Multimodal Fusion and Coherence Modeling for Video Topic Segmentation

Hai Yu, Chong Deng, Qinglin Zhang, Jiaqing Liu, Qian Chen, Wen Wang
Speech Lab, Alibaba Group
{yuhai.yu, w.wang}@alibaba-inc.com

Abstract

The video topic segmentation (VTS) task segments videos into intelligible, non-overlapping topics, facilitating efficient comprehension of video content and quick access to specific content. VTS is also critical to various downstream video understanding tasks. Traditional VTS methods using shallow features or unsupervised approaches struggle to accurately discern the nuances of topical transitions. Recently, supervised approaches have achieved superior performance on video action or scene segmentation over unsupervised approaches. In this work, we improve supervised VTS by thoroughly exploring **multimodal fusion** and **multimodal coherence modeling**. Specifically, (1) we enhance multimodal fusion by exploring different architectures using Cross-Attention and Mixture of Experts. (2) To generally strengthen multimodality alignment and fusion, we pre-train and fine-tune the model with multimodal contrastive learning. (3) We propose a new pre-training task tailored for the VTS task, and a novel fine-tuning task for enhancing multimodal coherence modeling for VTS. We evaluate our proposed approaches on *educational videos*, in the form of *lectures*, due to the vital role of topic segmentation of educational videos in boosting learning experiences. Additionally, to promote research in VTS, we introduce a large-scale Chinese lecture video dataset to augment the existing English lecture video datasets. Experiments on both English and Chinese lecture datasets demonstrate that our model achieves superior VTS performance compared to competitive unsupervised and supervised baselines[1].

1 Introduction

The proliferation of digital video content over the last few decades has underscored the importance of efficient content navigation and comprehension. As the unstructured nature of videos poses significant challenges for users seeking to quickly grasp or reference specific topics, Video Topic Segmentation (VTS) has emerged as a vital tool in addressing these demands. By delineating videos into coherent non-overlapping topics, VTS not only facilitates intuitive understanding of video content but also enables swiftly pinpointing and accessing topics of interest. This is particularly pertinent for the furtherance of various video understanding tasks, where VTS serves as a foundational component.

Traditional VTS approaches predominantly hinge on shallow features (Gandhi et al., 2015; Soares and Barrére, 2018b; Ali et al., 2021) and unsupervised methods (Gupta et al., 2023), due to scarcity of labeled data. These methods often fall short in capturing the semantic cues that signal topical shifts in video streams, hence suffer from limited precision. Recent advancements in supervised learning paradigms have achieved notable performance improvements in multi-modal task (Yang et al., 2022; Tu et al., 2022, 2023; Zhang et al., 2022) and various video segmentation tasks, such as video action segmentation (Zhou et al., 2018; Tang et al., 2019), scene segmentation (Huang et al., 2020; Islam et al., 2023), and topic segmentation (Wu et al., 2023; Wang et al., 2023; Xing et al., 2024), surpassing unsupervised methods. Performance of supervised approaches can be further enhanced by pre-training on vast volumes of unlabeled data (Xu et al., 2021; Mun et al., 2022) or initializing models from pre-trained models (Yan et al., 2023) and then fine-tuning the model. Hence, in this work, we focus on further improving supervised methods for VTS.

Compared to text topic segmentation (Koshorek et al., 2018; Xing and Carenini, 2021; Yu et al., 2023), videos contain rich and diverse multimodal contextual information. Fully utilizing multimodal information, such as visual cues and textual data (e.g., screen text and subtitles), could facilitate

[1]Our code and dataset are publicly available at https://github.com/alibaba-damo-academy/SpokenNLP/

Findings of the Association for Computational Linguistics: ACL 2025, pages 17577–17593

more detailed content understanding and in turn more accurate semantic segmentation than relying on text only. Our case studies in Appendix H demonstrate the great challenges posed by VTS, particularly to unsupervised approaches or supervised methods that rely solely on either visual or textual modality. The complexity inherent in video content—where multimodal signals must be effectively integrated—accentuates the difficulty. Also, coherence is essential for understanding logical structures and semantics. Enhancing coherence modeling has achieved significant improvements in long text topic segmentation (Yu et al., 2023). Therefore, we improve supervised VTS methods by thoroughly exploring **multimodal fusion** and **multimodal coherence modeling**. We enhance multimodal fusion from the perspectives of model architecture and pre-training and fine-tuning tasks. Specifically, we compare various multimodal fusion architectures built upon Cross-Attention and Mixture-of-Experts (MoE). We investigate the effect of multimodal contrastive learning for general pre-training and fine-tuning for strengthening cross-modal alignment. For enhancing multimodal coherence modeling, we propose a new pre-training task tailored for the VTS task, and a novel fine-tuning task by elevating intra-topic multimodal feature similarities and inter-topic multimodal feature differences. The proposed approaches are extensively evaluated on educational videos, in the form of lectures, due to the pivotal contributions of topic segmentation of educational videos in bolstering the learning experiences.

Our contributions can be summarized as follows.

- We propose a supervised multimodal sequence labeling model for VTS, denoted **MMVTS**. We explore various multimodal fusion architectures, and apply multimodal contrastive learning for strengthening cross-modal alignment. We also propose a new self-supervised pre-training task tailored to the VTS and a novel fine-tuning task for enhancing multimodal coherence modeling.
- We introduce a large-scale **C**hinese **L**ecture **V**ideo **T**opic **S**egmentation dataset (**CLVTS**) to promote the research of VTS.
- Experiments show that our model sets new state-of-the-art (SOTA) performance on both English and Chinese lecture video datasets, outperforming competitive unsupervised and supervised baselines. Comprehensive ablation study further confirms the effectiveness of our approaches.

2 Related Work

Text Topic Segmentation Text topic segmentation aims to automatically partition text into topically consistent, non-overlapping segments (Hearst, 1994). By automatically mining clues of topic shifts from large amounts of labeled data (Koshorek et al., 2018; Arnold et al., 2019), contemporary supervised models (Lukasik et al., 2020; Somasundaran et al., 2020; Zhang et al., 2021; Yu et al., 2023) demonstrate superior performance compared to unsupervised approaches (Riedl and Biemann, 2012; Solbiati et al., 2021). Notably, supervised models that excel at modeling long sequences (Zhang et al., 2021; Yu et al., 2023) are capable of capturing longer contextual nuances and thereby achieve better topic segmentation performance, compared to models that model local sentence pairs or block pairs (Wang et al., 2017; Lukasik et al., 2020). In addition, recent works (Somasundaran et al., 2020; Xing et al., 2020; Yu et al., 2023) show that strengthening coherence modeling can improve text topic segmentation performance. Inspired by these findings, in this work, we explore enhancing coherence modeling for video topic segmentation under the *multimodal* configurations.

Video Topic Segmentation For video topic segmentation, some approaches, such as BaSSL Mun et al. (2022), explore visual-only information. However, many recent works have achieved enhanced semantic understanding of videos by leveraging multimodal data. Gupta et al. (2023) introduced UnsupAVLS, which uses the TWFINCH algorithm to cluster video clips into topics based on visual and text features. Wang et al. (2023) proposed SWST, which concatenates visual and text features for language models; however, it may suffer from discrepancies between pre-training of the language model and fine-tuning. Wu et al. (2023) focused on hierarchical modeling of scene, story, and topic, without further exploring how to better integrate multimodal features. Xing et al. (2024) employed asymmetric cross-modal attention for obtaining text-aware visual representations. It may be most related to our work. However, our work differs from Xing et al. (2024) as we explore symmetric cross-modal attention and also investigate the Mixture-of-Experts mechanism, as well as introducing topic-level Contrastive Semantic Similarity Learning into fine-tuning for enhanced coherence modeling in the multimodal framework.

Dataset	Videos	Hours	Topics/Video	Clips/Topic	Seconds/Clip	Domain	Language	Available
NPTEL10 (Gandhi et al., 2015)	12	-	-	-	-	Education	English	×
Videoaula (Soares and Barrére, 2018a)	44	26.4	-	-	-	Education	Portuguese	✓
CS80 (Soares and Barrére, 2019)	80	-	-	-	-	Education	English	✓
MOOC100 (Das and Das, 2019)	100	100	6.9	-	-	Education	English	✓†
Coursera37 (Chand and Oğul, 2021)	37	2.8	16.5	-	-	Education	English	×
VSTAR (Wang et al., 2023)	8159	4625	61.2	0.4	90	Television	English	✓†
NewsNet (Wu et al., 2023)	1000	946	8.5	-	-	News	English	×
MultiLive (Qiu et al., 2023)	1000	1300	8.8	-	-	Livestream	English	×
AVLecture (Gupta et al., 2023)	350	297.5	5.4	46.2	12.3	Education	English	✓
YouTube (Xing et al., 2024)	5422	858.5	6.7	16	5.3	Diverse	English	×
Behance (Xing et al., 2024)	575	1225.2	5.2	248	6.0	Livestream	English	×
CLVTS (Ours)	510	395	10.1	35.7	7.7	Education	Chinese	✓

Table 1: Comparison between our **CLVTS** dataset and existing video datasets for the video topic segmentation task. † indicates that the data is not entirely open source. Prior to our work, AVLecture is the only publicly available large-scale video dataset supporting supervised VTS methods.

