# Multi-Modal Video Summarization Based on Two-Stage Fusion of Audio, Visual, and Recognized Text Information

Zekun Yang * Jiajun He † and Tomoki Toda *
* Information Technology Center, Nagoya University, Japan
yang.zekun@g.sp.m.is.nagoya-u.ac.jp, tomoki@icts.nagoya-u.ac.jp
† Graduate School of Information Science, Nagoya University, Japan
jiajun.he@g.sp.m.is.nagoya-u.ac.jp

***Abstract*—Multi-modal video summarization task provides a text summary by combining information from different inputs. Previous methods usually generate the summary from video and ground truth text inputs, but seldom study how audio, video, and ASR text influence the results. This work introduces a novel multi-modal video summarization approach that leverages video and audio inputs, along with speech-recognized text. Our method initiates by conducting speech recognition using a whisper ASR model, followed by representation of visual, textual, and audio information using video transformer, BART, and whisper, respectively. Subsequently, two-stage fusion between different modalities is applied. The fused features are fed into a BART decoder to generate text summaries. Experiments on the how2 dataset show that our method outperforms previous baselines. It is also revealed that incorporating audio information helps generate a better summary even if the texts contain certain errors.**

## I. Introduction

E-learning has become a crucial way for people to acquire knowledge. With numerous educational videos being created and shared online every day, knowing their main contents and recommending videos based on the audience's interest become necessary. To address this issue, video summarization tasks have been proposed to retrieve essential information and generate a short and readable text summary [1]–[3].

In recent years, large-scale generative pre-trained language models [4]–[6] based on Transformers [7] have been proposed and applied in video summarization tasks, both in single modality and in multiple modalities. In single modality approaches, Sharma et al. [8], [9] utilizes audio input to transcribe speech into text and generate the summary. For multiple modalities, Yu et al. [3], Xu et al. [10], Qiao et al. [11], and Yuan et al. [12] leverage both visual and language features, employing information fusion techniques to produce summary texts.

Automatic speech recognition (ASR) techniques have made significant advancements by developing larger and more robust models [13]–[16] and making use of multi-modal information [17]–[21]. Moreover, ASR-oriented feature representation has been adopted in summarization [8], speech translation [22], emotion recognition [23] etc., showing its potential to provide necessary information for various applications. However, previous multi-modal summarization tasks tend to focus on video and ground truth text information, few researchers study how audio, video, and ASR text information jointly influence the quality of generated summary texts. Unfortunately, online videos usually do not contain subtitle texts. Since textual information is highly beneficial for video summarization [8], it is necessary to carry ASR to obtain audio transcripts and provide richer information beyond just audio and video inputs.

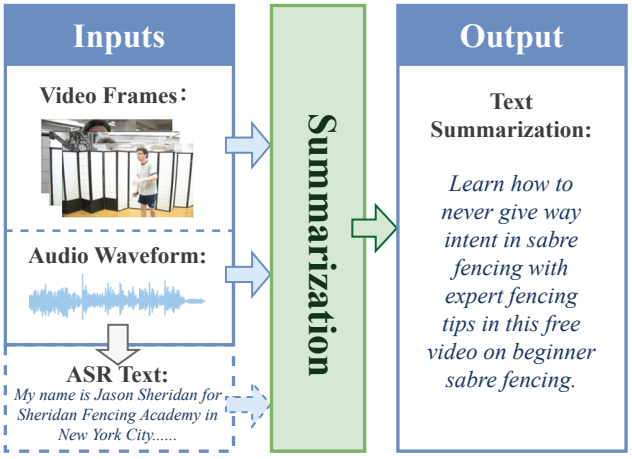


Fig. 1. The concept of the proposed video summarization method. Our method takes video and audio as inputs. We carry speech recognition and present two-stage fusion of visual, audio, and text to obtain text summaries.

