Video Content Summarization with Large Language-Vision Models

Kelley Lynch, Bohan Jiang, Ben Lambright, Kyeongmin Rim, James Pustejovsky
Department of Computer Science
Brandeis University
Waltham, Massachusetts
{kmlynch*CA,bohanjiang,blambright,krim,jamesp}@brandeis.edu

***Abstract*—We present a modular pipeline for summarizing broadcast news videos using large language and vision models, specifically integrating Whisper for ASR, TransNetV2 for shot segmentation, LLaVA for image captioning, and LLaMA for generating structured summaries. Implemented within the CLAMS platform using the Multimedia Interchange Format (MMIF) for component interoperability, our approach combines ASR transcriptions and image captions to enhance metadata extraction. We evaluated our pipeline with automated metrics based on user-generated Youtube video descriptons as well as human assessments. Our analysis highlights challenges with automated metrics and emphasizes the value of human evaluation for nuanced assessment. This work demonstrates the effectiveness of multimodal summarization for video metadata extraction and paves the way for enhanced video accessibility.**

***Index Terms*—Computer Vision, Speech Recognition, LLM, Summarization, Metadata Extraction, Audiovisual Indexing**

I. Introduction

Large multimodal models have recently demonstrated impressive capabilities across a wide range of tasks, including natural language understanding, computer vision, and cross-modal reasoning. In the realm of broadcast news, extracting meaningful metadata from videos is crucial for content management, retrieval, and enhancing user engagement. In this paper, we investigate the application of such models for metadata extraction from broadcast news videos by leveraging user-generated chapter-level summaries available on YouTube.

Specifically, we explore a pipeline that integrates **Whisper** [7] for automatic speech recognition (ASR), shot segmentation techniques, **LLaVA** [4] for shot-level image captioning, and **LLaMA** [10] for generating coherent video chapter summaries. Rather than evaluating ASR or captioning tools independently, we utilize these components collectively within a pipeline aimed at generating accurate and informative chapter summaries. By combining ASR transcriptions and image captions derived from shot segmentation, we aim to enhance video chapter summarization through structured generation techniques.

Our approach involves processing broadcast news videos uploaded to YouTube, where users have provided chapter-level summaries. These user-generated summaries serve as a benchmark to evaluate the effectiveness of our pipeline. We apply shot segmentation to divide the videos into meaningful segments, use ASR to transcribe spoken content, and generate image captions for each shot. These multimodal outputs are then integrated and fed into **LLaMA** to produce structured chapter summaries.

We implement our pipeline within the Computational Linguistics Applications for Multimedia Services (CLAMS) platform [8], demonstrating how our approach can be operationalized in a scalable and modular framework. To assess the performance of our system, we conduct evaluations on curated datasets of broadcast news videos, analyzing the impact of integrating these multimodal outputs on the quality of the generated summaries. In addition to automated metrics, we perform human evaluations to assess the coherence, relevance, and accuracy of the generated chapter summaries, providing a more comprehensive understanding of the pipeline's effectiveness.

Our results highlight the potential of zero-shot metadata extraction from broadcast news, offering valuable insights into the effectiveness and limitations of different summarization strategies for archiving and cataloguing use cases. This study contributes to the understanding of how multimodal models can be effectively combined in a pipeline to improve metadata extraction that can enhance the accessibility and discoverability of video content, paving the way for more efficient content management and user engagement in multimedia platforms.

II. Prior Work

VidChapters-7M has advanced the field of video chaptering by providing a large-scale dataset of user-annotated chapters across over 800,000 videos, covering diverse categories and enabling the development of models for temporal segmentation and chapter title generation [12]. The dataset leverages user-provided timestamps from online videos to align chapters with ASR transcripts, creating structured navigation for videos with inherent timestamps. In contrast, our work focuses on archival broadcast news videos, which lack these pre-existing timestamps and face unique challenges, including interspersed commercials, recording discontinuities, and varied digitization quality. Building on the foundation established by VidChapters-7M, we apply chaptering techniques to historical media, manually annotating timestamps to accommodate the

This work has been supported by grants from the Andrew W. Mellon Foundation.

complexities of broadcast news archives and enhancing accessibility to this unique video category.

