25.9	-	-
	I	-	-	-	28.12	71.23	-	-	-	62.12	80.63
	T+I	36.4	8.72	33.51	29.45	73.7	35.49	7.21	32.08	65.37	83.66
	T+I+M (ours)	**36.67**	**9.5**	**33.8**	**33.03**	**74.05**	**35.55**	**7.23**	**32.19**	**68.63**	**86.08**

Table 3: Ablation study on our modules in CMT-Sum, Cross Fusion Module (CFM) and Multi-Objective Generator (MOG). We compare them with (CFM) or without CFM (w/o CFM), and their performance on training with different tasks: text generation (**T**), image selection (**I**), text and image tasks (**T**+**I**) and text and image tasks with the image-text matching (**T**+**I**+**M**).

are designed, with CFM present or removed from the full model (**W/o CFM** v.s., **with CFM**) and MOG performing individual tasks on text generation (**T**), image selection (**I**), text and image tasks (**T**+**I**) and text and image tasks with the image-text matching (**T**+**I**+**M**):

Table 3 presents our results. Here, we observe that our model performs better when CFM is equipped. Using self-attention and cross-attention (as described in eq. 7), the CFM effectively learns both the intra-modality semantic and the inter-modality correlation between image and text. That way, CFM grounds the image content on the text segments and fuses the text information into individual images, producing an image-aware text representation and a text-aware image representation for the text generation and image selection tasks (resp.). In contrast, when the two types of contents are combined/concatenated directly (i.e., w/o CFM), the model can not effectively learn the modality interaction, which accordingly affects the performance.

In our multi-task ablation experiment, we observed a notable improvement in the ROUGE scores by incorporating the image selector into our model (i.e., Task T v.s., T+I). This finding highlights the essential role of learning the visual subtasks (in eq. 20 and 22) in enhancing the performance of the textual subtask. Indeed, scholarly papers have diverse types of images, covering overview figures, tables, charts, etc. Such diversity introduces noise and irrelevant information. Thus, it is not sufficient to merely fuse the image and text, and assume that all images are beneficial for the summary without considering the potential interference of irrelevant images. In this regard, the inclusion of an image selector (in eq. 20) becomes crucial in ef-

fectively filtering out noisy images and ensuring that only key images and their relevant visual content contribute to the text summarization process. In addition, we also need an effective matching strategy to learn a comprehensive multi-modal representation. Hence, when incorporating the image-text matcher (Task I+T+M), the model can further be enhanced to align multi-modal information (computed as eq. 22), yielding the best overall scores in our experiment. By combining text summarization, image selection, and image-text matching, CMT-Sum effectively learns the multimodal semantics in a more comprehensive manner. In the $\text{AVIATE}_{\text{SMSMO}}$ dataset, the pseudo image reference we construct (using the key-image heuristic) helps generate a better text summary, which indirectly leads to the improvement of ITM (image-text matching). In the $\text{Pubmed}_{\text{SMSMO}}$ dataset, our full CMT-Sum (with all modules included) outperforms others, indicating that it can effectively improve the multimodal summarization when a large-scale dataset with real multimodal reference is available.

6.3. Module Visualization

In this part, we will demonstrate the role of our Cross Fusion Module (CFM) and Multi-Objective Generator (MOG) in the model by evaluating and visualizing their effects.

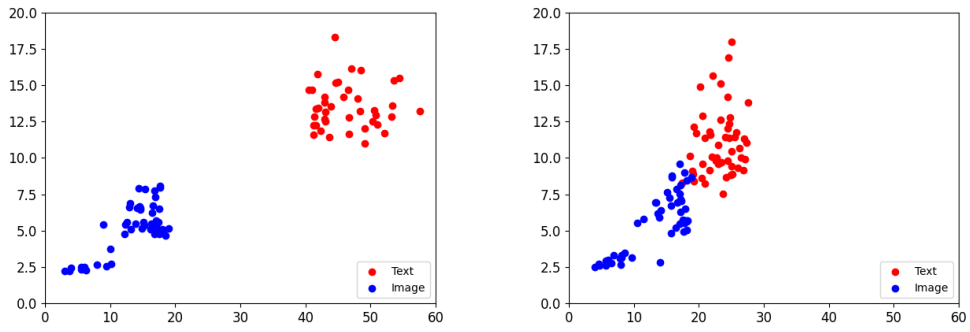
First, we evaluate our CFM module in ensuring a better fusion between image and text output. Here, we compute the Euclidean Distance on the representation of the text summary output (y_t) and the image output (y_i), as follows:

$$ED(text, image) = \sqrt{\left(\frac{1}{m}\sum_{j}^{m} y_t - \frac{1}{n}\sum_{i}^{n} y_i\right)^2} \tag{24}$$

where m and n denote the total number of text and image summaries (resp.); y_t and y_i are the text and image representations produced by our CMT-Sum model. In Table 4, we report the total Euclidean Distance for the multimodal summary on the train, valid and test part of the $AVIATE_{SMSMO}$ dataset. It is evident that via the CFM, text and image features exhibit a relatively consistent fusion across all data subsets, and their feature distance is shorter (i.e., more semantically similar) compared to those without CFM. In Figure 3, we visually present the Euclidean distance measurements for the text and image samples within the semantic space. Specifically, for the same set of samples in the validation dataset, Figure 3a illustrates the representation of the summary text and image generated by a model lacking the CFM module, while Fig. 3b showcases the same sample sets generated by a model with CFM. The representation was processed with PCA for the dimensionality reduction to be displayed in the same semantic space. Thanks to the self-attention and cross-attention (as computed in eq. 7), the CFM learn how much information to integrate from multimodal source and how much information to retain from the original modality. Consequently, it strengthens the correlation between text and image representation, as reflected in Figure 3b. By integrating relevant multimodal information with our CFM, the representations of the text and image samples are closer (i.e., more semantically similar) compared to those without CFM (Figure 3a). It suggests that CFM effectively learn to link text and images, helping the model better understand and process multimodal information.

