Space Time Crop And Attend: Improving Cross-Modal Video Representation Learning

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

ICCV 2021

* equal contribution

UNIVERSITY OF OXFORD FACEBOOK

Noise contrastive learning

Key idea: *discriminate augmentations from other images. (NPID, MoCo, CMC, SimCLR)*

$$f\left(\text{  }\right) = f\left(\text{  }\right)$$

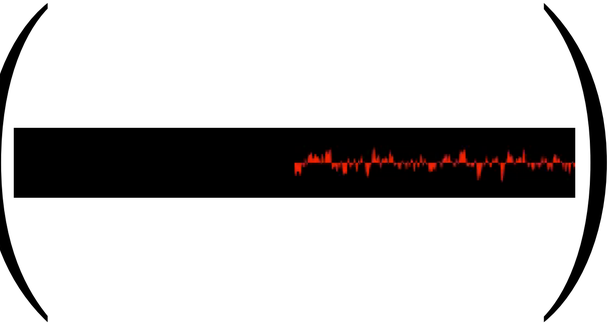
$$\neq f\left(\text{  }\right)$$

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

Images: https://upload.wikimedia.org/wikipedia/commons/thumb/c/c7/Tabby_cat_with_blue_eyes-3336579.jpg/399px-Tabby_cat_with_blue_eyes-3336579.jpg, https://upload.wikimedia.org/wikipedia/commons/thumb/4/47/American_Eskimo_Dog.jpg/306px-American_Eskimo_Dog.jpg

Multi-Modal Noise Contrastive Learning

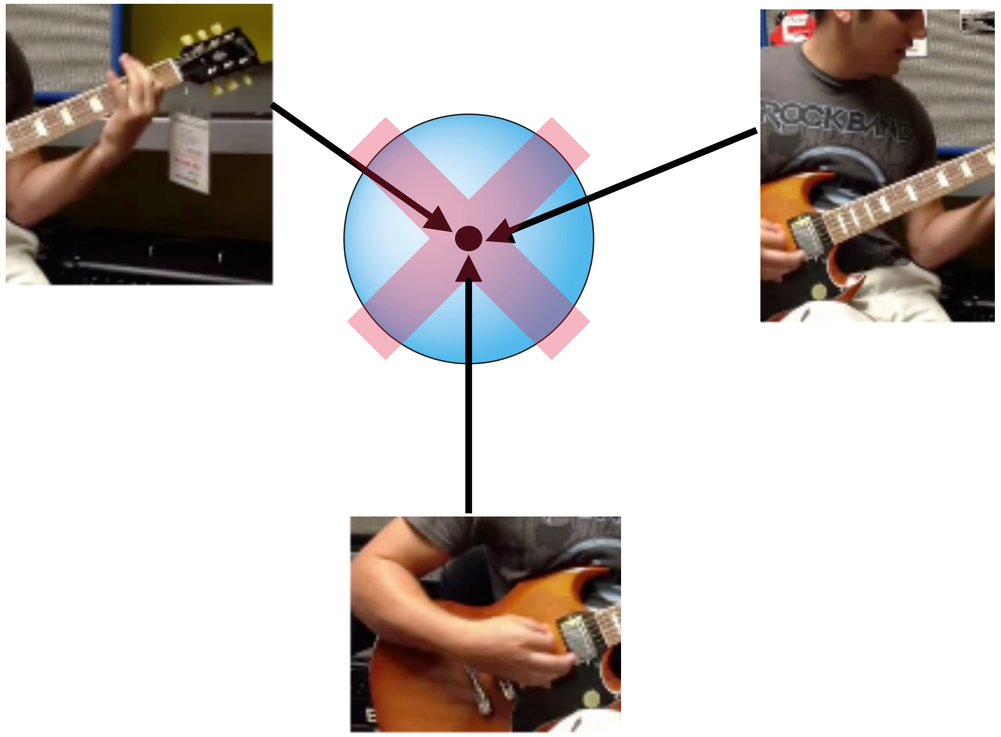
Key idea: *discriminate cross-modal pairs. (GDT, AVID, MMV)*

