Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

Automatic Chapter Generation for Hindi-English YouTube Videos

Ashana Agarwal*, Avani Gupta**, and Dr Rakhi Gupta***

[1]Department of Computer Science, Manipal University India, ashanaagarwal100@gmail.com

[2]Department of Computer Science, Manipal University, India

[3]Department of Economics, Apex University, India

ARTICLE INFO

Received: 11 Oct 2024
Revised: 26 Nov 2024
Accepted: 15 Dec 2024

ABSTRACT

Chaptering long-form multilingual videos into semantically meaningful chapters is crucial to making the content more accessible and navigable, especially in educational and informative videos. Previous work such as VidChapters-7M has investigated large-scale chaptering for English videos, but little research has ventured into bilingual or code-switched videos, especially for underrepresented languages such as Hindi. In this paper, we propose a new pipeline for automatic chapter generation on Hindi-English YouTube videos by utilizing both audio transcripts and semantic information. We construct a dataset of code-switched videos and assess our chapter generation using BLEU scores, readability (with Flesch Reading Ease), and temporal coherence metrics. We also include human evaluation to assess relevance and coherence of chapters. Our experiments show encouraging performance, and we highlight the challenges and opportunities of bilingual video understanding. We declare that this paper opens new avenues for multilingual video segmentation, paving the way to the development of inclusive AI systems and improved content navigation tools for the Indian subcontinent and beyond.

Keywords: segmentation, subcontinent, temporal, bilingual

Introduction

Exponential growth in online video material, particularly on sites such as YouTube, has changed the manner in which individuals view information, entertainment, and education. With millions of hours of video material uploaded every day, effectively navigating through lengthy videos is an urgent problem. Chapter segmentation—a mechanism of splitting videos into semantically significant chunks—is an improvement in terms of user experience since it enables users to find specific parts of a video rapidly. YouTube's own chaptering system is presently restricted in terms of scalability and language support, strongly English-centric [6]. This leaves a huge accessibility gap for non-English language videos or for multilingual environments.

A high proportion of Indian content creators produce video content by codeswitching between a mix of Hindi and English, commonly known as Hinglish, which reflects everyday linguistic practice across India [12]. However, existing chaptering settings are not well suited to support code-switching or the complex structural attributes that go with this kind of content. Thus, there is a need for a multilingual, culturally sensitive chapter generation solution to reach a wider audience and assist creators of local language content.

To tackle this, we introduce an automatic chaptering pipeline specifically designed for Hindi-English YouTube videos. Our method leverages ASR-transcripts to produce chapters without any human annotations [1]. In contrast to existing work like VidChapters-7M, which is limited to English videos and supervised training with human references, we follow a reference-free evaluation approach with light human validation [4].

Our approach includes extracting temporal and semantic information from transcripts, creating segment boundaries, and giving topic-representative titles. To quantify the quality of our chapters, we use metrics such as BLEU score for title-summary relevance [13], Flesch Reading Ease for readability [14], and temporal coherence to quantify structural balance. We also add semantic cohesion and intra-chapter coherence by using transformer-based embeddings [2]. Additionally, we conducted a human evaluation of the chapters produced by collecting feedback from five independent raters. The chapters were rated on a 5-point Likert scale for relevance, informativeness, and coherence.

To the best of our knowledge, this is the first attempt at solving the problem of automatic chapter generation for big-volume bilingual (Hindi-English) material. Our method is language-independent, robust, and may be generalized to other multilingual and code-switched environments. We believe that this work has wide-ranging implications for improving accessibility, optimization of video recommendation algorithms, and navigation of content in the Indian digital landscape.

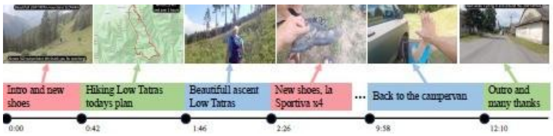


Figure 1: Youtube Video with Mannual Chapters and Timestamps

Section	Subsections
1. Introduction	–
2. Related Work	Chaptering and Segmentation of Temporal Video, Understanding Multilingual and Code-Switched Text, Evaluation of Chapter Quality
3. A Medium-Scale HindiEnglish Dataset for Multilingual Video Chaptering	Data Collection, Data Pre-processing, Data Analysis (Summary Word Count per Segment and Start Time, Distribution of Segment Lengths, Distribution of Video Lengths)
4. Methodology and Experimentation	Transcript Generation using Whisper, Chapter Refinement and Title Generation, Unified Output Format, Evaluation (Automated Assessment, Human Evaluation)
5. Conclusion, Limitations, Social Impacts and Future Work	Conclusion, Constraints, Societal Implications, Future Work

