ICCV OCT 19-23, 2025 HONOLULU HAWAII

Aligning Moment in Time using Video Queries

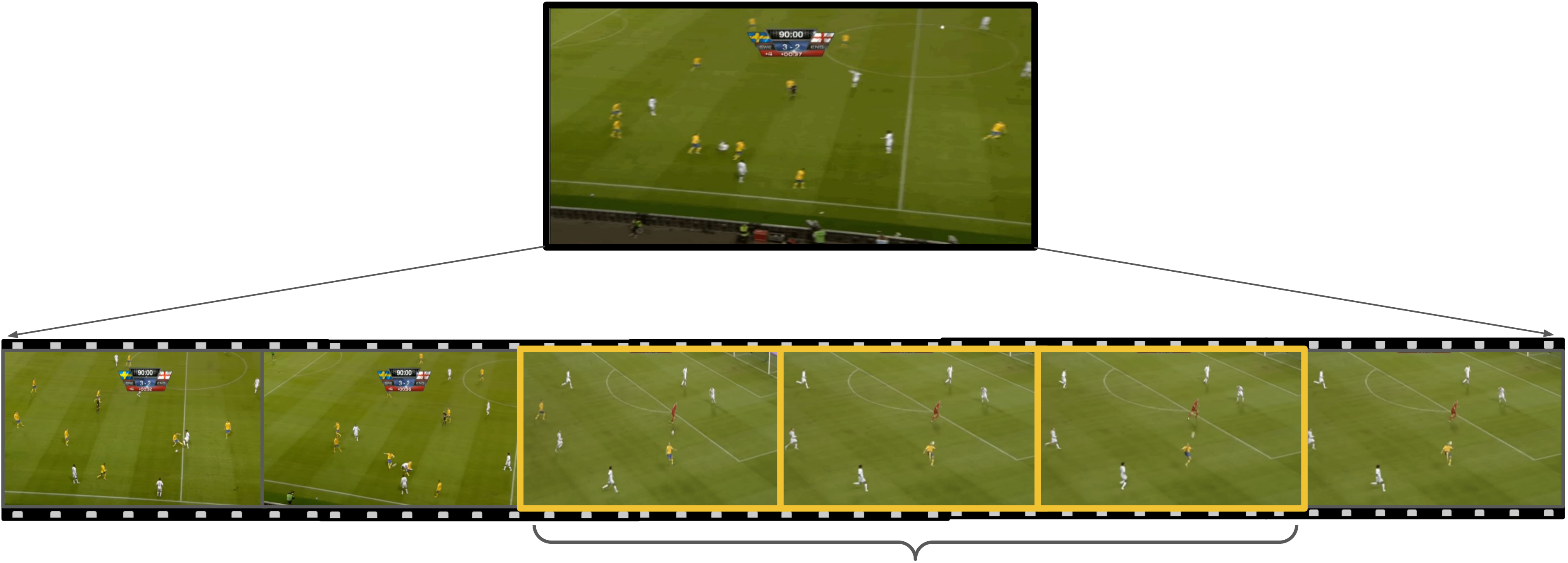
Yogesh Kumar[1*], Uday Agarwal[1*], Manish Gupta[2], Anand Mishra[1]

*Equal Contribution

1 भारतीय प्रौद्योगिकी संस्थान जोधपुर
Indian Institute of Technology Jodhpur

2 Microsoft

Video Moment Retrieval

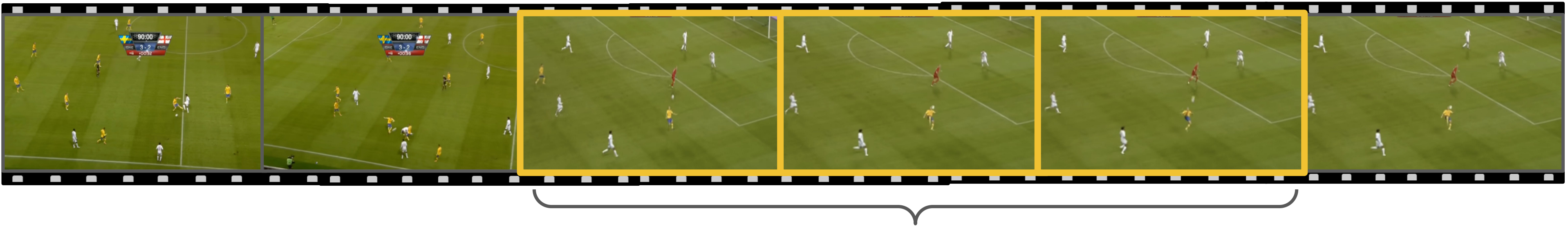


Segment of user Interest

How to Represent a Query?

***Option 1**: Text Query*

"A man positions himself beneath the ball, leaping into the air, bending his knees and arching his back. As he flips backward, he connects with the ball using his foot."



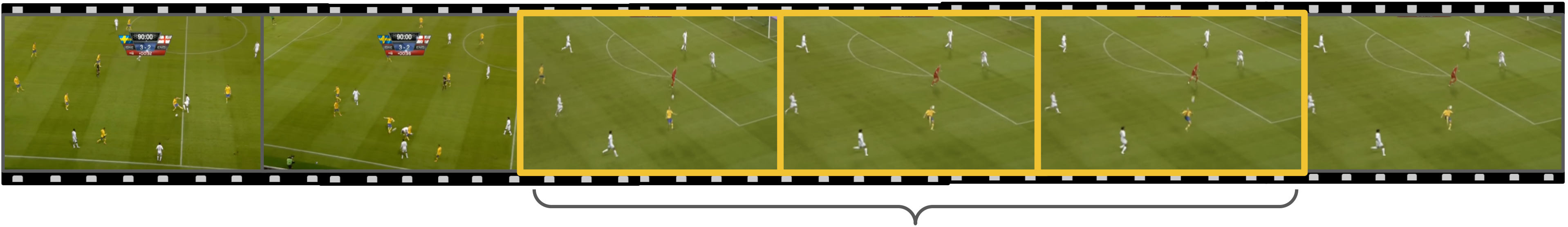
Segment of user Interest

How to Represent a Query?

