Multimodal Frame-Scoring Transformer for Video Summarization

Anonymous submission

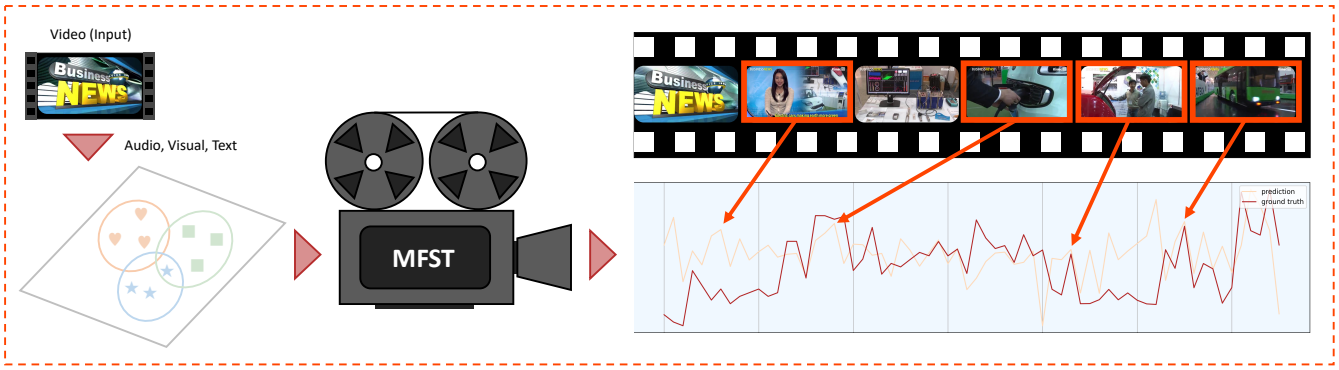

Figure 1: Video summarization aims to capture key frames using predicted importance score given a video and ground truth scores. In this paper, we propose the Multimodal Frame-Scoring Transformer (MFST), a framework to extract audio-visual-text features and to predict importance scores using multimodal representation based on extracted features. Note that unlike existing methods, MFST exploits learned multimodal feature extractors and frame-scoring transformer.

Abstract

As the number of video content has mushroomed in recent years, automatic video summarization has come useful when we want to just peek at the content of the video. However, there are two underlying limitations in generic video summarization task. First, most previous approaches read in just visual features as input, leaving other modality features behind. Second, existing datasets for generic video summarization are relatively insufficient to train a caption generator used for extracting text information from a video and to train the multimodal feature extractors. To address these two problems, this paper proposes the Multimodal Frame-Scoring Transformer (MFST), a framework exploiting visual, text, and audio features and scoring a video with respect to frames. Our MFST framework first extracts each modality features (audio-visual-text) using pretrained encoders. Then, MFST trains the multimodal frame-scoring transformer that uses multimodal representation based on extracted features as inputs and predicts frame-level scores. Our extensive experiments with previous models and ablation studies on TVSum and SumMe datasets demonstrate the effectiveness and superiority of our proposed method by a large margin in both F1 score and Rank-based evaluation.

Introduction

Our Intuition about Video Summarization. When humans watch a video on YouTube or Netflix, they perceive visual, linguistic, and audio information through various sense organs and know which parts of the video are interesting. To consider whether a scene in a movie is absorbing, for example, we observe characters' facial expressions and actions, recognize background and situation with language, and listen to the characters' utterances and sound effects. Intuitively, humans have access to well-defined scoring function in their mind, using versatile sensory systems, while video summarization models from previous studies did not.

Two Underlying Research challenges. Video summarization aims to capture key frames using predicted frame-wise importance scores, given datasets as shown in Figure 1. Despite its importance and convenience, video summarization has two inherent challenges: (i) Most previous approaches exploit just visual features, leaving other modality features behind (Zhao, Li, and Lu 2018; Rochan, Ye, and Wang 2018; Zhang, Grauman, and Sha 2018; Zhou, Qiao, and Xiang 2018; Jung et al. 2019; Rochan and Wang 2019; Park et al. 2020; Jung et al. 2020; Ghauri, Hakimov, and Ewerth 2021). (ii) Since the existing datasets (Gygli et al. 2014; Song et al. 2015), which consist only of videos and frame-level ground truths, relatively insufficient to train a caption generator used for extracting text information from the datasets and to train audio-visual-text feature extractors. Note that frame-wise human-scored video dataset is expensive to obtain compared to other common video datasets.

Language-attended methods exploit extracted text information from videos and predict importance score based on visual and text features. Bor-Chun Chen and Chen (2017) jointly combines video summarization and video caption model and trains the recurrent network in end-to-end manner. Narasimhan, Rohrbach, and Darrell (2021) leverages language-guided video summarization model given videos and corresponding user query or automatically generated video captions. Haopeng et al. (2022) collects video titles and descriptions for pre-training the language-attended self-supervised learning model.

Though language-guided approaches alleviate modality issue somewhat, there are underlying constraints in conveying vivid audio features into a video summarization model (e.g., we know there are limits to expressing a beautiful song just with the text *"beautiful song"*).

Our Solutions. This paper proposes Multimodal Frame-Scoring Transfomer (MFST) to handle different modality combinations and frame-level scoring for video summarization. Note that existing datasets for generic video summarization (Gygli et al. 2014; Song et al. 2015) are relatively insufficient to pretrain a dense caption generator and audio, visual, and text feature extractors. We investigate a new multimodal setting where it can mitigate the lack of human-scored videos used for training video summarization model.

