Video Moment Retrieval: A Survey of DETR-based Detection, LLM Integration, and Multi-task Learning Approaches (2020-2025)

Shuqi Wang
Beijing Normal - Hong Kong Baptist University (BNBU)
Zhuhai City, China
s230004040@mail.uic.edu.cn

Bingqing QU
Assistant Professor, Head of MA Programme (AIDM)
Beijing Normal - Hong Kong Baptist University (BNBU)
Zhuhai City, China
bingqingqu@uic.edu.cn

Abstract

With the dramatic surge in video content production, there is an increasing demand for advanced video understanding systems that enable precise moment-level content retrieval. Video Moment Retrieval (VMR), also known as temporal sentence grounding, enables users to locate specific temporal segments within videos using natural language queries [1]. This survey provides a comprehensive analysis of VMR advances from 2020 to 2025, focusing on three dominant technical paradigms: DETR-based detection frameworks that adapt object detection for end-to-end temporal localization, LLM-based methods that leverage large language models for enhanced semantic understanding, and unified multi-task frameworks that integrate multiple video understanding tasks for improved efficiency. This survey systematically categorizes and critically analyzes current methodologies, evaluates their strengths and limitations, and identifies key challenges, including long-form video understanding, robustness, and real-time processing requirements. My analysis reveals a field's trajectory toward foundation models that balance accuracy, efficiency, and generalization for practical deployment in intelligent multimedia systems.

1. Introduction

The exponential growth of video content has fundamentally transformed the digital landscape, with over 500 hours of video uploaded to YouTube every minute [2], and the global video streaming market projected to reach $2660.88 billion by 2032 [3]. This unprecedented scale of multimedia content creation has created an urgent need for intelligent video navigation systems. Users increasingly seek to locate specific moments within vast video archives, whether searching for particular cooking techniques in culinary tutorials, identifying key plays in sports broadcasts, or finding specific scenes in educational lectures. These diverse use cases highlight the critical importance of precise temporal content retrieval in our media-rich digital environment.

However, traditional keyword-based search methods, while effective for text retrieval, fail to capture the temporal dynamics and multimodal nature of video content. These approaches cannot effectively bridge the semantic gap between natural language descriptions and the rich spatiotemporal information present in video sequences, particularly when users describe visual events using nuanced natural language queries. This fundamental limitation has catalyzed the emergence of **Video Moment Retrieval**, or temporal sentence grounding [1]. This technology enables users to locate specific temporal segments within videos using natural language descriptions.

Formally, given an untrimmed video $V = \{v_1, v_2, ..., v_T\}$ consisting of T temporal frames and a natural language query Q describing a specific activity, event, or scene, the video moment retrieval task aims to predict the temporal boundaries $[t_s, t_e]$ where $1 \leq t_s < t_e \leq T$ that best corresponds to the semantic content described in the query. Mathematically, this can be formulated as an optimization problem:

$$[t_s^*, t_e^*] = \arg\max_{t_s, t_e} f(V_{t_s:t_e}, Q) \tag{1}$$

where $f(\cdot, \cdot)$ represents a cross-modal similarity function that measures the semantic alignment between the video seg-

Table 1. Comparative analysis of common VMR datasets.

Dataset	#Videos	#Queries	Domain	Avg. Length (s)
ActivityNet Captions	19,994	71,957	Open	117.6
Charades-STA	6,672	16,128	Indoor	30.6
TACoS	127	18,818	Cooking	287.1
QVHighlights	10,148	73,958	Open	150.2
YouCook2-Temporal	2,000	14,000	Cooking	316.8

ment $V_{t_s:t_e}$ and the textual query Q. The pioneering works of Gao *et al*. [1] (TALL) and Hendricks *et al*. [4] (MCN) in 2017 first formalized this problem, establishing two foundational technical paradigms: proposal-based localization through cross-modal alignment scoring and direct similarity computation between video segments and textual queries. These foundational approaches established two key technical challenges that remain central to ongoing research: developing better ways to align video content with text descriptions and improving the accuracy of temporal boundary detection.

It is important to note that VMR differs fundamentally from related video understanding tasks. Unlike video summarization, which extracts key segments without query guidance, or video question answering, which generates textual responses rather than temporal localization, VMR requires precise semantic-temporal alignment between natural language descriptions and corresponding video segments [5]. This task demands fine-grained temporal precision that distinguishes it from broader video understanding approaches, as it must simultaneously achieve semantic comprehension and accurate boundary localization within potentially hours-long video content.

