IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 July 2024, accepted 19 August 2024, date of publication 22 August 2024, date of current version 4 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3447737

RESEARCH ARTICLE

VATMAN: Integrating Video-Audio-Text for Multimodal Abstractive SummarizatioN via Crossmodal Multi-Head Attention Fusion

DOOSAN BAEK, JIHO KIM, AND HONGCHUL LEE
School of Industrial Management Engineering, Korea University, Seoul 02841, Republic of Korea

Corresponding author: Hongchul Lee (hclee@korea.ac.kr)

This work was supported by the BK21 FOUR funded by the Ministry of Education of Korea and National Research Foundation of Korea.

ABSTRACT The paper introduces VATMAN (Video-Audio-Text Multimodal Abstractive summarizatioN), a novel approach for generating hierarchical multimodal summaries utilizing Trimodal Hierarchical Multi-head Attention. Unlike existing generative pre-trained language models, VATMAN employs a hierarchical attention mechanism that hierarchically attends to visual, audio, and text modalities. However, in the existing literature, there is a lack of cross-modal attention at the block level. In light of this, we propose a block-level cross-modal attention mechanism, termed Blockwise Cross-modal Multi-head Attention (BCMA), to enhance the summarization performance. This attention mechanism enables the model to simultaneously capture context information from visual, audio, and text modalities, providing a more comprehensive understanding of the input data. In terms of performance, our VATMAN model outperforms the state-of-the-art trimodal model based on RNN in the How2 dataset. Specifically, it achieves a Rouge-1 improvement of 7.53% and Rouge-L improvement of 2.19%, demonstrating superior summarization quality. In addition, compared to uni-modal and di-modal baseline transformer models, VATMAN exhibits significant improvements in Rouge-L scores by 11.12% and 3.85%, respectively, highlighting its effectiveness in capturing hierarchical relationships across modalities. Furthermore, we evaluated our generated abstractive summaries using various metrics, including BLEU, METEOR, CIDEr, ContentF1, and BERTScore. Our proposed model consistently outperformed others across most metrics, demonstrating its effective performance in qualitative assessments.

INDEX TERMS Generative AI, natural language generation (NLG), large language model (LLM), multimodal abstractive summarization, How2, hierarchical crossmodal multi-head attention.

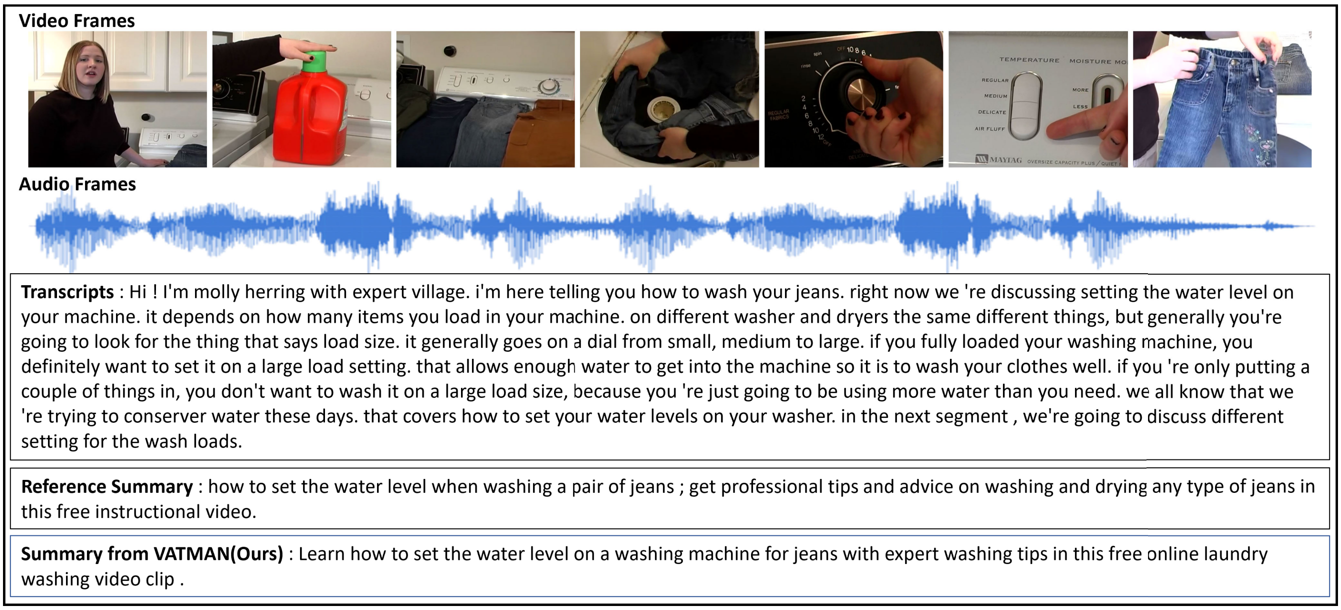
I. INTRODUCTION

Text summarization is a natural language processing task aimed at providing concise and easily readable summaries of given corpora, enabling users to quickly grasp important information. With the exponential growth of textual data, such as documents, articles, and news, the importance of text summarization is increasingly recognized [1]. Summarization tasks are classified into extractive summarization, which extracts the most important sentences or paragraphs based on statistical or linguistic features from the original document, and abstractive summarization, which semantically understands the entire content of the original document and generates summaries in a new way [2], [3], [4].

Research in abstractive summarization includes methods based on recurrent neural networks (RNN) and transformer-based approaches. Previous studies in the RNN-based methods include "Abstractive sentence summarization with attentive recurrent neural networks" [2], "Bottom-up attention" [5], and "MAST" [6]. More recently, transformer-based [7] sequence-to-sequence (seq2seq) pre-trained language models such as BART [8], T5 [9], PEGASUS [10], ProphetNet [11], GPT-3 [12], and GPT-4 [13] have significantly advanced text generation research, including abstractive summarization.

Recent textual information on the web is rarely generated in isolation but often accompanied by images, videos, audio,

The associate editor coordinating the review of this manuscript and approving it for publication was Chuan Li.



Transcripts : Hi ! I'm molly herring with expert village. i'm here telling you how to wash your jeans. right now we 're discussing setting the water level on your machine. it depends on how many items you load in your machine. on different washer and dryers the same different things, but generally you're going to look for the thing that says load size. it generally goes on a dial from small, medium to large. if you fully loaded your washing machine, you definitely want to set it on a large load setting. that allows enough water to get into the machine so it is to wash your clothes well. if you 're only putting a couple of things in, you don't want to wash it on a large load size, because you 're just going to be using more water than you need. we all know that we 're trying to conserver water these days. that covers how to set your water levels on your washer. in the next segment , we're going to discuss different setting for the wash loads.

