Supplementary Materials for "Anchor-Aware Similarity Cohesion in Target Frames Enables Predicting Temporal Moment Boundaries in 2D"

Jiawei Tan[1], Hongxing Wang[1,2]*, Junwu Weng[3], Jiaxin Li[1], Zhilong Ou[1], Kang Dang[4]

[1]School of Big Data and Software Engineering, Chongqing University, China

[2]Key Laboratory of Dependable Service Computing in Cyber Physical Society (Chongqing University), Ministry of Education, China

[3]ByteDance Intelligent Creation

[4]School of AI and Advanced Computing, XJTLU Entrepreneur College (Taicang), Xi'an Jiaotong-Liverpool University, Suzhou, China

{jwtan, ihxwang}@cqu.edu.cn, we0001wu@e.ntu.edu.sg, jiaxin_li@cqu.edu.cn, zlou@stu.edu.cn, Kang.Dang@xjtlu.edu.cn

We supply more studies of our proposed method for video moment retrieval (VMR) on the QVHighlights *val* dataset. Experimental evaluations depend on the challenging R1@.7 and mAP@Avg metrics.

1. Additional Quantitative Analysis

Ablation of the anchor-aware mask generated by query-embedded visual features. Table S1 shows that using only the original visual features for mask generation causes R1@.7 to drop from 64.8% to 60.8%, and mAP@avg to drop from 63.0% to 60.6%, highlighting the benefits of the query-embedded visual features.

Performance of different VLMs as backbones. Table S2 provides the results of different VLMs as backbones. The more advanced VLMs like InternVideo2-1B[12] are used as backbones, the better our method performs.

Statistical evidence about "The most relevant one almost always appears in the ground-truth". A frame exhibits high-similarity with the query if its CLIP similarity to the query ranks in the top 10%. Only 37.6% of the ground-truth (GT) frames meet this criterion, confirming that many GT frames lack high similarity. But the probability that the most matched frame to the query is within the GT is as high as 86.0%, confirming that the most relevant frame almost always appear in the GT.

Generalization of the F^{2}SGD module. Table S3 provides the R1@.7 results on F^{2}SGD. The first row shows the results of training and test on non-overlapping splits of the same dataset. The second row shows the results of training on ActivityNet Captions[3] (ACT) but test on QVHighlights[4] (QVH) and Charades-STA[1] (CHA). Due to the gap between different data distributions, cross-dataset zero-shot test is not better than in-dataset test. As a common challenge also faced by others[6, 10], further improvement is our shared goal.

	R1@.7	mAP@avg
original visual features	60.8	60.6
query-embedded visual features	64.8	63.0

Table S1. Impact of using original visual features or query-embedded visual features to generate the anchor-aware mask.

Backbone	R1@.7	mAP@avg
CLIP[9]	64.8	63.0
BLIP-2[5]	66.1	65.6
InternVideo2-1B[12]	68.2	66.7

Table S2. Impact of different VLM backbones.

	ACT	QVH	CHA
In-Dataset Test	24.7	39.6	27.7
Zero-Shot Test	-	27.9	18.2

Table S3. Results of generalizing the F^{2}SGD module trained on ActivityNet Captions (ACT) to QVHighlights (QVH) and Charades-STA (CHA).

Metrics for samples with low inter-frame similarity within a matching segment. We define segments with an average inter-frame similarity below 0.75 as highly distinct visual cases. Our method achieves 45.1% mAP@avg on such samples, surpassing the previous SOTA[13] at 21.9%. In addition, Fig. S1 illustrates our approach's ability to handle diverse visual representations (backgrounds) with similar texture descriptions ("tours a Guinness museum").

Computation cost analysis. Under identical hardware conditions, Fig. S2 compares the number of training parameters (M) and inference throughput measured in samples per sec-

*Corresponding author: Hongxing Wang.