[5] present a modular approach for multimodal summarization of TV shows, which highlights the flexibility of modular designs by dividing complex tasks into specialized subtasks. Their system includes distinct modules for scene boundary detection, scene reordering, visual captioning, dialogue summarization, and high-level summarization, allowing independent improvements in each module and adaptability across tasks. Inspired by this approach, our work similarly adopts a modular framework for summarizing archival broadcast news videos, where components like image captioning, ASR, and chapter summarization can operate separately. This modular structure supports greater interpretability and adaptability.

[2] address the challenges of evaluating LLM-based spoken document summarization systems by proposing a comprehensive human evaluation framework tailored to generative AI content. They highlight the limitations of automated metrics like ROUGE [3] or BERTScore [13], which may not adequately capture the nuances of summaries produced by large language models. By drawing on methodologies from the social sciences, they develop detailed evaluation criteria and best practices to ensure robustness, replicability, and trustworthiness in human evaluation studies. Their work includes case studies demonstrating the implementation of these methods in an industrial setting, emphasizing the importance of human-in-the-loop evaluations for assessing aspects such as fluency, coherence, relevance, and correctness of generated summaries. Inspired by their approach, we incorporate human evaluations in our study to assess the coherence, relevance, and accuracy of the chapter summaries produced by our pipeline. This allows us to obtain deeper insights into the effectiveness of our multimodal summarization strategy beyond what automated metrics can provide.

III. Task definition and experiment design

A. Data Annotation

Our dataset is based on 24 recorded news broadcasts sourced from The Museum of Classic Chicago Television's database, covering a range of regional and national news stations from the 1970s-1980s. The majority of these broadcasts are between 30-35 minutes in length, though five longer videos range from 45 to 100 minutes, resulting in an average video length of 41 minutes. Each video includes 25-50 lines of pre-annotated summary text that provides descriptions of video segments in chronological order, including both news content and commercials. Using each line of summary text as an individual chapter, we manually annotated the precise start and end times for each segment within the video. Given the straightforward nature of this alignment task, only one annotator was required per sample.

In total, our dataset comprises 1,061 timestamped chapters, with an average of 46.13 chapters per video and a typical chapter length of 54.57 seconds. The dataset includes both commercial and non-commercial content, with segment lengths and summary word counts varying significantly by type. Commercial and promotional segments average 29.48 seconds with 7.69 words in their summaries, while non-commercial news segments average 74.52 seconds with 16.11 words per summary. Across all chapters, the average summary length is 12.38 words. Figure 1 provides a visual summary of these chapter statistics, illustrating the distribution of chapter lengths and content types across the dataset.

IV. Implementation

In this section, we describe the implementation details of our multimodal summarization pipeline. We compare two pipeline configurations to evaluate the impact of integrating image captions and providing examples to **LLaMA** during summarization. Both pipelines leverage the CLAMS platform to ensure interoperability between components.

A. Pipeline Configurations

We implemented and compared two pipeline configurations:

1) **Monomodal Pipeline**: This baseline pipeline utilizes only the ASR transcriptions generated by **Whisper** as input to **LLaMA** for generating summaries by chapter. We employed the Whisper tiny model, setting the `no_speech_threshold` parameter to 0.6 and `condition_on_previous_text` to `True`.
2) **Multimodal Pipeline**: This pipeline incorporates both the ASR transcriptions and shot-level image captions. We performed shot segmentation using **TransNetV2** [9] model, which identifies shot boundaries within the video. From each detected shot interval, we extracted the middle frame as a representative key frame. We generated image captions for these key frames using **LLaVA-Next 1.6 7B**, specifically the model available on HuggingFace as `llava-hf/llava-v1.6-mistral-7b-hf`. For faster processing, we used the 4-bit quantized version of the model. The following parameters were set for caption generation: `num_beams` = 5, `max_length` = 200, `min_length` = 1, `repetition_penalty` = 1.5, `length_penalty` = 1.0, and `temperature` = 1. The combined textual data (ASR transcriptions and image captions) was then fed into **LLaMA** for summarization.