3 Methodology

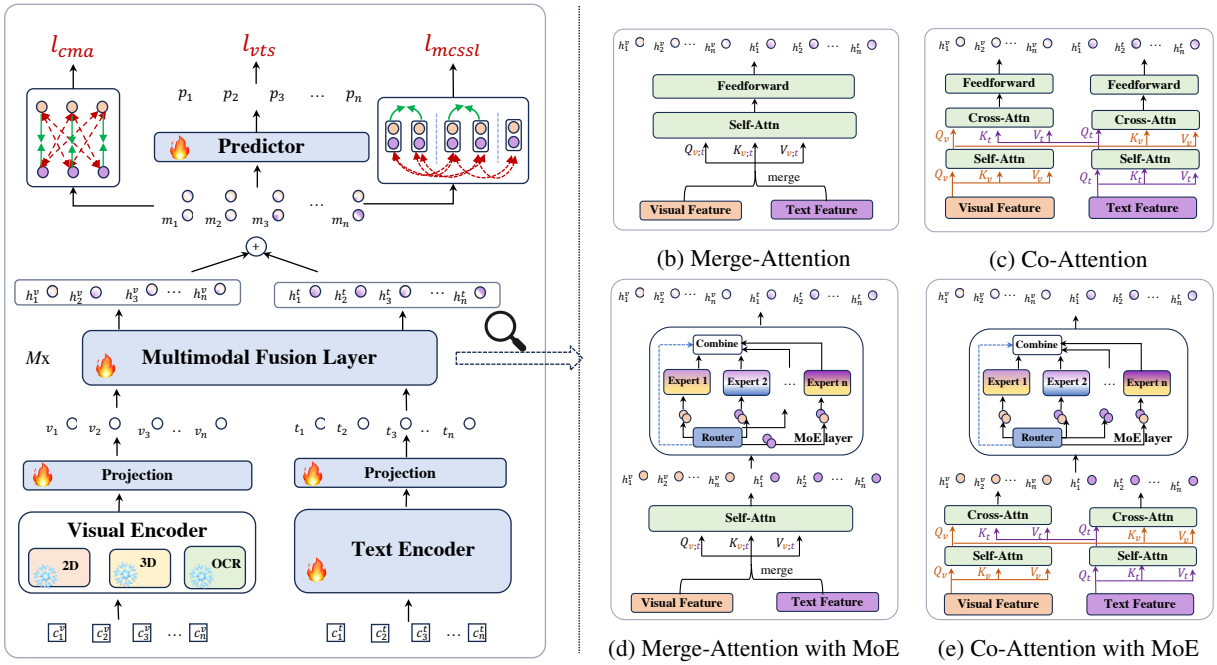
Figure 1 depicts the overall architecture of our MMVTS model. Section 3.1 presents the problem definition of multimodal VTS and the overall model architecture. We enhance multimodal fusion from the perspectives of model architecture, pre-training, and fine-tuning tasks. Specifically, we compare different multimodal fusion architectures built upon Merge- and Cross-Attention, and Mixture-of-Experts (Section 3.1). We explore multimodal contrastive learning for cross-modality alignment and propose a new pre-training task tailored for VTS (Section 3.2). For fine-tuning, we also propose a novel task for multimodal coherence modeling (Section 3.3).

3.1 MultiModal Video Topic Segmentation

Overall Architecture. Following prior works (Zhang et al., 2021; Wu et al., 2023), we define video topic segmentation as a clip-level sequence labeling task and propose our **M**ulti**M**odal **V**ideo **T**opic **S**egmentation (MMVTS) model. As illustrated in Figure 1a, we apply unimodal pre-trained encoders for the vision and text modality, respectively, and then fuse multimodal information at the intermediate representation level (i.e., *middle fusion* (Xu et al., 2023)) through the **Multimodal Fusion Layer**. Given a video, we transcribe it with a competitive automatic speech recognition (ASR) system[2] and use ASR 1-best as the text modality. We then divide the video into n clips $(c_i^v, c_i^t)_{i=1}^n$, with clips segmented at the sentence boundaries predicted on ASR 1-best. $c_i^v = \{f_1^i, ..., f_k^i\}$ denotes evenly sampled k frames within the i-th clip and is fed into a visual encoder $\boldsymbol{E_v}$ to extract visual features. $c_i^t = \{w_1^i, ..., w_{\|s_i\|+1}^i\}$ denotes the sequence of words from ASR 1-best within the i-th clip, where w_1^i is the inserted special token [BOS] and $\|s_i\|$ denotes the number of words in the i-th clip. c_i^t is fed into a text encoder $\boldsymbol{E_t}$ and the last hidden representation of [BOS] for the i-th clip is used as the text representation of the clip. After extracting the unimodal features, we first apply trainable projection matrices W_v and W_t to convert unimodal features into the same dimension, resulting in the visual feature sequence $\boldsymbol{v} = \{v_1, ..., v_n\}$ and the textual feature sequence $\boldsymbol{t} = \{t_1, ..., t_n\}$ (Eq. 1). Then we fuse the multimodal information with M Multimodal Fusion Layers $\boldsymbol{MFL}_M$ and obtain the updated visual features $\boldsymbol{h^v} = \{h_1^v, ..., h_n^v\}$ and textual features $\boldsymbol{h^t} = \{h_1^t, ..., h_n^t\}$ (Eq. 2), which are then concatenated into the multimodal features $\boldsymbol{m} = \{m_1, ..., m_n\}$ (Eq. 3). Finally, the multimodal features $\boldsymbol{m}$ are fed into the predictor consisting of a linear layer W_p to obtain the probability of binary classification $\boldsymbol{p} = \{p_1, ..., p_n\}$ (Eq. 4), where p_i indicates whether the i-th clip is at a topic boundary. We use the standard binary cross-entropy loss (Eq. 5) to train the model, where $y_i \in \{0, 1\}$ is the label. The last clip is excluded from loss computation. **Considering computational complexity, we freeze the visual encoder while keeping all other parameters trainable**.

Compared to *late fusion* where no cross-modal interaction happens until after independent predictions by each unimodal model, *middle fusion* and *early fusion* are found to generally outperform late fusion (Nagrani et al., 2021), probably because early and middle fusion aligns better with human perception where multimodal fusion happens early in sensory processing. On the other hand, compared to early fusion, middle fusion yields superior or comparable performance (Nagrani et al., 2021)

[2] https://tingwu.aliyun.com/home



(a) Overall Model Architecture

(b) Merge-Attention

(c) Co-Attention

(d) Merge-Attention with MoE

(e) Co-Attention with MoE

Figure 1: The overall architecture of our **MMVTS model** and four distinct architectures of the Multimodal Fusion Layers in (a). In the overall architecture, the snowflake symbol indicates that the parameters of the certain module are frozen; whereas, the flame symbol signifies a trainable module. The blue dotted lines in the l_{mcssl} module denote the topic boundaries. The green solid lines in the l_{cma} module depict the features being brought closer, while the red dashed lines depict the features being pushed apart.

and is much less computationally expensive since we could freeze some strong pre-trained unimodal encoders with a large number of parameters and only train a small number of parameters.

$$\begin{aligned} \boldsymbol{v} &= W_v \cdot \boldsymbol{E_v}(\{c_1^v, ..., c_n^v\}) \\ \boldsymbol{t} &= W_t \cdot \boldsymbol{E_t}(\{c_1^t, ..., c_n^t\}) \end{aligned} \tag{1}$$

$$\boldsymbol{h^v}; \boldsymbol{h^t} = \boldsymbol{MFL}_M(\boldsymbol{v}; \boldsymbol{t}) \tag{2}$$

$$m_i = h_i^v; h_i^v \tag{3}$$

$$\boldsymbol{p} = W_p \cdot \boldsymbol{m} \tag{4}$$

$$l_{vts} = -\sum_{i=1}^{n-1} [y_i \ln p_i + (1-y_i) \ln(1-p_i)] \tag{5}$$

Multimodal Fusion Layer (MFL). We compare four distinct cross-modal interaction mechanisms for Multimodal Fusion Layers. We investigate the Merge-Attention and Co-Attention multimodal fusion layers proposed in (Dou et al., 2022; Yang et al., 2023b). With ***Merge-Attention*** (Figure 1b), features from unimodal encoders are concatenated sequentially and then input into a standard transformer encoder layer (Vaswani et al., 2017), which *shares* attention parameters across modalities. A feed forward layer is added on top to produce the final output representation. In contrast, with ***Co-Attention*** (Figure 1c), features from each unimodal encoder first go through self-attention with *modality-specific* attention parameters, then we perform **symmetric cross-attention** to integrate information from all other modalities to enhance the representation of the considered modality, followed by a feed forward layer.

