This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Towards Efficient and Robust Moment Retrieval System: A Unified Framework for Multi-Granularity Models and Temporal Reranking

Huu-Loc Tran [1] Tinh-Anh Nguyen-Nhu [2*] Huu-Phong Phan-Nguyen [1*]
Tien-Huy Nguyen[1] Nhat-Minh Nguyen-Dich[3] Anh Dao [4]
Huy-Duc Do [5] Quan Nguyen [6] Hoang M. Le [7] Quang-Vinh Dinh [8]

[1] University of Information Technology, VNU-HCM, Vietnam
[2] Ho Chi Minh University of Technology, VNU-HCM, Vietnam
[3] Hanoi University of Science and Technology, Hanoi, Vietnam
[4] Michigan State University, USA
[5] National Economics University, Hanoi, Vietnam
[6] Posts and Telecommunications Institute of Technology, Hanoi, Vietnam
[7] York University, Canada
[8] AI VIETNAM Lab
22520567@gm.uit.edu.vn

Abstract

Long-form video understanding presents significant challenges for interactive retrieval systems, as conventional methods struggle to process extensive video content efficiently. Existing approaches often rely on single models, inefficient storage, unstable temporal search, and context-agnostic reranking, limiting their effectiveness. This paper presents a novel framework to enhance interactive video retrieval through four key innovations: (1) an ensemble search strategy that integrates coarse-grained (CLIP) and fine-grained (BEIT3) models to improve retrieval accuracy, (2) a storage optimization technique that reduces redundancy by selecting representative keyframes via TransNetV2 and deduplication, (3) a temporal search mechanism that localizes video segments using dual queries for start and end points, and (4) a temporal reranking approach that leverages neighboring frame context to stabilize rankings. Evaluated on known-item search and question-answering tasks, our framework demonstrates substantial improvements in retrieval precision, efficiency, and user interpretability, offering a robust solution for real-world interactive video retrieval applications.

1. Introduction

Recent advances in deep learning have significantly improved core vision tasks such as recognition, domain adaptation, and visual question answering [33–35, 37, 38], paving the way for more capable video understanding systems. However, most existing approaches [10, 19, 49, 52] are designed for short video clips—typically only seconds to a few minutes long—making them ill-suited for real-world scenarios involving long-form videos [54]. This gap presents a major challenge for applications like content-based video retrieval and surveillance, where processing hours-long content is essential.

Interactive video retrieval [20, 26, 30] has emerged as a promising solution by combining automated analysis with human input. Recent advancements, fueled by deep learning and validated in competitions like TRECVID [31] and VBS [28], demonstrate that human-computer collaboration not only refines automated results in real time but also substantially enhances search effectiveness.

The core challenge lies in translating a user's information need into an efficient search process that identifies relevant segments within massive video archives. Current state-of-the-art systems employ a variety of techniques. Text-based retrieval, for instance, matches user queries with metadata, transcripts, or deep-learning-generated embeddings (e.g., CLIP [41], W2VV++ [25]), thereby facilitating semantic search, though often at the expense of interpretability. Concept-based methods [5, 8, 9, 15, 18, 42] use

*All authors contributed equally to this paper.
This research is fully supported by AI VIETNAM [1].

object detection and scene classification models to annotate frames with high-level semantics, yet they are limited by predefined categories and annotation inconsistencies. Similarly, content-based visual search retrieves [2, 13, 44] visually similar frames using CNN-based descriptors [48] but may return semantically irrelevant results without a suitable reference. Even sketch-based and spatial queries, which offer a visual means to express search intent, are challenged by high interpretation variability.

Despite recent advances, current interactive video retrieval systems still exhibit key limitations. First, relying on a single model restricts their ability to capture both broad semantics and fine-grained details [40]. Second, indexing every frame creates redundancy, leading to storage and search inefficiencies. Third, unstable temporal search methods make it difficult to accurately locate event sequences. Finally, conventional reranking overlooks temporal context, resulting in inconsistent rankings. To overcome these challenges, we propose a comprehensive framework featuring four innovations:

1. **Ensemble Search:** By combining coarse-grained models that capture broad semantics with fine-grained models that focus on intricate details, our ensemble approach yields more robust retrieval outcomes.
2. **Storage Optimization:** We reduce redundancy by selecting representative keyframes through intelligent deduplication, significantly cutting storage needs without sacrificing search quality.
3. **Temporal Search:** Our dual-query mechanism-capturing both start and end points-accurately localizes video segments in chronological order, ensuring stable and interpretable results.
4. **Temporal Reranking:** Leveraging contextual information from neighboring frames, our reranking strategy refines candidate orderings to maintain structural coherence and relevance.

These contributions address critical shortcomings in existing systems, leading to more precise, efficient, and user-friendly interactive video retrieval. The following sections detail each component, present experimental results, and discuss future implications.

2. Related Work

Recent research in video retrieval has evolved along two complementary lines: localized moment retrieval within single videos and corpus-level retrieval that jointly identifies the relevant video and the corresponding moment. In this section, we review key methods spanning both Single Video Moment Retrieval (SVMR) and Video Corpus Moment Retrieval (VCMR), as well as interactive retrieval systems that incorporate human feedback.

2.1. Video Moment Retrieval

Video Moment Retrieval aims at localizing a target segment in an untrimmed video based on a natural language query. Early work in this area typically followed a proposal-based paradigm, generating candidate segments which are then ranked based on their relevance to the query [3, 4, 6, 11, 27, 47]. In contrast, proposal-free approaches directly regress the temporal boundaries by leveraging iterative attention between video frames and query words, often benefiting from transformer architectures [7, 12, 22, 29, 32, 39, 43, 46, 55, 57, 58, 62].

