ChapVidMR: Chapter-based Video Moment Retrieval using Natural Language Queries

Uday Agarwal[*]
Indian Institute of Technology
Jodhpur, Rajasthan, India
agarwaluday@iitj.ac.in

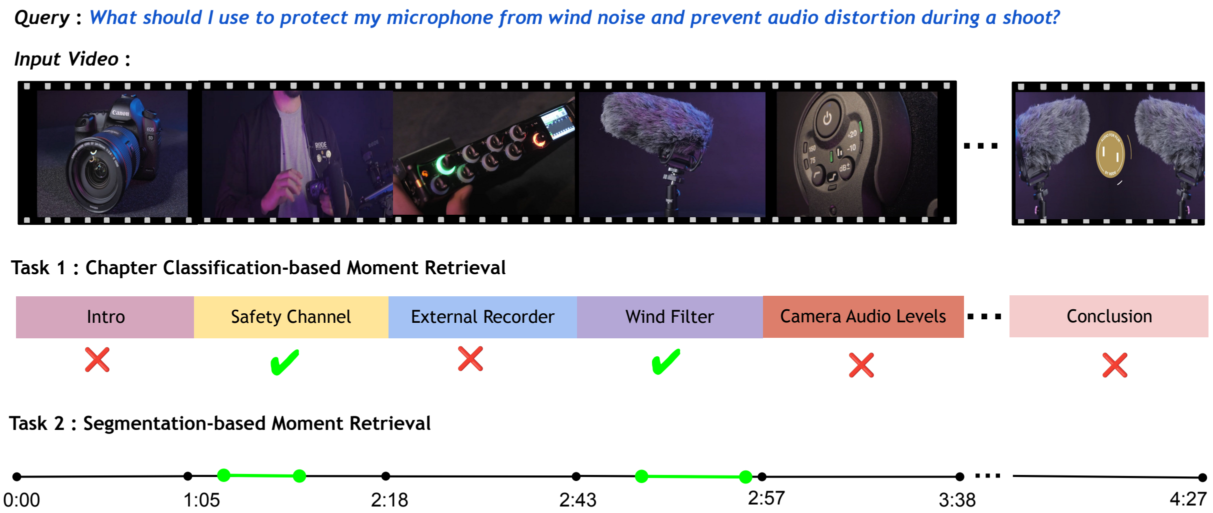
Yogesh Kumar[*]
Indian Institute of Technology
Jodhpur, Rajasthan, India
kumar.204@iitj.ac.in

Abu Shahid[*]
Indian Institute of Technology
Jodhpur, Rajasthan, India
shahid.3@iitj.ac.in

Prajwal Gatti[**]
University of Bristol
Bristol, UK
prajwal.gatti@bristol.ac.uk

Manish Gupta
Microsoft
Hyderabad, India
gmanish@microsoft.com

Anand Mishra
Indian Institute of Technology
Jodhpur, Rajasthan, India
mishra.iitj.ac.in



Query : What should I use to protect my microphone from wind noise and prevent audio distortion during a shoot?
Input Video :
Task 1 : Chapter Classification-based Moment Retrieval
Intro
Safety Channel
External Recorder
Wind Filter
Camera Audio Levels
Conclusion
Task 2 : Segmentation-based Moment Retrieval
0:00
1:05
2:18
2:43
2:57
3:38
4:27

Figure 1: Illustration of the proposed ChapVidMR dataset and tasks. The upper part of the figure visualizes a single sample of our dataset illustrating a video covering a technical review of a digital camera, segmented into chapters, with the entire video associated with a query. The bottom part of the figure highlights the two tasks enabled by the proposed dataset. **Task 1**: In the *Chapter Classification-based retrieval* task, the model classifies which of the chapters associated with a video are most relevant (highlighted by green checkmarks) to the query. **Task 2**: *Segmentation-based retrieval* task, where input videos are divided into multiple segments, and the highest-ranked ones, according to the query, are retrieved as desired moments (indicated by green lines).

ABSTRACT

Video Moment Retrieval (VMR) is the task of linking a query with a relevant moment from a video. Although, recently, there has been work on the VMR task where a query is linked to a single moment, the corresponding task where the query needs to be linked to multiple moments has been understudied. In this paper, we aim to work on the VMR task primarily by leveraging *chapters* of YouTube videos, i.e., video segments. YouTube chapters provide a meaningful segmentation of videos annotated by content authors. These annotated segmented regions help the viewer to navigate to a specific segment of the long video in which the user is interested. We present the ChapVidMR (Chapter-based Video Moment Retrieval) dataset, containing 10.8K user queries (obtained using GPT4) formed using multiple chapter names and other metadata extracted from videos using the YouTube API. Furthermore, we benchmark the proposed dataset on two VMR tasks: *chapter classification-based* VMR and *segmentation-based* VMR. In the chapter classification-based VMR task, the model classifies which of the chapters associated with a video are most relevant to the query. We represent a chapter by exploiting text (subtitles), audio, and visual

[*] These authors contributed equally to this work.
[**] This work was done while Prajwal Gatti was affiliated with IIT, Jodhpur.

ICVGIP'24, December 2024, Bangalore, India
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

(captions, video) modalities using state-of-the-art feature representation techniques and experiment with an exhaustive ablation for each modality. In segmentation-based VMR, the video is divided into various segments, and the segments most likely to answer the query are identified and returned. We benchmark our dataset on the state-of-the-art methods for segmentation tasks. We find that for the chapter classification-based VMR task, Sentence-BERT with subtitles and visual captions leads to the best results, while for the segmentation-based VMR, UniVTG is the most accurate. We make our code and data publicly available[1].

CCS CONCEPTS

• **Computing methodologies → Neural networks**; **Computer vision tasks**; **Visual content-based indexing and retrieval**; • **Information systems → Multimedia and multimodal retrieval**.

