A TWO-STAGE FRAMEWORK TO GENERATE VIDEO CHAPTER

Anonymous authors
Paper under double-blind review

ABSTRACT

We aim to address the problem of video chapter generation. Compared to traditional video activity analysis, this task is significantly different. The videos in chapter generation are much longer and contain many complex temporal structures. Moreover, the association between video frames and narrations plays a crucial role in expressing underlying information. To facilitate the research along this direction, we introduce a large-scale dataset called ChapterGen, which consists of approximately $10k$ user-generated videos with annotated chapter descriptions. Our data collection procedure is fast, scalable and does not require any additional manual annotation. On top of this dataset, we propose a two-stage framework to perform chapter localization and chapter title generation. This framework captures two aspects of a video, including visual dynamics and narration text. To parse the whole video efficiently, we build the framework based on a flexible clip sliding window. Our experiments demonstrate that the proposed framework achieves superior results over existing methods on both accuracy and efficiency.

1 INTRODUCTION

Video chapter generation aims to understand the whole of video and summarize its content to a couple of chapters, just like chapters in a book. This task is an important step towards building high-level semantic understanding of videos. It also offers many valuable industrial applications such as video preview, quick searching, content-driven video retrieval and video editing.

Over the past decade, extensive studies have been devoted to video understanding, e.g. action recognition and localization Wang et al. (2016); Carreira & Zisserman (2017); Lin et al. (2018; 2019a), event detection Sultani et al. (2018) and video summarization Song et al. (2015); Gygli et al. (2014). However, methods developed for these tasks are not applicable for video chapter generation. Specifically, the videos in these tasks consist of simple scenes and its durations are usually short. Visual features are sufficient for achieving good results. In contrast, the videos for chapter generation are much longer and contain many complex temporal structures. Recently, several datasets Huang et al. (2020b) and methods Rao et al. (2020); Chen et al. (2021) are developed on movies, which consist of many shots and scenes. However, movies made by various narrative techniques usually lack a clear temporal structure for chapter generation. Moreover, the movie scene analysis methods Rao et al. (2020); Chen et al. (2021) which developed their approaches on a couple of keyframe sequences may not suitable for video chapter generation.

To facilitate the research in video chapter generation, we need a new large and diverse dataset. More importantly, it should allow high-level semantics and clear temporal structures to be extracted and analyzed. Instead of manually annotation, in this work, we explore a different source of supervision to obtain the videos with labeled chapter descriptions. We observe that plenty of user videos are available on YouTube. And some are with uploaded video chapter descriptions which contain chapter beginning timestamps and chapter titles. To leverage this rich source of data, we construct a new large-scale dataset containing 9631 videos that depict various activities and topics including entertainment, education, business, etc.

On top of this new dataset, we develop a new framework to generate video chapters. Rather than processing extracted keyframes or trimmed fragments, we consider all video frames, which contain complete video context for chapter generation. Specifically, our framework contains two stages: (1) **Chapter localization.** The chapter localization is helpful for organizing relevant video content

in the same segmentation. Specifically, given an input video, we localize chapter beginning times. It calculates the temporal boundary for each potential chapter. Intuitively, the chapter temporal boundary highly depends on the semantic context. We develop a two-stream model which captures visual and narration text clues simultaneously. To address this problem efficiently, we propose a clip sliding window mechanism with skip sliding step. This mechanism is more efficient than the traditional one-step sliding window without sacrificing localization accuracy. (2) **Chapter title generation.** After chapter localization, we get the temporal boundary for each chapter. To generate a concise and conclusive chapter descriptive title, we need to summarize the video content within each boundary. We observed that the narration text, which explains what happens in videos, contains rich information for generating chapter titles. So we formulate the chapter title generation as a text summarization problem. However, the extractive summarization Narayan et al. (2018); Zhong et al. (2020), which merely copies informative fragments from the original text, is not suitable for title generation. Because titles are highly conclusive and probably contain novel words which do not appear in the original text. Inspired by abstractive summarization, we apply an encoder-decoder Transformer architecture Vaswani et al. (2017) to generate the target title.