Reference Summary : how to set the water level when washing a pair of jeans ; get professional tips and advice on washing and drying any type of jeans in this free instructional video.

Summary from VATMAN(Ours) : Learn how to set the water level on a washing machine for jeans with expert washing tips in this free online laundry washing video clip .

FIGURE 1. How2 dataset.

or other modalities [14], [15], [16], [17]. This has led to various research efforts in natural language processing tasks, utilizing not only text but also information from other modalities to obtain richer information and enhance performance [6], [18], [19], [20].

However, there is a significant lack of cases applying giant pre-trained language models to multimodal abstractive summarization tasks, integrating video, audio, and text information simultaneously [21]. Therefore, this paper proposes VATMAN (Video-Audio-Text Multimodal Abstractive Summarization), a generative summarization model that fuses information from video, audio, and text based on giant pre-trained language models. The model demonstrates outstanding performance in generative summarization through experiments on the How2 dataset. By injecting video and audio information simultaneously into a pre-trained giant language model, VATMAN contributes significantly to the tri-modal generative summarization, outperforming unimodal and dimodal baseline models by 11.12% and 3.85%, respectively, according to Rouge-L metrics. Fig.1 is a brief overall description of the video(images), audio, and text in the How2 dataset. The details constituting the How2 dataset will be discussed in detail in Section IV-A. The contributions of this study are summarized as follows:

- We introduce a novel approach for summarizing the content of videos into text using an LLM-based generative model, which comprehensively understands the content of the videos.
- We develop a fusion block that integrates features from video, audio, and text modalities simultaneously, enabling machines to effectively learn and generate content.
- We validate the text generation capability of the proposed method under diverse conditions, we verify it using the How2 dataset and demonstrate its effectiveness through 11 evaluation techniques, including the latest state-of-the-art method, BERTScore.

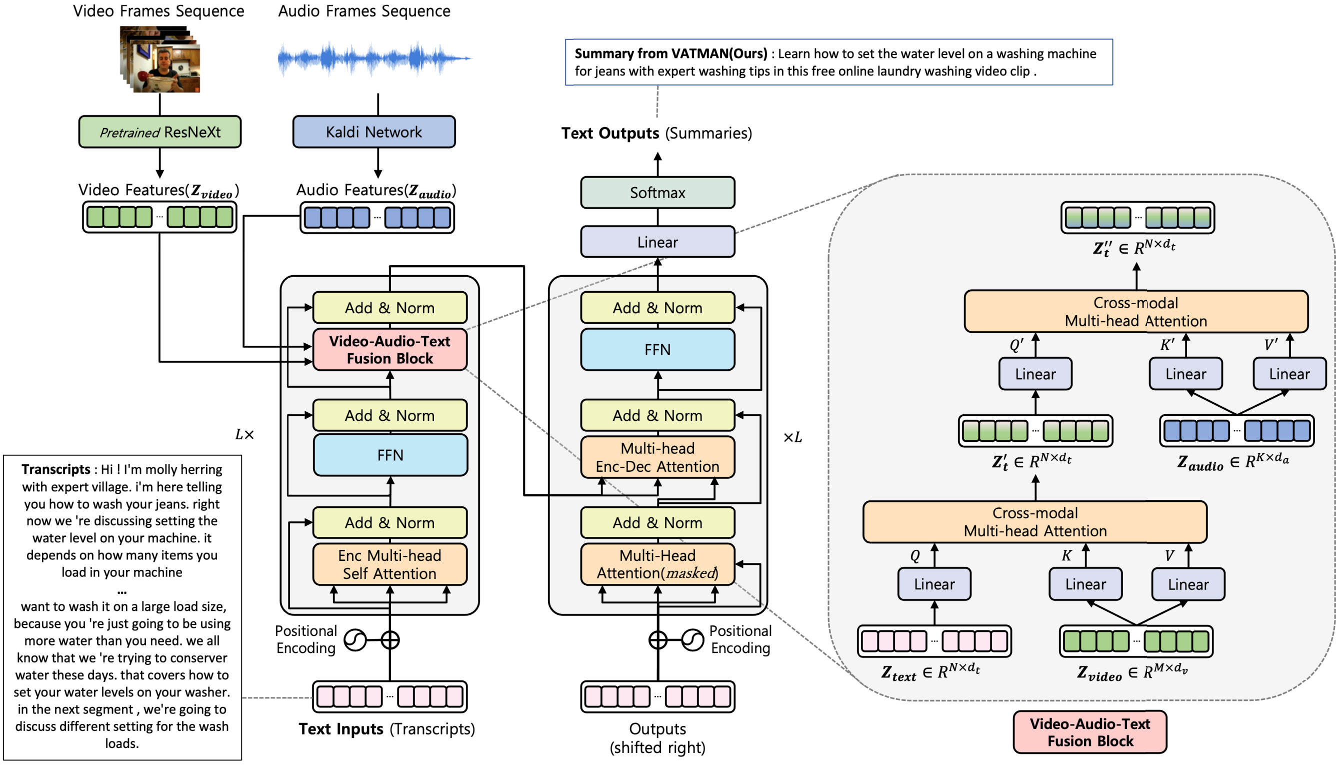
The paper is organized as follows. Section II reviews the relevant literature and Section III describes the proposed solution in detail. Section IV delves into the experimental results and provides a detailed discussion. Section V concludes the paper and discusses the limitations and future work.

II. RELATED WORK

A. UNI-MODAL ABSTRACTIVE SUMMARIZATION

Uni-modal Abstractive Summarization aims to comprehend the most crucial information from input documents and generate concise, readable summaries. In this field, where only one modality, namely text, is involved, significant advancements have been achieved due to the progress in sequence-to-sequence models, attention mechanisms, and the development of deep neural networks. Reference [22] introduced the coverage mechanism to address issues like inaccuracies and repetitive word usage in generated summaries. Reference [23] tackled repetitive and inconsistent structures by introducing a novel neural network model that processes input and output separately, coupled with supervised word prediction and reinforcement learning. Reference [24] proposed a sentence-level semantic similarity contrastive learning method for abstractive text summarization, which distinguish salient information from documents in terms of the semantic similarity.

