Support-Set Bottlenecks for Video-Text Representation Learning

Mandela Patrick*, Po-Yao Huang*, Yuki M. Asano*,
Florian Metze, Alexander Hauptmann, Joao Henriques, Andrea Vedaldi

ICLR 2021 spotlight

* equal contribution

UNIVERSITY OF OXFORD

CARNEGIE MELLON UNIVERSITY PITTSBURGH PENNSYLVANIA 1900

FACEBOOK

Noise contrastive learning

$f($  $)$ $=$ $f($  $)$ $\neq$ $f($  $)$

Key idea:
features should encode image’s core information.
learn this by comparing augmentations against other images.

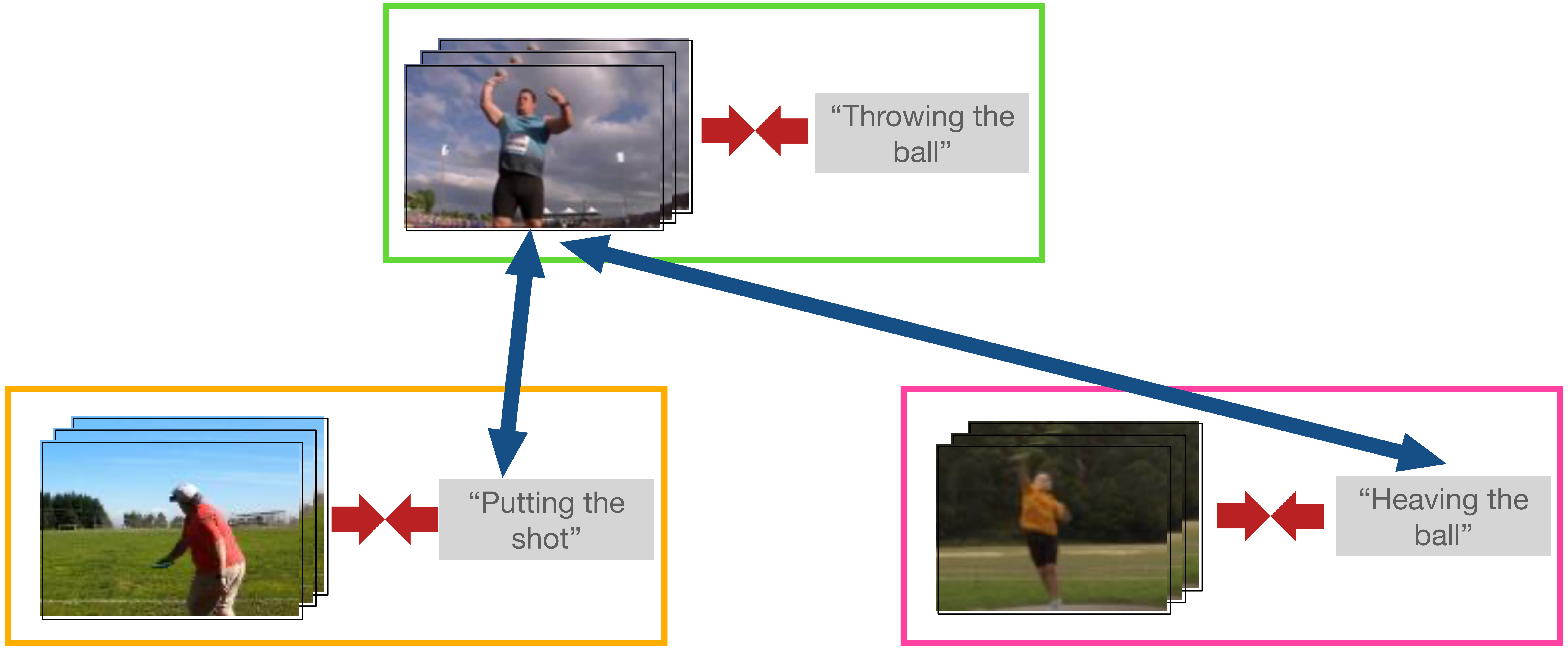
Examples: NPID, MoCo, CMC, SimCLR…

[Wu et al., CVPR 2018; He et al., CVPR 2020; Tian et al., ECCV 2020; Chen et al., ICML 2020]

Noise contrastive learning for Video-Text Representation Learning

Multi-modal contrastive formulation:

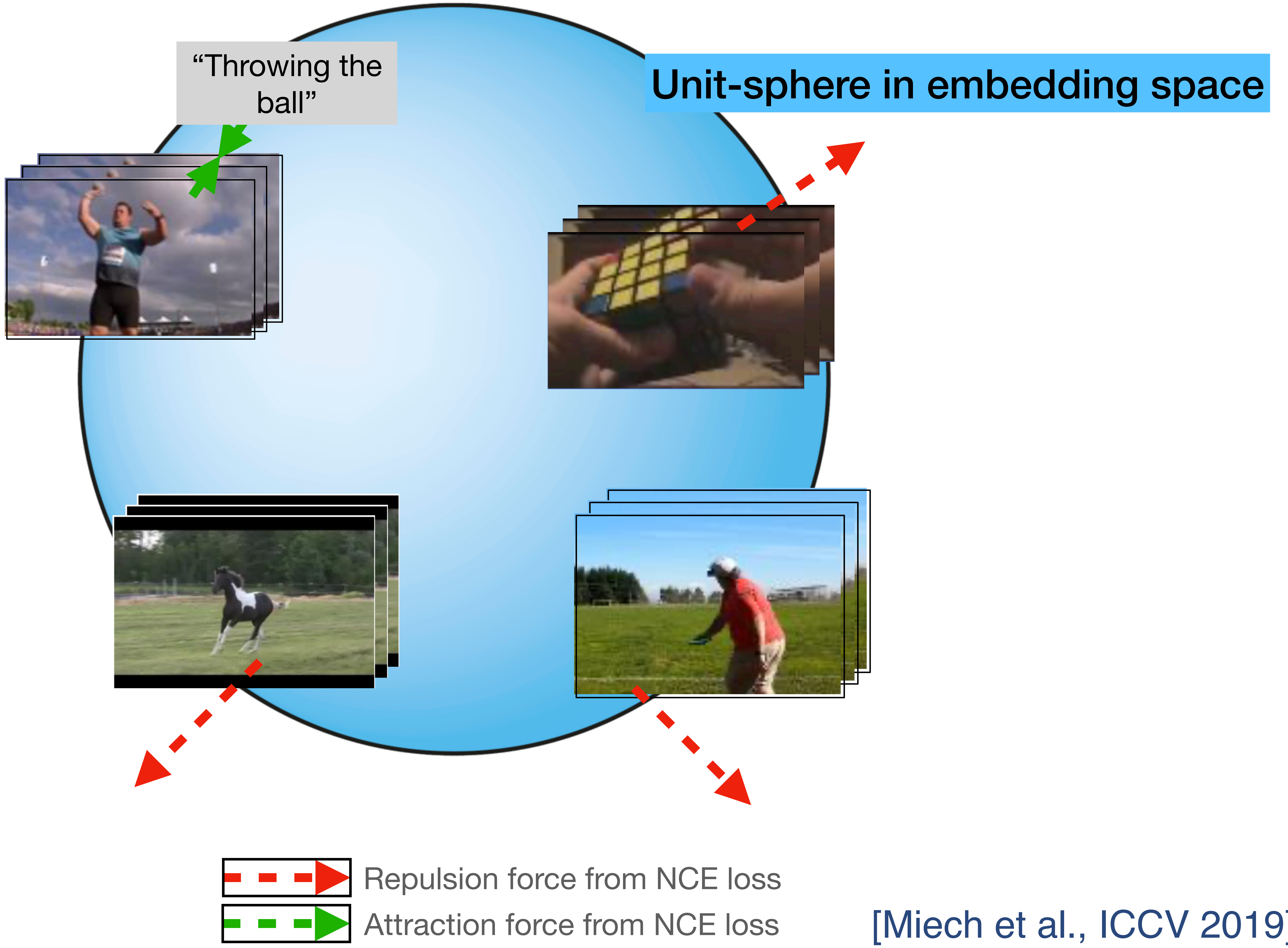
- **Pull** together videos and **their** captions
- **Push** apart videos and **other** captions



[Miech et al., ICCV 2019; Miech et al., CVPR 2020]

Curse of Noise contrastive learning: Faulty Negatives

This can incorrectly push apart videos with the same content

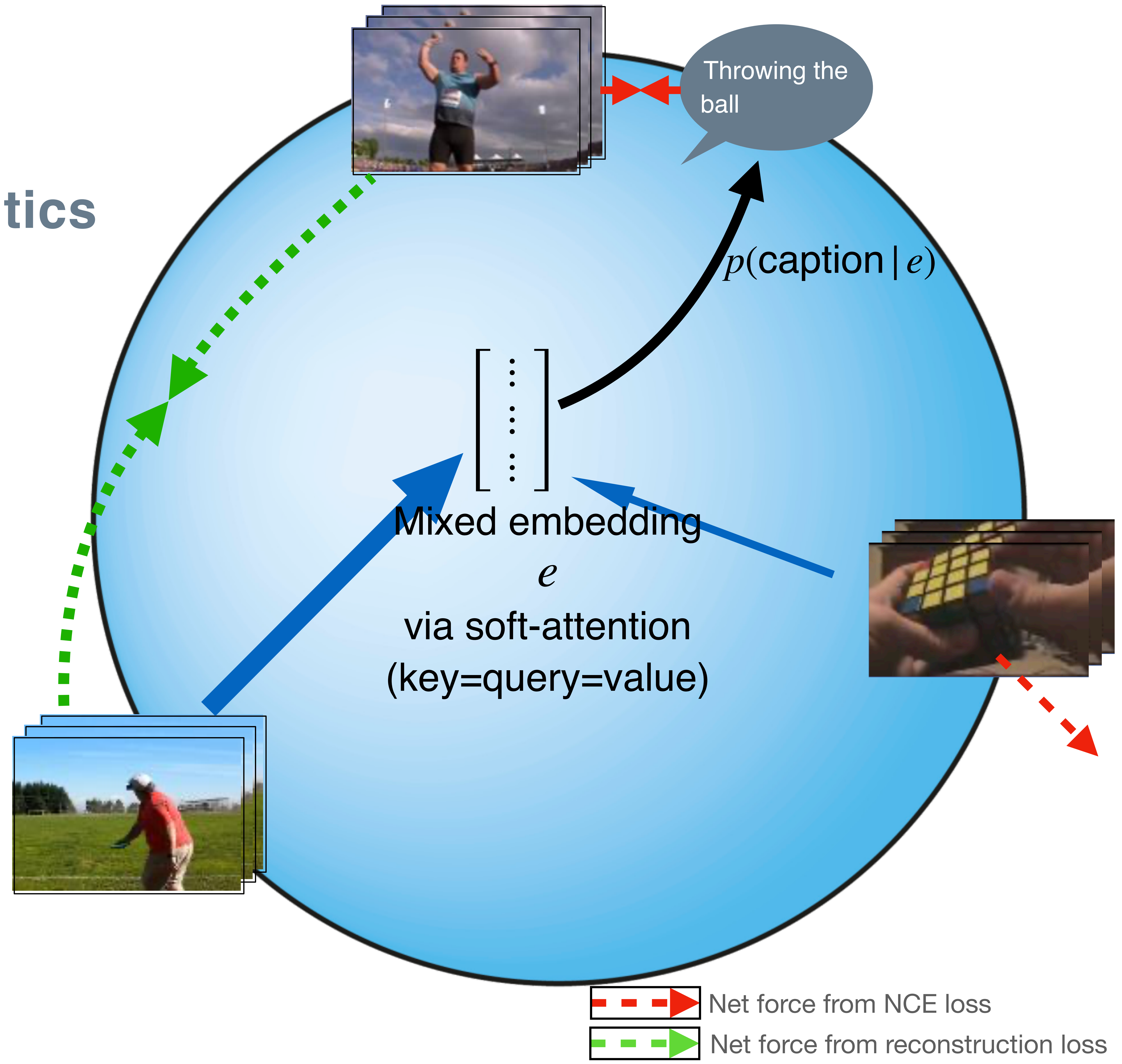


[Miech et al., ICCV 2019]