Content Index

Related Work

1.1 Chaptering and Segmentation of Temporal Video

The issue of generating automatic video chapters has gained much attention in the past few years owing to its relevance to improving the navigability of content as well as the user experience. The VidChapters-7M dataset [4] provided a full English dataset for chaptering under supervision, including multimodal features like ASR transcripts and visual features. Other works, such as YouCook2 [11] and HowTo100M [10], concentrated on tutorial videos, employing dense captions and narration transcripts to align content with chapter boundaries. These datasets are, however, primarily English-only content and do not account for multilingual or code-switched data.

1.2 Understanding Multilingual and Code-Switched Text

The emergence of code-switched and multilingual communication, particularly on social and casual video platforms, has introduced new challenges to the field of natural language processing. Although research in multilingual models has made progress, there is a dearth of studies focused specifically on the chaptering of videos that involve mixed languages, like Hindi-English. Studies such as [12] have examined code-switching behavior in Indian speakers, but to the best of our knowledge, this work is the first attempt to tackle the task of generating and evaluating chapters for code-switched content on YouTube in a systematic process.

1.3 Evaluation of Chapter Quality

Earlier research has employed automated evaluation metrics such as BLEU [13], METEOR [15], and ROUGE [16] to quantify chapter title-summary consistency and text quality. These metrics have problems with evaluating short texts or multilingual speech. To overcome these limitations, existing research [4] has complemented automated assessment with human evaluation methods to establish chapter coherence, informativeness, and usefulness. In this study, we also combine both automated and human evaluation to obtain a better representation of chapter quality.

2 A Medium-Scale Hindi-English Dataset for Multilingual Video Chaptering

Lacking a publicly accessible benchmark for bilingual Hindi-English YouTube video chaptering, we constructed a new dataset of 437 videos which is about 20GB in size from a variety of content categories, including history, interviews, commentary, and discussion. Each video demonstrates natural code-switching between Hindi and English characteristic of authentic multilingual usage patterns.

To build the dataset, we used a multi-stage pipeline:

- **Video Selection:** We collected publicly available YouTube videos that illustrate code-switched speech with clear audio and smooth switches between Hindi and English.
- **Audio Extraction:** Audio streams were pulled out of the videos using common video-to-audio conversion software with temporal alignment preserved relative to the original video timestamps.
- **Transcript Generation:** We used the Whisper Medium ASR model [1] to generate timestamped transcripts. As an open-source, multilingualsupported model, it assisted us in successfully capturing the intricate nuances of Hindi-English mixed speech.
- **Division into Chapters:** The transcripts generated were separated into chapters according to a range of linguistic and prosodic features, including:

– Pauses and silence durations,

– Tonal shifts and points of emphasis,

– Semantic topic shift detection with embeddings.

- **Chapter Formatting:** A chapter entry consists of a start_time, a title, and a summary that are generated automatically by applying a mix of rule-based heuristics and language models.

The generated dataset is organized in the form of JSON files, each of which corresponds to a single video and includes a collection of chapter annotations. To our knowledge, this is the first dataset to be generated with the exclusive intention of the chaptering of bilingual Hindi-English YouTube videos.

2.1 Data Collection

While creating our medium-sized bilingual corpus, we made sure to gather a diverse set of Hindi-English YouTube videos that already had chapters written by hand. We obtained the videos mainly from podcasts, educational content creators, and long discussion forums where bilingual narration is common.

We utilized the official YouTube Data API to simplify the process of data gathering automation. In particular, we searched through videos with chapters by querying metadata fields such as timestamps, segment titles, and video duration. Results collected from the API were filtered to contain only the videos that have chapter markers visible within their metadata or description.

To ensure the creation of high-quality and relevant content, we carefully curated a list of popular YouTube channels and their corresponding channel IDs that are renowned for sharing bilingual content. The list included prominent Indian podcast channels, educational platforms, and infotainment services. In addition, only videos longer than a set duration (typically 10 minutes) were selected to allow for meaningful segmentation.

This filtering process allowed us to gather a combination of content rich in code-switching and tonal variety, making the dataset suitable for the task of Hindi-English automated chapter generation.