B. Summarization with LLaMA

For both pipelines, we used **LLaMA** to generate chapter summaries. Specifically, we employed an 8-bit quantized version of **Meta-Llama-3.1-8B-Instruct** for summarization. We also evaluated the impact of providing examples to **LLaMA** as part of the prompt. By including exemplar summaries within the prompt, we leveraged the in-context learning capabilities of **LLaMA**, aiming to improve the quality and consistency of the generated summaries.

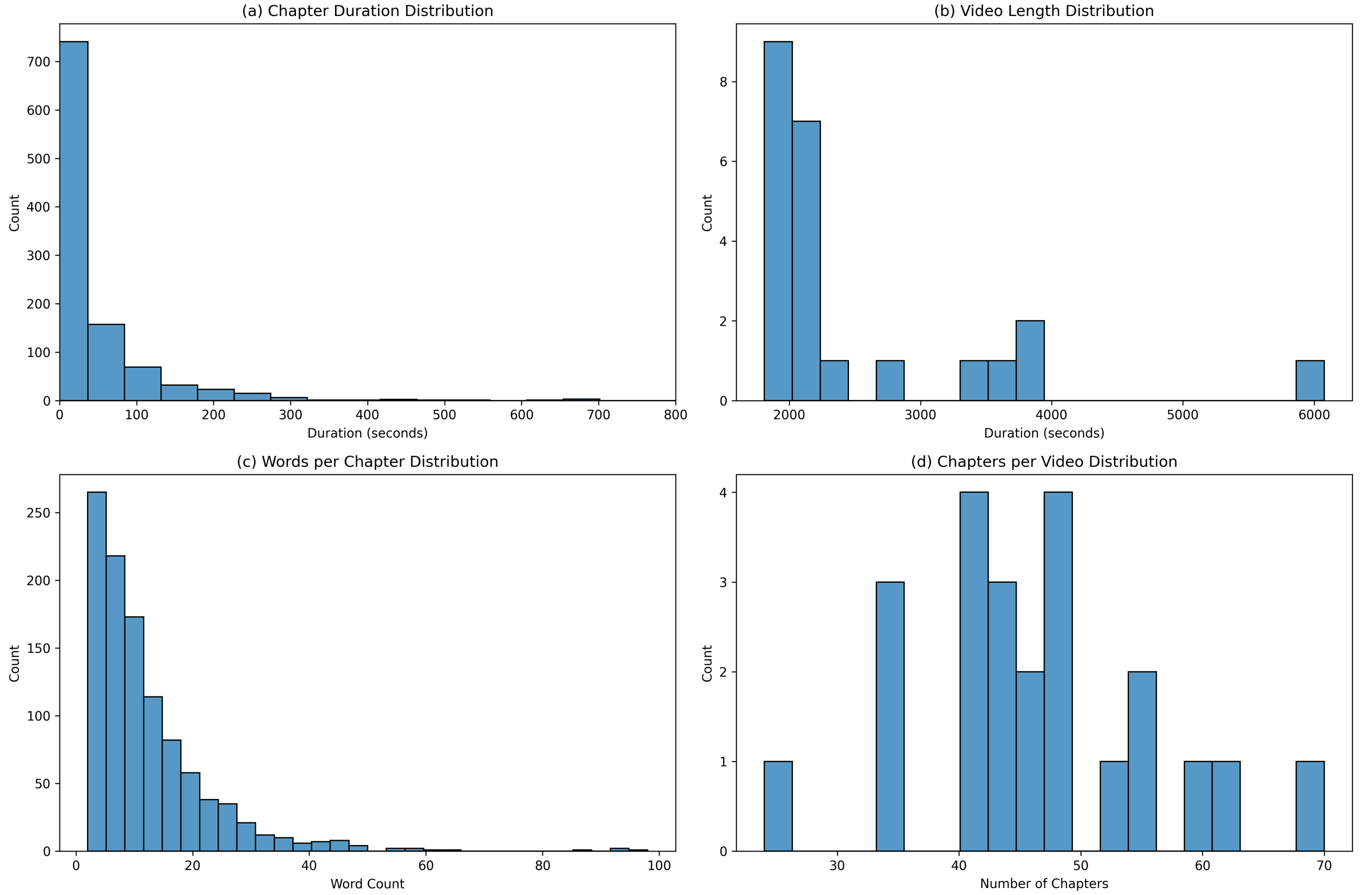

Fig. 1: A visual summary of chapter statistics, illustrating the distribution of chapter lengths and content types across the dataset.