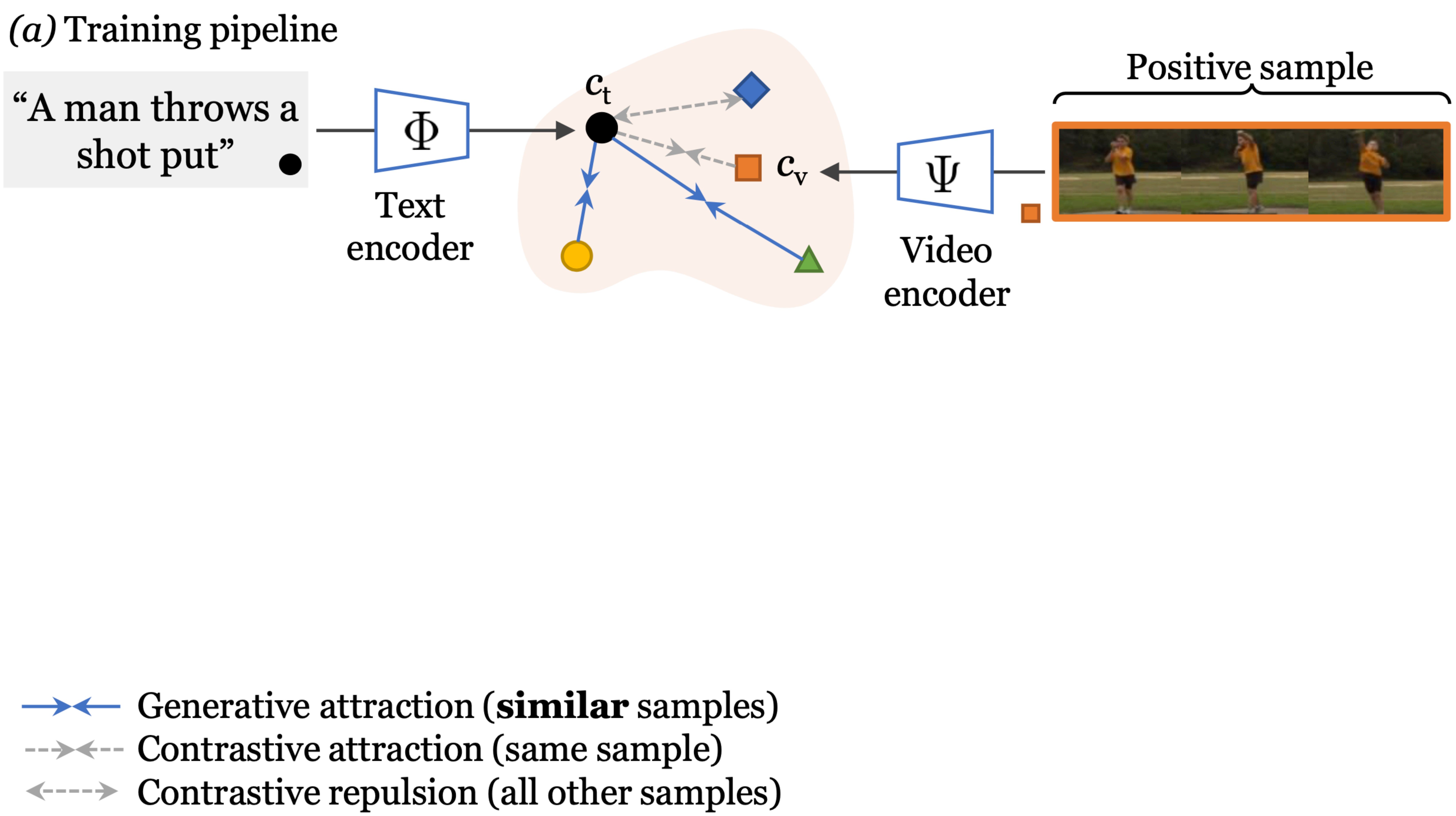
Key Insight:
Attention for Shared Semantics

We assume that, for each video, the batch always contain **at least one more congruent video**

We then learn to predict a video's caption based on the other videos that the network thinks are the most related