KEYWORDS

video moment retrieval, ChapVidMR, VMR, chapter-based retrieval

ACM Reference Format:
Uday Agarwal*, Yogesh Kumar*, Abu Shahid*, Prajwal Gatti**, Manish Gupta, and Anand Mishra. 2021. ChapVidMR: Chapter-based Video Moment Retrieval using Natural Language Queries. In *Proceedings of 15th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'24).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In the field of video content analysis [11, 15–18], the ability to search and retrieve specific segments from extensive video material has become increasingly important. This capability is crucial for various applications, including content moderation, educational resource management, and media production. The necessity for this video moment retrieval (VMR) task is driven by the vast amount of video content available and the need for efficient ways to navigate and extract relevant information. As the volume of video data continues to grow exponentially, traditional methods of manual searching and browsing have become impractical and time-consuming. This work introduces a different approach to this challenge by utilizing "chapters" in YouTube videos. Chapters in YouTube videos are predefined segments that the video owners have annotated, making it easier to find specific content within a video. These annotations not only enhance the viewing experience but also provide a structured way to segment videos, which is highly beneficial for automated processes. The use of chapters creates a natural index within videos, potentially simplifying the task of moment retrieval and improving the accuracy of search results.

We hypothesize that users are more likely to ask questions about specific segments in videos, especially those defined as chapters. This idea is based on the structured nature of chapters, which organize content into clear sections, making it easier for users to find and refer to particular topics or scenes in a video. The chapter structure mimics the organization of written content, allowing viewers to navigate long-form videos with greater ease and precision. This hypothesis is supported by initial observations of user behavior and feedback from content creators who have implemented chapter annotations. By focusing on these chapters, we aim to align our retrieval methods with how users typically search for information in video content. Additionally, leveraging chapter information could reduce computational complexity in VMR tasks, as it provides pre-defined semantic boundaries within the video content.

In this paper, we introduce the Chapter-based VMR (ChapVidMR) dataset, which is specifically designed to leverage chapter annotations. The dataset comprises user queries that are based on the names of multiple video chapters and other metadata obtained from the YouTube API. This approach allows us to capture the multi-faceted nature of user information needs, which often span across different segments of a video. The inclusion of chapter-based metadata provides a rich context for each query, enabling context-aware retrieval models. The unique aspect of ChapVidMR is that it extends beyond the traditional task of retrieving a single video segment. Instead, it focuses on identifying multiple relevant chapters that together provide a comprehensive answer to a user's query. This multi-chapter retrieval paradigm represents a significant shift from conventional VMR tasks, addressing the complexity of real-world information-seeking behavior. This approach mimics how users often seek information, making it a valuable resource for developing more effective video retrieval systems. Multi-moment retrieval not just enables natural queries but also helps users in finding more comprehensive and contextually relevant answers to their queries. To create ChapVidMR, we selected videos from a diverse set of sources and used GPT4 [19] to generate queries that relate to multiple chapters of a video. The selection process ensured a wide range of topics and video styles, enhancing the dataset's applicability to various domains. By utilizing GPT4, we were able to generate a large number of linguistically diverse and contextually relevant queries, simulating the variability in real user questions. This process ensured that the queries were natural and representative of genuine user inquiries. The queries required combining information from different chapters, reflecting the complex nature of real-world information needs. This complexity in the dataset encourages the development of more sophisticated retrieval algorithms that can understand and process multi-part queries, potentially leading to more robust and versatile VMR systems.

Based on ChapVidMR, we benchmarked it on two tasks: chapter classification-based VMR and segmentation-based VMR. In the chapter classification task, the model classifies which of the chapters associated with a video are most relevant to the query. Our proposed solution utilizes state-of-the-art multimodal features [7, 20, 29], including text, audio, and video data, to represent each chapter. The integration of these features is critical as it allows the model to assess the relevance of each chapter from different sensory inputs, thereby enhancing the accuracy of the classification. This is particularly evident in scenarios where audio cues or visual elements are significant indicators of the content's relevance to the query. Our results show that this approach is precise, particularly when leveraging subtitles and visual captions. On the other hand, segmentation-based VMR [15, 16, 18] involves first dividing the video into several segments and then determining segments that best address the user's query. This task tests the model's ability to finely segment video content based on the query's requirements.

In summary, our main contributions are as follows. (1) We introduce the ChapVidMR dataset that utilizes YouTube video chapters for generating user queries linked to multiple chapter names and

[1] https://github.com/vl2g/ChapVidMR

YouTube metadata. (2) Using ChapVidMR, we provide baseline solutions to the chapter classification-based VMR and segmentation-based VMR tasks, employing multimodal features. (3) We find that SentenceBERT and UniVTG lead to best results for the two tasks respectively. We make the code and dataset publicly available[1].

2 RELATED WORK

Video Moment Retrieval (VMR) systems [6, 8, 14–16, 18] localize temporal segments in a video given a textual description. Two popular strategies adopted by existing VMR systems are that of one-stage approaches and two-stage approaches. The two-stage approaches [5, 8, 21, 22, 26–28] entail generating a list of potential candidate moments which are later ranked according to various scores. These approaches typically employ a filtering mechanism to narrow down the extensive list of moments to those most likely to match the textual query, thus prioritizing efficiency and precision. Subsequently, more sophisticated algorithms assess the relevance of each candidate moment, enhancing the overall accuracy of the retrieval. The one-stage approaches [13–15, 18] process the entire video directly and return the start and end timestamps of the localized temporal segments. This method bypasses the preliminary selection phase, instead utilizing powerful models to parse and understand the video content in one go, which often results in faster response times but at the cost of computational intensity. In this work, we focused on the understudied task of retrieving multiple (two or more) segments that best answer a textual query.

