Archives available at journals.mriindia.com

International Journal on Advanced Electrical and Computer Engineering

ISSN: 2349-9338

Volume 15 Issue 01s, 2026

Educational Video Time-Stamped based on Texttiling with Sentence-BERT Embeddings

[1]Ahlam Enan, [2]Muneer Alsurori, [3]Akram Alsubari

[1 2 3]CS & IT Department, Ibb University, Ibb, Yemen

Email: [1]ahlam.enan@ibbuniv.edu.ye, [2]muneer.alsurory@ibbuniv.edu.ye, [3]akram.alsubari@ibbuniv.edu.ye

Peer Review Information

Submission: 05 Dec 2025

Revision: 25 Dec 2025

Acceptance: 10 Jan 2026

Keywords

Video segmentation, educational video analysis, Semantic TextTiling, BERT embeddings, Time-stamped chapter detection.

Abstract

Automatically segmenting long educational videos into coherent thematic units is essential for improving content navigation, enabling targeted review, and supporting downstream applications such as interactive video indexing and summarization. In this work, we propose an unsupervised, transcript-based segmentation method that enhances the classical TextTiling algorithm by incorporating contextual Sentence-BERT embeddings to detect topic transitions as drops in semantic coherence between consecutive segments, without requiring labeled training data. To evaluate our approach, we introduce EVTS (Educational Video Timestamp Segmentation), a large-scale dataset of 1,553 real-world educational YouTube videos, each annotated with creator-provided timestamps marking the start of distinct instructional segments. Experimental results show that our method aligns well with these human-defined boundaries, achieving an F1 score of 0.665, a precision of 0.735, and a Pk of 0.365 under an adaptive configuration. These findings indicate that the proposed approach effectively captures meaningful topic shifts and produces accurate, time-stamped segments suitable for real-world educational applications, even when evaluated against noisy, non-expert annotations.

1. Introduction

Video has become one of the most popular forms of online learning materials, serving as a rich media for delivering lectures, tutorials, and demonstrations. Through sites such as YouTube and Udemy, among others, educational videos have become increasingly available, allowing learners to view content anywhere, anytime. [1,2] This availability has led to the increasing use of video-based learning in both formal and informal learning environments. However, the rapid growth of video content has posed new challenges in educational contexts. As the number of lecture videos continues to grow, navigating and retrieving relevant information from them has become increasingly complex. Even with their widespread use, it remains a barrier to navigate through them efficiently. Unlike television shows or news stories, educational video clips often have subtle visual cuts, and explicit visual cues do not always initiate a change in meaning between topics. As a result, students may need to spend a long time watching entire videos to find particular information, reducing the meaning of their learning. Instructional lecture videos are more likely to cover a wide range of topics. Users may not be interested in all of this content, but rather in specific issues [3,4]. Therefore, automatically segmenting videos based on topical changes is promising for many downstream applications, such as Information Retrieval [5], table-of-contents generation [6], summarization and chaptering [7], and others. [8,9] A segmented video lecture can

enhance learning across learning platforms by providing students with structured video content so they can navigate instantly from one topic to another. Whenever they choose, at their own learning pace [10].

Despite the advantages of segmenting video lectures into topics, it is not easy. Human segmentation has high accuracy, but the time required to perform this task manually is almost impractical, especially for long lectures [11]. Although previous studies in video segmentation have used visual or audio features for entertainment or monitoring, this approach is not suitable for lecture videos, which exhibit limited visual variety and low semantic contrast in the written text. Recent advances in natural language processing (NLP) and semantic embedding models, such as Sentence-BERT [12], have opened new possibilities for interpreting the underlying form of spoken knowledge in videos. By combining text-based coherence methods such as TextTiling [13] with deep sentence embeddings, it is now possible to detect topic boundaries more accurately and meaningfully. In this work, we propose an educational video segmentation framework that leverages TextTiling with Sentence-BERT embeddings to identify semantic topic boundaries within lecture transcripts. Our approach aims to enhance the structural understanding of educational videos to facilitate applications such as automatic indexing, retrieval, and summarization. Furthermore, in this work, EVTS (Educational Video Timestamp Segmentation) was introduced, a novel dataset specialized in the Information Technology (IT) domain. This domain was deliberately selected due to the prevalence of long-form, topic-dense technical lectures on platforms such as YouTube—covering subjects like networking, cybersecurity, and data structures—where instructors frequently provide manual timestamps to structure their content. Such naturally occurring temporal annotations, combined with the conceptual complexity and hierarchical organization of IT lectures, offer a realistic and challenging setting for evaluating semantic video segmentation approaches. EVTS comprises synchronized audio-visual recordings, human-verified transcripts (VTT format), and fine-grained timestamps aligned with topical boundaries and segment titles, serving as a high-fidelity benchmark to advance research in educational video understanding.