The scope of video retrieval has further expanded to Video Corpus Moment Retrieval (VCMR), where the task is to first identify candidate videos from a large collection and subsequently localize the relevant moment within the selected video. Methods for VCMR are generally divided into one-stage and two-stage approaches [16, 21, 23, 56, 59–61]. One-stage methods jointly perform retrieval and localization in an end-to-end manner; for instance, HERO [23] employs a hierarchical transformer-based encoder that integrates visual and textual cues across modalities. Two-stage methods, on the other hand, first retrieve a set of candidate videos based on global text–video similarity and then apply fine-grained localization on each candidate. CONQUER [16] exemplifies this approach by introducing a query-aware ranking mechanism that benefits from enhanced interactions between the query and video content, thereby offering improved scalability for large-scale video retrieval.

Moreover, the emergence of large language models has influenced both SVMR and VCMR. Recent studies integrate video understanding and moment retrieval into a next-token prediction framework [14, 17, 50, 53]. Additionally, generative techniques have been explored; for instance, MomentDiff [24] formulates moment retrieval as a diffusion process that iteratively refines random temporal proposals into the correct segment.

2.2. Interactive Retrieval Systems

While fully automated retrieval pipelines have made significant progress, they often struggle with complex queries and long video inputs. To address these limitations, interactive retrieval systems incorporate human-in-the-loop strategies. Benchmarks such as the Video Browser Showdown (VBS) have spurred research on multimodal interactive methods that combine text, sketches, filters, and example frames. For example, [30] proposed a reinforcement learning-based framework that learns from user feedback to navigate large video corpora, while [26, 36] introduced a question-answering-based interactive system that simulates user interactions using a VideoQA model. These interactive approaches demonstrate that iterative refinement can substantially improve retrieval performance, making them particularly promising for real-world applications.

3. Method

3.1. Problem Definition

Holistic video understanding remains challenging, as most existing methods focus on short video segments despite the growing prevalence of long-form content spanning minutes to hours. Given a corpus of untrimmed videos $\mathcal{V}$ and a textual query q, the objective of Video Corpus Moment Retrieval (VCMR) is to identify the specific moment $m^* = (t^s, t^e)$ that best aligns with q, where t^s and t^e denote the start and end timestamps, respectively:

$$m^* = \arg\max_{m} P(m \mid q, \mathcal{V}). \tag{1}$$

The VCMR process comprises two stages: (i) retrieving candidate moments m from videos in the corpus, and (ii) accurately localizing the optimal moment m^* within the selected video v^*. To further enhance the system's capabilities, we incorporate a Question Answering (QA) component, allowing users to interact with localized moments. This facilitates refined responses by leveraging additional context from the targeted segment.

Traditional video-level retrieval methods often struggle to capture the fine-grained temporal details necessary for accurate localization. To overcome this, we shift to an image-level retrieval approach, treating individual frames as the basic retrieval units. Although this method lacks immediate temporal context, we mitigate this shortcoming using a combination of reranking (**Section** 3.3), ensemble search (**Section** 3.4), and temporal modeling (**Section** 3.5). These components collectively ensure that the retrieved frames maintain both semantic relevance and temporal coherence.

3.2. Data storage

For efficient image-level video retrieval, an optimized data storage strategy is critical for efficient retrieval performance. By selecting a minimal but representative set of frames, we reduce redundancy, making retrieval more efficient without compromising accuracy. In this section, our data storage comprising three stages: keyframe selection **3.2.1**, feature extraction **3.2.2**, and storage optimization **3.2.3**.

3.2.1. Keyframe Selection

To extract keyframes from videos, we utilize TransNetV2[45], a fast and accurate deep learning model designed for scene transition detection. TransNetV2[45] is a CNN and RNN hybrid that excels in detecting hard cuts and gradual transitions within video sequences. It processes input frames sequentially and identifies transition probabilities, enabling precise segmentation of scenes.

Once scene transitions are detected, we select keyframes for each scene. Instead of storing all frames, we sample four evenly spaced frames within each detected scene based on their frame indices. This strategy ensures a diverse yet compact representation of the scene, reducing storage requirements while preserving essential visual information.

3.2.2. Feature Extraction

To enhance retrieval accuracy, we employ powerful feature extractors, specifically BEIT3[51] and CLIP[41]. These models have demonstrated superior performance across multiple benchmark datasets, making them well-suited for video retrieval tasks. These models are particularly well-suited for our task due to their ability to generate robust, high-dimensional feature embeddings, capturing intricate visual relationships that traditional CNN-based models might overlook.

3.2.3. Storage Optimization

Despite selecting only four keyframes per scene, redundant frames may still exist, leading to suboptimal memory usage and slower retrieval times. We implement a duplicate removal algorithm based on feature similarity to address this.

We compute the cosine similarity between keyframes within each scene using the extracted feature embeddings from BEIT3 and CLIP. If a frame has a cosine similarity score greater than 0.9 with any other frame in the same scene, it is considered a near-duplicate and is removed. This process significantly reduces storage redundancy while maintaining retrieval efficiency. The **Algorithm** 1 describes all of the above.

By integrating efficient keyframe selection, feature extraction, and storage optimization, our approach ensures a highly scalable and accurate image-level video retrieval system. The effectiveness of our storage optimization strategy is visually demonstrated in the **Figure** 3, highlighting the impact of duplicate removal on the dataset efficiently.

