Instructional Video Summarization with Transformers: A Curriculum Learning Approach for ASR-Generated Transcripts

1st Mridul Banik
Department of Computer Science
Colorado State University
Fort Collins, CO, USA
mridul.banik23@alumni.colostate.edu

2nd Ismail Hossain
Department of Computer Science
George Mason University
Fairfax, VA, USA
ihossai4@gmu.edu

***Abstract*—This paper addresses the challenge of abstractive summarization for instructional video transcripts. Utilizing a document-level encoder rooted in transformer architectures, the proposed methodology enhances the fluency and generalizability of generated summaries across diverse video content. A unique dataset of over 5,000 extracted transcripts supports the training process, employing specific fine-tuning and order-preserving techniques. Assessments based on metrics such as Content F1 and human evaluations confirm that the synthesized narratives achieve quality comparable to human-authored text, providing concise and informative overviews for online educational platforms.**

***Index Terms*—abstractive summarization, BERT, instructional video summarization, automatic speech recognition (ASR), curriculum learning, natural language generation, transformer models, educational technology, pretrained language models, conversational transcript summarization, human evaluation, ROUGE scoring, Content F1, fine-tuning, natural language processing**

I. Introduction

In the current digital landscape, online instructional content has become a dominant form of knowledge dissemination, especially through video-sharing platforms such as YouTube, Coursera and Udemy. As educational material increasingly shifts to multimedia formats, users face a critical challenge: sifting through lengthy and often unstructured videos to locate relevant information. Instructional videos typically contain step-by-step guidance, demonstrations and spoken explanations that span several minutes to hours. Despite their value, viewers often lack the time or patience to watch entire videos, especially when the content lacks a clear overview or structured breakdown. This growing demand for efficient information retrieval and content accessibility has led to the need for accurate, automated summarization tools tailored specifically to the instructional video domain.

Traditionally, video summarization research has focused on visual content summarization—identifying and extracting keyframes or short video clips to create a condensed version of the video. However, while this method may be effective for summarizing cinematic or surveillance footage, it fails to capture the semantic richness embedded in the spoken narrative of instructional videos. In such contexts, the audio component, particularly the instructional dialogue, carries the most crucial information. As a result, textual summarization based on transcripts generated from automatic speech recognition (ASR) systems has gained traction as a more meaningful alternative.

Currently, many video creators manually write descriptions or titles that accompany their videos. These textual elements are primarily crafted for search engine optimisation (SEO), aiming to attract viewers rather than accurately reflect the video's content. Consequently, they are often vague, overly generic or misleading, providing little actual insight into the instructional steps or goals discussed in the video. This disconnect between the video's substance and its textual metadata presents a serious barrier to effective content discovery and comprehension.

To address these limitations, we propose an automated framework that generates abstractive summaries from ASR-generated transcripts of instructional videos. Our approach leverages a fine-tuned BERT-based model, which has shown state-of-the-art performance in a variety of text generation and summarization tasks. Unlike extractive models that merely select key sentences from the source text, our model constructs new sentences that encapsulate the core ideas in a fluent and coherent manner. This enables the generation of high-quality summaries that are not only informative but also concise and readable.

Moreover, our approach introduces several novel contributions. First, we curate and utilise a diverse training corpus by combining datasets such as CNN/DailyMail, WikiHow and How2 video summaries, covering a broad range of domains and summary styles. Second, we implement a robust preprocessing pipeline to enhance the quality of ASR transcripts, which are often plagued by noise, improper punctuation and irrelevant introductory remarks. By cleaning and standardising these transcripts, we significantly improve the input quality for our summarization model. Third, we apply a strategic training regimen that gradually exposes the model to increasingly complex and domain-specific content, improving its generalisation

capabilities.

Finally, the impact of our work extends beyond academic exploration. From an application standpoint, the ability to automatically generate reliable and user-friendly summaries has the potential to revolutionise how users interact with video-based learning materials. Educators, content creators and digital platforms can benefit from scalable solutions that enhance video indexing, improve content accessibility for users with time constraints or disabilities and support the development of intelligent recommendation systems. By bridging the gap between spoken instructions and textual understanding, our model contributes to a more accessible and engaging digital learning ecosystem.