Recently, research efforts such as BART [8], T5 [9], PEGASUS [10], ProphetNet [11], GPT-3 [12], and Z-Code++


Video Frames Sequence
Audio Frames Sequence
Pretrained ResNeXt
Kaldi Network
Video Features(Z_{video})
Audio Features(Z_{audio})
Add & Norm
Video-Audio-Text Fusion Block
Add & Norm
FFN
Add & Norm
Enc Multi-head Self Attention
$L\times$
Positional Encoding
Text Inputs (Transcripts)
Text Outputs (Summaries)
Softmax
Linear
Add & Norm
FFN
Add & Norm
Multi-head Enc-Dec Attention
Add & Norm
Multi-Head Attention(masked)
$\times L$
Positional Encoding
Outputs (shifted right)
$Z''_t \in R^{N\times d_t}$
Cross-modal Multi-head Attention
Q'
K'
V'
Linear
Linear
Linear
$Z'_t \in R^{N\times d_t}$
$Z_{audio} \in R^{K\times d_a}$
Cross-modal Multi-head Attention
Q
K
V
Linear
Linear
Linear
$Z_{text} \in R^{N\times d_t}$
$Z_{video} \in R^{M\times d_v}$
Video-Audio-Text Fusion Block

Summary from VATMAN(Ours) : Learn how to set the water level on a washing machine for jeans with expert washing tips in this free online laundry washing video clip .

Transcripts : Hi ! I'm molly herring with expert village. i'm here telling you how to wash your jeans. right now we 're discussing setting the water level on your machine. it depends on how many items you load in your machine

...

want to wash it on a large load size, because you 're just going to be using more water than you need. we all know that we 're trying to conserver water these days. that covers how to set your water levels on your washer. in the next segment , we're going to discuss different setting for the wash loads.

FIGURE 2. VATMAN (Framework of proposed method).

[25] have widely employed pre-trained large language models utilizing massive corpora. These models exhibit excellent performance in document generation summarization, showcasing the effectiveness of leveraging massive pre-training in this domain [26]. Furthermore, they utilize attention mechanism [27] to allocate importance to various parts of the input sequence, which proves beneficial for processing long sentences or documents.

B. MULTI-MODAL ABSTRACTIVE SUMMARIZATION

Multimodal Abstractive Summarization differs from Single-Modal Abstractive Summarization by involving two or more modalities in the input. In this context, multimodal abstractive summarization receives input not only in the form of text documents but also incorporates various modalities such as images, videos, and audio. Reference [28] collected a multimodal news article corpus containing 500 English news articles with accompanying videos and annotations. Reference [29] introduced the How2 dataset, consisting of approximately 2,000 hours of short educational videos, each summarized in 2-3 sentences. Reference [18] proposed a multimodal sequence-to-sequence model with hierarchical attention to integrate diverse modality information into consistent summaries. Reference [30] presented a multi-stage fusion network with a fusion forget gate module, capable of modeling fine-grained interactions between multiple-source modalities. Reference [6] proposed a sequence-to-sequence model with trimodal hierarchical attention, based on recurrent neural networks, to utilize information from text, video, and audio modalities. [31] introduced a model that combines video information with a pre-trained large language model through an additional attention layer. Reference [20] proposed two auxiliary tasks and employ multi-task learning to guide the model to learn the paragraph-level vision and language semantic alignment. GPT-4 [13] employs a large-scale multimodal approach, utilizing the Transformer architecture to process both image and text inputs to generate text outputs. This showcases human-level performance on various benchmarks, including a simulated bar exam, and leveraging infrastructure and optimization methods for predictable behavior across scales.

While prior studies have demonstrated excellent performance in generative summarization, there has been a lack of transformer-based abstractive summarization cases that utilize video, audio, and text information simultaneously. Therefore, in this study, we aim to build a pre-trained language model-based abstractive summarization model that leverages information from both text and video/audio modalities.

III. PROPOSED METHOD

In this paper, we leverage the exceptional text generation capabilities of a pre-trained sequence-to-sequence language model and apply it to multimodal abstractive summarization. To achieve this, we propose a model named VATMAN (Video-Audio-Text Multimodal Abstractive SummarizationN), which utilizes a pre-trained language model-based trimodal hierarchical multi-head attention technique. This approach aims to extend the model's capabilities beyond text summarization to include various modalities such as video and audio. VATMAN represents an innovative step towards effective multimodal summarization, harnessing the strengths of pre-trained language models and trimodal attention mechanisms.

A. PRE-TRAINED LANGUAGE MODELS FOR ABSTRACTIVE SUMMARIZATION

The structure of the Transformer-based sequence-to-sequence language model in this study is identical to the one depicted in Fig.2, excluding the proposed Video-Audio-Text Fusion Block. The input text of the language model undergoes tokenization, followed by transformation into a sequence of token embeddings, denoted as $X_{text} \in \mathbb{R}^{N \times d_t}$. Here, N represents the length of the sequence, and d_t represents the dimension of the features. To preserve positional information within the token embeddings, the positional embedding $E_{pe} \in \mathbb{R}^{N \times d_t}$' is added, resulting in the value Z_0^{enc} which is then used as the input to the encoder, as defined in (1).

$$Z_0^{enc} = X_{text} + E_{pe} \tag{1}$$

The encoder is composed of L encoder layers, where each encoder layer consists of: Encoder Multi-head Self-Attention($EMSAtt$) and Feed-Forward Network(FF). $EMSAtt$ is designed to capture relationships and dependencies within the input sequence through multi-head self-attention. FF processes and transforms the information captured by the self-attention mechanism using a feed-forward network. Additionally, after passing through each sub-layer, the output undergoes a residual connection and is subjected to layer normalization (LN). This step, as shown in (2), involving residual connection and layer normalization enhances the stability and efficiency of information flow through the encoder layers.

$$\begin{aligned} Z_t^{enc} = LN(FF(EMSAtt(Z_{l-1}^{enc}) + Z_{l-1}^{enc}) \\ + LN(EMSAtt(Z_{l-1}^{enc}) + Z_{l-1}^{enc})) \end{aligned} \tag{2}$$

Similar to the encoder, the decoder is also composed of L decoder layers. However, there are two key differences. Firstly, the multi-head self-attention ($MSAtt$) is masked, preventing it from referencing future words during the prediction of the next word. Secondly, an additional sub-layer, the multi-head encoder-decoder attention, is introduced to integrate information encoded during the decoding process. This sub-layer utilizes decoder embeddings to combine with the output embeddings from the encoder. In this study, we employ the BART model [8] as the backbone model, which introduces a novel pre-training task into the conventional Transformer architecture.

B. FEATURE EXTRACTION

Video data consists of 16 frames per second, following prior research [6], [18], [29], [31]. Utilizing the Kinetics dataset [32], we extract features using a pre-trained 3D ResNeXt-101 network [33], resulting in 2,048-dimensional features per frame as illustrated in Fig.3.

For audio data, we employ Kaldi [34] to extract 40-dimensional filter bank features and 3-dimensional pitch features. These features are combined, resulting in 43-dimensional audio features. Additionally, to incorporate speaker variability per video, we apply Cepstral Mean and Variance Normalization (CMVN).

The thus-extracted 2,048-dimensional video features and 43-dimensional audio features serve as input for the video-audio-text fusion methodology.