$$f\left(\text{  }\right) = f\left(\text{  }\right)$$

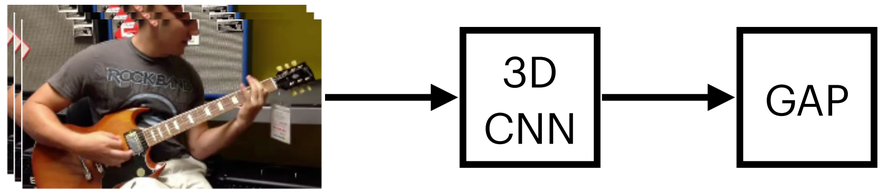
$$\neq f\left(\text{  }\right)$$

[Patrick et al., ICCV 2021; Morgado et al., CVPR 2021; Alayrac et al., NeurIPS 2020]

Problems with Multi-Modal Video Contrastive Learning Formulation



Within modal spatial invariance are not learned.



High-level temporal information is discarded.

Contribution 1: Feature-Crop Augmentation

Comparing differently cropped versions of an images improves self-supervised learning (SwAV)

- Expensive for video: extra temporal dimension, additional modalities, larger networks

Feature-Crop: get large number of crops for within-modality noise contrastive comparison.

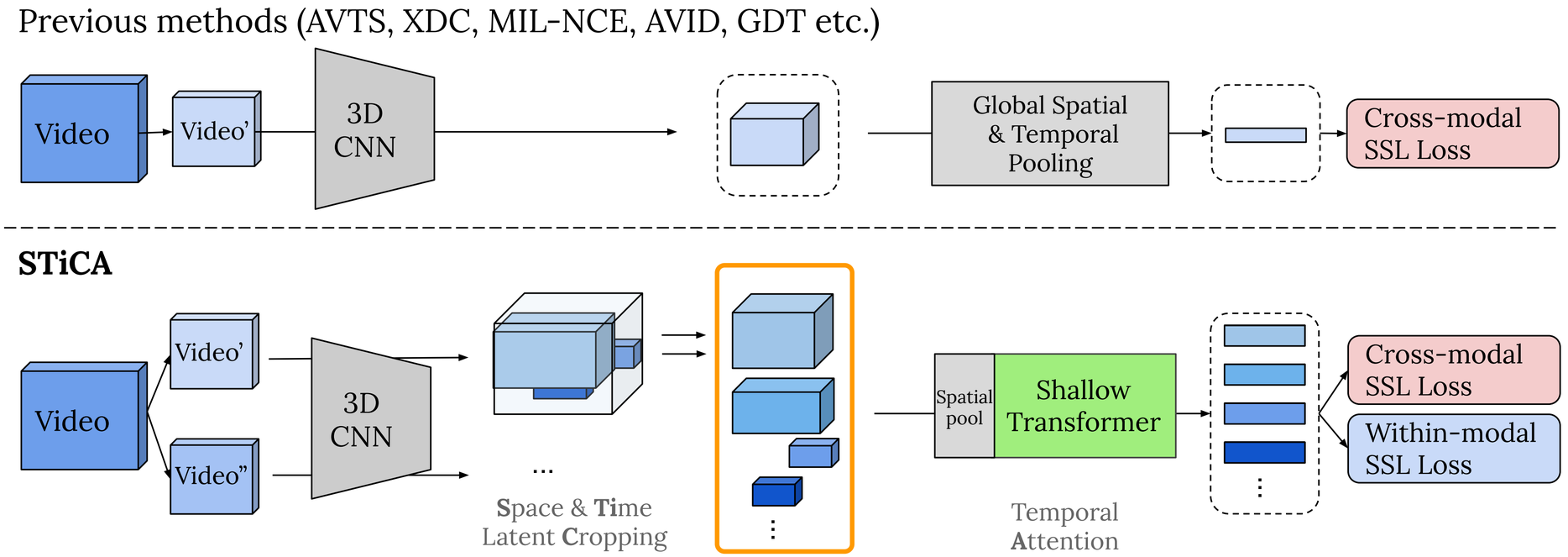
[Caron et al. NeurIPS 2020]