Inspired by (Mustafa et al., 2022), which interleaves MoE encoder layers and standard dense encoder layers for image-text multimodal models, we also investigate two new architectures by replacing the traditional single feed-forward layers in Figure 1b and 1c with a MoE module (Shazeer et al., 2016; Lepikhin et al., 2020). The resulting architectures are depicted by Figure 1d and 1e. The motivation is that adding MoE on top of the fused representations may *facilitate deeper cross-modal integration of information* and *improve model capacity without a proportional increase in computational complexity*. Specifically, *experts* are MLPs activated depending on the input. Firstly, we concatenate fused features output from self-attention or cross-attention. Then, we implement the Noisy Top-k Gating mechanism (Shazeer et al., 2016) to select K experts from a total of E candidates (Eq. 6 - 8), where $SN()$ denotes the standard normal dis-

tribution, W_n denotes tunable Gaussian noise to help load balancing, W_g is a trainable weight matrix, K and E are hyper-parameters. Finally, the outputs of the K activated experts are linearly combined with the learned gating weights (Eq. 9). For the MoE training objective, we sum the *importance* loss and the *load* loss (Shazeer et al., 2016) to balance expert utilization as in Eq. 10.

$$G(x) = Softmax(KeepTopK(H_x, k)) \tag{6}$$

$$H(x)_i = (W_g \cdot x)_i + SN() \cdot Softplus((W_n \cdot x)_i) \tag{7}$$

$$KeepTopK(x,k)_i = \begin{cases} x_i \text{ if } x_i \text{ is in top-k.} \\ -\infty \text{ otherwise.} \end{cases} \tag{8}$$

$$MoE(x) = \sum_{e=1}^{K} G(x)_e \cdot MLP_e(x) \tag{9}$$

$$l_{balance} = l_{importance} + l_{load} \tag{10}$$

3.2 Pre-training with Unlabeled Data

Prior works have demonstrated that standard self-supervised denoising pre-training (even only using the downstream task data) (Amos et al., 2023) or pre-training adapted to the downstream task (Gururangan et al., 2020) often perform substantially better than randomly initializing the parameters. Therefore, to better initialize the parameters of the Multimodal Fusion Layers, we explore pre-training with unlabeled video data before supervised fine-tuning. Firstly, we introduce **a general cross-modality alignment pre-training task** to learn the multimodal representation. We use contrastive learning loss to adjust the features learned by the Multimodal Fusion Layers, by maximizing the cosine similarity of the visual features and textual features of the same clip, while reducing the similarity of the modality features between different clips, as show in Eq. 11, where ϵ is used to prevent division by 0 and τ is a temperature hyper-parameter to scale the cosine similarity.

$$l_{cma} = -\frac{1}{n} \frac{\sum_{i=1}^{n} e^{sim(h_i^v, h_i^t)}}{\sum_{i=1}^{n} \sum_{j=1}^{n} e^{sim(h_i^v, h_j^t)} + \epsilon} \tag{11}$$

$$sim(x_1, x_2) = \frac{x_1^{\mathrm{T}} \cdot x_2}{\|x_1\| \cdot \|x_2\|} / \tau \tag{12}$$

Secondly, we introduce **a novel pre-training task tailored for the VTS task**, focusing on utilizing unlabeled data for learning pseudo topic boundaries and also enhancing modality alignment. We apply a Kernel Density Estimation (KDE) (Davis et al., 2011) model to estimate the topic duration distribution within the labeled training set. Videos are segmented based on KDE-sampled durations. For each segment, with equal probability, we: insert a random segment from other videos, replace it with another, or retain it. These modified segments serve as distinct topics, allowing the model to learn pseudo topic boundaries during pre-training. This task-adaptive pre-training task has the same l_{vts} objective as shown in Eq. 5. The overall pre-training objective is shown in Eq. 13, where α and β are hyper-parameters to adjust the loss weights.

$$l_{pretrain} = l_{vts} + \alpha l_{cma} + \beta l_{balance} \tag{13}$$

3.3 Fine-tuning with Multimodal Coherence Modeling

For fine-tuning, we introduce two auxiliary tasks to enhance multimodal coherence modeling. The cross modal alignment task is the same as the task in Eq. 11 used in pre-training. This continuity ensures that the modalities retain their coherence through both pre-training and fine-tuning stages, fostering a consistent interplay between different modalities. In addition, we adapt the Contrastive Semantic Similarity Learning (CSSL) task proposed by Yu et al. (2023), which leverages the inherent characteristics of topic-related coherence, to the multimodal context. We adopt the same strategy for selecting positive and negative sample pairs (Yu et al., 2023), but extend the features to the multimodal representations, as shown in Eq. 14, where k_1 and k_2 are hyper-parameters that determine the number of positive and negative pairs. For each clip's multimodal representation m_i, $m_{i,j}^+$ denotes the multimodal representation of the j-th similar clip in the same topic as clip i, while $m_{i,j}^-$ denotes the multimodal representation of the j-th dissimilar clip in a different topic from clip i. We hypothesize that this extension could improve multimodal representation learning by identifying relative consistency relations within topics and across topics.

$$l_{mcssl} = -\frac{1}{n} \sum_{i=1}^{n} log \frac{\sum_{j=1}^{k_1} e^{sim(m_i, m_{i,j}^+)}}{\sum_{j=1}^{k_1} e^{sim(m_i, m_{i,j}^+)} + \sum_{j=1}^{k_2} e^{sim(m_i, m_{i,j}^-)}} \tag{14}$$

The overall fine-tuning objective combines Eq. 5, 10, 11, and 14, as shown in Eq. 15, where σ, θ, and γ are hyper-parameters to adjust loss contribution. When the Multimodl Fusion Layers do not contain MoE structure, β in Eq. 13 and σ are set to zero.

$$l_{finetune} = l_{vts} + \sigma l_{balance} + \theta l_{mcssl} + \gamma l_{cma} \tag{15}$$

Model			Modality	AVLecture					CLVTS				
				F_1	*BS@30*	F_1*@30*	*mIoU*	*Avg*	F_1	*BS@30*	F_1*@30*	*mIoU*	*Avg*
UnsupAVLS (Gupta et al., 2023)			V+T	-	56.00‡	-	70.86‡	-	-	-	-	-	-
BaSSL (Mun et al., 2022)			V	-	43.94	-	46.95	-	-	-	-	-	-
LongFormer (Yu et al., 2023)			T	52.91	69.25	60.38	67.54	62.52	34.42	52.19	47.77	52.87	46.81
LongFormer$_{cssl}$ (Yu et al., 2023)			T	54.02	71.56	62.40	68.39	64.09	34.77	53.07	47.51	53.15	47.12
Llama-3-8B$_{Generative}$			T	40.00	57.55	56.52	62.8	54.22	27.50	40.58	43.71	50.38	40.54
Llama-3-8B$_{Discrete}$			T	39.27	68.8	62.55	70.43	60.26	31.47	60.40	**54.64**	**58.86**	**51.34**
SWST$_{seq}$ (Wang et al., 2023)			V+T	53.45	70.95	59.73	65.21	62.33	34.55	52.77	48.08	52.67	47.02
PT	**FT-Coh**	**MMVTS Models (Ours)**	**Modality**										
×	×	Baseline$_1$	V+T	55.19	71.76	61.19	66.39	63.63	37.32	49.75	47.07	50.51	46.16
×	✓	Baseline$_2$		56.72	72.56	63.03	67.97	65.07	37.29	48.48	47.62	51.73	46.28
✓	✓	Baseline$_3$		58.77	72.55	67.26	71.52	67.52	36.54	50.67	48.81	52.56	47.15
✓	✓	Merge-Attn	V+T	57.36	74.96	65.30	70.15	66.94	38.17	55.52	50.69	54.84	49.80
✓	✓	Co-Attn		**60.01**	73.88	67.27	**72.32**	68.37	38.49	57.23	50.59	54.47	50.20
✓	✓	Merge-Attn with MoE		57.54	73.48	64.36	70.43	66.45	38.77	**61.05**	51.10	54.41	51.33
✓	✓	Co-Attn with MoE		59.77	**75.01**	**67.94**	71.69	**68.61**	**39.98**	58.96	51.41	54.71	51.27

Table 2: Performance of baselines and our MMVTS models on AVLecture and CLVTS test sets. ‡ denotes the leakage of the ground-truth topic number. **V** and **T** under **Modality** denote Vision and Text modality, respectively. MMVTS Baseline$_{1,2,3}$ denote our MMVTS model w/o Multimodal Fusion Layers. Attn denotes Attention. **PT** denotes pre-training the model on unlabeled data (Section 3.2 Eq. 13) before fine-tuning. **FT-Coh** denotes adding the two auxiliary multimodal coherence modeling tasks during fine-tuning (Section 3.3 Eq. 15); w/o FT-Coh refers to fine-tuning with the standard l_{vts} (Eq. 5). For each metric, the best result among all models is boldfaced while the best result in each group is underscored.

4 Experiments

4.1 Experimental Setup

Datasets. Table 1 summarizes the statistics of various VTS datasets. It clearly shows that *prior to our work, AVLecture (Gupta et al., 2023) is the only publicly available large-scale labeled video dataset facilitating supervised VTS methods*. To promote the research in VTS, we introduce a large-scale labeled Chinese Lecture Video Topic Segmentation dataset (**CLVTS**). Both AVLecture and CLVTS are sourced from educational videos, where VTS significantly enhances learning experiences. In terms of differences, ***in addition to the linguistic distinctness from the English lecture dataset AVLecture, CLVTS is characterized by its natural and uninterrupted long videos, a stark contrast to AVLecture***, since nearly two-thirds of AVLecture are reassembled pre-segmented short videos. As shown in Table 1, CLVTS features a higher average number of topics per video than AVLecture. Details of the data collection and annotation procedure and analysis of the CLVTS dataset are in Appendix A. ***Importantly, we put careful ethical considerations for the datasets used in this research in Appendix A.3***.

