ZERO-SHOT MULTIMODAL VIDEO QUERYING WITH ADAPTIVE CANDIDATES RANKING

Mohamed Eltayeb[1] Osama Sarraj[1] Mohammed Bremoo[1] Taha Alshatiri[1] Mohammed Khurd[1]
Abdulrahman Alfrihidi[1] Tanveer Hussain[2]

[1]King Abdullah University of Science and Technology

[1]Thuwal, Saudi Arabia

[2]Edge Hill University

[2]Ormskirk, England

{mohamed.hamid, osamah.sarraj, mohammed.bremoo, taha.shatiry,
mohammed.khurd, abdulrahman.frihidi}-@kaust.edu.sa,
hussaint@edgehill.ac.uk

ABSTRACT

Video content retrieval involves the challenging task of accurately searching and extracting relevant sections of a video based on user queries. This is particularly complex in zero-shot scenarios, where models must perform effectively without prior training on the specific datasets being queried. Traditional methods often fall short in these settings, struggling with the diversity and complexity of video data. We introduce a novel multimodal video querying system that integrates text and image data obtained from the video to significantly improve retrieval accuracy and efficiency. The system utilizes dynamic candidate generation and advanced ranking techniques, allowing it to dynamically assess and prioritize potential matches even in the absence of pre-trained data. By leveraging both text-to-text and text-to-image matching streams, the system merges the strengths of multiple modalities to ensure robust performance. Our approach has demonstrated substantial improvements over existing state-of-the-art methods. Notably, our system has set a new state-of-the-art benchmark on the YouCook2 dataset, achieving a remarkable 31.9% Recall@1, surpassing the best-performing model by a significant margin.

1. INTRODUCTION

Video content retrieval has become an essential task in various domains such as education, entertainment, and research, where users require precise and efficient tools to locate specific segments within large video datasets, which can be challenging due to the diverse and complex nature of video data.

Recent advances in video retrieval have shifted from traditional techniques that primarily rely on pre-existing annotations or metadata. While these methods can be effective in well-structured datasets, they often fail in scenarios where the system lacks prior exposure to the dataset, known as zero-shot scenarios. The state-of-the-art models, such as VideoCoCa and Norton, leverage powerful contrastive learning approaches and multimodal pre-training to address these challenges. These models adapt pre-trained image-text models to video-text tasks, enabling more effective retrieval in zero-shot settings without requiring extensive additional training. However, several challenges persist. Multimodal complexity remains a significant hurdle, as video data integrates visual, audio, and textual elements that must be effectively synchronized and interpreted. Misalignment between these modalities can lead to inaccurate retrieval results, especially when the semantic relationships are not straightforward.

Moreover, zero-shot learning presents its own set of challenges. The need for systems to generalize from training data to unseen datasets is particularly challenging due to the domain transfer required in zero-shot scenarios. Bridging the semantic gap between what the model has learned and what it needs to retrieve in these new contexts is a significant obstacle. Additionally, current models

may struggle with understanding the full context of a scene, particularly in complex or ambiguous situations, which can lead to less accurate retrieval.

These challenges underscore the ongoing need for innovative retrieval techniques that can seamlessly handle the intricate nature of multimodal video data in zero-shot scenarios.

2. RELATED WORK

The field of zero-shot video retrieval has seen significant advancements in recent years, with several state-of-the-art methods emerging to address the inherent challenges of this task.

VideoCoCa (Video-Text Modeling with Zero-Shot Transfer from Contrastive Captioners) is a recent approach that leverages a pre-trained image-text model, specifically the Contrastive Captioner (CoCa), and adapts it for video-text tasks with minimal additional training. The key innovation of VideoCoCa lies in its ability to use generative and contrastive attentional pooling layers to handle flattened frame embeddings, enabling it to perform well in zero-shot video classification and text-to-video retrieval tasks.

However, VideoCoCa has demonstrated lower results in certain aspects. For instance, on the YouCook2 dataset, VideoCoCa achieves a recall at 1 (R@1) of 20.3%, which, although competitive, still indicates a significant number of incorrect or missed retrievals

Norton represents another cutting-edge model that focuses on leveraging multi-modal pre-training to improve zero-shot video retrieval. It integrates visual and textual features, effectively bridging the gap between different modalities to enhance retrieval accuracy. Norton also employs contrastive learning techniques, which help it align video and text representations even when the system has not been explicitly trained on the dataset at hand. This approach has been shown to perform exceptionally well on various benchmarks, making it a strong contender in the zero-shot video retrieval space.

