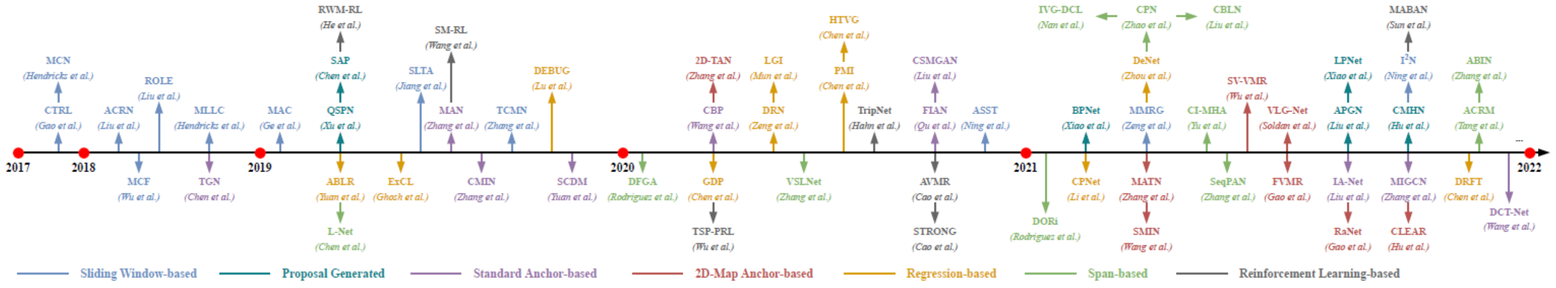
Video Moment Retrieval

Problem, Dataset, and Solution

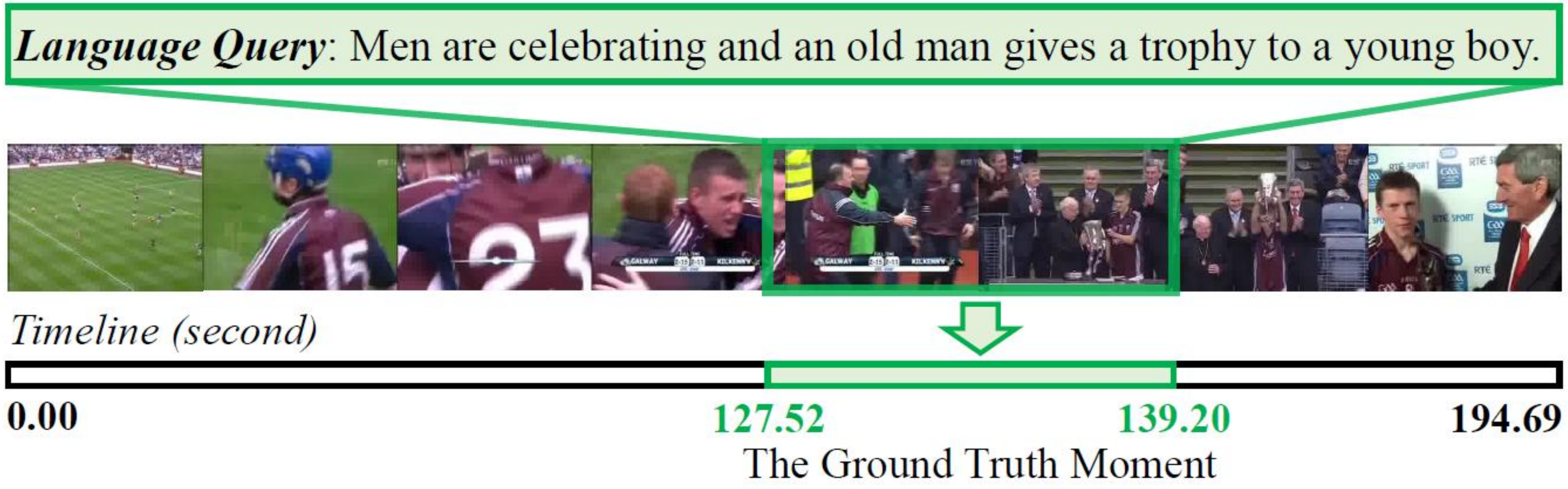
Dr. Sun, Aixin 孙爱欣

NTU Singapore



(Video, Query, Moment) vs (Document, Question, Answer)

Given **an untrimmed video** and **a text query**, *Video Moment Retrieval* (**VMR**) is to **locate a matching span** from the video that semantically corresponds to the query.



Language Query: Men are celebrating and an old man gives a trophy to a young boy.
Timeline (second)
0.00
127.52
139.20
194.69
The Ground Truth Moment

VMR is also known as temporal sentence grounding in videos (**TSGV**), or natural language video localization (**NLVL**)

Span-based Localizing Network for Natural Language Video Localization (Zhang et al., ACL 2020)

A High-level Overview

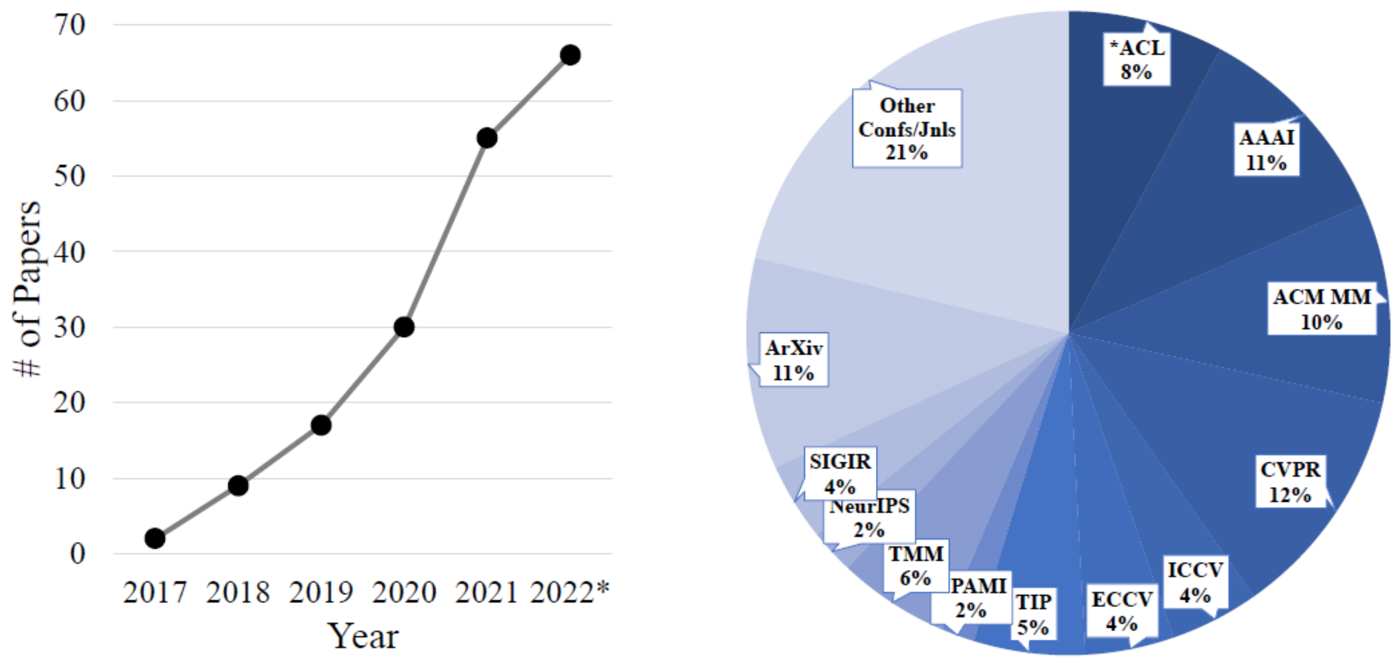
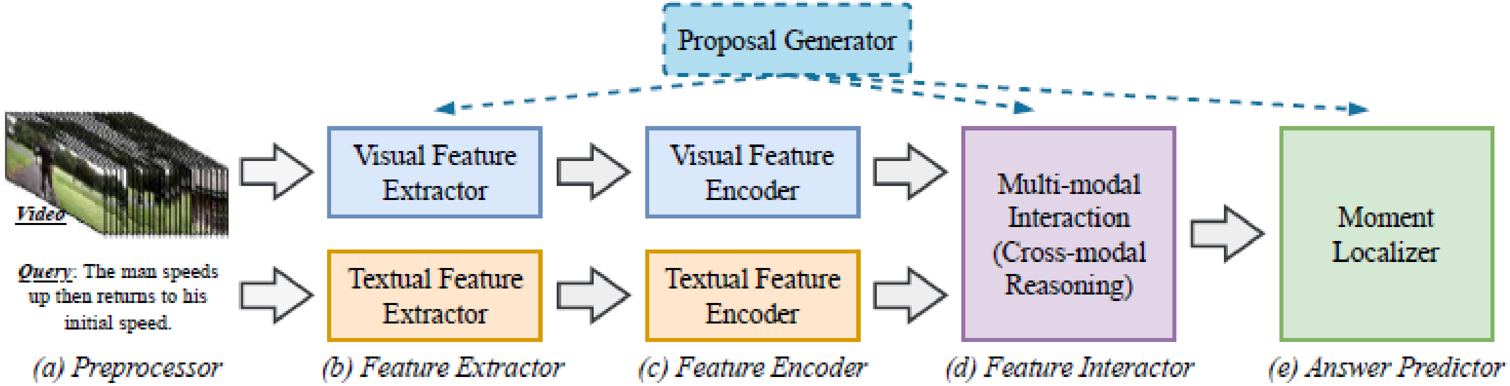