	Train	Valid	Test
With CFM	29.52	25.64	24.85
W/o CFM	46.91	42.76	36.12

Table 4: The total Euclidean distance on the text and image samples in the $AVIATE_{SMSMO}$ dataset.



(a) Without CFM

(b) With CFM

Figure 3: Euclidean distance measurement on a set of text and image summaries on the validation set of $AVIATE_{SMSMO}$. For the same set of samples, Figure 3a show the text and image representation without processing through the CFM modules, while Figure 3b does. For visualization, the representation was processed using the dimensionality reduction by PCA.

Next, we explore the effectiveness of the MOG. We take the example shown in Figure 1 and visualize the attention maps generated from the hierarchical attention mechanism by our Visual-aware Text Generator (in eq. 15). The map is shown in Figure 4. It displays the weights connecting image and text semantics at each summary generation step, demonstrating how the two modalities complement each other. We inspect the maps generated with both text generation and image selection performed (green row), in comparison with a baseline model which only performs text generation (red row). We randomly pick five source words that appear in the summary text to show. From Figure 4, it can be observed that the baseline model (red row) exhibits a more "sparse" attention distribution, where the word weights are spread across different images (e.g., the word "pre-defined" attends across F2 and F4). Conversely, our model's attention map focuses more on the images corresponding to the text in that section (e.g., the word "boundary" attends mostly on F1), thanks to the integration of the image selector. Particularly at each decoding step, when the text generator summarises the content relating to a particular section (e.g., Methodology), the image selector simultaneously identifies the most relevant visual content (e.g., a schematic diagram). This real-time, step-by-step alignment ensures the visual context closely matches the textual content being generated. It reinforces the semantic context of the current section and helps the text generator maintain focus on the specific section/theme being summarized. In cases where textual content might be ambiguous, the selected images (or attended image representation) can provide clarifying information, guiding the text generator towards more precise language and descriptions. Later on, when the text generator moves from

one section (or decoding step) to another, the changing image selections help shift its focus accordingly. This adaptive mechanism ensures that the generated summary maintains relevance throughout its length. Additionally, since the final image summary is selected based on the last decoding context, the image summary will include complete semantics of the decoded context.

In Figure 5, we explore the effectiveness of the ITM by observing the relevance of the section text and images shown in Figure 1 (i.e., S1 to S3 and F1 to F4). Each colour block denotes the cosine similarity between the image-aware text representation (S_j^1) and the text-aware image representation (v_j^1) of each section. The darker colour refers to a higher similarity in the heatmap. By training with the ITM alignment loss (eq. 22), our model learns a multi-modal semantic alignment representation for the section's text and image content. From Figure 5a, we can see that by aligning multimodal relevant information with our ITM, the image-text similarity is more relatively concentrated along each section (e.g., F1-S1, F2-S2) as compared to the one without ITM (Figure 5b).

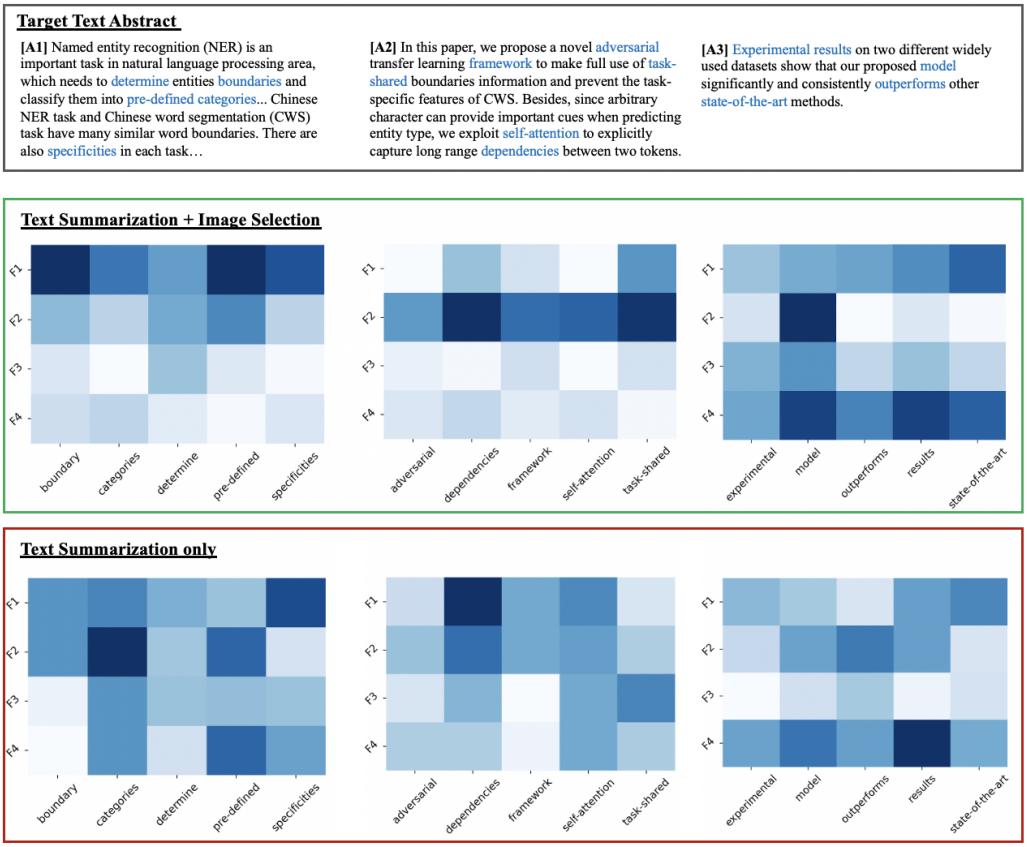
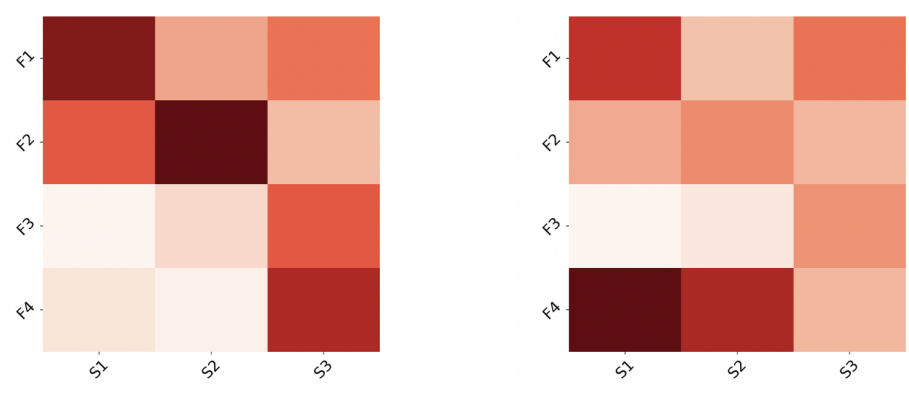