This survey provides a comprehensive analysis of VMR advances from 2020 to 2025, focusing on three dominant technical paradigms that have emerged: DETR-based detection frameworks, large language model integration approaches, and unified multi-task architectures. The contributions include: (1) systematic categorization and critical analysis of current methodologies with evaluation of their strengths and limitations, and (2) identification of key challenges and promising future research directions. The remainder of this survey is organized as follows: Section 2 reviews related works including datasets, evaluation metrics, and the evolution of technical approaches, Section 3 examines current challenges and future directions, and Section 4 concludes with key insights and recommendations.

2. Related Works

The evolution of video moment retrieval from 2020-2025 has been characterized by three dominant paradigms that represent distinct philosophical approaches to the problem. **DETR-based approaches** adapt object detection frameworks for end-to-end temporal localization, **LLM-based methods** leverage large language models for enhanced semantic understanding, and **unified multi-task frameworks** integrate multiple video understanding tasks for improved efficiency. Each paradigm addresses different aspects of the fundamental challenges in video-text alignment and temporal boundary detection.

Before examining these paradigms, we first review the datasets and evaluation metrics that have shaped VMR research, as they provide essential context for understanding the development and assessment of different approaches.

2.1. Datasets and Evaluation metrics

The development of VMR methods has been significantly influenced by the availability of standardized datasets and evaluation metrics. Several benchmark datasets have become the cornerstone for evaluating VMR approaches, each with distinct characteristics addressing different aspects of the task.

Datasets: The field has relied on several key datasets that vary in scale, domain, and complexity. ActivityNet Captions [6] contains approximately 20,000 videos with 100,000 sentence descriptions, each aligned with specific video segments. Charades-STA [1] was specifically designed for moment retrieval, containing 16,128 temporal annotations of 6,672 videos depicting daily indoor activities. TACoS [7] offers fine-grained annotations of cooking activities with multiple descriptions for the same video segment, emphasizing sequential actions. More recently, QVHighlights [5] introduced a dual-annotation approach that provides both moment boundaries and highlight scores, enabling multi-task evaluation. YouCook2-Temporal [8] focuses specifically on cooking procedures with step-by-step instructions temporally aligned with video segments.

Table 1 presents a comparative analysis of these key VMR datasets, revealing the diversity in domain focus and temporal scope that researchers must address when developing robust retrieval methods.

Evaluation Metrics: VMR performance is primarily evaluated using two key metrics based on temporal Intersection over Union (tIoU):

$$\text{tIoU} = \frac{|P \cap G|}{|P \cup G|} = \frac{\min(P_e, G_e) - \max(P_s, G_s)}{\max(P_e, G_e) - \min(P_s, G_s)} \quad (2)$$

where $P = [P_s, P_e]$ represents the predicted temporal boundary and $G = [G_s, G_e]$ represents the ground-truth temporal boundary.

- **R@1,IoU**: The percentage of queries where the top-1 retrieved moment has a tIoU with ground truth above a threshold (typically 0.3, 0.5, or 0.7):

$$\text{R@1,IoU=m} = \frac{1}{N} \sum_{i=1}^{N} 1(\text{tIoU}(P_1^i, G^i) > m) \quad (3)$$

where N is the total number of queries and $1(\cdot)$ is the indicator function.

- **mAP (mean Average Precision)**: Evaluates ranking quality by measuring how well the model ranks relevant moments:

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|\mathcal{R}^i|} \sum_{k=1}^{K} P(k) \cdot \text{rel}(k) \quad (4)$$

where $P(k)$ is precision at rank k, $\text{rel}(k)$ indicates whether the k-th result is relevant, and $|\mathcal{R}^i|$ is the number of relevant moments for query i.

R@1,IoU focuses on whether the model can accurately localize the most relevant moment, while mAP assesses how well the model ranks all potentially relevant moments in order of relevance. Together, this dual evaluation ensures models excel at both precise localization and effective retrieval ranking.

