Website

ICCV HONOLULU HAWAII
OCT 19-23, 2025

# THE DEVIL IS IN THE SPURIOUS CORRELATIONS:
## Boosting Moment Retrieval with Dynamic Learning

Xinyang Zhou*[1], Fanyue Wei*[2], Lixin Duan[1], Angela Yao[2], Wen Li[1],
[1]University of Electronic Science and Technology of China,
[2]National University of Singapore

Code: https://github.com/xyangzhou/TD-DETR
Website: https://xyangzhou.github.io/TD-DETR

Paper ID: 6639
Exhibit Hall I #1933

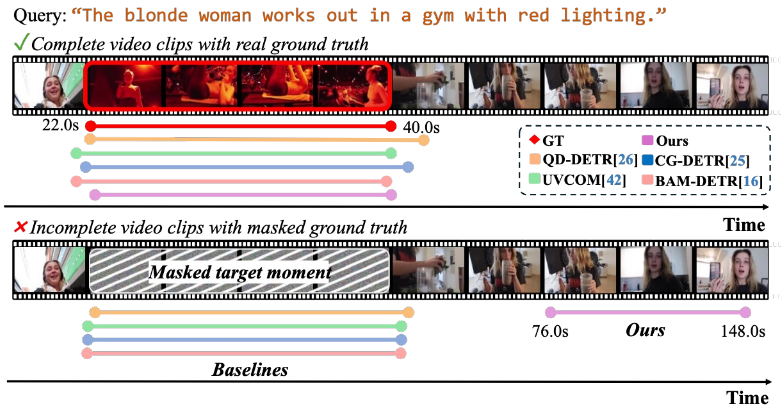
* Equal contribution

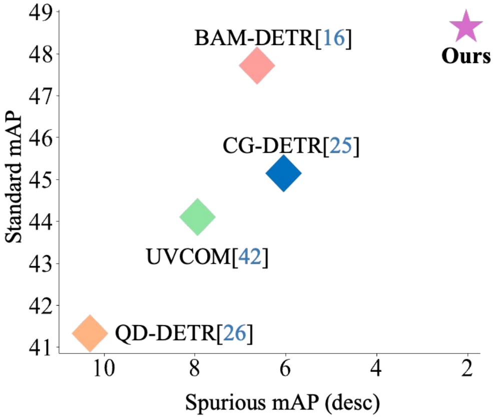
## Video Moment Retrieval


- Browsing through entire videos is time-consuming.
- Tools to retrieve corresponding moments automatically by textual description is widely needed

## Motivation

- The model makes predictions by overly associating queries with background frames rather than distinguishing target moments.