To this end, our framework consists of three stages: (i) Generating dense video captions using learned caption generator and extracting each modality features (audio-visual-text) using feature encoders. (ii) Language-guided attention on the modality space using attention layer. (iii) Multimodal frame-scoring transformer that reads in audio-visual-text representations as inputs and predicts frame-level scores.

Our extensive experiments with previous models and ablation studies on TVSum and SumMe datasets demonstrate the effectiveness and superiority of our proposed method by a large margin in both F1 score and Rank-based evaluation.

Our Contributions. The main contributions of this work are summarized as below:

- To the best of our knowledge, Our MFST is the first to introduce frame-scoring transformer exploiting multimodal features (audio-visual-text) for generic video summarization task.
- We investigate a new multimodal setting where it can mitigate the lack of human-scored videos used for training generic video summarization model exploiting pretrained modules.
- Our empirical study on generic video summarization datasets (TVSum and SumMe) demonstrates the effectiveness and superiority of our approach.

Related Works

Generic Video Summarization. In video summarization task, there are two broad categories of methods: (i) generic video summarization (Park et al. 2020; Jung et al. 2020; Ghauri, Hakimov, and Ewerth 2021; Narasimhan, Rohrbach, and Darrell 2021; Haopeng et al. 2022) and (ii) query-guided video summarization (Narasimhan, Rohrbach, and Darrell 2021; Wu, Lin, and Silva 2022; Liu et al. 2022; Jiang and Mu 2022).

The First category of methods aim to extract representative frames from original videos using well-defined frame-wise scoring function. Existing models focus on both supervised learning and unsupervised learning. Zhou, Qiao, and Xiang (2018) designed a reward function which determines diversity and representativeness of generated summaries based on end-to-end reinforcement learning framework. Rochan, Ye, and Wang (2018) tried to solve video summarization as a sequence labeling problem based on fully convolutional sequence models. Zhang, Grauman, and Sha (2018) proposed the retrospective encoder to embed both predicted summary and original video. Zhao, Li, and Lu (2018) integrated shot-level segmentation and video summary into a hierarchical RNN. Jung et al. (2019) proposed variance loss with variational autoencoder and generative adversarial networks. Rochan and Wang (2019) learned a mapping function between raw videos and summarized videos because this kind of dataset is much easier to gain. Yuan et al. (2019) proposed a cycle-consistent adversarial networks which consist of frame selector and evaluator. Jung et al. (2020) exploited global and local input decomposition to capture the interdependencies of video frames. To represent a relation graph, Park et al. (2020) leveraged recursive graph modeling networks. Noted that, these methods just employs visual features, leaving other modality features behind. In this paper, we investigate a new multimodal setting for video summarization including audio-visual-text features.

Query-Guided Video Summarization. The second category of methods find relevant moments according to user-defined query. Unlike generic video summarization, most query-guided models take Query-Focused Video Summarization (Sharghi, Gong, and Shah 2016) dataset, UT Egocentric (Lee, Ghosh, and Grauman 2012) dataset, and QVHighlights (Lei, Berg, and Bansal 2021) dataset as input. Sharghi, Laurel, and Gong (2017) introduced a parametrized memory network into query-focuesd video summarization. Wei et al. (2018) proposed a semantic attended network which consists of frame selector and video descriptor. Kanehira et al. (2018) investigated how to divide videos into groups under the assumption that video summaries extracted from similar videos should be similar. Narasimhan, Rohrbach, and Darrell (2021) leverages a framework for handling both generic model and query-guided model given videos and corresponding user query or automatically generated video captions. To effectively address generic queries from different modalities Wu, Lin, and Silva (2022) introduced a graph convolutional networks which is used for both summary module and intent module. Liu et al. (2022) proposed the unified multimodal transformer to cover different input modality combinations. Jiang and Mu (2022) jointly leveraged video summarization model and moment localization model.

Note that though query-driven approach is necessary because defining salient scenes is often subjective, it is difficult to apply if we do not know the contents of the video or if we do not need a subjective summary (e.g., YouTube video

