

Align and Attend: Multimodal Summarization with Dual Contrastive Losses

Bo He^{1*}, Jun Wang¹, Jieliu Qiu², Trung Bui³, Abhinav Shrivastava¹, Zhaowen Wang³

¹University of Maryland, College Park ²Carnegie Mellon University ³Adobe Research
{bohe, abhinav}@cs.umd.edu, junwang@umd.edu, jieliuq@andrew.cmu.edu, {bui, zhawang}@adobe.com

Abstract

The goal of multimodal summarization is to extract the most important information from different modalities to form summaries. Unlike unimodal summarization, the multimodal summarization task explicitly leverages cross-modal information to help generate more reliable and high-quality summaries. However, existing methods fail to leverage the temporal correspondence between different modalities and ignore the intrinsic correlation between different samples. To address this issue, we introduce *Align and Attend Multimodal Summarization (A2Summ)*, a unified multimodal transformer-based model which can effectively align and attend the multimodal input. In addition, we propose two novel contrastive losses to model both inter-sample and intra-sample correlations. Extensive experiments on two standard video summarization datasets (TVSum and SumMe) and two multimodal summarization datasets (Daily Mail and CNN) demonstrate the superiority of A2Summ, achieving state-of-the-art performances on all datasets. Moreover, we collected a large-scale multimodal summarization dataset BLiSS, which contains livestream videos and transcribed texts with annotated summaries. Our code and dataset are publicly available at <https://boheumd.github.io/A2Summ/>.

1. Introduction

With the development in multimodal learning, multimodal summarization has drawn increasing attention [1–9]. Different from traditional unimodal summarization tasks, such as video summarization [10–17] and text summarization [18–22], multimodal summarization aims at generating summaries by utilizing the information from different modalities. With the explosive growing amount of online content (e.g., news, livestreams, vlogs, etc.), multimodal summarization can be applied in many real-world applications. It provides summarized information to the users, which is especially useful for redundant long videos such as livestream

*Part of this work was done when Bo was an intern at Adobe Research.

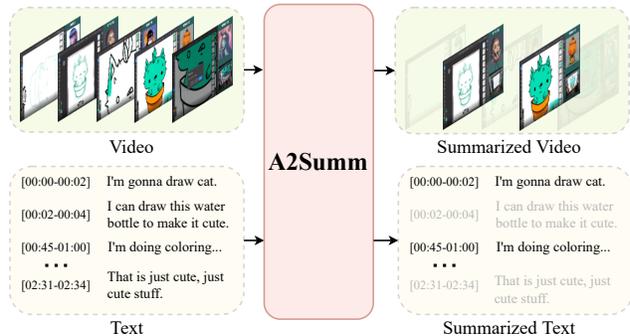


Figure 1. A2Summ is a unified multimodal summarization framework, which aligns and attends multimodality inputs while leveraging time correspondence (e.g., video and transcript) and outputs the selected important frames and sentences as summaries.

and product review videos.

Previous multimodal summarization methods [2, 4, 23, 24] leverage the additional modality information but can only generate the main modality summary, i.e., either a video summary or a text summary, severely limiting the use of complementary benefits in the additional modality. Recently, multimodal summarization with multimodal output (MSMO) has been explored in several studies [1, 6, 25, 26], which aim at generating both video and text summaries using a joint model. Compared to previous methods, which only produce a unimodal summary, MSMO provides a better user experience with an easier and faster way to get useful information. However, we find that the existing MSMO methods still have the following limitations. First, even if both modalities are learned together, the correspondence between different modalities is not exploited. For example, given a video and its transcripts, which are automatically matched along the time axis, no existing method utilizes the mutual temporal alignment information and treats the two modalities separately. Second, previous works adopt simple strategies to model the cross-modal correlation by sequence modeling and attention operation [1, 4, 25, 25, 26, 26], which requires a large number of annotated multimodal data which is hard to obtain.

Motivated by the above observations, we propose a novel architecture for multimodal summarization based on a uni-

fied transformer model, as shown in Figure 1. First, to leverage the alignment information between different modalities, we propose alignment-guided self-attention module to align the temporal correspondence between video and text modalities and fuse cross-modal information in a unified manner. Second, inspired by the success of self-supervised training [27–29], which utilizes the intrinsic cross-modality correlation within the same video and between different videos, we propose dual contrastive losses with the combination of an inter-sample and an intra-sample contrastive loss, to model the cross-modal correlation at different granularities. Specifically, the inter-sample contrastive loss is applied across different sample pairs within a batch, which leverages the intrinsic correlation between each video-text pair and contrasts them against remaining unmatched samples to provide more training supervision. Meanwhile, the intra-sample contrastive loss operates within each sample pair, which exploits the mutual similarities between ground-truth video and text summaries and contrasts the positive features against hard-negative features.

To facilitate the research of long video summarization with multimodal information, we also collected a large-scale livestream video dataset from the web. Livestream broadcasting is growing rapidly, and the summarization of livestream videos is still an unexplored area with great potential. Previous video summarization datasets consist of short videos with great variations in scene transitions. On the contrary, livestream videos are significantly longer (in hours as opposed to minutes) and the video content changes much more slowly over time, which makes it even harder for the summarization task. Besides, there has been a lack of annotated datasets with focus on transcript summarization, which can be a great complement to the livestream video summarization. Therefore, we collect a large-scale multimodal summarization dataset with livestream videos and transcripts, which are both annotated with ground-truth summaries by selecting important frames and sentences.

To summarize, our contributions include:

- We propose A2Summ, a unified transformer-based architecture for multimodal summarization. It can handle multimodal input with time correspondences which previous work neglects.
- We present dual contrastive losses that account for modeling cross-modal information at different levels. Extensive experiments on multiple datasets demonstrate the effectiveness and superiority of our design.
- A large-scale **Behance LiveStream Summarization (BLiSS)** dataset is collected containing livestream videos and transcripts with multimodal summaries.

2. Related Work

Video Summarization. Current techniques for video summarization can be divided into two categories, unsuper-

vised and supervised. Unsupervised learning approaches, including [11, 30–33, 33–39] utilize different hand-crafted features to score and select the video frames without the human-annotated summaries. DR-DSN [11] explores an unsupervised reward function to tackle video summarization. GLRPE [39] attempts to apply self-attention with relative position representation for unsupervised video summarization. With the help of the annotated video summarization datasets [40, 41], numerous supervised learning methods [14, 16, 17, 17, 42–45] have been proposed in recent years to summarize videos. Among them, DSNet [14] formulates supervised video summarization as a temporal interest detection process. RSGN [16] utilizes LSTM and GCN to model frame-level and shot-level dependencies. iPTNet [17] jointly trains the video summarization task and correlated moment localization task to utilize additional moment localization data samples to boost the video summarization performance.