In the domain of video moment retrieval (VMR), several datasets have been proposed, including DiDeMo [8], ActivityNet Captions [10], CharadesSTA [6], and TVR [14]. These datasets generally pair a single query with a video and often show a temporal bias, with moments more frequently tagged in the beginning than throughout the video. Such limitations may not fully replicate the diversity and complexity of real-world video querying scenarios. In response, the creation of the ChapVidMR dataset is designed to enhance the scope and utility of VMR research by incorporating multiple moments and queries for each video derived from YouTube chapters. This design choice reflects a more dynamic and varied set of use cases, from short informational clips to longer narrative sessions, ensuring that the dataset is representative of the typical user experience on video platforms. Each video in ChapVidMR serves as a more comprehensive resource for developing VMR systems, supporting nuanced interactions that require recognizing and correlating multiple thematic elements. To address these limitations, we introduce the ChapVidMR dataset, which utilizes YouTube chapters to collect multiple moments and corresponding queries per video. This approach allows for a broader scope of retrieval tasks, mimicking more realistic user interactions with video content. ChapVidMR includes videos that average 322.8 seconds in length and support multiple queries per video, each query averaging 17.7 words. This setup provides a richer dataset for VMR tasks, allowing for detailed query generation and robust segmentation and classification of video content. By doing so, ChapVidMR facilitates a more granular examination of how effectively VMR systems can handle varied and complex video contexts. Furthermore, the framework outlined in Fig. 3 ensures the scalability of ChapVidMR, facilitating an increase in the number of videos and queries. This scalability is crucial for developing robust VMR models that are adaptable to various video lengths and complexities.

The advent of Detection Transformers (DETR) and transformer-based pre-training [1, 4, 12], a model that employs the transformer encoder-decoder architecture and views object detection as a direct set prediction pipeline, enabled viewing the task of moment retrieval in the same light [15]. It simplifies the pipeline of localizing temporal segments in a video most relevant to a query by eliminating the need for anchor generation and non-maximum suppression. This streamlining not only reduces the complexity of the retrieval process but also enhances the precision by focusing directly on the relevancy of segments to the query. Additionally, the DETR framework allows for end-to-end training, which further aligns the model's learning with the specific nuances of the moment retrieval task. Despite these improvements, integrating DETR into video moment retrieval poses challenges, particularly in handling the vast amounts of video data and the diverse range of queries. The application of DETR-based approaches for downstream tasks has risen since then in both the domains, images [2, 3, 9] and videos [12, 18, 23]. This rise is indicative of the versatile capability of DETR to adapt to different media formats and extraction tasks, proving its utility beyond the initial scope of object detection. Despite its success, DETR-based approaches have their downside of slow convergence during training time. This issue is particularly pronounced in complex video datasets where the temporal dynamics and contextual variety demand extensive training iterations to achieve optimal performance. In this work, we experiment with state-of-the-art DETR-based approaches and report the results of their performance on the tasks of chapter classification-based VMR and boundary segmentation-based VMR.

3 CHAPVIDMR: CHAPTER-BASED VMR DATASET

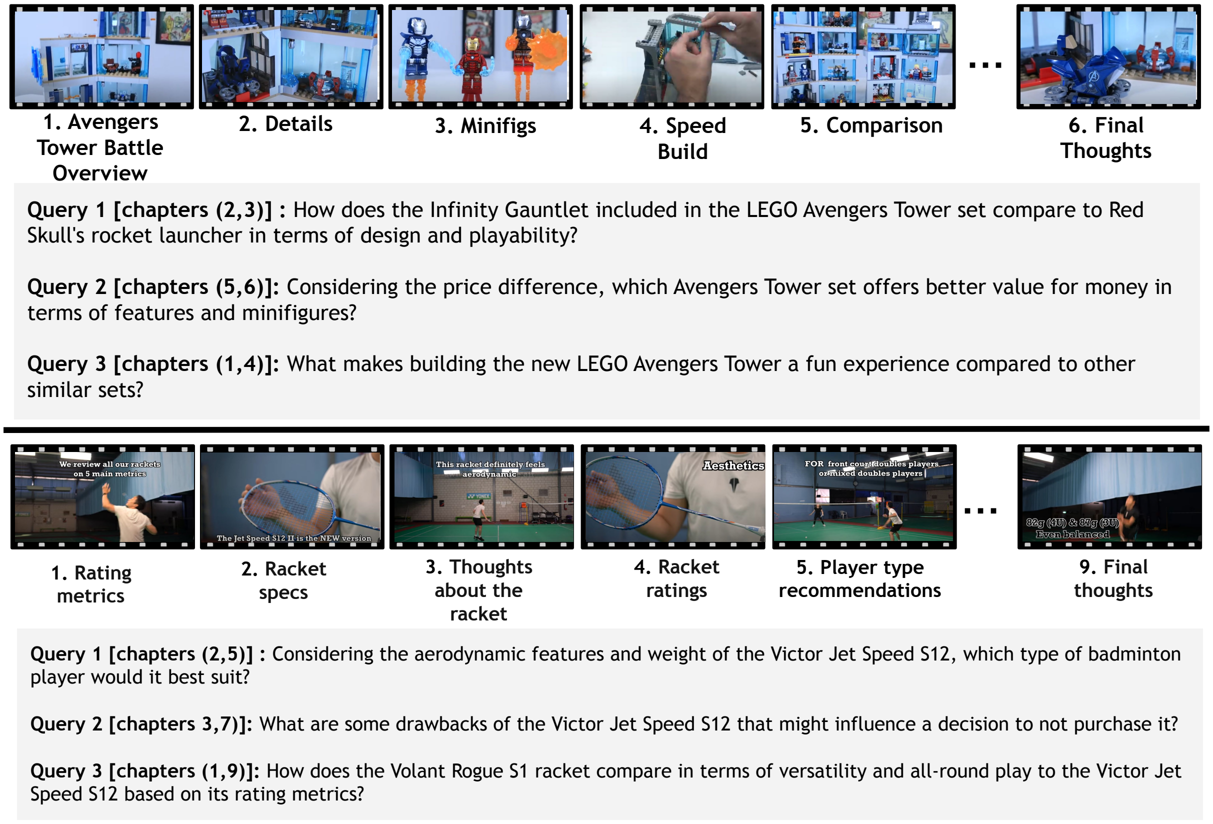
We introduce the ChapVidMR dataset, a collection of video segment-query pairs spanning various domains. This dataset is compiled using videos from the VidChapters-7M dataset [24] and YouTube. ChapVidMR is constructed using videos that leverage chapter information, focusing on classifying chapters that are most relevant to the queries. We use ChapVidMR to benchmark two tasks: Chapter-based Classification retrieval and Boundary Segmentation based retrieval. Table 1 presents a comparison of the size of ChapVidMR with existing moment retrieval datasets, highlighting that both queries and videos in ChapVidMR are longer. Additionally, our dataset typically contains multiple chapters or segments per video and leverages structured chapter information, providing a more realistic scenario where a single temporal segment may not sufficiently address the query.