Figure 4: Visualization of the attention map in our Visual-aware Text Generator, generated with both text generation and image selection performed (in green row), in comparison with a baseline model which only performs text generation (in red row). For the abstract segment A1 to A3 we presented in Figure 1, we show five words (in blue) that appear in the source text and their associated attention map with the images in the paper. The baseline shows a "sparse" distribution across different images and words. In contrast, our model shows a more concentrated distribution on related images and text corresponding to the abstract section/theme.



(a) With ITM **(b) W/o ITM**

Figure 5: The heatmap demonstrates the cosine similarity among the representations of images (F1 to F4) and sections (S1 to S3) presented in Figure 1. Map (a) is generated by a model with ITM incorporated, whereas map (b) lacks ITM.

6.4. Human Evaluation

A human evaluation is designed to analyze the summary output from three aspects. Informativeness (**Inf**) assesses whether the summary contains sufficient and necessary information from the input. Coherence (**Coh**) assesses whether the summarized content is presented in a coherent order. Accuracy (**Acc**) assesses the correctness and clarity of the summary content. From our AVIATE$_{\text{SMSMO}}$ dataset, we randomly selected 50 examples generated by our comparison methods. Three graduate students volunteered to evaluate the output. They are tasked to score each output from 0 to 5, where 0 and 5 indicate the lowest and highest scores of corresponding metrics. The final results are averaged across subjects.

Table 5 presents the results of the human evaluation. Our CMT-Sum achieves the highest scores in all metrics. The results indicate that summaries

		Inf	Coh	Acc
SSO	MemSum	1.42	1.6	1.25
	Long-T5	1.24	2.15	1.78
	Pegasus-X	1.6	2.10	2.12
	DYLE	1.11	2.14	2.01
MSSO	VG-BART	1.74	2.13	2.01
	MAST	1.86	2.42	2.51
	CFSum	2.02	2.45	2.52
MSMO	MSMO	2.34	2.41	2.51
	MOF	2.71	2.52	2.67
	UNMHG	2.84	2.63	2.15
	SITransformer	2.81	2.02	2.14
	A2sum	2.76	2.15	2.86
	MLASK	2.85	2.57	2.96
	CMT-Sum (ours)	**3.31**	**3.06**	**3.12**

Table 5: Human evaluation results of generated summary.

generated by CMT-Sum are more informative and cohesive, with high accuracy on both text and image information summarized. Compared with the SSO and MSSO models with text-only output, CMT-Sum provides a relevant image, which can include diagrams, graphs, or illustrations that complement the textual summary, offering additional information that might not be explicitly stated in the text. Additionally, our model's CFM Module (in eq. 7 plays a role in enhancing the integration of information across different levels. Particularly, when our text generator produces each word, it can draw upon both sectional (intra-modality) and visual (inter-modality) information. This integration allows for a more comprehensive understanding of the content, potentially leading to more accurate and detailed summaries. Finally, our step-by-step alignment of text generation with image selection (in

eq. 20) helps maintain temporal coherence in the summary, which ensures that visual references in the text accurately correspond to the content being discussed.