Baselines and Implementation Details. The implementation details are in Appendix B. We carefully select the following representative baselines.

- UnsupAVLS (Gupta et al., 2023) is an unsupervised approach that clusters video clips into a predefined number of topics, based on visual and text embeddings learned from matching the narration with the temporally aligned visual content.

- Visual-only BaSSL (Mun et al., 2022) is initially proposed for *video scene segmentation*. We use their released checkpoints to initialize our model and fine-tune on the VTS task to evaluate the performance of a *visual-only* model.

- Text-only LongFormer is evaluated on long document topic segmentation by Yu et al. (2023). We fine-tune LongFormer and LongFormer$_{cssl}$ in (Yu et al., 2023) to evaluate the performance of a *text-only* model w/o and w/ Contrastive Semantic Similarity Learning (CSSL) on the VTS task.

- Llama-3-8B Our pre-training (Section 3.2) is conducted on the relatively limited unlabeled videos of AVLectures and CLVTS datasets. To investigate the effect of fine-tuning a powerful pre-trained text large language model (LLM) on VTS, we fine-tune Llama-3-8B[3] with 8B parameters, using two different prompts (see Appendix E for details).

- SWST$_{seq}$ is our adapted version of the multimodal video scene and topic segmentation model (Wang et al., 2023) with a pre-trained LongFormer (Beltagy et al., 2020) as the backbone to VTS, for comparing performance between their *early fusion* and our *middle fusion* strategy on VTS.

In this work, we choose not to utilize pre-trained vision-language models such as (Yang et al., 2023a; Nguyen et al., 2024) due to their limitations in processing long video content as the case of educational videos, although these models demonstrate

[3] https://huggingface.co/meta-llama/Meta-Llama-3-8B

strong performance on short video clips lasting several seconds. In future work, we plan to enhance the long video understanding capabilities of large multimodal models (Zou et al., 2024; Zhou et al., 2024) to enable their applications to VTS.

Evaluation Metrics. We adopt four commonly used metrics, including positive F_1 (Zhang et al., 2021) (denoted as $\boldsymbol{F}_1$ for brevity), **BS@k** (Gupta et al., 2023), ***mIoU*** (Mun et al., 2022), and $\boldsymbol{F}_1$***@k***. Definitions of the four metrics are in Appendix C. Following Gupta et al. (2023), we set k to 30 seconds. We compute the average of these four metrics, denoted by ***Avg***, to measure the overall performance of a model.

4.2 Results and Analysis

Table 2 compares the performance of baselines (the first group) and variants of our MMVTS models (the second and the third group).

Unimodal performance. For BS@30 on AVLecture, the text-only Longformer (Row 3) outperforms the visual-only BaSSL (Row 2) by a large gain (+25.31), and also surpasses the unsupervised UnsupAVLS by a notable gain (+13.25). Such results are expected since ***the text modality inherently conveys more precise information for VTS than the vision modality***. Notably, the high *mIoU* of the unsupervised method is attributable to the leakage of the ground-truth number of topics.

Mutimodal performance. As shown in Table 2, ***Avg*** (the average of F_1, *BS@k*, *mIoU*, F_1*@k*) of the multimodal model $SWST_{seq}$ is only comparable to Avg of the text-only LongFormer, suggesting that more data may be necessary to mitigate the discrepancy between pre-training of the language model and fine-tuning with early fusion (as in $SWST_{seq}$), in order to fully exploit the potential of the early fusion strategy. Our MMVTS Baseline$_{1,2,3}$ simply concatenate unimodal features to predict topic boundaries. Without pre-training on unlabeled data (PT, Eq. 13) and the two auxiliary fine-tuning tasks to enhance multimodal coherence modeling (FT-Coh, Eq. 15), on F_1, MMVTS Baseline$_1$ outperforms the text-only Longformer by 2.28 and 2.90 on AVLecture and CLVTS respectively, while on *Avg* score, MMVTS Baseline$_1$ outperforms Longformer by 1.1 on AVLecture yet slightly underperforms Longformer by 0.65 on CLVTS. These results suggest that ***simply concatenating unimodal features to predict topic boundaries does not bring consistent gains over unimodal models***. The third group in Table 2 evaluates our MMVTS models with the four Multimodal Fusion Layer (MFL) architectures and with pre-training (PT) and fine-tuning (FT-Coh). ***Overall, after PT and FT-Coh, on AVLecture, our MMVTS model using Co-Attention with MoE as MFLs outperforms all the competitive unsupervised and supervised visual-only and text-only baselines as well as the multimodal $SWST_{seq}$ and achieves the best Avg (4.52 absolute and 7.05% relative gain over previous SOTA), and the bestBS@30 and F_1@30 results, setting new SOTA on AVLecture. On CLVTS, our MMVTS model achieves the best F_1 (5.21 absolute and 14.98% relative gain over previous SOTA) and yields the Avg score comparable to the best Avg score.*** On AVLecture, both gains on *Avg* from our MMVTS model with Co-Attention MoE after PT and FT-Coh over the previous SOTA LongFormer$_{cssl}$ and MMVTS Baseline$_1$ are statistically significant ($p < 0.05$). Moreover, our MMVTS model significantly outperforms the multimodal model $SWST_{seq}$ on both datasets, demonstrating the effectiveness of our ***middle fusion*** strategy and new pre-training and fine-tuning methods. Table 5 shows that our MMVTS model with Co-Attention with MoE has 192M trainable parameters while LongFormer$_{cssl}$ has 130M parameters. It is also notable that the performance of models on CLVTS is generally much lower than that on AVLecture, with the best *Avg* on AVLecture and CLVTS differing by 17.50 (68.84 versus 51.34), indicating ***a greater challenge to VTS from our CLVTS dataset than the AVLecture dataset***.

Comparison with fine-tuning text LLMs. Table 2 also shows that on CLVTS, fine-tuning the powerful pre-trained text LLM Llama-3-8B with our *Discrete* prompt (Appendix E) achieves the best *Avg* score 51.34, probably attributable to Llama-3-8B's extensive pre-training and vast knowledge base; still, the performance of our MMVTS model w/ Merge-Attn and MoE and Co-Attn and MoE after PT and FT-Coh is nearly the same, with Avg **51.33** and 51.27 respectively. However, on AVLecture, fine-tuning Llama-3-8B performs much worse than MMVTS model (60.26 versus our 68.61). Particularly, F_1 on both AVLecture and CLVTS from Llama-3-8B are much worse than MMVTS model as we find that Llama-3-8B's predicted boundaries often have a clip offset. ***These results underscore the value of our proposed small VTS models, since the efficiency and flexibility of our competitive to superior small models make them indispensable in many real-world applications***. Future research

PT	Model	F_1	*BS@30*	F_1*@30*	*mIoU*	*Avg*
×	Merge-Attn	56.56	73.28	64.03	70.06	$65.98_{0.90}$
×	Co-Attn	*57.71*	72.20	65.40	70.20	$66.38_{1.29}$
×	Merge-Attn with MoE	56.80	72.26	63.44	69.65	$65.54_{1.64}$
×	Co-Attn with MoE	*57.71*	*74.30*	*65.53*	*71.45*	$67.25_{0.56}$
✓	Merge-Attn	57.36	74.96	65.30	70.15	$66.94_{0.60}$
✓	Co-Attn	**60.01**	73.88	67.27	**72.32**	$68.37_{0.52}$
✓	Merge-Attn with MoE	57.54	73.48	64.36	70.43	$66.45_{0.14}$
✓	Co-Attn with MoE	59.84	**75.62**	**67.69**	72.21	$\mathbf{68.84}_{0.93}$

Table 3: Ablation studies of the pre-training tasks on AVLecture test set. The two auxiliary coherence modeling tasks are added in fine-tuning (Eq. 15). For **Avg**, we report mean and standard deviation from three runs with different random seeds.

Model	F_1	*BS@30*	F_1*@30*	*mIoU*	*Avg*
Co-Attn with MoE	59.84	**75.62**	**67.69**	72.21	**68.84**
w/o l_{cmal}	58.96	74.62	67.39	72.17	68.29
w/o l_{mcssl}	59.47	74.53	66.74	**72.24**	68.25
w/o l_{cmal} & l_{mcssl}	**60.57**	73.36	66.52	70.42	67.72

Table 4: Ablation studies of the two auxiliary coherence modeling fine-tuning tasks on AVLecture test set. Models are initialized from pre-training (Eq. 13).

could continue exploring how to integrate strengths of LLMs and multimodal approaches for VTS.

Incorporate audio modality. We also explore adding the audio modality (A) to the vision and text modalities (V+T) for our MMVTS model and find that V+T+A slightly improves the Avg score over V+T by 1.09% and 3.32% relatively, as shown in Appendix I. Our ongoing research explores different audio features as well as directly employing visual and audio cues (i.e., V+A) for VTS.