3.1 ChapVidMR Dataset Analysis

We collected 2, 500 videos from VidChapters-7M [24] and prompted GPT4 to generate five queries per video. As Large Language Models can be prone to errors, they might return chapters that do not exist or return corrupted responses. After further cleaning, we are left with 2, 324 videos. We split them as 1, 862 (80.2%) in the train set and 4, 62 (19.8%) in the test set. We have a total of 10, 168 queries from these videos: 8, 132 in train (79.9%) and 2, 036 (20.1%) in the

Dataset	Multi-moments	Chapter used	#Videos	#Queries	Avg. video len (sec)	Avg. query len (words)
DiDeMo [8]	✗	✗	10.6K	41.2K	29.3	8.0
ActivityNet Captions [10]	✗	✗	15.0K	72.0K	117.6	14.8
CharadesSTA [6]	✗	✗	6.7K	16.1K	30.6	7.2
TVR [14]	✗	✗	21.8K	109K	76.2	13.4
VidChapter-7M [24]	✗	✓	817K	-	1380.0	-
QV Highlights [15]	✓	✗	10.2K	10.3K	150.0	11.3
ChapVidMR (ours)	✓	✓	2.3K	10.8K	322.8	17.7

Table 1: Comparison of ChapVidMR with other video moment retrieval datasets. Our dataset leverages chapter information for query generation and multi-moment retrieval. Our proposed dataset features videos with an average length that surpasses that of existing datasets for moment retrieval. Additionally, our dataset demonstrates enhanced query length.


1. Avengers Tower Battle Overview
2. Details
3. Minifigs
4. Speed Build
5. Comparison
...
6. Final Thoughts
We review all our rackets on 5 main metrics
The Jet Speed S12 II is the NEW version
This racket definitely feels aerodynamic
Aesthetics
FOR front court doubles players or mixed doubles players |
82g (4U) & 87g (3U) Even balanced
1. Rating metrics
2. Racket specs
3. Thoughts about the racket
4. Racket ratings
5. Player type recommendations
...
9. Final thoughts

Query 1 [chapters (2,3)] : How does the Infinity Gauntlet included in the LEGO Avengers Tower set compare to Red Skull's rocket launcher in terms of design and playability?

Query 2 [chapters (5,6)]: Considering the price difference, which Avengers Tower set offers better value for money in terms of features and minifigures?

Query 3 [chapters (1,4)]: What makes building the new LEGO Avengers Tower a fun experience compared to other similar sets?

Query 1 [chapters (2,5)] : Considering the aerodynamic features and weight of the Victor Jet Speed S12, which type of badminton player would it best suit?

Query 2 [chapters 3,7)]: What are some drawbacks of the Victor Jet Speed S12 that might influence a decision to not purchase it?

Query 3 [chapters (1,9)]: How does the Volant Rogue S1 racket compare in terms of versatility and all-round play to the Victor Jet Speed S12 based on its rating metrics?

Figure 2: Visualization of queries and the corresponding ground truth chapters (in parenthesis) used to generate the queries in the ChapVidMR dataset.

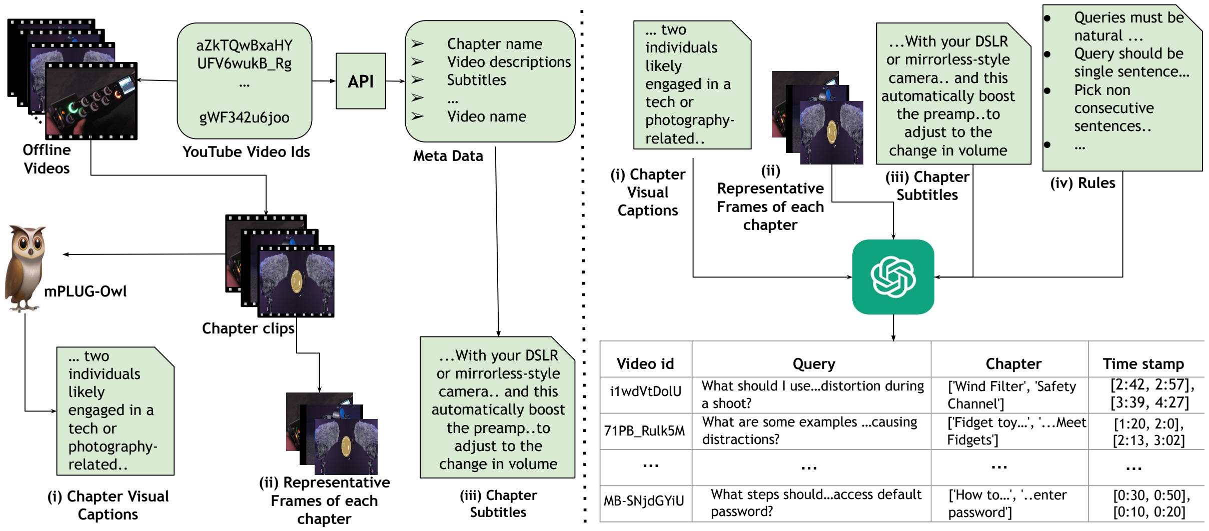
test set. We provide a detailed summary of the dataset statistics in Table 2. Fig. 2 shows a few samples from the dataset. The videos fall under the following categories: education, people and blogs, sports, news and politics, HowTo and style, science and technology, entertainment, film and animation, etc.