Fig. 2. Statistics of the collected papers in this survey. Left: number of papers published each year (till September 2022). Right: distribution of papers by venue, where *ACL denotes the series of conferences hosted by the Association for Computational Linguistics.

- A fast-growing area
- Multidisciplinary
  - ACL, AAAI, MM, CVPR …
- Practical impact

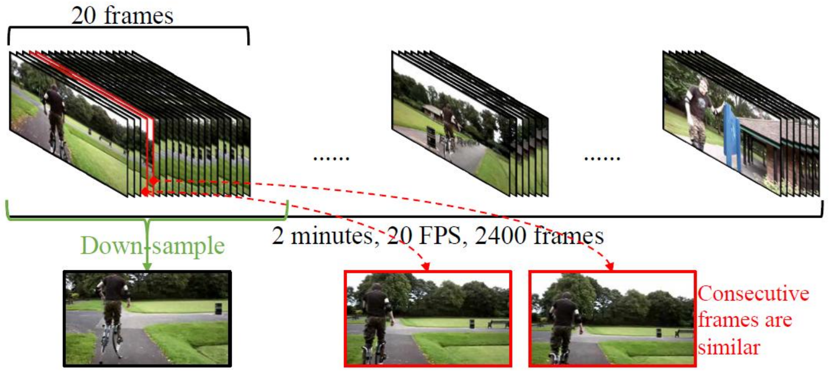
Hao Zhang et al 2023. [Temporal Sentence Grounding in Videos: A Survey and Future Directions]. IEEE TPAMI 45, 8 (Aug. 2023)

A General Pipeline for VMR



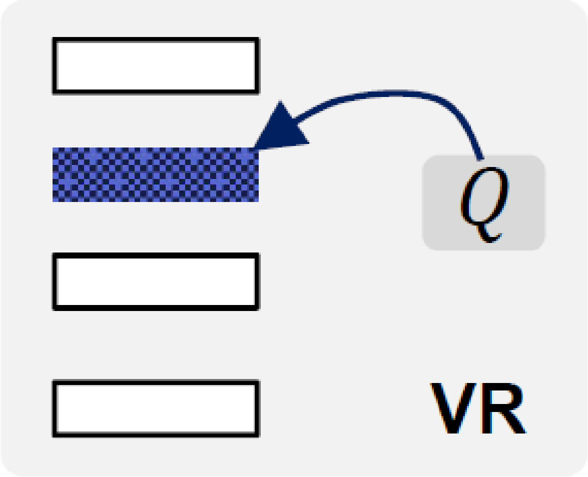
- A **proposal** can be considered as a **candidate answer moment**, e.g., a video segment sampled from the input video.
- Proposal-free methods predict answers directly without the need of generating candidate answers.

A key challenge or limitation: Video is a series of still images and the number of frames can be very large.

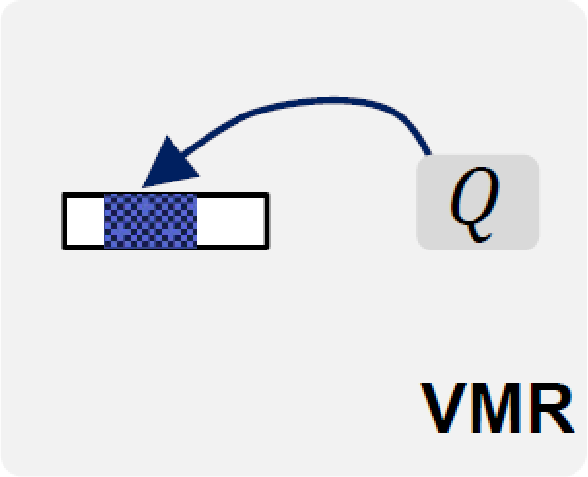


20 frames
......
......
2 minutes, 20 FPS, 2400 frames
Down-sample
Consecutive frames are similar

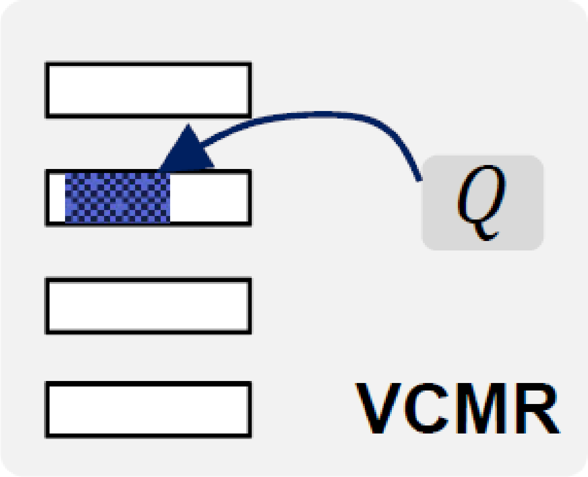
The Related Tasks: VR vs VMR → VCMR



Video Retrieval



Video Moment Ret.