Algorithm 1 Frame Filtering

```
for each group, values in grouped_videos do
    Load scene boundaries from file
    Initialize id = 0 and frame_in_scene = []
    Set S ← 0.9
    for each key, path in values do
        if current frame is within the scene boundary then
            Add frame to frame_in_scene
        else
            Compute embeddings: f frame embedding
            Compute similarity: E = sim(f1, f2)
            Identify redundant frames (if E > S)
            Remove redundant frames
            Move to the next scene (id += 1)
            Reset frame_in_scene
        end if
    end for
end for
```

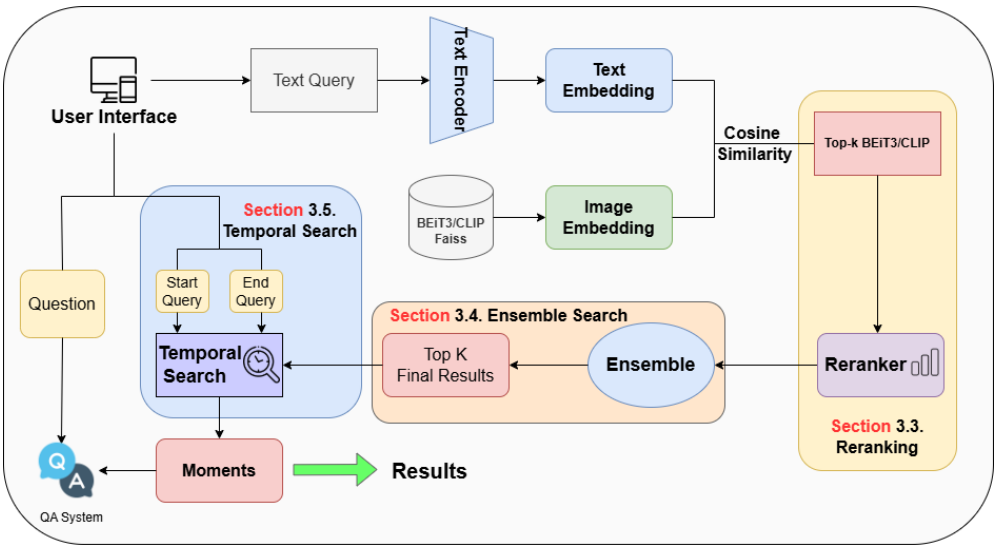


Figure 1. Overview of our iteractive retrieval system. The system ranks the top-k results through a reranking module (**Section** 3.3 before passing them to an ensemble module (**Section** 3.4) for final selection. The temporal search module (**Section** 3.5) refines the results by identifying the most relevant time segments, ensuring the retrieval aligns with the query's temporal context. The final output consists of the most relevant moments, providing accurate answers based on the keyframe range.

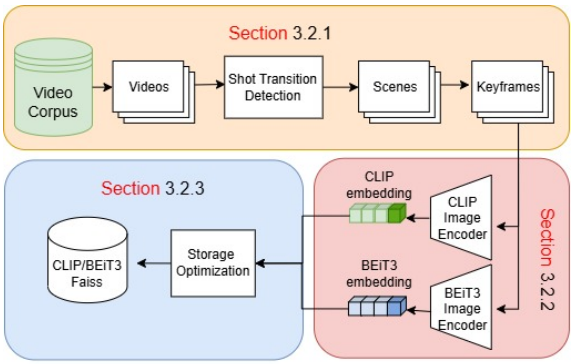


Figure 2. Videos are segmented, deduplicated with cosine embeddings, and stored in the FAISS index.

3.3. Reranking

As described in **Algorithm** 2, the neighbor score aggregation method should be seen as a strong keyframe selection approach due to the enhancement of both stability and temporal relevance. Since a shot detection model has previously been employed to retrieve the keyframes, we believe that the areas surrounding a keyframe are most likely to possess visual and semantic features in common or correspond to temporal shifts within a shot. Hence, this approach guarantees reliable keyframe selection when local neighbors show stable visual similarity and ensures temporal coherence when the queried content depicts motion that covers some of the adjacent frames.

The trade-off of stability and temporal relevance is depicted in the **Algorithm** 2. The aggregated score reinforces

Figure 3. Before and after frame filtering.

the candidate keyframe on stable neighbors - both semantically and scores-wise responsible for a strong representation of the scene. Aggregated against the neighbors based on a temporally descriptive query, it explicitly expresses the context of any dynamic for which a query illustrates. Thus, extreme situations like sudden changes in the scene would be handled by the conditional score check that would filter out the irrelevant or missing part of the contribution and thus the method could achieve some decent robustness.

Algorithm 2 Neighbor Score Aggregation

Require: Indcies I, Query Q,

function AGGREGATENEIGHBORSCORES(I, Q)
  Initialize dictionary aggregated_score A
  for each idx in I **do**
    $key \leftarrow$ Convert idx to integer
    neighbors $\leftarrow$ GETNEIGHBORS(key)
    $total_score \leftarrow 0$
    for each $neighbor$ N in neighbors **do**
      score $\leftarrow$ COMPUTESCORE(N, Q)
      if $score \neq$ None **then**
        Append $neighbor$ to indices
        $total_score \leftarrow total_score + score$
      end if
    end for
    UPDATESCORES(A, key, $total_score$)
  end for
  sorted_scores $\leftarrow$ SORT(A, descending)
  return sorted_scores
end function

3.4. Ensemble Search

Image-text retrieval requires both fine-grained visual understanding and coarse-grained conceptual alignment. BEiT-3 [51] excels in detailed image-text alignment but may overlook broader context, while CLIP [41] performs well in zero-shot retrieval but struggles with fine-grained distinctions. These differences stem from their training: BEiT-3 focuses on comprehensive semantic understanding, whereas CLIP learns condensed representations from limited text. To balance precision and recall, we propose an ensemble search method that combines their strengths, improving retrieval robustness across diverse queries.