We also performed a manual evaluation of the quality of the dataset. Three annotators evaluated 100 samples. Each annotator judged all the 100 queries on two measures: (1) Is the query natural? (2) Do ground truth chapters answer the query sufficiently? For each measure, annotators were asked to report scores on a Likert scale[2] of 0–4. The average scores were found to be 3.12 and 3.27 for the two measures, respectively. These results indicate the superior quality of our proposed dataset.

3.2 Dataset Curation Pipeline

In this section, we outline the pipeline used to curate the dataset, as shown in Fig. 3, including the collection of candidate videos and the processing to extract constituent chapters and associated metadata.

[2] https://en.wikipedia.org/wiki/Likert_scale


aZkTQwBxaHY
UFV6wukB_Rg
...
gWF342u6joo
YouTube Video Ids
Offline Videos
API
Chapter name
Video descriptions
Subtitles
...
Video name
Meta Data
mPLUG-Owl
Chapter clips
... two individuals likely engaged in a tech or photography-related..
(i) Chapter Visual Captions
(ii) Representative Frames of each chapter
...With your DSLR or mirrorless-style camera.. and this automatically boost the preamp..to adjust to the change in volume
(iii) Chapter Subtitles
Queries must be natural ...
Query should be single sentence...
Pick non consecutive sentences..
...
(iv) Rules

Video id	Query	Chapter	Time stamp
i1wdVtDolU	What should I use...distortion during a shoot?	['Wind Filter', 'Safety Channel']	[2:42, 2:57], [3:39, 4:27]
71PB_Rulk5M	What are some examples ...causing distractions?	['Fidget toy...', '...Meet Fidgets']	[1:20, 2:0], [2:13, 3:02]
...	...	...	...
MB-SNjdGYiU	What steps should...access default password?	['How to...', '..enter password']	[0:30, 0:50], [0:10, 0:20]

Figure 3: Data Generation Pipeline. The left part of the figure shows the generation of (i) Chapter Visual captions, (ii) Representative frame of each chapter, and (iii) Chapter Subtitles. We use a video captioner (mPLUG-Owl [25]) to generate chapter-wise visual captions, and we extract chapter subtitles using YouTube APIs. The right part shows the data generation using GPT4. We feed (i) Chapter Visual Captions, (ii) Representative frames of chapters, (iii) Chapter subtitles, and (iv) a set of rules to GPT4. Finally, GPT4 returns queries and ground truth chapters used to form the final dataset.

	Splits	μ	m	σ	Max	Min
Video	Full Dataset	326.2	296.0	146.6	600	34
	Train Set	326.2	296.0	146.5	600	34
	Test Set	326.3	361.0	209.8	563	55
Query	Full Dataset	17.9	18.0	3.5	41	7
	Train Set	17.9	18.0	3.5	41	7
	Test Set	17.6	18.0	2.6	22	11

Table 2: Statistics of the proposed ChapVidMR dataset, measured in terms of video duration (seconds) and query length (words). μ=Mean Length; m=Median Length; σ=Standard Deviation

Video Curation: We collect videos from the YouTube and VidChapters-7M dataset [24] and follow the following set of constraints: (i) duration of the video must not exceed ten minutes; (ii) the video must contain at least three user-generated chapters; (iii) exclude videos belonging to the categories of Gaming and Music as these videos often do not contain information that a user would naturally search for; and (iv) the video must have garnered at least 300 views. We find that imposing these conditions sufficiently filters out videos that can be considered noise (for example, videos generated by bots, videos with poor or misleading meta-data, and unsafe videos). Next, we split the candidate videos into clips based on the chapter time boundaries.

Generating Metadata: Once the videos have been split into their constituent chapters, we extract the available meta-data: subtitles, description, and additionally generate visual captions for each chapter of each video using the mPLUG-Owl [25]. Visual captions provide additional understanding of the video’s content and context.

Generating Queries: We use GPT4 (visual variant) [19] to generate user-like queries as detailed in Fig. 3. To this end, we prompt GPT4 to generate queries that require multiple chapters for an effective and complete answer. To guide GPT4 toward generating natural-sounding queries, we provide the following contextual data: the title, description, and category of the YouTube video. Additionally, the chapter-wise metadata encompassing subtitles, visual captions, and chapter names is provided. We supplement this rich textual context by providing GPT4 with a representative image (middle frame from the chapter), allowing GPT4 to leverage the visual domain. To ensure a good response from GPT4, we coded up a set of rules as part of the prompt, as shown in Table 3.

Generating Finer Boundaries: For the task of segment-based retrieval, we further refined the chapter boundaries. We pass queries along with their associated chapter subtitles, which are numbered according to timestamps, to GPT4. The GPT4 is tasked with selecting the most appropriate sentences to answer the query. Finally, we determine the timestamps of finer boundaries based on the selected sentences within the chapter boundaries.

4 BENCHMARKING CHAPVIDMR USING STATE-OF-THE-ART APPROACHES

4.1 Tasks enabled by ChapVidMR

Chapter Classification-based VMR Task: In this task, the goal is to extract multiple chapters given a text query, as shown in Fig. 4 (left). The retrieved chapters are ranked according to relevance, and the top ones are reported as the chapters required to answer the given query effectively. Classical VMR methods return the start and end timestamps of the video segment most relevant to the given text. However, *chapters* provide more contextual information as compared to segments characterized by timestamps since they are typically created based on a natural content division within the

S.No.	Rule
1	Queries must be natural in a way that a human would ask.
2	Pick two chapters with which you feel you can generate meaningful queries. Generate one if possible; else return 'NONE'.
3	Query should be a single sentence. Avoid using the word "and" or similar conjunctions to combine two independent queries that can be answered by the two video chapters individually.
4	Queries should be natural, in the way a human would write, and the queries can differ in their details, granularity, and which chapters they focus on.
5	These captions were generated using APIs, and there can be noise. Therefore, be cautious in interpreting them.
6	Do not assume something you are not sure about.
7	Generate queries using visual context provided via visual captions & video frames given in chapter annotations. Subtitles and chapter names are only for your context, and the user asking or answering the queries doesn't have access to them.
8	Avoid using introductory and outro chapters for framing questions.

Table 3: Rules Provided to GPT for Query Generation.