C. CLAMS Platform and MMIF Integration

The CLAMS platform facilitated the integration of these components by providing standardized data exchange formats and processing pipelines. Each component was wrapped as a CLAMS app, allowing us to construct modular and reusable pipelines. This modularity ensures interoperability between different tools and simplifies the process of experimenting with various pipeline configurations by easily swapping or modifying individual components.

Furthermore, our use of the CLAMS platform and SDK enhances interoperability between components through the Multimedia Interchange Format (MMIF). MMIF is an annotation format for audiovisual media and associated text like transcripts and closed captions. It is a JSON format used to transport data between CLAMS apps and is inspired by, and partially based on, JSON-LD specifications. MMIF consists of two formal components: the JSON schema and the vocabularies (type hierarchies) [1]. The JSON schema defines the syntactic elements of MMIF, while the vocabularies define concepts and their ontological relations.

By adhering to the MMIF format, each component in our pipeline can produce and consume metadata in a consistent manner, facilitating seamless integration and communication between modules. Along with the formal specifications and documentation, a reference implementation of MMIF is provided through the CLAMS SDK, developed in Python and distributed via the Python Package Index [2]. This SDK functions as a software development kit that helps developers easily utilize various features of MMIF in their applications.

The use of MMIF ensures that outputs from one component can be directly utilized as inputs for another without the need for extensive data conversion or custom interfaces. This standardization not only streamlines the integration process but also enhances the scalability and extensibility of the pipeline, enabling easy addition or replacement of components as needed. Overall, the combination of the CLAMS platform and MMIF significantly enhances the modularity, interoperability, and reusability of our multimedia processing workflows.

[1] Formal specification and definitions are available at https://mmif.clams.ai

[2] Source code and documentation are available at https://sdk.clams.ai/

V. Evaluation

To assess the effectiveness of our multimodal summarization pipeline, we conducted both automated metric evaluations and human evaluations. The automated metrics provide quantitative measurements of summary quality based on n-gram overlaps and semantic similarity with reference summaries, while the human evaluations offer qualitative insights into the coherence, relevance, and accuracy of the generated summaries.

A. Automated Metrics

We employed standard automated metrics commonly used in text generation tasks to evaluate the quality of the generated chapter summaries. Specifically, we used BLEU [6] to measure n-gram overlaps, ROUGE [3] to evaluate recall-oriented overlaps with the reference texts, and Semantic Textual Similarity (STS) scores calculated using the SentenceTransformers model `all-MiniLM-L6-v2` [11]. These metrics provide a quantitative assessment of how closely the generated summaries match the reference summaries provided by users on YouTube. All metrics yield a single real number ranging between 0 and 1, where a larger number indicates higher similarity between the text pair.

To ensure a fair comparison, we set the summaries generated by **LLaMA** using only the ASR transcriptions as our baseline performance. We then compared the BLEU, ROUGE, and STS scores of this baseline with those of our proposed pipelines, which integrate ASR transcriptions and image captions derived from shot segmentation. Given that the lengths of summaries between the pipeline-generated summaries and the reference summaries vary significantly—the human-annotated references are often more concise than the model-generated summaries—we focused on unigram-level measurements for both BLEU and ROUGE metrics. This approach emphasizes the presence of relevant words and reduces the impact of length discrepancies on the evaluation.