C. MULTI-MODAL FUSION

As shown in Fig.2, We insert a fusion unit block at the end of the encoder block. This sub-layer includes a fusion mechanism for video-audio-text, along with residual connections and layer normalization. We propose a video-audio-text fusion mechanism depicted on the right side of Fig.2. Given the embeddings from each modality text input $Z_{text} \in \mathbb{R}^{N \times d_t}$, video input $Z_{video} \in \mathbb{R}^{M \times d_t}$, and audio input $Z_{audio} \in \mathbb{R}^{K \times d_a}$, the output of the first fusion (video-text fusion) mechanism $Z_t' \in \mathbb{R}^{N \times d_t}$ passes through a hidden layer with dimensions identical to the text input. It encapsulates both the attention information from text and image modalities.

Next, the output of this hidden layer Z_t serves as the query, while the audio input Z_{audio} serves as the key and value inputs for the second fusion (video-audio-text fusion) mechanism. At this point, the final hidden layer value $Z_t'' \in \mathbb{R}^{N \times d_t}$ is produced. Since this matches the dimension of the text input Z_{text} before passing through the video-audio-text fusion layer, the dimension remains consistent even when stacking multiple layers. This method hierarchically attends to video, audio, and text modalities.

1) CROSS-MODAL MULTI-HEAD ATTENTION(CMA)

Attention mechanism assigns weights to each element of an input sequence to focus on important information, commonly used in natural language processing and computer vision for selective processing of data [7].

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

Here, Q, K, and V as shown in (3), represent the Query, Key, and Value matrices respectively, where d_k denotes the dimensionality of the Key vectors.

In the framework of Cross-modal Multi-head Attention (CMA) [28], we conceptualize the query as text (modality α) and the key and value as video (modality β), though other modalities are also feasible. The mechanism computes weighted sums of values based on the similarity between modality α and β. We operationalize information fusion from β to α by embedding the features of α into Query and the

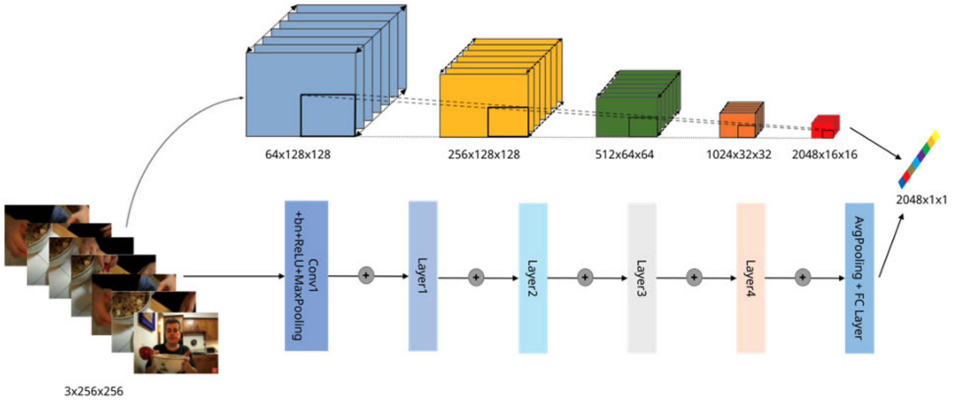


FIGURE 3. Video features through ResNeXt-101 network.

features of β into Key and Value. This latent adaptation from β to α is embodied in the $\mathrm{CMA}(Q_\alpha, K_\beta, V_\beta)$. These quantities are described in (4).

$$\begin{aligned}\mathrm{CMA}(Q_\alpha, K_\beta, V_\beta) &= \mathrm{softmax}\left(\frac{Q_\alpha K_\beta^T}{\sqrt{d_k}}\right)V_\beta \\ &= \mathrm{softmax}\left(\frac{X_\alpha W_{Q_\alpha} X_\beta^T W_{K_\beta}^T}{\sqrt{d_k}}\right)X_\beta W_{V_\beta}\end{aligned} \tag{4}$$

W_Q, W_K, and W_V are weight matrices for each head, linearly transforming the features before applying the attention mechanism. By leveraging this approach, the model effectively captures cross-modal interactions between different types of information, facilitating comprehensive understanding and learning across modalities. Building upon this CMA technique, we introduce a method that fuses text and video from the rear end, followed by fusing the resulting hidden vector with audio.

2) VIDEO-TEXT CMA FUSION

In the first video-text fusion block, the query (Q) is a feature vector containing the text information (Z_{text}) passed through the previous layer. This feature vector is linearly transformed to a common dimension for Cross-modal Multi-head Attention (CMA) operations as shown in (5). The keys (K) and values (V) are feature vectors obtained by linearly transforming the last 2,048 dimensions of visual features from video data to the common dimension as shown in (6), (7). Subsequently, cross-modal multi-head attention is applied, fusing the query text features with the key and value visual features, resulting in the output vector (O) as expressed in (8). This output vector is then linearly transformed to the dimension of the original input text features (Z_{text}) and concatenated with the initial text features, yielding the resulting vector Z_t'. This concatenated feature vector is expressed in (9).

$$Q = Z_{\text{text}} W_q \qquad Q \in \mathbb{R}^{N \times d_c} \tag{5}$$
$$K = Z_{\text{video}} W_k \qquad K \in \mathbb{R}^{M \times d_c} \tag{6}$$
$$V = Z_{\text{video}} W_v \qquad V \in \mathbb{R}^{M \times d_c} \tag{7}$$
$$O = CMA(Q, K, V) \qquad O \in \mathbb{R}^{N \times d_c} \tag{8}$$
$$Z_t' = Concat(Z_{text}, O)W_t \tag{9}$$

The CMA operation employed here utilizes multiple heads to learn and model relationships between different modalities, such as text, video, and audio features [35]. Q, K, and V represent features extracted from different modalities, like text and video. This technique combines features from diverse modalities to capture relationships between them.

3) VIDEO-AUDIO-TEXT CMA FUSION

The feature vector (Z_t') passing through the video-text fusion block undergoes a linear transformation to become the query (Q') as shown in (10). The keys (K') and values (V') are feature vectors obtained from the audio features, representing the last 43 dimensions transformed into a common dimensionality (common dimension) as shwon in (11), (12). Subsequently, the second cross-modal multi-head attention is applied, where the query representing the fusion of text and video features is combined with the keys and values representing audio features, resulting in the output vector (O') as expressed in (13). Finally, the feature vector (Z_t') passing through the video-text fusion block is concatenated and linearly transformed to the dimensions of the original text, yielding the final result vector (Z_t''), as expressed in (14).

$$Q' = Z_{\text{text}}^{'W_q} \qquad Q' \in \mathbb{R}^{N \times d_c} \tag{10}$$
$$K' = Z_{\text{audio}} W_k \qquad K' \in \mathbb{R}^{M \times d_c} \tag{11}$$
$$V' = Z_{\text{video}} W_v \qquad V' \in \mathbb{R}^{M \times d_c} \tag{12}$$
$$O' = CMA(Q', K', V') \qquad O' \in \mathbb{R}^{N \times d_c} \tag{13}$$
$$Z_t'' = Concat(Z_{text}', O')W_t \tag{14}$$

TABLE 1. Training hyperparameter.