We conduct extensive ablation studies to validate effectiveness of the proposed Multimodal Fusion Layers, pre-training and fine-tuning tasks.

(1) Effect of Multimodal Fusion Layers. Comparing the third group and Baseline$_3$ in Table 2 shows that with the same pre-training and fine-tuning, the best performing architecture using Multimodal Fusion Layers always substantially outperforms Baseline$_3$. Specifically, with PT and FT-Coh, on AVLecture, both Co-Attention and Co-Attention with MoE notably outperform MMVTS Baseline$_3$ by 1.32 on *Avg*; on CLVTS, all four Multimodal Fusion Layer architectures achieve remarkable gains on *Avg* over MMVTS Baseline$_3$, from 2.65 to 4.18. These results demonstrate that **deep cross-modal interaction has notable advantage for multimodal fusion over simple unimodal feature concatenation for VTS**. Moreover, adding MoE on top consistently improves Co-Attention on both AVLecture and CLVTS, by 0.47 and 1.07 on *Avg*; whereas, the effect of MoE on top of Merge-Attention is inconsistent, with a slight degradation on AVLecture and 1.53 gain on *Avg* on CLVTS. In addition, we conduct more analysis of the effect of Co-Attn with MoE with different numbers of multimodal fusion layers in Appendix G.

(2) Effect of Pre-training tasks. Table 2 shows that for simple concatenation of unimodal features, pre-training before fine-tuning (as Baseline$_3$) outperforms Baseline$_2$ (w/o PT). We conduct ablation studies of the proposed pre-training and fine-tuning tasks on AVLecture. We apply the same fine-tuning with multimodal coherence modeling (FT-Coh, Eq. 15) and compare (a) random initialization of parameters for Multimodal Fusion Layers (w/o pre-training) (b) pre-training the model on unlabeled data. Table 3 shows that w/o pre-training, Co-Attn slightly improves *Avg* over Merge-Attn by 0.4, and MoE further improves *Avg* by 0.87. ***Pre-training improves the performance on all four MFL architectures, with the average Avg score increased by 1.36 (66.29 → 67.65)***. Pre-training also ***improves model stability*** as the standard deviations of all w/ PT experiments are less than 1. Table 6 in Appendix further compares the fine-tuning performance after applying different pre-training tasks. Compared to using both pre-training tasks, removing l_{cma} or l_{vts} degrades *Avg* by 0.4 and 1.64 respectively, indicating that the pretraining task aligned more closely with the downstream task, i.e., l_{vts}, yields greater gains.

(3) Effect of Fine-tuning tasks. Table 4 compares different fine-tuning tasks after pre-training. Compared to the standard l_{vts}, adding the two auxiliary losses l_{cma} and l_{mcssl} notably improves *Avg* by 1.12, while decreasing F_1 by 0.73. These results suggest that ***while multimodal coherence modeling may slightly compromise the precision of exact matches, it enhances the overall contextual comprehension of a model for VTS***. Adding l_{cma} or l_{mcssl} individually improves *Avg* by 0.53 and 0.57, with improvements mainly on *BS@30* and *mIoU*, suggesting that feature alignment at different granularities may improve fuzzy matching.

5 Conclusion

We propose a novel supervised VTS model by thoroughly exploring multimodal fusion and coherence modeling. We also introduce a large-scale labeled Chinese Lecture dataset for VTS. Extensive experiments demonstrate the superiority of our model and effectiveness of critical algorithmic designs.

Limitations

Our MMVTS model leverages the vision modal information encoded by the visual encoder. Considering the computational complexity, we keep the visual encoder frozen while keeping all other parameters trainable. This particular design may result in suboptimal utilization of the extensive multimodal information inherent in the dataset. We plan to further investigate different audio features in multimodal fusion (that is, V+T+A) as well as directly employing visual and audio cues (that is V+A since the current text modality is just ASR 1-best of the audio), to improve the VTS performance. We will continue enhancing the integration of general pre-trained multimodal models and large language models, which may offer a more holistic and effective exploration of multimodal information for VTS. Additionally, we will conduct more explorations and investigate approaches to make Co-Attn with MoE more stable in future work.

References

Mohammed Mahmood Ali, Mohammad S Qaseem, and Altaf Hussain. 2021. Segmenting lecture video into partitions by analyzing the contents of video. In *2021 International Conference on Data Analytics for Business and Industry (ICDABI)*, pages 191–196. IEEE.

Ido Amos, Jonathan Berant, and Ankit Gupta. 2023. Never train from scratch: Fair comparison of long-sequence models requires data-driven priors. In *The Twelfth International Conference on Learning Representations*.

Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A Gers, and Alexander Löser. 2019. Sector: A neural model for coherent topic segmentation and classification. *Transactions of the Association for Computational Linguistics*, 7:169–184.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Dipesh Chand and Hasan Oğul. 2021. A framework for lecture video segmentation from extracted speech content. In *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 000299–000304. IEEE.

Ananda Das and Partha Pratim Das. 2019. Automatic semantic segmentation and annotation of mooc lecture videos. In *International Conference on Asian Digital Libraries*, pages 181–188. Springer.

Richard A Davis, Keh-Shin Lii, and Dimitris N Politis. 2011. Remarks on some nonparametric estimates of a density function. *Selected Works of Murray Rosenblatt*, pages 95–100.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. 2022. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176.

Ankit Gandhi, Arijit Biswas, and Om Deshmukh. 2015. Topic transition in educational videos using visually salient words. *International Educational Data Mining Society*.

Anchit Gupta, CV Jawahar, Makarand Tapaswi, et al. 2023. Unsupervised audio-visual lecture segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5232–5241.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Marti A Hearst. 1994. Multi-paragraph segmentation expository text. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 9–16.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. Movienet: A holistic dataset for movie understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 709–727. Springer.

Md Mohaiminul Islam, Mahmudul Hasan, Kishan Shamsundar Athrey, Tony Braskich, and Gedas Bertasius. 2023. Efficient movie scene detection using state-space transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18749–18758.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 469–473.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020.

Gshard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*.

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

Michal Lukasik, Boris Dadachev, Kishore Papineni, and Gonçalo Simões. 2020. Text segmentation by cross segment attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4707–4716.

Jonghwan Mun, Minchul Shin, Gunsoo Han, Sangho Lee, Seongsu Ha, Joonseok Lee, and Eun-Sol Kim. 2022. Bassl: Boundary-aware self-supervised learning for video scene segmentation. In *Proceedings of the Asian Conference on Computer Vision*, pages 4027–4043.

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. 2022. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*, 34:14200–14213.

Thong Nguyen, Yi Bin, Junbin Xiao, Leigang Qu, Yicong Li, Jay Zhangjie Wu, Cong-Duy Nguyen, See-Kiong Ng, and Luu Anh Tuan. 2024. Video-language understanding: A survey from model architecture, model training, and data perspectives. *arXiv e-prints*, pages arXiv–2406.

Jielin Qiu, Franck Dernoncourt, Trung Bui, Zhaowen Wang, Ding Zhao, and Hailin Jin. 2023. Liveseg: Unsupervised multimodal temporal segmentation of long livestream videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5188–5198.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Martin Riedl and Chris Biemann. 2012. Topictiling: a text segmentation algorithm based on lda. In *Proceedings of ACL 2012 student research workshop*, pages 37–42.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2016. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*.

Mike Zheng Shou, Stan Weixian Lei, Weiyao Wang, Deepti Ghadiyaram, and Matt Feiszli. 2021. Generic event boundary detection: A benchmark for event segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8075–8084.

Eduardo R Soares and Eduardo Barrére. 2018a. Automatic topic segmentation for video lectures using low and high-level audio features. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*, pages 189–196.

Eduardo R Soares and Eduardo Barrére. 2018b. A framework for automatic topic segmentation in video lectures. In *Anais Estendidos do XXIV Simpósio Brasileiro de Sistemas Multimídia e Web*, pages 31–36. SBC.

Eduardo R Soares and Eduardo Barrére. 2019. An optimization model for temporal video lecture segmentation using word2vec and acoustic features. In *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web*, pages 513–520.

Alessandro Solbiati, Kevin Heffernan, Georgios Damaskinos, Shivani Poddar, Shubham Modi, and Jacques Cali. 2021. Unsupervised topic segmentation of meetings with bert embeddings. *arXiv preprint arXiv:2106.12978*.

Swapna Somasundaran et al. 2020. Two-level transformer and auxiliary coherence modeling for improved text segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7797–7804.

Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. 2019. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216.

Yunbin Tu, Liang Li, Li Su, Shengxiang Gao, Chenggang Yan, Zheng-Jun Zha, Zhengtao Yu, and Qingming Huang. 2022. I 2 transformer: Intra-and inter-relation embedding transformer for tv show captioning. *IEEE Transactions on Image Processing*, 31:3565–3577.

Yunbin Tu, Liang Li, Li Su, Zheng-Jun Zha, Chenggang Yan, and Qingming Huang. 2023. Self-supervised cross-view representation reconstruction for change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2805–2815.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Liang Wang, Sujian Li, Yajuan Lü, and Houfeng Wang. 2017. Learning to rank semantic coherence for topic

segmentation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1340–1344.