In this work, we present a novel method that produces text summaries based on video and audio inputs, which is shown in Fig. 1. In this method, text is first generated by a whisper ASR model. Then, video transformer [7], BART [4] and frozen whisper [14] are used to extract features from visual, text, and audio modalities respectively. Two methods based on two-stage fusion are introduced to combine information from three different modalities (i.e. visual, text, and audio) together. A BART decoder is finally adopted to obtain the text summaries. We conduct experiments on the how2 dataset [24] to evaluate our method. The results show that our method outperforms previous baselines. We perform further experiments and prove

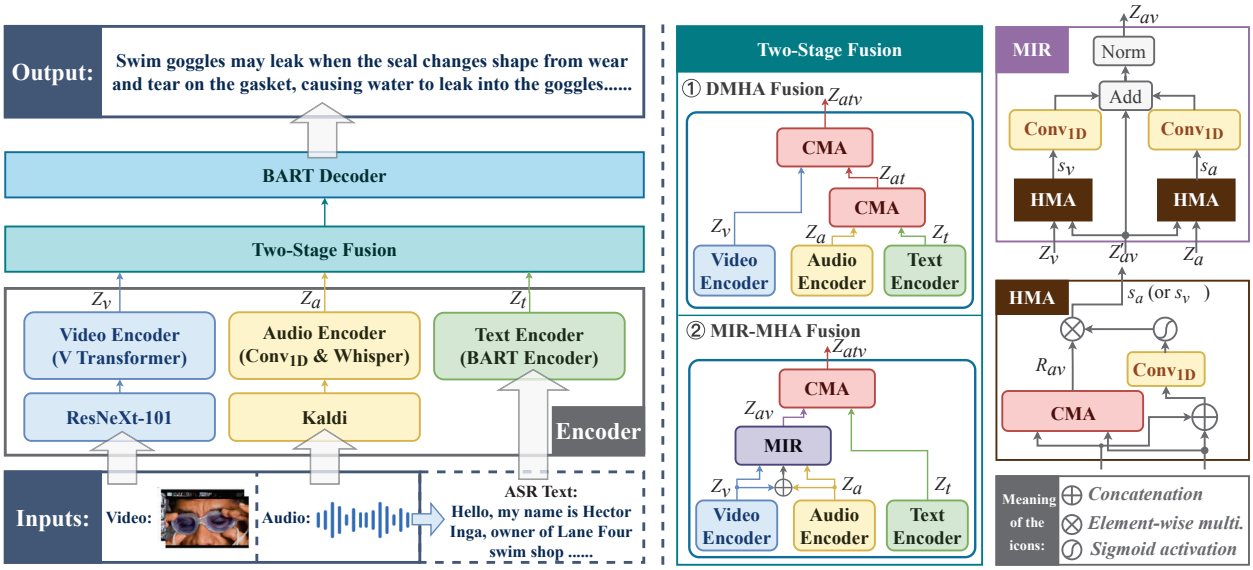


Fig. 2. The workflow of the proposed video summarization method. We propose two kinds of two-stage fusion methods, named DMHA fusion and MIR-MHA fusion, to combine information from different input modalities and generate the summary.

that audio input helps generate a better summary even if the input texts have certain errors.

## II. Proposed Method

The workflow of our proposed method is shown in Fig. 2. This method takes video frames and audio waveforms as inputs and produces text summaries as the output. For the given video frames and audio waveform inputs, visual features, audio spectrograms, and ASR texts are first extracted from the corresponding input modalities. Then, video transformer [7], BART [4] and frozen whisper [14] are used to represent visual, text, and audio information respectively. Next, two-stage fusion strategies are introduced to combine information from video, audio, and text modalities together. The fused features are conveyed into a BART decoder to obtain the text summary. The detailed method will be introduced in the rest of this section.

### A. Video and Audio Inputs

For video features, we follow previous work [2], [24] to retrieve 2048D representation by a 3D ResNeXt-101 model [25] pre-trained on the Kinetics dataset [26]. We further process this 2048D feature with a visual transformer [7] encoder, incorporating positional embeddings. The output from this process serves as our visual input.

For audio information, we follow previous work [24] to retrieve audio spectrogram from audio waveforms using Kaldi [27]. We extract 40D filter bank features from raw speech. Then, we obtain the audio feature representation using pre-trained whisper [14]. We set up a multi-layer convolution network to convert the input filter bank feature to 80D to fit the whisper inputs. We freeze the encoder and obtain the last hidden state from its output as the audio features.

### B. ASR Text

Literature [8] has demonstrated that the text summary can be generated by fine-tuning an ASR model, which suggests that leveraging ASR texts contributes to the generation of text summaries. Motivated by this finding, we employ a whisper model [14] to generate ASR texts. We freeze the parameters in its encoder and decoder and just let whisper recognize speech and generate the corresponding text. Then, we use a BART encoder [4] to encode the generated texts.