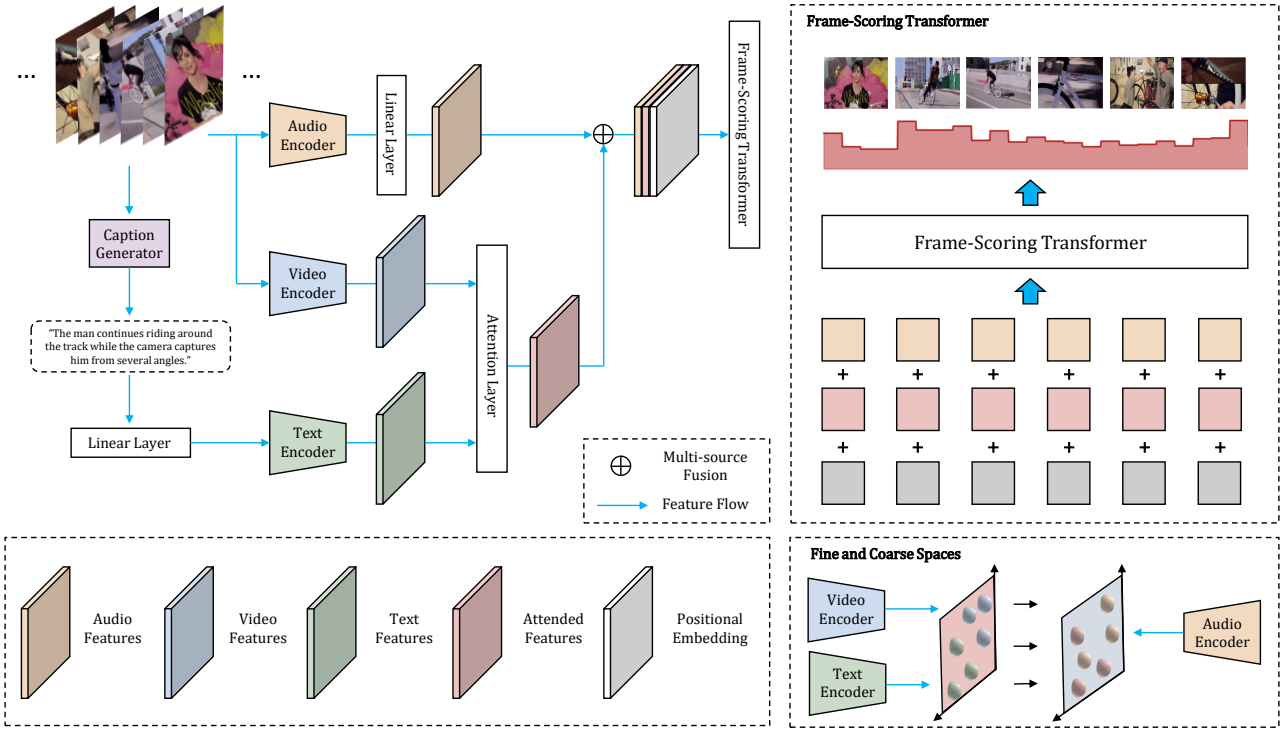

Figure 2: Schematic depiction of the proposed Multimodal Frame-Scoring Transformer (MFST) for video summarization. Given a video, we first generate dense video captions using learned caption generator. To mitigate the lack of human-scored videos, we extract each modality features (audio-visual-text) from videos and generated captions, exploiting learned feature extractors. Then, we calculate text-attended visual representation using attention layer and compute Multi-source fusion across the text-attended features and audio features. Finally we feed the fused representation to frame-scoring transformer with positional encodings at the bottom of transformer encoder and decoder stacks. We modify the transformer so that it can use modality-fused representations as input and output predicted frame-level scores.

previews). In this paper, we aim to tackle the first category based on a novel multimodal frame-scoring framework.

Frame-Scoring Transformer. Some recent works apply transformer (Vaswani et al. 2017) for generating video summaries. Narasimhan, Rohrbach, and Darrell (2021) employed transformer for predicting importance score using language-attended representation. Liu et al. (2022) leveraged modality encoder, query generator, and query decoder using transformer framework. Inspired by Narasimhan, Rohrbach, and Darrell (2021), but unlike this work, we propose frame-scoring transformer exploiting audio-visual-text modality represenatation.

Approach

Preliminaries. In this section, we present the proposed MFST framework (as shown in Figure 2) with extracted modality features. we consider the standard generic video summarization setting, where given a set of videos V and ground truth frame scores S_{gt}, the goal is to minimize the loss $\mathcal{L}_\theta$ with respect to predicted frame scores S:

$$\hat{\theta} = \arg\min_{\theta}(\mathcal{L}_\theta(S_{gt}, S)) \tag{1}$$

Note that existing datasets for generic video summarization (Gygli et al. 2014; Song et al. 2015) are relatively insufficient to train a caption generator and audio, visual, and text feature extractors. Considering that large-scale videos scored by humans are not available, we exploit a pretrained caption generator and feature extractors for each modality:

$$C = g_c(V) \tag{2}$$
$$\mathcal{T} = f_t(C) \tag{3}$$
$$\mathcal{V} = f_v(V) \tag{4}$$
$$\mathcal{A} = f_a(V) \tag{5}$$

, where C denotes a set of dense video captions, g_c is a dense video caption generator (Iashin and Rahtu 2020a), f_t is a CLIP-based text feature extractor (Xu et al. 2021), f_v is a CLIP-based visual feature extractor (Xu et al. 2021), and f_a is a learned audio feature extractor in (Baevski et al. 2020).

Fine and Coarse Spaces. To represent text-attended visual representation, we consider fine and coarse modality spaces. Inspired by (Alayrac et al. 2020), but unlike this study, we first propose to learn visual-text representation. Though video caption methods based on automatic speech recognition are useful (Hessel et al. 2019; Alayrac et al. 2020; Iashin and Rahtu 2020b), most videos in generic video summarization rarely have human dialogues. In this paper, we employ the feature-based video cation model g_c (Iashin and Rahtu 2020a) to extract dense video captions C =

$\{C_1, C_2, ..., C_N\}$. Then, we leverage a fine-grained embedding space where visual features $\mathcal{V} = \{\mathcal{V}_1, \mathcal{V}_2, ..., \mathcal{V}_N\}$ and text features $\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_N\}$ lie. Note that (Liang et al. 2022) demonstrates "multimodal video-text pretraining" paradigm can not solve the *modality gap* phenomenon completely which causes performance degradation. In this work, we compute text-attended visual representation using attention layer:

$$h_{\mathcal{VT}} = \text{Attention}(h_{\mathcal{V}}, h_{\mathcal{T}}) \tag{6}$$

, where h stands for feature embedding space. Intuitively, in the fine-grained embedding space, each $v_{i,l}$ from $\mathcal{V}_i$ chooses the most relevant caption t_i^* in the $\mathcal{T}_i$ using attention mechanism. Lastly, we project the fine-grained embedding space into the coarse-grained embedding space by modality fusion:

$$h_{\mathcal{AVT}} = \mathcal{F}_M(h_{\mathcal{VT}}, h_{\mathcal{A}}) \tag{7}$$

, where $\mathcal{F}_M$ denotes function of fusion and $h_{\mathcal{AVT}}$ represents coarse-grained feature representation. The complete process is summarized in Algorithm 1.