Algorithm 3 Ensemble Search

Require: query (string), model_configs (list of (model_name, weight, use_flag))
Ensure: ranked_results (list of (index, score))

Initialize score_dict as empty dictionary
Normalize weights: $\sum w_i = 1$ for selected models
for all (model_name, w, use_flag) $\in$ model_configs **do**
  if use_flag **then**
    Load model and processor for model_name
    $e \leftarrow$ EncodeText(model_name, query)
    $I, S \leftarrow$ Search(model_name_idx, e, $M = 50$)
    $S_{\max} \leftarrow \max(S)$
    for all $(i, s) \in (I, S)$ **do**
      score_dict$[i]$ += $\frac{s}{S_{\max}} \times w$
    end for
  end if
end for
ranked_results $\leftarrow$ Sort(score_dict, descending)
return ranked_results

Our ensemble search methodology is formally outlined in **Algorithm 3**. The process begins by encoding the query text with both models using their tokenization methods and normalizing the results to unit length to produce text embeddings. These model-specific query embeddings are then used to retrieve the top M results from each corresponding index. To effectively combine these results, we normalize the similarity scores from each model by their respective maximum values, thereby preventing disparities in scale. By applying weighting factors during ensemble search, the approach reflects the relative contribution of each model. The weighted scores are aggregated by each image identifier, with scores being summed when an image appears in both result sets. The final step incorporates sorting these aggregated scores in descending order to produce a comprehensive list that both precisely detailed and broadly relevant to the query, overcoming the limitations of each model in isolation.

3.5. Temporal Search

Temporal video search is challenging due to the dynamic nature of visual content and frame dependencies. Unlike static image retrieval, it must identify not just relevant frames but also their temporal range for accurate event localization. A single frame often gives only a rough estimate without clear temporal boundaries. To address this issue, we propose a bidirectional temporal search strategy that refines retrieval precision by leveraging additional queries, depicted in **Algorithm 4**. We assume that the initially retrieved and reranked input frame corresponds to the correct reference frame, and our objective is to localize the segment that best aligns with the given query. To achieve this, we conduct a bidirectional search, extending to the left of the input frame index until either 20 relevant frames are identified or the similarity score falls below an acceptable threshold. The same approach is applied symmetrically to the right. Following this, we determine the optimal frame pair by selecting two frames that exhibit the highest similarity scores with the respective queries while ensuring that their temporal distance, including the input frame, does not exceed a predefined constraint, denoted as gap_C. This approach effectively refines temporal retrieval by incorporating local context while maintaining alignment with the query semantics.

3.6. Interactive Video Retrieval System

In this section, we introduce our *Interactive Video Retrieval System*, designed to handle two core tasks: *Moment Retrieval* and *Video Question Answering (QA)*. The system features a user-friendly interface where queries can be entered, search strategies can be selected, and results can be iteratively refined. **Figures 4a, 4b, and 4c** illustrate various stages of user interaction within the system.

Algorithm 4 Temporal Frame Pair Selection

function FINDBESTFRAMEPAIR(query_1, query_2, input_frame, index, img_path, gap_C)
    Identify the video and frame ID from *input_frame*
    Initialize a list for relevant **left** and **right** frames
    Set a similarity **threshold** to filter frames
    while Frame is relevant and not exceeding limit **do**
        Compute similarity with *query_1*
        Stop if similarity is too low
        Add frame to **left list** and move left
    end while
    while Frame is relevant and not exceeding limit **do**
        Compute similarity with *query_2*
        Stop if similarity is too low
        Add frame to **right list** and move right
    end while
    Collect all candidate frames from left, current, and right
    Find the best pair of frames that maximizes combined similarity
    Ensure the frames satisfy the temporal constraint (gap_C)
    return the best matching frame pair
end function

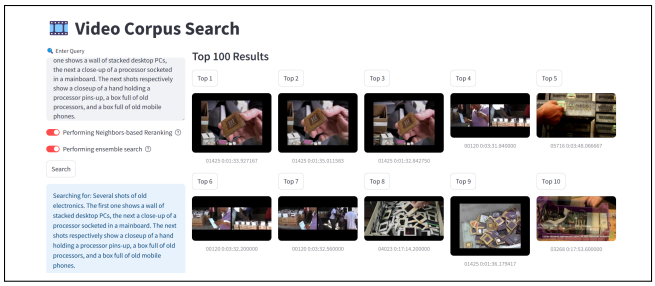
3.6.1. Moment Retrieval

As shown in **Figure 4a**, the user can enter a free-form text query in the designated text box. Two retrieval strategies, *neighbors-based reranking* and *ensemble search*, are provided to adjust how the system prioritizes candidate frames. Neighbors-based reranking refines the initial results by leveraging local similarities between frames, while ensemble search combines multiple search strategies to enhance retrieval robustness.

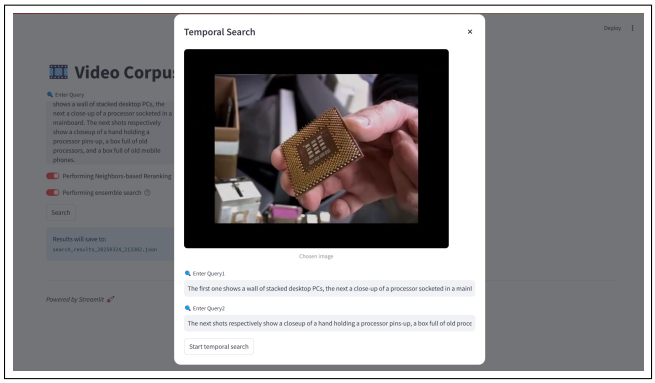
Top-100 Keyframe Display. After the user submits the query, the system returns up to 100 keyframes that best match the textual description (**Figure 4a**). Each keyframe is displayed with a timestamp, allowing users to quickly assess and compare different candidates.