Type	BART, T5(Uni-modal)	BART, T5(Di-modal)	VATMAN(Tri-modal)
Input	Text	Text + Vision or Text + Audio	Text + Vision + Audio
Backbone architecture	Transformer based Enc-Dec	Transformer based Enc-Dec	BART based Enc-Dec
Attention type in fusion block	Dot-product attention	Cross-modal Multi-head Attention	Cross-modal Multi-head Attention
Layers(L)	6	6	6
Heads	-	4	4
Common dimension(d_c)	256	256	256
Batch	32	32	16
Epochs	150	150	150
Learning rate	3×10^{-5}	3×10^{-5}	3×10^{-5}
Optimizer	Adam	Adam	Adam

The final result vector Z_t'' retains the dimensions of the original text features while encompassing both video and audio features in a single vector. This vector serves as the input sequence for the multi-head encoder-decoder attention block in the transformer's decoder.

IV. EXPERIMENT

A. HOW2 DATASET

The How2 Dataset consists of a total of 79,114 video, transcripts, summary data, divided into a set of 2,000 hours with various domains such as cooking, music, indoor, outdoor activities, and sports. Additionally, there are 13,445 audio, video, transcripts, summary data, forming a set of 300 hours. This dataset, as shown in Fig.1, comprises short instructional videos spanning diverse domains, where transcripts accompanying the videos provide textual representation converted from the speaker's audio, conveying the overall content of the visual material. This dataset was compiled by downloading videos from YouTube, accompanied by various metadata, including accurate subtitles and summaries in English, authored by the video creators. The summaries condense abstracted overviews into 2-3 sentences, demonstrating an understanding of both video and audio modalities simultaneously by humans. During the summarization process, if access is limited to text only, understanding whether the term "green" in the subtitles refers to "green" or signifies the "the surface of a golf course" remains unclear. However, with additional visual context (video modality) of a flagpole in green grass or auditory context (audio modality) of outdoor sounds associated with hitting a golf ball, a multimodal model can accurately interpret the expression as "the surface of a golf course" [29]. Therefore, the How2 dataset is employed for experiments that aim to maximize the utilization of visual and auditory contexts.

B. IMPLEMENTATION DETAILS

1) DATA PREPROCESSING

In this study, performance needs to be evaluated when injecting auditory context (audio) to simultaneously leverage three modalities. To achieve this, 300 hours of text/video/audio data are sampled based on the IDs of 300 hours of audio data. A total of 13,445 data points, consisting of 12,798 training samples, 520 validation samples, and 127 test samples, are utilized for experimentation. Subtitles are normalized to lowercase and undergo punctuation filtering. After tokenization, the subtitles are preprocessed by either segmenting into 512-token sequences or padding. Video and audio data have features extracted as described in Section III-B. Feature Extraction resulting in feature vectors of dimensions 2048 and 43, respectively.

2) SOFTWARE AND HARDWARE

In this study, we employ the PyTorch deep learning framework [36] for code implementation and leverage PyTorch-Lightning for distributed training. Additionally, the experiments are conducted using four Nvidia GeForce RTX 3090 GPUs.

3) HYPERPARAMETER SELECTION

Hyperparameters are variables that tune and control machine learning models or deep learning algorithms. These variables, set by the user before training the model, consist of values that affect the model's structure, learning process, and more. Table 1 presents the final selection of hyperparameters used in the training. As mentioned in Fig.2, key hyperparameters include the number of layers (L) in the encoder and decoder, the backbone architecture, attention type in the fusion block, the number of heads, common dimensionality used in fusion, batch size, epochs, learning rate, and optimizer. For models generating unimodal summaries where the input is text only, the attention type utilizes Dot-product attention. However, for the proposed tri-modal model (VATMAN) designed for generating summaries from three input modalities, with an updateable parameter count around 220 million, a memory issue occurs when using a batch size of 32. Consequently, the batch size is adjusted to 16.

C. EVALUATION METRICS AND MODEL PERFORMANCE COMPARISON

ROUGE-N [37] measures the ratio of N-grams in the model-generated summary that overlap with the reference summary. Specifically, ROUGE-1 represents the unigram overlap, ROUGE-2 represents the bigram overlap, and ROUGE-L represents the ratio of the longest common subsequence

TABLE 2. Evaluation results(Rouge, BLEU) of baselines and our proposed model(VATMAN) on test data of How2 300hours.

Input modality	Method	Rouge-1 (%)	Rouge-2 (%)	Rouge-L (%)	BLEU-1 (%)
Transcripts (Unimodal)	RNN [6]	15.8475	2.5922	18.4049	0.399
	Transformer [7]	41.5003	20.5349	39.7585	17.2905
	T5 [9]	45.0766	21.5857	37.0916	39.7570
	BART [8]	49.0718	26.7150	41.6582	41.2951
Transcripts + Video (Dimodal)	RNN	33.6737	17.6605	34.2591	29.9351
	Transformer	43.3903	23.109	42.5416	19.0915
	T5	47.4345	23.3531	39.9249	42.7832
	BART	51.7293	29.0403	43.7583	49.4860
Transcripts + Audio (Dimodal)	RNN	37.5794	18.98	37.2308	23.9267
	Transformer	41.5403	20.5006	41.1485	17.8997
	T5	46.4088	23.4022	39.2089	38.9838
	BART	50.1433	26.5348	42.0547	45.1529
Transcripts + Video + Audio (Trimodal)	RNN	48.85	29.51	43.23	-
	VATMAN(Ours.)	**52.5315**	**29.4880**	**44.1799**	**49.8411**

TABLE 3. Evaluation results(Rouge, BLEU, METEOR, CIDEr, content F1, BERTScore) of T5, BART, and VATMAN on test data of How2 300hours.