Option 1: *Text Query*

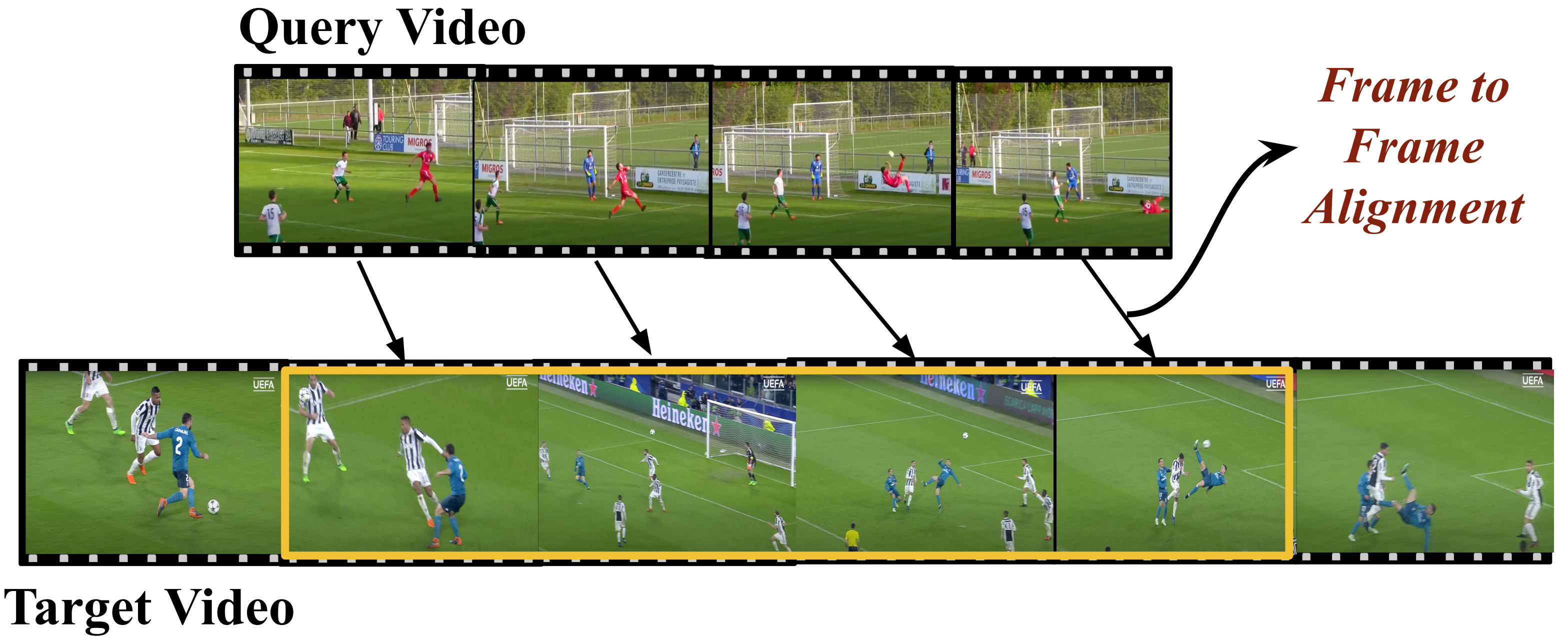
"A man positions himself beneath the ball, leaping into the air, bending his knees and arching his back. As he flips backward, he connects with the ball using his foot."