Text Summarization. In general, text summarization can be categorized into two groups: (i) Extractive summarization [19–22, 46] generates output summary by identifying the salient parts of the input document. NN-SE [19] develops a neural attention model to select sentences or words of the input document as the output summary. SummaRuNNer [20] employs RNN for extractive summarization. Miller [21] adopts clustering algorithm in the feature space to select important sentences. (ii) Abstractive summarization [18, 47–50] performs the summarization by paraphrasing the important parts of the input document. Lead3 [18] applies the attentional encoder-decoder RNN for the task of abstractive text summarization. However, those approaches are designed for pure unimodal summarization that doesn’t consider cross-modal alignment and fusion. Recently, StreamHover [51] presents an unsupervised model for transcript summarization and collects a livestream transcript summarization dataset. Inspired by it, we collect a new livestream dataset with a much larger scale and richer multimodal annotations for the multimodal summarization task.

Multimodal Summarization. Existing work [2, 4, 23, 24, 52] commonly utilize additional complementary modalities to enhance the feature representation for the primary modality, however, they typically generate summaries from a single modality. For example, CLIP-It [2] builds a language-guided framework to obtain a summary video conditioned on the text. MMS [23] learns joint representations of text and images and outputs text summaries. Recently, multimodal summarization with multimodal output (MSMO) has been explored in several studies. Zhu et al. [1] propose the first MSMO model and collect a multimodal summarization dataset with text and image modalities. Li et al. [25] extend it with video-based news articles and adopt conditional self-attention for text and video fusion. Recently, Fu et al. [6] collect a multimodal dataset with more modalities included such as audio and transcript.

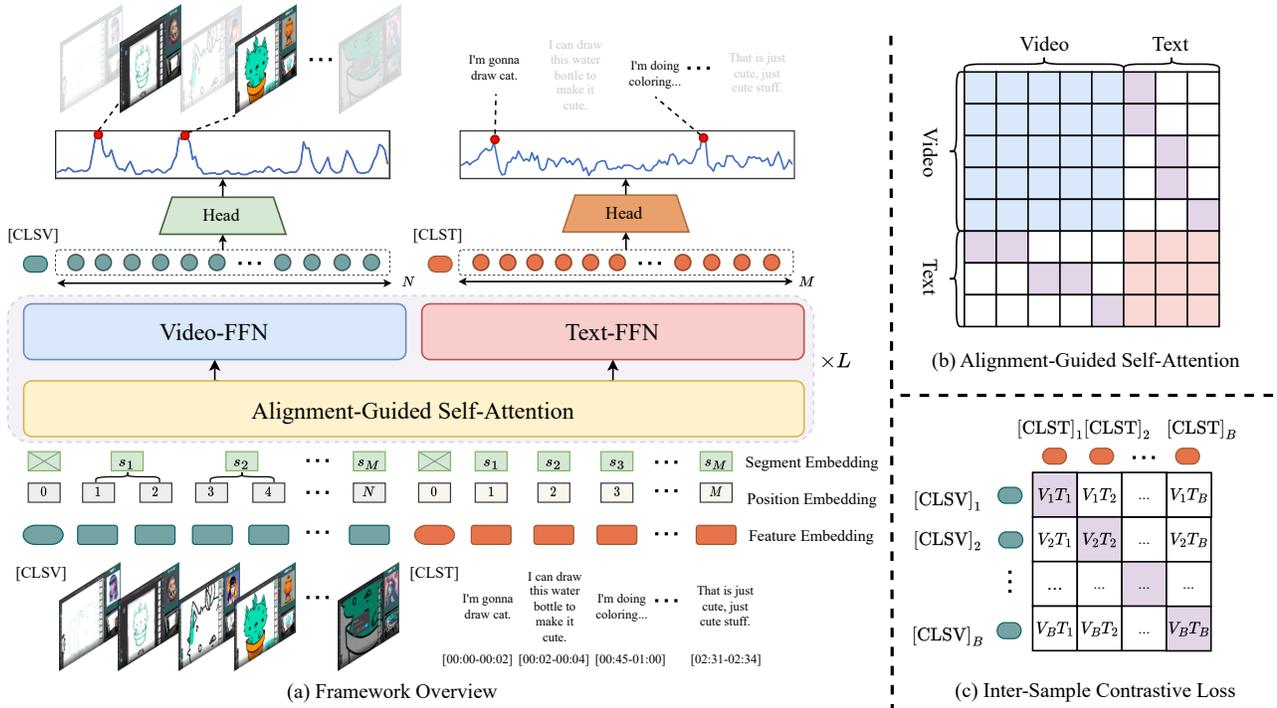


Figure 2. (a) The overview of A2Summ framework. Given N video frames and M text sequences as input, A2Summ predicts the important frames and sentences as multimodal summaries. (b) Alignment-guided self-attention module is applied to align and fuse each video and text pair. (c) Inter-sample contrastive loss is calculated by maximizing the similarity of [CLSV] and [CLST] tokens from the same pair while minimizing the similarity of tokens from different pairs. B is the batch size. Best viewed in color.

3. Method

3.1. Overview

Given the untrimmed multimodality input (*e.g.*, video, text, and sound), the multimodal summarization task aims at selecting the most important parts from each modality. Figure 2(a) illustrates an overview of our A2Summ framework. The input to our model is the multi-modality (*e.g.*, video and transcribed text in our case) with N video frames and M sentences. Since each transcribed sentence has its start time and end time, the video and text modalities can be automatically aligned by the corresponding timestep. The overall architecture can be divided into three parts: the input embedding (Sec. 3.2), the multimodal alignment and fusion (Sec. 3.3), and the loss function (Sec. 3.4).

3.2. Input Embedding

Similar to previous work [10, 14, 17, 53], we use pre-trained feature extraction models (*e.g.*, GoogleNet [54] and RoBERTa [55]) to extract deep neural features for each frame and sentence. After feature extraction, features from different modalities are projected into a common C -dimensional embedding space by a linear fully connected (FC) layer. Specifically, we denote the generated video and text features as $F \in \mathbb{R}^{N \times C}$ and $S \in \mathbb{R}^{M \times C}$, respectively.