II. Related Work

Automatic text summarization has been a prominent research area in natural language processing (NLP) for several decades. Early research predominantly focused on extractive summarization methods, where systems identified and selected representative sentences directly from the source text [1]. These approaches employed statistical techniques like term frequency–inverse document frequency (TF–IDF) and graph-based algorithms such as TextRank [2] or LexRank [3]. While effective in some contexts, these methods often failed to generate coherent and contextually meaningful summaries for complex inputs.

The emergence of deep learning introduced transformative approaches to summarization. Sequence models based on recurrent neural networks (RNNs) and long short-term memory (LSTM) units enabled the development of abstractive summarization systems [4]. These models could generate novel phrases and sentence structures, offering greater abstraction than extractive techniques. However, they struggled with long-term dependencies and exhibited high computational demands during training [5].

A significant enhancement to sequence-to-sequence models came with the attention mechanism [6], which allowed models to dynamically focus on relevant parts of the input during decoding. This led to substantial improvements in the fluency and coherence of summaries. Building on this, the introduction of the Transformer architecture by Vaswani *et al.* [7] revolutionised NLP by replacing recurrence with self-attention and enabling highly parallelisable training.

The Transformer laid the groundwork for pre-trained models like BERT [8], GPT [9] and T5 [10]. BERT, in particular, proved effective for various NLP tasks, including summarization [11]. Models like BERTSUM adapted BERT for document-level summarization tasks by adding positional embeddings and segment-level classification heads [11].

Despite these advancements, most summarization models are designed and trained on structured text, such as news articles, legal documents or research papers [12]. In contrast, instructional video transcripts generated via automatic speech recognition introduce significant challenges—such as disfluencies, colloquial speech and lack of punctuation [13]. These characteristics reduce the effectiveness of traditional models trained on grammatically clean text.

Moreover, past research in video summarization has largely focused on visual content—selecting keyframes or scenes based on motion or audio cues [14]. While beneficial for visual-heavy domains, such methods are less suitable for instructional content where verbal narration contains the core instructional value.

Recent work has begun to explore multimodal summarization that incorporates both audio and visual cues [15]. Datasets like How2 [16] and HowTo100M [17] provide large-scale multimodal instructional data. However, high-quality human-labelled summaries remain scarce, limiting model performance on real-world ASR-generated transcripts.

Our approach bridges this gap by fine-tuning BERT-based summarization models on hybrid datasets that span structured text (e.g., CNN/DailyMail [12]), semi-structured text (e.g., WikiHow [18]) and noisy transcripts (e.g., How2 [16]). Through domain-specific preprocessing and curriculum learning, we improve semantic quality and context awareness in the generated summaries.

III. Methodology

Our methodology consists of four core components: (1) data collection and dataset creation, (2) preprocessing and cleaning of transcripts, (3) fine-tuning summarization models and (4) evaluation using both automated and human-based scoring methods.

A. Data Collection

To train and evaluate our summarization model effectively, we aggregated diverse and complementary datasets encompassing both textual and audio-visual instructional content. The datasets used are as follows:

- **CNN/DailyMail:** This corpus contains over 300 000 news articles paired with human-written summaries. Each article averages approximately 119 words, offering structured journalistic content useful for pretraining.
- **WikiHow:** A large-scale repository of more than 200 000 how-to articles with stepwise summaries. The data are semi-structured and cover a wide spectrum of instructional domains.
- **How2 Dataset:** Comprising approximately 8 000 YouTube videos with an average duration of 90 seconds and human-written summaries, this dataset includes both speech transcripts and paired descriptions created by video authors.
- **HowTo100Million:** We sampled and processed 5 195 videos from the publicly available HowTo100M dataset. These videos span 140+ instructional categories, and we extracted auto-generated speech-to-text transcripts for use in training and evaluation.

Table I provides a summary of key statistics across all datasets.