Yuxuan Wang, Zilong Zheng, Xueliang Zhao, Jinpeng Li, Yueqian Wang, and Dongyan Zhao. 2023. Vstar: A video-grounded dialogue dataset for situated semantic understanding with scene and topic transitions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5036–5048.

Haoqian Wu, Keyu Chen, Haozhe Liu, Mingchen Zhuge, Bing Li, Ruizhi Qiao, Xiujun Shu, Bei Gan, Liangsheng Xu, Bo Ren, et al. 2023. Newsnet: A novel dataset for hierarchical temporal segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10669–10680.

Linzi Xing and Giuseppe Carenini. 2021. Improving unsupervised dialogue topic segmentation with utterance-pair coherence scoring. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 167–177.

Linzi Xing, Brad Hackinen, Giuseppe Carenini, and Francesco Trebbi. 2020. Improving context modeling in neural topic segmentation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 626–636.

Linzi Xing, Quan Tran, Fabian Caba, Franck Dernoncourt, Seunghyun Yoon, Zhaowen Wang, Trung Bui, and Giuseppe Carenini. 2024. Multi-modal video topic segmentation with dual-contrastive domain adaptation. In *International Conference on Multimedia Modeling*, pages 410–424. Springer.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800.

Peng Xu, Xiatian Zhu, and David A Clifton. 2023. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Shen Yan, Xuehan Xiong, Arsha Nagrani, Anurag Arnab, Zhonghao Wang, Weina Ge, David Ross, and Cordelia Schmid. 2023. Unloc: A unified framework for video localization tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13623–13633.

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. 2023a. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10714–10726.

Cheng-Fu Yang, Yao-Hung Hubert Tsai, Wan-Cyuan Fan, Russ R Salakhutdinov, Louis-Philippe Morency, and Frank Wang. 2022. Paraphrasing is all you need for novel object captioning. *Advances in Neural Information Processing Systems*, 35:6492–6504.

Ziyi Yang, Yuwei Fang, Chenguang Zhu, Reid Pryzant, Dongdong Chen, Yu Shi, Yichong Xu, Yao Qian, Mei Gao, Yi-Ling Chen, et al. 2023b. i-code: An integrative and composable multimodal learning framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10880–10890.

Hai Yu, Chong Deng, Qinglin Zhang, Jiaqing Liu, Qian Chen, and Wen Wang. 2023. Improving long document topic segmentation models with enhanced coherence modeling. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5592–5605.

Miaoran Zhang, Marius Mosbach, David Adelani, Michael Hedderich, and Dietrich Klakow. 2022. Mcse: Multimodal contrastive learning of sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5959–5969.

Qinglin Zhang, Qian Chen, Yali Li, Jiaqing Liu, and Wen Wang. 2021. Sequence model with self-adaptive sliding window for efficient spoken document segmentation. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 411–418. IEEE.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*.

Luowei Zhou, Chenliang Xu, and Jason Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Heqing Zou, Tianze Luo, Guiyang Xie, Fengmao Lv, Guangcong Wang, Juanyang Chen, Zhuochen Wang, Hansheng Zhang, Huaijian Zhang, et al. 2024. From seconds to hours: Reviewing multimodal large language models on comprehensive long video understanding. *arXiv preprint arXiv:2409.18938*.

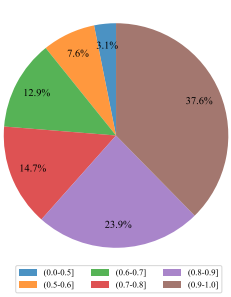
A The CLVTS Dataset

A.1 Data Collection and Annotation

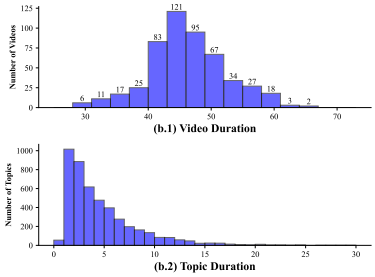
The CLVTS dataset is primarily sourced from educational videos, in the form of lectures, from Public Video Platforms[4][5]. Specifically, videos are first

[4] https://study.163.com/

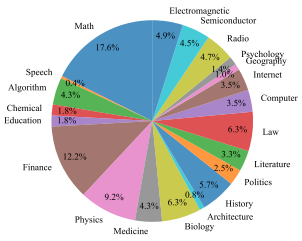
[5] https://www.bilibili.com/v/knowledge



(a) Distribution of Consistency Scores.



(b) Video Duration and Topic Duration (in minutes).



(c) Distribution of Subjects of Lectures.

Figure 2: Statistics of our CLVTS dataset.

transcribed by a competitive Automatic Speech Recognition (ASR) system[6]. We then ask annotators to combine visual and textual (ASR 1-best) information to mark the timestamp (in seconds) at the end of each topic. We ensure high accuracy and reliability of annotations from three aspects, including **annotator training**, **hierarchical topic labeling**, and **a multi-annotator strategy**.

Firstly, before the actual annotation process starts, the annotators take two rounds of training. Each annotator needs to annotate 5 videos in each round; at the end of each round, we assess the annotation quality, provide feedback, and ensure that the annotators address the issues and understand the annotation guideline clearly, at the end of training.

Secondly, during annotation, to help the annotators thoroughly understand the lecture content, we ask the annotators to annotate topics hierarchically, that is, they label both coarse-grained topic (large topic) and fine-grained topic (small topic) boundaries while the large topic boundaries are a subset of the small topic boundaries. We take the **small** topic boundaries as the final topic boundary labels for supervised VTS modeling.

Thirdly, we employ a multi-annotator strategy. All data is annotated in batches, with two annotators annotating each sample independently. The third annotator reviews the annotations by the first two annotators, rectifies errors, and provides the final annotations. After the annotations of each batch, we randomly select 5% of a batch for quality assessment. If the unacceptable rate (the proportion of the wrongly annotated topic boundaries) is lower than 10%, the data are deemed satisfactory and accepted; otherwise, after communicating quality assessment results and possible reasons for errors, the annotators are requested to re-annotate the batch based on the feedback. This quality control process is repeated until the unacceptable rate is lower than 10% for all batches. In this work, we finished quality control of all data within 3 iterations.

[6] https://tingwu.aliyun.com/home

A.2 Dataset Analysis

To evaluate the inter-annotator agreement on VTS annotations on the CLVTS dataset, following (Shou et al., 2021), we compute the $F_1@k$ score (defined in Appendix C) based on the absolute distance between two topic boundary sequences, varying the threshold k from 0 to 8 seconds with a step size of 2 seconds, where 8 seconds are approximately the average duration of a video clip. By averaging the F_1 scores across all three pairs of annotated topic boundaries from three annotators on the same video, we obtain the **consistency score**. The more similar the annotations from all annotators on the same video are, the higher the consistency score is. Figure 2a shows that the consistency scores of the majority of videos in our CLVTS dataset exceed 0.5, indicating a decent degree of consensus for VTS annotations (Shou et al., 2021).

Table 1 compares our CLVTS dataset against existing VTS datasets. Both AVLecture and CLVTS are sourced from educational contexts, where VTS significantly enhances learning experiences. Table 1 highlights that CLVTS features a higher average number of topics per video. Among annotated videos in CLVTS, 47% are presentations showing slides, 34% are blackboard demonstrations, and 19% are miscellaneous types. **We also collect 1027 hours of unlabeled videos from the same sources for pre-training.** Figure 2b and 2c show a diverse distribution of video durations and topic durations and a broad spectrum of subjects in labeled CLVTS.

Model	Numbers
LongFormer$_{cssl}$	130M
MMVTS Model (Ours)	
Baseline$_{1,2,3}$	154M
Merge-Attn	161M
Co-Attn	173M
Merge-Attn with MoE	175M
Co-Attn with MoE	192M

Table 5: The number of trainable parameters of the baseline LongFormer$_{cssl}$ and variants of our MMVTS models. The model names conform to the model names in Table 2.

A.3 Ethical Considerations

The dataset used in this research is strictly for academic and non-commercial purposes. We implemented several measures to ensure compliance with ethical standards, as follows.

- **Data Transparency and Anonymization.** We only provide ASR transcripts after rigorous text anonymization processes, visual features of video clips, our annotations, and links to the original videos, to ensures transparency regarding the data sources and their usage while maintaining anonymity.
- **Data Access Compliance.** To further ensure ethical use of the dataset, we require researchers to contact us via emails to confirm their compliance with ethical guidelines and the conditions outlined in our data usage declaration, before granting them access to the dataset. This procedure includes ensuring that they are aware of and adhere to the Personal Information Protection Law (PIPL) and any relevant legal frameworks regarding personal data usage.
- **Authorization.** Any personal data should be used only with express authorization, ensuring lawful and fair processing in accordance with applicable laws.

B Implementation Details

We partition the labeled data within AVLecture and CLVTS into 70% for training, 10% for validation, and 20% for testing, respectively. The unlabeled data of AVLectures and CLVTS are used for pre-training for each dataset, respectively.

Our experiments are implemented with the *transformers* package[7]. We use the same maximum sequence length 2048 as in Yu et al. (2023) for a fair comparison with the text-only models. **All results are the mean values over three runs with different random seeds.**

For video with the number of clips greater than the max sequence length, we use sliding window and take the last clip of the prior sample as the first clip of the next sample. All supervised models use a threshold strategy, where clips with scores above a threshold 0.5 are predicted as topic boundaries.