For each modality, there is a special token “[CLS]”

prepended at the start of the feature sequences, which enables a holistic representation. Following BERT [56], we add a learnable position embedding to each feature sequence so that the order information can be incorporated. To utilize the time correspondence information between the video frames and text sentences, we add an additional learnable segment-based positional embedding at the input stage. More precisely, each sentence has its own timestep information denoted as $[t_s, t_e]$, where t_s and t_e denote the start and the end time index of each sentence. We note that a single text sentence usually corresponds to several video frames, making $M \leq N$. For all frames inside each time index window $\{F_i\}_{i \in [t_s, t_e]}$, the segment embedding is shared across these frames and the corresponding sentence. After adding these positional embeddings, the input sequences to the multimodal transformer from both modalities are concatenated along the time axis, denoted as $X \in \mathbb{R}^{(M+N) \times C}$.

3.3. Multimodal Alignment and Fusion

Alignment-Guided Self-Attention. A core component of A2Summ is the alignment-guided self-attention module which allows us to exploit the time correspondence between video and text modalities. Inspired by the superior advantages of Transformers [57] in modeling different modalities (*e.g.*, visual, language, and audio) on various multimodal tasks (*e.g.*, visual question answering [58–62],

vision-language pre-training [63–65]), we adopt the transformer architecture to align and fuse our multimodal input. However, for the multimodal summarization task, the inputs are often untrimmed videos and text sentences, which are dominated by irrelevant backgrounds. Directly applying global self-attention across inputs from all modalities may introduce extra noise to the multimodal fusion process.

Motivated by this observation, we propose the alignment-guided self-attention module to fuse the input across different modalities. We formulate this process by using a masked self-attention operation in Figure 2(b). Specifically, an attention mask $A \in \mathbb{R}^{(N+M) \times (N+M)}$ initialized with all 0 is defined to indicate the timestep alignment information, where N and M denote the length of the video and text feature sequences, respectively. For the intra-modality modeling, we follow the standard procedure with global attention operation, where features from the same modality can attend to each other such that all entries corresponding to intra-modality attention are filled with value 1 in the attention mask. For the cross-modality attention between video and text input, we only fill in the entries from the same segment with value 1. For example, suppose the k^{th} sentence S_k corresponding to the time index window $[t_s, t_e]$. We consider the frames which also lie into the same time window $[t_s, t_e]$ to be the same segment, denoted as $\{F_i\}_{i \in [t_s, t_e]}$. Then, we assign the elements of attention mask as follows $A[N+k, t_s : t_e] = 1$. The attention mask is then applied to the attention matrix computed by the standard self-attention approach [57, 66–68]:

$$Q = XW_Q, K = XW_K, V = XW_V, \quad (1)$$

$$D_{i,j} = \frac{A_{i,j} \exp(Q_i K_j^T / \sqrt{D})}{\sum_k A_{i,k} \exp(Q_i K_k^T / \sqrt{D})}, \quad (2)$$

$$Z = X + DVW_O, \quad (3)$$

where $i, j \in [1, M + N]$ are the entry indices of the matrix, X is the concatenated input from video and text modalities, and $W_Q, W_K, W_V, W_O \in \mathbb{R}^{C \times C}$ are the linear projection matrices for generating the query, key, value, and the output. Multi-head attention [57] is also adopted to improve the capacity of the attention module. In this way, we explicitly utilize the alignment correspondence between different modalities, avoiding the negative impacts caused by noisy background frames or irrelevant sentences.

Mixture-of-Modality-Experts. Based on the mixture-of-modality-experts transformer [61] in the multimodal tasks, after the self-attention layer, we introduce two different experts to jointly model features from different modalities including the video expert (Video-FFN), and text expert (Text-FFN) rather than the standard shared FFN [57].

Score Prediction. Finally, on top of the cascaded transformer blocks, two separate score prediction branches assign relevance scores to each frame and each sentence. Based

on predicted scores, two different procedures are followed to generate the final summary. For the standard video summarization datasets (e.g., SumMe [41] and TVSum [40]), based on the pre-processed KTS [69] segmentation results, segment-level scores are computed from frame-level scores, and the final video summary is generated by selecting top 15% of video durations by Knapsack algorithm. For the multimodal summarization datasets (e.g., Daily Mail [6]), the frames and sentences with the top highest scores are selected to generate the final summary prediction for the video and text modalities separately.

3.4. Loss Function

We employ three different loss functions to train our model, including the classification loss and the novel dual contrastive losses, which consist of the inter-sample contrastive loss and the intra-sample contrastive loss.

Classification Loss. We apply the focal loss [70] for the importance score classification, which handles the class imbalance issue by down-weighting losses for well-classified samples. The details are shown below:

$$\mathcal{L}_{\text{cls}_m} = -\frac{1}{N} \sum_{i=1}^N \begin{cases} -\alpha(1-p_i)^\gamma \log(p_i), & \text{if } y_i=1 \\ -(1-\alpha)p_i^\gamma \log(1-p_i), & \text{if } y_i=0 \end{cases} \quad (4)$$

$$\mathcal{L}_{\text{cls}} = \mathcal{L}_{\text{cls}_{\text{video}}} + \mathcal{L}_{\text{cls}_{\text{text}}} \quad (5)$$

where m could be either video or text, and p_i is the predicted score for each frame/sentence while y_i is the ground-truth label. If $y_i=1$, it indicates the i^{th} frame/sentence is the key-frame/key-sentence. The final classification loss is the sum of the two single modality losses.

Inter-Sample Contrastive Loss. Driven by the success of contrastive learning in the image-language pre-training tasks [71–73], we want to utilize the intrinsic relationships between each input video and text pair. As shown in Figure 2(c), given a batch of B sample pairs, we design an auxiliary inter-sample contrastive loss to predict which of the B^2 possible video-text pairs across a batch correctly matches and belongs to the same sample. Specifically, we use the pre-pended [CLS] token as a holistic representation for each video and text sample. Similar to CLIP [71], we maximize the cosine similarity of the video embedding [CLSV] and the text embedding [CLST] from B real pairs in the batch while minimizing the cosine similarity of embeddings from the $B^2 - B$ incorrect pairs. Specifically, the inter-sample contrastive loss is calculated as

$$\begin{aligned} \mathcal{L}_{\text{inter}} = & \mathbb{E}_{z \sim [\text{CLSV}]_j, z^+ \sim [\text{CLST}]_j, z^- \sim \mathcal{I}_{k \neq j} [\text{CLST}]_k} \ell(z, z^+, z^-) \\ & + \mathbb{E}_{z \sim [\text{CLST}]_j, z^+ \sim [\text{CLSV}]_j, z^- \sim \mathcal{I}_{k \neq j} [\text{CLSV}]_k} \ell(z, z^+, z^-) \end{aligned} \quad (6)$$