In summary, our contributions lie in three aspects:

- We construct a large dataset on 9631 videos, which contain annotated associations between chapter beginning timestamps and chapter titles. The data collection is fast, scalable and does not require expensive manual annotation. This dataset can be used as a standard benchmark for video chapter study, which we believe is a promising direction towards high-level video understanding.
- We develop a two-stage framework to address the video chapter generation problem. Specifically, a two-stream model with a novel sliding window mechanism is developed for chapter localization. An encoder-decoder Transformer is applied for chapter title generation.
- We conduct experiments to show that our proposed framework can achieve better results than existing competitors on both accuracy and efficiency. We wish that this work can motivate future study to video chapter generation and other high-level video understanding tasks.

2 Related Work

We collect a new dataset for video chapter generation and propose a two-stage framework which consists of chapter localization and chapter title generation. We briefly review the most relevant works below.

Video understanding dataset. In recent years, with the increasing interest in video understanding, many video understanding tasks and datasets have been proposed. The low-level video understanding datasets like Youtube-8M Abu-El-Haija et al. (2016) for video classification, Activitynet Caba Heilbron et al. (2015), Kinetics-400 Kay et al. (2017) for action recognition. These datasets have plentiful short or trimmed videos consisting of several action sequences. For high-level video understanding tasks, TVSum Song et al. (2015) contains 50 video sequences for video summarization, MPII-MD Rohrbach et al. (2015) has 94 videos for video captioning. The recent MovieNet dataset Huang et al. (2020a) provides 1100 movies, partially annotated with story scene boundaries, cinematic styles and movie synopses. Our collected dataset differs from the existing datasets in two aspects: (1) Our data collection is fast and scalable without expensive manual annotation. Therefore, we can easily expand the dataset in terms of data size. (2) The collected videos cover diverse topics and activities. The videos are created by different authors with different styles. Hence, the data have more complex spatial-temporal structures and they are more practical in industrial applications.

Video temporal localization. Video temporal localization is the problem of identifying some target locations in videos with its beginning and end times. This problem covers a wide range of tasks such as video action temporal localization, video shot detection and movie scene segmentation. Video action temporal localization Shou et al. (2016); Lin et al. (2018; 2019b) is to detect some specific actions start and end from the dense frame sequences. Video shot detection is to find shots based on some low-level features, including color histogram Rui et al. (1998), camera motion parameters Rasheed & Shah (2003) and audio clues Sidiropoulos et al. (2011). All these problems are low-

level video understanding. The recent movie scene segmentation Rao et al. (2020); Chen et al. (2021) is to cluster video segments based on visual features and high-level story semantics. In spite of the effectiveness, these methods are developed on a shot detection preprocessing and make an assumption that the movie scenes are composed of several consecutive shots. This assumption could be not applicable in video chapter generation problem (see experiments in Section 5.2). As a comparison, our two-stream model is more general in handling the video chapter localization problem.

Text summarization. Given a long text, text summarization is to understand the meaning of the provided long text and output some concise and conclusive sentences. The text summarization approaches can be categorized as extractive summarization Narayan et al. (2018); Zhong et al. (2020) and abstrastive summarization Song et al. (2019); Zhang et al. (2020). Extractive summarization which directly copies informative fragments from the input, while abstractive summarization attempt to generate new words or sentences. In video chapter generation, chapter titles are very short (i.e. smaller than 10 words). And the words in a title may not appear in the given narration text. Therefore, the flexible abstractive summarization is more proper for our problem. In inspired by seq2seq translation Sutskever et al. (2014), we apply the state-of-the-art Transformer encoder-decoder architecture Vaswani et al. (2017) to address the chapter title generation problem.

3 CHAPTERGEN DATASET

We construct a new dataset ChapterGen on 9631 videos which consist of various topics like entertainment, education, business, daily life, etc. The comparison between our dataset and existing high-level video understanding datasets is summarized in Table 1.

Dataset	Video num	Avg. length (min)	Task
TVSum Song et al. (2015)	50	4.2	Video summarization
MPII-MD Rohrbach et al. (2015)	94	26.1	Video captioning
MovieNet Huang et al. (2020b)	1100	118.8	Movie analysis
MovieScenes Rao et al. (2020)	150	118.8	Movie scene detection
ChapterGen	9631	13.6	Video chapter generation

Table 1: The comparison between ChapterGen and existing datasets

Each video in ChapterGen is associated with uploaded chapter descriptions. The chapter descriptions provide a list of chapter beginning timestamps and chapter titles for introducing the main content of video parts. We use the chapter descriptions as the training ground-truth for video chapter generation.