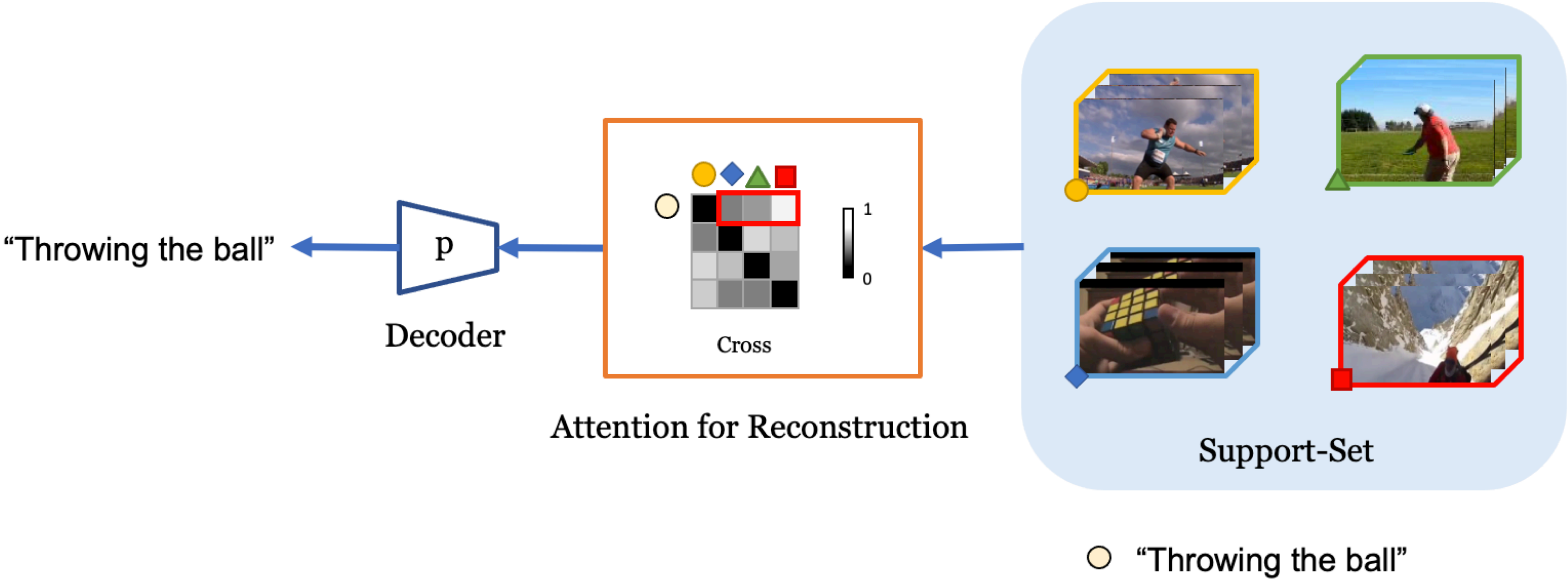


Our Approach: Support-set Bottlenecks



Cross Attention for Reconstruction

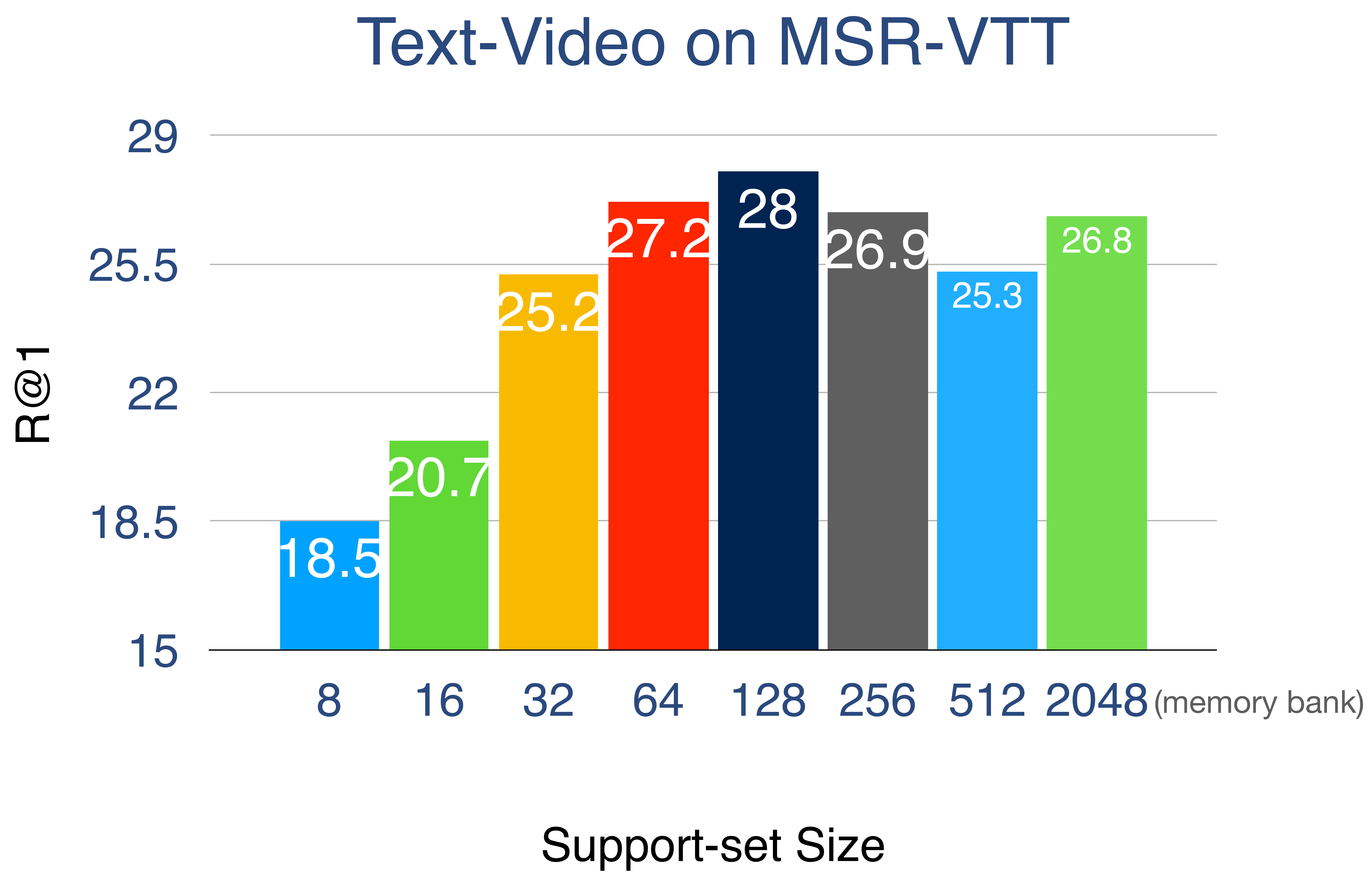
- Learn to reconstruct caption as weighted combination of videos in the support-set
- Implicitly pulls together videos with similar captions