Temporal Search with Dual Queries. To accurately pinpoint the desired video segment, users provide **two separate textual descriptions** (**Figure 4b**): one describing the **start** of the moment and another describing the **end**. Based on these two mini-queries, the system suggests a start frame and an end frame within the relevant video. **Figure 4c** shows an example interface where the system highlights the proposed start frame in green and the proposed end frame in red. Users can review these suggestions and adjust them if necessary to refine the moment boundaries.

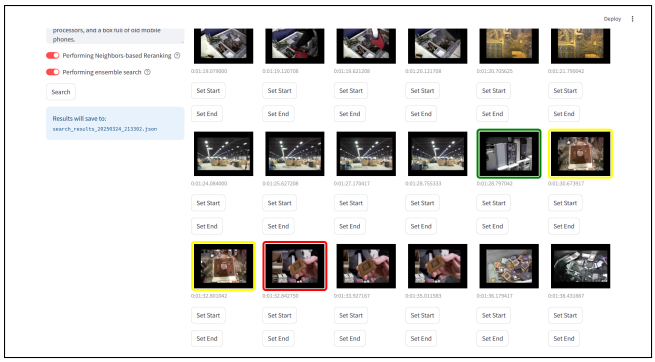
Finalizing the Moment. Once satisfied with the suggested or adjusted frames, the user confirms the selection. The system then extracts and logs the identified segment, which can subsequently be used for more detailed examination or for tasks such as QA.



(a) Overall User Interface



(b) Temporal Search User Interface



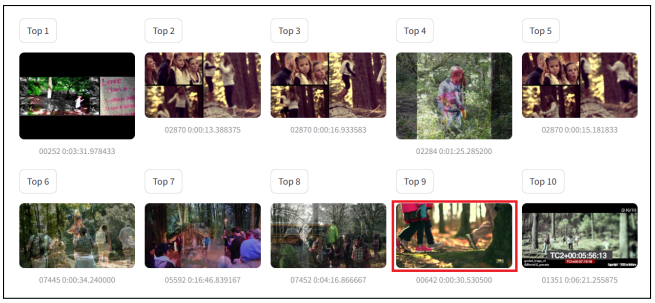
(c) Boundary Selection

Figure 4. The UI for Interactive Retrieval System. The selected start frame will be annotated as a frame with green border, whereas the end frame will be annotated as a frame with red border

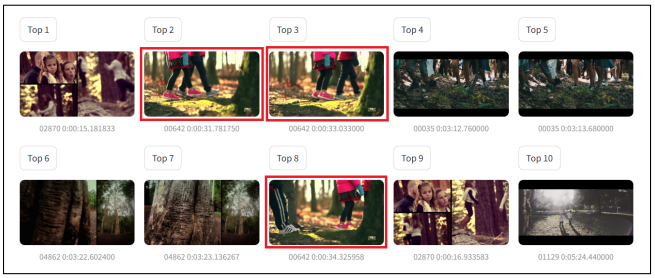
3.6.2. Video Question and Answering (QA)

After the system identifies the relevant segment via *temporal search*, users can perform **Video Question Answering (QA)** by observing the extracted frames. Unlike automated QA systems, our approach relies on users to examine the displayed segment and derive the most suitable answer based on their own observations.

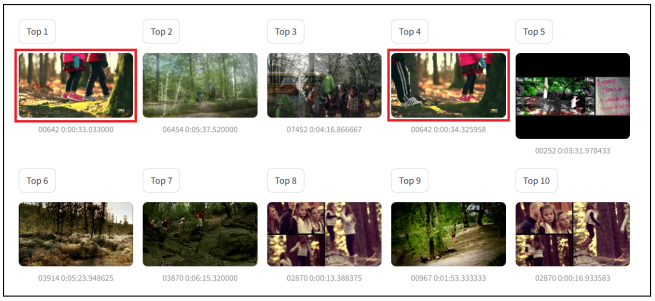
Observation-based QA has a number of benefits, such as increased accuracy through the use of direct observation instead of assumptions, increased reliability through real-time data gathering, and greater flexibility in adapting to changing environments. It also helps detect subtle anomalies that VideoQA models might miss, leading to more effective de-



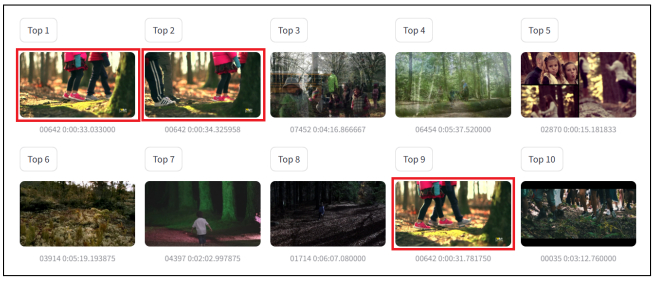
(a) Using only BeIT-3 features (no rerank or ensemble).



(b) Neighbors-based reranking approach.



(c) Ensemble search (BeIT-3 + OpenCLIP).



(d) Combining both ensemble and reranking strategies.

Figure 5. Experimental results for Known-Item Search. The actual target frame is highlighted with a red box.

cision making and better overall quality assurance.