### C. Information Fusion

Information fusion at a deeper layer of the encoder tends to improve the quality of text summary [3]. Hence, we conduct two-stage fusion in the last layer of the encoder to combine information in video, audio, and text modalities together. We take ASR text as the main input flow to ensure the length is the same as that of the text. We employ two kinds of fusion methods called Dual Multi-Head-Attention (DMHA) fusion and Modality-Invariant-Representation Multi-Head-Attention (MIR-MHA) fusion respectively to combine the input information.

**DMHA Fusion** Yu et al [3] uses multi-head-attention to fuse text and visual information. We wonder how such a structure performs when audio information is also involved. Hence, we present DMHA fusion, which contains two multi-head-attention fusion. The fusion first carries A-T fusion which fuses audio and text information. Then it carries V-AT fusion which fuses the results of A-T fusion and the video information.

In A-T fusion, we use the last hidden state of BART encoder $Z_t$ as the query and use the audio feature $Z_a$ as both key and value. Linear projections $W_Q$, $W_K$ and $W_V$ are made to obtain $Q_{AT}$, $K_{AT}$ and $V_{AT}$, respectively.

$$Q_{AT} = Z_t W_Q \tag{1}$$

TABLE I
EXPERIMENTAL RESULTS OF OUR METHOD AND OTHER METHODS, WHERE GT MEANS USING GROUND TRUTH TEXTS, ASR MEANS USING SPEECH-RECOGNIZED TEXTS.

| Method | Input | | | Fusion | | Rouge-1 | Rouge-2 | Rouge-L |
|---|---|---|---|---|---|---|---|---|
| | Modality | Text | Punctuation | Method | Order | | | |
| **End-to-End** [9] | A | – | – | – | – | 60.9% | 43.0% | 55.9% |
| **BASS** [9] | A | – | – | – | – | 64.0% | 49.0% | 60.1% |
| **MCR** [12] | V+T | GT | GT | MHA | (in parallel) | 61.7% | 45.2% | 59.0% |
| **VG-VPLM**[*][3] | V+T | GT | GT | MHA | V-T | 65.3% | 49.3% | 60.4% |
| | V+T | ASR | Fullstop | MHA | V-T | 63.8% | 47.2% | 59.1% |
| | V+T | ASR | – | MHA | V-T | 55.6% | 36.7% | 49.7% |
| **Ours** | V+A+T | GT | GT | DMHA | A-T+V-AT | 64.7% | 48.6% | 59.5% |
| | V+A+T | ASR | Fullstop | DMHA | A-T+V-AT | 66.0% | 50.0% | 61.3% |
| | V+A+T | ASR | Fullstop | DMHA | V-T+A-VT | 64.9% | 48.9% | 60.8% |
| | V+A+T | ASR | – | DMHA | A-T+V-AT | 59.6% | 41.2% | 53.7% |
| | V+A+T | GT | GT | MIR-MHA | A-V+AV-T | 64.8% | 48.7% | 59.9% |
| | V+A+T | ASR | Fullstop | MIR-MHA | A-V+AV-T | **66.4%** | **50.2%** | **61.6%** |
| | V+A+T | ASR | – | MIR-MHA | A-V+AV-T | 58.8% | 40.4% | 53.1% |

*Based on our experiments.

$$K_{AT} = Z_a W_K \tag{2}$$

$$V_{AT} = Z_a W_V \tag{3}$$

We then set up cross-modal multi-head attention (CMA) to get the audio features $R_{AT}$ based on the text query.

$$R_{AT} = \mathrm{CMA}(Q_{AT}, K_{AT}, V_{AT}) \tag{4}$$

To remove redundant information and noise, we use a forget gate on $R_{AT}$ by setting up a forget gate mask $M$ and doing element-wise multiplication (shown with $\otimes$) with $R_{AT}$ to output the updated $R'_{AT}$:

$$M = \mathrm{sigmoid}(\mathrm{Concatenate}(R_{AT}, Z_t)W_m) \tag{5}$$

$$R'_{AT} = M \otimes R_{AT} \tag{6}$$

We concatenate $Z_t$ and $R'_{AT}$, and linearly project it to specific dimensions with $W_{AT}$ to obtain $Z'_{at}$.

$$Z'_{at} = \mathrm{Concatenate}(Z_t, R'_{AT})W_{AT} \tag{7}$$

We then obtain $Z_{at}$ by adding $Z'_{at}$ and $Z_t$.

$$Z_{at} = Z'_{at} + Z_t \tag{8}$$

We do a similar step for V-AT fusion. We use $Z_{at}$ to obtain $Q_{ATV}$, and visual feature $Z_v$ to obtain $K_{ATV}$ and $V_{ATV}$ respectively.