Example Videos and Attentions

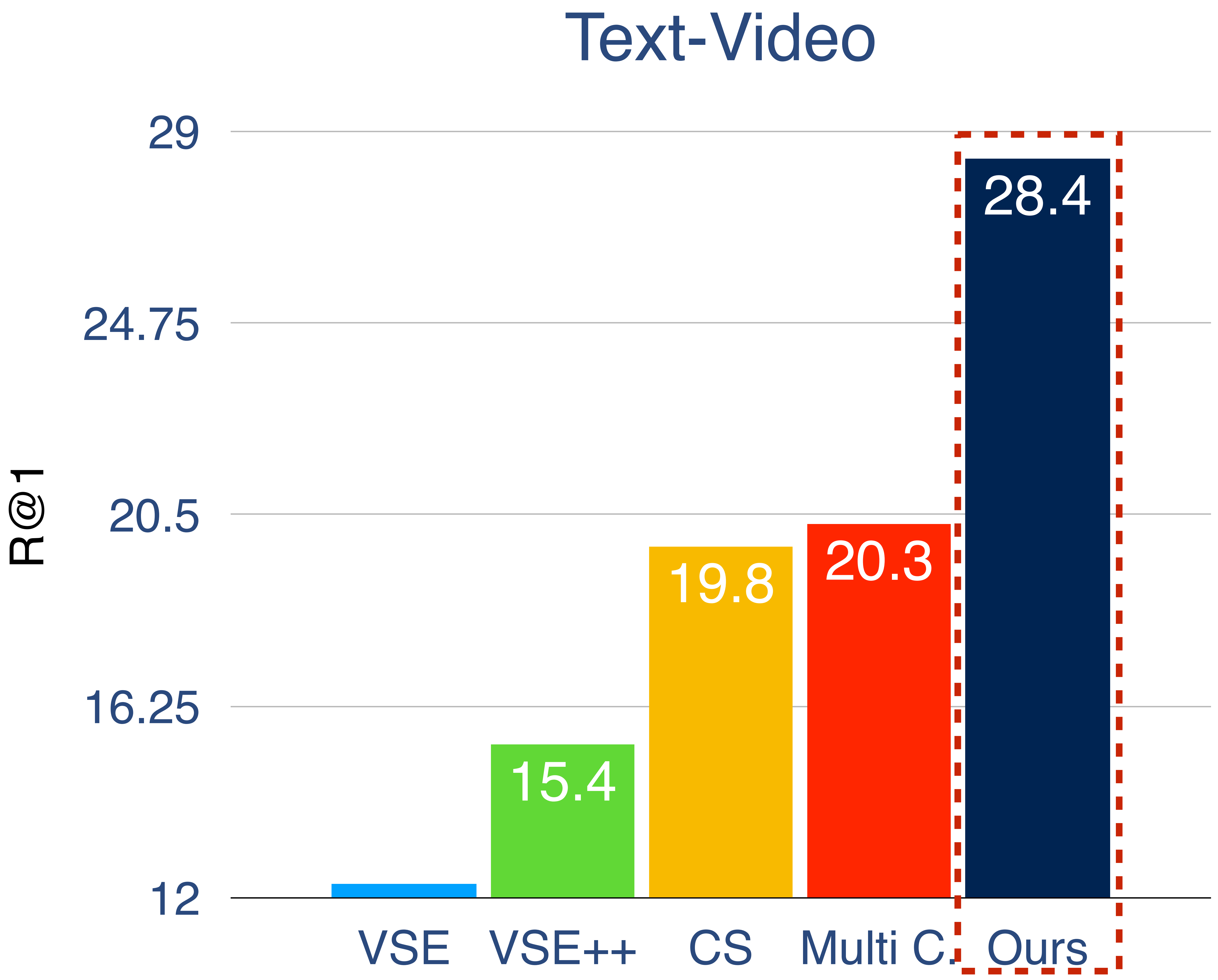
- Attention is highly-focused (top-left square)
- Attention acts as bottleneck (tries to relate semantic concepts across videos)