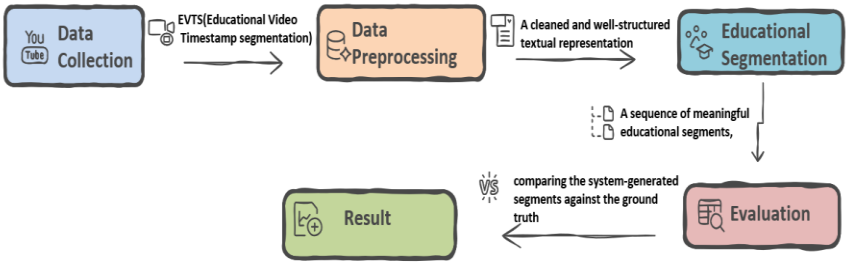

Fig. 1. Overview of the proposed unsupervised segmentation pipeline for educational videos.

As shown in Fig 1, the proposed pipeline consists of four main stages: (1) data collection, where educational videos are obtained from the YouTube platform, (2) preprocessing, which converts raw VTT files into clean, structured text, (3) educational segmentation, where our unsupervised model identifies topic boundaries, and (4) evaluation, which compares the system-generated segmentations with the segmentations generated by content creators.

The rest of this paper is organized as follows: Section 2 related work; Section 3 introduces the EVTS dataset, including the collection process, preprocessing steps, and data statistics; Section 4 explains the proposed segmentation method; Section 5 presents the results and discusses; and Section 6 concludes the paper with future directions.

2. Related Work

Video segmentation is a pivotal process in video analysis, enabling efficient browsing, indexing, and summarization of content. The research community has pursued four primary forms of video segmentation techniques: frame segmentation, audio segmentation, transcript segmentation, and multimedia segmentation. Research in this area has developed across four main approaches: frame-based segmentation, audio segmentation, text segmentation, and multimodal segmentation. However, frame-based segmentation approaches are often ineffective for educational videos, where visual changes are minimal and semantic shifts occur primarily through spoken discourse rather than visual cues. Therefore, this section will focus on research based on audio, text, and multimodal segmentation approaches to

identify the underlying thematic structure of lecture content.

2.1 Audio-based Segmentation

Early efforts relied on acoustic features to detect structural shifts. For instance, Delacourt et al. (2000) [14] used the Bayesian Information Criterion (BIC) to identify speaker changes from raw audio. Later, Soares and colleagues [15–17] demonstrated that combining acoustic cues (e.g., silence, pitch) with semantic information from ASR transcripts significantly improves boundary detection in educational settings. More recently, Adi et al. (2025) [18] leveraged Speech Language Models (SLMs) to model acoustic-semantic shifts (e.g., emotion, speaker role), achieving state-of-the-art unsupervised segmentation. While effective, these methods remain sensitive to audio quality and may overlook purely textual topic shifts.

2.2 Multimodel segmentation

To overcome the limitations of single-modality approaches, several studies fused visual, audio, and textual signals. Ghauri et al. (2020) [19] proposed a deep multimodal framework but found that fusion does not consistently outperform unimodal baselines. Others, such as Das et al. (2020) [20] and Gupta et al. (2023) [22], integrated slide content (OCR) with transcripts and audio, showing gains in MOOC-style lectures. Recent work by Yu et al. (2024) [23] and Vasuki et al. (2024) [24] further advanced this line of work by employing cross-attention and Mixture-of-Experts architectures. However, multimodal systems are computationally expensive, require synchronized modalities, and often depend on synthetic or small-scale datasets that lack real-world complexity. Overall,

These studies highlight the evolution of multimodal video segmentation, from early multimodal fusion approaches to large-scale datasets and advanced embedding-based models. While multimodal information can improve segmentation, transcript-based segmentation remains an efficient and scalable alternative. In the following section, we focus on transcript segmentation, leveraging semantic embeddings to detect topic boundaries in educational video lectures.