**MIR-MHA Fusion** Modality-Invariant Representation (MIR) [19], [23] captures the shared information in different modalities to ease the fusion process. In this work, we introduce MIR-MHA fusion based on MIR fusion. Our MIR-MHA fusion first carries A-V fusion for audio and video information using the MIR technique and then carries AV-T fusion to fuse the results of A-V fusion and the text information.

In A-V fusion, we first linearly project the audio feature $Z_a$ and visual feature $Z_v$ into the same dimension with $W_a$ and $W_v$ to obtain $Z'_a$ and $Z'_v$, respectively.

$$Z'_a = Z_a W_a \tag{9}$$

$$Z'_v = Z_v W_v \tag{10}$$

We vertically concatenate $Z'_a$ and $Z'_v$ to obtain the shared information $Z'_{av}$.

$$Z'_{av} = \mathrm{Concatenate}(Z'_a; Z'_v) \tag{11}$$

Then, we convey $Z'_a$, $Z'_v$ and $Z'_{av}$ into the MIR-generator, where a hybrid-modal attention (HMA) is set up to extract information in each modality-specific representations.

$$s_m = \mathrm{HMA}(Z'_m, Z'_{av}), m \in \{a, v\} \tag{12}$$

where the details of the HMA module will be described later.

Next, the resulted $s_m$ ($m \in \{a,v\}$) features are added to input sequence $Z'_{av}$ to obtain modality-invariant representation $Z^{inv}_{av}$. 1×1 convolution with PReLU activation [28] and layer normalization [29] are used here.

$$Z^{inv}_{av} = \mathrm{Norm}(Z'_{av} + \sum_{m \in \{v,a\}} \mathrm{conv}(s_m)) \tag{13}$$

Finally, the modality-specific information and representations are concatenated to get the representations of MIR fusion $Z_{av}$.

$$Z_{av} = \mathrm{Concatenate}(s_v, s_a, Z^{inv}_{av}) \tag{14}$$

Where $s_m$ contains information from both audio and visual modalities.

The HMA module in the MIR generator extracts representation for input audio and video features respectively using multi-head-attention. Here, $Z'_{av}$ is used as query, audio (video) feature $Z'_a$ ($Z'_v$) is used as both key and values.

$$R_{AV} = \mathrm{CMA}(Z'_{av}, Z'_m, Z'_m), m \in \{a, v\} \tag{15}$$

A parallel convolutional network is then used to learn the mask for the modality-specific information.

$$s_m = R_{AV} \otimes \sigma(\mathrm{conv}(\mathrm{Concatenate}(Z'_m, Z'_{av}))), m \in \{a, v\} \tag{16}$$

Where $\otimes$ means element-wise multiplication.

To fuse $Z_{av}$ with text information on AV-T fusion, we set up another multi-head-attention fusion. We use the last hidden state of the original BART encoder $Z_t$ as the query and use $Z_{av}$ as both key and value. This multi-head-attention is similar to the one proposed in DMHA fusion.

### D. Decoding and Generation

We use a BART decoder to decode the sequence. The features obtained by two-stage fusion are fed into the decoder, which generates potential hypotheses and their corresponding probabilities through beam search. The decoding proceeds in a left-to-right direction until the decoded token reaches an `<end>` mark or the maximum length of the generated summary. Finally, the hypothesis with the highest probability is chosen as the definitive summary.

## III. Experimental Evaluations

### A. Method Implementation

**Dataset** We conducted our experiments using the how2 dataset [24], which comprises 72,980 instructional videos spanning a total duration of 2,000 hours. This dataset provides video, audio, and text information. The audio information includes the sentence-level filter bank and the video-level filter bank, while the text information contains video transcripts and short text summaries. The audio filter bank in the dataset is derived from 16 kHz raw speech with a 25 ms time window and a 10 ms frame shift using Kaldi [27]. The videos in the dataset have a wide range of domains, with 68,333 for training, 2,520 for validation, and 2,127 for test. The corresponding sentence-level filter bank features are 1,013,715, distributed as 950,026 for training, 34,687 for validation, and 29,002 for test.