Support Set acts as bottleneck



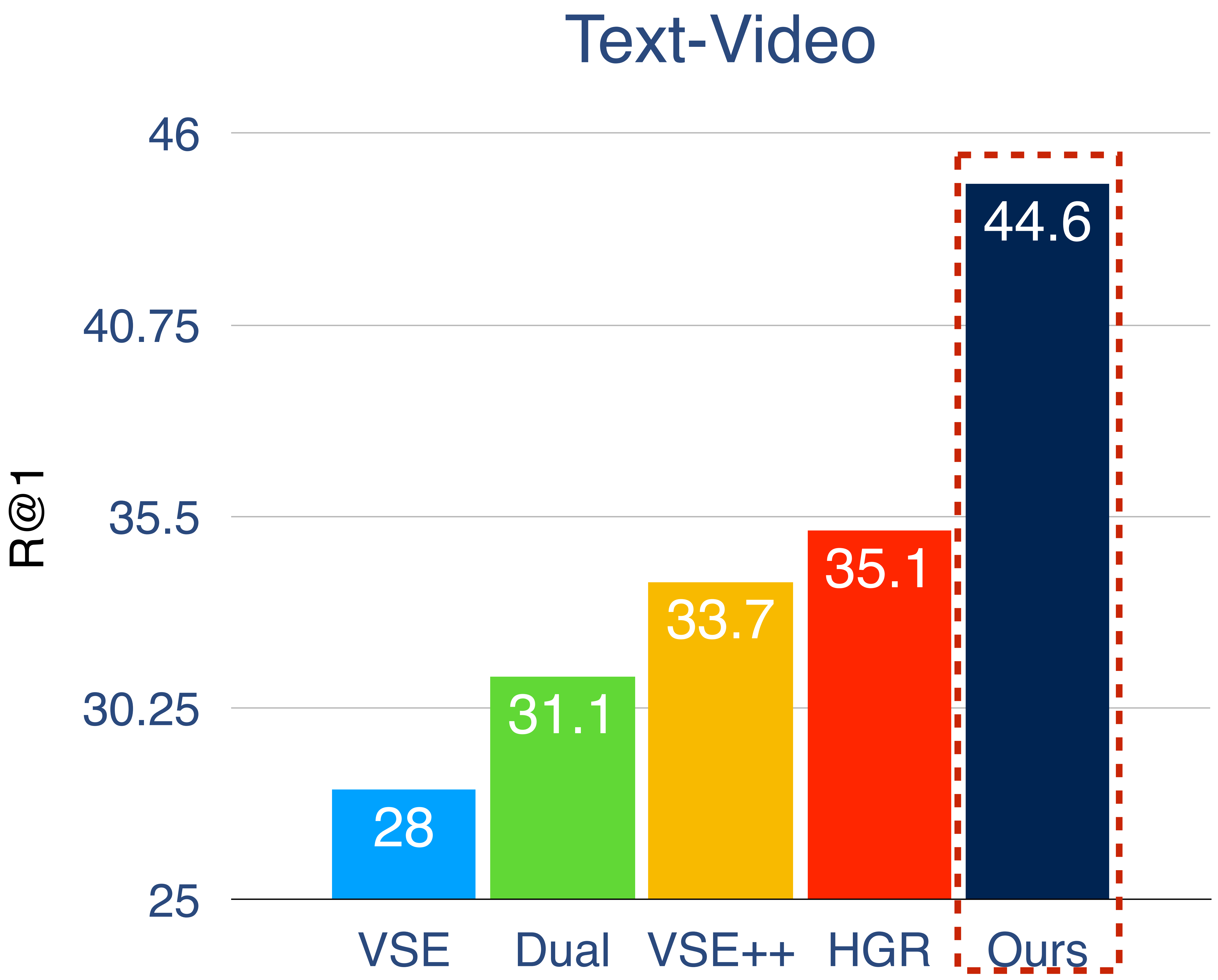
Comparison to State-of-the-Art

SOTA on MSVD Text-Video Retrieval



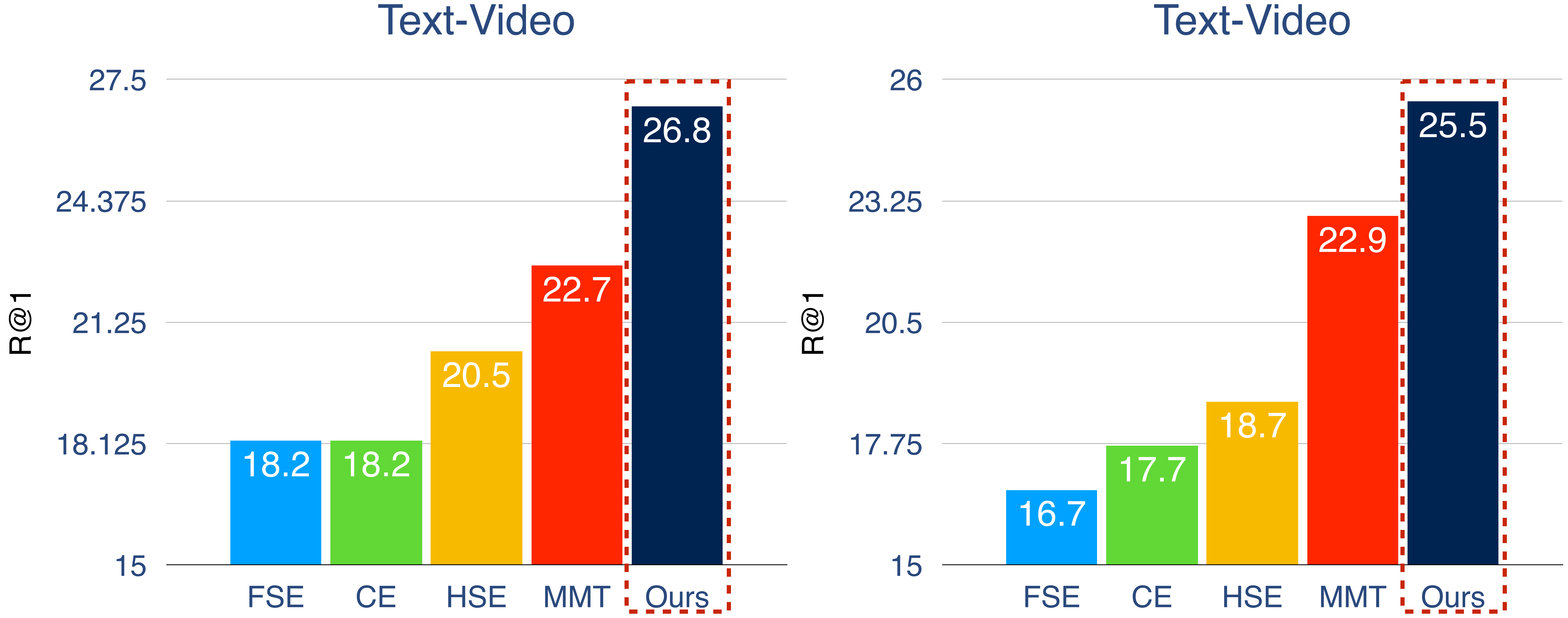
[Kiros et al., arXiv; Faghri et al., BMVC 2018; Liu et al., BMVC 2019; Mithun et al., ICMR 2018]