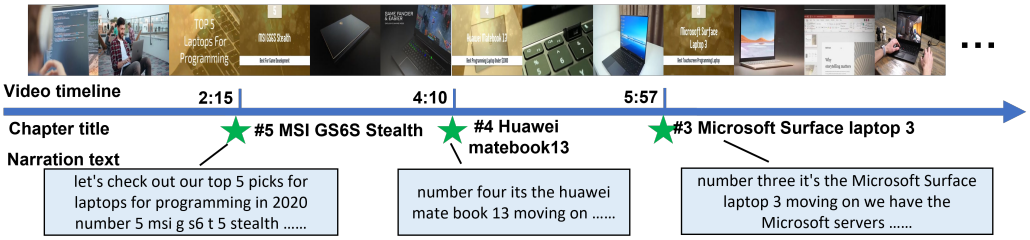
Except for video frames, we notice that the narration text plays an important role in understanding video content. To leverage narration information, we choose videos with English subtitles which are automatically generated by YouTube ASR. The videos that have less than 0.5 word/second in average are ignored and discarded.

A video contains a lot of duplicated frames. We downsample them by 1 frame/second but keep all narrations. In this way, the model can quickly parse the whole video without loss of chapter generation accuracy. We also intentionally use videos whose duration is less than 30 minutes, so that we can train our model more efficiently. Note that our approach is not limited by video duration. It is able to directly handle longer videos (e.g. a couple of hours) without any modification.

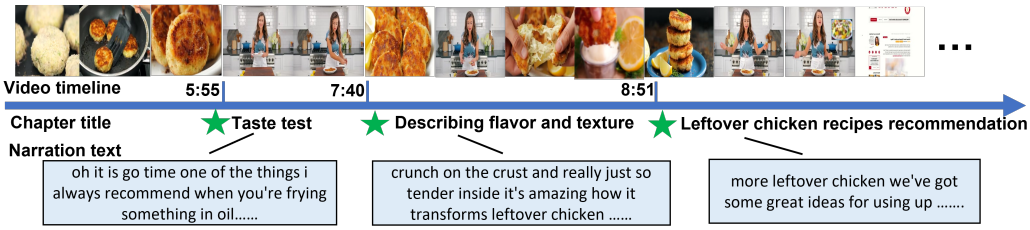
Since videos are automatically collected, there are some anomaly data. We improve the quality and consistency of the dataset by applying the following criteria. (1) We discard videos whose duration is less than 100 seconds. Because we think the short video contains insufficient information to generate valid chapters. (2) We also rule out videos in which the chapter duration is less than 8 seconds. If a chapter is too short, it is either incorrectly written by an amateurish video author or lacks useful information for chapter generation.

We observe that there is no uniform rule or guidance when video authors write their chapter descriptions. Some videos have clear content and temporal structures. But some have relatively vague structures. To evaluate performance thoroughly, we manually inspect the collected videos and mark

each video as an easy or hard case. The easy case represents the video has clear structures. It is easy to distinguish chapter boundaries and other video content. The hard case means that the video has vague structures and there are no obvious visual or narration features to identify chapter boundaries. Figure 1 gives some examples to demonstrate this concept.



(a) Easy case



(b) Hard case

Figure 1: We show the easy and hard cases in our dataset. (a) shows an easy case in which each chapter has clear temporal structures. (b) shows a hard case. There are no clear visual or narration features to determine where the chapter beginning is without watching most parts of the video.

4 METHODOLOGY

In this section, we present our framework to generate video chapters. Formally, in ChapterGen dataset, a video contains a series of image frames X and narration text S. The chapter descriptions contain a list of chapter beginning timestamps B and chapter titles D. We formulate the video chapter generation into a two-stage framework. First, a chapter localization function g is learned to map video content X, S to chapter beginning timestamps $B = g(X, S)$. Then, chapter titles D are summarized by using timestamps B, which provides chapter boundaries, and video context S. This procedure is learned by a chapter title generation function $D = h(B, S)$. Figure 2 demonstrates the two-stage framework. Next, we give the details of chapter localization and chapter title generation respectively.