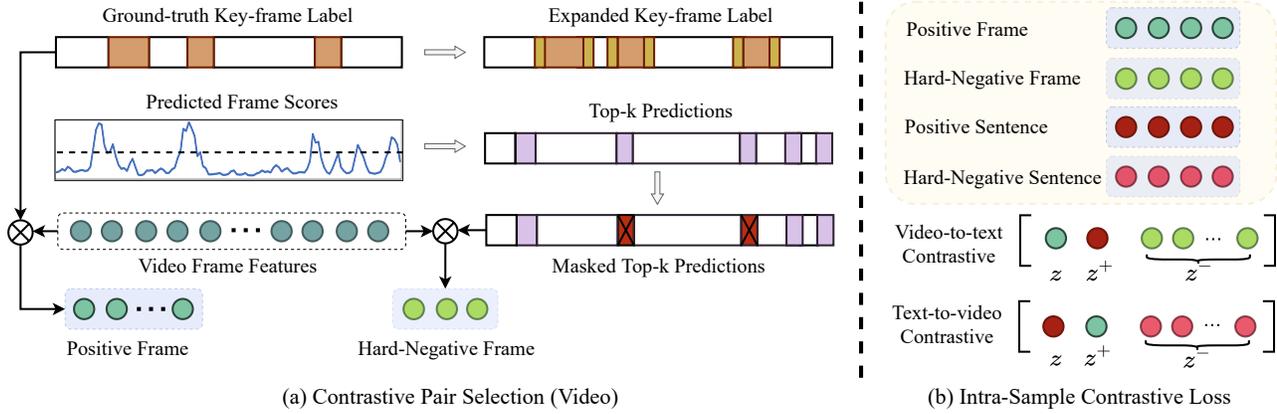


Figure 3. (a) Contrastive pair selection process for selecting positive and hard-negative video frame features. The same procedure is applied to the text modality. The crossed red boxes denote the top predicted time steps masked out by the expanded key-frame label. (b) Intra-sample contrastive loss is applied between the selected video and text pairs. Best viewed in color.

where $\ell(z, z^+, z^-)$ is the standard contrastive loss [72] with the following equation:

$$\begin{aligned} \ell(z, z^+, z^-) &= -\log \left(\frac{\exp(z^T \cdot z^+ / \tau)}{\exp(z^T \cdot z^+ / \tau) + \sum_k \exp(z^T \cdot z_k^- / \tau)} \right) \end{aligned} \quad (7)$$

and τ is a learnable temperature parameter.

Intra-Sample Contrastive Loss. While the above inter-sample contrastive loss only considers the relationship across different samples, however, for the summarization task, to correctly detect the key-frames and key-sentences from each untrimmed video and text input, more fine-grained information modeling, in particular, is crucial. It would require the model to accurately distinguish the key-frames and key-sentences from the background frames and less-related sentences. Intuitively, the human-annotated key-frames and key-sentences share mutual information with each other. Meanwhile, they both should reveal the most salient parts from the original untrimmed video and text sequences. For instance, for a cooking recipe video with transcribed text, the annotated key-frames and key-sentences should clearly reveal the instructions for each step. More importantly, these key-frames and key-sentences should be deeply correlated with each other and share similar high-level semantic meanings. Motivated by this observation, we propose the intra-sample contrastive loss which is calculated within each video and text pair sample rather than across different sample pairs.

Specifically, we assign features associated with the predefined ground-truth key timesteps as positive pairs for both modalities. To form the contrastive pairs, as pointed out by [74, 75], the quality of the negative samples is of vital importance for the effectiveness of contrastive learning. Therefore, we need to select the hard negative samples for video and text separately. Specifically, since the pre-annotated

non-key timesteps are negative samples, based on the prediction scores for each frame ($p_{i=1}^N$) and sentence ($q_{i=1}^M$), we argue that the wrongly classified timesteps with highest prediction scores are hard-negative samples. Intuitively, for a long untrimmed video, due to the time dependencies, the frames adjacent to the key-frames have very similar visual contents and should also be treated as the key-frames. However, if these frames are selected as the hard-negative samples, it tends to confuse the model and may hurt the final performance. Therefore, we exclude those timesteps before selecting the hard-negative samples.

As shown in Figure 3(a), given the ground truth (GT) key-frame label, we first expand the key-frame segments on both sides to include more adjacent frames as dummy key-frames. Then, based on the predicted scores for each timestep, we select timesteps with top- k highest scores but not in the expanded GT key-frame labels as hard-negative samples. Here, $k_{\text{video}} = \lfloor \frac{N}{r} \rfloor$, $k_{\text{text}} = \lfloor \frac{M}{r} \rfloor$, r is a hyper-parameter controlling the total number of selected hard-negative samples. In this way, we form contrastive pairs for both video and text modalities. Formally, we denote the positive frames, hard-negative frames, positive sentences, and hard-negative sentences as \mathcal{I}_{PF} , \mathcal{I}_{HNF} , \mathcal{I}_{PS} , and \mathcal{I}_{HNS} , respectively. As shown in Figure 3(b), the proposed intra-sample contrastive loss is applied as follows:

$$\begin{aligned} \mathcal{L}_{\text{intra}} &= \mathbb{E}_{z \sim \mathcal{I}_{\text{PF}}, z^+ \sim \mathcal{I}_{\text{PS}}, z^- \sim \mathcal{I}_{\text{HNF}}} \ell(z, z^+, z^-) \\ &+ \mathbb{E}_{z \sim \mathcal{I}_{\text{PS}}, z^+ \sim \mathcal{I}_{\text{PF}}, z^- \sim \mathcal{I}_{\text{HNS}}} \ell(z, z^+, z^-) \end{aligned} \quad (8)$$

where ℓ follows the same contrastive equation as Eq. 7.

Overall Loss. The final loss is the combination of the above three losses,

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \beta \cdot \mathcal{L}_{\text{inter}} + \lambda \cdot \mathcal{L}_{\text{intra}} \quad (9)$$

where β and λ are hyper-parameters controlling the trade-off between three loss components.