6.5. Case Study and Relevance Visualization

Table 6 and 7 present the summary outputted by different models. We also include the original abstract for reference (top line in Table 6). Table 6 displays the text summary generated by the SSO models (Seq2seq and DYLE) and MSSO models (MAST and CFSum). We can see that the SSO models, which only incorporate text information, neglect some concepts that presented in the image (e.g., the nil-aware passage extractor). In contrast, MAST, CFSum, and our CMT-Sum all utilize multimodal input, allowing them to consider both text and image semantics. However, MAST and CFSum focus on either global or local correspondences, but not both. Particularly, MAST maps the entire document and its images into a single shared space, capturing overall themes but potentially overlooking important details. For example, the model mentioned the "nill-aware answer extraction framework" and the "evidence-decomposition", but there is not much description of them. On the other hand, CFSum focuses on word-/phrase-level fusion, which good at capturing details but potentially overlooking the overall context. We can see that CFSum described a lot of model detail (e.g., "matching the first Q tokens with the second Q tokens", which is the model details of the nill-aware answer extractor). By comparison, our CMT-Sum designs a fusion approach that operates at both word and section levels, incorporating intra-text and inter-text-image fusion through a hierarchical structure. In brief, the unimodal encoders (in eq. 2 and 3) capture the intra-modality features

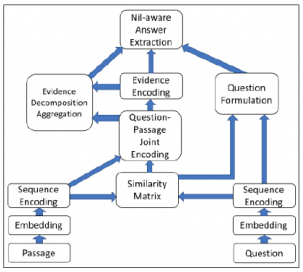
Model	Output	
Ground-truth	Recently, there has been a surge of interest in reading comprehension-based (RC) question answering (QA). However, current approaches suffer from an impractical assumption that every question has a valid answer in the associated passage. A practical QA system must possess the ability to determine whether a valid answer exists in a given text passage. In this paper, we focus on developing QA systems that can extract an answer for a question if and only if the associated passage contains an answer. If the associated passage does not contain any valid answer, the QA system will correctly return Nil. We propose a nil-aware answer span extraction framework that is capable of returning Nil or a text span from the associated passage as an answer in a single step. We show that our proposed framework can be easily integrated with several recently proposed QA models developed for reading comprehension and can be trained in an end-to-end fashion. Our proposed nil-aware answer extraction neural network decomposes pieces of evidence into relevant and irrelevant parts and then combines them to infer the existence of any answer. Experiments on the NewsQA dataset show that the integration of our proposed framework significantly outperforms several strong baseline systems that use pipeline or threshold-based approaches.	
Seq2seq (Transformer, text-only)	Question answering (QA) is a challenging problem in natural language understanding. QA systems are typically built where each step is a question and its qa. However, most existing QA approaches have difficulty scaling over the QA problem. In this paper, we propose a novel end-to-end neural network model that enables multi-hop reasoning over long short-term memory representations. We conducted on the recently proposed cross-hop reading comprehension dataset, which demonstrates its usefulness as a benchmark for multi-passage QA model. Multi-choice reading comprehension shows that our approach significantly outperforms state-of-the-art QA models. Moreover, it was proposed for the QA framework including a large margin. Ablation studies find whether positional information is inherently encoded in the corresponding question. We propose multi-answer QA datasets containing human judgments of multihop reasoning. We evaluate our proposed framework on the benchmark dataset and the multi-hop	
DYLE	In this paper, we present a new machine reading comprehension system for machine comprehension. The system is based on a large cloze-style MC dataset, where the answer span of the associated passage contains a valid answer for each question. In this system, the question span of a given passage is modeled as a single-choice answer for every question given by the relevant passage. We show that this system can answer questions about the context of a given question, such as whether the answer is correct or not. We also show that the baseline system can be used to answer the question by answering the question with a single sentence. Our results show that our system outperforms other systems in the MRC dataset.	
MAST	We propose a nil-aware answer extraction framework for multiple-choice question answering (QA). In this paper, we propose is based on evidence decomposition-based evidence-decomposition (EDD). In this paper, we explore the problem of determining whether a text passage contains a valid answer for a question. We investigate whether the answer to a given question can be found in the context of a single text passage. In addition, we examine whether the answers to the same question are valid for the same text passage, and if the answer is not valid for all questions. We also propose a large cloze-style multiple-choice dataset, which returns Nil as answer. We show that our framework is capable of answering multiple questions at the same time.	
CFSum	In this paper, we consider the problem of answering a question in a machine reading comprehension (MRC) dataset, in which a question is answered by a single candidate answer (i.e., the first k tokens). We propose a novel approach for answering the question in the context of the context, which is based on the fact that the question has only a single token-limited answer (e.g., s 1). We show that this approach is able to answer the question with the first K tokens, and the second k tokens of the sentence, by matching the first Q tokens of a given sentence with the second Q token of the given sentence. The first k token of a sentence, which has a single Q token, can be easily attended to by a system that can easily attend to it by watching word overlaps. In addition to this, we show that it is possible to obtain the correct answer by watching the second K tokens of other sentences, and that it can also attend to the third K token of another sentence. We demonstrate that this is the case in a multi-sentence context, where only one Q token appears in other sentences. In this context, the question can be answered with only one answer, that is, by looking at the	

Table 6: Summary output comparison between CMT-Sum and SSO (Seq2seq and DYLE) and MSSO (MAST and CFSum) baselines. In the comparison, valid and relevant contexts w.r.t the ground truth are highlighted in blue, while irrelevant or incorrect contexts are highlighted in red.