4.1 CHAPTER LOCALIZATION

This stage takes into account the video frames X and narration text S to localize every chapter beginning time in a specific second. We consider the video context in time windows. The sliding window mechanism can be applied to determine where the chapter beginning time is. Note that our sliding window mechanism is different from the ones in temporal action localization Shou et al. (2016); Lin et al. (2018). Instead of using sliding windows to fit target intervals Shou et al. (2016), we use sliding windows to gather information and decide if the chapter begins in a time window. Lin et al. (2018) applies temporal convolutional layers to find action starting and ending time. But those layers are developed on a pretrained feature sequence and its time window slide second by second. By contrast, our sliding window mechanism is more general and efficient when handling long-duration videos.

The video context in a time window can be defined as a video clip. Each clip contains some images $X_{t_a} = \{x_t\}_{t=a}^{a+k}$ and narration words $S_{t_a} = \{s_t\}_{t=a}^{a+k}$, where k is the size of time window and a is

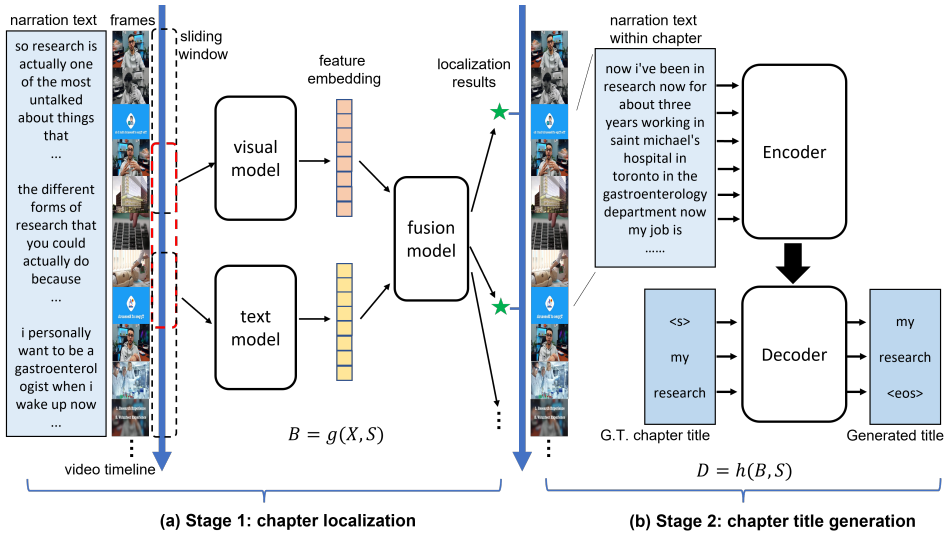


Figure 2: The framework of our approach. (a) Chapter localization stage. A two-stream network with a sliding window mechanism is developed for extracting visual and text features. (b) Chapter title generation stage. The Transformer encoder-decoder is used for generating chapter title.

the clip start time. Given a clip context X_{t_a}, S_{t_a}, we would like to evaluate the chapter beginning probability p_{t_a}. We formulate chapter localization function as a binary classifier $p_{t_a} = g(X_{t_a}, S_{t_a})$. The binary logistic regression loss function L_b can be defined for training:

$$L_b = \sum_{t_a} (l_{t_a} \cdot \log(p_{t_a}) + (1 - l_{t_a}) \cdot \log(1 - p_{t_a})) \tag{1}$$

where t_a denotes the clip start time. l_{t_a} is binary 0-1 label which represents if a chapter begins in a clip. We need to localize the chapter beginning time in a specific second. To fulfill it, we simply use the middle time of positive clips as the final localization results $\{b_i\} \in B$:

$$b_i = t_a + \frac{k}{2} \tag{2}$$

The Eq. 1 provides a training target to decide where the chapter beginning time is. Nevertheless, sliding clip windows second by second is inefficient, especially when processing long-duration videos. Since the video in ChapterGen dataset consists of hundreds and thousands of seconds of content. A few seconds offset in localization result is negligible in practice. We allow a small time offset o round the groundtruth chapter beginning time $\hat{t}$. All $\hat{t}$ can be sampled by sliding windows when setting the sliding step $u \leq 2o$. We adopt the maximum $u = 2o$ as the skip sliding step to achieve the best efficiency without losing any potential chapter beginning times. Under this skip sliding window, the label l_{t_a} can be defined as

$$l_{t_a} = \begin{cases} 0 & |t_a + \frac{k}{2} - \hat{t}| > o \\ 1 & |t_a + \frac{k}{2} - \hat{t}| \leq o \end{cases} \tag{3}$$

Figure 3 shows this skip sliding window mechanism.