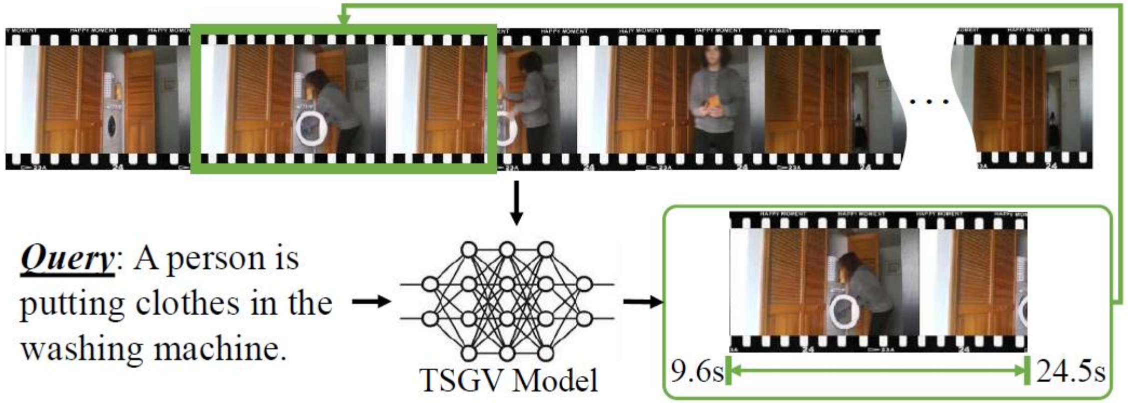


Video Corpus Moment Ret.

- **VR**: retrieve a video from a collection for a given query based on visual content, i.e., video search
- **VCMR**: retrieve a moment from a collection of movies for a query; a direct extension of the VMR (e.g., by adding videos containing no answers)

Is the VMR Task Reasonable?

- **Input**: A video, a query
- **Output**: a matching moment
- **Training/Evaluation**:
    - A few datasets available with manual annotation of ground truth moments to queries



Query: A person is putting clothes in the washing machine.

- **Under what application scenario?**
    - A user would like to query **ONE given video**
    - Expect exact (or at most) **one** answer
    - With a very **detailed description of the moment**

Is the VMR Task Reasonable?

- **Under what application scenario?**
  - A user would like to query **ONE given video**
  - Expect exact (or at most) **one** answer
  - With a very **detailed description of the moment**

- The task is defined by the **Data Annotation**
  - Annotators watch a video, provide textual descriptions of meaningful video moments in video.
  - Each description serves as the query to retrieve the corresponding moment.
  - A query typically describes one specific moment precisely
- **Hidden Assumption**: A model trained on these datasets can assume the existence of the moment to be searched for, and all queries are from users who possess a good understanding of the source video.

To Search for Video Moment in Reality

- User **may not have good knowledge** of the source videos to be searched for
- Queries from user **may not be a precise description** of a moment
- There are likely **many videos** in a search setting
- There are likely **many moments matching the query** with different **levels of relevance**

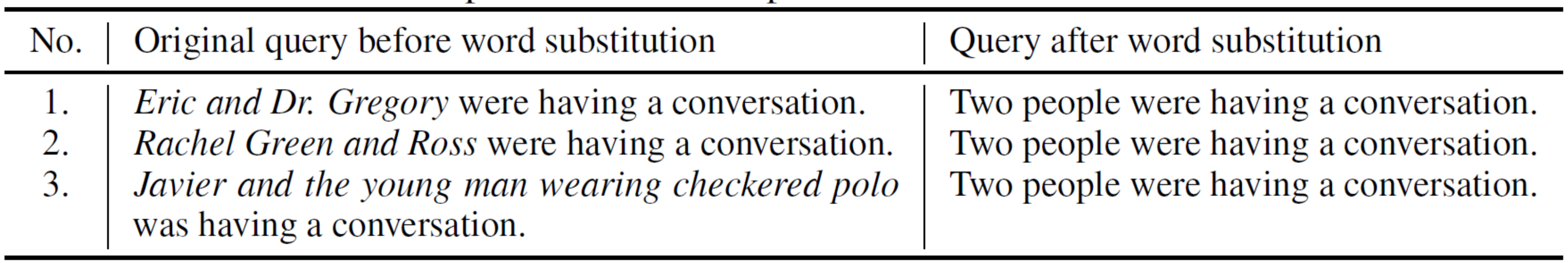
→Ranked Video Moment Retrieval (RVMR)

- To retrieve **a ranked list of moments** matching **an imprecise query** from a **collection of videos**.

The TVR-Ranking Dataset

- The **TVR dataset** contains video clips from six different TV series
  - Created for the VMR task
  - Moments in videos are annotated with precise descriptions
- The **TVR-Ranking** dataset Annotation
  - Convert the precise moment descriptions to imprecise descriptions (we call them *moment captions*) as queries
  - Annotate **level of relevance** between query and candidate moments

Table 2: Three example moment descriptions before and after word substitution.