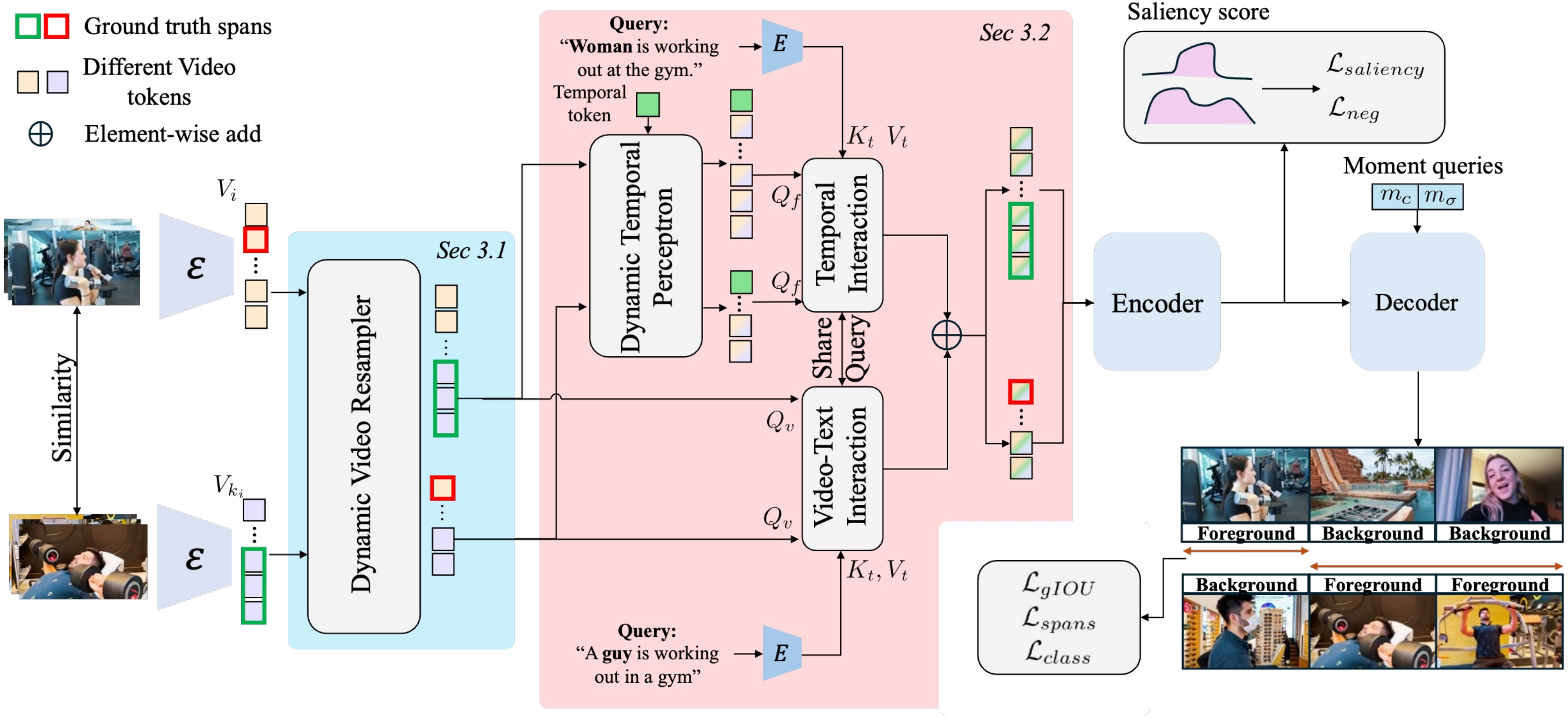



  - Even when the target moment is masked, the existing method still predicts a similar span.
  - Such issues lead to a sub-optimal performance.