4.2 Chapter Title Generation

We generate chapter titles after chapter localization. Instead of video frames, we use narration text to generate chapter titles, as narration text has plenty of high-level semantic information, while images usually contain low-level visual features. Moreover, both narration text and chapter titles are textual information. It is easier for the model to learn the potential relationship in the same modality.

The localization results provide chapter boundaries that can be used for selecting relevant video content. We adopt the standard Transformer encoder-decoder architecture Vaswani et al. (2017).

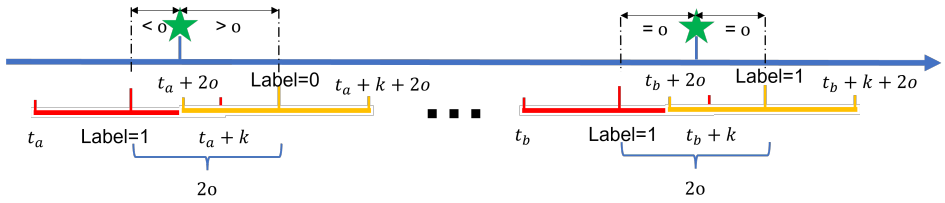


Figure 3: The skip sliding window mechanism. The video timeline is denoted by blue arrow with G.T. chapter beginning time marked by green stars. The red and yellow segments represent two consecutive sliding windows. If a small time offset o around G.T. is acceptable, the sliding window step $u = 2o$ can achieve best efficiency without missing any G.T. chapter beginning time.

Specifically, the goal of encoder-decoder Transformer is to estimate the conditional probability $p(d_1, ..., d_{n'}|s_1, ..., s_n)$ where $(s_1, ..., s_n)$ is the narration text sequence within a chapter boundary and $(d_1, ..., d_{n'})$ is its corresponding chapter title sequence whose length $n^{'}$ could much shorter than n. The Transformer architecture computes this conditional probability by first obtaining the fixed dimensional representation r of the input sequence $(s_1, ..., s_n)$ given by the last hidden state of the encoder part, and then computing the probability of $(d_1, ..., d_{n'})$ with the decoder part which uses the representation r to generate target words one by one:

$$p(d_1, ..., d_{n'}) = \prod_{i=1}^{n'} p(d_i|r, d_1, ..., d_{i-1}) \tag{4}$$

where every $p(d_i|r, d_1, ..., d_{i-1})$ distribution can be calculated by using a softmax over all the words. We initialize the encoder and decoder with pretrained parameters on large text corpora Zhang et al. (2020) and finetune the model on ChapterGen dataset to generate chapter titles. We also follow the word tokenization rule in Zhang et al. (2020) and use the special symbol "<s>", "<eos>" as the generation procedure start and end flags respectively. The chapter title generation procedure is shown in Figure 2(b).

5 Experiments

We conduct experiments on the ChapterGen dataset. First, we present comparative results on chapter localization and chapter title generation to show our approach competence. And then, a detailed ablation study is conducted to evaluate different settings in chapter localization and the performances on easy and hard cases. Finally, we demonstrate the qualitative results of the whole framework and show saliency maps to study what the model learns.

5.1 Experiment Setup

Dataset Details. The ChapterGen dataset is randomly split into train and test subsets with 7705, 1926 samples respectively. Unless mentioned, we use the sliding window hyperparameters $k = 16, o = 2$ in all experiments. Based on these hyperparameters, videos can be divided into a lot of clips for training and inference. Table 2 reports a detailed statistic of train/test subsets.

Dataset	Videos	Clips	Images	Words
Train	7705	1542783	6282810	18097041
Test	1926	388743	1583021	4551809

Table 2: The statistics of train/test split.