2.2 Data Pre-processing

After compiling an initial list of video URLs from the YouTube API, we filtered our dataset stringently. Specifically, we eliminated all videos with a runtime exceeding two hours. This was done to facilitate efficient functioning of the audio-to-text transcription model and to avoid creating overly long transcripts that could result in segmentation noise.

For the remaining set of videos, we applied a combination of pytube and ffmpeg to extract their respective audio streams. The audio files were organized in a structured local directory for subsequent processing.

We used OpenAI's Whisper Medium Automatic Speech Recognition (ASR) model [1] to transcribe the audio. This model was chosen for its strong performance in handling multilingual audio and its ability to generate precise timestamps for every utterance. The model also captured acoustic cues, such as pauses and tone changes, which aided in establishing semantically meaningful sentence boundaries.

The initial transcripts were subjected to rigorous cleaning using a variety of pre-processing techniques. This included the removal of filler words, timestamp artifacts, and irregular punctuation. We also focused on normalizing sentence structure, improving readability, and preserving the semantic fidelity of the original spoken material.

This organized pipeline ensured that every video transcript was clean, properly timestamped, and linguistically uniform—critical prerequisites for successful chapter generation and evaluation.

2.3 Data Analysis

To gain insight into structural features and trends within our dataset, we conducted preliminary exploratory data analysis on the video durations and corresponding segment characteristics. This step validated the homogeneity and consistency of data used for chapter construction.

2.3.1 Summary Word Count per Segment and Start Time

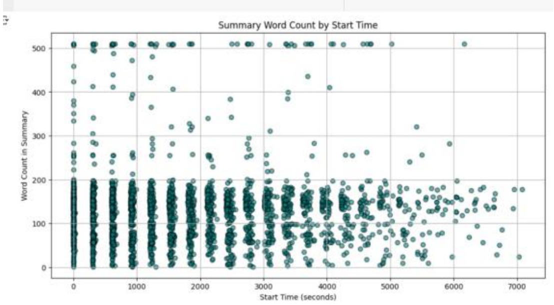


Figure 2: Start Time vs. Word Count in Summary

Figure 2 shows summary word counts against their corresponding start positions for all chapters. From this graphical representation, we can identify structural trends with regards to information density on the video timeline. As postulated in [2], early segments of videos are made up of repetitive or introductory content, while subsequent segments are more informative. Our figure verifies this trend for Hinglish videos and guided our process of condensing and combining chapters.

2.3.2 Distribution of Segment Lengths

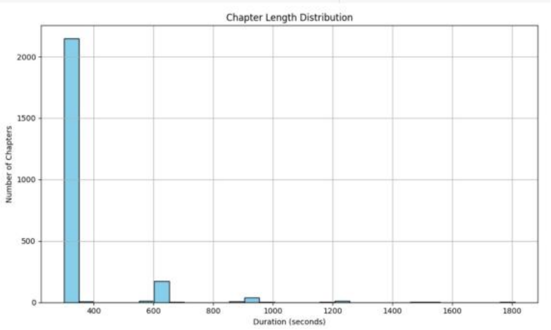


Figure 3: Duration vs. Number of Chapters

Figure 3 is a histogram of chapter lengths over the dataset, which indicates the majority of segments to be 4–6 minutes. This aligns with earlier chaptering strategies, which try to balance shortness and semantic coherence [4][6]. Our adaptive merging algorithm was tuned to have a lower bound chapter length of approximately 5 minutes, to balance readability and usability.

2.3.3 Distribution of Video Lengths

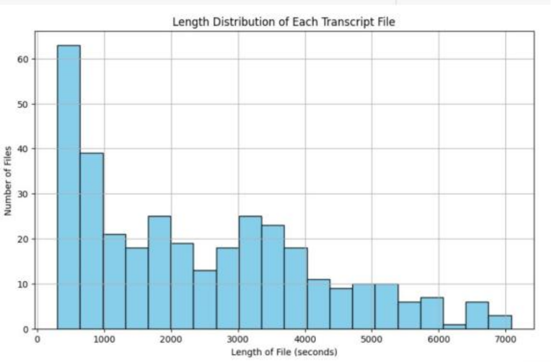


Figure 4: Length of Videos vs. Number of Files

The distribution of video lengths in our dataset is depicted in Figure 4. Most videos are between 10 and 30 minutes, with very few videos being longer than 60 minutes. This selection method is consistent with the best practice of previous datasets like VidChapters-7M , in which the shortest and longest video lengths were removed to reduce segmentation noise and improve ASR performance at preprocessing.