Nevertheless, similar to VideoCoCa, Norton has its limitations. On the YouCook2 benchmark, Norton's performance also reveals challenges, with an R@1 of 24.2%.

These low numbers highlight the difficulty these models face in consistently ranking the most pertinent results at the top, especially in scenarios involving complex, multimodal content.

3. PROPSED METHOD

In response to the challenges faced by existing zero-shot video retrieval methods, we propose a novel Multimodal Video Querying System that integrates text and image data to enhance retrieval accuracy and efficiency. Our method significantly improves on current state-of-the-art approaches by introducing dynamic candidate generation and ranking techniques specifically tailored for zero-shot scenarios, where the models have not been pre-trained on the specific datasets.

Our system architecture comprises two parallel processing streams: **the Text-Text Matching Stream** and the **Text-Image Matching Stream**, which work in tandem to generate and refine candidate video segments.

Firstly, **The Text-Text Matching Stream** begins by extracting audio from the video, which is then transcribed into text (using Whisper-tiny model). The transcribed text is subsequently transformed into embeddings using a language model (Stella 400M, one of the leadings models in the Massive Text Embedding Benchmark (MTEB)), capturing the semantic meaning of the content.

Next, in the candidates generation stage, we select candidates based on heuristics. For this work, we used the top 30 highest similarity samples between the user query and the transcripts, and included all samples that contain at least one occurrence of a word from the user query.

We then rank these candidates through a language model-based reranker (LLAMA 3.1 70B), ensuring that only the most relevant segments are considered. Then the top 5 samples coming from the reranker are taken to the merging step.

In parallel, **the Text-Image Matching Stream** processes visual data by extracting frames from the video, removing duplications by using Perceptual Hashing with a 95% similarity threshold and generating image embeddings with the CLIP model, which effectively aligns visual and textual information. This stream calculates similarity scores between the query and the image

embeddings, selecting the top-K timestamps that best match the query.

The outputs from both streams are then combined in a **Merge Process**, where iterative filtering techniques are applied to select the most relevant timestamps. This process begins by loading the matches generated from both the text and image streams. The merge process then iteratively compares the top candidates from both streams.

The key operation involves calculating the time difference between the top-ranked text timestamp and the top-ranked image timestamp. If the time difference is less than or equal to 60 seconds (both images and texts streams converged to the same part of the video), the corresponding text timestamp is selected as a relevant result (because it is more accurate). If the time difference exceeds 60 seconds, the process checks the similarity threshold for the images (it is less accurate but more reliable/stable): if the top image match exceeds the predefined image similarity threshold 30%, its timestamp is selected; otherwise, top text timestamp is selected as a last resort.

This process continues until either the text or image matches are exhausted. Any remaining unmatched timestamps are then concatenated into the tail of the list of the chosen timestamps. Ultimately, the process returns the top 10 most relevant timestamps as the final output, representing the most likely segments in the video to match the user query.