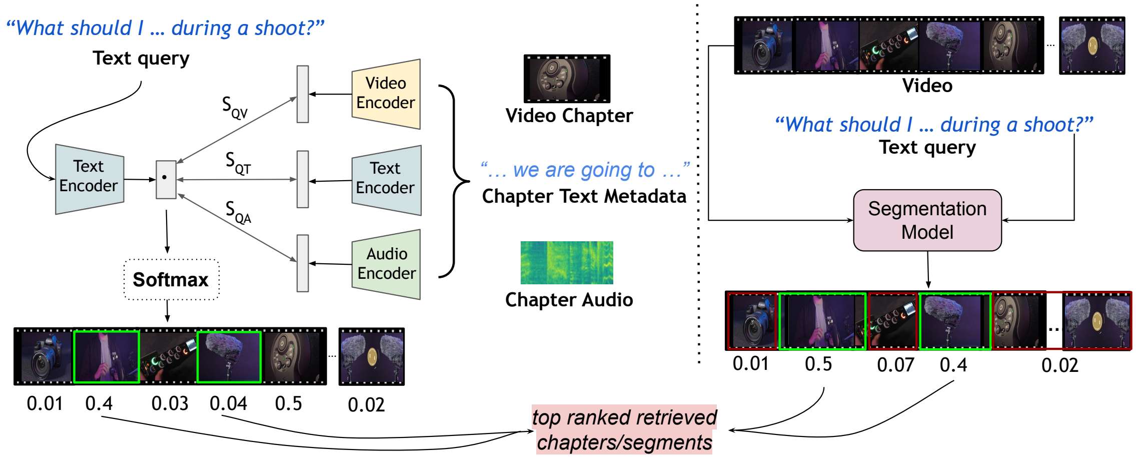


Figure 4: Illustration of Methods. The left part of the figure shows the classification-based moment retrieval process. Each video chapter is processed by three encoders, one per modality. The similarity between the chapter and the text query is calculated using a dot product. Chapters are then ranked and retrieved based on their similarity scores. The right part of the image illustrates the segmentation-based retrieval framework. Here, both the video and text query are input into a video moment segmentation model. The top-ranked segmented regions are then retrieved as output segments.

video, thereby being more intuitive and information-rich. Titles of the chapter provide immediate context, allowing viewers to understand the content of the video segment at a glance and directly jump to the relevant chapters without having to search for information in long videos.

Boundary Segmentation-based VMR Task: Similar to the Moment Retrieval task [15, 16, 18] as shown in Fig. 4 (right), we benchmark our dataset on the Finer Boundary segmentation task. In this task, given a video and text query, the model segments the videos into different parts and returns the most appropriate segments that answer the query. For this task, we further refined our chapter boundaries to return tight boundaries within the chapters as detailed in Section 3.2.

4.2 Benchmarking

Chapter classification-based VMR Methods: We experiment with LanguageBind [29], ImageBind [7], UniVTG [16] and Sentence-BERT [20] to classify and retrieve the multiple top segments required for answering the given query. Our experimental setup includes first segmenting the videos into given chapters and subsequently extracting embeddings of their respective vision, language, and audio modalities. ImageBind [7] is an approach that shows that only *image-paired data* instead of all combinations of paired data, is required to learn a joint-embedding space effectively binding other modalities and enabling novel emergent applications including cross-modal retrieval, cross-modal detection, and generation, etc. LanguageBind [29], a multi-modal pretraining framework that extends video-language pretraining to encompass multiple modalities, including audio. Taking language as the bind, all modalities are mapped to a unified embedding space, thereby enabling effective semantic alignment. ImageBind relies on images as intermediaries, while LanguageBind dispenses with this requirement altogether and instead directly aligns all modalities to the language space, thus enhancing its applicability to additional modalities in downstream tasks. Sentence-BERT [20] is a modified version of the well-known BERT [4] model, which uses siamese and triplet network structures to derive textual queries that can be compared using cosine similarity in an efficient and accurate manner. It adds a pooling operation

Method	Modalities	Avg IoU	Avg Precision	Avg Recall
Random Baseline	-	25.5	35.4	31.1
ImageBind [7]	Audio	25.3	35.6	33.4
ImageBind [7]	Video	35.9	47.0	44.7
ImageBind [7] + mPLUG-Owl [25]	Subtitles + Visual Captions	27.2	36.2	37.5
LanguageBind [29]	Audio	26.7	36.1	35.5
LanguageBind [29]	Video	35.9	47.0	44.7
LanguageBind [29] + mPLUG-Owl [25]	Subtitles + Visual Captions	36.0	46.7	45.7
Sentence-BERT [20] + mPLUG-Owl [25]	Subtitles + Visual Captions	**43.5**	**53.7**	**53.3**

Table 4: Chapter Classification VMR Results. When chapters are enhanced with subtitles and visual captions, Sentence-BERT yields the best results compared to other methods.

to the output of BERT specifically, three pooling experiments are tried out, which include 1) using the output of the CLS-token, 2) taking the mean of all output vectors, and 3) computing a max-over-time for the output vectors, with the default setting being mean. In the case of Sentence-BERT, only the embeddings of the subtitles and visual captions are extracted. A dot product between each segment's embedding and the query's representation is computed. Then, the top two chunks are retrieved according to each modality - audio, text, and vision.

Chunked Text Methods for Chapter Classification VMR: The tokenizers of ImageBind [7] and LanguageBind [29] process only 77 tokens (approx. 58 words) which leads to context truncation. Hence, we first divide the chapter text into chunks of 30 words each. We then perform experiments using subtitles only, visual captions only, and subtitles and visual captions concatenated together. We compute the score per chapter as an aggregate over consistent chunks in the following three ways. (1) Mean Score: We first extract the embeddings of each 30 word string and compute the dot product of each of these with the query. The chapter score is then computed as an average of these scores. (2) Mean Pool: We mean pool the embeddings of all the chunks of a single chapter and then compute the dot product with the query. (3) Max Score: We compute the dot product scores of each chunk embedding with the query embedding and select the maximum score as chapter score.