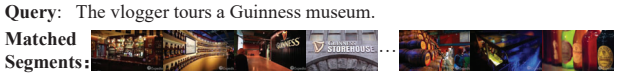


Figure S1. An example of low similarity between frames in a matching segment.

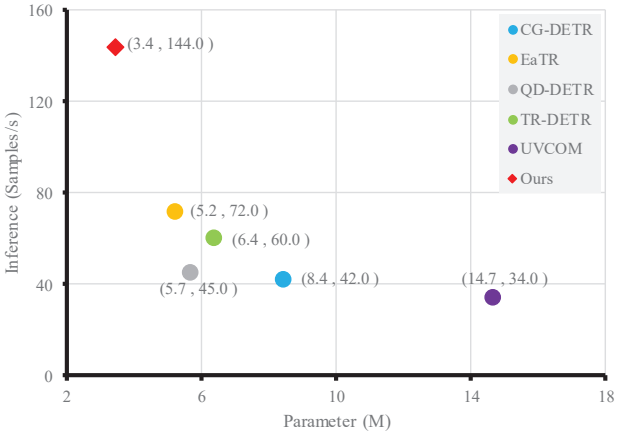


Figure S2. Comparison with the SOTA methods, CG-DETR [7], EaTR [2], QD-DETR (BR) [8], TR-DETR [11], and UVCOM [13], in terms of training parameters (M) and inference throughput measured in samples per second (Samples/s).

Reweight w/ Mask	Align w/ Mask	R1@.7	mAP@Avg
-	-	34.4	37.1
✓	-	38.2	41.8
-	✓	58.2	57.0
✓	✓	64.8	63.0

Table S4. Impact of w/o and w/ Mask on "Reweight" and "Align" stages of A^{2}FA. ✓ signifies "included" mask, while - "excluded".

ond (Samples/s) across cutting-edge methods [2, 7, 8, 11, 13]. For a fair comparison, all the involved methods utilize CLIP [9] as the encoding backbone. As can be seen, our method outperforms others by achieving faster inference speed with fewer model parameters.

Impact of anchor-aware mask at different stages of A^{2}FA. Table S4 shows the effect on VMR of incorporating the anchor-aware mask (generated by Eq. (6) as given in the main manuscript) into "reweight" and "align" stages in A^{2}FA. Without masks in both "reweight" and "align", the model achieves only 34.4% in R1@0.7 and 37.1% in mAP@Avg. Adding masks to either "reweight" or "align" improves performance, with "align" benefiting more from the mask usage. This is because the mask in the "reweight" stage focuses on evaluating frame importance, while in the "align" stage, it aligns query-related frames with the query, highlighting its critical role in feature alignment for VMR. In the case that masking is used in both stages, we achieve the best performance, demonstrating the complementary nature of "reweight" and "align".

The ratio of the result of formula (5) to the result of formula (15). We count the number of cases where the results of Eq. (5) completely cover the results of Eq. (15), which account for 91.6% of the total. This indicates that the results of Eq. (15) are nearly entirely encompassed by those of Eq. (5).