Option 2: *Video Query*



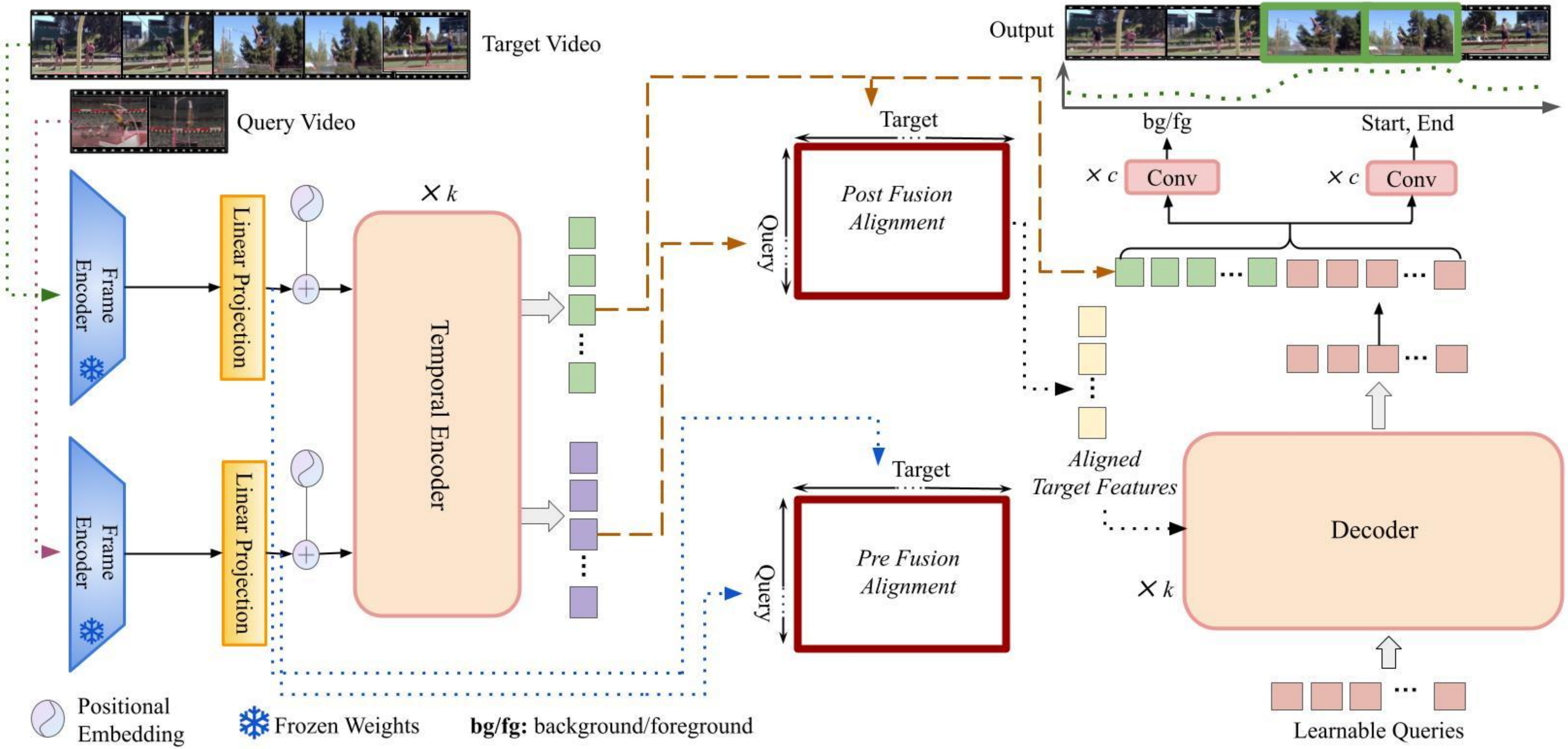
Segment of user Interest

Advantage of Video Query

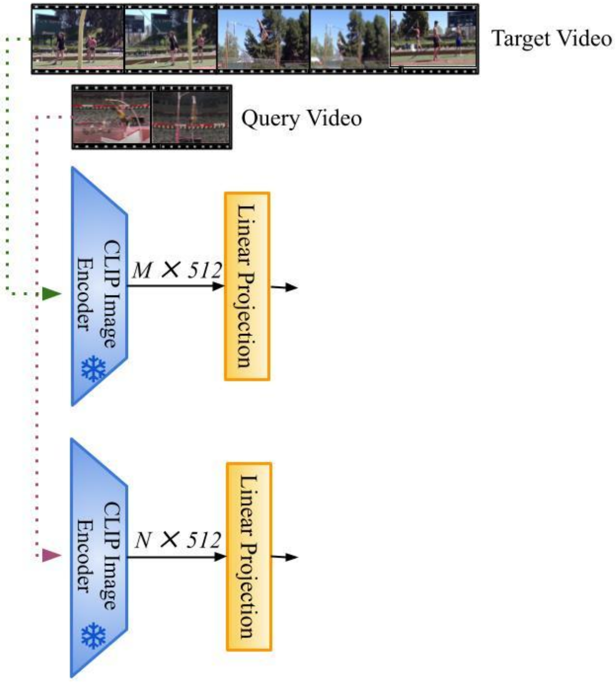


- *Enables fine-grain alignment and correlation Learning*
- *Capture rich spatial-temporal cues directly from video query*

Proposed Method: MATR (Moment Alignment TRansformer)