Input modality	Method	Rouge-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	Content F1	BERTScore
Transcripts	T5 [9]	37.0916	39.7570	27.2397	19.7924	14.3322	19.1181	6.54438	88.73	91.39
	BART [8]	41.6582	41.2951	30.0475	23.4626	18.5903	20.8227	9.55826	89.17	91.69
Transcripts + Video	T5	39.9249	42.7832	29.8261	22.0866	16.4179	20.0638	7.47168	88.90	91.54
	BART	43.7583	49.4860	36.5272	28.6499	23.7186	23.3517	**14.06855**	89.35	91.53
Transcripts + Audio	T5	39.2089	39.9838	27.2016	20.1025	14.9351	19.1734	6.64870	88.91	91.75
	BART	42.0547	45.1529	32.2971	25.1008	19.9448	21.6267	10.59267	89.13	91.53
Transcripts +Video + Audio	**VATMAN(Ours.)**	**44.1799**	**49.8411**	**36.8026**	**29.0786**	**23.6499**	**23.7069**	14.03920	**89.39**	**91.90**

between the model summary and the reference summary. This can be generalized as in (15).

$$\text{ROUGE-N} = \frac{\text{Number of overlapped n-gram}}{\text{Total words in reference summary}} \tag{15}$$

BLEU-N [38], similar to ROUGE, is an N-gram count-based metric used for evaluating natural language processing performance. It measures how many overlapping words exist between reference summaries (R) and generated summaries (S), calculating precision and recall, with results ranging between 0 and 1, which can be defined as in (16).

$$\text{BLEU} = \min\left(1, \frac{S \text{ length}}{R \text{ length}}\right)\left(\prod_{n=1}^{4} \text{n-gram}_{precision}\right) \tag{16}$$

Content F1 Score, which can be defined as in (17), is a comprehensive evaluation metric calculated as the harmonic mean of Precision and Recall. It evaluates the semantic alignment between the given input sentence and the generated sentence in sentence generation tasks. Recall measures the ratio of overlapping content words in the summary to the total number of words in the original summary, while Precision is an indicator of the proportion of overlapping content in the machine-generated result or summary compared to the reference sentence.

$$\text{Content F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

METEOR [39] complements the limitations of BLEU and is based on a generalized concept of unigram matching between machine-generated translation and human-made reference translation. Fmean is the harmonic mean of accuracy (precision) and recall. The penalty is calculated by determining the chunks where the candidate translation and the reference translation match, and it is smallest when both translations perfectly match, i.e., when the chunk is 1. METEOR is the product of these two values as shown in (18).

$$\text{METEOR} = F\text{mean} \times (1 - \text{Penalty}) \tag{18}$$

CIDEr [40] is provided as an evaluation metric for tasks involving the description of images as sentences. It is a technique used to automatically assess how well candidate sentences for an image align with the aggregate of image descriptions. For encoding, it performs TF-IDF(Term Frequency-Inverse Document Frequency) [41] weighting for each N-gram.

BERTScore [42] involves passing reference and generated sentences through the BERT model to extract contextual embeddings. It then calculates precision and recall based on the cosine similarity between the contextual embeddings of the two sentences. Reference [42] also proposed a method that incorporates IDF values into the similarity calculation. BERTScore is advantageous due to its precision, recall, and correlation with qualitative evaluations, thanks to its refined

TABLE 4. Comparison of generative summary results for gifferent modalities on VATMAN.

ID: fZM3IcM2Xs4	
Reference	watch and learn from our **seafood expert** about mashing potatoes for **christmas dinner** with bacon scallops in this **free recipe video** on making bacon wrapped scallops .
Text to Text	how to mash potatoes ; get professional tips and advice from an expert chef on methods , techniques , and products for making traditional mexican food in this free cooking video .
Video-Text to Text	learn how to mash potatoes for **christmas dinner** in this free holiday recipe video .
Audio-Text to Text	how to mash potatoes ; get professional tips and advice from an expert chef on making traditional american food recipes in this free cooking video .
Video-Audio-Text to Text	learn how to mash potatoes for **christmas dinner** with **expert cooking tips** in this **free holiday recipe video clip** .
ID: g27MVuMkWs4	
Reference	**to view lyrics on an ipod** , turn itunes on , choose a song , **enter the lyrics in the lyrics tab** , sync the ipod , play the song on the ipod and **use the scrubber** to get the lyrics that were entered in itunes . song lyrics can be cut and pasted into itunes from many web sites with tips from a computer consultant in this free video on technology .
Text to Text	To see lyrics on your iPod, start iTunes , type in the lyrics, sync your iPod from iTunes . You can also copy and paste lyrics into iTunes from various websites , as explained by a computer consultant in a free technology video .
Video-Text to Text	To check lyrics on an iPod , use iTunes to find the song with lyrics . Discover how to view lyrics with tips from a computer specialist in this free video on computers .
Audio-Text to Text	Gain expertise on viewing lyrics on an iPod , including tips on iPod accessories and features , in this free video clip .
Video-Audio-Text to Text	**To view lyrics on an iPod** , activate iTunes , choose a song , **input the lyrics in the designated tab** , synchronize the iPod , play the song on the device , and **use the scrubber** to view the lyrics entered in iTunes . Tips from a computer consultant in a free technology video demonstrate how to copy and paste song lyrics into iTunes from different websites .
ID: g38AmwPAYvg	
Reference	Becoming a portrait photographer involves studying **light , both natural and studio lights ,** to understand how it **reacts with human skin** , and experimenting with different methods of portraiture to build up a portfolio . become a portrait photographer with information from a professional art and commercial photographer in this free video on photography .
Text to Text	To excel as a portrait photographer, one should delve into the intricate study of lighting and understand its nuanced interaction with human skin. Valuable tips for success can be acquired from a professional photographer through a freely available video on photography .
Video-Text to Text	A prospective portrait photographer must dedicate time to studying lighting intricacies and comprehending how lights react with human skin. Valuable guidance from a professional art and commercial photographer is accessible in a free video resource on photography .
Audio-Text to Text	Achieving proficiency as a portrait photographer involves a thorough study of lighting and its effects on human skin. Valuable tips from a professional photographer are readily available in a freely accessible video resource on photography .
Video-Audio-Text to Text	Becoming a portrait photographer involves studying **natural light and studio lighting** to understand how it **reacts to human skin** and experiment with **different methods of portrait to build a portfolio.** In this free video for photography, you become a portrait photographer with information from professional and commercial photographers

embeddings, contextual understanding, syntactic structure reflection, and low-frequency word weighting.

Table 2 compares the performance of the proposed multimodal summarization model with prior research [6], [7], [8], [9] using the same test dataset. Based on the quantitative experimental results, the proposed method achieved the best performance, surpassing the existing methods across all evaluation metrics. Specifically when compared to the existing trimodal model based on RNN, our model achieves a Rouge-1 improvement of 7.53% and Rouge-L improvement of 2.19%, demonstrating superior summarization quality. the highest ROUGE-1, 2, and L scores reflect our method's ability to match human-generated summaries either locally or globally. Additionally, the highest BLEU score indicates our machine-generated translations are closer to human translations.