TABLE I
SUMMARY OF DATASETS USED FOR TRAINING AND EVALUATION

Dataset	Number of samples	Average length (words)	Summary type
CNN/DailyMail	300 000+	119 (article), 83 (summary)	Human-written
WikiHow	200 000+	Varied	Instructional, structured
How2	8 000	291 (transcript), 33 (summary)	Human-written
HowTo100M (sample)	5 195	259–859 (transcript)	ASR-based

B. Preprocessing

Given the noisy nature of ASR-generated transcripts, particularly from the HowTo100M and How2 datasets, extensive preprocessing was performed. This stage included:

- **Punctuation restoration:** ASR transcripts often lack proper sentence boundaries and punctuation. We used SpaCy and Stanford CoreNLP to segment sentences accurately and restore syntactic structure.
- **Entity detection and removal:** Many transcripts begin with informal introductions (e.g., "Hi, I'm John"), mention personal names or irrelevant greetings. We removed such elements to focus solely on instructional content.
- **Sentence boundary correction:** We employed rule-based heuristics and pre-trained models to delineate sentence boundaries reliably.
- **Filler removal:** We removed filler words (e.g., "um", "uh"), repeated phrases and off-topic remarks using lexical filters and part-of-speech tagging.
- **Text normalisation:** Abbreviations, contractions and non-standard spellings were expanded to ensure consistency.

This preprocessing pipeline significantly improved the quality and coherence of transcripts before feeding them into our summarization model.

C. Fine-Tuning Summarization Models

We employed a BERT-based encoder–decoder architecture to generate abstractive summaries. The encoder was initialised from the BERT base (uncased) model, while the decoder was a Transformer with six layers initialised randomly. Our fine-tuning process followed a curriculum-based strategy detailed in Section IV.

D. Evaluation

To measure the effectiveness of the summarization model, we used both automatic and human evaluations. Automatic metrics included ROUGE-1, ROUGE-L and the Content F1 score. Human evaluations involved blind assessments of fluency, informativeness, conciseness, realism and clarity, as described in Section V.

IV. IMPLEMENTATION

The implementation of our abstractive summarization model followed a carefully designed multi-stage training strategy, with the goal of transferring knowledge from well-structured domains (e.g., news articles) to less structured, informal instructional video transcripts. This staged curriculum learning approach allowed the model to incrementally adapt to increasing levels of complexity in language, structure and domain-specific terminology.

A. Training Phases

The training was divided into three major stages, each targeting a specific dataset designed to incrementally improve the model's generalisation and semantic coherence:

- **Stage 1: CNN/DailyMail** – We began by training the model on the CNN/DailyMail dataset, which contains formal, well-structured news articles paired with summaries. This stage helped initialise the model with general language understanding and summarization skills.
- **Stage 2: WikiHow** – In the second stage, we fine-tuned the model using the WikiHow dataset, which introduces a shift in domain towards instructional and how-to content. These articles feature a semi-structured format and domain-specific language.
- **Stage 3: How2 + HowTo100M** – Finally, the model was trained on ASR-generated transcripts from the How2 dataset and our sampled subset of the HowTo100M dataset. These transcripts are conversational, noisy and lack formal structure, representing real-world conditions for transcript-based summarization.

The effectiveness of this order-preserving training strategy was evident in the model's ability to handle noisy and informal language more fluently. We observed that reversing or randomising the dataset order significantly degraded performance, emphasising the importance of a curriculum-based fine-tuning approach.

B. Training Configuration

All training experiments were conducted using a 4-GPU Linux-based system (NVIDIA Tesla V100), leveraging PyTorch and the HuggingFace Transformers library. We used the BERT-base-uncased model as the encoder and a Transformer decoder initialised randomly.

TABLE II
TRAINING CONFIGURATION SUMMARY

Parameter	Value
Total training steps	210 000
Batch size	50
Epochs per stage	10–20
Optimizer	Adam
Learning rate (encoder)	0.002
Learning rate (decoder)	0.2
GPU configuration	4 × NVIDIA V100
Pretrained encoder	BERT base (uncased)
Decoder architecture	Transformer (6 layers)

C. Multi-Stage Training Pipeline

Figure 1 illustrates the complete training flow of our model. The input datasets are progressively introduced in a specific sequence, each contributing to a different learning goal — from structural coherence to domain specificity to transcript fluency.