2. Additional Visualizations

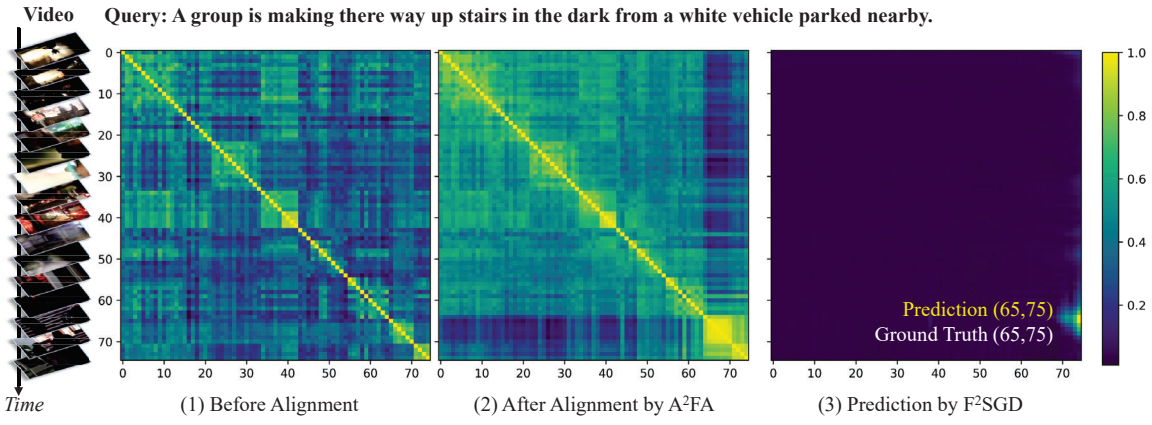
Evolution of the 2D similarity space at different stages. We present in Fig. S3 the evolution of the 2D similarity space at different stages of our method: (1) similarity map between frame features before alignment, (2) similarity map between frame features after alignment by our proposed A^{2}FA, and (3) prediction map made by our proposed F^{2}SGD. To illustrate various scenarios, we include cases where a query corresponds to either a single matching interval or multiple matching intervals. From the figure, we can see that alignment by our A^{2}FA significantly enhances query-related frame-frame similarity. This makes our subsequent F^{2}SGD easy to highlight the upper-right corners of high-similarity blocks, enabling the precise detection of single or multiple matching interval boundaries.

Query relevance scores r generated by Eq. (9). Based on Eq. (9) in the main manuscript, Fig. S4 visualizes the query relevance scores across different video scenarios. Results show that query-relevant frames within the semantically related interval (SRI) $[b^{\mathrm{L}}, b^{\mathrm{R}}]$ (Eq. (5) in the main manuscript) are assigned high query relevance scores, while query-irrelevant frames receive low scores. This confirms that the "Reweight" stage in our A^{2}FA can effectively reduce the importance of query-irrelevant frames in the SRI.

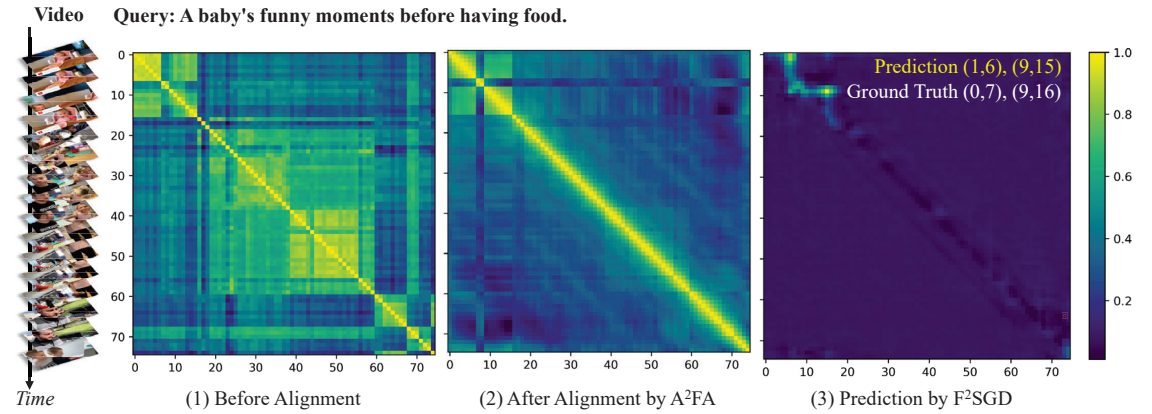
Various visualizations of retrieval results. In Fig. S5, we present various VMR examples. Besides results of our approach, we also provide ground-truth annotations for reference and include the previous state-of the-art methods, TaskWeave [14] and UVCOM [13], for comparison. For diverse query-video pairs, our approach exhibits its superiority, which consistently localizes the queried moments with higher overlap with ground-truth intervals in comparison with strong competitors.