Multimodal Frame-Scoring Transformer. Frame-scoring transformer (Narasimhan, Rohrbach, and Darrell 2021) takes feature representation as input and predict importance scores. Note that previous approach exploits just text-visual representation, leaving audio features behind. MFST introduces frame-scoring transformer to video summarization, which is modified to predict frame-level importance scores S, based on coarse-grained feature representation $h_{\mathcal{AVT}}$:

$$\text{Multimodal - Attn.}(h_{\mathcal{AVT}}) = \text{Concat}(\text{h}_1, ..., \text{h}_\text{h})W^{O_{\mathcal{AVT}}} \tag{8}$$

$$\text{h}_\text{i} = \text{Attn.}(Q_{\mathcal{AVT}}W_i^{Q_{\mathcal{AVT}}}, K_{\mathcal{AVT}}W_i^{K_{\mathcal{AVT}}}, V_{\mathcal{AVT}}W_i^{V_{\mathcal{AVT}}}) \tag{9}$$

$$\text{Attn.}(h_{\mathcal{AVT}}) = \text{softmax}(\frac{Q_{\mathcal{AVT}}K_{\mathcal{AVT}}^T}{\sqrt{d_k}})V_{\mathcal{AVT}} \tag{10}$$

, where $W^{Q_{\mathcal{VTA}}}$, $W^{K_{\mathcal{VTA}}}$, and $W^{V_{\mathcal{VTA}}}$ denote parameter matrices and d_k is the dimension of $K_{\mathcal{AVT}}$.

Finally, we feed $h_{\mathcal{AVT}}$ to frame-scoring transformer with positional encodings at the bottom of transformer encoder and decoder stacks. Given ground truth frame scores S_{gt} of N frames from a video, we train MFST using the mean square error:

$$\mathcal{L}_\theta(S_{gt}, S) = \frac{1}{N}\left\|S_{gt} - S\right\|_2^2 \tag{11}$$

The complete process is summarized in Algorithm 2.

Algorithm 1: Fine-to-Coarse Space Projection

Require: $V = \{V_1, V_2, ..., V_N\}$
for all V_i **do**
  Generate dense video captions $C_i = g_c(V_i)$
  , where $C_i = \{c_1^i, c_2^i, ..., c_M^i\}$
end for
for all V_i and C_i **do**
  $\mathcal{T}_i = f_t(C_i)$, $\mathcal{V}_i = f_v(V_i)$, and $\mathcal{A}_i = f_a(V_i)$
  , where $\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_N\}$, $\mathcal{V} = \{\mathcal{V}_1, \mathcal{V}_2, ..., \mathcal{V}_N\}$, and $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_N\}$
end for
$h_{\mathcal{VT}}$ = Concat($h_{\mathcal{V}_1\mathcal{T}_1}, h_{\mathcal{V}_2\mathcal{T}_2}, ..., h_{\mathcal{V}_N\mathcal{T}_N}$)
for all $\mathcal{T}_i$, $\mathcal{V}_i$ and $\mathcal{A}_i$ **do**
  $h_{\mathcal{T}_i} := \mathcal{T}_i$, $h_{\mathcal{V}_i} := \mathcal{V}_i$, $h_{\mathcal{A}_i} := \mathcal{A}_i$,
  Calculate fine-grained modality space $h_{\mathcal{V}_i\mathcal{T}_i}$
  $h_{\mathcal{V}_i\mathcal{T}_i}$ = Attention($h_{\mathcal{V}_i}, h_{\mathcal{T}_i}$)
  Project fine-to-coarse embedding space $h_{\mathcal{A}_i\mathcal{V}_i\mathcal{T}_i}$
  $h_{\mathcal{A}_i\mathcal{V}_i\mathcal{T}_i} = \mathcal{F}_M(h_{\mathcal{V}_i\mathcal{T}_i}, h_{\mathcal{A}_i})$
end for
$h_{\mathcal{AVT}}$ = Concat($h_{\mathcal{A}_1\mathcal{V}_1\mathcal{T}_1}, h_{\mathcal{A}_2\mathcal{V}_2\mathcal{T}_2}, ..., h_{\mathcal{A}_N\mathcal{V}_N\mathcal{T}_N}$)

Algorithm 2: Multimodal Frame-Socring Transformer

Require: $h_{\mathcal{AVT}} = \{h_{\mathcal{A}_1\mathcal{V}_1\mathcal{T}_1}, h_{\mathcal{A}_2\mathcal{V}_2\mathcal{T}_2}, ..., h_{\mathcal{A}_N\mathcal{V}_N\mathcal{T}_N}\}$
for each epoch **do**
  for i in range(h) **do**
    Calculate self-attention for each i
    Attn.($h_{\mathcal{AVT}}$) = softmax($Q_{\mathcal{AVT}}K_{\mathcal{AVT}}^T/\sqrt{d_k}$)$V_{\mathcal{AVT}}$
    Calculate single head attention h_i
    h_i = Attn.($Q_{\mathcal{A}_i\mathcal{V}_i\mathcal{T}_i}, K_{\mathcal{A}_i\mathcal{V}_i\mathcal{T}_i}, V_{\mathcal{A}_i\mathcal{V}_i\mathcal{T}_i}$)
  end for
  Multimodal-Attn. = Concat($\text{h}_1, ..., \text{h}_h$)
  S = Linear(Multimodal-Attn.)
  Compute loss between ground truth scores S_{gt} and predicted scores S
  $\mathcal{L}_\theta(S_{gt}, S) = \frac{1}{N}\left\|S_{gt} - S\right\|_2^2$
end for

Experiments

Dataset Description

We conduct our video summarization experiments on two benchmarks: **TVSum** (Song et al. 2015) and **SumMe** (Gygli et al. 2014).