**ASR Text Generation** We used original filter bank features in the training set to generate ASR text. Since the input of whisper needs to be an 80D audio spectrogram, we establish a multi-layer convolution network preceding the pre-trained whisper model. We padded the filter bank features to fit the 3000 frame whisper input and trained the convolution layer before the whisper encoder, with the filter bank extracted from the top 2.5% (approximately 23,000) sentences. Subsequently, we employed this trained model to generate ASR text for the entire sentence-level dataset. To acquire ASR text at the video level, we concatenated the sentence-level ASR texts corresponding to each video.

**Data Preprocessing** We set the max length to 256 for video features, and 512 for text. The video/text sequences were padded or truncated to the corresponding max length before being conveyed into the model.

For audio features, video-level filter banks usually contain 6000-9000 frames [8], exceeding the capacity of the whisper input (3000 frames). To address this, we utilized whisper to generate audio features for every 3000 frames from the video-level filter bank. These generated frames were concatenated and down-sampled every 3 frames to fit the 3000-frame input of the whisper encoder.

**Hyper Parameters** We generated ASR text using pre-trained `whisper-tiny.en` model[1] with 4 layers in both encoder and decoder. We also generated the audio feature from the last hidden state of the encoder. We modeled the text feature using pre-trained `bart-base` model[2] with 6 layers in both encoder and decoder. Both models are not case-sensitive. For the video transformer, we used a 4-layer encoder with 8 attention heads and 2048 feed-forward dimensions. We set the max length of our generated summary to 64, the batch size to 8, and the learning rate to 3e-4. Adam optimizer [30] was used during optimization. The model was trained for 60 epochs.

**Hardware and Software** We performed our experiments on a computer with Intel Xeon Gold 6142 CPU, 128GB RAM, 3T HDD, and Nvidia Titan GPU. Our program was made using Python 3.9 and Pytorch 1.13.1 on Ubuntu 20.02.

### B. Results and Analysis

We take End-to-End [8] and BASS [9] which use audio input only, VG-GPLM [3] and MCR [12] which use both video and ground truth text as our baseline. We adopt Rouge-1, Rouge-2, and Rouge-L [31] to evaluate the summarization results. We use a pre-trained fullstop model [32] to give punctuation to ASR text. We also test our method under the ground truth text with ground truth punctuation. Our results are in table I. Note that since whisper is not designed to be fine-tuned, we did not conduct experiments for video and audio inputs.

From the results, when we use V+A+T features with ASR text, our method has a better performance compared with VG-GPLM [3] that uses V+T features with ASR text, and End-to-End method [8]. Notably, our approach even outperforms models using ground truth text and punctuation. This means that information from different modalities (for V+T, it is A; for A, it is V+T) helps a lot to generate the summary.

We further compare results with and without punctuation, revealing better quality of the summary texts when punctuation is included. This is also reasonable since punctuation plays a crucial role in sentence segmentation and helps eliminate ambiguity in the input text.

## IV. Discussion

**Fusion Order** From table I, we find that when we use DMHA fusion, the proposed method, which first fuses A-T features and then fuses V-AT features, generates a better text summary than the one that first fuses V-T features and then fuses A-VT features. The reason is considered as that, audio and text share similar information which will strengthen the representation of the fused feature when faced with video information. We also find that using MIR fusion with punctuation slightly improves the rouge score in summarization. This means that MIR gives a stronger feature representation when fusing video and text information.

**The Contribution of Errors in Transcripts** We find that our approach outperforms the model using ground truth text and punctuation. We wonder why this happens and conduct a deeper investigation on this problem using our model with DMHA fusion and VG-GPLM [3] model.

The first finding is that there is a space before each punctuation in the ground truth text, making the encoding of the same punctuation different between ground truth and ASR text. We remove space to form a GT-Fixed set and run experiments again. The results are in the middle of table II.