2.3 Transcript-based Segmentation

With the rise of accurate ASR, text-only segmentation has emerged as a lightweight yet powerful alternative. Early methods used noun phrases [25] or lexical similarity [26] but struggled with semantic continuity. The advent of contextual embeddings changed this landscape: Galanopoulos & Mezaris (2018) [29] used word embeddings for unsupervised segmentation, while Freisinger et al. (2023–2025) [31,32] introduced hierarchical relevance measures and LoRA-tuned LLMs for multilingual content. Retkowski & Waibel (2024) [30] and Gklezakos et al. (2024) [33] further improved scalability with real-time and hierarchical clustering systems. Despite these advances, a critical gap remains: most methods are evaluated on artificial, small-scale, or synthetically segmented datasets that do not reflect the nuanced topic transitions in real educational videos. Moreover, even unsupervised text-based approaches often lack fine-grained temporal alignment with human-defined boundaries.

Our work addresses this gap by introducing EVTS, a dataset of 1,570 real-world educational videos annotated with creator-provided timestamps that mark natural topic boundaries. Building on this resource, we enhance the classical TextTiling algorithm with BERT sentence embeddings to detect drops in semantic coherence, enabling accurate, unsupervised, and temporally aligned segmentation without labeled data.

3. Data Set

As part of this work, we present a new educational dataset for evaluating text segmentation systems on less-structured content and more-overlapping topics, focused on IT educational videos. The dataset comprises 1553 English-language YouTube videos, including their transcripts, audio, metadata, and timestamps. We processed the data to adapt it to the text segmentation task.

3.1 Collection

For research purposes, we developed a dataset (EVTS) to facilitate video text segmentation research in Information Technology (IT). We generated the dataset from YouTube by selecting learning videos on various IT topics and technologies. We downloaded the videos and corresponding transcripts, along with related metadata (timestamps and video metadata), using yt-dlp. The data included only videos in English with transcripts and timestamps, provided by the creators themselves. A set of domain-specific seed terms was selected to identify relevant content, including "Introduction to Information Technology," "Networking," "Cybersecurity," "IT Basics," "Data Structures and Algorithms," "Software Development," and "Database Management." The use of these terms was to identify educational and lecture-based content more relevant to the research. During the preliminary stage, YouTube search results were restricted to videos that included transcripts and timestamps, available and provided by the creators themselves. Subsequently, we conducted a manual evaluation and selected YouTube channels that met established quality criteria for speech clarity, transcript accuracy, and relevance to information technology education. As illustrated in Fig.2, our data collection pipeline follows a clear, sequential workflow

from keyword identification to final folder organization, ensuring high-quality, consistent, and research-ready data.

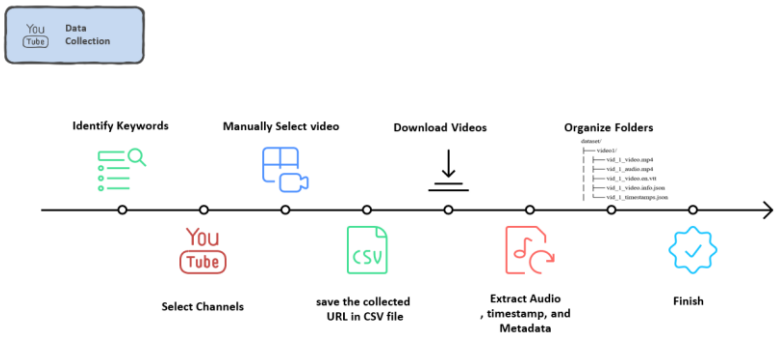

Fig. 2. Data scraping for the EVTS dataset.