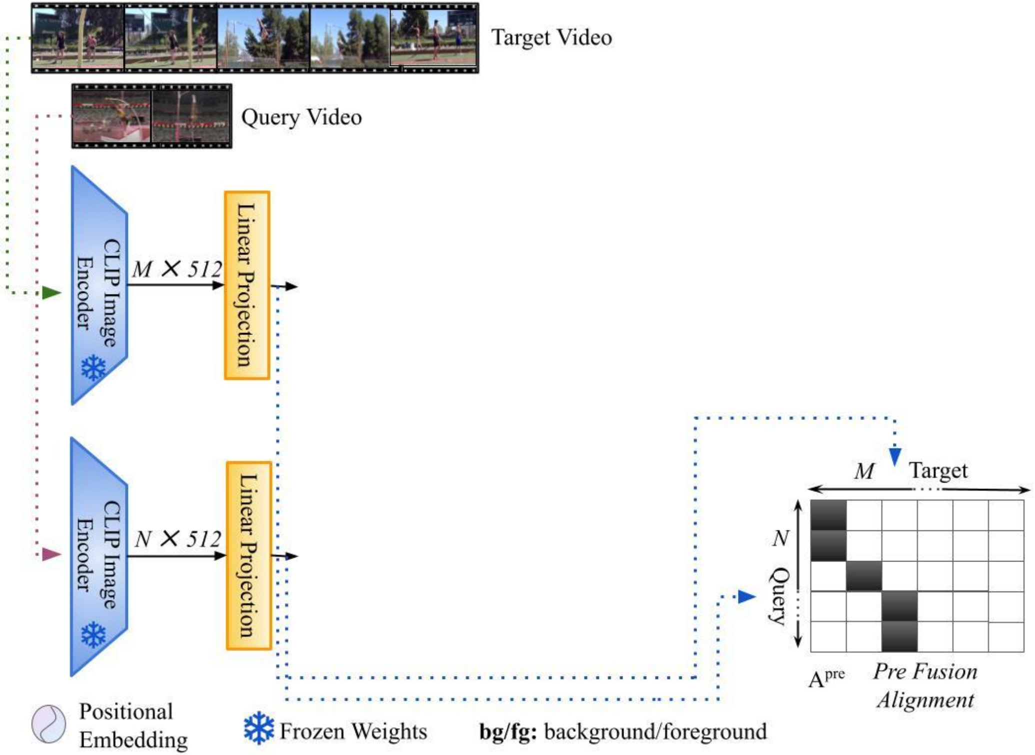


MATR: *Video Feature Extraction*

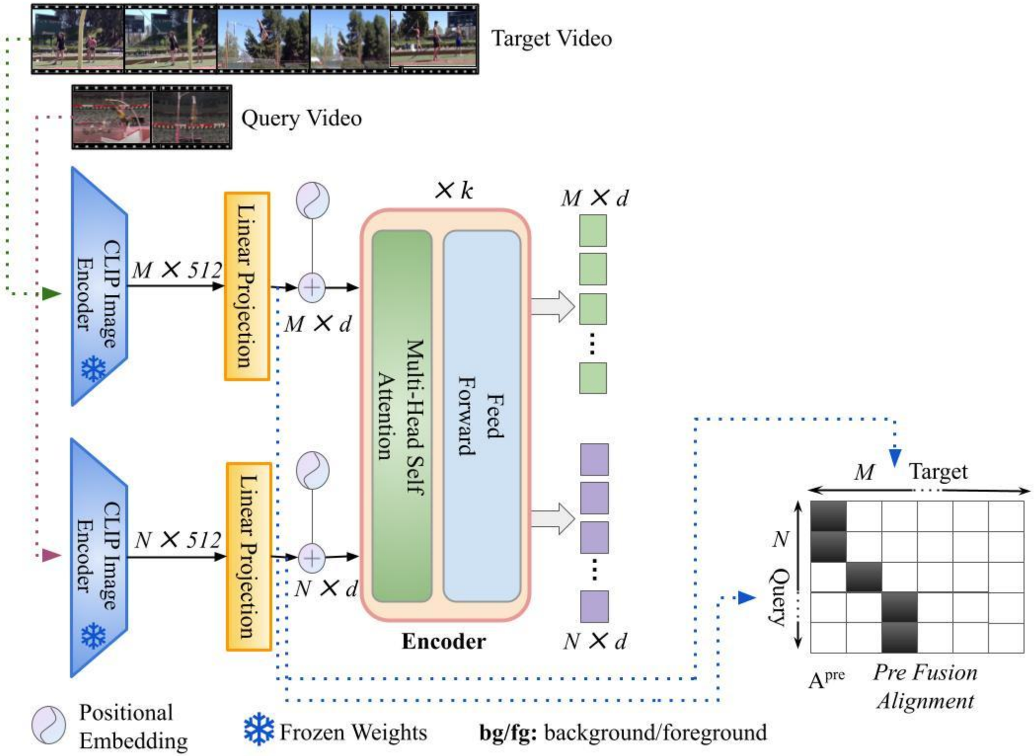


Positional Embedding ❄ Frozen Weights **bg/fg:** background/foreground

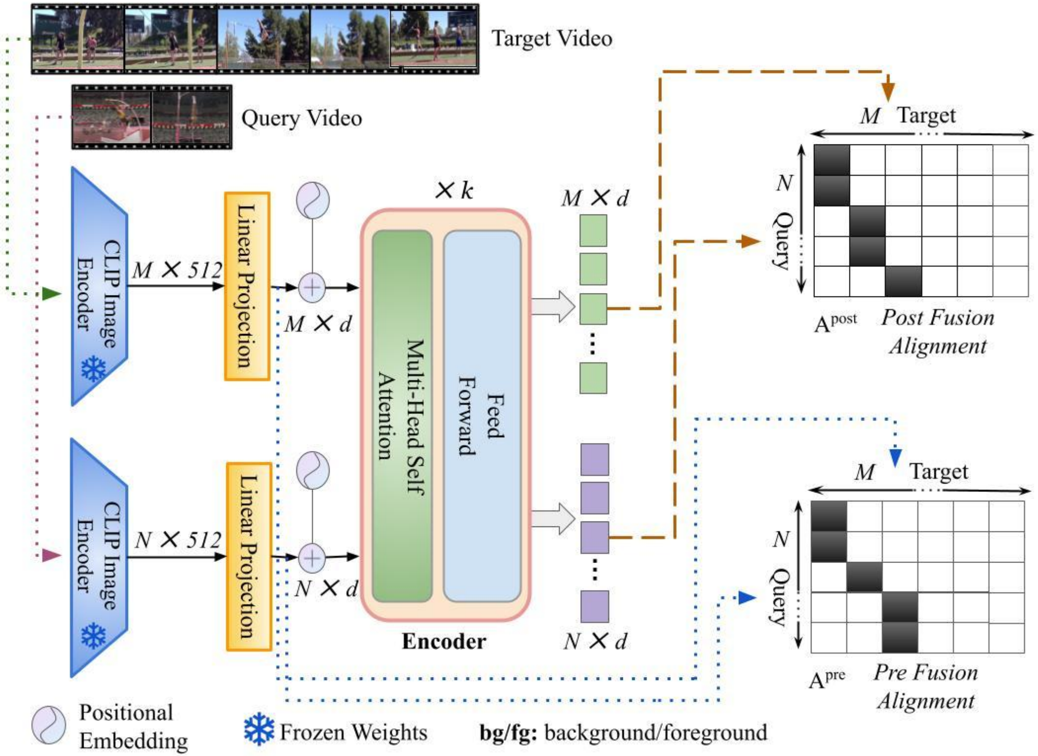
MATR: *Fre-fusion Alignment*