Methodology and Experimentation

Our proposed system for Hindi-English bilingual video automated chapter extraction is a multi-step pipeline process from audio extraction to final structured chapter outputs. The major building blocks of the methodology are as follows:

2.4 Transcript Generation using Whisper

We used OpenAI's Whisper Medium model, one of the latest automatic speech recognition (ASR) models, to transcribe audio from selected YouTube videos [1]. For each video, the audio was downloaded locally and saved. Whisper was then used to transcribe the audio, producing both .txt and .json files for each input. The transcriptions included timestamps, with segmentations determined by pause durations, tone changes, and other acoustic features detected by the model.

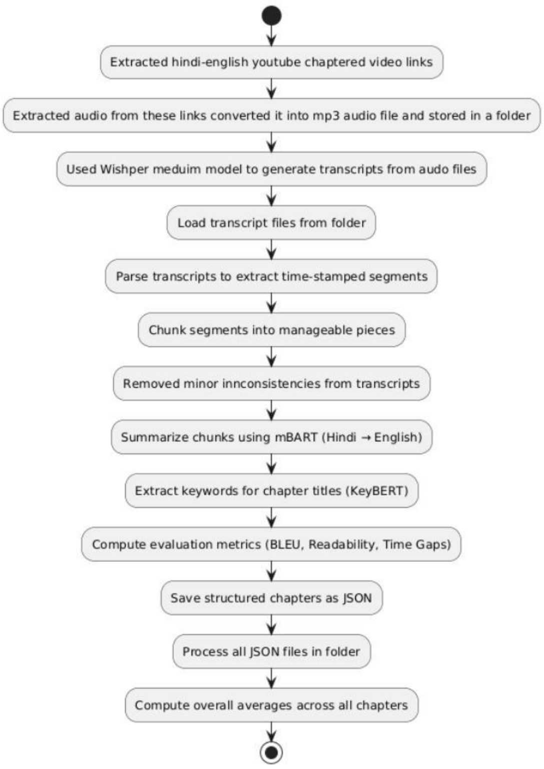

Extracted hindi-english youtube chaptered video links
Extracted audio from these links converted it into mp3 audio file and stored in a folder
Used Wishper meduim model to generate transcripts from audo files
Load transcript files from folder
Parse transcripts to extract time-stamped segments
Chunk segments into manageable pieces
Removed minor innconsistencies from transcripts
Summarize chunks using mBART (Hindi → English)
Extract keywords for chapter titles (KeyBERT)
Compute evaluation metrics (BLEU, Readability, Time Gaps)
Save structured chapters as JSON
Process all JSON files in folder
Compute overall averages across all chapters

Figure 5: Data collection, methodology and Evaluation Flowchart

The early results of this phase were chapters predominantly in Hindi, typically only a few seconds long. These introductory passages occasionally showed structural dissonance, repetitive usage, or jarring transitions—often due to overlapping speech or irrelevant content.

2.5 Chapter Refinement and Title Generation

To improve and further process these raw transcripts, we introduced a text processing and summarization module. The steps taken are:

- **Deletion of inconsistencies:** Rule-based cleaning methods and transformerbased language models [2], including MBartForConditionalGeneration and MBart50TokenizerFast, were used to correct grammatical mistakes, remove disfluencies, and improve sentence structure coherence in chapter abstracts.

- **Translation:** Since the initial abstracts were in Hindi, we translated them into English using pretrained multilingual transformer models for broader accessibility and evaluation.
- **Dynamic chapter merging:** For improved semantic coherence and usability, chapters shorter than 5 minutes were systematically merged with adjacent ones to ensure a minimum contextual length.
- **Title extraction:** We used KeyBERT [3], a BERT-based keyword extraction technique, to create short English titles from the cleaned summaries of each chapter.

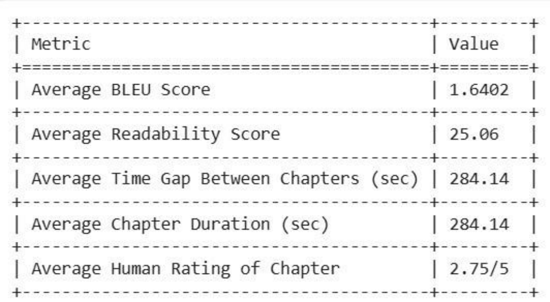
2.6 Unified Output Format

Each input video resulted in a final structured .json file containing:

- **start-time:** Timestamp in HH:MM:SS format
- **title:** Brief English chapter title
- **summary:** Refined English summary of the chapter content

All finalized outputs were saved in a designated directory and used as the foundation for our evaluation and analysis.