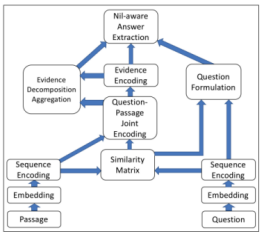
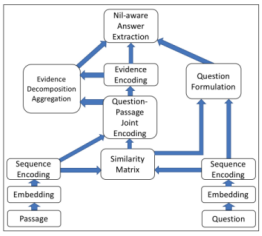
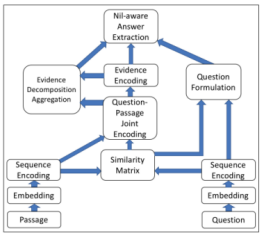
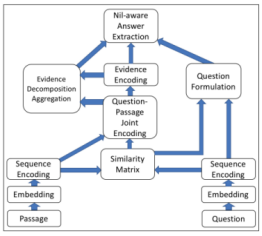
Model	Output	
MSMO	We present a scalable answer extraction answering upper bound for reading comprehension which simultaneously learning the passage of model (referred by the representation QA rate of the input sentences used in a document). We show that this new encoding provides intuitive, related classification and smoothing preferable classifiers by reducing syntax extraction for span information. Furthermore, we demonstrate its generalization over different span types of span recognition, and our method called several benchmark	<table><tr><th colspan="2">Dataset</th><th>#Passages</th><th>#Questions</th></tr><tr><td rowspan="2">Train</td><td>NewsQA</td><td rowspan="2">10,938</td><td>92,549</td></tr><tr><td>+Nil Qs</td><td>107,673</td></tr><tr><td rowspan="2">Dev</td><td>NewsQA</td><td rowspan="2">638</td><td>5,166</td></tr><tr><td>+Nil Qs</td><td>5,988</td></tr><tr><td rowspan="2">Test</td><td>NewsQA</td><td rowspan="2">632</td><td>5,126</td></tr><tr><td>+Nil Qs</td><td>5,971</td></tr></table>
MOF	In this paper, we propose a multi-choice question answering (QA) framework based on CNN and Daily Mail news articles summaries. The framework is based on a large cloze-style MC dataset, where the answer span of the associated passage contains a valid answer for each question. In this model, the question span of a given passage is modeled as a single-choice answer for every question given by the relevant passage. In contrast to previous models, existing QA models rely on multiple-choice questions given by a single passage. We propose a model that can be trained on multiple passages without accessing the original full text. In addition, the model is trained on the relevant passages, which are annotated by a higher-level encoding layer and annotated using relevant and relevant relevant components. Experiments show that our approach consistently outperforms current state-of-the-art, even when no answer is available.	
UNMHG	The introduction of large scale machine comprehension datasets with machine comprehension models that are highly accurate and efficient in answering questions given raw texts has been proposed recently. While conventional machine comprehension models were given a paragraph that always contains an answer to a question, some researchers have extended the models to an open-domain setting where relevant documents have to be searched from an extremely large knowledge source such as Wikipedia. We introduce a new neural model for QA that is able to infer the answer of each question given a single sentence. The model is capable of inferring the answer for any question given the associated passages. However, it is difficult to infer whether the answer is valid for any given passage given a sentence.	
MLASK	Machine comprehension (MC) systems process the process of reading comprehension (RC) by answering questions after understanding natural language text. Several datasets and resources have been developed recently. However, none of the models considered nil questions, although it is crucial for a practical QA system to be able to determine whether a text passage questions a valid answer for a question. In this paper, we focus on developing QA systems that extract an answer for a question if and only if the associated passage contains a valid. In this paper, we propose a nil-aware answer extraction framework which returns Nil or a span of text as answer, when integrated with end-to-end neural MC models. Our proposed framework is able to extract an answer for a question if and only if the associated passage contains a valid answer. In this work, we show that our framework is capable of extracting an answer from a text passage. We demonstrate that our proposed framework can be used to extract the answer from text passages.	
CMT-Sum (Ours)	Question-answering (QA) is a challenging task. In this paper, we propose a new end-to-end neural reading comprehension model which aims to perform a question-answering task by enforcing complete evidence from a paragraph that contains a question under relevant passages. We present a novel framework for jointly extractive QA-named entity recognition and question answering, where the question is represented by a passage-aware question answering (AMA). To this end, we extend a neural network framework to train models for the-art question-answering models. Experimental results on several benchmark datasets show that our method achieves state-of-the-art performance, where the proposed method significantly outperforms the baseline by a large margin for the proposed multi-hop reading comprehension datasets. The experimental results demonstrate that our proposed method outperforms several baselines on several large-scale wikihop QA datasets.	

Table 7: Summary output comparison between CMT-Sum multimodal output baselines (MSMO, MOF, UNMHG, MLASK and ours). In the comparison, valid and relevant contexts w.r.t the ground truth are highlighted in blue, while irrelevant or incorrect contexts are highlighted in red.

separately for text and image inputs. Then, the CFM in eq. 7 fuse the inter-modality feature of the two (feature) sets at a sectional level. Finally, with the hierarchical attention in eq. 11 and 12, the model generates the target summary by jointly considering both local-level features (word-specific) and global-level features (section-specific) through a hierarchical attention mechanism. As can be seen in Table 7, the summary outputted by our model incorporates both global structure (e.g., Problem scope/aim, model description and experiment results), as well as local detail description (e.g., describe the purpose of the evidence aggregation).