4. Experiments

4.1. Datasets and Implementation Details

Datasets. We evaluate A2Summ on two standard video summarization datasets (SumMe [41] and TVSum [40]), two multimodal summarization datasets (Daily Mail [6] and CNN [6]), and a new Behance LiveStream Summarization (BLiSS) dataset. **TVSum** dataset consists of 50 videos pertaining to 10 categories. **SumMe** dataset consists of 25 videos capturing multiple events. **Daily Mail** dataset contains 1,970 samples and **CNN** dataset contains 203 samples, which are crawled from the news website including video, images, text articles, and captions. We follow the same data split as [6]. The **BLiSS** dataset consists of 13,303 pairs of livestream videos and transcribed text, with annotated summaries for both modalities.

Evaluation Metrics. For SumMe and TVSum datasets, following previous work [2, 10, 12, 14, 17, 33], we evaluate the video summarization dataset by the F1 score metric. However, as pointed out by [76], the performance of F1 evaluation is mostly determined by the pre-processing segmentation step, and a random method is able to reach similar performance scores. As a result, they propose to utilize rank order statistics (Kendall’s τ [77] and Spearman’s ρ [78]) as alternative evaluation metrics which are more reliable compared to F1 score. For multimodal summarization datasets, we evaluate the generated text summary by ROUGE [79] following previous works [6, 26, 47, 80]. Specifically, R-1, R-2, and R-L represent ROUGE-1, ROUGE-2, and ROUGE-L F1 scores, respectively, which are widely used to calculate the n-grams overlapping between the output text summary and ground truth text summary. Same as [6, 25, 26], the cosine image similarity is measured between the features of the predicted video summary and ground-truth video summary.

Implementation Details. For standard video summarization datasets (SumMe and TVSum), we follow previous work [10, 14, 17, 53] and use the pre-extracted GoogLeNet [54] feature as the video input. To collect the corresponding text modality, we adopt the pre-trained image caption model GPT-2 [81]¹ to generate the caption for each frame. Next, for all the text modality input, we apply the pre-trained RoBERTa [55]² to extract textual features for each sentence. For multimodal summarization datasets (Daily Mail and CNN), we use the same feature extractor as [6, 26]. For the BLiSS dataset, pre-trained CLIP [71] and RoBERTa [55] are adopted to extract features for each frame and each sentence. The focal loss [70] with $\alpha = 0.25$ and $\gamma = 2.0$ is adopted for the classification loss. More dataset-specific training/testing details and hyper-parameter choosing are described in the supplementary material.

¹<https://huggingface.co/nlpconnect/vit-gpt2-image-captioning>

²<https://huggingface.co/distilroberta-base>

4.2. BLiSS Dataset

Behance³ is a public website with a large amount of livestream videos created by artists showing their work process. The videos are generally hours long and accompanied with transcripts of streamers’ speeches. We follow previous work StreamHover [51] to expand their dataset to a much larger scale with more modalities.

Data Collection. We collected 674 livestream videos with transcripts and other metadata. Each video was further divided into 5-minute long clips for human annotation. Annotators were instructed to select the key sentences from transcripts, and write the text summary and key phrases in their own words for the entire clip. For each video, we also obtained its thumbnail animation from the website, and selected the most similar frame to the thumbnail from each clip as ground truth key-frames. More details about the collection process are elaborated in the supplementary material.

Comparison with Existing Multimodal Datasets. The BLiSS dataset is much larger than the standard video summarization datasets (SumMe and TVSum) and multimodal summarization datasets (Daily Mail and CNN). BLiSS has 13,303 data samples and 1,109 total video hours, which is much longer than TVSum (3.5 hours) and Daily Mail (44.2 hours). For the text modality, the total number of text tokens is 5.4M (BLiSS), greater than 1.3M (Daily Mail) and 0.2M (CNN). There are other multimodal summarization datasets for the abstractive text summarization task with additional image or video modalities. For example, MSMO [1], MMSS [24], VMSMO [25] and How2 [83]. However, none of them have aligned video and text modalities. Furthermore, we keep some metadata, including the title, streamer information, and audio modality for further potential research.

4.3. Results

SumMe and TVSum Datasets. We compare the proposed method A2Summ with the previous state-of-the-art (SOTA) methods on SumMe [41] and TVSum [40] datasets in Table 2. We first observe that A2Summ achieves the best performance on both datasets. Except for the F1 score metric, our A2Summ is slightly worse than CLIP-It [2] but still higher than it for the other two metrics on the TVSum dataset. CLIP-It also adopts transformer architecture to fuse different modalities by cross-attention, which takes in the generated video caption as text modality. However, it ignores the time correspondence between video and text modalities. Instead, our A2Summ aligns cross-modality information and exploits the intrinsic correlation between the video and text at different granularities by our inter-sample and intra-sample contrastive losses. In addition, the state-of-the-art method iPTNet [17] utilizes an additional moment localization dataset Charades-STA [85] to help address the data

³<http://behance.net/>

Table 1. Comparison with state-of-the-art methods on the CNN [6] and Daily Mail [6] datasets.

Category	Method	CNN			Daily Mail			
		R-1	R-2	R-L	R-1	R-2	R-L	Cos(%)
Video	VSUMM [30]	–	–	–	–	–	–	68.74
	DR-DSN [11]	–	–	–	–	–	–	68.69
	CLIP-It [2]	–	–	–	–	–	–	69.25
Text	Lead3 [18]	–	–	–	41.07	17.87	30.90	–
	SummaRuNNer [20]	–	–	–	41.12	17.92	30.94	–
	NN-SE [19]	–	–	–	41.22	18.15	31.22	–
Multimodal	MM-ATG [1]	26.83	8.11	18.34	35.38	14.79	25.41	69.17
	Img+Trans [5]	27.04	8.29	18.54	39.28	16.64	28.53	–
	TFN [82]	27.68	8.69	18.71	39.37	16.38	28.09	–
	HNNatTI [4]	27.61	8.74	18.64	39.58	16.71	29.04	68.76
	M ² SM [6]	27.81	8.87	18.73	41.73	18.59	31.68	69.22
Ours	Video-only	–	–	–	–	–	–	69.30
	Text-only	29.39	10.85	26.11	42.77	19.19	34.60	–
	A2Summ	30.82	11.40	27.40	44.11	20.31	35.92	70.20

Table 2. Comparison with state-of-the-art methods on the SumMe [41] and TVSum [40] datasets with F1 scores, Kendall’s τ [77] and Spearman’s ρ [78] metrics. We include the results of methods using GoogleNet [54] features for a fair comparison. **Bold** and underline represent the top-1 and top-2 results.