4. Experimental Results

4.1. Known-Item Search

4.1.1. Reranking and Ensemble Search

We evaluate our system on a *Known-Item Search* task, aiming to retrieve specific frames from a video based on a user-defined query. Three retrieval strategies are examined:

1. **Single-Model (BeIT-3):** Relying solely on the BeIT-3 feature extractor to match frames against the query.
2. **Neighbors-Based Reranking:** Refining the initial candidate list by leveraging stable local neighborhoods, thereby boosting frames that are contextually consistent.
3. **Ensemble (BeIT-3 + OpenCLIP):** Combining two feature extractors to incorporate multiple "views" of the query for greater robustness.
4. **Combining Ensemble and Reranking:** Integrating both ensemble search and reranking to maximize robustness, handling variations in difference perspectives.

Query Example: *"Two scenes in a forest. In the first shot, several people are walking as sunlight shines on the ground, and only the lower body of the people is visible. On the right side of the frame there is a tree covered in green moss. In the second shot, we know that they are children walking through the woods."*

Single-Model Baseline (Figure 5a). Using only BeIT-3 features provides a baseline for performance. While the system identifies one correct frame, it is ranked relatively low, implying that a single-model approach struggles with nuanced scene details-such as partially visible bodies or the presence of a moss-covered tree. This limitation underscores the need for strategies that capture both fine-grained and coarse-grained cues.

Neighbors-Based Reranking (Figure 5b). To address the shortcomings of the baseline, we introduce a neighbor-based reranking step. By examining local neighborhoods in feature space, the system promotes frames that share stable visual cues, thus better aligning with the query context. The results show a notable increase in top-ranked correct frames, indicating that spatial/temporal consistency plays a crucial role in distinguishing truly relevant frames from visually similar distractors.

Ensemble Search (Figure 5c). Next, we integrate BeIT-3 and OpenCLIP into an ensemble to combine multiple "views" of the data. BeIT-3 excels at detailed, fine-grained matching, while OpenCLIP offers strong semantic alignment. Merging these representations often propels correct frames into top-1 or top-5 positions, enhancing the system's ability to detect subtle scene elements and high-level thematic cues simultaneously. In practice, this synergy is particularly beneficial for queries that describe both specific objects (e.g., mossy tree) and overarching context (children walking).

Combining Ensemble and Reranking (Figure 5d). Finally, uniting both ensemble search and neighbors-based reranking yields the most robust results. By first leveraging diverse feature extractors and refining through local coherence, the system effectively handles challenges such as partial occlusions, lighting shifts, and complex motion. Across all trials, frames marked with red bounding boxes consistently surface at the top ranks, demonstrating the combined advantage of ensemble and reranking over the single-model

baseline in terms of precision and stability.

4.1.2. Temporal Search

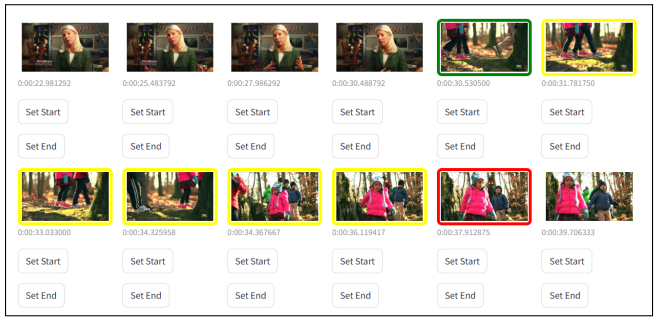


Figure 6. Temporal Search Example. The start and end frames are enclosed by a green and red box, respectively.

Building on the enhanced list of candidate frames from the image-based retrieval stage, we introduce a **temporal search** mechanism. Unlike single-frame retrieval, temporal search aims to locate a continuous video segment that evolves from a "start" description to an "end" description, thus capturing the natural progression of events. Specifically, we use two distinct textual queries:

- **Start Query:** "In the first shot, several people are walking, and on the right side of the frame there is a tree covered in green moss; the camera only shows their lower bodies."
- **End Query:** "In the second shot, we learn that they are children walking through the woods."

By incorporating temporal constraints-such as the chronological order and semantic linkage between two events-our system more accurately reconstructs the full narrative of the user's query. This approach is grounded in the notion that many video concepts (e.g., characters entering or leaving the frame, environment changes) are inherently sequential and cannot be fully represented by a single static image. Consequently, temporal search delivers a more holistic view, allowing users to observe how a scene unfolds between the designated start and end points.

As demonstrated in **Figure 6**, this dual-query approach consistently yields a focused segment containing the relevant frames between the start and end descriptions. Even in videos featuring gradual transitions or subtle camera movements, the temporal search mechanism effectively localizes the moment of interest. In practice, users reported that specifying two queries not only helped them retrieve more accurate results but also made the retrieval process feel more natural, as it closely mirrored how people describe events in everyday conversation (i.e., "It starts when X happens and ends when Y occurs.").

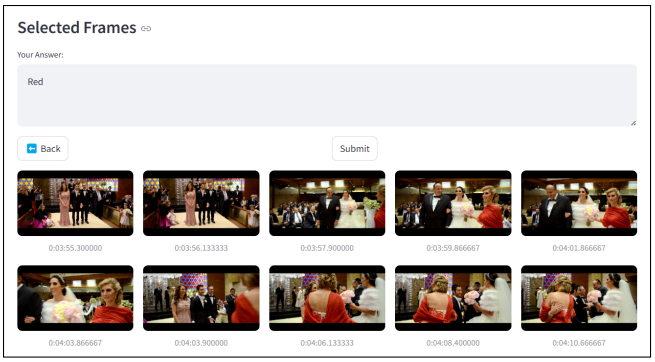


Figure 7. Example QA interface. After retrieving the relevant moment, the system displays a sequence of frames that help the user answer the posed question.