Prior to computing the evaluation scores, we performed specific pre-processing steps on both the reference summaries and the generated summaries to ensure an accurate comparison. We stripped any leading or trailing punctuation and removed quotation marks. Additionally, we converted all text to lowercase. These steps eliminate trivial discrepancies that could skew evaluation metrics due to insignificant differences in punctuation, casing, or the presence of quotation marks.

Furthermore, we addressed the interchangeable use of the terms “promo” and “commercial” found in the reference summaries. To standardize the terminology and prevent semantic mismatches during evaluation, we substituted all instances of “promo” with “commercial” in both the reference and generated summaries. This normalization ensures that synonymous terms do not adversely affect the assessment of summary quality, allowing the evaluation to focus on the substantive content rather than lexical variations.

We analyzed a total of 3,180 segments to assess the effectiveness of our summarization pipeline. The overall performance of the system achieved an average BLEU score of 0.1219 (±0.1577), an average STS score of 0.4713 (±0.2640), and average ROUGE scores of 0.2300 (±0.2779) for ROUGE-1, 0.3214 (±0.2718) for ROUGE-2, and 0.2183 (±0.2161) for ROUGE-3.

When differentiating between commercial and non-commercial segments, we observed that commercial segments (1,410 in total) had slightly higher BLEU scores (0.1267 ± 0.1985) compared to non-commercial segments (1,770 in total), which had BLEU scores of 0.1181 ± 0.1153. In terms of STS scores, non-commercial segments had higher scores (0.4794 ± 0.2464) compared to commercial segments (0.4612 ± 0.2844). Regarding ROUGE scores, commercial segments outperformed non-commercial segments in ROUGE-1 (0.3171 ± 0.3600 vs. 0.1607 ± 0.1570) and ROUGE-3 (0.2555 ± 0.2682 vs. 0.1887 ± 0.1570), while ROUGE-2 scores were comparable between the two categories (0.3108 ± 0.2943 for commercial vs. 0.3299 ± 0.2521 for non-commercial). These findings indicate that our pipeline performs consistently across different types of segments, with a slight advantage in handling commercial content.

Table I presents a tabular view of the automated evaluation results for different pipeline configurations.

B. Human Evaluation

While automated metrics provide useful quantitative insights, they may not fully capture the nuances of summary quality, especially for content generated by large language models. To gain a deeper understanding of the effectiveness of our pipeline, we conducted human evaluations based on a comprehensive rubric inspired by [2], who outline criteria and a methodology for human evaluations of LLM-based video summaries.

We annotated each segment summary and assigned a score based on four main evaluation categories:

1) **Quantity**: text length and content coverage.
2) **Quality**: content precision and model hallucination.
3) **Relevance**: salience and repetition.
4) **Manner**: text coherence and matching tones.

By adopting this rubric, we assessed the summaries on multiple dimensions, including the appropriateness of length, accuracy of information, relevance of content, and linguistic coherence. This systematic approach allowed us to capture aspects of summary quality that automated metrics might overlook, providing a more holistic evaluation of our pipeline’s performance.

Logistically, we used 4 university students to rate each segment-wise, short passage from the three pipelines. We allowed raters to see the reference summary on the side, but accompanying videos were not provided as part of the task. Raters were directed to give 0 or 1 for each rubric category, and then the overall rating of a passage was computed by adding all those scores. Hence, the final rating is a single number between 0 and 4.

After initial training sessions, raters were randomly assigned a certain number of video episodes. To measure the reliability of the rating process, at least 2 raters were assigned to a video.

TABLE I: Summary of Evaluation Scores for Different Pipeline Configurations

Pipeline Configuration	**Avg BLEU**	**Std BLEU**	**Avg STS**	**Std STS**	**Avg R-1**	**Std R-1**	**Avg R-2**	**Std R-2**	**Avg R-L**
Transcript + Caption	0.1125	0.1498	0.4805	0.2571	0.1917	0.2464	0.3458	0.2795	0.1985
Transcript + Caption + Examples	0.1146	0.1462	0.4516	0.2656	0.2103	0.2667	0.3103	0.2684	0.2030
Transcript + Examples	0.1386	0.1745	0.4819	0.2685	0.2881	0.3076	0.3082	0.2656	0.2534