- TVSum (Song et al. 2015) contains 50 videos, including the topics of news, documentaries. The duration of each video varies from 1 to 10 minutes. 20 annotators provide frame-level importance scores for each video.
- SumMe (Gygli et al. 2014) consists of 25 user videos, covering various topics (e.g., holidays and sports). Each video ranges from 1 to 6 minutes. 15 to 18 persons annotated multiple ground truth summaries for each video.

Metric Description

We follow the same experimental metrics used in existing works: F-score and Rank-based evaluation. *True positive* means highlight overlaps between model-generated summary V_m and human-generated summary V_h based on importance scores. The precision and recall are calculated as follows:

$$\text{Precision} = \frac{\mid V_m \cap V_h \mid}{\mid V_m \mid}, \text{Recall} = \frac{\mid V_m \cap V_h \mid}{\mid V_h \mid}. \tag{12}$$

Table 1: Experimental results on SumMe under the Canonical, Augment, and Transfer settings (F-score).

Methods	SumMe		
	Can	*Aug*	*Tran*
vsLSTM (Zhang et al. 2016)	0.376	0.416	0.407
SGAN (Mahasseni, Lam, and Todorovic 2017)	0.387	0.417	—
SGAN$_s$ (Mahasseni, Lam, and Todorovic 2017)	0.417	0.436	—
H-RNN (Zhao, Li, and Lu 2017)	0.421	0.438	—
DRDSN (Zhou, Qiao, and Xiang 2018)	0.421	0.439	0.426
HSA-RNN (Zhao, Li, and Lu 2018)	0.423	0.421	—
ACGAN (He et al. 2019)	0.460	0.470	0.445
WS-HRL (Chen et al. 2019)	0.436	0.445	—
re-S2S (Zhang, Grauman, and Sha 2018)	0.425	0.449	—
S-FCN (Rochan, Ye, and Wang 2018)	0.475	0.511	0.441
VASNet (Fajtl et al. 2018)	0.497	0.510	—
CSNet$_s$ (Jung et al. 2019)	0.513	0.521	0.451
GLRPE (Jung et al. 2020)	0.502	—	—
SumGraph (Park et al. 2020)	0.514	0.529	0.487
RSGN (Zhao et al. 2021)	0.450	0.457	0.440
RSGN$_{uns}$ (Zhao et al. 2021)	0.423	0.436	0.412
MSVA (Ghauri, Hakimov, and Ewerth 2021)	0.545	—	—
CLIP-It (Narasimhan, Rohrbach, and Darrell 2021)	0.542	0.564	0.519
SSPVS (Haopeng et al. 2022)	0.501	—	—
iPTNet (Jiang and Mu 2022)	0.545	0.569	0.492
MFST (Ours)	**0.595**	**0.655**	**0.576**

Table 2: Experimental results on SumMe (Kendall's τ and Spearman's ρ).

Methods	SumMe	
	τ	ρ
Random	0.000	0.000
Human	0.205	0.213
Ground Truth	1.000	1.000
SGAN (Mahasseni, Lam, and Todorovic 2017)	—	—
WS-HRL (Chen et al. 2019)	—	—
DRDSN (Zhou, Qiao, and Xiang 2018)	0.047	0.048
dppLSTM (Zhang et al. 2016)	—	—
CSNet$_s$ (Jung et al. 2019)	—	—
GLRPE (Jung et al. 2020)	—	—
HSA-RNN (Zhao, Li, and Lu 2018)	0.064	0.066
RSGN (Zhao et al. 2021)	0.083	0.085
RSGN$_u$ (Zhao et al. 2021)	0.071	0.073
SumGraph (Park et al. 2020)	—	—
SSPVS (Haopeng et al. 2022)	0.123	0.170
MSVA (Ghauri, Hakimov, and Ewerth 2021)	0.200	**0.230**
MFST (Ours)	**0.229**	0.229

Rank-based evaluations (Otani et al. 2019) compute Kendall's τ and Spearman's ρ which measure non-parametric rank correlations:

$$\tau = Kendall(S_{gt}, S) \tag{13}$$
$$\rho = Spearman(S_{gt}, S) \tag{14}$$

Performance Comparison

Experimental Settings. We compare MFST with existing models in three different experimental settings:

- In Canonical setting, we selects the dataset (e.g., TVSum or SumMe) and randomly splits the dataset into training and evaluation.
- In Augment setting, we merges the two datasets into one and randomly splits the dataset into training and evaluation.
- In Transfer setting, we trains a model using one dataset and evaluates the trained model on the other dataset.

As we follow the experimental protocol in proposed by existing studies, in all experimental settings, we conduct experiments over 5 splits and average the results. Each experiment randomly selects 20% of the dataset for evaluation.

Implementation Details. We employ VideoCLIP (Xu et al. 2021) for extracting both visual features and text features. To exploit audio features, we leverage feature extractor using Wav2Vec2 (Baevski et al. 2020). We train our model on 8 NVIDIA GeForce TITAN GPUs for 20 epochs. The batch size are selected based on available GPU memory. We use Adam optimizer with learning rate 1e-4 and weight decay of 1e-3.