3.2 Preprocessing

Educational videos often come with auto-generated subtitles in WebVTT format. While useful, these transcripts are messy and not ready for semantic analysis out of the box. They contain timing tags, repeated lines, and word-by-word fragments, making it hard to understand the actual flow of ideas. To fix this, we built a simple yet effective cleaning pipeline that converts raw captions into clean, time-aligned text segments, ready for topic segmentation. First, let's show what's wrong with raw WebVTT files: Markup clutter: Lines are filled with tags like <c>, <i>, or even word-level timestamps like <00:01:23.450>. These aren't part of the spoken content; they're just formatting artifacts. Repeated text: The same sentence often appears 2–3 times across consecutive captions (e.g., while the speaker is still talking). Word-level fragmentation: Instead of complete sentences, you get one word per line, each with its own timestamp. This breaks the natural flow of speech. No sentence structure: Spoken language is full of pauses, fillers ("um", "so"), and run-on phrases, so clear sentence boundaries rarely exist. Our cleaning approach. We handle these issues in three straightforward steps: Strip all tags and timestamps. We remove every <...> tag, whether it's for styling (<c>) or timing (<00:01:23.450>). Only the actual words remain. Reconstruct a clean word sequence with timing. We go through the original captions and record each word once, along with the time it first appeared. This gives us a clean, non-redundant stream of words, each tied to its start time, and groups words into short, coherent segments. Since individual words aren't meaningful on their own, we bundle every consecutive word into a brief phrase. The segment keeps the timestamp of its first word, so it stays aligned with the video.

The output is a list of clean textual segments, each with a precise start time.

All prepressing steps are shown in Figure 3.

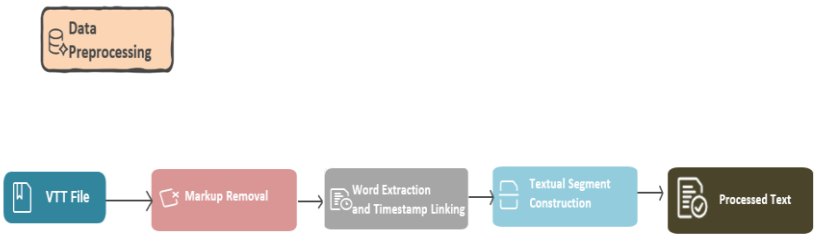

Fig. 3. Overview of the data preprocessing pipeline

3.3 Data statistics

The EVTS dataset comprises 1,553 educational videos collected from publicly available YouTube sources, amounting to 5,966,062 seconds (approximately 1,657.2 hours) of content. The average video duration is 3,812.2 seconds (about 63.5 minutes), with durations ranging from 103 seconds (1.7 minutes) to 42,004 seconds (11.7 hours), reflecting the natural variability of real-world educational lectures. The majority of videos (1,317, or 84.2%) are 30 minutes or shorter (≤ 1,800 seconds), while 248 videos (15.8%) ex-

ceed this duration. The dataset was initially compiled from over 6,000 candidate video URLs, filtered based on content relevance, availability of creator-provided transcripts and timestamps, and practical constraints such as storage capacity. Only videos that met these criteria were retained to ensure reliable ground truth for segmentation. Table 1 summarizes the overall dataset statistics, including total duration, video count, and the distribution of durations.

In terms of temporal annotations, EVTS contains 15,065 segments across all videos, corresponding to 15,065 topic boundaries. On average, each video includes 9.6 boundaries, ranging from 2 to 64. The average segment duration is 351.2 seconds (≈5.9 minutes), though there is considerable variation: 4,673 segments (31.0%) are shorter than 60 seconds, while 218 segments (1.4%) exceed 3,600 seconds (1 hour). The shortest segment is 0.0 seconds, and the longest reaches 14,598.0 seconds (243.3 minutes).

Table 2 provides a detailed breakdown of the ground-truth boundary and segment duration statistics.

Table 1. General statistics of the EVTS dataset

Characteristic	Value
Total Videos (The total count of videos in the dataset.)	1,553
Total Duration The cumulative duration of all videos, reported in seconds and hours.	5,966,062 seconds (1657.2 hours)
Average Duration The mean duration per video, expressed in seconds and minutes.	3812.2 seconds (63.5 minutes)
Longest Video The duration of the single longest video in the dataset.	42004 seconds (11.7 hours)
Shortest Video The duration of the shortest video in the dataset.	103 seconds (1.7 minutes)
Videos > 30 minutes The number and percentage of videos whose duration exceeds 30 minutes.	248 (15.8%)
Videos ≤ 30 minutes The number and percentage of videos with a duration of 30 minutes or less.	1317 (84.2%)

Table 2. Ground Truth Timestamps Statistics

Characteristic	Value
Number of Videos with Timestamps The overall count of temporal segments across all videos.	1,553
Total Segments The overall count of segments in the entire dataset.	15065
Avg Segments per Video The mean number of segments per individual video.	9.6
Max Segments in a Video The highest number of segments observed in any single video.	64
Min Segments in a Video The lowest number of segments observed in any single video.	2

These statistics are based on vidX.info.json files, which exist for all videos, even if they don't have Timestamps.json.