Table 3 compares the performance of our proposed model, which utilizes video, audio, and text data, with unimodal and dimodal giant language models BART and T5. Compared to the unimodal and dimodal baseline transformer models, VATMAN shows significant improvements in Rouge-L scores by 11.12% and 3.85%, respectively, highlighting its effectiveness in capturing hierarchical relationships across modalities. This suggests that incorporating information from video or audio alongside text leads to better performance in abstractive summarization than relying solely on text. Furthermore, the most notable improvements are observed when information from video, audio, and text is simultaneously integrated. In addition, our experiments demonstrate that our proposed method outperforms others not only in quantitative evaluation metrics like ROUGE and BLEU but also in qualitative evaluation metrics (METEOR, CIDEr, Content F1, BERTScore). When one more input modality was added to the T5 model, the BERTScore increased by 0.16% and 0.39% compared to the single-modality baseline, respectively. Furthermore, we aim to confirm the superior performance of our proposed model. So, we compared its performance with that of previous models across various input modalities using the latest performance technique, BERTScore. When our model used all three modalities as input, the BERTScore improved by 0.56% compared to the T5 single-modality baseline and by 0.23% compared to the BART single-modality baseline. Particularly noteworthy is that, unlike previous studies, we used the state-of-the-art BERTScore technique, and the highest score we achieved indicates our method's proficiency in contextual understanding and reflecting syntactic structure.

D. QUALITATIVE ANALYSIS OF GENERATED ABSTRACTIVE SUMMARIES

Previous studies [6], [30], [31] have focused on qualitative human-evaluations between references and generated summaries. Human-evaluation is crucial for the advancement of Natural Language Generation (NLG) systems, and recently, various methods have been proposed to fulfill this purpose [43], [44], [45]. Table 4 compares the generated summaries from our proposed model, VATMAN. The first generated summary discusses a cooking video created by a seafood expert on Christmas Eve. While the text-only summary includes instructions on crushing potatoes, the summaries that incorporate audio and text data provide additional details on preparing holiday dishes during 'christmas dinner'. It is evident that the model understands and generates content related to the need for a 'video clip' during holidays when both video and audio information are utilized. The Second generated summary relates to methods for viewing lyrics on an iPod as outlined in the reference. While the uni and dimodal models lack specific information about the temporal sequence of the search process, the trimodal model, including video and audio details, certainly understands and explains the fine-grained act of using 'scrubber'. The third generated summary pertains to the methods of becoming a portrait photographer, as outlined in the reference. While the uni and dimodal models lack specific details about both natural and studio lighting, the trimodal model, when provided with video and audio information, accurately capture and explain these lighting conditions and how they interact with human skin.

From these summaries, it is evident that summaries generated with the inclusion of video or audio outperform those using only text data. Moreover, the combination of video and audio information results in more comprehensive summaries, surpassing the richness of content in previous modalities. Additionally, there is a notable enhancement in the inclusion of specific actions and methods in the summaries.

V. CONCLUSION

In this paper, we propose the VATMAN model, which is based on the pretrained language model (BART) and employs a Trimodal Hierarchical Multi-head Attention technique. While previous research in multimodal abstractive summarization tasks featured transformer-based multimodal models, they primarily focused on using text and video to generate summarized text, neglecting additional modalities such as audio. This limitation motivated us to assume that a multimodal structure capable of utilizing additional modalities simultaneously could enhance the performance of the summarization task by incorporating previously unexplored information. Indeed, in the related work, it has been confirmed that as additional modalities are incorporated, machines better comprehend and perform generation tasks more effectively [6], [18], [29], [30], [31]. To address this, we introduce an additional attention layer that allows the integration of audio information while maintaining the overall structure of the existing multimodal framework that utilizes text and video information concurrently. Moreover, we conducted a comprehensive evaluation of VATMAN, our proposed summarization model, leveraging the publicly available How2 dataset. We employed various evaluation metrics, including ROUGE, BLEU, Content F1, BERTScore, METEOR, and CIDEr, to assess performance effectively. The experimental results consistently demonstrate the superiority of our proposed model, which utilizes audio information in addition to text and video, across all evaluation metrics. Particularly noteworthy is the experimental verification of the importance of trimodality information through experiments that combine text, video, and audio information. Furthermore, we conducted a qualitative evaluation of our proposed VATMAN model by varying the input modalities, including Unimodal, Di-modal, and Tri-modal configurations. Through this evaluation, we confirmed that summaries generated by combining all trimodal information are richer and contain more informative content. As future research directions, we anticipate exploring alternative model structures that combine audio information or additional modalities. While our study demonstrates the effective improvement of summarization performance through the hierarchical structure, additional research is needed to determine the extent of integration for audio information or additional modalities. Expanding from the perspective of incorporating additional modality information, research on model structures considering integration levels can further enhance summarization performance. Furthermore, we aim to develop an integrated model trained on a broader range of data and test it on various datasets to explore a robust method.

REFERENCES

[1] V. Hassija, A. Chakrabarti, A. Singh, V. Chamola, and B. Sikdar, "Unleashing the potential of conversational AI: Amplifying chat-GPT's capabilities and tackling technical hurdles," *IEEE Access*, vol. 11, pp. 143657–143682, 2023.

[2] S. Chopra, M. Auli, and A. M. Rush, "Abstractive sentence summarization with attentive recurrent neural networks," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Human Lang. Technol.*, 2016, pp. 93–98.

[3] R. Nallapati, B. Zhou, C. dos Santos, C. Gulcehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNs and beyond," in *Proc. 20th Signal Conf. Comput. Natural Lang. Learn.*, S. Riezler and Y. Goldberg, Eds., Berlin, Germany: Association for Computational Linguistics, 2016, pp. 280–290. [Online]. Available: https://aclanthology.org/K16-1028

[4] H. Jang and W. Kim, "Reinforced abstractive text summarization with semantic added reward," *IEEE Access*, vol. 9, pp. 103804–103810, 2021.

[5] S. Gehrmann, Y. Deng, and A. Rush, "Bottom-up abstractive summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds., Brussels, Belgium: Association Computational Linguistics, Oct./Nov. 2018, pp. 4098–4109. [Online]. Available: https://aclanthology.org/D18-1443

[6] A. Khullar and U. Arora, "MAST: Multimodal abstractive summarization with trimodal hierarchical attention," in *Proc. 1st Int. Workshop Natural Lang. Process. Beyond Text*, G. Castellucci, S. Filice, S. Poria, E. Cambria, and L. Specia, Eds., Nov. 2020, pp. 60–69. [Online]. Available: https://aclanthology.org/2020.nlpbt-1.7

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[8] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Jul. 2020, pp. 7871–7880. [Online]. Available: https://aclanthology.org/2020.acl-main.703

[9] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 5485–5551, 2020.