Results on Video Summarization. Recent studies of generic video summarization have struggled with two fundamental research questions: *(1) How to learn diverse modality features to predict importance score better?* and *(2) How to alleviate data sparsity problem?* Note that these two questions are not separate, rather highly correlated.

We conduct experiments to answer the two questions. Table 1 and Table 2 show our extensive experiments with previous methods on SumMe dataset. As shown in Table 1, under the canonical, augment, and transfer settings, we demonstrate that MFST outperforms existing methods on the benchmark. Table 2 also shows that MFST outperforms by a large margin in Rank-based evaluation except MSVA. However, MSVA did not conduct experiments in Augment setting and Transfer setting, it is hard to say MSVA is compatible with our model.

It should be noted that the performance of models using two or more modalities is higher than models using unimodality and the results of MFST, which properly exploits the visual, text and audio modalities, demonstrate the effectiveness and superiority of our approach.

Table 3 and Table 4 also demonstrates our comprehensive experiments with existing models on TVSum dataset. As shown in Table 3, our model achieves state-of-the-art performance by a large margin in F1 score. Unlike Table 2, Our model outperforms all existing methods in Table 4. Note that experimental results demonstrate that methods which employ multimodal features and mitigate data sparsity problem get high a score than other methods.

Ablation Studies We further conduct ablation studies on SumMe and TVSum datsets. As shown in Table 5, to validate the effectiveness of our audio-attended representation, we compare two results, the one without audio features and the other with audio features. Results in Table 5 demonstrate our approach can be advanced when extracted audio features are added. Interestingly, our framework without audio features also outperforms existing methods on both SumMe and TVSum. We assume that MFST exploiting pretrained model can represent modality features in a embedding space leading higher prediction performance than other models.

Qualitative Results. Figure 3 shows visualized qualitative results on generated summary from MFST with human-generated summary. Note that model-generated score graph and ground truth graph are very similar, which means maximum points and minimum points in both graphs are fairly

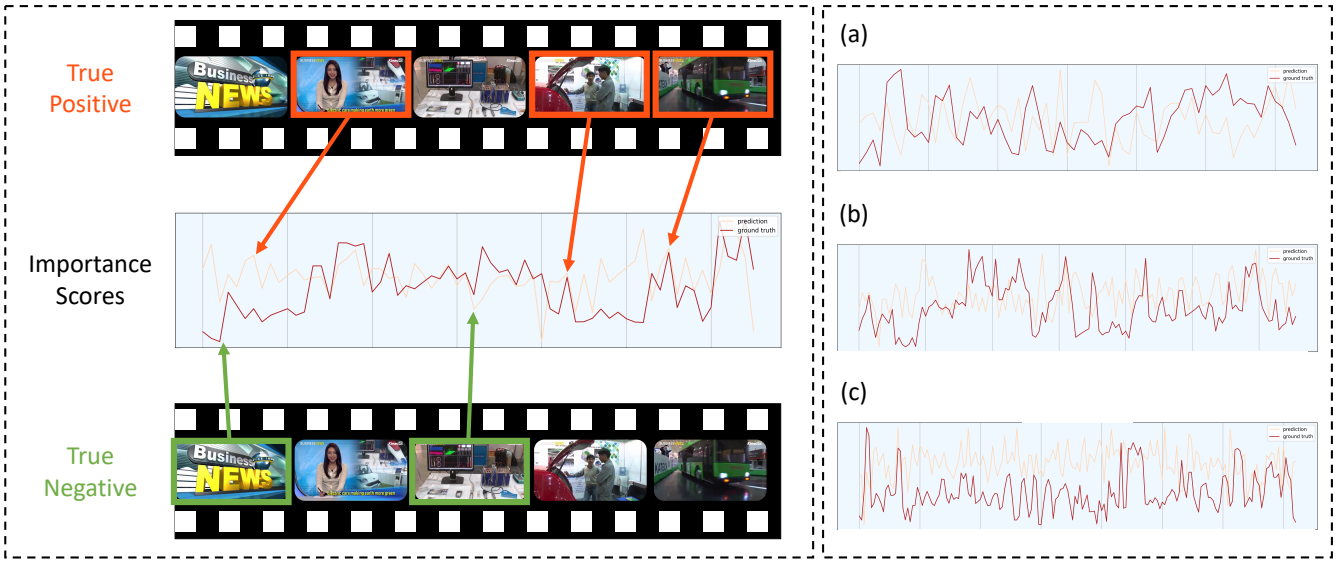

Figure 3: Comparison of predicted score and averaged ground truth scores from TVSum dataset. Note that model-predicted score graph and ground truth graph are fairly similar in that maxima and minima.

Table 3: Experimental results on TVSum under the Canonical, Augment, and Transfer settings (F-score).