[1] https://huggingface.co/openai/whisper-tiny.en
[2] https://huggingface.co/facebook/bart-base

TABLE II
EXPERIMENTAL RESULTS OF OUR METHOD USING DIFFERENT KINDS OF TEXT, WHERE ASR MEANS USING SPEECH-RECOGNIZED TEXTS, GT MEANS USING GROUND TRUTH TEXTS, GT-FIXED MEANS USING GROUND TRUTH TEXTS WITH FIXED PUNCTUATION, REMOVE SHORT AND ADD EXTRA MEANS THE SIMULATION OF ASR SHORT DELETION AND INSERTION ERRORS RESPECTIVELY.

| Text | Method | Rouge-1 | Rouge-2 | Rouge-L |
|---|---|---|---|---|
| ASR Text | Ours | 66.0% | 50.0% | 61.3% |
| | VG-GPLM | 63.8% | 47.2% | 59.1% |
| GT | Ours | 64.7% | 48.6% | 59.5% |
| | VG-GPLM | 65.3% | 49.3% | 60.4% |
| GT-Fixed | Ours | **66.2%** | 50.2% | 61.4% |
| | VG-GPLM | 66.1% | 50.1% | 61.3% |
| Remove Short | Ours | 66.1% | **50.3%** | **61.7%** |
| | VG-GPLM | 66.0% | 50.4% | 61.6% |
| Add Extra | Ours | 65.4% | 49.4% | 60.6% |
| | VG-GPLM | 64.5% | 48.5% | 59.5% |

We also wonder how ASR errors affect the results. We compare the ASR transcripts and ground truth transcripts and simulate ASR short deletion and insertion errors. Here, we define *short deletion* as no more than 10 tokens between ASR text and ground truth text in each transcript and we insert all the extra tokens, i.e., all ASR insertion errors to the ground truth text for the insertion error. We use `fullstop` model to rewrite the punctuation based on such simulations while keeping other transcripts the same as is. We show these results in the latter part in table II.

From the results, when we run the experiments on the GT-Fixed set, our model has comparable results with the one using ASR text. This means punctuation has a crucial influence when connecting audio and text information.

We find that when we use *Remove Short* text, the rouge score has no obvious difference between our method and VG-GPLM method. This means using audio or not has little influence on the quality of the generated summary.

We also find that when we use ASR text and *Add Extra* texts, the rouge score using our method is better than using VG-GPLM method. This means that the audio signal in such summarization task helps resist some errors caused by input audio in the texts and complements the information that cannot be captured in texts. This further implies that such summarization tasks can be finished with their original audio and video information.

**Prediction Example** We give an example of generating the summary using our method with MIR-MHA fusion and VG-GPLM [3] respectively in Fig. 3, where both methods use ASR texts. The texts in red are the matched texts between the ground truth summary and the summary generated from both methods. Note that we omitted some ASR texts in the figure.

From the example, our method using V+A+T input has more matched words compared with the one using V+T input. To analyze in detail, the proposed method correctly generates the word `expert` and `planting`, but fails to generate the word `spring`. The proposed method also continuously generates the words `get professional advice from an expert on` corresponding to the ground truth, while the method using V+T input generates `get advice from a professional on`. This means audio also provides effective information for a better summary.

| | |
|---|---|
| **Video Frames** |  |
| **ASR Text** | back to the seedling tomato plants for spring planting in the garden. the time that it takes for the seed to germinate are going to be largely determined by the average temperature of the pot …… you've got a tomato plant that is already established root and just needs to be kept with a certain. generally speaking, you can tell that by looking. |
| **GT Summary** | how to care for tomato seedlings for spring planting; get professional tips and advice from an expert on growing your own fruits and vegetables in this free gardening video. |
| **(Proposed) V+A+T Summary** | caring for tomato seedlings? get professional advice from an expert on tomato planting and caring in this free video. |
| **(VG-GPLM) V+T Summary** | looking for tomato seedlings? get advice from a professional on spring tomato caring in this free video. |

Fig. 3. An example showing the generated summary with ASR text using our method with MIR-MHA fusion and VG-GPLM. Texts in red are the matched texts between the ground truth and the summary generated from both methods.

## V. CONCLUSION AND OUTLOOK

In this work, we proposed a new method that generates text summaries based on video and audio inputs. We used whisper to generate text from the input audio. Then, we employed the video transformer, BART, and frozen whisper to represent input information. We designed two kinds of two-stage fusion methods to combine information from different modalities and adopted a BART decoder to obtain the final summary texts. Experiments showed that our method outperforms the previous baselines. Further experiments proved that audio information helps generate a better summary even if the texts contain errors. Next, we want to apply a general-purpose sound feature extractor for audio modality instead of ASR. We are also interested in generating key video frames to make our summary contain both graphics and text.