[10] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 11328–11339.

[11] W. Qi, Y. Yan, Y. Gong, D. Liu, N. Duan, J. Chen, R. Zhang, and M. Zhou, "ProphetNet: Predicting future N-gram for sequence-to-sequence pre-training," in *Proc. Findings Assoc. Comput. Linguistics (EMNLP)*, T. Cohn, Y. He, and Y. Liu, Eds., Nov. 2020, pp. 2401–2410. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.217

[12] T. B. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.

[13] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[14] D. Kiela, S. Bhooshan, H. Firooz, E. Perez, and D. Testuggine, "Supervised multimodal bitransformers for classifying images and text," 2019, *arXiv:1909.02950*.

[15] M. Wang, R. Hong, G. Li, Z.-J. Zha, S. Yan, and T.-S. Chua, "Event driven web video summarization by tag localization and key-shot identification," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 975–985, Aug. 2012.

[16] M. Alsaedi, F. A. Ghaleb, F. Saeed, J. Ahmad, and M. Alasli, "Multi-modal features representation-based convolutional neural network model for malicious website detection," *IEEE Access*, vol. 12, pp. 7271–7284, 2024.

[17] D. Kim, E. Lee, D. Yoo, and H. Lee, "Fine-grained human hair segmentation using a text-to-image diffusion model," *IEEE Access*, vol. 12, pp. 13912–13922, 2024.

[18] S. Palaskar, J. Libovický, S. Gella, and F. Metze, "Multimodal abstractive summarization for How2 videos," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, 2019, pp. 6587–6596. [Online]. Available: https://aclanthology.org/P19-1659

[19] D. Chen and R. Zhang, "Building multimodal knowledge bases with multimodal computational sequences and generative adversarial networks," *IEEE Trans. Multimedia*, vol. 26, pp. 2027–2040, 2023.

[20] C. Cui, X. Liang, S. Wu, and Z. Li, "Align vision-language semantics by multi-task learning for multi-modal summarization," *Neural Comput. Appl.*, pp. 1–14, May 2024.

[21] Z. Li, X. Xie, F. Ling, H. Ma, and Z. Shi, "Matching images and texts with multi-head attention network for cross-media hashing retrieval," *Eng. Appl. Artif. Intell.*, vol. 106, Nov. 2021, Art. no. 104475.

[22] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, R. Barzilay and M.-Y. Kan, Eds. Vancouver, BC, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1073–1083. [Online]. Available: https://aclanthology.org/P17-1099

[23] R. Paulus, C. Xiong, and R. Socher, "A deep reinforced model for abstractive summarization," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–13. [Online]. Available: https://openreview.net/forum?id=HkAClQgA-

[24] Y. Huang, Z. Li, Z. Chen, C. Zhang, and H. Ma, "Sentence salience contrastive learning for abstractive text summarization," *Neurocomputing*, vol. 593, Aug. 2024, Art. no. 127808.

[25] P. He, B. Peng, S. Wang, Y. Liu, R. Xu, H. Hassan, Y. Shi, C. Zhu, W. Xiong, M. Zeng, J. Gao, and X. Huang, "Z-Code++: A pre-trained language model optimized for abstractive summarization," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, ON, Canada: Association for Computational Linguistics, Jul. 2023, pp. 5095–5112.

[26] P. Maddigan and T. Susnjak, "Chat2VIS: Generating data visualizations via natural language using ChatGPT, Codex and GPT-3 large language models," *IEEE Access*, vol. 11, pp. 45181–45193, 2023.

[27] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent.*, Jan. 2015, pp. 1–15.

[28] H. Li, J. Zhu, C. Ma, J. Zhang, and C. Zong, "Multi-modal summarization for asynchronous collection of text, image, audio and video," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, M. Palmer, R. Hwa, and S. Riedel, Eds. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 1092–1102. [Online]. Available: https://aclanthology.org/D17-1114

[29] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze, "How2: A large-scale dataset for multimodal language understanding," 2018, *arXiv:1811.00347*.

[30] N. Liu, X. Sun, H. Yu, W. Zhang, and G. Xu, "Multistage fusion with forget gate for multimodal summarization in open-domain videos," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., Nov. 2020, pp. 1834–1845. [Online]. Available: https://aclanthology.org/2020.emnlp-main.144

[31] T. Yu, W. Dai, Z. Liu, and P. Fung, "Vision guided generative pre-trained language models for multimodal abstractive summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, M.-F. Moens, X. Huang, L. Specia, and S. W.-T. Yih, Eds. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3995–4007.

[32] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.

[33] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6546–6555.

[34] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, 2011.

[35] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. Conf. Assoc. Comput. Linguistics Meeting*, 2019, p. 6558.

[36] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019, pp. 8026–8037.

[37] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. Workshop Text Summarization ACL*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[38] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, P. Isabelle, E. Charniak, and D. Lin, Eds. Philadelphia, PA, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040

[39] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, Eds. Ann Arbor, MI, USA: Association for Computational Linguistics, Jul. 2005, pp. 65–72. [Online]. Available: https://aclanthology.org/W05-0909

[40] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.

[41] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *J. Documentation*, vol. 60, no. 5, pp. 493–502, Oct. 2004.

[42] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–43. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

[43] C. van der Lee, A. Gatt, E. van Miltenburg, and E. Krahmer, "Human evaluation of automatically generated text: Current trends and best practice guidelines," *Comput. Speech Lang.*, vol. 67, May 2021, Art. no. 101151.

[44] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 27730–27744.

[45] A. S. De Oliveira Góes and R. C. L. De Oliveira, "A process for human resource performance evaluation using computational intelligence: An approach using a combination of rule-based classifiers and supervised learning algorithms," *IEEE Access*, vol. 8, pp. 39403–39419, 2020.

DOOSAN BAEK received the B.S. degree in applied statistics from Korea University, where he is currently pursuing the M.S. degree with the Department of Industrial Management Engineering. His research interests include developing machine learning and AI algorithms for both structured and unstructured data, such as text and images generation and applying them to solving problems, such as NLP.

JIHO KIM received the B.S. degree in industrial and information systems engineering from Seoul National University of Science and Technology. He is currently pursuing the Ph.D. degree with the Department of Industrial Management Engineering, Korea University, Republic of Korea. His research interests include big data analytics, artificial intelligence, natural language processing, and business intelligence.

HONGCHUL LEE received the B.S. degree in industrial engineering from Korea University, the M.S. degree in industrial engineering from The University of Texas at Arlington, and the Ph.D. degree in industrial engineering from Texas A&M University. He is currently a Professor with the Department of Industrial Management Engineering, Korea University. His research interests include system engineering, system simulations, and artificial intelligence.