Evaluation Metrics. There are two stages in our framework and each stage has a different target. And thus, different evaluation metrics is needed. For chapter localization, we follow the metrics **AP** (Average Precision) and **Recall@c** Rao et al. (2020); Chen et al. (2021) for the sake of fair comparison. In ablation study, we report **F1-Score@c** results. The Recall@c and F1-Score@c

calculate the percentage of the ground truth chapter start time that is within c seconds of the predicted ones.

For chapter title generation, we adopt the commonly used metrics **ROUGE F1** scores (i.e. ROUGE-1, ROUGE-2, ROUGE-L) in text summarization works Narayan et al. (2018); Zhong et al. (2020).

Implementation Details. In chapter localization, we implement a binary classifier by a two-stream model to handle visual data and narration text. The visual feature is processed by standard pretrained ResNet-50 He et al. (2016) with TSM component Lin et al. (2019a). Compared with 3D CNN Tran et al. (2015), the TSM can be seamlessly integrated with 2D CNN to capture temporal features without introducing heavy computation burdens. We verify the effect of TSM in ablation study (see Section 5.3). The narration text feature is extracted by BERT Devlin et al. (2018). Before training, we finetune the BERT on ChapterGen dataset by unsupervised Masked Language Model (MLM) task. Finally, the extracted visual embedding and text embedding are concatenated and sent to a fully connected fusion layer to output final binary logits. All input images are resized to 224×224 resolution.

In chapter title generation, we apply Transformer encoder-decoder Vaswani et al. (2017) and initialize it with pretrained parameters Zhang et al. (2020). The maximum generated title length is set to 30 words. We train all models end-to-end using AdamW optimizer Loshchilov & Hutter (2017) with batch size 64. The learning rate is set to $1e^{-5}$ with linear warmup and cosine decay scheduler. All experiments are run on one NVIDIA Tesla V100 card.

5.2 Main Results

We evaluate chapter localization and chapter title generation respectively. In chapter localization, we follow the same evaluation metrics in Chen et al. (2021) using the Recall at 0.5 classification threshold (i.e. an input data is positive if its classification probability is larger than 0.5). As shown in Table 3, our method with both visual and text modalities achieves the best results in AP and Recall. This verifies the effectiveness of our model. Our method with only text modality has the fastest inference speed. Compared with the visual model, the text model has better performance. This may indicate that the narration text has more discriminative features than video frames in chapter localization. But visual clues are complementary for text. The combination of visual and text features can achieve the best results (AP 38.9% $\rightarrow$ 42.7%).

Rao et al. (2020); Chen et al. (2021) are the state-of-the-art methods on movie scene segmentation, which is related to our chapter localization problem. They are based on a shot detection preprocessing Sidiropoulos et al. (2011) and assume that the segmentation begins at shot boundaries. This assumption could be improper on our dataset which has complex temporal structures. As a result, their performances are inferior. Moreover, Chen et al. (2021) introduces unsupervised shot contrastive learning for pretraining. The performance of all layers finetuning is better than classifier head finetuning (i.e. freeze backbone). But both of them are worse than our visual model. This indicates that the simple shot contrastive pretraining may not be able to learn a good initialization on ChapterGen dataset.

Method	Modalities	Inference FPS	AP	Recall (0.5 thr.)	Recall@3s (0.5 thr.)	Recall@5s (0.5 thr.)
Random	-	-	-	1.1	7.4	11.1
Rao et al. (2020)	visual	2652.8	6.6	8.7	18.9	21.1
Chen et al. (2021)	visual (freeze)	2740.0	7.9	13.1	33.2	38.6
Chen et al. (2021)	visual (unfreeze)	2740.0	8.7	15.9	38.3	43.7
Our	visual	2874.2	26.3	18.1	43.5	58.6
Our	text	**5934.0**	38.9	21.9	54.2	71.6
Our	visual + text	1156.9	**42.7**	**25.6**	**60.8**	**75.7**

Table 3: The chapter localization results. The Inference FPS is the average model inferred frames per second which indicates the model efficiency. The bold fonts show the best results.

For chapter title generation, there is no related work for comparison. We compare our method with three heuristic strategies which are widely adopted in text summarization Zhong et al. (2020); Song

et al. (2019); Zhang et al. (2020). **Random:** uniformly select one sentence (i.e. 10 consecutive words) at random from the original narration text. **Lead:** select the first 10 words from the narration text as the chapter title. **Principal:** select top-1 scored sentence based on importance. The importance score is calculated by ROUGE-1 F-Score between the sentence and the rest of the narration text.