SOTA on VATEX Text-Video Retrieval

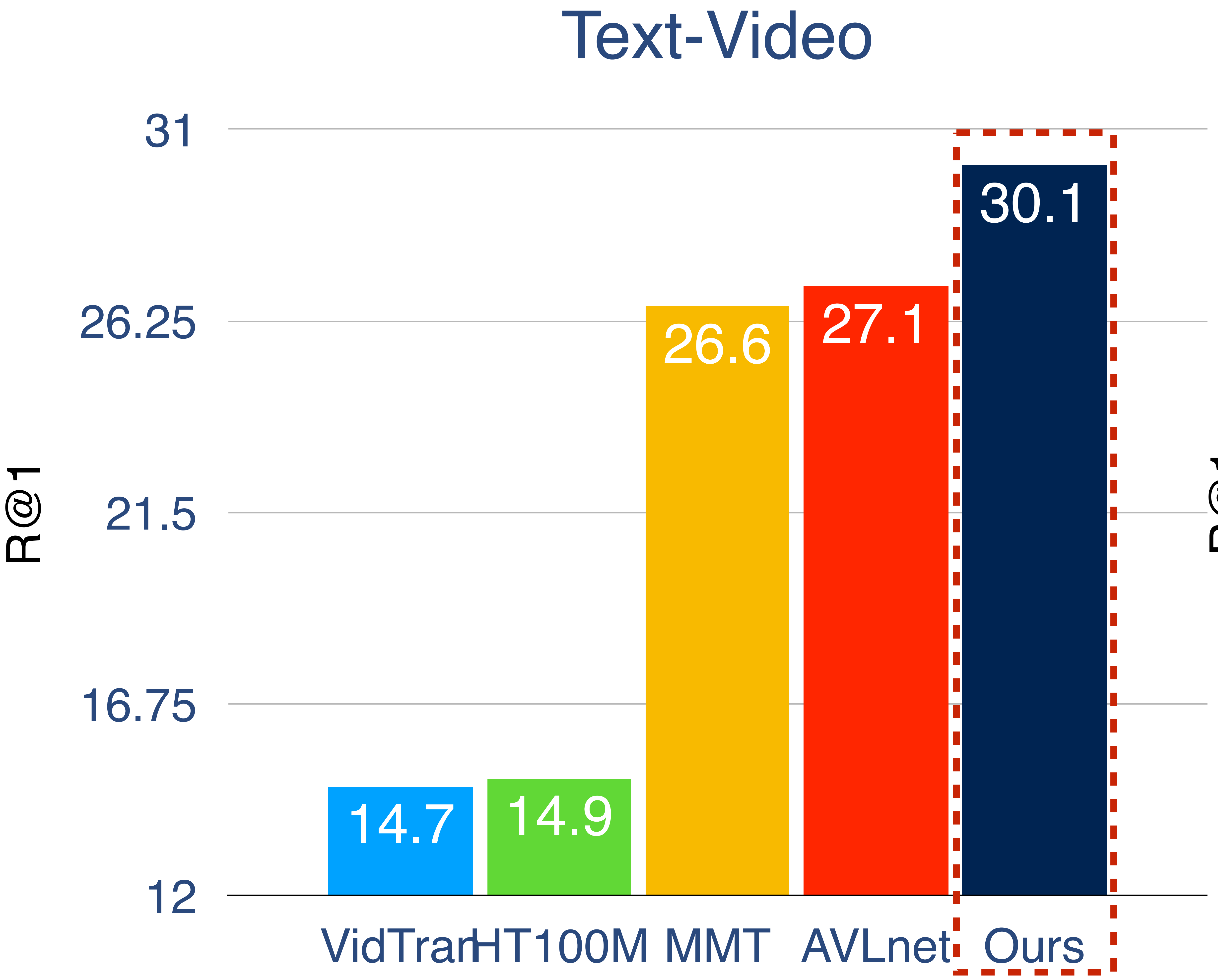


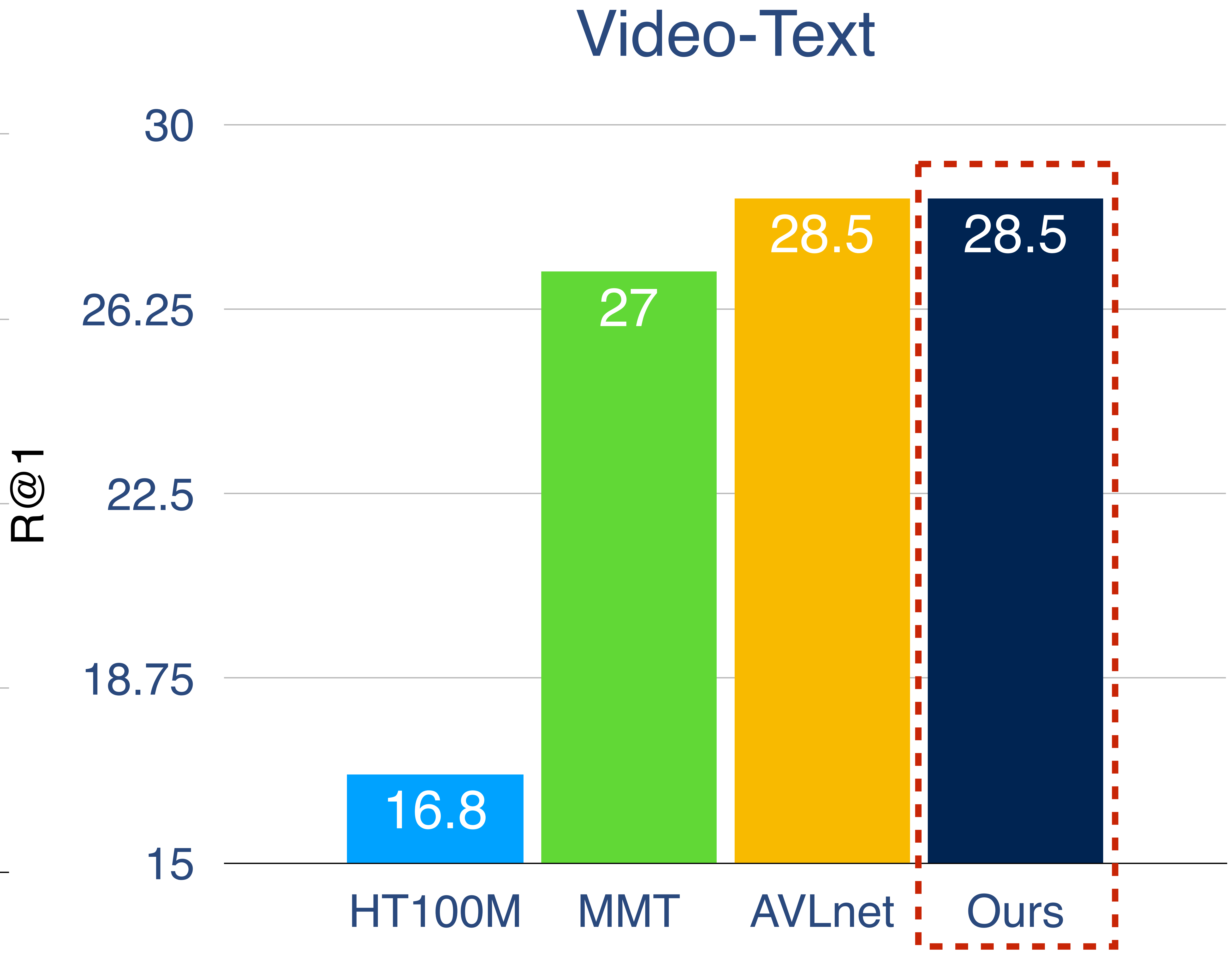
[Kiros et al., arXiv; Dong et al., CVPR 2019; Faghri et al., BMVC 2018; Chen et al., CVPR 2020]