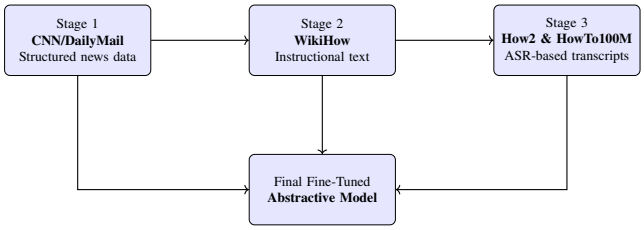

Fig. 1. Multi-stage training pipeline. Each stage progressively introduces data from increasingly informal domains, starting with structured news articles and culminating in noisy ASR-generated transcripts.

D. Performance Observations

During training, we monitored both training loss and ROUGE metrics on validation subsets after each stage. We observed that:

- The model trained solely on CNN/DailyMail achieved high scores on that domain but failed to generalise to informal or instructional data.
- Fine-tuning with WikiHow improved the model's performance on structured how-to tasks.
- Including the How2 and HowTo100M datasets significantly enhanced the model's fluency and realism when summarizing noisy ASR transcripts.

The progressive nature of training allowed the model to retain general language understanding from news and adapt it effectively to domain-specific and noisy instructional settings.

V. Results

In this section, we present both the quantitative and qualitative evaluation results of our summarization models across multiple datasets and configurations. Our primary metrics for automatic evaluation include ROUGE-1, ROUGE-L and Content F1 scores. We also conducted human evaluations to assess the fluency, informativeness, conciseness, realism and clarity of the generated summaries.

A. Quantitative Evaluation

The model trained using a curriculum learning approach—starting from CNN/DailyMail, followed by WikiHow and concluding with How2/HowTo100M transcripts—outperformed all other configurations. The highest performance was achieved when all datasets were included in a fixed training sequence without shuffling.

TABLE III
Performance comparison of various training configurations

Model configuration	ROUGE-1	ROUGE-L	Content F1
BertSum on CNN/DailyMail only	22.47	20.07	26.0
BertSum + CNN/DM + WikiHow	26.32	22.47	28.1
Full dataset + ordered training	**48.26**	**44.02**	**36.4**
Full dataset + shuffled	42.15	39.67	30.7
Pointer generator + WikiHow	28.53	26.54	29.3

As Table III shows, the use of multiple datasets in an ordered training pipeline significantly improved semantic quality and structural coherence. The BertSum model trained only on CNN/DailyMail failed to generalise well to conversational or instructional transcripts, while the fully trained model achieved ROUGE-1 of 48.26 and Content F1 of 36.4—demonstrating both high lexical overlap and semantic richness.

B. Human Evaluation

We conducted blind evaluations with 31 human participants. Each participant was presented with a set of 25 summaries—some machine-generated, others written by humans. Participants were asked to:

- Identify whether the summary was AI- or human-generated (Turing test).
- Rate summaries on five dimensions: fluency, informativeness, conciseness, realism and clarity (scale: 1–5).

The human evaluation results are summarised in Figure 2.

While human-written summaries scored marginally higher overall, AI-generated summaries were often indistinguishable in blind tests—indicating the model's effectiveness in generating fluent, human-like outputs.

VI. Discussion

The results of our experiments highlight several important observations regarding summarization model performance on instructional video transcripts:

A. Generalisation Challenges with ASR Transcripts

Although BERT-based models have demonstrated exceptional performance in many NLP tasks, applying them directly to ASR-generated transcripts presents specific challenges. These transcripts typically lack structure, include colloquialisms and often suffer from punctuation errors and background noise. Pretrained models like BertSum, which are optimised on clean, well-formatted text (e.g., news articles), struggle when exposed to such informal, noisy data without appropriate domain adaptation.

B. Importance of Dataset Diversity and Ordering

The success of our curriculum-based training strategy reinforces the hypothesis that the order and domain diversity of training data play a crucial role in performance. Training first on structured datasets allows the model to grasp syntactic consistency, while progressive exposure to informal instructional data helps refine the decoder's handling of less predictable input.