4.2. Question-Answering

In the **Question-Answering (QA)** task, the user first employs our Known-Item Search approach to pinpoint the specific moment of interest in a long video. Once that segment is located, the system presents a series of sequential frames spanning from the user-defined start to the end of the event. This detailed, frame-by-frame visualization allows the user to observe contextual and visual cues that would otherwise be lost in single-frame retrieval.

Query Example: *"The groom, flanked by family, awaits his bride with excitement and anticipation. The radiant bride walks down the aisle, escorted by her proud parents, carrying a bouquet of fresh flowers. The bride approaches the groom, as guests witness this special moment in a beautifully adorned ceremony hall. What color is the bride's mother's dress?"*

Once the user has inspected the relevant frames as shown in **Figure 7**, they input their final answer (e.g., *"Red"* for the mother's dress color) into the answer box. The system then records the user's response along with any selected frames. This mechanism not only boosts transparency but also aids future analysis, allowing users or evaluators to verify how a particular answer was derived.

5. Conclusion

In this work, we propose a unified framework for interactive video retrieval that addresses the challenges of long-form content. By integrating ensemble search, storage optimization, temporal search, and temporal reranking, our approach overcomes the limitations of existing systems, enhancing both accuracy and efficiency. Through a combination of coarse- and fine-grained retrieval models, our method ensures precise content identification while minimizing redundancy. Our framework demonstrates the promise of human-computer collaboration, offering a scalable solution for content-based video search and multimedia analysis, with strong performance on known-item and question-answering tasks.

References

[1] AI VIETNAM — aivietnam.edu.vn. 1

[2] Giuseppe Amato, Paolo Bolettieri, Fabio Carrara, Franca Debole, Fabrizio Falchi, Claudio Gennaro, Lucia Vadicamo, and Claudio Vairo. The visione video search system: Exploiting off-the-shelf text search engines for large-scale video retrieval, 2021. 2

[3] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 2

[4] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 162–171, 2018. 2

[5] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4969–4978, 2019. 1

[6] Shaoxiang Chen and Yu-Gang Jiang. Semantic proposal for activity localization in videos via sentence query. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8199–8206, 2019. 2

[7] Shaoxiang Chen, Wenhao Jiang, Wei Liu, and Yu-Gang Jiang. Learning modality interaction for temporal sentence localization and event captioning in videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 333–351. Springer, 2020. 2

[8] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10638–10647, 2020. 1

[9] Sheng Fang, Shuhui Wang, Junbao Zhuo, Qingming Huang, Bin Ma, Xiaoming Wei, and Xiaolin Wei. Concept propagation via attentional knowledge graph reasoning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4789–4800, 2022. 1

[10] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 1

[11] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 2

[12] Aleksandr Gordeev, Vladimir Dokholyan, Irina Tolstykh, and Maksim Kuprashevich. Saliency-guided detr for moment retrieval and highlight detection. *arXiv preprint arXiv:2410.01615*, 2024. 2

[13] Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus. End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision*, 124(2):237–254, 2017. 2

[14] Yongxin Guo, Jingyu Liu, Mingda Li, Dingxin Cheng, Xiaoying Tang, Dianbo Sui, Qingbin Liu, Xi Chen, and Kevin Zhao. Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. *arXiv preprint arXiv:2405.13382*, 2024. 2

[15] Nico Hezel, Konstantin Schall, Klaus Jung, and Kai Uwe Barthel. Video search with sub-image keyword transfer using existing image archives. In *MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II*, page 484–489, Berlin, Heidelberg, 2021. Springer-Verlag. 1

[16] Zhijian Hou, Chong-Wah Ngo, and Wing Kwong Chan. Conquer: Contextual query-aware ranking for video corpus moment retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3900–3908, 2021. 2

[17] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14271–14280, 2024. 2

[18] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 30:2340–2349, 2021. 1

[19] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 1

[20] Minh-Dung Le-Quynh, Anh-Tuan Nguyen, Anh-Tuan Quang-Hoang, Van-Huy Dinh, Tien-Huy Nguyen, Hoang-Bach Ngo, and Minh-Hung An. Enhancing video retrieval with robust clip-based multimodal system. In *Proceedings of the 12th International Symposium on Information and Communication Technology*, page 972–979, New York, NY, USA, 2023. Association for Computing Machinery. 1

[21] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 447–463. Springer, 2020. 2

[22] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858, 2021. 2

[23] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020. 2

[24] Pandeng Li, Chen-Wei Xie, Hongtao Xie, Liming Zhao, Lei Zhang, Yun Zheng, Deli Zhao, and Yongdong Zhang. Momentdiff: Generative video moment retrieval from random to real. *Advances in neural information processing systems*, 36:65948–65966, 2023. 2

[25] Xirong Li, Chaoxi Xu, Gang Yang, Zhineng Chen, and Jianfeng Dong. W2vv++ fully deep learning for ad-hoc video

search. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1786–1794, 2019. 1

[26] Kaiqu Liang and Samuel Albanie. Simple baselines for interactive video retrieval with questions and answers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11091–11101, 2023. 1, 2

[27] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 15–24, 2018. 2

[28] Jakub Lokoč, Patrik Veselý, František Mejzlík, Gregor Kovalčík, Tomáš Souček, Luca Rossetto, Klaus Schoeffmann, Werner Bailer, Cathal Gurrin, Loris Sauter, Jaeyub Song, Stefanos Vrochidis, Jiaxin Wu, and Björn þóR Jónsson. Is the reign of interactive search eternal? findings from the video browser showdown 2020. 17(3), 2021. 1