2.2. DETR-based Approaches

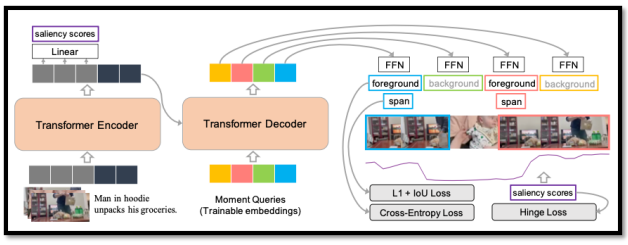

Figure 1. Moment-DETR model overview [5]. The architecture employs a transformer encoder-decoder with three prediction heads for predicting saliency scores, foreground/background scores, and moment coordinates. The video and text feature extractors process multimodal inputs for cross-modal alignment.

DETR-based approaches reconceptualize video moment retrieval as an end-to-end detection problem. These methods adapt the successful Detection Transformer architecture from object detection to temporal localization. This represents a significant shift from traditional proposal-based methods to direct set prediction. This shift eliminates hand-crafted post-processing steps. As shown in Figure 1, the approach enables simultaneous moment detection and boundary regression through learnable query embeddings, treating temporal segments as detection targets. Moment-DETR [5] pioneered this direction with a transformer encoder-decoder architecture featuring specialized prediction heads for saliency scoring, foreground classification, and moment coordinate regression. The end-to-end framework offers clear advantages by jointly optimizing semantic understanding and temporal localization, while set-based prediction naturally handles variable-length outputs and multiple moment detection.

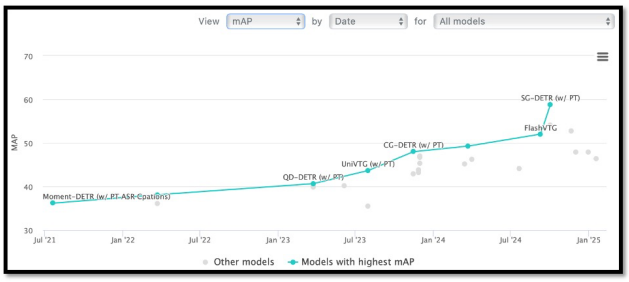

Figure 2. mAP Performance comparison on QVHighlights leaderboard from Papers with Code, showing DETR-based methods dominating the top rankings for video temporal grounding tasks.

Subsequent DETR-based methods have systematically addressed key limitations through progressive innovations targeting cross-modal alignment. QD-DETR [9] introduced a query-dependent video representation that dynamically adapts visual features based on textual queries through cross-attention mechanisms, fundamentally changing how visual features are contextualized. BAM-DETR [10] tackled boundary imprecision by incorporating multi-scale boundary-aware attention that leverages temporal context at multiple resolutions with explicit boundary regression heads. CG-DETR [11] advanced the field through correlation-guided query calibration, addressing the semantic gap between video content and query representation via adaptive correlation modeling. Recent innovations include LA-DETR [12], which introduced length-aware mechanisms for handling diverse moment durations, and SG-DETR [13], which incorporated saliency guidance to achieve 58.80 mAP on QVHighlights through joint saliency-moment modeling. As evident from the Papers with Code leaderboard [0] shown in Figure 2, DETR-based approaches now dominate the top performance rankings, with these progressive innovations collectively driving substantial improvements and demonstrating the paradigm's continued potential for advancing video temporal grounding.

[0] https://paperswithcode.com/sota/moment-retrieval-on-qvhighlights

2.3. LLM-based Approaches

LLM-based approaches represent a paradigm shift from pattern recognition to semantic understanding. It leverages large language models to bridge the semantic gap in video-text alignment through sophisticated reasoning [14]. This transition reflects a movement toward methods that handle the complexities of natural language understanding, particularly temporal relationships, causal reasoning, and abstract concept grounding. Unlike traditional approaches relying on visual feature matching, LLM-based systems harness pre-trained linguistic knowledge to interpret complex queries and establish sophisticated video-text correspondences. As shown in Figure 3, the Chrono framework exemplifies this through its model-agnostic design that adapts pre-trained multimodal language models for temporal localization. By interleaving frame embeddings with timestamps and constructing comprehensive prompts, these systems demonstrate how language models' understanding of temporal concepts and contextual nuances enhances moment retrieval accuracy, especially for queries involving abstract concepts.