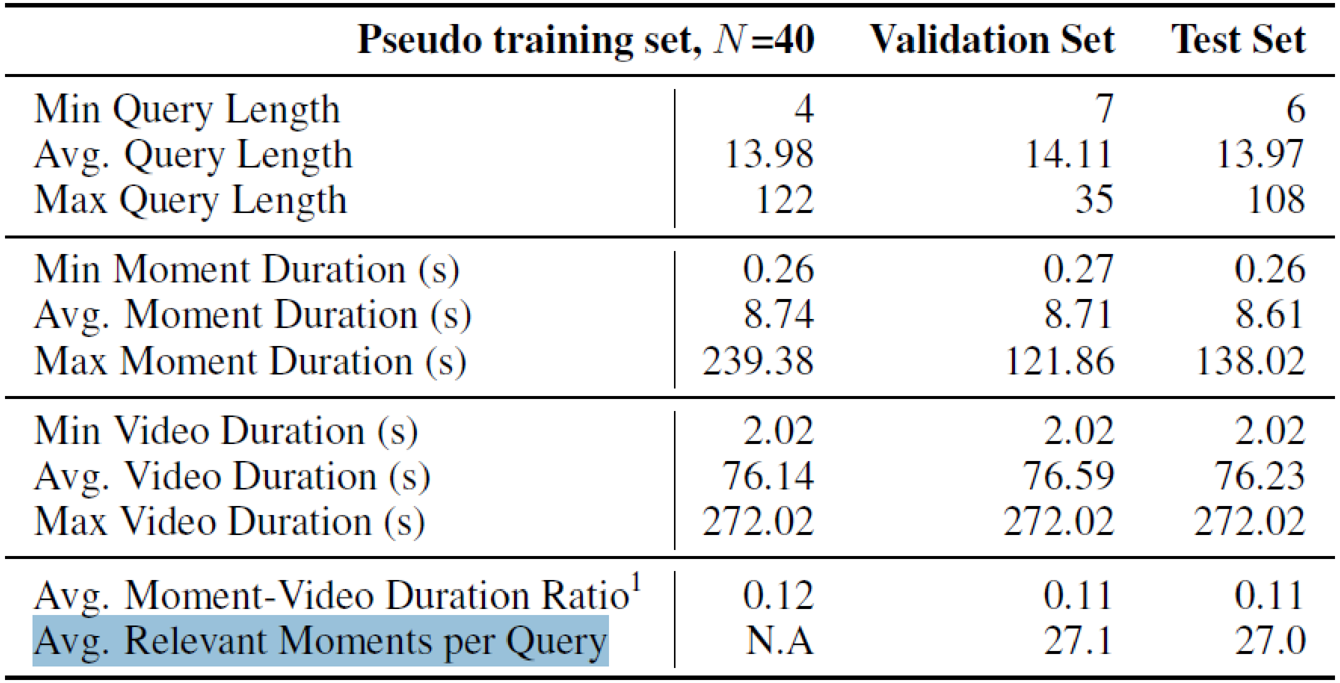


No.	Original query before word substitution	Query after word substitution
1.	*Eric and Dr. Gregory* were having a conversation.	Two people were having a conversation.
2.	*Rachel Green and Ross* were having a conversation.	Two people were having a conversation.
3.	*Javier and the young man wearing checkered polo* was having a conversation.	Two people were having a conversation.

https://www.arxiv.org/abs/2407.06597

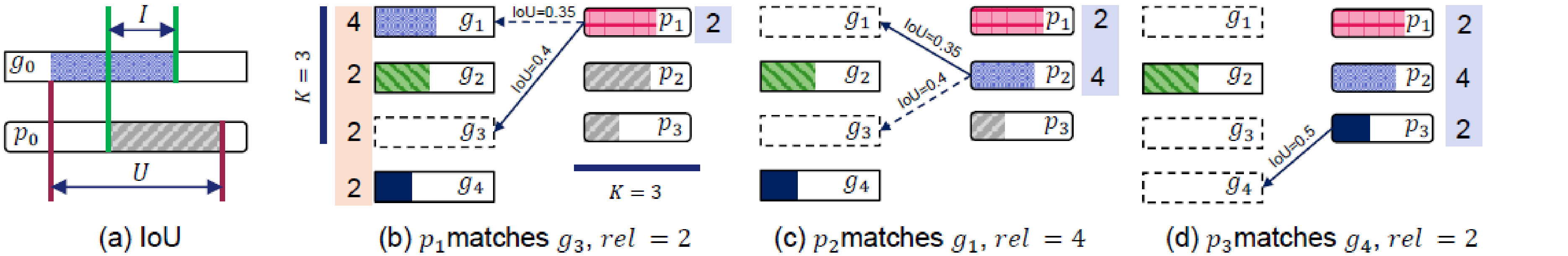
TVR-Ranking

- From the 72,842 moment captions, we randomly select
  - 500 and 2781 moment captions as queries in validation and test sets respectively, for **manual annotation**.
  - The remaining moment captions are used to construct **a pseudo training set**



	Pseudo training set, N=40	Validation Set	Test Set
Min Query Length	4	7	6
Avg. Query Length	13.98	14.11	13.97
Max Query Length	122	35	108
Min Moment Duration (s)	0.26	0.27	0.26
Avg. Moment Duration (s)	8.74	8.71	8.61
Max Moment Duration (s)	239.38	121.86	138.02
Min Video Duration (s)	2.02	2.02	2.02
Avg. Video Duration (s)	76.14	76.59	76.23
Max Video Duration (s)	272.02	272.02	272.02
Avg. Moment-Video Duration Ratio[1]	0.12	0.11	0.11
Avg. Relevant Moments per Query	N.A	27.1	27.0

Evaluation Metric for RVMR: NDCG@K, IoU ≥ μ



(a) IoU (b) $p_1$matches g_3, $rel = 2$ (c) $p_2$matches g_1, $rel = 4$ (d) $p_3$matches g_4, $rel = 2$

Figure 2: Illustration (a) IoU, and (b)–(d) for $NDCG@3$, $\mu = 0.3$. (b) p_1 matches g_3 with $rel = 2$ for the larger IoU, above the 0.3 threshold. (c) p_2 matches g_1 as g_3 is no longer available. (d) p_3 matches g_4, with $rel = 2$.

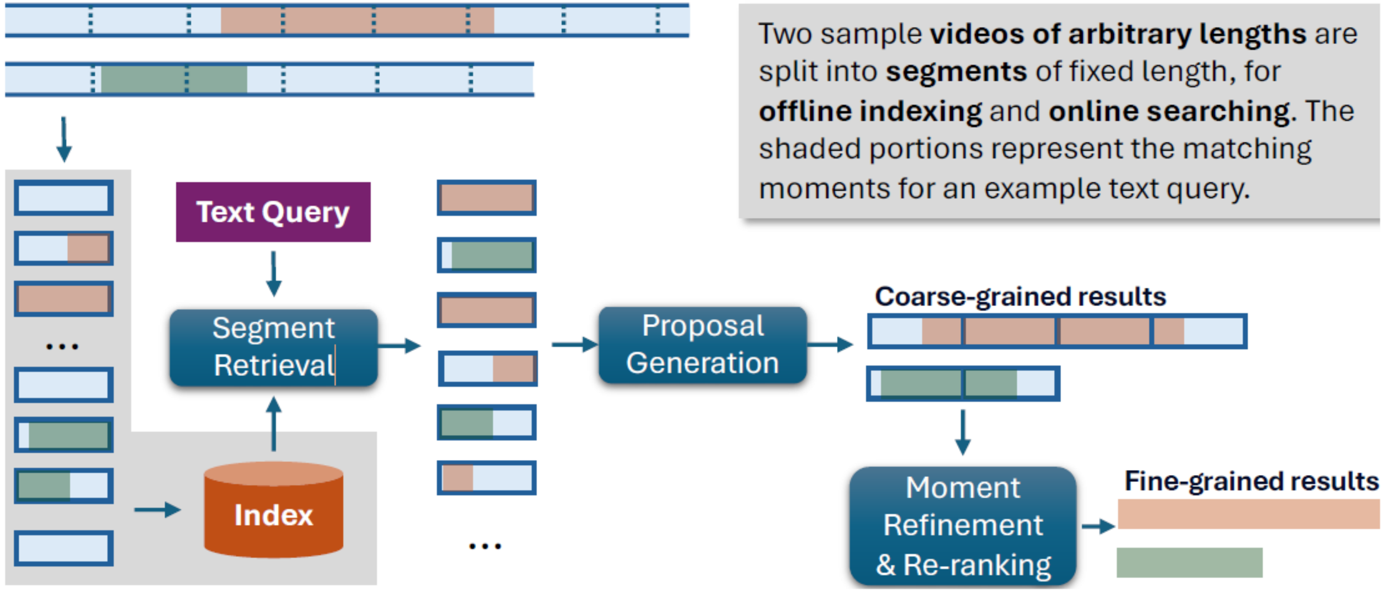
- If a predicted moment has lower IoU than μ with any ground truth, then it is considered zero relevance
- If IoU ≥ μ, then the relevance level of the best match is assigned
- NDCG is computed from the ranking

Where Are We?