Methods	TVSum		
	Can	*Aug*	*Tran*
vsLSTM (Zhang et al. 2016)	0.542	0.579	0.569
SGAN (Mahasseni, Lam, and Todorovic 2017)	0.508	0.589	—
SGAN$_s$ (Mahasseni, Lam, and Todorovic 2017)	0.563	0.612	—
H-RNN (Zhao, Li, and Lu 2017)	0.579	0.619	—
DRDSN (Zhou, Qiao, and Xiang 2018)	0.581	0.598	0.589
HSA-RNN (Zhao, Li, and Lu 2018)	0.587	0.598	—
ACGAN (He et al. 2019)	0.585	0.589	0.578
WS-HRL (Chen et al. 2019)	0.584	0.585	—
re-S2S (Zhang, Grauman, and Sha 2018)	0.603	0.639	—
S-FCN (Rochan, Ye, and Wang 2018)	0.568	0.592	0.582
VASNet (Fajtl et al. 2018)	0.614	0.623	—
CSNet$_s$ (Jung et al. 2019)	0.588	0.590	0.592
GLRPE (Jung et al. 2020)	0.591	—	—
SumGraph (Park et al. 2020)	0.639	0.658	0.605
RSGN (Zhao et al. 2021)	0.601	0.611	0.600
RSGN$_{uns}$ (Zhao et al. 2021)	0.580	0.591	0.597
MSVA (Ghauri, Hakimov, and Ewerth 2021)	0.628	—	—
CLIP-It (Narasimhan, Rohrbach, and Darrell 2021)	0.663	0.690	0.655
SSPVS (Haopeng et al. 2022)	0.607	—	—
iPTNet (Jiang and Mu 2022)	0.634	0.642	0.598
MFST (Ours)	**0.737**	**0.779**	**0.691**

Table 4: Experimental results on TVSum (Kendall's τ and Spearman's ρ).

Methods	TVSum	
	τ	ρ
Random	0.000	0.000
Human	0.177	0.204
Ground Truth	0.364	0.456
SGAN (Mahasseni, Lam, and Todorovic 2017)	0.024	0.032
WS-HRL (Chen et al. 2019)	0.078	0.116
DRDSN (Zhou, Qiao, and Xiang 2018)	0.020	0.026
dppLSTM (Zhang et al. 2016)	0.042	0.055
CSNet$_s$ (Jung et al. 2019)	0.025	0.034
GLRPE (Jung et al. 2020)	0.070	0.091
HSA-RNN (Zhao, Li, and Lu 2018)	0.082	0.088
RSGN (Zhao et al. 2021)	0.083	0.090
RSGN$_u$ (Zhao et al. 2021)	0.048	0.052
SumGraph (Park et al. 2020)	0.094	0.138
SSPVS (Haopeng et al. 2022)	0.169	**0.231**
MSVA (Ghauri, Hakimov, and Ewerth 2021)	0.190	0.210
MFST (Ours)	**0.222**	0.224

Table 5: The results of ablation studies in multimodality.

Methods	SumMe			TVSum		
	Can	*Aug*	*Tran*	*Can*	*Aug*	*Tran*
MFST	0.542	0.629	0.553	0.708	0.753	0.659
MFST + Audio (Ours)	**0.595**	**0.655**	**0.576**	**0.737**	**0.779**	**0.691**

overlapped. Results in Figure 3 represents our model predicts parts that humans find interesting or not.

Conclusion

We pose two underlying research questions for generic video summarization: (1) How to learn diverse modality features to predict importance score better? and (2) How to alleviate data sparsity problem? In this paper, we propose MFST, a simple and effective frame-scoring framework given videos. Unlike existing methods, MFST exploits audio-visual-text features using learned feature extractors and frame-scoring multimodal transformer. We first generate text-attended representation on fine-grained embedding space using attention mechanism. Then, we project fine-grained space to coarse-grained space based on modality fusion. Though our comprehensive comparisons with previous approaches and ablation studies demonstrate the effectiveness and superiority of our proposed method, lack of deep insights about modality representations and utilizing large-scale models are biggest limitations of our work. For example, we could not discover precise reasons why our framework is superior with respect to modality representations. We leave them as future works.

References

Alayrac, J.-B.; Recasens, A.; Schneider, R.; Arandjelović, R.; Ramapuram, J.; Fauw, J. D.; Smaira, L.; Dieleman, S.; and Zisserman, A. 2020. Self-Supervised Multimodal Versatile Networks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 12449–12460. Curran Associates, Inc.

Bor-Chun Chen, Y.-Y. C.; and Chen, F. 2017. Video to Text Summary: Joint Video Summarization and Captioning with Recurrent Neural Networks. In Tae-Kyun Kim, G. B., Stefanos Zafeiriou; and Mikolajczyk, K., eds., *Proceedings of the British Machine Vision Conference (BMVC)*, 118.1–118.14. BMVA Press. ISBN 1-901725-60-X.

Chen, Y.; Tao, L.; Wang, X.; and Yamasaki, T. 2019. Weakly supervised video summarization by hierarchical reinforcement learning. In *Proceedings of the ACM Multimedia Asia*, 1–6.

Fajtl, J.; Sokeh, H. S.; Argyriou, V.; Monekosso, D.; and Remagnino, P. 2018. Summarizing videos with attention. In *Asian Conference on Computer Vision*, 39–54. Springer.

Ghauri, J. A.; Hakimov, S.; and Ewerth, R. 2021. SUPERVISED VIDEO SUMMARIZATION VIA MULTIPLE FEATURE SETS WITH PARALLEL ATTENTION.

Gygli, M.; Grabner, H.; Riemenschneider, H.; and Van Gool, L. 2014. Creating Summaries from User Videos. In Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision – ECCV 2014*, 505–520. Cham: Springer International Publishing. ISBN 978-3-319-10584-0.

Haopeng, L.; Qiuhong, K.; Mingming, G.; and Rui, Z. 2022. Video Summarization Based on Video-text Modelling.

He, X.; Hua, Y.; Song, T.; Zhang, Z.; Xue, Z.; Ma, R.; Robertson, N.; and Guan, H. 2019. Unsupervised video summarization with attentive conditional generative adversarial networks. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2296–2304.