4. Segmentation Method

To automatically identify topic boundaries in educational lecture transcripts, we adopt an unsupervised semantic segmentation approach that leverages TextTiling, enhanced with contextual sentence embeddings. Unlike traditional TextTiling, which relies on lexical co-occurrence, our method leverages BERT sentence embeddings to capture more profound semantic coherence between textual units. The core idea is that topic shifts correspond to drops in semantic similarity between consecutive text blocks. By sliding a window over the sequence of embedded sentences and computing cosine similarity between adjacent blocks, we detect local minima that indicate potential boundaries. This approach requires no labeled training data and operates directly on cleaned transcripts, making it scalable and domain-agnostic.

The whole procedure is formalized in Algorithm 1.

Algorithm 1: BERT-Enhanced TextTiling for Topic Segmentation

Input:
A sequence of cleaned sentences
Pre-trained sentence embedding model (e.g., all-MiniLM-L6-v2)
Window size (number of sentences per block)
Depth threshold (minimum depth to accept a boundary)
Output:

A list of boundary indices, where each

```
E←[M(s1),M(s2),...,M(sn)]          Transform sentences into dense semantic vectors.
blocks←[mean(E[i:i+w]) for i in 0,w,2w,...]        Group embeddings into
                                                   non-overlapping blocks of size
sims←[cosine_sim(blocks[j], blocks[j+1]) for j=0 to |blocks|−2]
                          Measure the semantic similarity between consecutive blocks
B←∅
for i = 1 to |sims| − 2 do
  if sims[i] < sims[i−1] and sims[i] ≤ sims[i+1] then
    left_peak ← max(sims[0:i])
    right_peak ← max(sims[i+1:])
    depth ← (left_peak − sims[i]) + (right_peak − sims[i])
    if depth ≥ τ then
      b ← (i+1) × w
      if b < n then
        B. append(b)
return B
```

Algorithm 1 outlines our unsupervised topic segmentation approach. It begins by converting each cleaned sentence into a dense semantic vector using a pre-trained Sentence-BERT model. These vectors are then grouped into non-overlapping blocks, and cosine similarity is computed between consecutive blocks to measure semantic coherence. Topic boundaries are identified at local minima (valleys) in the similarity curve, but only if the drop in similarity is sufficiently deep, ensuring that minor fluctuations are ignored while meaningful topic shifts are captured. This method requires no labeled data and operates directly on the transcript, making it suitable for real-world educational videos where visual cues are minimal and topic transitions are often subtle.

5. Results and Discussion:

5.1 Ground Truth Alignment

The ground truth for each video consists of human-defined timestamps (in seconds) indicating the start of each thematic segment, as provided in the timestamps.json files. Since our segmentation model operates on textual units and outputs sentence indices, we first align these timestamps to the preprocessed transcript. Specifically, for each reference timestamp t, we identify the sentence whose start time is closest to t and record its index. This yields a list of reference boundary indices. The predicted boundaries from our model are already expressed as sentence indices. All evaluation metrics are then computed on these aligned indices, following the standard protocol of a ±3 sentence tolerance window [34].

5.2 Metrics

To assess the performance of our unsupervised segmentation method, we adopt standard metrics from the text segmentation literature [34]. All metrics are computed on sentence indices, with a tolerance window of ±3 sentences to account for minor alignment discrepancies. Precision is the proportion of predicted boundaries that correctly match a ground-truth boundary. Recall, the proportion of ground-truth boundaries that are successfully detected. F1 Score, the harmonic meaning of Precision and Recall, serving as the primary performance indicator Pk Metric: Estimates the probability that the system and ground truth disagree on whether two nearby sentences belong to the same segment.

$$Pk = \left(\frac{1}{(N - k)} \sum_{i=1}^{N-k} [1(ref\_seg\_i \neq pred\_seg\_i)]\right.$$

N: total number of sentences
K: sliding window size
1[
·]: $indicator\ function\ (1\ if\ the\ condition\ is\ true, 0\ otherwise)$

ref_{seg_i}: $indicator\ function$
(1 $if\ the\ condition\ is\ true$, 0 $otherwise$)
$pred_{seg_i}$:
$whether\ a\ boundary\ exists\ within\ window$
$in\ the\ predicted\ segmentation$