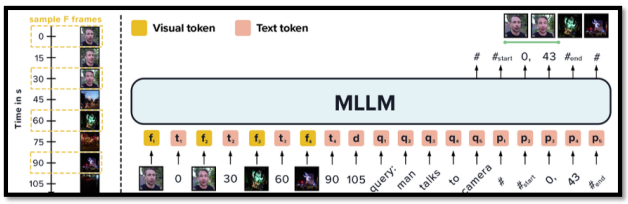

Figure 3. Chrono model overview [14]: this system interleaves frame embeddings with timestamps, followed by video duration and moment retrieval query. The MLLM outputs retrieved moments by predicting global BOS/EOS tokens and respective start/end times for each temporal window.

The evolution follows a clear progression from enhanced text understanding to sophisticated video-text reasoning, with breakthroughs in multimodal fusion and temporal reasoning. Early implementations focused on improving textual query comprehension through pre-trained language encoders, achieving meaningful gains in semantic alignment. Video-ChatGPT [15] marked a breakthrough by introducing conversational reasoning for complex temporal queries. LLaVA-MR [16] advanced specialized training by fine-tuning multimodal language models for temporal localization, achieving 52.73 mAP and **76.59 R@1, IoU=0.5** through instruction tuning. Additionally, LLMEPET [17] introduced knowledge enhancement through pseudo-event generation, leveraging LLMs to create additional training signals. Video-Mamba-Suite [18] explored efficient architectures through state space models, addressing computational challenges while maintaining sophisticated understanding capabilities. This evolution demonstrates a transition from simple text processing to sophisticated multimodal reasoning, with gains particularly pronounced for complex, semantically rich queries.

2.4. Unified Multi-task Frameworks

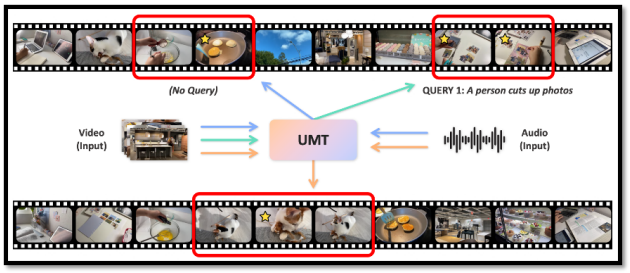

Figure 4. UMT unified framework architecture [19]. The system handles different input modality combinations and outputs both video moment retrieval (marked by red rectangles) and highlight detection results (marked by golden stars). Different colored arrows denote various input-output combinations, demonstrating the framework's flexibility.

Unified multi-task frameworks emerge from recognizing that video moment retrieval and related tasks like highlight detection share complementary objectives and benefit from joint optimization through shared representations. This paradigm responds to computational complexities of deploying multiple specialized models while addressing the insight that temporal localization and saliency detection are inherently correlated. As shown in Figure 4, the UMT framework demonstrates this through its flexible architecture, handling different input modalities while producing both moment retrieval and highlight detection results simultaneously. The approach leverages the observation that temporal localization and saliency detection provide mutual supervision signals, where highlighted information guides moment boundary detection through attention reweighting, and precise temporal understanding enhances highlight quality assessment. This synergy enables efficient parameter utilization and improved performance compared to independent models.

The architectural evolution has progressed from simple multi-task learning to comprehensive video understanding systems integrating multiple levels of temporal reasoning. UMT [19] established the foundational approach through shared transformer encoders and task-specific decoders, demonstrating efficiency improvements while maintaining competitive performance through joint attention mechanisms enabling cross-task knowledge transfer. UniVTG [20] advanced unified video-language temporal grounding by integrating moment retrieval, highlight detection, and video summarization, showcasing superior transfer learning through unified temporal representations. UVCOM [21] broadened the scope to include creation capabilities, incorporating video generation through shared video-language encoders, enabling bidirectional understanding. InternVideo2 [22] represents the culmination as a large-scale foundation model

achieving state-of-the-art performance across benchmarks through progressive multi-task training and unified temporal-spatial-semantic representations. This trajectory illustrates the field's movement toward holistic video intelligence systems that efficiently handle multiple related tasks.