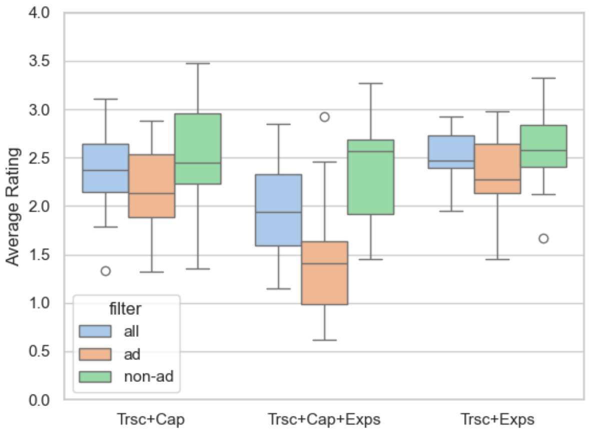

Fig. 2: A visual summary of human evaluation statistics, illustrating the distribution of ratings across the generated summaries.

Once all segments in a video were rated, we computed the arithmetic mean of the ratings to give an overall rating for the video. Then we used Krippendorff’s α [1] with interval-scaling distance function to measure the inter-annotator agreement (IAA). Likewise, IAA was computed at each segment-level and then averaged at the video-level. The average α was 0.637, indicating moderate reliability of the rating rubrics[3].

Table II shows the average ratings by different pipelines and content elements. Figure 2 is a visualization of the human ratings using a boxplot scheme. α scores from the table indicate moderate agreement across the board. Annotators generally agreed whether each summary was good (3-4) or bad (0-1). As for the scores, the pipeline including the transcript, caption, and examples performed the worst in all circumstances. On average, the pipeline including the transcript and examples best met our human evaluation metrics. Additionally, there was a slight increase in the quality of non-ad segment summaries compared to ads (2.520 versus 2.025, respectively). With an average human evaluation score per segment of 2.288, the pipeline was able to correctly summarize some aspects of the segments.

VI. Discussion

A. *Challenges with Reference Summaries*

An additional challenge in evaluating our summarization pipeline arises from the presence of extraneous information in the reference summaries. These summaries often include details that are irrelevant to the core content, such as the names of voiceover artists, segments removed due to copyright restrictions, and links to other videos. For example:

> Commercial: Annie at Arie Crown Theatre (through June 9th)

In this instance, the date is unlikely to be included in a summary focused on content. Another example includes:

> Commercial: Yellow Pages - Get in the Yellow (featuring Joe Mantegna) (sung to the tune of ”We’re in the Money”)

Here, the musical elements and the specific phrasing of the commercial name ”Get in the Yellow” are not typically relevant to the summary generation, especially since our prompt instructs to generate summaries in the format “Commercial: [Brand/Product Name].” Additionally, some user-generated summaries contain hyperlinks, such as:

> Commercial: Pursettes Premenstrual Tablets (posted separately here: https://youtu.be/bgimjaxuk44)

And examples of copyright removal include:

> Spotlight on Truman College art exhibit showcasing depression-era artists; Buddy Black interviews WPA artists Frances Badger and William Carter about their trials, tribulations, and careers (snippet of Dave Brubeck version of ”Brother, Can You Spare a Dime” had to be futzed with for copyright reasons)

These elements introduce noise into the evaluation process, as they do not contribute to the informative content of the summaries and can potentially mislead automated metrics.

B. *Impact of Shot Captions on ASR Errors*

Our multimodal pipeline leverages shot-level image captions to mitigate errors inherent in Automatic Speech Recognition (ASR). Specifically, commercials often present challenges where ASR transcripts fail to capture the full product or brand

[3] $\alpha = 1$ means perfect agreement, $\alpha = 0$ means agreement equivalent to chance, and negative $\alpha < 0$ means the annotation is less reliable than random score assignment

TABLE II: Summary of Human Evaluation Scores for Different Pipeline Configurations