C. Role of Preprocessing in Improving Fluency

Preprocessing pipelines, including sentence boundary correction, filler removal and entity normalisation, were instrumental in enhancing output quality. Our experiments showed that even minimal preprocessing can improve ROUGE scores by 2–4 points and significantly reduce semantic hallucination in generated summaries.

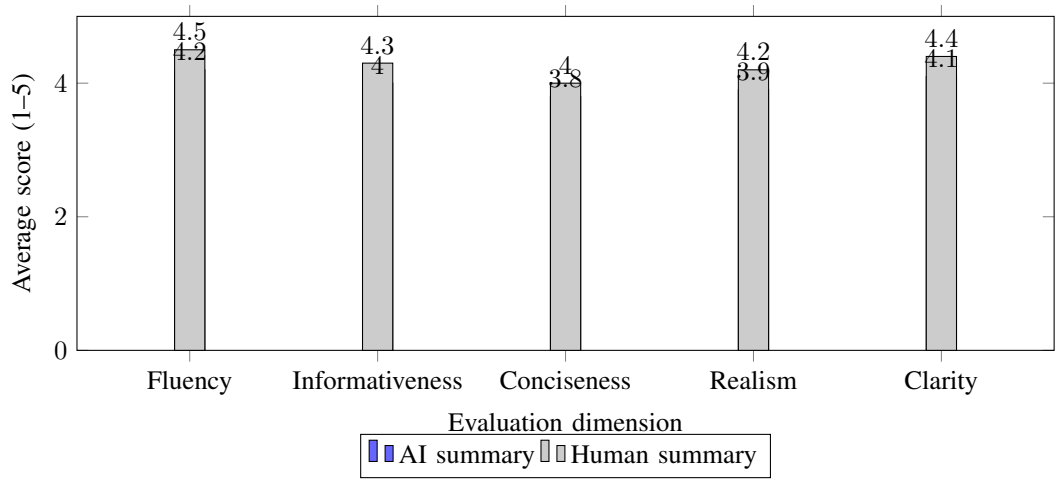

Fig. 2. Average human evaluation scores for AI-generated versus human-written summaries across five dimensions.

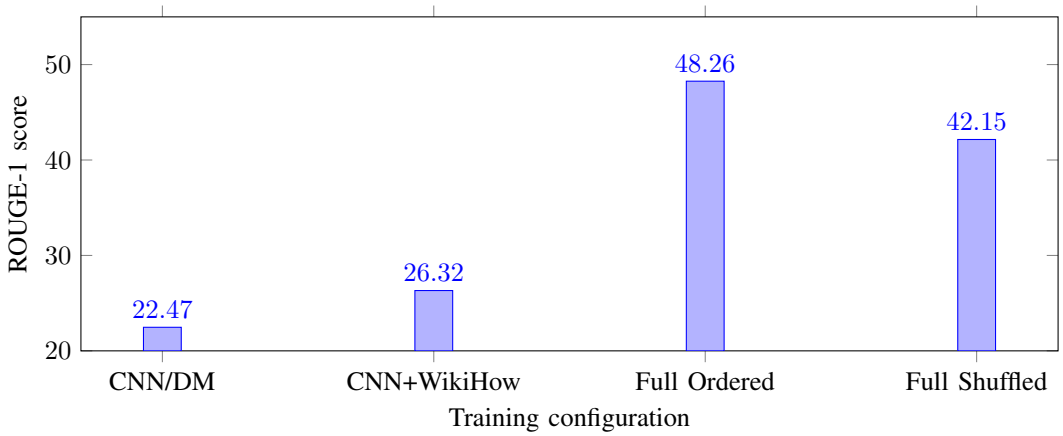

Fig. 3. Comparison of ROUGE-1 scores across different training configurations. Ordered training yields the highest ROUGE-1 scores.

D. Model Comparison Insights

Figure 3 visually compares the performance trends across different model configurations.

From the chart, it is evident that model robustness is maximised through domain progression, whereas random training can lead to overfitting or poor generalisation.