2.5. Comparative Analysis and Method Selection

The choice between these paradigms depends critically on specific application requirements and constraints, as demonstrated in Table 2. DETR-based approaches excel in scenarios requiring precise temporal boundary detection with superior localization accuracy, making them ideal for video editing workflows where boundary precision is paramount. LLM-based methods demonstrate clear superiority for semantically complex queries involving abstract concepts and temporal reasoning, achieving impressive performance on reasoning-heavy benchmarks, but suffer from significant computational overhead and longer inference times. Unified frameworks provide optimal resource efficiency with substantial parameter reduction compared to specialized models, making them ideal for industrial deployment and real-time applications where multiple tasks must be performed simultaneously, though they may underperform on individual tasks due to the classic accuracy-efficiency trade-off. This review paper believes that selection should be guided by specific needs: DETR-based methods for precision-critical scenarios, LLM-based approaches for semantic complexity, and unified frameworks for deployment constraints prioritizing efficiency and comprehensive capability.

3. Future Research Directions

The field of video moment retrieval faces several critical challenges that will shape future research directions. **Long-form video understanding** emerges as a primary frontier, as current methods struggle with hour-long content where temporal dependencies span extended sequences and computational complexity scales unfavorably. While recent works like CONE [23] and TimeLoc [24] have begun addressing this through hierarchical processing and efficient attention mechanisms, significant advances in memory-efficient architectures remain essential for practical deployment.

Robustness and generalization present equally important challenges, particularly addressing dataset biases that limit real-world performance and improving cross-domain transfer capabilities. Current DETR-based methods, while achieving impressive performance on standard benchmarks, often struggle when applied to domains significantly different from their training data. The emerging challenge of **negative query handling** [25] - determining when requested moments don't exist in videos - requires developing uncertainty quantification and rejection mechanisms that current end-to-end frameworks lack.

Lastly, **Efficiency versus capability trade-offs** remain central to practical deployment. While LLM-based approaches demonstrate superior semantic understanding, their computational overhead limits real-time applications. Conversely, unified frameworks offer resource efficiency but may sacrifice individual task performance. Future work must bridge this gap through architectural innovations that maintain LLM-level semantic understanding while achieving DETR-level efficiency.

4. Conclusion

This survey has examined the rapid evolution of video moment retrieval from 2020 to 2025, highlighting the emergence of three dominant paradigms that have fundamentally transformed both the theoretical understanding and practical capabilities of the field. DETR-based approaches have established new standards for precise temporal localization through end-to-end learning, achieving substantial performance improvements while eliminating the complexity of traditional multi-stage processing pipelines. Concurrently, LLM-based methods have revolutionized semantic understanding capabilities, particularly excelling in complex reasoning scenarios that require deep contextual comprehension, though this advancement comes with increased computational demands. Meanwhile, unified multi-task frameworks have demonstrated the compelling practical benefits of joint optimization, achieving significant resource efficiency while maintaining competitive performance across multiple related tasks.

Based on my analysis, the field's trajectory indicates a convergence toward foundation models that must carefully balance accuracy, efficiency, and generalization for real-world deployment. As video content continues to proliferate exponentially across digital platforms, these technological advances position VMR as a critical enabling technology for the next generation of intelligent multimedia systems. Success depends on solving these technical challenges that are fundamental to practical applications.

References

[1] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "Tall: Temporal activity localization via language query," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5277–5285, 2017. doi: 10.1109/ICCV.2017.563. (document), 1, 1, 2.1

[2] Statista, "Youtube: hours of video uploaded every minute 2022," 2022. 1

[3] Statista, "Video streaming worldwide - statistics & facts," 2024. 1

[4] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, "Localizing moments in video with natural language," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5803–5812, 2017. doi: 10.1109/ICCV.2017.618. 1

[5] J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, and J. Liu, "Qvhighlights: Detecting moments and highlights in videos via natural language queries," in *Advances in Neural Information Processing Systems*, vol. 34, pp. 5537–5549, 2021. doi: 10.48550/arXiv.2107.09609. 1, 2.1, 1, 2.2

[6] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 706–715, 2017. doi: 10.1109/ICCV.2017.83. 2.1

[7] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, "Grounding action descriptions in videos," *Transactions of the Association for Computational Linguistics*, vol. 1, pp. 25–36, 2013. doi: 10.1162/tacl$_{\alpha 0}$0207. 2.1