2.7 Evaluation



```
+----------------------------------------+---------+
| Metric                                 | Value   |
+========================================+=========+
| Average BLEU Score                     | 1.6402  |
+----------------------------------------+---------+
| Average Readability Score              | 25.06   |
+----------------------------------------+---------+
| Average Time Gap Between Chapters (sec) | 284.14  |
+----------------------------------------+---------+
| Average Chapter Duration (sec)         | 284.14  |
+----------------------------------------+---------+
| Average Human Rating of Chapter        | 2.75/5  |
+----------------------------------------+---------+
```

Figure 6: Evaluation Metrics

To determine the coherence and quality of the chapters produced, we used a combination of automated and manual assessment techniques. Our aim was to evaluate both the linguistic consistency of chapter titles and summaries, and the practical usability of the chapter segmentation in a real-world viewing context.

2.7.1 Automated Assessment

We computed the following quantitative metrics across the produced chapters:

- **BLEU Score:** BLEU (Bilingual Evaluation Understudy) score was calculated between each chapter title and its corresponding summary to assess semantic similarity [13]. The average BLEU score across all files was **1.6402**.

  Since BLEU was initially developed to assess monolingual text, its use to measure our bilingual summaries has some scoring limitations. This creates the need to supplement BLEU with human evaluation to more effectively judge chapter quality within code-switched contexts.

- **Readability Score:** Using the Flesch Reading Ease score [14], we evaluated the readability of each summary. The dataset had an average readability of **25.06**, indicating moderate complexity suitable for academic or professional audiences.

- **Temporal Gaps:** The average time between two successive chapters was approximately **284.14 seconds (4.7 minutes)**, showing that our adaptive merging technique produced compact, coherent chapters.
- **Chapter Lengths:** The average final chapter length was also **284.14 seconds**, validating the success of our duration normalization strategy.

It should be noted that this is the **first documented study of automatic video chaptering from Hindi-English bilingual data**. Prior work, such as VidChapters-7M [4], focused exclusively on English content. Therefore, direct comparison of BLEU, readability, and temporal metrics is not entirely applicable due to the linguistic diversity and complexity introduced by code-switching and bilingual content.

2.7.2 Human Evaluation

In addition to automated metrics, we conducted a human evaluation study to assess the practical quality and usefulness of the generated chapters. Five independent volunteers rated a random sample of chapters based on two criteria:

- **Quality of Chapter Summary:** How well the summary reflects the segment content.
- **Title Relevance:** How appropriate and concise the chapter title is.

Each aspect was rated on a scale from 1 to 5. The average rating across all evaluators was **2.75 out of 5**, indicating a baseline quality level with ample room for improvement using more advanced language models or fine-tuned summarization approaches.

Conclusion, Limitations, Social Impacts and Future Work

2.8 Constraints

Although the findings were positive, our research is not without some limitations:

- **Chapter Length Consistency:** Most generated chapters fall within a typical duration (5–10 minutes), which, while consistent, sometimes fails to align naturally with topic boundaries in the video content.
- **Title Contextuality:** While we employ summarization and keyword extraction techniques like KeyBERT [3], generated titles may sometimes lack contextual richness or appear overly generic.
- **Bilingual Complexity:** Handling phonetically diverse Hindi-English code-switched speech presents inherent challenges. Some semantic nuances can be lost or mistranslated during the transcription or translation phases.

These limitations highlight areas for further research, particularly the integration of more adaptive segmentation methods and advanced fine-tuned large language models.

2.9 Societal Implications

This project exemplifies the growing capability of AI to address real-world accessibility issues. By automating a time-consuming but valuable task—video chaptering—this system empowers content creators, especially in low-resource or multilingual environments, to deliver better organized and accessible content effortlessly.

From the viewer's perspective, chaptered videos enhance learning efficiency and engagement by enabling selective consumption of relevant content. Users can jump to sections aligned with their interests or needs, thereby improving information retention and time management.

Furthermore, our system and bilingual dataset lay the groundwork for future research into inclusive AI tools that enhance comprehension and accessibility across languages, regions, and platforms.