Method	SumMe			TVSum		
	F1	τ	ρ	F1	τ	ρ
Random [76]	41.0	0.000	0.000	57.0	0.000	0.000
Human [76]	54.0	0.205	0.213	54.0	0.177	0.204
DR-DSN [11]	42.1	–	–	58.1	0.020	0.026
HSA-RNN [45]	42.5	0.064	0.066	44.1	0.082	0.088
CSNet [33]	48.6	–	–	58.5	0.025	0.034
VASNet [84]	49.7	–	–	61.4	–	–
DSNet-AB [14]	50.2	0.051	0.059	62.1	0.108	0.129
DSNet-AF [14]	51.2	0.037	0.046	61.9	0.113	0.138
RSGN [16]	45.0	0.083	0.085	60.1	0.083	0.090
CLIP-It [2]	51.6	–	–	64.2	0.108	0.147
iPTNet [17]	<u>54.5</u>	<u>0.101</u>	<u>0.119</u>	63.4	<u>0.134</u>	<u>0.163</u>
A2Summ	55.0	0.108	0.129	<u>63.4</u>	0.137	0.165

scarcity problem but results in a much longer training time, however, without utilizing extra datasets, our A2Summ can still outperform it on all the metrics, which strongly justifies the superiority of our design.

Daily Mail and CNN Datasets. As shown in Table 1, we also compare our A2Summ with previous methods on the CNN [6] and Daily Mail [6] datasets. Since the text modality of CNN and Daily Mail datasets do not have

time information, we only apply the inter-sample and intra-sample contrastive losses without the alignment-guided self-attention module. We first observe that A2Summ can indeed greatly benefit from leveraging the multimodal information, which boosts the text summary metric by 1-2% ROUGE F1 score and increases the video summary cosine similarity by 0.9%. Compared to the state-of-the-art multimodal method M²SM [6], which utilizes additional transcript extracted from videos as the bridge between video and text modality, A2Summ is better by 3% and 2.4% in ROUGE-1 F1 score on two datasets respectively. For the video summarization, our transformer-based A2Summ can outperform multimodal summarization method M²SM [6] and state-of-the-art video summarization model CLIP-It [11] by 1%.

BLISS Dataset. We validate A2Summ on the livestream videos from the BLISS dataset by comparing it with existing video and text summarization methods. As shown in Table 3, when comparing with video summarization methods DSNet-AF [14] and CLIP-It [2], our A2Summ achieves the best results on the video cosine similarity metric. Although CLIP-It also utilizes the additional text modalities by the cross-attention operation, A2Summ can still outperform it by 1%. Compared to the extractive text summarization method SummaRuNNer [20] and the abstractive text summarization method BART [86], A2Summ outperforms both of them by at least 3% on all the ROUGE scores. It further demonstrates the superior effectiveness of A2Summ on livestream videos.

4.4. Ablation Studies

To further investigate the contribution of each component in A2Summ, we conduct ablation studies in Table 4. We first observe that adding the text modality input can en-

Table 3. Comparison results on the BLiSS dataset. Note that BART [86] is abstractive summarization based method, and the rest are extractive summarization based.

Method	R-1	R-2	R-L	Cos(%)
DSNet-AF [14]	–	–	–	62.70
CLIP-It [2]	–	–	–	63.58
Miller [21]	40.90	26.48	39.14	–
BART [86]	49.11	38.59	48.08	–
SummaRuNNer [20]	49.70	38.00	48.51	–
A2Summ	52.50	41.65	51.44	64.57

Table 4. Contribution of each component on the SumMe dataset. “Align.”, “Inter.”, and “Intra.” represent the alignment-guided self-attention, inter-sample and intra-sample contrastive loss, respectively.

Inputs	Align.	Inter.	Intra.	F1	τ	ρ
Video-Only				49.8	0.070	0.084
Multimodal				50.5	0.083	0.096
	✓			51.5	0.089	0.104
	✓	✓		52.5	0.905	0.110
	✓		✓	54.0	0.102	0.121
	✓	✓	✓	55.0	0.108	0.129

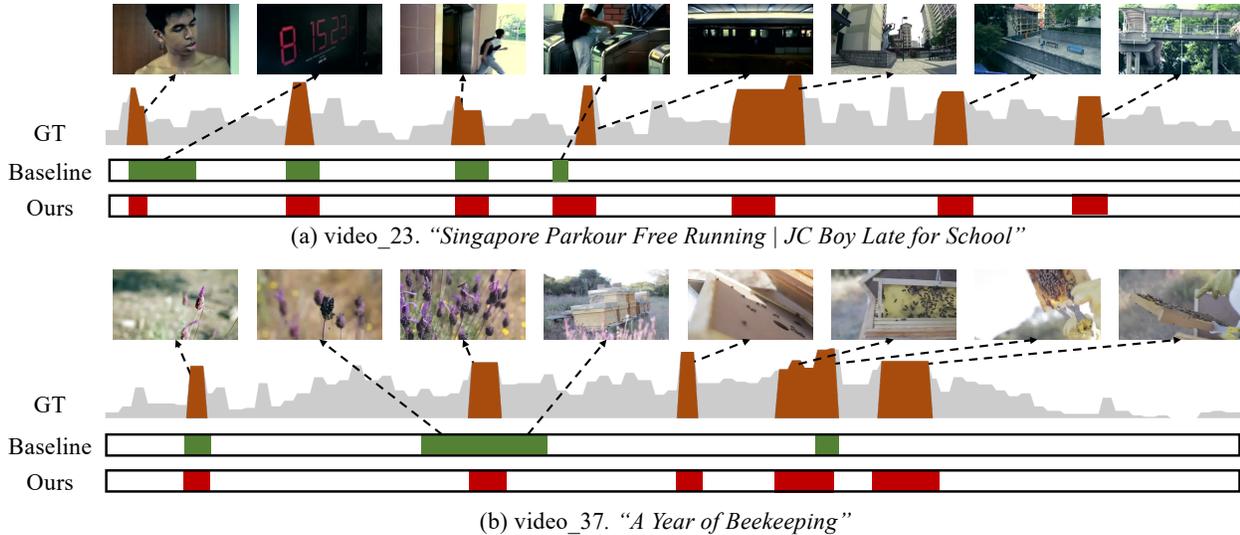


Figure 4. Example summarization results for the TVSum dataset. Titles are shown for each video. “Baseline” denotes our A2Summ without the alignment module and dual contrastive losses. The gray histogram shows the ground-truth importance scores for each frame.

hance the final results of video summarization. However, as we mentioned before, without alignment of video and text modalities, directly applying global attention between untrimmed video and text input tends to introduce too much noise and result in inferior performance. After we align and attend the video and text modalities with the proposed alignment-guided self-attention module, we can improve the F1 score by 1%. Furthermore, for the dual contrastive losses, it is obvious that a consistent gain can be achieved by adding either one of these two losses. In particular, introducing the intra-sample contrastive loss significantly increases the performance by 2.5%. It proves that exploring the intrinsic mutual correlation between video and text and mining hard negative samples can greatly enhance the ability to localize the important frames and sentences. In addition, two contrastive losses are complementary to each other. When incorporating all three proposed components together, our approach boosts the final performance from 50.5% to 55.0%.