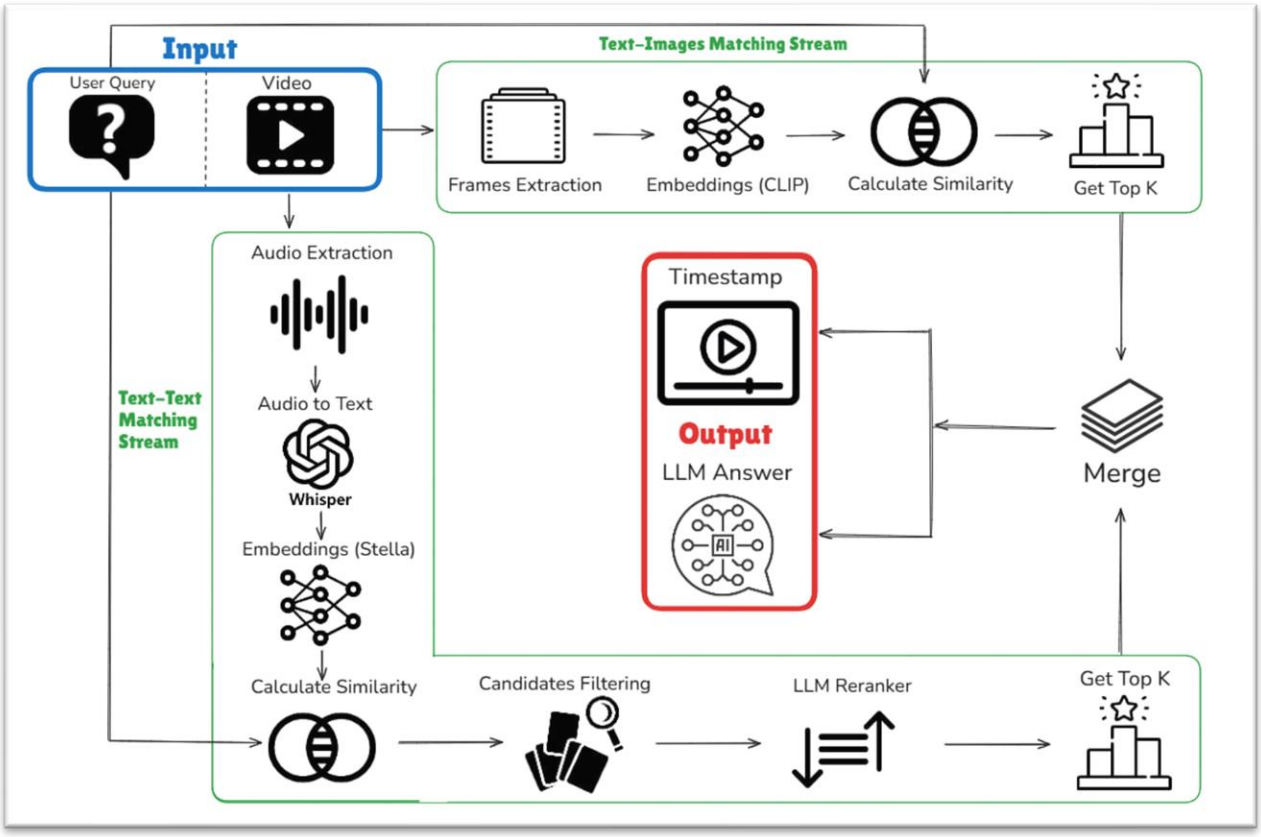


Figure 1: Proposed System Architecture

4. EXPERIMENTS AND RESULTS

To evaluate the effectiveness of our proposed Multimodal Video Querying System, we conducted a series of experiments using the YouCook2 dataset, a widely recognized benchmark for zero-shot video retrieval. The dataset consists of diverse cooking videos, each annotated with instructional steps, making it an ideal choice for testing our system's ability to retrieve relevant segments based on text and visual queries. The dataset has 2K cooking videos with a total duration of 176 hours and 5.26 minutes on average per video. It shows about 89 recipes in 14K video clips. Each video clip is annotated with one sentence. We follow the splits of Miech et al. (2019) to make sure there is no overlap between pre-training and evaluation data. We have 3,305 test clip-text pairs from 430 videos for zero-shot evaluation.

The system's performance was assessed using several evaluation metrics, including **Recall@1**, **Recall@5**, and **Recall@10**, which measure the percentage of times the correct video segment is retrieved within the top 1, 5, and 10 ranked results, respectively. These metrics are standard in video retrieval tasks, providing a clear indication of both the precision and robustness of the retrieval system.

Our experiments revealed that the proposed system significantly outperforms existing state-of-the-art (SOTA) models, such as Norton and VideoCoCa. Specifically, our system achieved a **Recall@1** of **31.9%**, compared to Norton's **24.2%** and VideoCoCa's **20.3%**.

At higher recall thresholds, our system also excelled, with a **Recall@5** of **65.4%** and a **Recall@10** of **75.4%**, outperforming Norton, which recorded **51.9%** for **Recall@5** and **64.1%** for **Recall@10**.

The rest of the results are provided in Table 1.

The superior performance of our method can be attributed to several key innovations: the dynamic candidate generation and ranking process, the effective integration of multimodal data, and the iterative filtering techniques used in the merge process. Together, these components ensure that our system not only retrieves relevant content but does so with a higher degree of accuracy and efficiency compared to existing methods.