Following (Gupta et al., 2023), for each video clip, we extract three visual feature types: *OCR*, *2D* and *3D*. Specifically, the *OCR* features are derived by encoding the textual output obtained from the *OCR* API[8] of the clip's central image. Encoding the textual output from OCR is performed using the BERT-based sentence transformer model[9], where *all-mpnet-base-v2* and *paraphrase-multilingual-mpnet-base-v2* models are employed for experiments on the English and Chinese datasets, respectively. The *3D* features are extracted using the same video feature extraction pipeline as in Gupta et al. (2023). The *2D* features are extracted by sampling three frames from each clip, subsequently encoding these frames with visual encoder, and applying max pooling. Specifically, we choose the visual encoder of CLIP (Radford et al., 2021), which is pre-trained to predict if an image and a text snippet are paired together. The images from AVLecture and CLVTS are processed to extract 2D features using CLIP$_{ViT-B/16}$[10] and CN-CLIP$_{ViT-B/16}$[11], respectively.

After extracting the visual features, we concatenate them as shown in Eq. 16 to get v_i, which will then be fed into the following projection layer. During pre-training and fine-tuning, the parameters of visual encoders are kept frozen. The learning rate is $5e-5$ and dropout probability is 0.1. AdamW (Loshchilov and Hutter, 2017) is used for optimization. The batch size is 8 and the epoch for pre-training and fine-tuning is 1 and 5, respectively. The loss weight α and γ for l_{cma} is 0.5, β and σ for $l_{balance}$ is 1.0 when MoE is in Multimodal Fusion Layers, θ for l_{mcssl} is 0.5. k_1 and k_2 of Eq. 14 are 1 and 3, following Yu et al. (2023). We comprehensively compare different types of fusion structure

[7] https://github.com/huggingface/transformers

[8] https://help.aliyun.com/zh/viapi/developer-reference/api-sy75xq

[9] https://www.sbert.net/docs/sentence_transformer/pretrained_models.html

[10] https://github.com/openai/CLIP

[11] https://github.com/OFA-Sys/Chinese-CLIP

Model	F_1	*BS@30*	F_1*@30*	*mIoU*	*Avg*
Co-Attn with MoE	59.84	**75.62**	67.69	**72.21**	**68.84**
w/o l_{vts}	57.19	74.64	66.04	70.91	67.20
w/o l_{cma}	**60.23**	73.54	**67.86**	72.12	68.44

Table 6: Ablation experiments of pre-training task on AVLecture. The model parameters derived from these distinct pre-training tasks served as the initial parameters for subsequent fine-tuning of the model. Additionally, the coherence modeling tasks are incorporated during the fine-tuning phase.

using one Multimodal Fusion Layer, then use the best performing fusion structure for the remaining experiments. We also investigate the impact of different numbers of Multimodal Fusion Layers in Figure 3. As to the MoE layer in Multimodal Fusion Layers, we choose 4 candidate experts and activate 2 experts for each input feature. The intermediate size of expert is 3072. The total number of trainable parameters is shown in Table 5.

$$v_i = v_i^{2d}; v_i^{3d}; v_i^{ocr} \tag{16}$$

C Evaluation Metrics

$\boldsymbol{F_1}$ is a metric used to evaluate the accuracy of text topic segmentation (Lukasik et al., 2020; Zhang et al., 2021). It focuses on the performance of **exact matching** of the positive class and balances the precision and recall rates.

BS@k (Gupta et al., 2023) is the average number of predicted boundaries matching with the ground truth boundaries within a k-second interval, which can be considered as the recall rate based on **fuzzy matching**.

$\boldsymbol{F_1}$***@k*** denotes the F_1 score calculated based on matching predicted boundaries and ground truth boundaries within k seconds. Considering subjectivity and uncertainty in VTS annotations, we introduce $F_1@k$ as a supplement to $BS@k$ to enabling a more comprehensive assessment of model performance in dealing with ambiguous (uncertain) boundaries.

mIoU is commonly used in video action segmentation (Zhou et al., 2018) and video scene segmentation (Mun et al., 2022). While the F_1, $BS@k$ and $F_1@k$ metrics focus on the accuracy of positive predictions (either exact match or fuzzy match), the *mIoU* metric measures the overlapping area between predicted segments and ground truth segments, hence providing a generalized assessment of how well the model's predicted segments match the actual segments on the segmentation task.

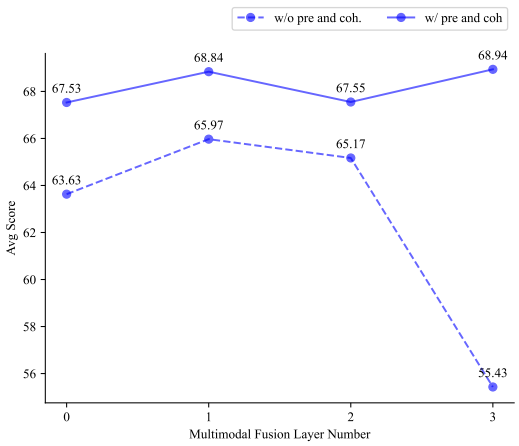


Figure 3: Performance (**Avg**) of fine-tuning our MMVTS model on AVLecture test set, w/o pre-training and multimodal coherence modeling tasks (denoted by w/o pre and coh) and w/ pre-training and multimodal coherence modeling tasks (denoted by w/ pre and coh), with different numbers of Multimodal Fusion Layers, where the fusion structure is Co-Attention with MoE.

Our implementation of *BS@30* draws upon the code published by Gupta et al. (2023)[12], while the approach to implement *mIoU* is guided by Mun et al. (2022)[13]. We have relied on the *scikit-learn* package[14] to compute F_1, following the implementation by Yu et al. (2023), to ensure fair comparisons. The definitions provided in the aforementioned sources also inform our implementation of $F_1@k$.

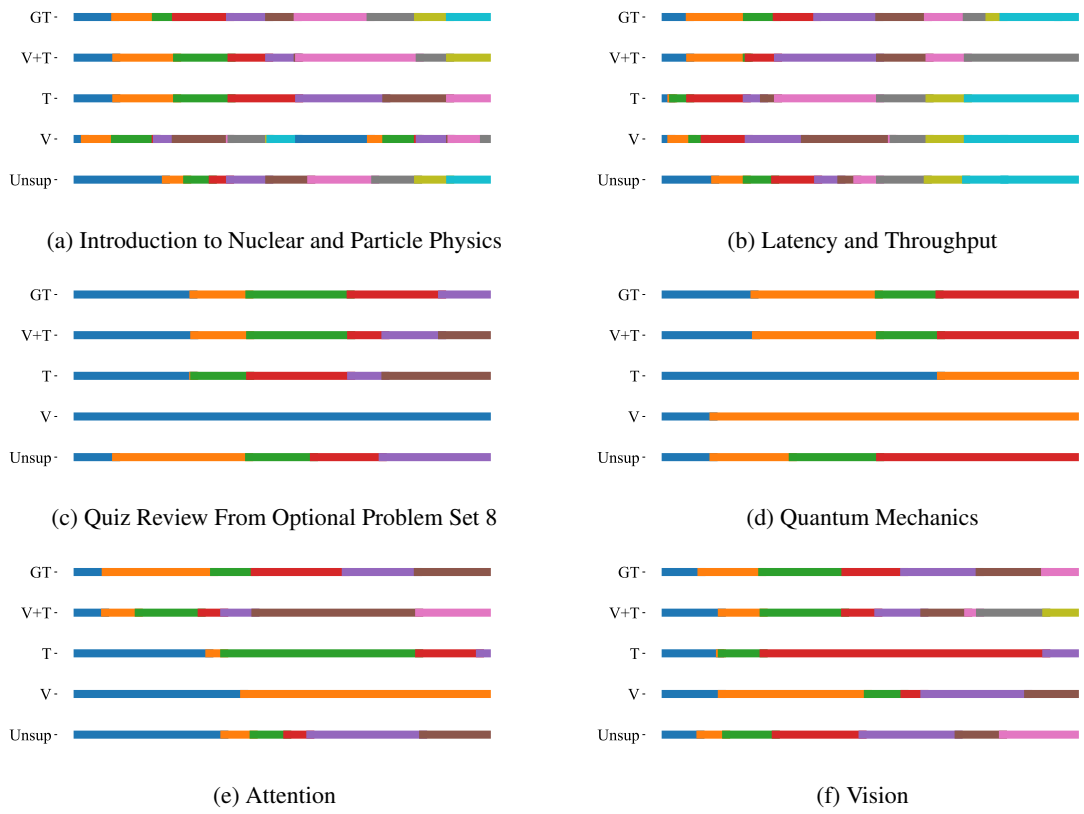
D Artifact Use Consistent With Intended Use

All of our use of existing artifacts is consistent with their intended use, provided that it was specified. For the CLVTS data set we created, its license will be for research purpose only.