[8] L. Zhou, C. Xu, and J. J. Corso, "Towards automatic learning of procedures from web instructional videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, pp. 7590–7598, 2018. doi: 10.48550/arXiv.1703.09788. 2.1

[9] W. Moon, S. Hyun, S. Park, D. Park, and J.-P. Cho, "Query-dependent video representation for moment retrieval and highlight detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23023–23033, 2023. doi: 10.48550/arXiv.2303.13874. 2.2

[10] P. Lee and H. Byun, "Bam-detr: Boundary-aligned moment detection transformer for temporal sentence grounding in videos," *arXiv preprint arXiv:2312.00083*, 2023. doi: 10.48550/arXiv.2312.00083. 2.2

[11] W. Moon, S. Hyun, S. Lee, and J.-P. Heo, "Correlation-guided query-dependency calibration for video temporal grounding," *arXiv preprint arXiv:2311.08835*, 2023. doi: 10.48550/arXiv.2311.08835. 2.2

[12] S. Park, J. Choi, K. Baek, and H. Shim, "Length-aware detr for robust moment retrieval," *arXiv preprint arXiv:2412.20816*, 2024. doi: 10.48550/arXiv.2412.20816. 2.2

[13] A. Gordeev, V. Dokholyan, I. Tolstykh, and M. Kuprashevich, "Saliency-guided detr for moment retrieval and highlight detection," *arXiv preprint arXiv:2410.01615*, 2024. doi: 10.48550/arXiv.2410.01615. 2.2

[14] B. Meinardus, H. Rodriguez, A. Batra, A. Rohrbach, and M. Rohrbach, "Chrono: A simple blueprint for representing time in mllms," *arXiv preprint arXiv:2406.18113*, 2024. doi: 10.48550/arXiv.2406.18113. 2.3, 3

[15] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, "Video-chatgpt: Towards detailed video understanding via large vision and language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. doi: 10.48550/arXiv.2306.05424. 2.3

[16] W. Lu, J. Li, A. Yu, M.-C. Chang, S. Ji, and M. Xia, "Llava-mr: Large language-and-vision assistant for video moment retrieval," *arXiv preprint arXiv:2411.14505*, 2025. doi: 10.48550/arXiv.2411.14505. 2.3

[17] Y. Jiang, W. Zhang, X. Zhang, X.-Y. Wei, C. W. Chen, and Q. Li, "Prior knowledge integration via llm encoding and pseudo event regulation for video moment retrieval," *arXiv preprint arXiv:2407.15051*, 2024. doi: 10.48550/arXiv.2407.15051. 2.3

[18] G. Liu, H. Zhao, B. Fan, Y. Liu, Y. Zhong, L. Xu, J. Liu, and H. Jiang, "Video-mamba-suite: State space model as a versatile alternative for video understanding," *arXiv preprint arXiv:2403.09626*, 2024. doi: 10.48550/arXiv.2403.09626. 2.3

[19] Y. Liu, S. Li, Y. Wu, C. W. Chen, Y. Shan, and X. Qie, "Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3042–3051, 2022. doi: 10.48550/arXiv.2203.12745. 4, 2.4

[20] K. Q. Lin, A. J. Wang, M. Soldan, M. Wray, R. Yan, E. Zhong, D. Wang, P. Torr, and G. Bertasius, "Univtg: Towards unified video-language temporal grounding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2794–2804, 2023. doi: 10.1109/ICCV51070.2023.00262. 2.4

[21] Y. Xiao, Z. Luo, Y. Liu, Y. Ma, H. Bian, Y. Ji, Y. Yang, and X. Li, "Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 1024–1032, 2023. doi: 10.48550/arXiv.2311.16464. 2.4

[22] Y. Wang, K. Li, Y. Li, Y. He, B. Huang, Z. Zhao, H. Zhang, J. Xu, Y. Liu, Z. Wang, S. Xing, G. Chen, J. Pan, J. Yu, Y. Wang, L. Wang, and Y. Qiao, "Internvideo2: Scaling foundation models for multimodal video understanding," *arXiv preprint arXiv:2403.15377*, 2024. doi: 10.48550/arXiv.2403.15377. 2.4