WindowDiff, A refined version of Pk that reduces penalties for repeated errors in the same region, making it fairer for long or dense transcripts.

$$WindowDiff = \frac{1}{|R|}\sum_{r\in R} 1[ref\_count(r) = pred\_count(r)]$$

R: set of reference boundary indices
$1[\cdot]$: $indicator\ function$
(1 $if\ the\ condition\ is\ true$, 0 $otherwise$)
$ref_{count(r)}$: number of boundaries in a window centered at reference boundary in the ground truth
$pred_{count(r)}$: number of boundaries in the same window in the predicted segmentation

To evaluate the robustness of our BERT-enhanced TextTiling pipeline, we compare two evaluation strategies on the full EVTS dataset (1,553 videos): Fixed Configuration: A baseline using static hyperparameters (chunk_size=12, window_size=2, depth_threshold=0.02). These values were selected based on preliminary experiments on a development subset and represent a balanced, general-purpose setup that avoids extreme sensitivity or over-smoothing. Adaptive Configuration: A content-aware strategy that dynamically selects the best hyperparameters per video by optimizing F1 score over a grid of chunk_size ∈ {8, 12, 16}, window_size ∈ {2, 3, 4}, depth_threshold ∈ {0.02, 0.05, 0.08, 0.10}.
The adaptive approach is motivated by a key observation: educational videos vary significantly in topic density.
We hypothesize that no single configuration can optimally handle both extremes. Therefore, we estimate content density as:

$$\text{Density} = \frac{Video\ duration(seconds)}{Number\ of\ cleaned\ sentences}$$

and select hyperparameters that maximize F1 for each video. In practice, this means:

High-density videos: chunk=8, depth=0.02 (high sensitivity).

Low-density videos: chunk=16, depth=0.10 (noise suppression).

Table 3, the adaptive strategy yields consistent improvements:

Metric	Fixed	Adaptive	Δ
F1	0.561	0.665	+ 0.104
Precision	0.628	0.735	+ 0.107
Recall	0.588	0.654	+ 0.066
Pk	0.419	0.365	– 0.054
WindowDiff	0.689	0.696	– 0.007

The F1 score gain shows that selecting content-conscious hyperparameters significantly improves segmentation quality. The greater improvement in accuracy (+0.107 vs. +0.066 in recall) indicates that the adaptive approach primarily reduces excessive fragmentation, a common problem in content-dense lectures when using fixed, highly sensitive parameters.
Notably, the pk decreased by 0.054, confirming that the adaptive approach produces fragmentation that is more structurally consistent with human-defined boundaries. The WindowDiff quasi-stability (from 0.689 to 0.696) indicates that while local boundary errors decrease, the global distribution of segments remains stable, demonstrating robust, context-sensitive behavior. Our method remains completely unsupervised and does not require any baseline facts at inference time. In practice, the fixed configuration (segment = 12, window = 2, depth = 0.02) provides a strong, general baseline (F1 = 0.561), while the adaptive strategy reveals the maximum potential for adjusting conscious content. In practice, this means that after identifying optimal sentence-level boundaries via the adaptive pipeline, we map each segment's starting sentence index back to its original timestamp in the cleaned VTT transcript. This produces a final list of semantically grounded temporal markers, suitable for applications such as video navigation, summarization, or interactive learning interfaces. Our method remains fully unsupervised and requires no ground truth at inference time. Yet, it achieves markedly higher alignment with human-annotated thematic structure when hyperparameters are tuned to content characteristics.

6. Conclusion and Future Work

In this work, a robust, fully unsupervised pipeline for thematic segmentation of educational video transcripts was proposed, based on a BERT-enhanced Text-Tiling approach. By aligning predicted sentence-level boundaries with human-annotated timestamps and evaluating within a

standard ±3-sentence tolerance window, it was demonstrated that content-aware hyperparameter tuning significantly outperforms fixed configurations.