SOTA on ActivityNet Text-Video and Video-Text Retrieval



SOTA on MSR-VTT Text-Video and Video-Text Retrieval



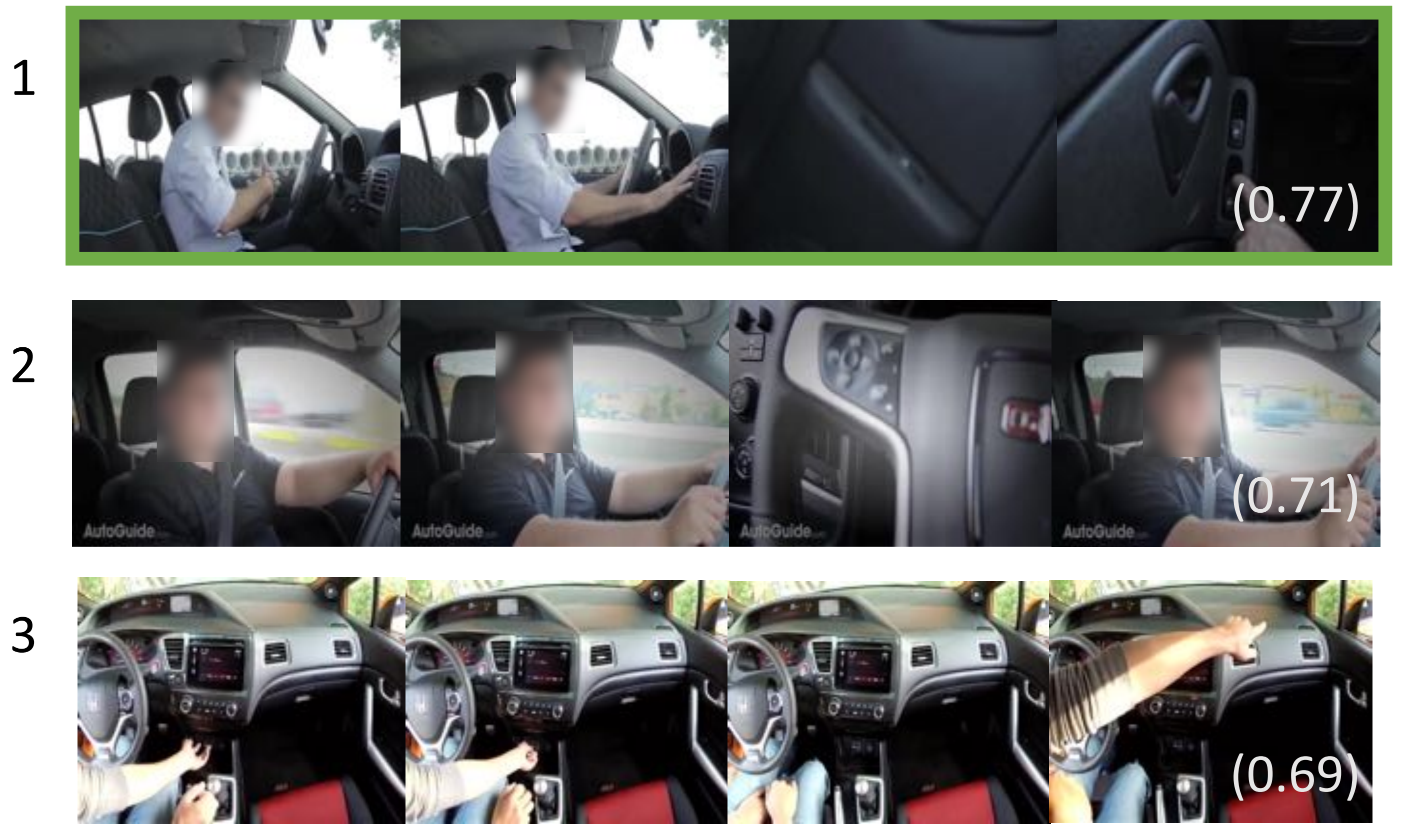


[Bruno et al., arXiv; Miech et al., ICCV 2019; Gauber et al., ECCV 2020; Rouditchenko et al., arXiv]

Qualitative Results and Limitations

A Person is swimming in some white water rapids.

A man is showing the interior of a car.



A Jeep or other off-road vehicle is driving slowly through a very narrow valley without any road

Conclusion

Noise contrastive learning uses instance discrimination to learn effective representations.

Instance discrimination naturally produces faulty negatives that hurt representations.

We propose to alleviate this using a generative objective that implicitly pulls together semantically related videos.

We set SOTA on all video-text and text-video retrieval benchmarks.