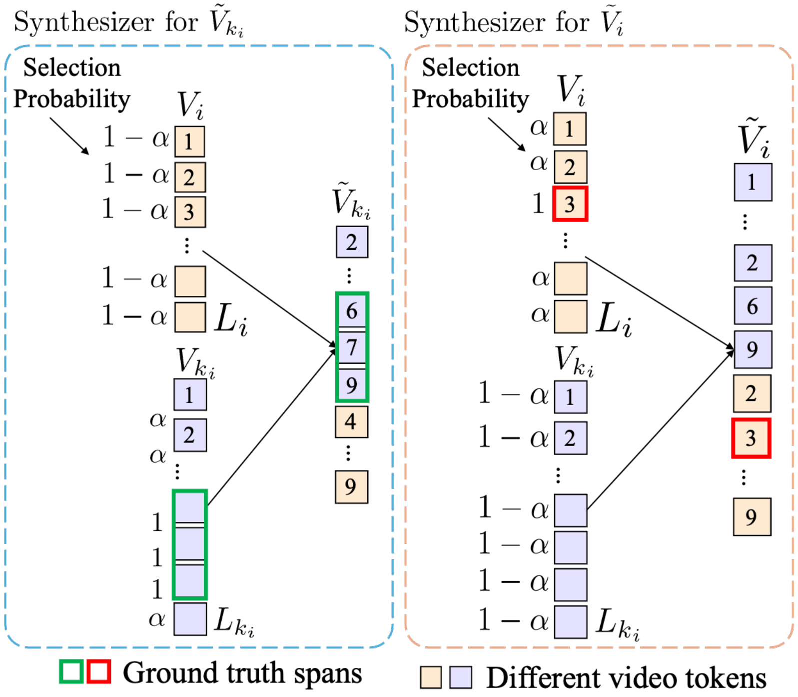
# Overview

- Learning Temporal Dynamics utilizing DETR



## ■ Video Synthesizer for Dynamic Context

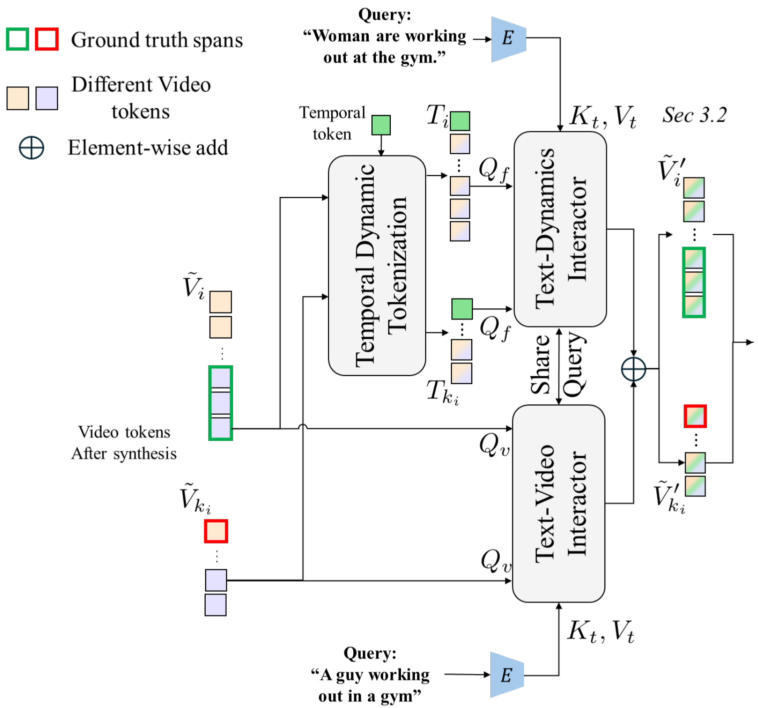
- Spurious correlations stem from linking the moment's **context** to the text **query**.
- Synthesizing new samples for the target moments with more dynamic contextual variations.
- Enforcing model to attend to the target moment **corresponding** to the text query, even with in a dynamic context.



# Approach

## ■ Temporal Dynamics Enhancement

- The attention of DETR-like architecture tends to emphasize background frames.
- Align text queries with temporal **dynamic** representations.
- Establishing a **stand-up** correlation between the query-related moment and its context.



# Experiments

## Results on Standard Evaluation

### QVHighlights dataset

| Method | MR-R1 | | MR-mAP | | |
|---|---|---|---|---|---|
| | @0.5 | @0.7 | @0.5 | @0.75 | Average |
| MCN [2] *ICCV'17* | 11.41 | 2.72 | 24.94 | 8.22 | 10.67 |
| CAL [6] *arXiv'19* | 25.49 | 11.54 | 23.40 | 7.65 | 9.89 |
| XML [17] *ECCV'20* | 41.83 | 30.35 | 44.63 | 31.73 | 32.14 |
| XML+ [18] *NIPS'21* | 46.69 | 33.46 | 47.89 | 34.67 | 34.90 |
| SnAG† [27] *CVPR'24* | 59.79 | 48.10 | 58.63 | 44.37 | 42.71 |
| *SnAG /w TD-DETR* | **66.48** | **52.93** | **63.71** | **49.11** | **46.75** |
| Moment-DETR [18] *NIPS'21* | 52.89 | 33.02 | 54.82 | 29.40 | 30.73 |
| UMT [23] *CVPR'22* | 56.23 | 41.18 | 53.83 | 37.01 | 36.12 |
| MomentDiff [19] *NIPS'23* | 57.42 | 39.66 | 54.02 | 35.73 | 35.95 |
| QD-DETR [26] *CVPR'23* | 62.40 | 44.98 | 62.52 | 39.88 | 39.86 |
| UniVTG [20] *ICCV'23* | 58.86 | 40.86 | 57.60 | 35.59 | 35.47 |
| CG-DETR[25] *arXiv'23* | 65.40 | 48.40 | 64.50 | 42.80 | 42.90 |
| UVCOM [42] *CVPR'24* | 63.55 | 48.70 | 64.47 | 44.01 | 43.27 |
| BAM-DETR[16] *ECCV'24* | 64.53 | 48.64 | 64.57 | 46.33 | 45.36 |
| *TD-DETR* (**Ours**) | **64.53**$_{\pm 0.62}$ | **50.37**$_{\pm 0.53}$ | **66.21**$_{\pm 0.21}$ | **47.32**$_{\pm 0.53}$ | **46.69**$_{\pm 0.26}$ |