- We have a well-defined task: ranked video moment retrieval (RVMR)
- We have a manually annotated dataset TVR-Ranking for validation and testing
- We have defined a new evaluation metric NDCG@K, IoU ≥ μ

- **The main challenges:**
  - Moment retrieval is from a large collection of videos
  - Videos can be in different length
  - There are multiple matching moments with different levels of relevance
  - The retrieval needs to be both efficient and effective

A Flexible and Scalable Framework for Video Moment Search



Two sample **videos of arbitrary lengths** are split into **segments** of fixed length, for **offline indexing** and **online searching**. The shaded portions represent the matching moments for an example text query.

Figure 1: The Segment-Proposal-Ranking (SPR) framework. All videos are divided into non-overlapping, equal-length segments (*e.g.*,, 4 seconds) for indexing and searching. The final results are computed based on the relevant segments retrieved.

- **Segment**
  - Handling videos of arbitrary length
- **Proposal**
  - Reducing target moment candidates over efficient dense retrieval
- **Ranking**
  - Refining proposals and computing the relevance score

https://arxiv.org/abs/2501.05072

Segment Retrieval

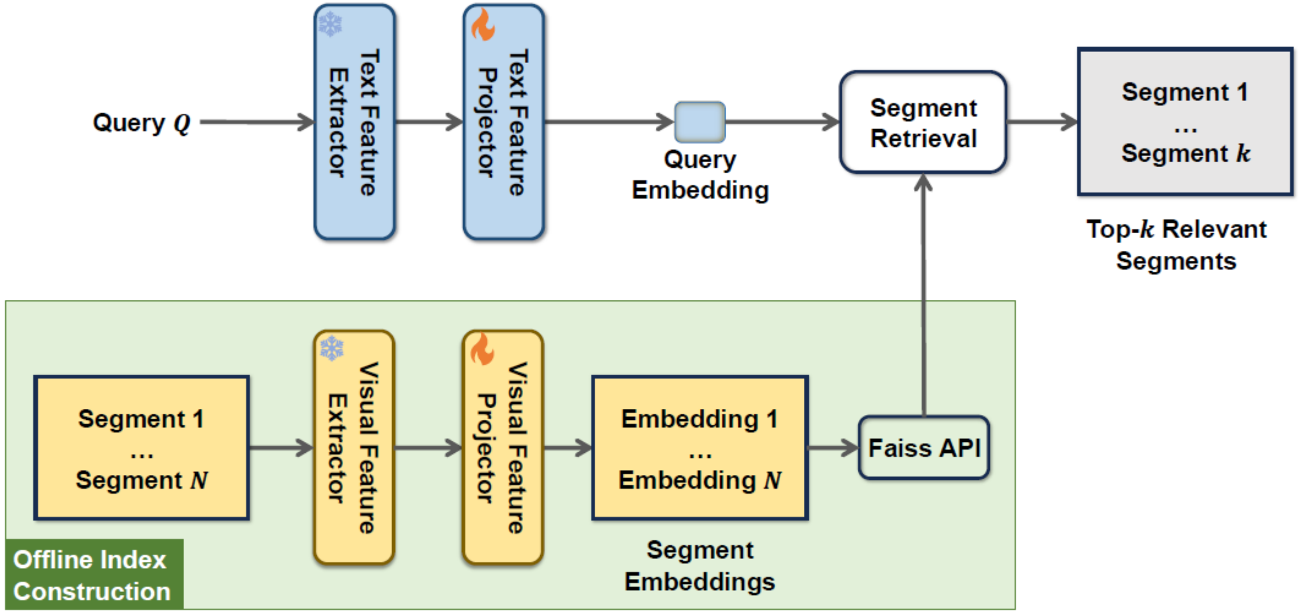


Figure 2: Segment retrieval. With the offline constructed index, the online search/inference takes less than 0.2 seconds to retrieve 100-200 relevant segments for a given query.

- Offline index for dense representations of segments
- Both query and segment features are projected to the same space
- Open to advancements in both text and video feature extraction/projection

Moment Refinement and Re-ranking

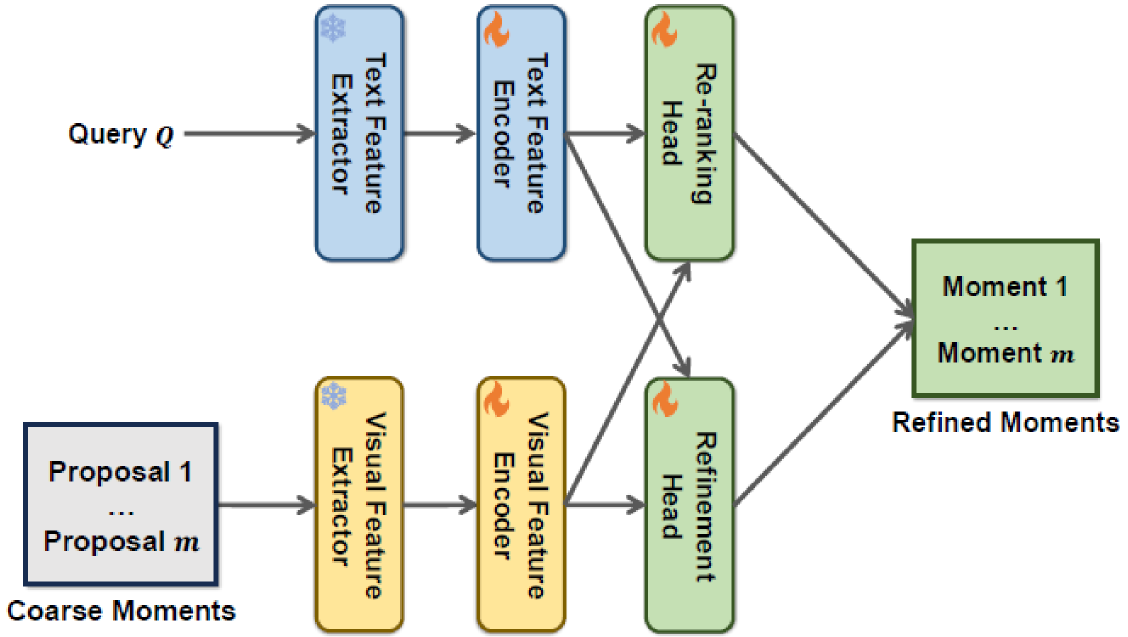
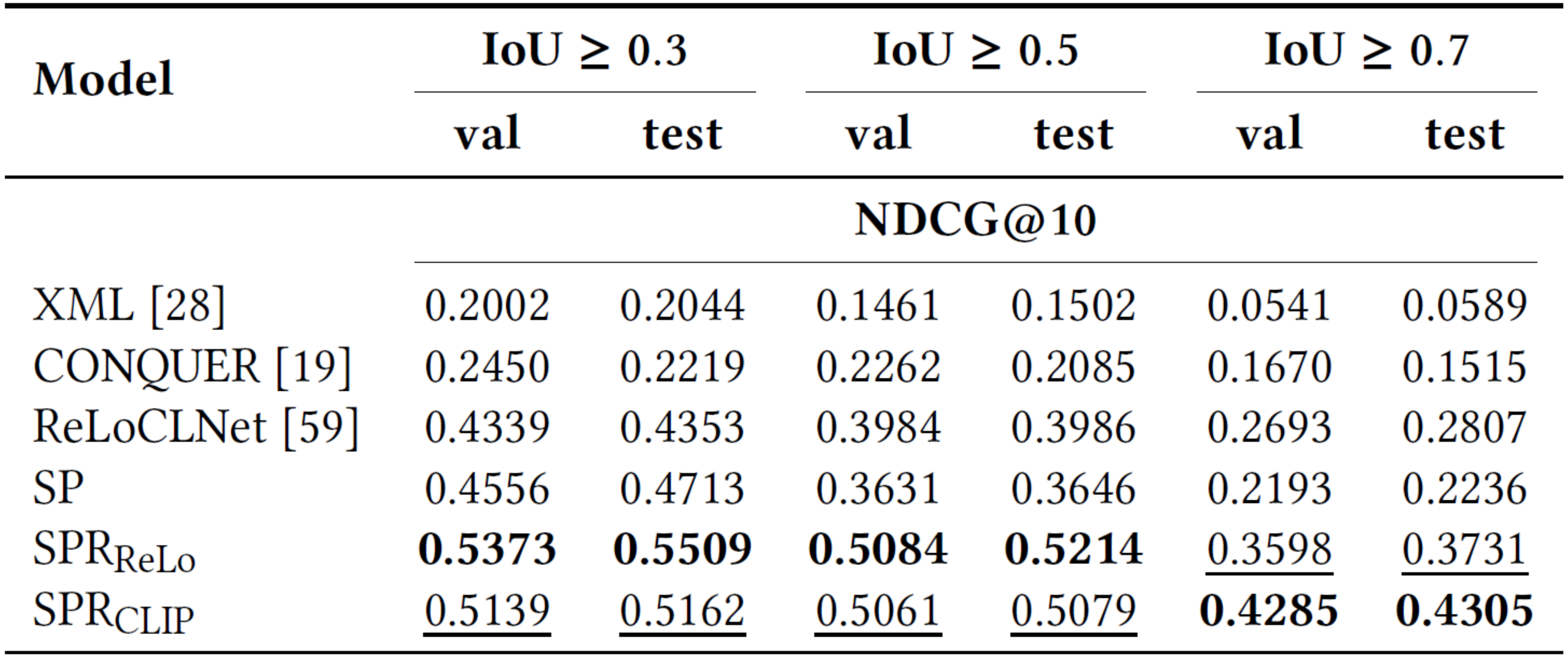