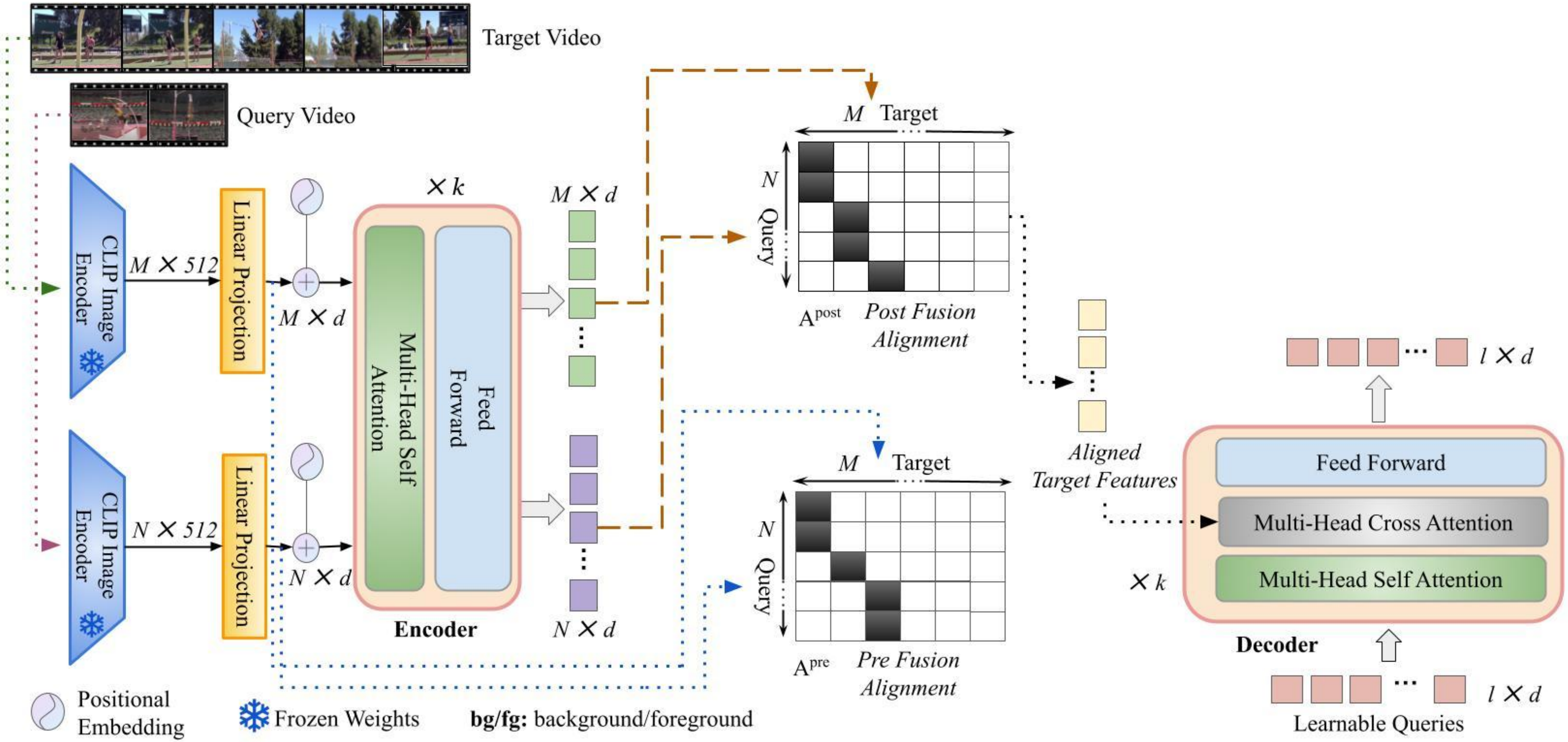
MATR: *Fre-fusion Alignment*



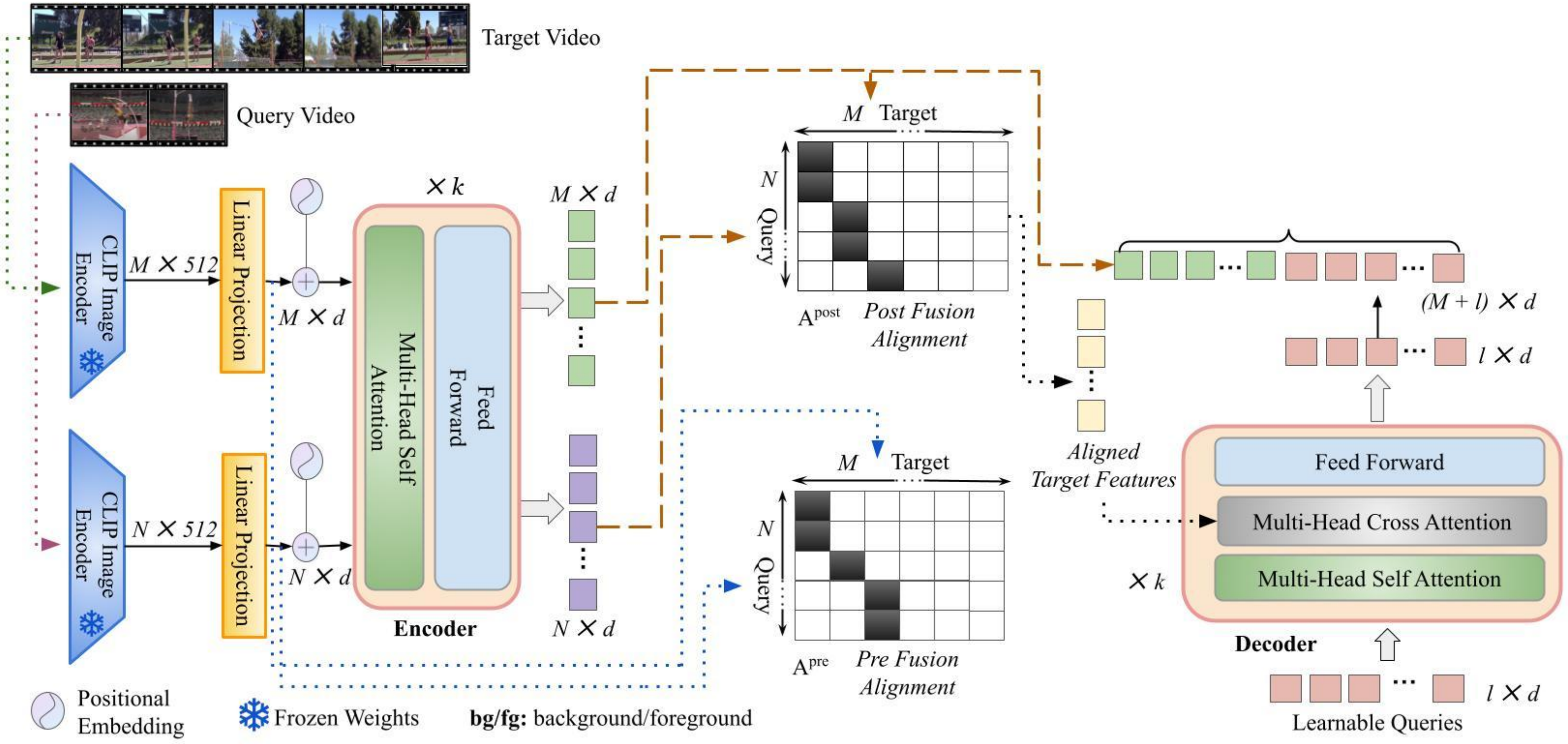
MATR: *Post-fusion Alignment*



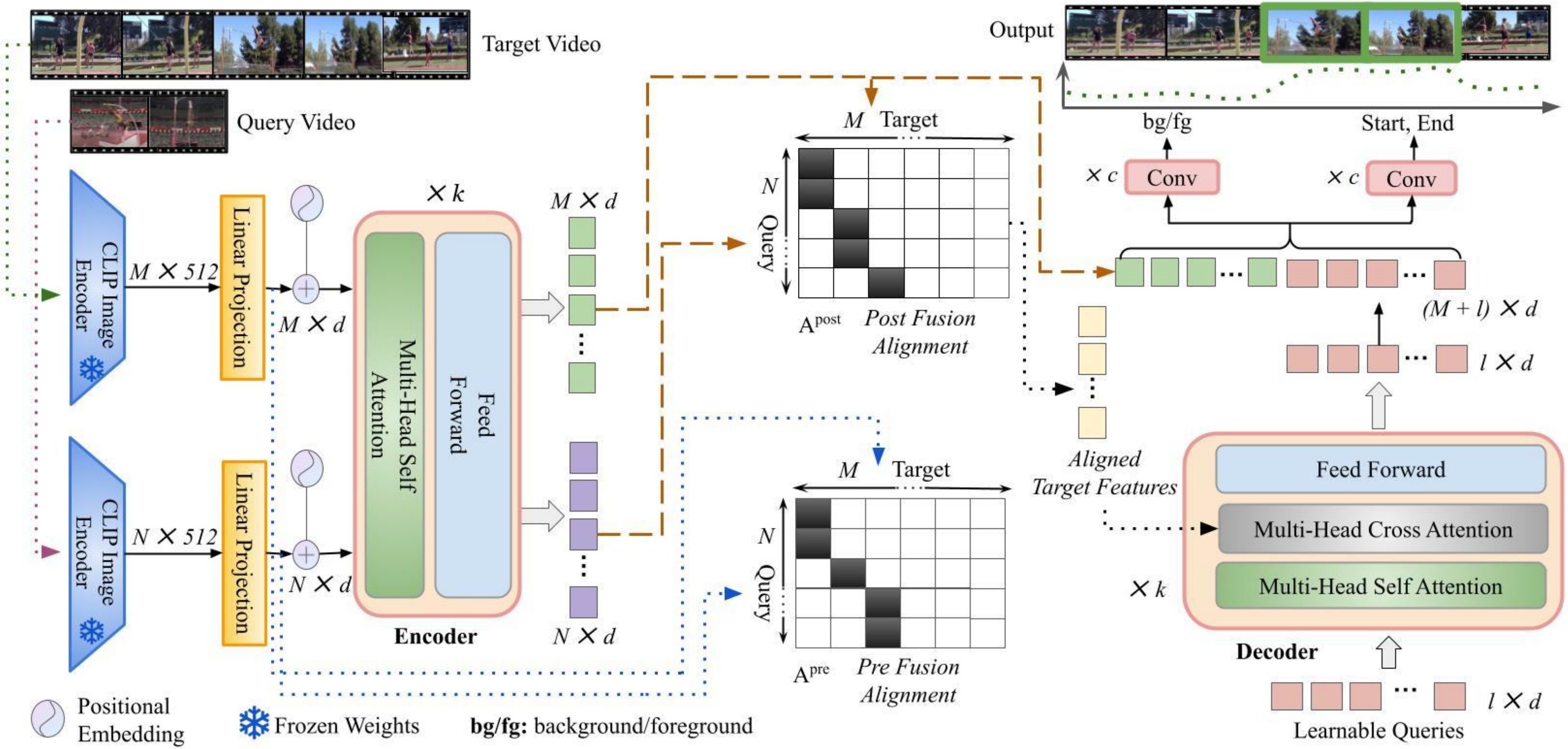
MATR: *Passing Aligned Target Feature to Decoder*