[23] Z. Hou, W. Zhong, L. Ji, D. Gao, K. Yan, W.-K. Chan, C.-W. Ngo, Z. Shou, and N. Duan, "Cone: An efficient coarse-to-fine alignment framework for long video temporal grounding," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023. doi: 10.48550/arXiv.2209.10918. 3

[24] C.-L. Zhang, L. Sui, S. Liu, F. Mu, Z. Wang, and B. Ghanem, "Timeloc: A unified end-to-end framework for precise timestamp localization in long videos," *arXiv preprint arXiv:2503.06526*, 2025. doi: 10.48550/arXiv.2503.06526. 3

[25] K. Flanagan, D. Damen, and M. Wray, "Moment of untruth: Dealing with negative queries in video moment retrieval," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025. doi: arXiv:2502.08544. 3

5. Appendix

Table 2 presents a comprehensive performance comparison of recent VMR methods on the QVHighlights dataset, showcasing the evolution and effectiveness of different paradigms from 2021 to 2024. All performance metrics are sourced from Papers with Code QVHighlights leaderboard [1].

[1] https://paperswithcode.com/sota/moment-retrieval-on-qvhighlights

Table 2. Performance comparison of recent VMR methods on QVHighlights dataset.

Model	Paradigm	Year	mAP ↑	R@1,IoU=0.5 ↑	R@1,IoU=0.7 ↑	mAP@0.5 ↑	mAP@0.75 ↑
SG-DETR (w/ PT)	DETR-based	2024	**58.80**	74.20	60.40	**76.20**	**60.80**
SG-DETR	DETR-based	2024	54.10	72.20	56.60	73.20	55.80
LLaVA-MR	LLM-based	2024	52.73	**76.59**	**61.48**	69.41	54.40
FlashVTG	DETR-based	2024	52.00	70.69	53.96	72.33	53.85
InternVideo2-6B	Unified	2024	49.24	71.42	56.45	–	–
CG-DETR (w/ PT)	DETR-based	2023	47.97	68.48	53.11	69.40	49.12
VideoLights-B-pt	Unified	2024	47.94	70.36	55.25	69.53	49.17
LA-DETR	DETR-based	2024	47.93	63.94	51.10	65.65	49.44
BAM-DETR (w/ audio)	DETR-based	2023	46.91	64.07	48.12	65.61	47.51
BAM-DETR (w/ PT ASR)	DETR-based	2023	46.67	63.88	47.92	66.33	48.22
LD-DETR	DETR-based	2025	46.41	66.80	51.04	67.61	46.99
R²-Tuning	Transfer	2024	46.17	68.03	49.35	69.04	47.56
BAM-DETR	DETR-based	2023	45.36	62.71	48.64	64.57	46.33
Video-Mamba-Suite	Mamba-based	2024	45.18	66.65	52.19	64.37	46.68
LLMEPET	LLM-based	2024	44.05	66.73	49.94	65.76	43.91
UVCOM (w/ PT ASR)	Unified	2023	43.80	64.53	48.31	64.78	43.65
UniVTG (w/ PT)	Unified	2023	43.63	65.43	50.06	64.06	45.02
UVCOM	Unified	2023	43.18	63.55	47.47	63.37	42.67
CG-DETR	DETR-based	2023	42.86	65.43	48.38	64.51	42.77
QD-DETR (w/ PT)	DETR-based	2023	40.62	64.10	46.10	64.30	40.50
QD-DETR (w/ audio)	DETR-based	2023	40.19	63.06	45.10	63.04	40.10
BM-DETR	DETR-based	2023	40.08	60.12	43.05	63.08	40.18
QD-DETR (w/ ASR)	DETR-based	2023	40.00	63.20	45.20	63.40	40.40
QD-DETR	DETR-based	2023	39.86	62.40	44.98	62.52	39.88
UMT (w/ audio + ASR)	Unified	2022	38.08	–	–	–	–
Moment-DETR (w/ ASR)	DETR-based	2021	36.14	59.78	40.33	60.51	35.36
UMT	Unified	2022	36.12	–	–	–	–
UniVTG	Unified	2023	35.47	58.86	40.86	57.60	35.59
SeViLA-Localizer	Retrieval	2024	32.30	54.50	36.50	–	–

PT: Pre-training; ASR: Automatic Speech Recognition;
Bold values indicate the best performance in each metric;
"–" indicates results not reported in the original paper.