Figure 3: Refinement and re-ranking. This module computes precise timestamps of matching moments and re-ranks them by their relevance to the given query.

- Proposal generation from retrieved segments → Rule-based
- Moment refinement and re-rank
  - Given a proposal and query, get the best match→ the original NLVL task
  - But at a smaller scale: answers are limited to the small number of matching proposals

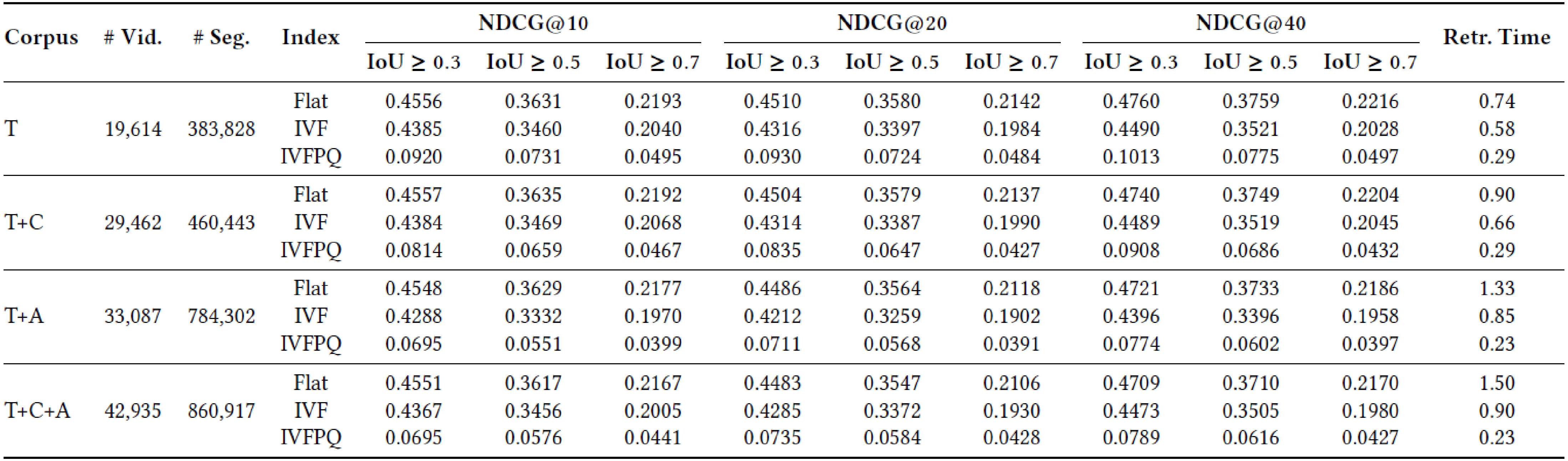
Evaluation Against Baselines



Model	IoU ≥ 0.3		IoU ≥ 0.5		IoU ≥ 0.7	
	val	test	val	test	val	test
	NDCG@10					
XML [28]	0.2002	0.2044	0.1461	0.1502	0.0541	0.0589
CONQUER [19]	0.2450	0.2219	0.2262	0.2085	0.1670	0.1515
ReLoCLNet [59]	0.4339	0.4353	0.3984	0.3986	0.2693	0.2807
SP	0.4556	0.4713	0.3631	0.3646	0.2193	0.2236
SPR_{ReLo}	**0.5373**	**0.5509**	**0.5084**	**0.5214**	0.3598	0.3731
SPR_{CLIP}	0.5139	0.5162	0.5061	0.5079	**0.4285**	**0.4305**

- SP: Segment and Proposal (without refinement and re-ranking)
- SPR: with refinement and reranking
- CLIP vs ReLo: different feature extractors

Scalability Test on TVR-Ranking Validation Dataset



Corpus	# Vid.	# Seg.	Index	NDCG@10			NDCG@20			NDCG@40			Retr. Time
				IoU ≥ 0.3	IoU ≥ 0.5	IoU ≥ 0.7	IoU ≥ 0.3	IoU ≥ 0.5	IoU ≥ 0.7	IoU ≥ 0.3	IoU ≥ 0.5	IoU ≥ 0.7	
T	19,614	383,828	Flat	0.4556	0.3631	0.2193	0.4510	0.3580	0.2142	0.4760	0.3759	0.2216	0.74
			IVF	0.4385	0.3460	0.2040	0.4316	0.3397	0.1984	0.4490	0.3521	0.2028	0.58
			IVFPQ	0.0920	0.0731	0.0495	0.0930	0.0724	0.0484	0.1013	0.0775	0.0497	0.29
T+C	29,462	460,443	Flat	0.4557	0.3635	0.2192	0.4504	0.3579	0.2137	0.4740	0.3749	0.2204	0.90
			IVF	0.4384	0.3469	0.2068	0.4314	0.3387	0.1990	0.4489	0.3519	0.2045	0.66
			IVFPQ	0.0814	0.0659	0.0467	0.0835	0.0647	0.0427	0.0908	0.0686	0.0432	0.29
T+A	33,087	784,302	Flat	0.4548	0.3629	0.2177	0.4486	0.3564	0.2118	0.4721	0.3733	0.2186	1.33
			IVF	0.4288	0.3332	0.1970	0.4212	0.3259	0.1902	0.4396	0.3396	0.1958	0.85
			IVFPQ	0.0695	0.0551	0.0399	0.0711	0.0568	0.0391	0.0774	0.0602	0.0397	0.23
T+C+A	42,935	860,917	Flat	0.4551	0.3617	0.2167	0.4483	0.3547	0.2106	0.4709	0.3710	0.2170	1.50
			IVF	0.4367	0.3456	0.2005	0.4285	0.3372	0.1930	0.4473	0.3505	0.1980	0.90
			IVFPQ	0.0695	0.0576	0.0441	0.0735	0.0584	0.0428	0.0789	0.0616	0.0427	0.23