Method	Setting	ROUGE-1	ROUGE-2	ROUGE-L
Random	All	4.9	0.9	4.7
Lead	All	10.7	4.1	10.3
Principal	All	9.4	3.4	9.0
Our	Easy	**31.5**	**11.6**	**31.2**
	Hard	23.4	6.0	23.0
	All	28.5	9.5	28.2

Table 4: The chapter title generation results.

As shown in Table 4, Lead strategy is the best among all heuristic strategies. This may suggest that the first couple of sentences in each chapter contain the most conclusive information. Our title generation model outperforms all heuristic strategies with a large margin. This verifies the effectiveness of the Transformer encoder-decoder architecture in chapter title generation.

5.3 Ablation Study

We present ablation studies for the chapter localization stage to verify the choices of different sliding window size hyperparameters k, the influence of the TSM component, and the results on easy and hard cases.

Setting	Modalities	AP	F1-Score (0.5 thr.)	F1-Score@3s (0.5 thr.)	F1-Score@5s (0.5 thr.)
window size $k = 8$	visual + text	41.7	13.9	31.8	39.4
window size $k = 12$	visual + text	42.2	13.7	32.1	40.1
window size $k = 16$	visual + text	**43.1**	**15.1**	**35.1**	**43.0**
window size $k = 20$	visual + text	42.6	14.4	34.3	42.5
window size $k = 24$	visual + text	42.9	14.9	35.1	42.9
with TSM	visual	**26.3**	**7.2**	**17.6**	**25.5**
without TSM	visual	21.5	6.9	17.5	24.5
Easy cases	visual	**33.0**	**8.4**	**20.0**	**26.3**
Hard cases	visual	15.8	5.4	13.2	19.4
All cases	visual	26.3	7.2	17.5	23.8
Easy cases	text	**45.0**	**14.3**	**34.4**	**43.8**
Hard cases	text	28.9	9.8	24.1	33.4
All cases	text	38.9	12.5	30.6	39.8
Easy cases	visual + text	**50.2**	**17.5**	**40.7**	**48.3**
Hard cases	visual + text	30.0	11.4	25.7	34.2
All cases	visual + text	42.7	15.1	35.0	43.0

Table 5: The ablation study results on different settings.

Choice of sliding window size k. The sliding window size determines how much context information is considered for chapter localization. The small window size is helpful for the model to focus on most related content. The large window size provides more plentiful content information but may introduce more noises and extra computation burden. Table 5 shows the results of our two-stream model with different sliding window sizes. Window size $k = 16$ achieves best trade-off.

Influence of TSM. We capture visual features in temporal by using TSM Lin et al. (2019a). Compared with 3D CNN Tran et al. (2015), TSM is lightweight and can be seamlessly integrated with 2D CNN. The performance comparison is shown in Table 5. The model with TSM outperforms no TSM in all F1-scores and AP (26.3% vs. 21.5%).

Results on easy and hard cases. As illustrated in Figure 1, some videos have clear content and temporal structures, while some are vague. We conduct experiments on easy, hard and all cases to investigate how different input data influence the performance. As shown in Table 5, no matter what modality it is, the results on easy cases are better than hard cases. When mixing up all easy and hard cases, the performance is in-between easy and hard. The similar conclusion can also be observed for chapter title generation in Table 4. Besides, text modality always outperforms visual modality. This verifies the previous conclusion that narration text may contain more discriminative features than frames in chapter localization.

5.4 Qualitative Results

We present qualitative results on both chapter localization and chapter title generation to further explore their effectiveness. Figure 4(a) visualizes the comparison between G.T. chapters and generated chapters. Our framework can localize the chapter beginning time and generate desirable chapter titles.