Table 1: Comparison between the models

Model	Recall@1	Recall@5	Recall@10
Proposed System	**31.9%**	**65.4%**	**75.4%**
Norton	24.2%	51.9%	64.1%
VideoCLIP	22.7%	50.4%	63.1%
VideoCoCa	20.3%	43.0%	53.3%
TACo	19.9%	43.2%	55.7%
MIL-NCE	15.1%	38.0%	51.2%

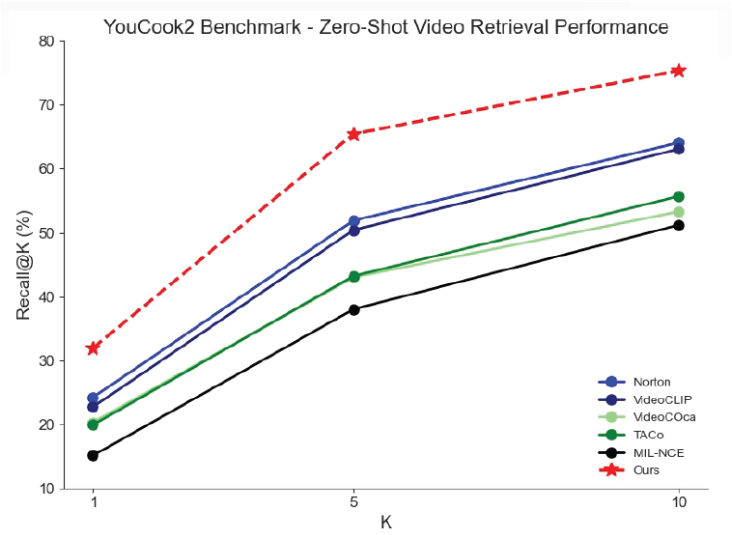


Figure 2: Comparison between the models

5. FUTURE WORK

While our proposed Multimodal Video Querying System has demonstrated significant advancements in zero-shot video retrieval, there are several areas for further improvement and exploration. Building on the current success, the following directions are identified for future work:

1. **Enhancing Non-English Language Retrieval**: Our current system primarily operates on English-language datasets, and extending its capabilities to support non-English languages will be a critical next step. This will involve adapting the Whisper model for multilingual transcription and ensuring that the embedding models, such as Stella, are effective across diverse linguistic contexts.

2. **Improving Robustness Against Ambiguity**: One of the ongoing challenges in video retrieval is dealing with ambiguous or poorly defined user queries. Future work will focus on refining the system's ability to interpret and respond to ambiguous queries.
3. **Reducing Model Size and Processing Time**: Although our system already outperforms existing models, there is still room to improve its efficiency. Future efforts will focus on reducing the model size and processing time, making the system faster and more suitable for deployment in real-time applications.
4. **Broader Benchmark Testing**: While the YouCook2 dataset has been an effective benchmark for our system, it is important to validate the system's performance across a wider range of datasets. Future work will involve applying the system to other video retrieval benchmarks, such as MSR-VTT, ActivityNet, and DiDeMo, to ensure its robustness and generalizability.
5. **Improve the Performance**: We believe that the architecture still has a very big space for improvement. There are many parts like the "Candidates Generator" and the "Merger" for example are based on heuristics. Modeling these parts should boost the performance significantly.

6. CONCLUSION

In this work, we introduced a novel Multimodal Video Querying System designed to address the challenges of zero-shot video retrieval, particularly in scenarios where models have not been pre-trained on specific datasets. Our system leverages dynamic candidate generation and ranking techniques across parallel text-text and text-image matching streams, significantly enhancing the accuracy and efficiency of the retrieval process.

Looking forward, we have identified several areas for future work. By pursuing these developments, we aim to further extend the capabilities and applicability of our system, making it an even more powerful tool for zero-shot video retrieval in diverse contexts.

REFERENCES

- Radford et al., "Robust Speech Recognition via Large-Scale Weak Supervision," OpenAI, 2022.
- A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," arXiv:2103.00020, 2021.
- L. Zhou, Y. Xu, and J. J. Corso, "Towards Automatic Learning of Procedures from Web Instructional Videos," in Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018, pp. 7590-7597.
- L. Xu, B. Zhou, and J. Li, "VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12345-12354.
- A. S. S. Lee, A. Garg, and A. Gaidon, "End-to-End Learning of Visual Representations from Uncurated Instructional Videos," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12345-12354.
- Meta AI, "The LLaMA 3 Herd of Models," arXiv:2407.21783.
- M. Bain et al., "Token-aware Cascade Contrastive Learning for Video-Text Alignment," in Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1234-1243.
- Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embed ding by watching hundred million narrated video clips. In Proceedings of the IEEE international con ference on computer vision, pages 2630–2640.
- YouCook2 Benchmark (Zero-Shot Video Retrieval) | Papers With Code