4.5. Visualization

Figure 4 shows the visual comparison between baseline and A2Summ. We observe that the typical errors of the

baseline model can be addressed by the proposed alignment module and dual contrastive losses, such as missing detection of important segments and inaccurate summary boundary prediction. It further verifies the effectiveness of A2Summ. More visualizations on the BLiSS dataset are provided in the supplementary material.

5. Conclusion

In this paper, we present A2Summ, a novel unified transformer-based framework for multimodal summarization. A2Summ is designed to align and attend different modalities by leveraging time correspondences that previous methods neglect. Also, we introduce dual contrastive losses to exploit the inter-sample and intra-sample cross-modality information. Extensive experiments on multiple datasets validate the effectiveness of our A2Summ. In addition, we collect a large-scale multimodal summarization dataset focusing on livestream videos and transcripts. We hope it can be beneficial for further research in this area.

Acknowledgements. This work was partially supported by DARPA SemaFor (HR001119S0085) program and gifts from Adobe.

References

- [1] Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. Msmo: Multimodal summarization with multimodal output. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4154–4164, 2018. [1](#), [2](#), [6](#), [7](#)
- [2] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization. *Advances in Neural Information Processing Systems*, 34:13988–14000, 2021. [1](#), [2](#), [6](#), [7](#), [8](#)
- [3] Aman Khullar and Udit Arora. Mast: Multimodal abstractive summarization with trimodal hierarchical attention. *arXiv preprint arXiv:2010.08021*, 2020. [1](#)
- [4] Jingqiang Chen and Hai Zhuge. Abstractive text-image summarization using multi-modal attentional hierarchical rnn. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4046–4056, 2018. [1](#), [2](#), [7](#)
- [5] Chiori Hori, Huda Alamri, Jue Wang, Gordon Wichern, Takaaki Hori, Anoop Cherian, Tim K Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, et al. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2352–2356. IEEE, 2019. [1](#), [7](#)
- [6] Xiyan Fu, Jun Wang, and Zhenglu Yang. Mm-avs: A full-scale dataset for multi-modal summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5922–5926, 2021. [1](#), [2](#), [4](#), [6](#), [7](#)
- [7] Jielin Qiu, Jiacheng Zhu, Mengdi Xu, Franck Dernoncourt, Trung Bui, Zhaowen Wang, Bo Li, Ding Zhao, and Hailin Jin. Mhms: Multimodal hierarchical multimedia summarization. *ArXiv*, abs/2204.03734, 2022. [1](#)
- [8] Jielin Qiu, Jiacheng Zhu, Mengdi Xu, Franck Dernoncourt, Trung Bui, Zhaowen Wang, Bo Li, Ding Zhao, and Hailin Jin. Semantics-consistent cross-domain summarization via optimal transport alignment. *ArXiv*, abs/2210.04722, 2022. [1](#)
- [9] Tsu-Jui Fu, William Yang Wang, Daniel McDuff, and Yale Song. Doc2ppt: Automatic presentation slides generation from scientific documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 634–642, 2022. [1](#)
- [10] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *European conference on computer vision*, pages 766–782. Springer, 2016. [1](#), [3](#), [6](#)
- [11] Kaiyang Zhou, Yu Qiao, and Tao Xiang. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. [1](#), [2](#), [7](#)
- [12] Mrigank Rochan, Linwei Ye, and Yang Wang. Video summarization using fully convolutional sequence networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 347–363, 2018. [1](#), [6](#)
- [13] Tsu-Jui Fu, Shao-Heng Tai, and Hwann-Tzong Chen. Attentive and adversarial learning for video summarization. In *2019 IEEE Winter Conference on applications of computer vision (WACV)*, pages 1579–1587. IEEE, 2019. [1](#)
- [14] Wencheng Zhu, Jiwen Lu, Jiahao Li, and Jie Zhou. Dsnnet: A flexible detect-to-summarize network for video summarization. *IEEE Transactions on Image Processing*, 30:948–962, 2020. [1](#), [2](#), [3](#), [6](#), [7](#), [8](#)
- [15] Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. Sumgraph: Video summarization via recursive graph modeling. In *European Conference on Computer Vision*, pages 647–663. Springer, 2020. [1](#)
- [16] Bin Zhao, Haopeng Li, Xiaoqiang Lu, and Xuelong Li. Reconstructive sequence-graph network for video summarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2793–2801, 2021. [1](#), [2](#), [7](#)
- [17] Hao Jiang and Yadong Mu. Joint video summarization and moment localization by cross-task sample transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16388–16398, 2022. [1](#), [2](#), [3](#), [6](#), [7](#)
- [18] Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*, 2016. [1](#), [2](#), [7](#)
- [19] Jianpeng Cheng and Mirella Lapata. Neural summarization by extracting sentences and words. *arXiv preprint arXiv:1603.07252*, 2016. [1](#), [2](#), [7](#)
- [20] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-first AAAI conference on artificial intelligence*, 2017. [1](#), [2](#), [7](#), [8](#)
- [21] Derek Miller. Leveraging bert for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*, 2019. [1](#), [2](#), [8](#)
- [22] Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. Extractive summarization as text matching. *arXiv preprint arXiv:2004.08795*, 2020. [1](#), [2](#)
- [23] Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing 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*, pages 1092–1102, 2017. [1](#), [2](#)
- [24] Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, Chengqing Zong, et al. Multi-modal sentence summarization with modality attention and image filtering. In *IJCAI*, pages 4152–4158, 2018. [1](#), [2](#), [6](#)
- [25] Mingzhe Li, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. Vmsmo: Learning to generate multimodal summary for video-based news articles. *arXiv preprint arXiv:2010.05406*, 2020. [1](#), [2](#), [6](#)
- [26] Xiyan Fu, Jun Wang, and Zhenglu Yang. Multi-modal summarization for video-containing documents. *arXiv preprint arXiv:2009.08018*, 2020. [1](#), [6](#)