3. Limitation

Although our proposed method excels in retrieving video moments, it occasionally fails to accurately locate some query-video pairs due to the exclusion of additional factors, such as action and audio. This observation encourages us to extend the current approach to handling multimodality scenarios in future works.



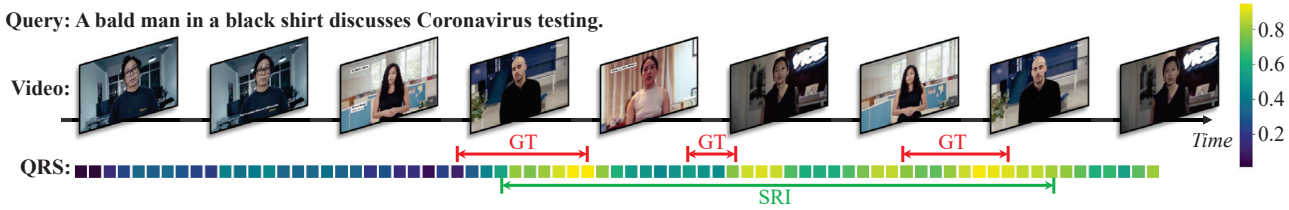
(a) A query text corresponds to one matching interval.



(b) A query text corresponds to multiple matching intervals.

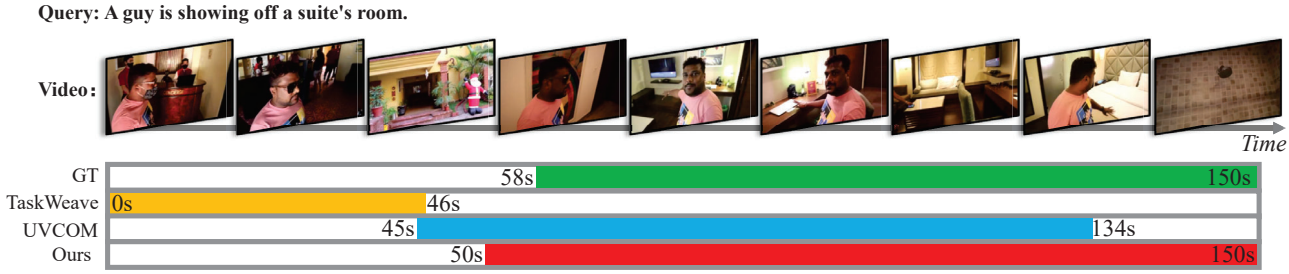
Figure S3. Intermediate results at different stages, including (1) similarity map between frame features before alignment, (2) similarity map between frame features after alignment by our proposed A^{2}FA, and (3) prediction map by our proposed F^{2}SGD.

(a) A query text corresponds to one matching interval.