†reproduced by the official code

### Charades-STA dataset

| Method | R1@0.5 | R1@0.7 |
|---|---|---|
| CAL [6] | 44.90 | 24.37 |
| 2D TAN [52] | 39.70 | 23.31 |
| VSLNet [49] | 47.31 | 30.19 |
| IVG-DCL [28] | 50.24 | 32.88 |
| SnAG† [27] | 65.72 | 37.32 |
| *SnAG /w TD-DETR* | **70.14** | **42.35** |
| Moment-DETR [18] | 53.63 | 31.37 |
| Moment-Diff [19] | 55.57 | 32.42 |
| UMT [23] | 48.31 | 29.25 |
| QD-DETR [26] | 57.31 | 32.55 |
| CG-DETR[25] | 58.40 | 36.30 |
| BAM-DETR[16] | 59.95 | 39.38 |
| *TD-DETR* (**Ours**) | **60.89** | **40.35** |

†reproduced by the official code

## Results on Spurious Correlation Evaluation

- We replace the target clips of video content with masks without changing the duration of the videos.
- To verify the issue of spurious correlation, we introduce the Spurious mAP as the metric.
- Our model achieves the best ratio of mAP to Spurious mAP.

| Method | Spurious R1 ↓ | | Spurious mAP ↓ | | Standard mAP ↑ | |
|---|---|---|---|---|---|---|
| | @0.7 | @0.9 | @0.75 | Avg. | @0.75 | Avg. |
| QD-DETR | 9.35 | 5.29 | 9.90 | 10.40 | 41.82 | 41.22 |
| Ours w/ QD | 8.26 | 3.68 | 7.46 | 8.15 | 49.86 | 49.05 |
| CG-DETR | 4.65 | 1.29 | 5.55 | 6.14 | 45.70 | 44.90 |
| Ours w/ CG | 2.58 | 0.39 | 3.38 | 4.41 | 49.16 | 48.38 |
| BAM-DETR | 7.16 | 1.87 | 6.30 | 6.72 | 48.56 | 47.61 |
| Ours w/ BAM | 1.61 | 0.52 | 1.73 | 1.98 | 49.62 | 48.67 |

## ■ Ablation study

### Analysis on the proposed components

| | VSDC | TDEM | QVHighlight R1↑ @0.5 | QVHighlight R1↑ @0.7 | QVHighlight mAP↑ @0.5 | QVHighlight mAP↑ @0.75 | QVHighlight mAP↑ Avg. | QVHighlight Spurious R1↓ @0.7 | QVHighlight Spurious R1↓ @0.9 | QVHighlight Spurious mAP↓ @0.75 | QVHighlight Spurious mAP↓ Avg. | Charades-STA R1↑ @0.5 | Charades-STA R1↑ @0.7 | Charades-STA Spurious R1↓ @0.7 | Charades-STA Spurious R1↓ @0.9 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (a) | | | 61.12 | 46.77 | 62.45 | 43.66 | 42.54 | 9.35 | 5.29 | 9.90 | 10.40 | 57.31 | 32.55 | 25.72 | 6.31 |
| (b) | ✓ | | 63.47 | 49.39 | 64.82 | 47.67 | 46.39 | 8.77 | 3.87 | 8.64 | 8.91 | 39.12 | 63.67 | 23.15 | 5.42 |
| (c) | | ✓ | 62.93 | 48.25 | 64.22 | 45.49 | 44.84 | 8.84 | 4.0 | 9.10 | 9.56 | 38.51 | 60.80 | 24.03 | 5.73 |
| (d) | ✓ | ✓ | **65.88** | **53.67** | **66.43** | **49.86** | **49.05** | **8.26** | **3.68** | **7.46** | **8.15** | **60.89** | **40.35** | **22.13** | **4.82** |

### Generalization on different baselines

| Method | QVHighlights val R1@0.7 | QVHighlights val mAP@0.75 | QVHighlights val mAP | Charades-STA test R1@0.5 | Charades-STA test R1@0.7 |
|---|---|---|---|---|---|
| CG | 52.10 | 45.70 | 44.90 | 58.40 | 36.30 |
| Ours w/ CG | 53.25 +1.15 | 49.16 +3.46 | 48.38 +3.48 | 59.35 +0.95 | 37.84 +1.54 |
| BAM | 51.61 | 48.56 | 47.61 | 59.95 | 39.38 |
| Ours w/BAM | 52.87 +1.26 | 49.62 +1.06 | 48.82 +1.21 | 60.92 +0.97 | 40.25 +0.87 |
| QD | 46.66 | 41.82 | 41.22 | 57.31 | 32.55 |
| Ours w/ QD | 53.67 +7.01 | 49.86 +8.04 | 49.00 +7.78 | 60.89 +3.58 | 40.35 +7.80 |
| SnAG | 48.10 | 44.37 | 42.71 | 65.72 | 37.32 |
| Ours w/ SnAG | 52.93 +4.83 | 49.11 +4.74 | 46.75 +4.04 | 70.14 +4.42 | 42.35 +5.03 |