Pipeline Configuration	**Avg** α [-1,1]	**Avg Rating** [0,4]	**Avg Ad-only**	**Avg Non-Ad**
Transcript + Caption	0.589	2.376	2.184	2.559
Transcript + Caption + Examples	0.662	1.943	1.479	2.387
Transcript + Examples	0.669	2.507	2.346	2.598
Average	0.637	2.288	2.025	2.520

names accurately. For instance, an ASR system might transcribe a commercial as “commercial: yellow pages” instead of the more precise “commercial: Yellow Pages - Get in the Yellow.” However, the corresponding image captions generated by **LLaVA** reliably include the complete product or brand names, ensuring that the generated summaries reflect the correct and full information. By integrating these image captions with the ASR transcriptions, our pipeline effectively corrects partial or inaccurate transcriptions, leading to more accurate and informative chapter summaries. This complementary use of visual information not only enhances the semantic accuracy of the summaries but also reduces the likelihood of omitting crucial details due to ASR limitations.

C. Limitations of Automated Metrics

One significant challenge we encountered with automated evaluation metrics like BLEU and ROUGE is their tendency to penalize summaries that differ in granularity and verbosity from the reference texts. These metrics rely heavily on n-gram overlap, which means that even semantically correct summaries can receive low scores if they use different wording or provide more detailed information than the reference. For instance, consider the following model-generated summary from our pipeline:

> The news segment discusses a special election in Detroit, where voters approved a 1% increase in the city’s income tax to address financial concerns. The measure passed with 54% approval and 45% opposition, with Mayor Young attributing the outcome to a tax revolt. The increase is expected to spare the city from severe cuts in services and prevent it from being forced into state receivership.

When compared to the more concise reference summary:

> Detroit voters approve 1% tax increase to offset budget shortfalls in that city,

the model’s output received a BLEU score of just 0.11. Despite the low score, the generated summary is semantically accurate and provides additional relevant details about the election results and their implications. This example highlights how automated metrics may not fully capture the quality of summaries that are longer or more detailed than the references, as they tend to favor lexical similarity over semantic correctness. This limitation motivates our use of human evaluations to more accurately assess the semantic content and overall quality of the generated summaries.

VII. Conclusion

This paper presented a multimodal pipeline for generating metadata-rich summaries of broadcast news videos, leveraging state-of-the-art models and modular design for adaptability. We compared an ASR-only pipeline with a multimodal variant that integrates ASR transcriptions and image captions, finding that the latter provides more accurate and informative summaries, particularly in identifying commercial content. Implementing the pipeline within the CLAMS platform using MMIF ensured efficient data interchange between components, enhancing flexibility and scalability.

Automated evaluations with BLEU and ROUGE provided baseline insights, though their limitations in capturing semantic accuracy highlighted the need for human assessments. Our human evaluation framework, inspired by [2], offered a comprehensive view of summary quality, addressing issues like ASR errors and extraneous information in reference summaries. This combination of automated and human evaluations enabled us to refine the pipeline, underscoring the importance of multimodal integration in improving metadata extraction.

Our approach showcases the potential for modular, multimodal pipelines in video summarization, opening avenues for further refinement. Future work will explore more advanced models and improved evaluation metrics to enhance semantic accuracy, enabling broader applications across diverse video content types and improving user engagement through accessible, metadata-rich video content.

Appendix A
Prompt Structures

In this appendix, we describe the structure of the prompts used in our summarization pipeline. We employed three different prompt templates to instruct **LLaMA** during the summarization process. These prompts were carefully designed to guide the model in generating summaries that align with the style of the summaries in our dataset.

To help the model produce summaries in the desired style, we included example summaries in some of the prompts. These examples were selected from a video not included in our evaluation dataset but from the same YouTube channel. This approach aimed to provide the model with stylistic cues similar to the target summaries without overlapping with the evaluation data.