- Adding videos from a different domain **C: Charades**, and **A: ActivityNet Captions**.
- When number of segments doubled, no much change on the accuracy, and search time increase linearly if Flat indexing is used, or sublinearly

The Bottleneck of Perfect VMR?

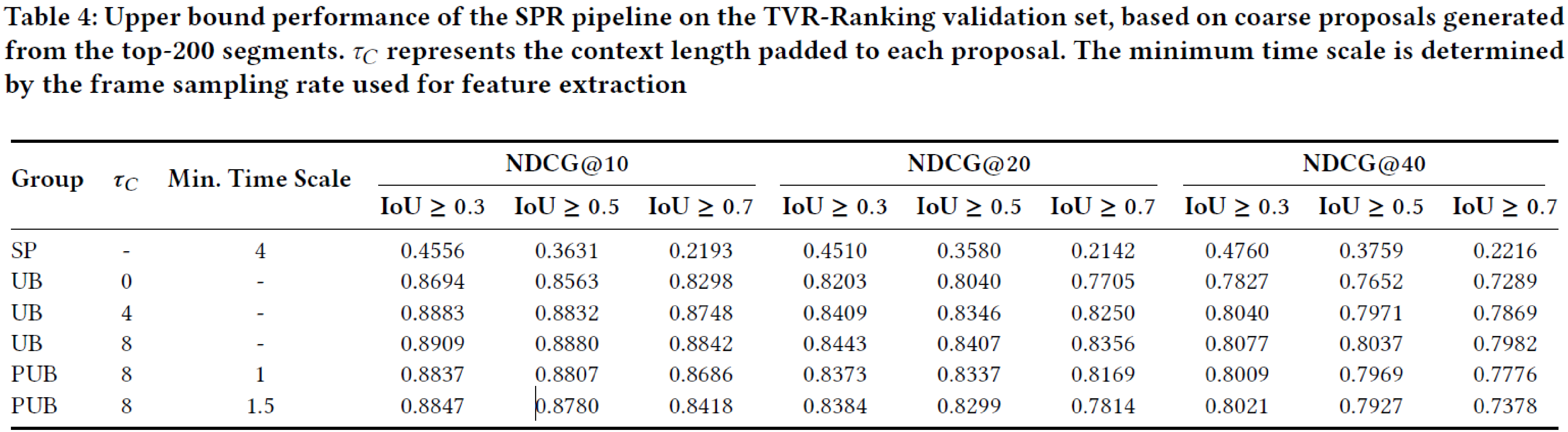


Table 4: Upper bound performance of the SPR pipeline on the TVR-Ranking validation set, based on coarse proposals generated from the top-200 segments. τ_C represents the context length padded to each proposal. The minimum time scale is determined by the frame sampling rate used for feature extraction

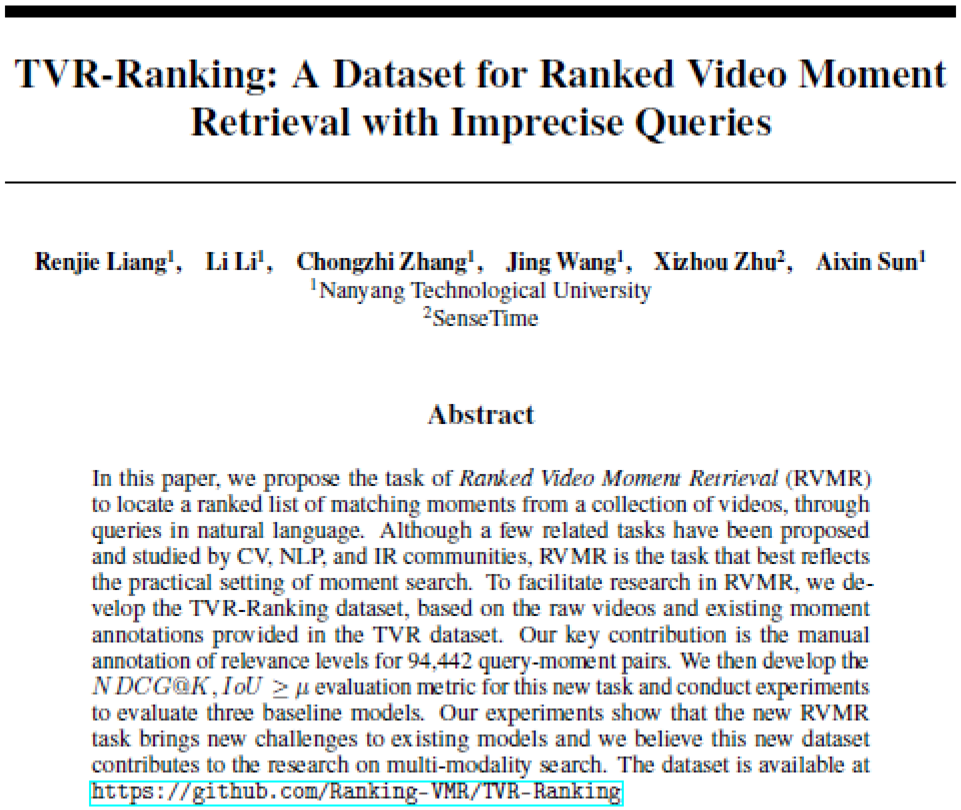
Group	τ_C	Min. Time Scale	NDCG@10			NDCG@20			NDCG@40		
			IoU $\geq$ 0.3	IoU $\geq$ 0.5	IoU $\geq$ 0.7	IoU $\geq$ 0.3	IoU $\geq$ 0.5	IoU $\geq$ 0.7	IoU $\geq$ 0.3	IoU $\geq$ 0.5	IoU $\geq$ 0.7
SP	-	4	0.4556	0.3631	0.2193	0.4510	0.3580	0.2142	0.4760	0.3759	0.2216
UB	0	-	0.8694	0.8563	0.8298	0.8203	0.8040	0.7705	0.7827	0.7652	0.7289
UB	4	-	0.8883	0.8832	0.8748	0.8409	0.8346	0.8250	0.8040	0.7971	0.7869
UB	8	-	0.8909	0.8880	0.8842	0.8443	0.8407	0.8356	0.8077	0.8037	0.7982
PUB	8	1	0.8837	0.8807	0.8686	0.8373	0.8337	0.8169	0.8009	0.7969	0.7776
PUB	8	1.5	0.8847	0.8780	0.8418	0.8384	0.8299	0.7814	0.8021	0.7927	0.7378

- Segment retrieval: Speed, and Proposal Quality
- Proposal: Set an upper bound (UB) for the refinement and re-ranking
- PUB: Practical UB due to the refinement module can only work on preset time scales, not a continuous time span

What do We Aim to Achieve and the Key Challenges?