EVTS (Educational Video Timestamps Segmentation) was, to the best of our knowledge, presented as one of the largest-scale, time-stamped datasets explicitly designed for both educational video segmentation and semantic timestamp generation. Built from 1,553 real-world educational videos, each annotated with creator-provided timestamps marking the start of distinct instructional segments, EVTS enables precise temporal grounding of thematic content. Across 1,553 videos in the EVTS dataset, our adaptive strategy selects optimal chunk_size, window_size, and depth_threshold for each video, achieving a state-of-the-art absolute improvement. Crucially, this gain was accompanied by consistent reductions in structural error metrics (Pk and WindowDiff), confirming that the detected boundaries better reflect human-defined thematic transitions. The practical value of this improvement lies in timestamp generation. Since each predicted segment is anchored to the start time of its first sentence in the cleaned VTT transcript, the refined boundaries directly yield semantically coherent temporal markers. These timestamps are not merely syntactic breaks but meaningful points of topic shift, making them highly suitable for applications such as video navigation, interactive summarization, or segment-based retrieval in educational platforms.

Future direction, the availability of high-quality ground truth in EVTS unlocks several promising supervised and hybrid directions. First, we plan to develop a fully supervised segmentation model that directly learns to predict segment boundaries. This would bypass heuristic hyperparameter tuning entirely and likely yield further gains in accuracy. explore semi-supervised or few-shot learning approaches that leverage the full EVTS dataset to generalize to new domains with minimal annotation. Finally, we intend to integrate our segmentation output into downstream educational applications, such as automatic video summarization, interactive Q&A systems, or adaptive learning pathways, and rigorously evaluate their pedagogical impact through user studies.

Reference:

[1] Chen, C. (2025). Entertainment social media based on deep learning and interactive experience application in English e-learning teaching system. Entertainment Computing, 52, 100846.

[2] Chand, D., & Ogul, H. (2020, March). Content-based search in lecture video: a systematic literature review. In 2020 3rd International Conference on Information and Computer Technologies (ICICT) (pp. 169-176). IEEE.

[3] Kishi, R. M., & Goularte, R. (2016, November). Video scene segmentation through an early fusion multimodal approach. In Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia) (pp. 41-46). SBC.

[4] Soares, E. R., & Barrére, E. (2018, October). A framework for automatic topic segmentation in video lectures. In Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia) (pp. 31-36). SBC.

[5] Huang, X., Peng, F., Schuurmans, D., Cercone, N., & Robertson, S. E. (2003). Applying machine learning to text segmentation for information retrieval. Information Retrieval, 6(3), 333-362.

[6] Freisinger, S., Seeberger, P., Ranzenberger, T., Bocklet, T., & Riedhammer, K. (2025). Towards Multi-Level Transcript Segmentation: LoRA Fine-Tuning for Table-of-Contents Generation. In Proc. Interspeech 2025 (pp. 276-280).

[7] Gummadi, V. P. K. (2022). MuleSoft API Manager: Comprehensive lifecycle management. Journal of Information Systems Engineering and Management, 7(4), 1–9.

[8] Retkowski, F., & Waibel, A. (2024). From Text Segmentation to Smart Chaptering: A Novel Benchmark for Structuring Video Transcriptions. *CoRR*..

[9] Ghinassi, I. (2021, May). Unsupervised Text Segmentation via Deep Sentence Encoders: a first step towards a common framework for text-based segmentation, summarization and indexing of media content. Proceedings of 2nd International Workshop on Data-driven Personalisation of Television.

[10] Yang, H., & Meinel, C. (2014). Content based lecture video retrieval using speech and video text information. IEEE transactions on learning technologies, 7(2), 142-154.

[11] Lin, M., Chau, M., Cao, J., & Nunamaker Jr, J. F. (2005). Automated video segmentation for lecture videos: A linguistics-based approach. International Journal of Technology and Human Interaction (IJTHI), 1(2), 27-45.

[12] Reimers, N., & Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.

[13] Hearst, M. A. (1997). Text tiling: Segmenting text into multi-paragraph subtopic passages. Computational linguistics, 23(1), 33-64.

[14] Delacourt, P., & Wellekens, C. J. (2000). DISTBIC: A speaker-based segmentation for audio data indexing. Speech communication, 32(1-2), 111-126.

[15] Soares, E. R., & Barrére, E. (2017, October). An approach for automatic segmentation of scenes in educational videos through the use of audio transcription and semantic annotation. In Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web (pp. 229-235).

[16] Soares, E. R., & Barrére, E. (2018, October). 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 (pp. 189-196).