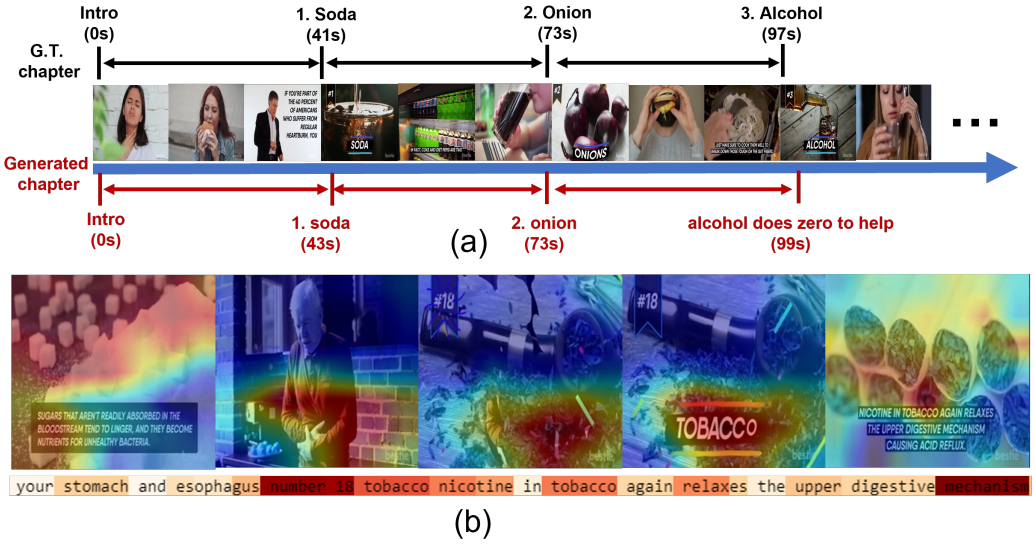


Figure 4: The chapter generation results and saliency maps. Our framework can precisely localize the chapter start time and generate semantically reasonable chapter titles.

Figure 4(b) demonstrates the saliency map on both video frames and narration text. It is a clip of images and narration text in a sliding window. We can observe that the regions of sugar, stomach and tobacco contribute the most in the visual modality. And "number 18", "tobacco", "mechanism" keywords in narration text are captured as important features. Based on this saliency map, we come to some empirical conclusions: (1) The visual model captures different regions (i.e. background or foreground) on different frames. It can focus on relevant regions to extract visual features. (2) The text model can grip attention on some important keywords for prediction.

6 Conclusion

In this work, we propose a two-stage framework for video chapter generation, which is a high-level video understanding task. The proposed framework contains chapter localization and chapter title generation. Chapter localization stage captures the visual dynamics and narration text to localize chapter start time. Chapter title generation takes the narration text within chapter boundary to generate chapter titles. To facilitate research for video chapter generation, we collect a dataset called ChapterGen. Our framework outperforms other state-of-the-art methods in chapter localization and improves the title generation results over traditional heuristic methods. Both quantitative and qualitative studies demonstrate that our framework can capture the association between complex narration text and temporal structures.

REFERENCES

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 961–970, 2015.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Shixing Chen, Xiaohan Nie, David Fan, Dongqing Zhang, Vimal Bhat, and Raffay Hamid. Shot contrastive self-supervised learning for scene boundary detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9796–9805, 2021.

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.

Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *European conference on computer vision*, pp. 505–520. Springer, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. 2020a.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *The European Conference on Computer Vision (ECCV)*, 2020b.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7083–7093, 2019a.

Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, 2018.

Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3889–3898, 2019b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018.

Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A local-to-global approach to multi-modal movie scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10146–10155, 2020.

Zeeshan Rasheed and Mubarak Shah. Scene detection in hollywood movies and tv shows. *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, 2:II–343, 2003.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3202–3212, 2015.

Yong Rui, Thomas S. Huang, and Sharad Mehrotra. Exploring video structure beyond the shots. *Proceedings. IEEE International Conference on Multimedia Computing and Systems (Cat. No.98TB100241)*, pp. 237–240, 1998.

Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1049–1058, 2016.

Panagiotis Sidiropoulos, Vasileios Mezaris, Yiannis Kompatsiaris, Hugo Meinedo, Miguel M. F. Bugalho, and Isabel Trancoso. Temporal video segmentation to scenes using high-level audiovisual features. *IEEE Transactions on Circuits and Systems for Video Technology*, 21:1163–1177, 2011.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*, 2019.

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*, pp. 5179–5187, 2015.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6479–6488, 2018.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, 2015.

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*, pp. 5998–6008, 2017.

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pp. 20–36. Springer, 2016.

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*, pp. 11328–11339. PMLR, 2020.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.