- A more practical task definition
- A meaningful annotated dataset
- An efficient and effective moment retrieval

- Video data is limited to existing datasets
- Query data is not directly collected from real users
- Video representation remains a key challenge for semantic retrieval and understanding

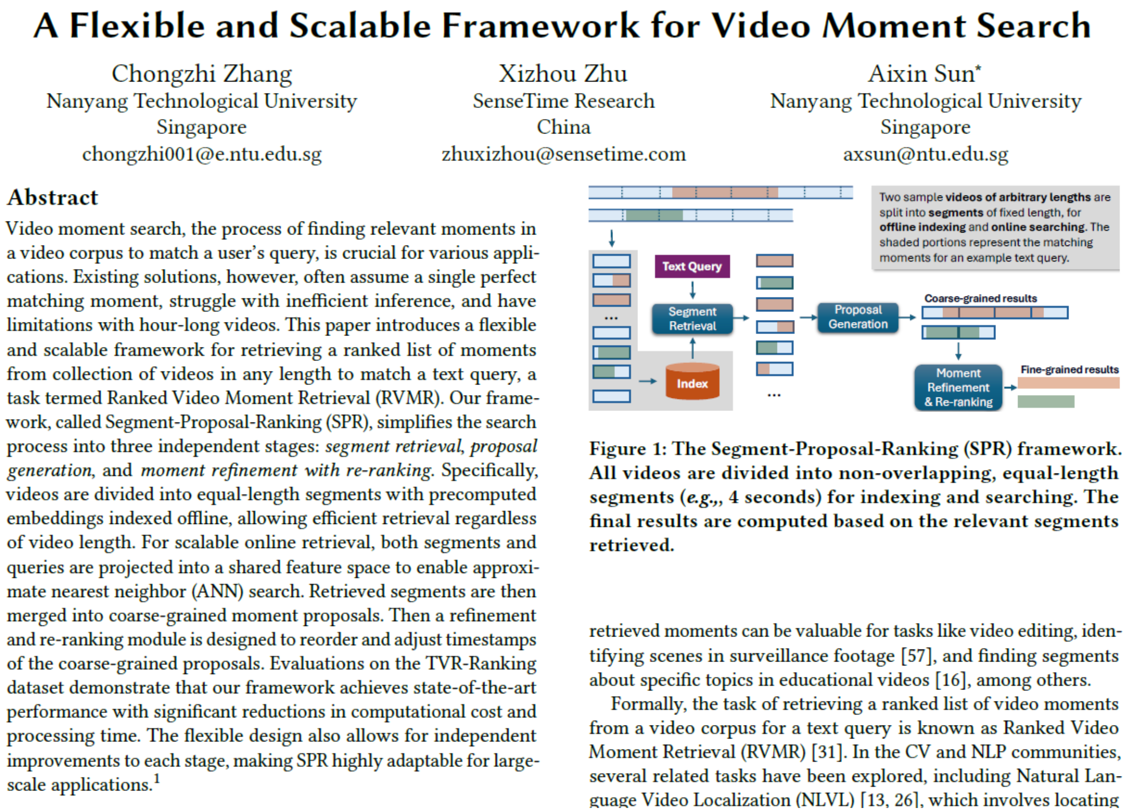


TVR-Ranking: A Dataset for Ranked Video Moment Retrieval with Imprecise Queries

Renjie Liang[1], Li Li[1], Chongzhi Zhang[1], Jing Wang[1], Xizhou Zhu[2], Aixin Sun[1]
[1]Nanyang Technological University
[2]SenseTime

Abstract

In this paper, we propose the task of *Ranked Video Moment Retrieval* (RVMR) to locate a ranked list of matching moments from a collection of videos, through queries in natural language. Although a few related tasks have been proposed and studied by CV, NLP, and IR communities, RVMR is the task that best reflects the practical setting of moment search. To facilitate research in RVMR, we develop the TVR-Ranking dataset, based on the raw videos and existing moment annotations provided in the TVR dataset. Our key contribution is the manual annotation of relevance levels for 94,442 query-moment pairs. We then develop the $NDCG@K, IoU \geq \mu$ evaluation metric for this new task and conduct experiments to evaluate three baseline models. Our experiments show that the new RVMR task brings new challenges to existing models and we believe this new dataset contributes to the research on multi-modality search. The dataset is available at `https://github.com/Ranking-VMR/TVR-Ranking`


Two sample **videos of arbitrary lengths** are split into **segments** of fixed length, for **offline indexing** and **online searching.** The shaded portions represent the matching moments for an example text query.
Text Query
Segment Retrieval
Index
...
...
Proposal Generation
Coarse-grained results
Moment Refinement & Re-ranking
Fine-grained results

A Flexible and Scalable Framework for Video Moment Search

Chongzhi Zhang
Nanyang Technological University
Singapore
chongzhi001@e.ntu.edu.sg

Xizhou Zhu
SenseTime Research
China
zhuxizhou@sensetime.com

Aixin Sun*
Nanyang Technological University
Singapore
axsun@ntu.edu.sg

Abstract

Video moment search, the process of finding relevant moments in a video corpus to match a user's query, is crucial for various applications. Existing solutions, however, often assume a single perfect matching moment, struggle with inefficient inference, and have limitations with hour-long videos. This paper introduces a flexible and scalable framework for retrieving a ranked list of moments from collection of videos in any length to match a text query, a task termed Ranked Video Moment Retrieval (RVMR). Our framework, called Segment-Proposal-Ranking (SPR), simplifies the search process into three independent stages: *segment retrieval*, *proposal generation*, and *moment refinement with re-ranking*. Specifically, videos are divided into equal-length segments with precomputed embeddings indexed offline, allowing efficient retrieval regardless of video length. For scalable online retrieval, both segments and queries are projected into a shared feature space to enable approximate nearest neighbor (ANN) search. Retrieved segments are then merged into coarse-grained moment proposals. Then a refinement and re-ranking module is designed to reorder and adjust timestamps of the coarse-grained proposals. Evaluations on the TVR-Ranking dataset demonstrate that our framework achieves state-of-the-art performance with significant reductions in computational cost and processing time. The flexible design also allows for independent improvements to each stage, making SPR highly adaptable for large-scale applications.[1]

Figure 1: The Segment-Proposal-Ranking (SPR) framework. All videos are divided into non-overlapping, equal-length segments (*e.g.,*, 4 seconds) for indexing and searching. The final results are computed based on the relevant segments retrieved.

retrieved moments can be valuable for tasks like video editing, identifying scenes in surveillance footage [57], and finding segments about specific topics in educational videos [16], among others.

Formally, the task of retrieving a ranked list of video moments from a video corpus for a text query is known as Ranked Video Moment Retrieval (RVMR) [31]. In the CV and NLP communities, several related tasks have been explored, including Natural Language Video Localization (NLVL) [13, 26], which involves locating

Thank you!

https://personal.ntu.edu.sg/axsun/