Segmentation-based VMR Methods: We use the following methods for finer boundary segmentation retrieval tasks: Moment-DETR [15], QD-DETR [18] and UniVTG [16]. Moment-DETR [15] is a transformer encoder-decoder model for VMR. It treats retrieval as a set prediction task, processing video and query representations to predict relevant video moment coordinates and saliency scores. This model operates without human priors and is enhanced by weakly supervised pretraining with ASR captions, enabling it to outperform traditional methods in identifying relevant video highlights. QD-DETR [18] extends Moment-DETR and integrates the text query context into video representation using cross-attention layers and enhances query utilization by training with negative video-query pairs for accurate matching. Additionally, it features an input-adaptive saliency predictor that tailors saliency scoring to specific video-query pairs. This approach offers a more query-responsive solution compared to Moment-DETR, which lacks mechanisms for adaptive saliency and explicit query integration. UniVTG [16] is a unified framework designed to handle various Video Temporal Grounding (VTG) tasks, including moment retrieval. It revisits and standardizes a broad spectrum of VTG tasks and labels into a cohesive formulation. This standardization facilitates the creation of scalable pseudo-supervision through a refined data annotation approach. UniVTG employs a versatile grounding model that not only performs well across different VTG [16] tasks but also enhances task-specific functionalities like moment retrieval.

5 EXPERIMENTS AND RESULTS

5.1 Implementation Details and Metrics

For the task of Chapter-based retrieval, we experiment with ImageBind [7], LanguageBind [29], and Sentence-BERT [20]. For ImageBind, we utilize the officially released *huge* variant of the model. In the case of LanguageBind, the representations of the respective modalities are extracted using the released model's fully fine-tuned versions of the video, audio, and text encoders. In the case of the text modality, OpenCLIP's text encoder, which has 768-dimensional 12 layers, is used. While experimenting with Sentence-BERT, we use the pretrained model from HuggingFace. Specifically, we use the DistilBert TAS-B variant, which projects sentences/paragraphs to a 768-dimensional vector space. For the finer boundary segmentation, the official implementation of the models (QD-DETR [18], Moment-DETR [15]) with pretrained weights were used. We conducted segmentation using a single Nvidia A6000 GPU on a server equipped with 256 GB of RAM.

For classification-based retrieval, we utilized the same server configuration, however, without the use of GPUs. For chapter classification VMR, we report the Average Intersection over Union (IoU), Average Precision, and Average Recall. Here, IoU is defined as the ratio of true positives (TP) to the combined count of true positives (TP), false negatives (FN), and false positives (FP). We use standard definitions for Recall and Precision. For Segmentation-based VMR, we use Recall@1 and mAP at various IoU thresholds following [15].

5.2 Results and Observations

Performance of Chapter Classification VMR Methods: Table 4 summarizes the results of methods applied to the chapter classification VMR task, utilizing different combinations of modalities. The Random Baseline method provided the lowest performance across all metrics. The methods employing the ImageBind [7] and LanguageBind [29] models showed similar performance when using video alone, both achieving an average precision and recall of

	Input Representation	Avg IoU		Avg Precision		Avg Recall	
		I-Bind	L-Bind	I-Bind	L-Bind	I-Bind	L-Bind
Mean Score	Subtitles	34.9	33.1	46.4	43.7	43.2	42.1
	Visual Captions	29.1	29.7	39.2	39.6	37.7	38.2
	Subtitles + Visual Captions	**37.8**	34.9	**46.4**	45.0	**49.1**	45.4
Mean Pool	Subtitles	34.9	33.1	46.4	43.7	43.2	42.1
	Visual Captions	29.1	29.7	39.2	39.6	37.7	38.2
	Subtitles + Visual Captions	**37.8**	34.9	**46.4**	45.0	**49.1**	45.4
Max Score	Subtitles	39.8	37.3	49.1	46.0	50.3	48.7
	Visual Captions	29.4	29.4	39.7	39.3	38.0	38.0
	Subtitles + Visual Captions	**41.4**	41.4	**50.6**	46.6	**52.5**	49.0

Table 5: Chapter Classification VMR results with chunked text for LanguageBind (L-Bind) and ImageBind (I-Bind) across three evaluation methods: Mean Score, Mean Pool, and Max Score.

Method	Chapter Classification VMR			Boundary Segmentation VMR			
	Avg Precision	Avg Recall	Avg IoU	R1@0.5	R1@0.7	mAP@0.5	mAP@0.75
Moment-DETR [15]	37.8	37.4	29.0	19.0	15.0	6.0	3.3
QD-DETR [18]	37.0	36.3	27.6	24.0	21.0	8.0	3.7
UniVTG [16]	**40.4**	**38.9**	**30.1**	**26.8**	**23.3**	**11.7**	**5.3**

Table 6: Comparative performance of Moment-DETR, QD-DETR, and UniVTG on Chapter Classification-based VMR and Segmentation-based VMR tasks. Metrics include Average Precision, Recall, and Intersection over Union (IoU) for chapter classification and R1@0.5, R1@0.7, mAP@0.5, and mAP@0.75 for boundary segmentation.

around 47% and 44.7%, respectively. When these methods were combined with mPLUG-Owl [25], which provides visual captions, performance metrics showed a noticeable improvement, particularly for the combination of LanguageBind [29] and mPLUG-Owl [25]. The inclusion of visual captions appears to provide a significant contextual anchor, enhancing the model’s ability to use relevant chapters within the video more accurately. This fusion between text and image understanding modules provides an improvement in retrieval performance, illustrating the potential of multimodal approaches in complex VMR tasks. However, the most effective method was the integration of Sentence-BERT with mPLUG-Owl [25], which significantly enhanced the performance, achieving the highest average IoU, precision, and recall rates of 43.5%, 53.7%, and 53.3%, respectively.