To the best of our knowledge, this is the first documented effort to automate chapter generation for code-switched Hindi-English video content—addressing a critical gap in both research and application.

2.10 Future Work

While the current system is promising, there are many possibilities to further enhance and support its functionality.

First, our aim is to optimize transformer models specifically for code-switched Hindi-English datasets. Such an optimization would make the summarization and title-generation functions better comprehend and retain contextual appropriateness across language divides, producing outputs that are more culturally and linguistically sensitive.

Secondly, we plan to explore several chapter segmentation algorithms that are capable of dynamically adjusting chapter length and chapter number based on the semantic density, speaker transitions, and narrative structure of each video. This would allow for a more natural segmentation approach tailored to the individual flow of each video, rather than employing constant thresholds or time-based merging.

Eventually, subsequent versions of the system will support multimodal cues (e.g., speaker, face, or scene change visual) in a quest to get a better boundary detection and make the user's experience better, particularly for conversational or visually engaged videos.

Conclusion

In this study, we proposed a novel automatic chapter generation framework for bilingual Hindi-English YouTube videos. Using Automatic Speech Recognition (ASR) via Whisper Medium [1], transformer-based summarization models [2], and semantic filtering techniques, we created an end-to-end pipeline that efficiently extracts, cleans, and segments transcripts into organized chapters. Our solution significantly improves the usability and accessibility of long-form content.

This system is particularly useful for North Indian content creators who produce educational, podcast-style, or commentary videos. It enables them to automatically create chapters with minimal manual input, saving time in postproduction and providing the audience with a more structured and engaging viewing experience.

Acknowledgements

We would like to express our sincere gratitude to the creators and maintainers of the Whisper ASR model, the HuggingFace and KeyBERT communities, and the developers of various open-source tools that made this research possible.

We are also thankful to the individuals who participated in the human evaluation process, providing valuable feedback that helped validate our approach.

Finally, we acknowledge the support of our mentors, peers, and the broader research community for their guidance and encouragement throughout this project.

References

[1] A. Radford, J. W. Kim, T. Xu, et al., "Whisper: Robust speech recognition via large-scale weak supervision," *OpenAI*, 2022. [Online]. Available: https://openai.com/research/whisper

[2] T. Wolf, L. Debut, V. Sanh, et al., "Transformers: State-of-the-art natural language processing," in *Proceedings of EMNLP*, 2020. [Online]. Available: https://huggingface.co

[3] M. Grootendorst, "KeyBERT: Minimal and easy keyword extraction with BERT," 2020. [Online]. Available: https://github.com/MaartenGr/ KeyBERT

[4] S. Wani, S. Roy, N. Rahman, et al., "VidChapters-7M: Video Chapters at Scale," in *NeurIPS Datasets and Benchmarks*, 2023. [Online]. Available: https://openreview.net/forum?id=JgFiMNbovO

[5] C. Sun, R. Panda, S. Chattopadhyay, et al., "VideoCLIP: Contrastive pretraining for zero-shot video-text understanding," in *Proceedings of CVPR*, 2022.

[6] C. Pfeiffer and N. Bregler, "Automatic video chaptering using temporal structure and content similarity," in *Proceedings of ACM Multimedia*, 2006.

[7] D. Xu, C. Liu, et al., "Coarse-to-Fine Feature Mining for Video Semantic Segmentation," in *Proceedings of ICCV*, 2021.

[8] C. Manning, M. Surdeanu, J. Bauer, et al., "The Stanford NLP toolkit," in *Proceedings of ACL*, 2008.

[9] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," in *Proceedings of NeurIPS*, 2017.

[10] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips," in *Proceedings of ICCV*, 2019.

[11] L. Zhou, X. Chenliang, and J. J. Corso, "Towards automatic learning of procedures from web instructional videos," in *Proceedings of AAAI*, 2018.

[12] A. Si and T. M. Ellison, "Inter-individual differences in Hindi–English codeswitching: A quantitative approach," *International Journal of Bilingualism*, vol. 26, no. 2, pp. 215–234, 2022.

[13] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of ACL*, pp. 311–318, 2002.

[14] R. Flesch, "A new readability yardstick," *Journal of Applied Psychology*, vol. 32, no. 3, pp. 221–233, 1948.

[15] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, 2005.

[16] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proceedings of the ACL Workshop on Text Summarization Branches Out*, pp. 74–81, 2004.