In comparison to the text-only output, the multimodal output in Table 7 provides additional information by including a selected image along with the text summary. This additional visual element serves to reinforce or clarify the textual content, and vice versa, leading to better overall comprehension. However, the effectiveness of this approach depends on the relevance of the chosen image to the textual content. If a model selects an image that is contextually irrelevant to the text summary, it may cause ambition and confuse readers (see e.g., the output from MSMO). MSMO selects the image summary by observing only the image's hidden state. In contrast, our image selector identifies the key image by considering the multi-modal semantic alignment representation (as computed by eq. 19). The generated text provides the image encoder with richer global representations, comprising the full semantic content of both the source image and text. Other than that, when compared to MLASK, our model incorporates a multi-task module that includes image-text matching, as computed in eq. 22. It enables both the text generator and image selector to better align and more effectively learn fused

features from the multimodal content. Particularly, the image-text matcher improves the alignment between the visual and textual elements of each section (see Figure 5), enabling the text generator to create summaries that are more accurately tailored to each particular section of the source material. From our case in Table 7, the "evidence aggregation module" overlooked in MLASK has been covered in our summary (i.e., enforcing complete evidence from a paragraph).

7. Conclusion

This paper introduces a new model for scientific summarization that leverages cross-modality and multi-task learning techniques. Our model effectively improves multimodal summary generation and the diversity of the generated summaries, encompassing both text and image information in scientific papers. The novelty of our paper lies in the finer-grained fusion of the two modalities through our cross-fusion module, as well as the generation of aligned multimedia summaries that capture the semantics of different modalities through our multi-objective generator. This approach distinguishes our work from existing studies in scientific NLP, which often handle modalities independently and primarily focus on text content. Our research complements current research, which mainly builds upon text-only corpora (and lacks multimodal semantics). Experimental results demonstrate that our multimodal model generates summaries that are more coherent, informative, and accurate, showcasing the effectiveness of our approach.

Acknowledgements

Funding: The work is supported by LEO Dr David P. Chan Institute of Data Science, the Hong Kong RGC ECS (LU23200223/130393), the Lam Woo Research Fund (LWP20018/871232), the Direct Grant (DR23A9/101194), the Faculty Research Grants (DB23B5/102083 and DB23AI/102070) and the Research Seed Fund (102241) of Lingnan University, Hong Kong.

Appendix A. Keywords used to identify key figures in $\text{AVIATE}_{\text{SMSMO}}$

Keywords
flow chart, flowchart, illustration, general block diagram, system structure, system architecture, overall, overview, framework, workflow, structure, flow, demonstration, graphic visualization, graphical (model), theoretical model

Table A.8: Here, we present the keywords that we use to identify the key figures in our $\text{AVIATE}_{\text{SMSMO}}$ dataset. The key image of individual papers is determined by the number of keywords each image caption contains. If there is a tie, the image that appears earlier in the paper will be taken. Images which can not align with any keywords are excluded.

References

[1] Stanford, Stanford ai index report, https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf (2023).

[2] J. Yoon, E. Chung, An investigation on graphical abstracts use in scholarly articles, International Journal of Information Management 37 (1) (2017) 1371–1379.

[3] S. T. Yang, P.-S. Lee, L. Kazakova, A. Joshi, B. M. Oh, J. D. West, B. Howe, Identifying the central figure of a scientific paper, in: 2019 International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2019, pp. 1063–1070.

[4] C. D. Paice, The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases, in: Proceedings of the 3rd annual ACM conference on Research and development in information retrieval, 1980, pp. 172–191.

[5] S. Teufel, M. Moens, Summarizing scientific articles: experiments with relevance and rhetorical status, Computational linguistics 28 (4) (2002) 409–445.

[6] X. Chen, H. Alamro, M. Li, S. Gao, R. Yan, X. Gao, X. Zhang, Target-aware abstractive related work generation with contrastive learning, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 373–383.

[7] Y. K. Atri, S. Pramanick, V. Goyal, T. Chakraborty, See, hear, read:

Leveraging multimodality with guided attention for abstractive text summarization, Knowledge-Based Systems 227 (2021) 107152.

[8] Q. Xie, J. A. Bishop, P. Tiwari, S. Ananiadou, Pre-trained language models with domain knowledge for biomedical extractive summarization, Knowledge-Based Systems 252 (2022) 109460.

[9] X. Chen, H. Alamro, M. Li, S. Gao, X. Zhang, D. Zhao, R. Yan, Capturing relations between scientific papers: An abstractive model for related work section generation, Association for Computational Linguistics, 2021.

[10] Y. Dong, A. Mircea, J. C. K. Cheung, Discourse-aware unsupervised summarization for long scientific documents, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021, pp. 1089–1102.

[11] I. Cachola, K. Lo, A. Cohan, D. Weld, Tldr: Extreme summarization of scientific documents, Findings of EMNLP (2020).

[12] K. Luu, X. Wu, R. Koncel-Kedziorski, K. Lo, I. Cachola, N. A. Smith, Explaining relationships between scientific documents, 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), 2021, pp. 2130–2144.

[13] M. La Quatra, L. Cagliero, Transformer-based highlights extraction from scientific papers, Knowledge-Based Systems 252 (2022) 109382.

[14] Y. Du, Q. Li, L. Wang, Y. He, Biomedical-domain pre-trained language model for extractive summarization, Knowledge-Based Systems 199 (2020) 105964.

[15] M. Krubiński, P. Pecina, Mlask: multimodal summarization of video-based news articles, in: Findings of the Association for Computational Linguistics: EACL 2023, 2023, pp. 910–924.

[16] B. M. Yao, A. Shah, L. Sun, J.-H. Cho, L. Huang, End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 2733–2743.