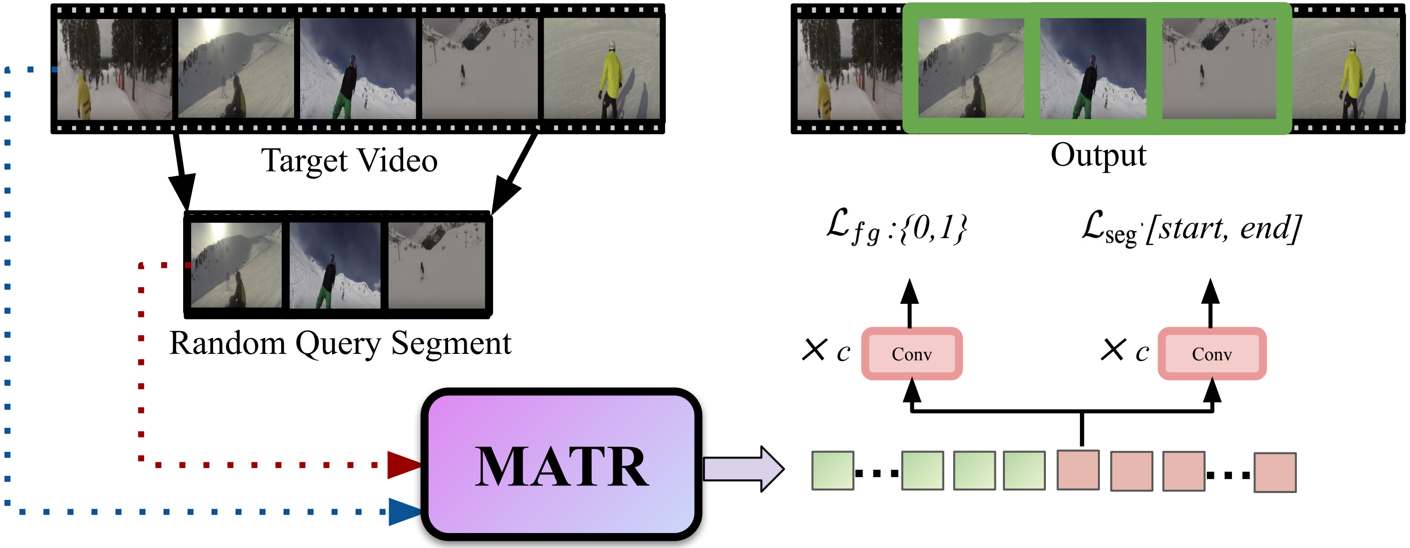
MATR: *Combining Query fused and Query aligned representations of target Video*



MATR: *Predicting moments using heads on combined target representation*



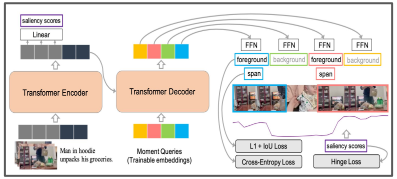
Pre-training: *How to better initialize MATR?*

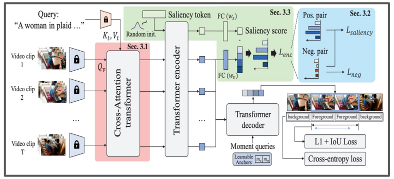


Random clip localization

Competitive Approaches

Text-VMR Methods

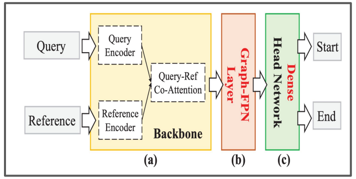


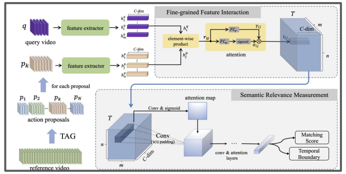


Moment-DETR QD-DETR

[Lei et al., NeurIPS’21;Moon et al., CVPR’23]

Video to Video VMR Methods

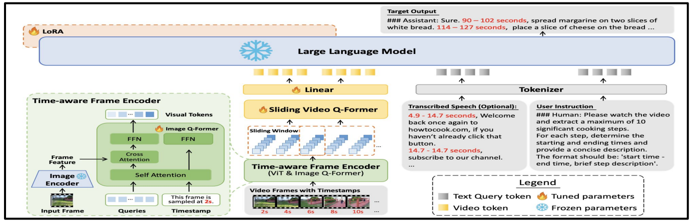




GDP FFI_SRM

[Chen et al., AAAI’20;Huo et al., TMM’23]

Vision Language Models



Target Output

Assistant: Sure. 90 – 102 seconds, spread margarine on two slices of white bread. 114 – 127 seconds, place a slice of cheese on the bread ...

Transcribed Speech (Optional):

4.9 - 14.7 seconds, Welcome back once again to howtocook.com, if you haven‘t already click that button.
14.7 - 14.7 seconds, subscribe to our channel.
...