E Fine-tune Llama-3-8B for Video Topic Segmentation

Considering the computational complexity and the data volume, we employ LoRA (Hu et al., 2021) for fine-tuning Llama-3-8B[15] on the training set of each AVLecture and CLVTS datasets, with a maximum sequence length of 2048. With LoRA, the number of trainable parameters is 3 million. Our training configuration includes a batch size of 32

[12] https://github.com/Darshansingh11/AVLectures/
[13] https://github.com/kakaobrain/bassl
[14] https://scikit-learn.org/stable/
[15] https://huggingface.co/meta-llama/Meta-Llama-3-8B



(a) Introduction to Nuclear and Particle Physics

(b) Latency and Throughput

(c) Quiz Review From Optional Problem Set 8

(d) Quantum Mechanics

(e) Attention

(f) Vision

Figure 4: Video Topic Segmentation examples for six lecture videos from AVLecture. (a)-(b) are of the **Slides** mode, (c)-(d) are **Blackboard** mode and (e)-(f) are **Mixed** mode. **GT** denotes ground truth, **V** refers to the visual-only baseline BaSSL which only uses visual features, **T** denotes the text-only LongFormer which only relies on the text modality, and **V+T** signifies our MMVTS model that integrates information from both visual and text modalities. **Unsup** denotes the unsupervised baseline UnsupAVLS, which also integrates visual and text features.

and a total of 6 epochs, utilizing a learning rate of $5e-5$ and *cosine* scheduler (Loshchilov and Hutter, 2016). Table 7 shows two prompts we used to fine-tuning Llama-3-8B on the text modality of VTS data. We first evaluate the *Generative* type prompt that Yu et al. (2023) has experimented with, due to its better zero-shot and one-shot text topic segmentation performance than the *Discriminative* type. However, compared with Longformer which also uses the textual modality, the *Avg* score of fine-tuning Llama-3-8B with the *Generative* prompting strategy (denoted by Llama-3-8B$_{Generative}$) is 8.3 and 6.27 absolutely worse than those from Longformer on AVLecture and CLVTS, respectively. We suspect that this is due to the inherent issue of sparse labels in binary classification for VTS, posing challenges to applying the large language model (LLM) Llama-3-8B in a generative manner. Inspired by this hypothesis, we refine the *Discriminative* prompt into the *Discrete* prompt for Llama-3-8B, as shown in Table 7. The results in Table 2 show that on **Avg**, Llama-3-8B$_{Discrete}$ consistently and markedly outperforms Llama-3-8B$_{Generative}$ on both AVLecture and CLVTS, by 6.04 and 10.8 absolute points, respectively. These results suggests that the *Discrete* prompt is more suitable for prompting Llama-3-8B for the VTS task than the Generative prompt.

However, with either Generative or Discrete prompt, Llama-3-8B$_{Discrete}$ does not provide consistent advantages over the text-only small model Longformer. Table 2 shows that F_1 results of Llama-3-8B$_{Discrete}$ on AVLecture and CLVTS are much worse than those from Longformer (we find that the prediction boundary of Llama-3-8B usually has a clip offset). The *Avg* score of Llama-3-8B$_{Discrete}$ on AVLecture is 2.26 absolute worse than LongFormer, although Llama-3-8B$_{Discrete}$ achieves the best *Avg* score on CLVTS, surpassing LongFormer by 4.53 absolute. Future work could further explore how to better use LLMs for the VTS task.

Type	Prompts for text topic segmentation.
Generative	Please identify several topic boundaries for the following document and each topic consists of several consecutive utterances. please output in the form of {topic i:[], ... ,topic j:[]} with json format, where the elements in the list are the index of the consecutive utterances within the topic, and output even if there is only one topic. document: [0]: s_1 [1]: s_2 ... [$n-1$]: s_n Please give the result directly in json format: output: {"topic_0": [0, 1, 2, . . . , k-1], "topic_1": [k, k+1, . . .], . . . }
Discrete	Please identify several topic boundaries for the following document. please output in the form of {topic_segment_ids:[xxx]} with json format, where the elements in the list are the index of the last sentence of every topic, if there is only one topic then the array is empty. document: [0]: s_1 [1]: s_2 ... [$n-1$]: s_n Please give the result directly in json format: output: {"topic_segment_ids": [x, x, x]}

Table 7: Our designed prompts for fine-tuning Llama-3-8B on the text data for video topic segmentation. n denotes the number of sentences and s_i denotes the i-th sentence in the document.

F More Analysis of the Baseline Results

Considering the baseline results in Table 2, the only seemingly similar scores between the unsupervised method and supervised baselines are mIoU scores, which are attributed to the leakage of the ground-truth topic number. If we use the topK probability to determine predictions, where K for each sample is the ground-truth topic number, mIoU and BS@30 of $SWST_{seq}$ (Row 7) will be 4.4 and 19.05 points higher than UnsupAVLS, respectively. However, this evaluation is not reasonable, as models should not disclose the ground-truth topic number during testing. Therefore, we opted for the threshold-based evaluation strategy commonly used in classification tasks

G Performance of Multimodal Fusion Layers with Varied Numbers of Layers

Using Co-Attention with MoE as the architecture of Multimodal Fusion Layers, we investigate the *Avg* performance of models featuring various numbers of Multimodal Fusion Layers (MFLs) on the AVLecture data set, as depicted in Figure 3. We find that directly fine-tuning our MMVTS model, without pre-training nor the two auxiliary tasks for coherence modeling, a single MFL yields the best performance, surpassing the no-layer configuration by 2.34 points. Adding more layers leads to degraded performance and training instability, particularly noticeable with three layers.

By incorporating the pre-training phase followed by fine-tuning with coherence modeling tasks, Avg performance enhancements of 4.0, 2.87, 2.37, and 13.51 points for 0, 1, 2, 3 MFL layers are observed, respectively. These results clearly demonstrate the substantial benefits from our pre-training and coherence modeling strategies in boosting the model's performance. Notably, our pre-training and coherence modeling reaches the convergence of the model with three MFL layers, achieving results that marginally exceed the performance of a single-layer model by 0.1 points.

Model	Modality	AVLecture	CLVTS
Longformer	T	62.52	46.81
Llama-3-8B$_{Discrete}$	T	60.26	51.34
MMVTS Model (Ours)	V+T	68.61	51.27
	V+T+A$_1$	67.49	51.98
	V+T+A$_2$	**69.36**	**52.97**

Table 8: The **Avg** score of integration of audio information into the MMVTS model with Co-Attn and MoE architecture, on AVLecture and CLVTS test sets. V+T+A$_1$ notes that during fine-tuning phase, the audio features are incorporated into the cross-modality alignment loss l_{cma}, while V+T+A$_2$ does not involve audio features in l_{cma}.

H Qualitative Analysis of Video Topic Segmentation Examples

We present the topic-segmented outputs for six lecture videos from three presentation modes including **slides**, **blackboard**, and **mixed**, in Figure 4. Relying solely on the visual modality, segmentation points are predominantly identified through the superficial cues associated with visual transitions. In contrast, using only the textual modality, topic boundaries are discerned based on semantic content; however, this approach has limitations, such as missing topic boundaries or the accumulation of topics. The integration of the visual modality with the text modality offers complementarity of information, thereby improving the overall VTS performance. The case studies in Figure 4 help illustrate the importance of multimodal fusion for the VTS task regardless of the video presentation modes.

I Integration of Audio Information

We use a pre-trained Whisper-small model[16] with 244 million parameters and use the Whisper encoder to extract audio features for video clips (the parameters of the Whisper audio encoder are frozen). Similar to visual features, we perform maximum pooling on the encoded audio features for each clip to obtain the clip-level audio features. Then, we feed the audio features into a learnable projection layer and then through the Multimodal Fusion Layers, and then concatenate them with visual features and text features for the final topic boundary prediction.

We initialize the parameters from pre-trained visual and textual MMVTS model with Co-Attn and MoE. During fine-tuning, we compare two configurations. The first configuration is denoted by V+T+A_1, which incorporates the audio feature into the cross-modality alignment loss l_{cma}. The second configuration is denoted by V+T+A_2, which does not involve audio features in l_{cma}. The pairwise cross-modality alignment loss weight is set to 0.33 in V+T+A_1.

The results are shown in Table 8. As can be seen from the table, integration of audio features by V+T+A_1 improves the *Avg* score by 0.71 on the CLVTS test set while decreasing the *Avg* socre by 1.12 on the AVLecture test set. However, we find that excluding audio from the l_{cma} (V+T+A_2) results in an absolute improvement of 0.75 and 1.7 in the *Avg* score on AVLecture and CLVTS test sets, respectively, which suggests that audio features can contribute to more accurate topic boundary prediction under certain conditions, but their role in modality alignment needs to be treated with caution and demands further exploration. We sample some videos and observe that the cosine similarity between audio features of adjacent clips shows relatively small differences, ranging from 0.97 to 0.98, while visual features between adjacent clips exhibit larger differences, ranging from 0.88 to 0.97. This disparity might complicate the task of simultaneously aligning text, visual, and pooled clip-level audio features. Future research could explore more nuanced integration of audio features to provide supplementary paralinguistic information, such as pitch, energy, and pause duration. Alternatively, direct use of audio and visual information for VTS may bypass the ASR step altogether (as ASR is used to obtain the text modality).

[16] https://github.com/openai/whisper/