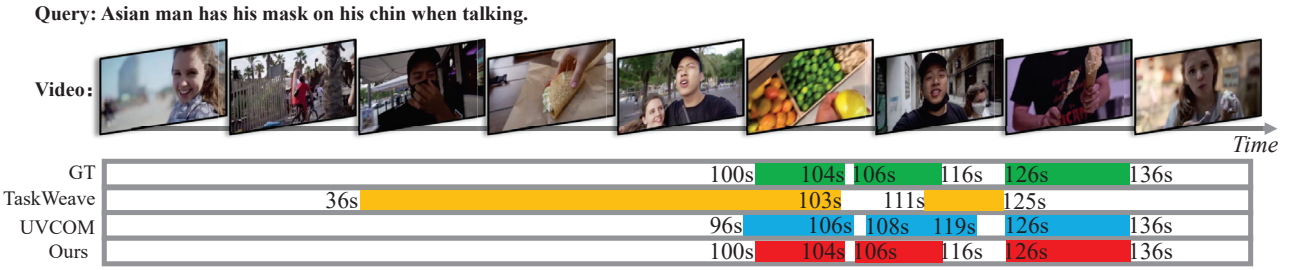


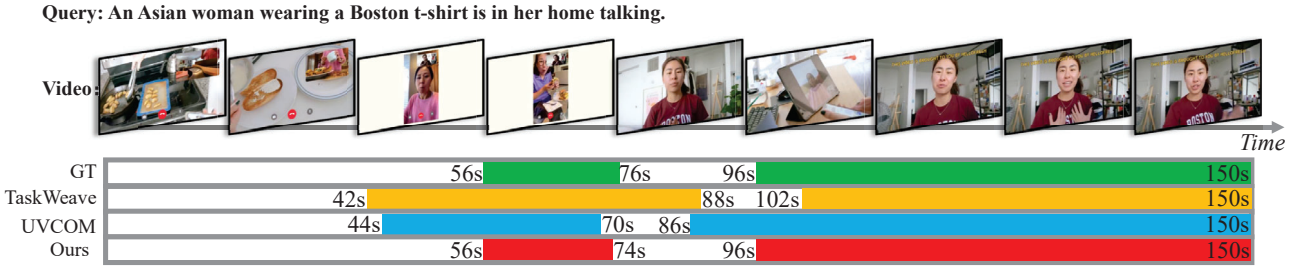
(b) A query text corresponds to multiple matching intervals.

Figure S4. Visualization of query relevance scores (QRS) obtained by Eq. (9) in the main manuscript. GT signifies the true matching interval, SRI refers to the semantically related interval determined by Eq. (5) in the main manuscript.



(a) A query text corresponds to one matching interval.





(b) A query text corresponds to multiple matching intervals.

Figure S5. Visualized comparisons of the proposed method with the SOTA methods TaskWeave [14] and UVCOM [13]. GT denotes ground-truth matching intervals for reference.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. Localizing moments in video with natural language. In *ICCV*, pages 5804–5813, 2017. 1

[2] Jinhyun Jang, Jungin Park, Jin Kim, Hyeongjun Kwon, and Kwanghoon Sohn. Knowing where to focus: Event-aware transformer for video grounding. In *ICCV*, pages 13800–13810, 2023. 2

[3] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715, 2017. 1

[4] Jie Lei, Tamara L. Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In *NeurIPS*, pages 11846–11858, 2021. 1

[5] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In

ICML, pages 19730–19742, 2023. 1

[6] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *ICCV*, pages 2782–2792, 2023. 1

[7] WonJun Moon, Sangeek Hyun, Su Been Lee, and Jae-Pil Heo. Correlation-guided query-dependency calibration in video representation learning for temporal grounding. *CoRR*, abs/2311.08835, 2023. 2

[8] WonJun Moon, Sangeek Hyun, Sanguk Park, Dongchan Park, and Jae-Pil Heo. Query - dependent video representation for moment retrieval and highlight detection. In *CVPR*, pages 23023–23033, 2023. 2

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1, 2

[10] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *CVPR*, pages 14313–14323, 2024. 1

[11] Hao Sun, Mingyao Zhou, Wenjing Chen, and Wei Xie. TR-DETR: task-reciprocal transformer for joint moment retrieval and highlight detection. In *AAAI*, pages 4998–5007, 2024. 2

[12] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2: Scaling foundation models for multimodal video understanding. In *ECCV*, pages 396–416, 2024. 1

[13] Yicheng Xiao, Zhuoyan Luo, Yong Liu, Yue Ma, Hengwei Bian, Yatai Ji, Yujiu Yang, and Xiu Li. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In *CVPR*, pages 18709–18719, 2024. 1, 2, 4

[14] Jin Yang, Ping Wei, Huan Li, and Ziyang Ren. Task-driven exploration: Decoupling and inter-task feedback for joint moment retrieval and highlight detection. In *CVPR*, pages 18308–18318, 2024. 2, 4