A. Prompts Used

1) **Transcript Only with Examples Included**

Prompt 1: Transcript Only with Examples

```
System Prompt:
"You are a helpful assistant for summarizing
    news stories."

User Prompt:
You will be shown a transcript for a single
    news segment. You should generate a short
    summary of the segment. If the segment is a
     commercial, say Promo followed by the
    product name.
{transcript}

Examples:
Gary Shepard on the rescue of two whales from
    the ice in Barrow, Alaska; interview with
    Ron Morris of NOAA Fisheries.
FAA proposes "major surgery" for Boeing 737
    jetliners (with a proposal to replace 7,200
     rivets in key joints of 291 737's), as
    reported by Bettina Gregory
31 members of House Armed Services Committee
    urge Reagan to improve safety procedures at
     nuclear weapons manufacturing facilities
```

2) **Transcript and Captions with Examples Included**

Prompt 2: Transcript and Captions with Examples

```
User Prompt:
You will be shown a transcript for a single
    news segment. You should generate a short
    summary of the segment. If the segment is a
     commercial, say Promo followed by the
    product name.
{transcript}
{frame_captions}

Examples:
Gary Shepard on the rescue of two whales from
    the ice in Barrow, Alaska; interview with
    Ron Morris of NOAA Fisheries.
FAA proposes "major surgery" for Boeing 737
    jetliners (with a proposal to replace 7,200
     rivets in key joints of 291 737's), as
    reported by Bettina Gregory
31 members of House Armed Services Committee
    urge Reagan to improve safety procedures at
     nuclear weapons manufacturing facilities
```

3) **Transcript and Captions without Examples**

Prompt 3: Transcript and Captions without Examples

```
User Prompt:
You will be shown frame level captions and a
    transcript for one segment of a news video.
     You should generate a short summary of the
     segment. If the segment is a commercial,
    say Promo followed by the product name. The
     content is for only one segment. If there
    is no content, say None.
{frame_captions}
{transcript}
```

B. Prompt Design Considerations

We carefully designed the prompts to guide the model in generating concise and relevant summaries. Including examples in the prompts was intended to provide the model with concrete instances of the desired output style and format.

The example summaries were chosen from a video not included in our evaluation dataset but from the same YouTube channel. This selection aimed to:

- **Provide stylistic consistency**: By using examples from the same channel, we ensured that the model received stylistic cues similar to those in the target summaries.
- **Enhance model guidance**: The examples demonstrated how to handle different types of segments, including news reports and commercials, instructing the model on when to generate summaries and when to indicate a promo.

In the prompts that included both transcripts and frame captions, we provided the model with multimodal information to improve the quality of the summaries, especially in cases where ASR transcripts might be incomplete or contain errors.

In the prompt without examples, we tested the model's ability to generalize the task instructions without explicit stylistic examples, relying solely on the provided content and the initial instructions.

Overall, the prompt variations allowed us to evaluate the impact of including examples and additional information on the model's summarization performance.

References

[1] Andrew F Hayes and Klaus Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89, 2007.

[2] Margaret Kroll and Kelsey Kraus. Optimizing the role of human evaluation in llm-based spoken document summarization systems. In *Interspeech 2024*, pages 1935–1939, 2024.

[3] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[4] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[5] Louis Mahon and Mirella Lapata. A Modular Approach for Multimodal Summarization of TV Shows, August 2024. arXiv:2403.03823 [cs].

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[7] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[8] Kyeongmin Rim, Kelley Lynch, and James Pustejovsky. Computational linguistics applications for multimedia services. In Beatrice Alex, Stefania Degaetano-Ortlieb, Anna Kazantseva, Nils Reiter, and Stan Szpakowicz, editors, *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 91–97, Minneapolis, USA, June 2019. Association for Computational Linguistics.

[9] Tomáš Souček and Jakub Lokoč. TransNet V2: An effective deep network architecture for fast shot transition detection, August 2020. arXiv:2008.04838 [cs].

[10] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[11] Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. MiniLMv2: Multi-Head Self-Attention Relation Distillation for Compressing Pretrained Transformers, June 2021. arXiv:2012.15828.

[12] Antoine Yang, Arsha Nagrani, Ivan Laptev, Josef Sivic, and Cordelia Schmid. VidChapters-7M: Video Chapters at Scale, September 2023. arXiv:2309.13952 [cs].

[13] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.