VII. Conclusion

In this study, we proposed a robust, scalable and domain-adaptive framework for abstractive summarization of instructional video transcripts using a fine-tuned BERT-based architecture. Our approach bridges the gap between traditional text summarization and the unique challenges posed by ASR-generated instructional transcripts, which are often informal, unstructured and linguistically noisy.

Through a curriculum-based training strategy that incorporated datasets like CNN/DailyMail, WikiHow, How2 and a curated subset of the HowTo100Million corpus, we successfully enhanced the model's generalisability from formal written language to spontaneous, conversational speech data. The ordered progression from clean to noisy domains allowed our model to retain structural understanding while adapting to domain-specific nuances.

Quantitative evaluations using ROUGE-1, ROUGE-L and Content F1 metrics demonstrated significant improvements when combining diverse datasets with careful preprocessing. The highest-performing configuration yielded a ROUGE-1 score of 48.26 and a Content F1 score of 36.4. Furthermore, human evaluations revealed that summaries generated by our model were often indistinguishable from human-authored ones in terms of fluency, informativeness and realism.

Our framework also includes a novel preprocessing pipeline to improve sentence boundary detection, remove non-essential utterances and normalise textual noise—all of which significantly contribute to better summary generation. Taken together, these components form a comprehensive solution for enhancing the accessibility and usability of instructional video content through automated summarization.

VIII. Future Work

While the results presented in this study are promising, several avenues exist for future exploration and enhancement:

- **Multimodal summarization:** Incorporating visual and audio cues alongside text can provide a more holistic understanding of instructional content. Leveraging video frames and speech intonation may improve summary contextualisation.
- **Real-time summarization:** Developing a low-latency version of our model suitable for live or real-time summarization could benefit educational livestreams, virtual classrooms and webinars.
- **Cross-lingual and multilingual summarization:** Extending the model to support transcripts in multiple languages, particularly those with limited training resources, will enhance the accessibility of educational content on global platforms.
- **Personalised summarization:** Integrating user preferences and prior learning history into the summarization process may lead to more customised, learner-centric summaries that adapt to different skill levels.
- **Robustness against ASR errors:** While our preprocessing techniques mitigate some noise, future research can focus on designing models inherently robust to ASR inaccuracies, such as mispronunciations, accent variations and background noise.
- **Deployment as a public API or plugin:** Packaging the trained summarizer as a user-friendly API or browser plugin could enable educational institutions, creators and learners to easily integrate summarization into their existing workflows.

By exploring these directions, we aim to further advance the effectiveness, scalability and inclusivity of automated summarization technologies, contributing to more accessible and structured digital learning environments.

REFERENCES

[1] A. Nenkova, "Automatic text summarization of newswire: Lessons learned from the document understanding conference," *AAAI*, 2005.

[2] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *EMNLP*, 2004, pp. 404–411.

[3] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of Artificial Intelligence Research*, vol. 22, pp. 457–479, 2004.

[4] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *NeurIPS*, 2014, pp. 3104–3112.

[5] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 1073–1083.

[6] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, 2015.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2019.

[9] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *OpenAI Blog*, vol. 1, no. 8, 2018.

[10] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *JMLR*, vol. 21, no. 140, pp. 1–67, 2020.

[11] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," *arXiv preprint arXiv:1908.08345*, 2019.

[12] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," *NeurIPS*, vol. 28, pp. 1693–1701, 2015.

[13] F. Ladhak, E. Durmus, K. McKeown, and D. Downey, "Exploring content selection in summarization of scientific documents," *arXiv preprint arXiv:2010.09174*, 2020.

[14] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *ECCV*, 2016, pp. 766–782.

[15] S. Palaskar, J. Libovický, S. Gella, and F. Metze, "Multimodal abstractive summarization for how2 videos," *arXiv preprint arXiv:1906.02081*, 2019.

[16] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze, "How2: A large-scale dataset for multimodal language understanding," *arXiv preprint arXiv:1811.00347*, 2018.

[17] 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," *arXiv preprint arXiv:1906.03327*, 2019.

[18] M. Koupaee and W. Y. Wang, "Wikihow: A large scale text summarization dataset," *arXiv preprint arXiv:1810.09305*, 2018.