**Acknowledgments** This work was supported in part by a project, JPNP20006, commissioned by NEDO, and JST CREST Grant Number JPMJCR22D1, Japan.

## REFERENCES

[1] N. Liu, X. Sun, H. Yu, W. Zhang, and G. Xu, "Multistage fusion with forget gate for multimodal summarization in open-domain videos," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 1834–1845.

[2] S. Palaskar, J. Libovickỳ, S. Gella, and F. Metze, "Multimodal abstractive summarization for how2 videos," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 6587–6596.

[3] T. Yu, W. Dai, Z. Liu, and 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.

[4] M. Lewis, Y. Liu, N. Goyal, *et al.*, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[5] C. Raffel, N. Shazeer, A. Roberts, *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[6] W. Qi, Y. Yan, Y. Gong, *et al.*, "Prophetnet: Predicting future n-gram for sequence-to-sequencepre-training," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 2401–2410.

[7] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[8] R. Sharma, S. Palaskar, A. W. Black, and F. Metze, "End-to-end speech summarization using restricted self-attention," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 8072–8076.

[9] R. Sharma, K. Zheng, S. Arora, S. Watanabe, R. Singh, and B. Raj, "Bass: Block-wise adaptation for speech summarization," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2023, 2023, pp. 1454–1458.

[10] Z. Xu, X. Meng, Y. Wang, *et al.*, "Learning summary-worthy visual representation for abstractive summarization in video," *arXiv preprint arXiv:2305.04824*, 2023.

[11] L. Qiao, C. Wu, Y. Liu, H. Peng, D. Yin, and B. Ren, "Grafting pre-trained models for multimodal headline generation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 2022, pp. 244–253.

[12] J. Yuan, J. Yun, B. Zheng, L. Jiao, and L. Liu, "Mcr: Multilayer cross-fusion with reconstructor for multimodal abstractive summarisation," *IET Computer Vision*, 2023.

[13] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[14] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*, 2023, pp. 28 492–28 518.

[15] Y. Zhang, W. Han, J. Qin, *et al.*, "Google usm: Scaling automatic speech recognition beyond 100 languages," *arXiv preprint arXiv:2303.01037*, 2023.

[16] Z. Yao, L. Guo, X. Yang, *et al.*, "Zipformer: A faster and better encoder for automatic speech recognition," *arXiv preprint arXiv:2310.11230*, 2023.

[17] P. H. Seo, A. Nagrani, and C. Schmid, "Avformer: Injecting vision into frozen speech models for zero-shot av-asr," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 922–22 931.

[18] M. Han, F. Chen, Z. Ni, *et al.*, "Vilas: Integrating vision and language into automatic speech recognition," *arXiv preprint arXiv:2305.19972*, 2023.

[19] Y. Hu, C. Chen, R. Li, H. Zou, and E. S. Chng, "Mirgan: Refining frame-level modality-invariant representations with adversarial network for audio-visual speech recognition," *arXiv preprint arXiv:2306.10567*, 2023.

[20] J. Hong, M. Kim, J. Choi, and Y. M. Ro, "Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 783–18 794.

[21] J. He, Z. Yang, and T. Toda, "Ed-cec: Improving rare word recognition using asr postprocessing based on error detection and context-aware error correction," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–6.

[22] M. Anwar, B. Shi, V. Goswami, W.-N. Hsu, J. Pino, and C. Wang, "Muavic: A multilingual audio-visual corpus for robust speech recognition and robust speech-to-text translation," *arXiv preprint arXiv:2303.00628*, 2023.

[23] J. He, X. Shi, X. Li, and T. Toda, "Mf-aed-aec: Speech emotion recognition by leveraging multimodal fusion, asr error detection, and asr error correction," *arXiv preprint arXiv:2401.13260*, 2024.

[24] R. Sanabria, O. Caglayan, S. Palaskar, *et al.*, "How2: A large-scale dataset for multimodal language understanding," in *NeurIPS*, 2018.

[25] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[26] W. Kay, J. Carreira, K. Simonyan, *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[27] D. Povey, A. Ghoshal, G. Boulianne, *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[29] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

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

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

[32] O. Guhr, A.-K. Schumann, F. Bahrmann, and H. J. Böhme, "Fullstop: Multilingual deep models for punctuation prediction," Jun. 2021.