[17] Soares, E. R., & Barrére, E. (2018, October). A framework for automatic topic segmentation in video lectures. In Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia) (pp. 31-36). SBC.

[18] Elmakies, A., Abend, O., & Adi, Y. (2025). Unsupervised speech segmentation: A general approach using speech language models. arXiv preprint arXiv:2501.03711.

[19] Ghauria, J. A., Hakimova, S., & Ewertha, R. (2020). Classification of Important Segments in Educational Videos using Multimodal Features.

[20] Das, A., & Das, P. P. (2020). Incorporating domain knowledge to improve topic segmentation of long MOOC lecture videos, CoRR abs/2012.07589 (2020). URL: https://arxiv.org/abs/2012.07589

[21] Dimitsas, M., & Leidner, J. L. (2023, September). Topic Segmentation of Educational Video Lectures Using Audio and Text. In European Conference on Artificial Intelligence (pp. 447-458). Cham: Springer Nature Switzerland.

[22] Gupta, A., Jawahar, S. F., and Tapaswi, M. (2023). Unsupervised Audiovisual Lecture Segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 5232–5241).

[23] Yu, H., Deng, C., Zhang, Q., Liu, J., Chen, Q., & Wang, W. (2024). Multimodal Fusion and Coherence Modeling for Video Topic Segmentation. *CoRR*

[24] .Vasuki, M., Gangadharan, M. A., Daniel, J. T., Sadashiv, A., Venugopal, V., & Vekkot, S. (2024, July). Multi-Modal Automatic Video Segmentation with Sentence Transformer Embeddings and KeyBERT-Based Subtopic Extraction. In 2024 2nd World Conference on Communication & Computing (WCONF) (pp. 1-6). IEEE.

[25] Lin, M., Nunamaker, J. F., Chau, M., & Chen, H. (2004, January). Segmentation of lecture videos based on text: a method combining multiple linguistic features. In 37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the (pp. 9-pp). IEEE.

[26] Tu, Y., Xiong, Y., Chen, W., & Brinton, C. (2018, November). A domain-independent text segmentation method for educational course content. In 2018 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 320-327). IEEE.

[27] Das, A., & Das, P. P. (2020). Incorporating domain knowledge to improve topic segmentation of long MOOC lecture videos, CoRR abs/2012.07589 (2020). URL: https://arxiv.org/abs/2012.07589.

[28] Chand, D., & Oğul, H. (2021, January). A framework for lecture video segmentation from extracted speech content. In 2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI) (pp. 000299-000304). IEEE.

[29] Galanopoulos, D., & Mezaris, V. (2018, December). Temporal lecture video fragmentation using word embeddings. In International Conference on Multimedia Modeling (pp. 254-265). Cham: Springer International Publishing.

[30] Retkowski, F., & Waibel, A. (2024). From text segmentation to smart chaptering: A novel benchmark for structuring video transcriptions. arXiv preprint arXiv:2402.17633.

[31] Freisinger, S., Schneider, F., Herygers, A., Georges, M., Bocklet, T., & Riedhammer, K. (2023). Unsupervised multilingual topic segmentation of video lectures: What can hierarchical labels tell us about the performance?. In SLaTE Workshop (pp. 141-145).

[32] Freisinger, S., Seeberger, P., Ranzenberger, T., Bocklet, T., & Riedhammer, K. (2025). Towards Multi-Level Transcript Segmentation: LoRA Fine-Tuning for Table-of-Contents Generation. In Proc. Interspeech 2025 (pp. 276-280).

[33] Gklezakos, D. C., Misiak, T., & Bishop, D. (2024). TreeSeg: Hierarchical topic segmentation of large transcripts. arXiv preprint arXiv:2407.12028.

[34] Choi, F. Y. (2000). Advances in domain independent linear text segmentation. arXiv preprint cs/0003083.

[35] Beeferman, D., Berger, A., & Lafferty, J. (1999). Statistical models for text segmentation. Machine learning, 34(1), 177-210.

[36] Li, J., Chiu, B., Shang, S., & Shao, L. (2020). Neural text segmentation and its application to sentiment analysis. IEEE Transactions on

Knowledge and Data Engineering, 34(2), 828-842.

[37] Pevzner, L., & Hearst, M. A. (2002). A critique and improvement of an evaluation metric for text segmentation. Computational Linguistics, 28(1), 19-36.