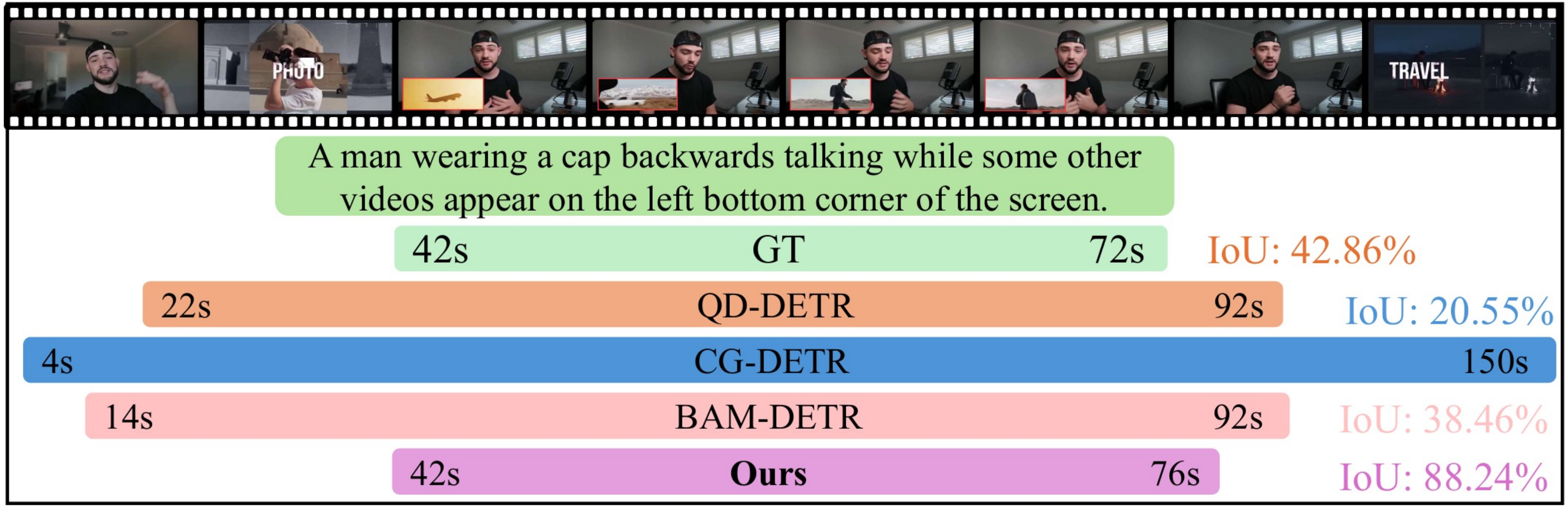
### Comparisons across different sampling strategies.

| Method | QVHighlights R1@0.7 | QVHighlights mAP@0.75 | QVHighlights mAP | Charades-STA R1@0.5 | Charades-STA R1@0.7 |
|---|---|---|---|---|---|
| baseline | 46.66 | 41.82 | 41.22 | 57.31 | 32.55 |
| w/ random | 51.29 | 47.82 | 47.56 | 58.66 | 37.98 |
| w/ similarity | **53.67** | **49.86** | **49.05** | **60.89** | **40.35** |

# Experiments

■ Qualitative Analysis

Example MR prediction for the given masked video.



# Conclusion

## ■ Contribution

- To the best of our knowledge, we are the first to investigate the **spurious correlation** in moment retrieval.

- We propose a **dynamic** learning approach that mitigates *spurious correlations*
  - Dynamically contextualizing target moments through novel video **synthesis**
  - Enhancing representations with **aligned** temporal dynamics.

- The proposed method achieves **state-of-the-art** performance across all benchmarks and provides a strong interpretation of *spurious correlations*.

# Thank you