- [27] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2
- [28] Yuki M. Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. In *Neural Information Processing Systems (NeurIPS)*, 2020. 2
- [29] Mandela Patrick, Yuki M. Asano, Polina Kuznetsova, Ruth Fong, João F. Henriques, Geoffrey Zweig, and Andrea Vedaldi. On compositions of transformations in contrastive self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021. 2
- [30] Sandra Eliza Fontes De Avila, Ana Paula Brandao Lopes, Antonio da Luz Jr, and Arnaldo de Albuquerque Araújo. Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern recognition letters*, 32(1):56–68, 2011. 2, 7
- [31] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1346–1353. IEEE, 2012. 2
- [32] Ehsan Elhamifar, Guillermo Sapiro, and S Shankar Sastry. Dissimilarity-based sparse subset selection. *IEEE transactions on pattern analysis and machine intelligence*, 38(11):2182–2197, 2015. 2
- [33] Yunjae Jung, Donghyeon Cho, Dahun Kim, Sanghyun Woo, and In So Kweon. Discriminative feature learning for unsupervised video summarization. In *Proceedings of the AAAI Conference on artificial intelligence*, pages 8537–8544, 2019. 2, 6, 7
- [34] Shiyang Lu, Zhiyong Wang, Tao Mei, Genliang Guan, and David Dagan Feng. A bag-of-importance model with locality-constrained coding based feature learning for video summarization. *IEEE Transactions on Multimedia*, 16(6):1497–1509, 2014. 2
- [35] Li Yuan, Francis EH Tay, Ping Li, Li Zhou, and Jiashi Feng. Cycle-sum: Cycle-consistent adversarial lstm networks for unsupervised video summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9143–9150, 2019. 2
- [36] Jielin Qiu, Franck Deroncourt, Trung Bui, Zhaowen Wang, Ding Zhao, and Hailin Jin. Liveseg: Unsupervised multi-modal temporal segmentation of long livestream videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5188–5198, 2023. 2
- [37] Bin Zhao and Eric P Xing. Quasi real-time summarization for consumer videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2513–2520, 2014. 2
- [38] Xufeng He, Yang Hua, Tao Song, Zongpu Zhang, Zhengui Xue, Ruhui Ma, Neil Robertson, and Haibing Guan. Un-supervised video summarization with attentive conditional generative adversarial networks. In *Proceedings of the 27th ACM International Conference on multimedia*, pages 2296–2304, 2019. 2
- [39] Yunjae Jung, Donghyeon Cho, Sanghyun Woo, and In So Kweon. Global-and-local relative position embedding for unsupervised video summarization. In *European Conference on Computer Vision*, pages 167–183. Springer, 2020. 2
- [40] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5179–5187, 2015. 2, 4, 6, 7
- [41] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *European conference on computer vision*, pages 505–520. Springer, 2014. 2, 4, 6, 7
- [42] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. *Advances in neural information processing systems*, 27, 2014. 2
- [43] Zhong Ji, Kailin Xiong, Yanwei Pang, and Xuelong Li. Video summarization with attention-based encoder–decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(6):1709–1717, 2019. 2
- [44] Aidean Sharghi, Jacob S Laurel, and Boqing Gong. Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4788–4797, 2017. 2
- [45] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7405–7414, 2018. 2, 7
- [46] Jun Wang. ESSumm: Extractive Speech Summarization from Untranscribed Meeting. In *Proc. Interspeech 2022*, pages 3243–3247, 2022. 2
- [47] Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*, 2017. 2, 6
- [48] Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*, 2017. 2
- [49] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR, 2020. 2
- [50] Yang Liu and Mirella Lapata. Text summarization with pre-trained encoders. *arXiv preprint arXiv:1908.08345*, 2019. 2
- [51] Sangwoo Cho, Franck Deroncourt, Tim Ganter, Trung Bui, Nedim Lipka, Walter Chang, Hailin Jin, Jonathan Brandt, Hassan Foroosh, and Fei Liu. StreamHover: Livestream transcript summarization and annotation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6457–6474, 2021. 2, 6
- [52] Quoc-Tuan Truong and Hady W Lauw. Vistanet: Visual aspect attention network for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 305–312, 2019. 2

- [53] Michael Gygli, Helmut Grabner, and Luc Van Gool. Video summarization by learning submodular mixtures of objectives. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3090–3098, 2015. 3, 6
- [54] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 3, 6, 7
- [55] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3, 6
- [56] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3
- [57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3, 4
- [58] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 3
- [59] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 3
- [60] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 3
- [61] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlm0: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021. 3, 4
- [62] Jun Wang, Mingfei Gao, Yuqian Hu, Ramprasaath R. Selvaraju, Chetan Ramaiah, Ran Xu, Joseph JaJa, and Larry Davis. Tag: Boosting text-vqa via text-aware visual question-answer generation. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. 3
- [63] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13041–13049, 2020. 4
- [64] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29, 2016. 4
- [65] Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9992–10002, 2020. 4
- [66] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 4
- [67] Bo He, Xitong Yang, Le Kang, Zhiyu Cheng, Xin Zhou, and Abhinav Shrivastava. Asm-loc: action-aware segment modeling for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13925–13935, 2022. 4
- [68] Bo He, Xitong Yang, Zuxuan Wu, Hao Chen, Ser-Nam Lim, and Abhinav Shrivastava. Gta: Global temporal attention for video action understanding. *British Machine Vision Conference (BMVC)*, 2021. 4
- [69] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. Category-specific video summarization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 540–555. Springer, 2014. 4
- [70] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 4, 6
- [71] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 4, 6
- [72] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 4, 5
- [73] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 4
- [74] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020. 5
- [75] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 33:21798–21809, 2020. 5
- [76] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. Rethinking the evaluation of video summaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7596–7604, 2019. 6, 7
- [77] Maurice G Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945. 6, 7

- [78] Daniel Zwillinger and Stephen Kokoska. *CRC standard probability and statistics tables and formulae*. Crc Press, 1999. 6, 7
- [79] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6
- [80] Xiuying Chen, Shen Gao, Chongyang Tao, Yan Song, Dongyan Zhao, and Rui Yan. Iterative document representation learning towards summarization with polishing. *arXiv preprint arXiv:1809.10324*, 2018. 6
- [81] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 6
- [82] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017. 7
- [83] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*, 2018. 6
- [84] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Summarizing videos with attention. In *Computer Vision—ACCV 2018 Workshops: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers 14*, pages 39–54. Springer, 2019. 7
- [85] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 6
- [86] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. 7, 8