Hessel, J.; Pang, B.; Zhu, Z.; and Soricut, R. 2019. A Case Study on Combining ASR and Visual Features for Generating Instructional Video Captions. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 419–429. Hong Kong, China: Association for Computational Linguistics.

Iashin, V.; and Rahtu, E. 2020a. A Better Use of Audio-Visual Cues: Dense Video Captioning with Bi-modal Transformer. In *British Machine Vision Conference (BMVC)*.

Iashin, V.; and Rahtu, E. 2020b. Multi-Modal Dense Video Captioning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 958–959.

Jiang, H.; and Mu, Y. 2022. Joint Video Summarization and Moment Localization by Cross-Task Sample Transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16388–16398.

Jung, Y.; Cho, D.; Kim, D.; Woo, S.; and Kweon, I. S. 2019. Discriminative Feature Learning for Unsupervised Video Summarization. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, 8537–8544. AAAI Press.

Jung, Y.; Cho, D.; Woo, S.; and Kweon, I. S. 2020. Global-and-Local Relative Position Embedding for Unsupervised Video Summarization. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV*, 167–183. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-030-58594-5.

Kanehira, A.; Gool, L. V.; Ushiku, Y.; and Harada, T. 2018. Viewpoint-Aware Video Summarization. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7435–7444.

Lee, Y. J.; Ghosh, J.; and Grauman, K. 2012. Discovering important people and objects for egocentric video summarization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 1346–1353.

Lei, J.; Berg, T. L.; and Bansal, M. 2021. Detecting Moments and Highlights in Videos via Natural Language Queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858.

Liang, W.; Zhang, Y.; Kwon, Y.; Yeung, S.; and Zou, J. 2022. Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. *arXiv preprint arXiv:2203.02053*.

Liu, Y.; Li, S.; Wu, Y.; Chen, C. W.; Shan, Y.; and Qie, X. 2022. UMT: Unified Multi-modal Transformers for Joint Video Moment Retrieval and Highlight Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mahasseni, B.; Lam, M.; and Todorovic, S. 2017. Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 202–211.

Narasimhan, M.; Rohrbach, A.; and Darrell, T. 2021. CLIP-It! Language-Guided Video Summarization. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 13988–14000. Curran Associates, Inc.

Otani, M.; Nakashima, Y.; Rahtu, E.; and Heikkila, J. 2019. Rethinking the evaluation of video summaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7596–7604.

Park, J.; Lee, J.; Kim, I.-J.; and Sohn, K. 2020. SumGraph: Video Summarization via Recursive Graph Modeling. In *ECCV*.

Rochan, M.; and Wang, Y. 2019. Video Summarization by Learning From Unpaired Data. In *2019 IEEE/CVF Confer-*

ence on Computer Vision and Pattern Recognition (CVPR), 7894–7903.

Rochan, M.; Ye, L.; and Wang, Y. 2018. Video Summarization Using Fully Convolutional Sequence Networks. In Ferrari, V.; Hebert, M.; Sminchisescu, C.; and Weiss, Y., eds., *Computer Vision – ECCV 2018*, 358–374. Cham: Springer International Publishing. ISBN 978-3-030-01258-8.

Sharghi, A.; Gong, B.; and Shah, M. 2016. Query-Focused Extractive Video Summarization. In *ECCV*.

Sharghi, A.; Laurel, J. S.; and Gong, B. 2017. Query-Focused Video Summarization: Dataset, Evaluation, and a Memory Network Based Approach. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2127–2136.

Song, Y.; Vallmitjana, J.; Stent, A.; and Jaimes, A. 2015. TVSum: Summarizing web videos using titles. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5179–5187.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wei, H.; Ni, B.; Yan, Y.; Yu, H.; and Yang, X. 2018. Video Summarization via Semantic Attended Networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press. ISBN 978-1-57735-800-8.

Wu, G.; Lin, J.; and Silva, C. T. 2022. IntentVizor: Towards Generic Query Guided Interactive Video Summarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10503–10512.

Xu, H.; Ghosh, G.; Huang, P.-Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; and Feichtenhofer, C. 2021. VideoCLIP: Contrastive Pre-training for
Zero-shot Video-Text Understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics.

Yuan, L.; Tay, F. E. H.; Li, P.; Zhou, L.; and Feng, J. 2019. Cycle-SUM: Cycle-consistent Adversarial LSTM Networks for Unsupervised Video Summarization. In *AAAI*.

Zhang, K.; Chao, W.-L.; Sha, F.; and Grauman, K. 2016. Video summarization with long short-term memory. In *European conference on computer vision*, 766–782. Springer.

Zhang, K.; Grauman, K.; and Sha, F. 2018. Retrospective Encoders for Video Summarization. In Ferrari, V.; Hebert, M.; Sminchisescu, C.; and Weiss, Y., eds., *Computer Vision – ECCV 2018*, 391–408. Cham: Springer International Publishing. ISBN 978-3-030-01237-3.

Zhao, B.; Li, H.; Lu, X.; and Li, X. 2021. Reconstructive Sequence-Graph Network for Video Summarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhao, B.; Li, X.; and Lu, X. 2017. Hierarchical recurrent neural network for video summarization. In *Proceedings of the 25th ACM international conference on Multimedia*, 863–871.

Zhao, B.; Li, X.; and Lu, X. 2018. HSA-RNN: Hierarchical Structure-Adaptive RNN for Video Summarization. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7405–7414.

Zhou, K.; Qiao, Y.; and Xiang, T. 2018. Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press. ISBN 978-1-57735-800-8.