Performance of Chunked Text Methods for Chapter Classification VMR: Table 5 compares the performance of LanguageBind (L-Bind) [29] and ImageBind (I-Bind) [7] in chapter classification VMR tasks, using subtitles, visual captions, or both. Across three experiments, Mean Score, Mean Pool, and Max Score, the combination of subtitles and visual captions yields the highest performance for both frameworks, with ImageBind [7] slightly outperforming LanguageBind [29]. The addition of visual captions, in particular, seems to provide a robust layer of contextual data that complements the text-based input from subtitles, leading to improved identification and retrieval of video chapters. The Max Score experiment shows the best results, particularly when using both subtitles and visual captions.

Performance of Segmentation VMR Methods: The comparative analysis of segmentation methods for classification-based and segmentation-based retrieval, as shown in Table 6, indicates that UniVTG consistently achieves the highest performance in both tasks. It leads in precision, recall, and Intersection over Union (IoU) scores, particularly excelling in the Boundary Segmentation task with higher mAP and recall metrics at stricter IoU thresholds. This superior performance can be attributed to the model’s ability to finely parse and interpret complex video segments, tailoring its responses to the nuanced demands of the query. Moment-DETR and QD-DETR perform similarly, although QD-DETR slightly surpasses Moment-DETR for Boundary Segmentation but trails in average precision and recall for Chapter Classification.

6 ETHICS STATEMENT

Our dataset curation relies on GPT-4, which may introduce biases from its pre-training data, affecting question generation and alignment with human expectations. These biases could propagate stereotypes or misrepresentations. While we critically assessed outputs to ensure alignment with research goals, further study is required to fully understand and mitigate these risks.

7 CONCLUSION

We introduced the ChapVidMR dataset, which leveraged YouTube video chapters for generating multi-moment user queries. The dataset was composed of queries based on multiple chapter names and metadata extracted from the YouTube API. We explored two main VMR tasks: chapter classification-based and segmentation-based. For the chapter classification task, queries were mapped to relevant chapters using multimodal features. In contrast, the segmentation-based VMR task involved splitting videos into segments and then locating the most relevant segments. Overall, the dataset and methodologies established baseline solutions for video moment retrieval using chapters as structured video segments.

Acknowledgment: This work is supported by Microsoft Academic Partnership Grant (MAPG) 2023. Yogesh Kumar is supported by a UGC fellowship, Govt. of India.

REFERENCES

[1] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

[2] B. Cheng, A. G. Schwing, and A. Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021.

[3] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022.

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

[5] V. Escorcia, M. Soldan, J. Sivic, B. Ghanem, and B. C. Russell. Temporal localization of moments in video collections with natural language. *ArXiv*, 2019.

[6] J. Gao, C. Sun, Z. Yang, and R. Nevatia. TALL: temporal activity localization via language query. In *ICCV*, 2017.

[7] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra. Imagebind one embedding space to bind them all. In *CVPR*, 2023.

[8] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. C. Russell. Localizing moments in video with natural language. In *ICCV*, 2017.

[9] K. Huang, T. Wu, H. Su, and W. H. Hsu. Monodtr: Monocular 3d object detection with depth-aware transformer. In *CVPR*, 2022.

[10] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. Dense-captioning events in videos. In *ICCV*, 2017.

[11] Y. Kumar and A. Mishra. Few-shot referring relationships in videos. In *CVPR*, 2023.

[12] Y. Kumar, S. Mallick, A. Mishra, S. Rasipuram, A. Maitra, and R. R. Ramnani. Qdetrv: Query-guided DETR for one-shot object localization in videos. In *AAAI*, 2024.

[13] J. Lei, L. Yu, T. L. Berg, and M. Bansal. TVQA+: spatio-temporal grounding for video question answering. In *ACL*, 2020.

[14] J. Lei, L. Yu, T. L. Berg, and M. Bansal. TVR: A large-scale dataset for video-subtitle moment retrieval. In *ECCV*, 2020.

[15] J. Lei, T. L. Berg, and M. Bansal. Detecting moments and highlights in videos via natural language queries. In *NeurIPS*, 2021.

[16] K. Q. Lin, P. Zhang, J. Chen, S. Pramanick, D. Gao, A. J. Wang, R. Yan, and M. Z. Shou. Univtg: Towards unified video-language temporal grounding. In *ICCV*, 2023.

[17] K. Q. Lin, P. Zhang, J. Chen, S. Pramanick, D. Gao, A. J. Wang, R. Yan, and M. Z. Shou. Univtg: Towards unified video-language temporal grounding, 2023.

[18] W. Moon, S. Hyun, S. Park, D. Park, and J. Heo. Query - dependent video representation for moment retrieval and highlight detection. In *CVPR*, 2023.

[19] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

[20] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*, 2019.

[21] D. Shao, Y. Xiong, Y. Zhao, Q. Huang, Y. Qiao, and D. Lin. Find and focus: Retrieve and localize video events with natural language queries. In *ECCV*, 2018.

[22] H. Xu, K. He, B. A. Plummer, L. Sigal, S. Sclaroff, and K. Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *AAAI*, 2019.

[23] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *CVPR*, 2022.

[24] A. Yang, A. Nagrani, I. Laptev, J. Sivic, and C. Schmid. Vidchapters-7m: Video chapters at scale. In *NeurIPS*, 2024.

[25] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, C. Li, Y. Xu, H. Chen, J. Tian, Q. Qi, J. Zhang, and F. Huang. mplug-owl: Modularization empowers large language models with multimodality. *CoRR*, abs/2304.14178, 2023.

[26] D. Zhang, X. Dai, X. Wang, Y. Wang, and L. S. Davis. MAN: moment alignment network for natural language moment retrieval via iterative graph adjustment. In *CVPR*, 2019.

[27] S. Zhang, H. Peng, J. Fu, and J. Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, 2020.

[28] M. Zheng, Y. Huang, Q. Chen, Y. Peng, and Y. Liu. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In *CVPR*, 2022.

[29] B. Zhu, B. Lin, M. Ning, Y. Yan, J. Cui, H. Wang, Y. Pang, W. Jiang, J. Zhang, Z. Li, W. Zhang, Z. Li, W. Liu, and L. Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *CoRR*, abs/2310.01852, 2023.