[17] K. Overbay, J. Ahn, J. Park, G. Kim, et al., mredditsum: A multimodal abstractive summarization dataset of reddit threads with images, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 4117–4132.

[18] J. Zhu, H. Li, T. Liu, Y. Zhou, J. Zhang, C. Zong, Msmo: Multimodal summarization with multimodal output, in: Proceedings of the 2018 conference on empirical methods in natural language processing, 2018, pp. 4154–4164.

[19] Editage, How are graphical abstracts effective in academic communication?, https://www.editage.com/services/graphical-abstract-design-visual-abstract-services (2024).

[20] J. Chen, H. Zhuge, Extractive summarization of documents with images based on multi-modal rnn, Future Generation Computer Systems 99 (2019) 186–196.

[21] K. Liu, Y. Li, N. Xu, P. Natarajan, Learn to combine modalities in multimodal deep learning, arXiv preprint arXiv:1805.11730 (2018).

[22] M. Xiao, J. Zhu, F. Zhai, Y. Zhou, C. Zong, Diusum: Dynamic image utilization for multimodal summarization, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, 2024, pp. 19297–19305.

[23] J. Zhu, Y. Zhou, J. Zhang, H. Li, C. Zong, C. Li, Multimodal summarization with guidance of multimodal reference, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 9749–9756.

[24] S. Phani, A. Abdul, M. K. S. Prasad, H. K. D. Sarma, Mmsft: Multilingual multimodal summarization by fine-tuning transformers, IEEE Access (2024).

[25] M. Krubiński, P. Pecina, Towards unified uni-and multi-modal news headline generation, in: Findings of the Association for Computational Linguistics: EACL 2024, 2024, pp. 437–450.

[26] L. Zhang, X. Zhang, J. Pan, Hierarchical cross-modality semantic correlation learning model for multimodal summarization, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 11676–11684.

[27] M. Xiao, J. Zhu, H. Lin, Y. Zhou, C. Zong, Cfsum: Coarse-to-fine contribution network for multimodal summarization, in: The 61st Annual Meeting Of The Association For Computational Linguistics, 2023.

[28] L. Jin, J. Chen, Self-supervised opinion summarization with multi-modal knowledge graph, Journal of Intelligent Information Systems 62 (1) (2024) 191–208.

[29] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, N. Goharian, A discourse-aware attention model for abstractive summarization of long documents, in: Proceedings of NAACL-HLT, 2018, pp. 615–621.

[30] Y. Guo, W. Qiu, Y. Wang, T. Cohen, Automated lay language summarization of biomedical scientific reviews, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 160–168.

[31] M. Yasunaga, J. Kasai, R. Zhang, A. R. Fabbri, I. Li, D. Friedman, D. R. Radev, Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 33, 2019, pp. 7386–7393.

[32] J. Pilault, R. Li, S. Subramanian, C. Pal, On extractive and abstractive neural document summarization with transformer language models, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 9308–9319.

[33] Y. Lu, Y. Dong, L. Charlin, Multi-xscience: A large-scale dataset for extreme multi-document summarization of scientific articles, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 8068–8074.

[34] Elsevier, How to produce a good visual abstract, tools and resources for authors, https://www.elsevier.com/authors/tools-and-resources/visual-abstract (2021).

[35] J. Im, M. Kim, H. Lee, H. Cho, S. Chung, 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), 2021, pp. 388–403.

[36] C. Zhu, R. Xu, M. Zeng, X. Huang, A hierarchical network for abstractive meeting summarization with cross-domain pretraining, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 194–203.

[37] H. Li, J. Zhu, C. Ma, J. Zhang, C. Zong, 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, 2017, pp. 1092–1102.

[38] M. Lu, Y. Liu, X. Zhang, A modality-enhanced multi-channel attention network for multi-modal dialogue summarization, Applied Sciences 14 (20) (2024) 9184.

[39] D. Argade, V. Khairnar, D. Vora, S. Patil, K. Kotecha, S. Alfarhood, Multimodal abstractive summarization using bidirectional encoder representations from transformers with attention mechanism, Heliyon 10 (4) (2024).

[40] X. Fu, J. Wang, Z. Yang, Multi-modal summarization for video-containing documents, arXiv preprint arXiv:2009.08018 (2020).

[41] K. Yu, C. Zhang, J. Ding, Y. Yue, Y. Wu, Multimodal dialogue response generation based on selective attention and gating mechanisms (2023).

[42] Z. Zhang, J. Wang, Z. Sun, Z. Yang, Lams: a location-aware approach for multimodal summarization (student abstract), in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 15949–15950.

[43] H. Li, P. Yuan, S. Xu, Y. Wu, X. He, B. Zhou, Aspect-aware multimodal summarization for chinese e-commerce products, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 8188–8195.

[44] L. Jing, Y. Li, J. Xu, Y. Yu, P. Shen, X. Song, Vision enhanced generative pre-trained language model for multimodal sentence summarization, Machine Intelligence Research 20 (2) (2023) 289–298.

[45] Y. Liu, L. Qiao, C. Lu, D. Yin, C. Lin, H. Peng, B. Ren, Osan: A one-stage alignment network to unify multimodal alignment and unsupervised domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3551–3560.

[46] C. Jiang, H. Xu, W. Ye, Q. Ye, C. Li, M. Yan, B. Bi, S. Zhang, F. Huang, J. Zhang, Copa: Efficient vision-language pre-training through collaborative object-and patch-text alignment, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 4480–4491.