[29] Chujie Lu, Long Chen, Chilie Tan, Xiaolin Li, and Jun Xiao. Debug: A dense bottom-up grounding approach for natural language video localization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5144–5153, 2019. 2

[30] Zhixin Ma and Chong Wah Ngo. Interactive video corpus moment retrieval using reinforcement learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 296–306, 2022. 1, 2

[31] Fotini Markatopoulou, Anastasia Moumtzidou, Damianos Galanopoulos, Konstantinos Avgerinakis, Stelios Andreadis, Ilias Gialampoukidis, Stavros Tachos, Stefanos Vrochidis, Vasileios Mezaris, Yiannis Kompatsiaris, and Ioannis Patras. Iti-certh participation in trecvid 2017. In *TREC Video Retrieval Evaluation*, 2017. 1

[32] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23023–23033, 2023. 2

[33] Ba Hung Ngo, Ba Thinh Lam, Thanh Huy Nguyen, Quang Vinh Dinh, and Tae Jong Choi. Dual dynamic consistency regularization for semi-supervised domain adaptation. *IEEE Access*, 2024. 1

[34] Khoi Anh Nguyen, Linh Yen Vu, Thang Dinh Duong, Thuan Nguyen Duong, Huy Thanh Nguyen, and Vinh Quang Dinh. Enhancing vietnamese vqa through curriculum learning on raw and augmented text representations. *arXiv preprint arXiv:2503.03285*, 2025.

[35] Thanh-Huy Nguyen, Thien Nguyen, Xuan Bach Nguyen, Nguyen Lan Vi Vu, Vinh Quang Dinh, and Fabrice MERIAUDEAU. Semi-supervised skin lesion segmentation under dual mask ensemble with feature discrepancy co-training. In *Medical Imaging with Deep Learning*. 1

[36] Tien-Huy Nguyen, Quang-Khai Tran, and Anh-Tuan Quang-Hoang. Improving generalization in visual reasoning via self-ensemble, 2024. 2

[37] Tho-Quang Nguyen, Huu-Loc Tran, Tuan-Khoa Tran, Huu-Phong Phan-Nguyen, and Tien-Huy Nguyen. Fa-yolov9: Improved yolov9 based on feature attention block. In *2024 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6, 2024. 1

[38] Xuan-Bach Nguyen, Hoang-Thien Nguyen, Thanh-Huy Nguyen, Nhu-Tai Do, and Quang Vinh Dinh. Emotic masked autoencoder on dual-views with attention fusion for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4784–4792, 2024. 1

[39] Zhaobo Qi, Yibo Yuan, Xiaowen Ruan, Shuhui Wang, Weigang Zhang, and Qingming Huang. Collaborative debias strategy for temporal sentence grounding in video. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2

[40] Guoping Qiu. Challenges and opportunities of image and video retrieval. *Frontiers in Imaging*, 1, 2022. 2

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 3, 5

[42] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, 2017. 1

[43] Cristian Rodriguez, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdong Li, and Stephen Gould. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2464–2473, 2020. 2

[44] Luca Rossetto, Ivan Giangreco, Claudiu Tanase, and Heiko Schuldt. vitrivr: A flexible retrieval stack supporting multiple query modes for searching in multimedia collections. In *Proceedings of the 24th ACM International Conference on Multimedia*, page 1183–1186, New York, NY, USA, 2016. Association for Computing Machinery. 2

[45] Tomás Soucek and Jakub Lokoc. Transnet v2: An effective deep network architecture for fast shot transition detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11218–11221, 2024. 3

[46] Hao Sun, Mingyao Zhou, Wenjing Chen, and Wei Xie. Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4998–5007, 2024. 2

[47] Xin Sun, Jialin Gao, Yizhe Zhu, Xuan Wang, and Xi Zhou. Video moment retrieval via comprehensive relation-aware network. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):5281–5295, 2023. 2

[48] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations, 2016. 2

[49] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 1

[50] Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. *arXiv preprint arXiv:2410.03290*, 2024. 2

[51] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186, 2023. 3, 5

[52] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 1

[53] Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. Hawkeye: Training video-text llms for grounding text in videos. *arXiv preprint arXiv:2403.10228*, 2024. 2

[54] Chao-Yuan Wu and Philipp Krähenbühl. Towards long-form video understanding, 2021. 1

[55] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1–10, 2021. 2

[56] Sunjae Yoon, Ji Woo Hong, Eunseop Yoon, Dahyun Kim, Junyeong Kim, Hee Suk Yoon, and Chang D Yoo. Selective query-guided debiasing for video corpus moment retrieval. In *European Conference on Computer Vision*, pages 185–200. Springer, 2022. 2

[57] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9159–9166, 2019. 2

[58] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10287–10296, 2020. 2

[59] Bowen Zhang, Hexiang Hu, Joonseok Lee, Ming Zhao, Sheide Chammas, Vihan Jain, Eugene Ie, and Fei Sha. A hierarchical multi-modal encoder for moment localization in video corpus. *arXiv preprint arXiv:2011.09046*, 2020. 2

[60] Hao Zhang, Aixin Sun, Wei Jing, Guoshun Nan, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. Video corpus moment retrieval with contrastive learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 685–695, 2021.

[61] Xuemei Zhang, Peng Zhao, Jinsheng Ji, Xiankai Lu, and Yilong Yin. Video corpus moment retrieval via deformable multigranularity feature fusion and adversarial training. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 2

[62] Zijian Zhang, Zhou Zhao, Zhu Zhang, Zhijie Lin, Qi Wang, and Richang Hong. Temporal textual localization in video via adversarial bi-directional interaction networks. *IEEE Transactions on Multimedia*, 23:3306–3317, 2020. 2