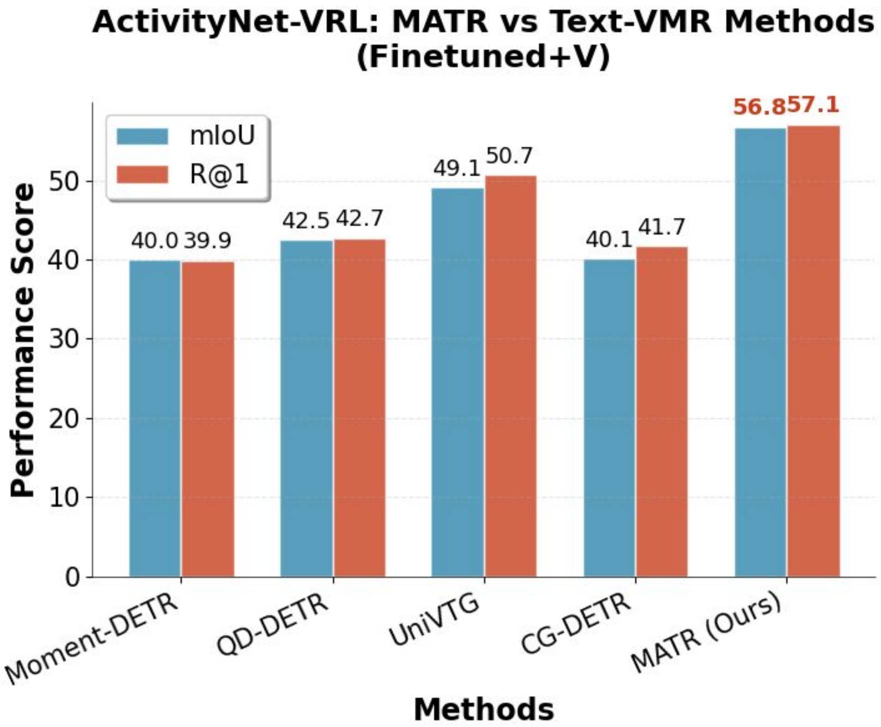
User Instruction

Human: Please watch the video and extract a maximum of 10 significant cooking steps.
For each step, determine the starting and ending times and provide a concise description.
The format should be: 'start time - end time, brief step description'.

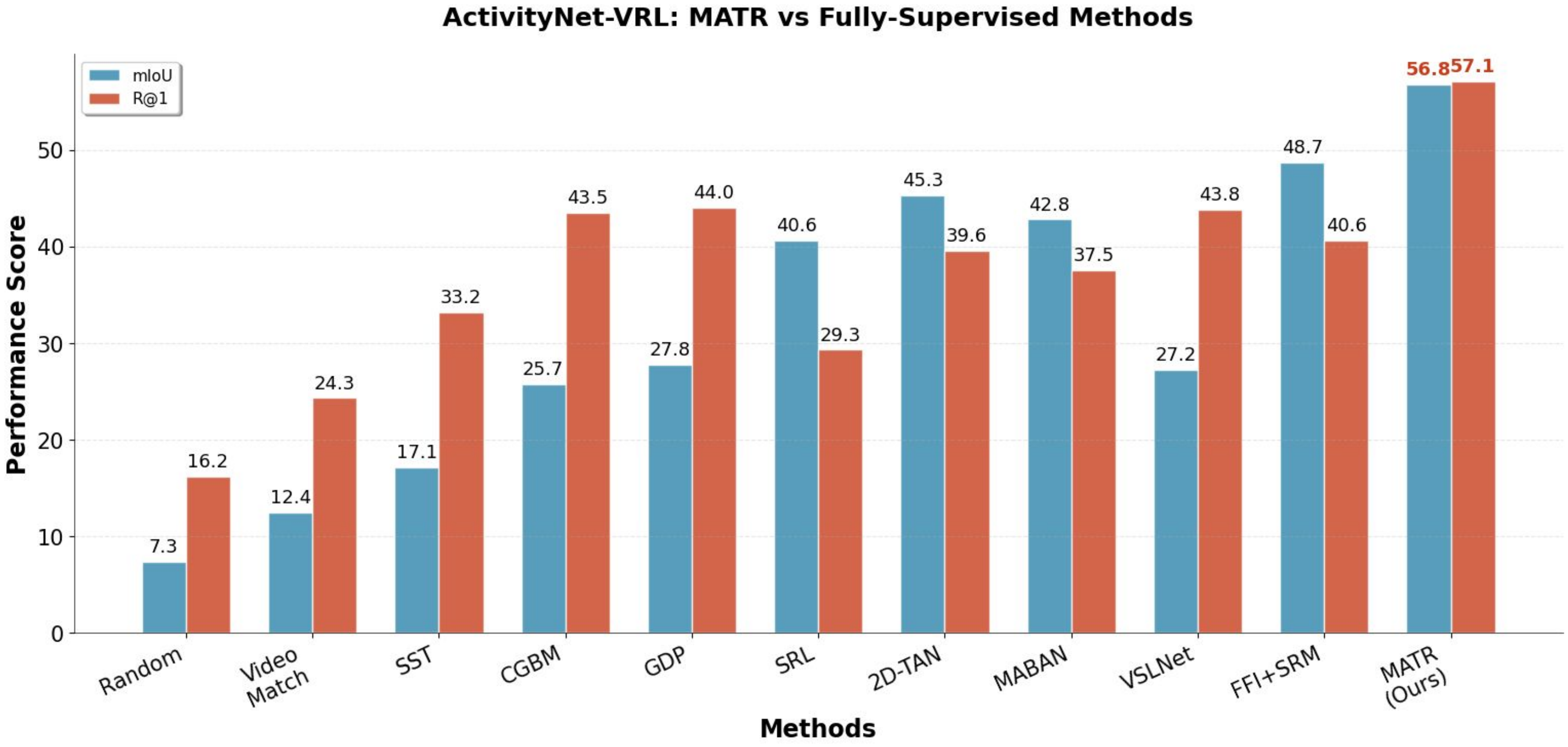
TimeChat

[Ren et al., CVPR 2023]

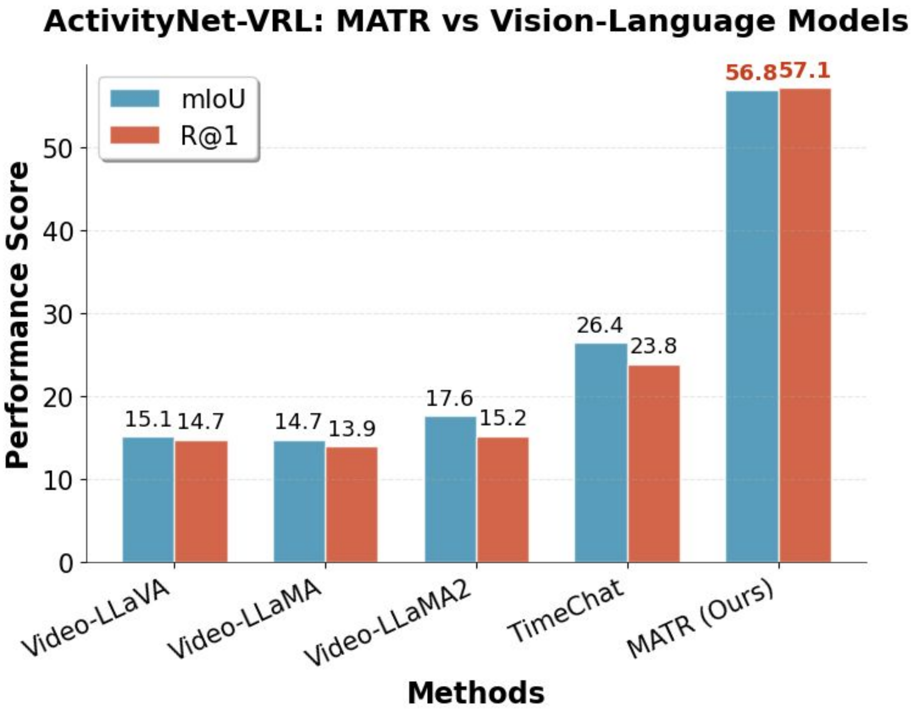
Results: *Comparison with Text-VMR Methods*