[47] H. Li, J. Zhu, J. Zhang, X. He, C. Zong, Multimodal sentence summarization via multimodal selective encoding, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 5655–5667.

[48] T. Yu, W. Dai, Z. Liu, P. Fung, Vision guided generative pre-trained language models for multimodal abstractive summarization, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 3995–4007.

[49] C. Suman, A. Naman, S. Saha, P. Bhattacharyya, A multimodal author profiling system for tweets, IEEE Transactions on Computational Social Systems 8 (6) (2021) 1407–1416.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[51] J. Qiu, J. Zhu, M. Xu, F. Dernoncourt, T. Bui, Z. Wang, B. Li, D. Zhao, H. Jin, Mhms: Multimodal hierarchical multimedia summarization, arXiv preprint arXiv:2204.03734 (2022).

[52] B. He, J. Wang, J. Qiu, T. Bui, A. Shrivastava, Z. Wang, Align and attend: Multimodal summarization with dual contrastive losses, in: Pro-

ceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 14867–14878.

[53] S. Liu, L. Wang, X. Zhu, X. Lu, Z. Wang, K. Hu, Sitransformer: Shared information-guided transformer for extreme multimodal summarization, arXiv preprint arXiv:2408.15829 (2024).

[54] Z. Zhang, X. Meng, Y. Wang, X. Jiang, Q. Liu, Z. Yang, Unims: A unified framework for multimodal summarization with knowledge distillation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 11757–11764.

[55] Y. K. Atri, V. Goyal, T. Chakraborty, Fusing multimodal signals on hyper-complex space for extreme abstractive text summarization (tldr) of scientific contents, KDD23: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data MiningAugust 2023 (2023).

[56] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[57] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150 (2020).

[58] Y. Liu, M. Lapata, Text summarization with pretrained encoders, 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, pp. 3730–3740.

[59] Grobid, Grobid parser, https://github.com/kermitt2/grobid (2020).

[60] R. Collobert, J. Weston, A unified architecture for natural language processing: Deep neural networks with multitask learning, in: Proceedings of the 25th international conference on Machine learning, 2008, pp. 160–167.

[61] R. Caruana, Multitask learning, Springer, 1998.

[62] C. Clark, S. Divvala, Pdffigures 2.0: Mining figures from research papers, in: Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, 2016, pp. 143–152.

[63] S. Rose, D. Engel, N. Cramer, W. Cowley, Automatic keyword extraction from individual documents, Text mining: applications and theory (2010) 1–20.

[64] R. Mihalcea, P. Tarau, Textrank: Bringing order into text, in: Proceedings of the 2004 conference on empirical methods in natural language processing, 2004, pp. 404–411.

[65] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[66] R. Paulus, C. Xiong, R. Socher, A deep reinforced model for abstractive summarization, in: International Conference on Learning Representations.

[67] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al., Google's neural machine translation system: Bridging the gap between human and machine translation, arXiv preprint arXiv:1609.08144 (2016).

[68] R. Nallapati, F. Zhai, B. Zhou, Summarunner: A recurrent neural network based sequence model for extractive summarization of documents, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 31, 2017.

[69] G. Erkan, D. R. Radev, Lexrank: Graph-based lexical centrality as salience in text summarization, Journal of artificial intelligence research 22 (2004) 457–479.

[70] N. Gu, E. Ash, R. Hahnloser, MemSum: Extractive summarization of long documents using multi-step episodic Markov decision processes, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 6507–6522. doi:10.18653/v1/2022.acl-long.450.
URL https://aclanthology.org/2022.acl-long.450

[71] J. Bian, X. Huang, H. Zhou, S. Zhu, Gosum: Extractive summarization of long documents by reinforcement learning and graph organized discourse state, arXiv preprint arXiv:2211.10247 (2022).

[72] S. Cho, K. Song, X. Wang, F. Liu, D. Yu, Toward unifying text seg-

mentation and long document summarization, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 106–118.

[73] M.-T. Luong, H. Pham, C. D. Manning, Effective approaches to attention-based neural machine translation, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1412–1421.

[74] A. See, P. J. Liu, C. D. Manning, Get to the point: Summarization with pointer-generator networks, in: R. Barzilay, M. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, Association for Computational Linguistics, 2017, pp. 1073–1083. doi:10.18653/v1/P17-1099.
URL https://doi.org/10.18653/v1/P17-1099

[75] M. Guo, J. Ainslie, D. Uthus, S. Ontanon, J. Ni, Y.-H. Sung, Y. Yang, Longt5: Efficient text-to-text transformer for long sequences, Findings of the Association for Computational Linguistics: NAACL 2022 (2022).

[76] J. Phang, Y. Zhao, P. J. Liu, Investigating efficiently extending transformers for long input summarization, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 3946–3961.

[77] Z. Mao, C. H. Wu, A. Ni, Y. Zhang, R. Zhang, T. Yu, B. Deb, C. Zhu, A. H. Awadallah, D. Radev, Dyle: Dynamic latent extraction for ab-

stractive long-input summarization, in: 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022, Association for Computational Linguistics (ACL), 2022, pp. 1687–1698.

[78] A. Khullar, U. Arora, Mast: Multimodal abstractive summarization with trimodal hierarchical attention, in: Proceedings of the First International Workshop on Natural Language Processing Beyond Text, 2020, pp. 60–69.

[79] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[80] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to sequence learning with neural networks, Advances in neural information processing systems 27 (2014).