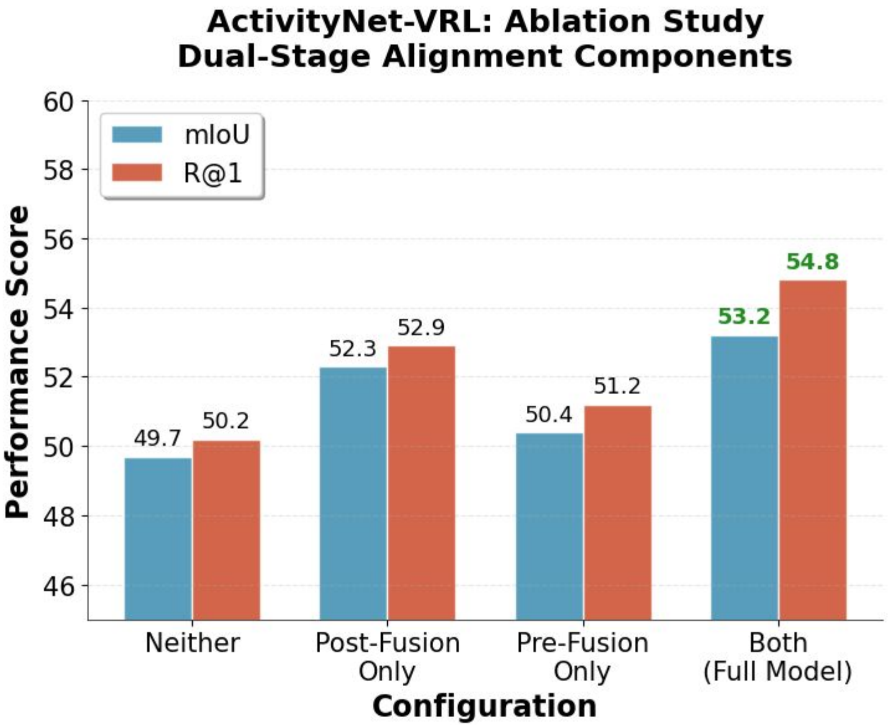
Results: *Comparison with Video-to-Video Methods*



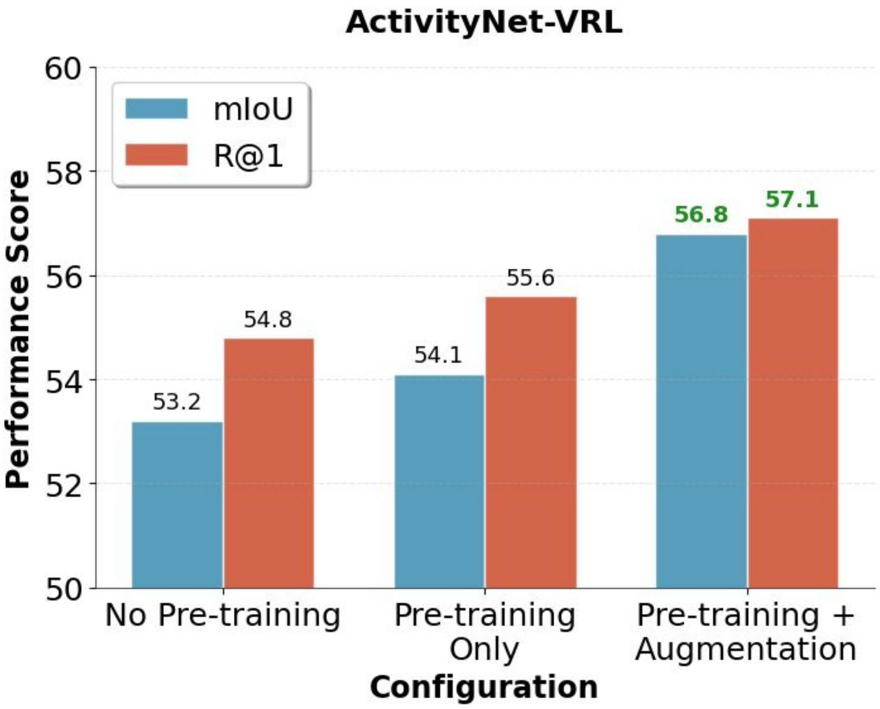
Results: *Comparison with Vision-Language Models*



Ablations: *Pre/Post Fusion Alignment*



Advantage of Pre-training



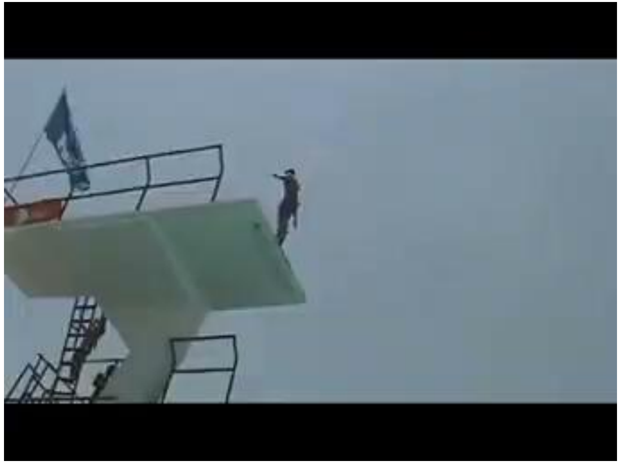
Qualitative Results (1/2)

Query Video

Target Video

Ground truth: [2.0, 12.7], **Prediction**: [1.7, 12.7]

Qualitative Results (2/2)



Query Video

Target Video

Ground truth: [13.3, 25.6], **Prediction**: [13.2, 25.8].

Conclusion

- MATR advances Video to Video moment retrieval via:
  - *Dual-stage alignment within transformer*
  - *Self-supervised pre-training*
  - *Strong performance across benchmarks*
- Future Directions
  - *Multi-Moment Extension*
  - *Multimodal Queries (Video + Text)*

https://github.com/vl2g/MATR

भारतीय प्रौद्योगिकी संस्थान जोधपुर
Indian Institute of Technology Jodhpur

Microsoft