Contribution 2: Transformer for Late Temporal Attention Modelling

- Most video networks (X3D, C3D, R3D, S3D, R(2+1)-D) use **spatio-temporal average pooling** to get fixed length feature vector representation.
  - $\Phi(\tilde{v}) = (\mathscr{P}_t \circ \mathscr{P}_s \circ \Psi)(\tilde{v})$
- We hypothesise that pooling in time is naive, and propose to use a **transformer** for temporal pooling.
  - $\Phi(\tilde{v}) = (\mathscr{P}_{tsf} \circ \mathscr{P}_s \circ \Psi)(\tilde{v})$

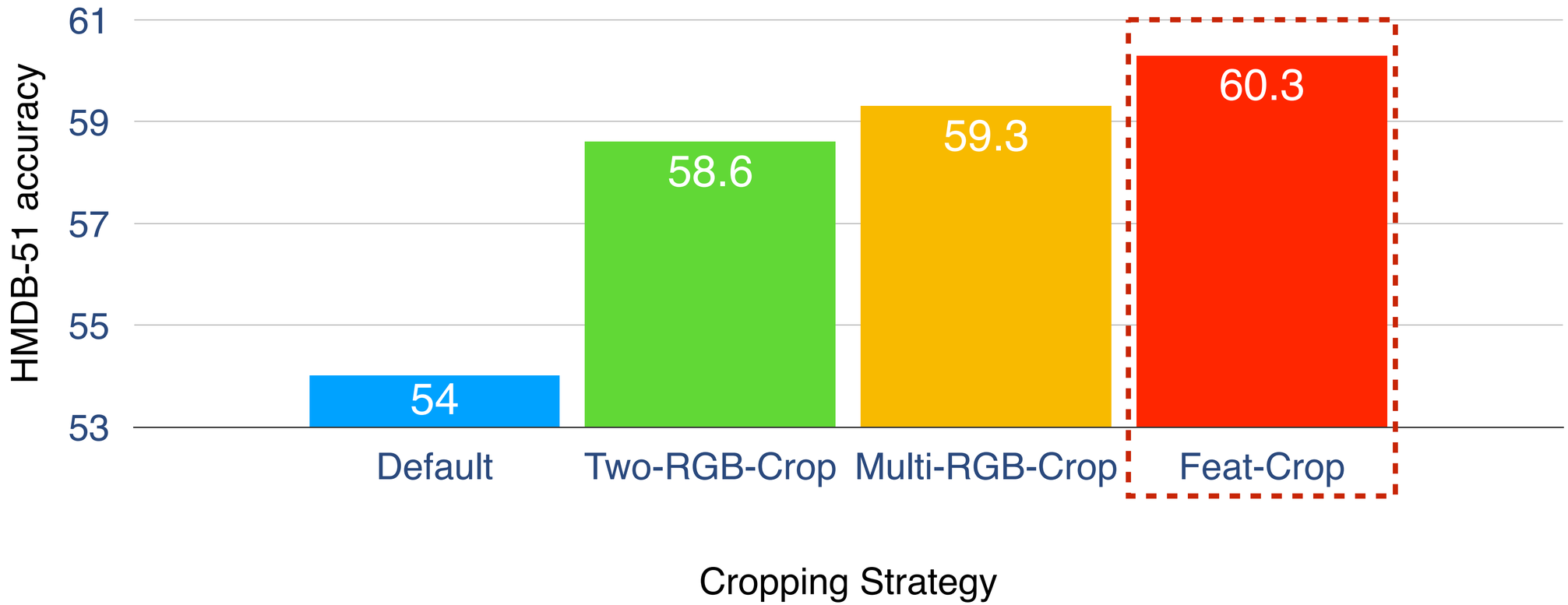
Proposed Approach: StiCA

Our proposed approach, **StiCA (Space-Time Crop and Attend)**, combines these two contributions to improve cross-modal video representation learning.

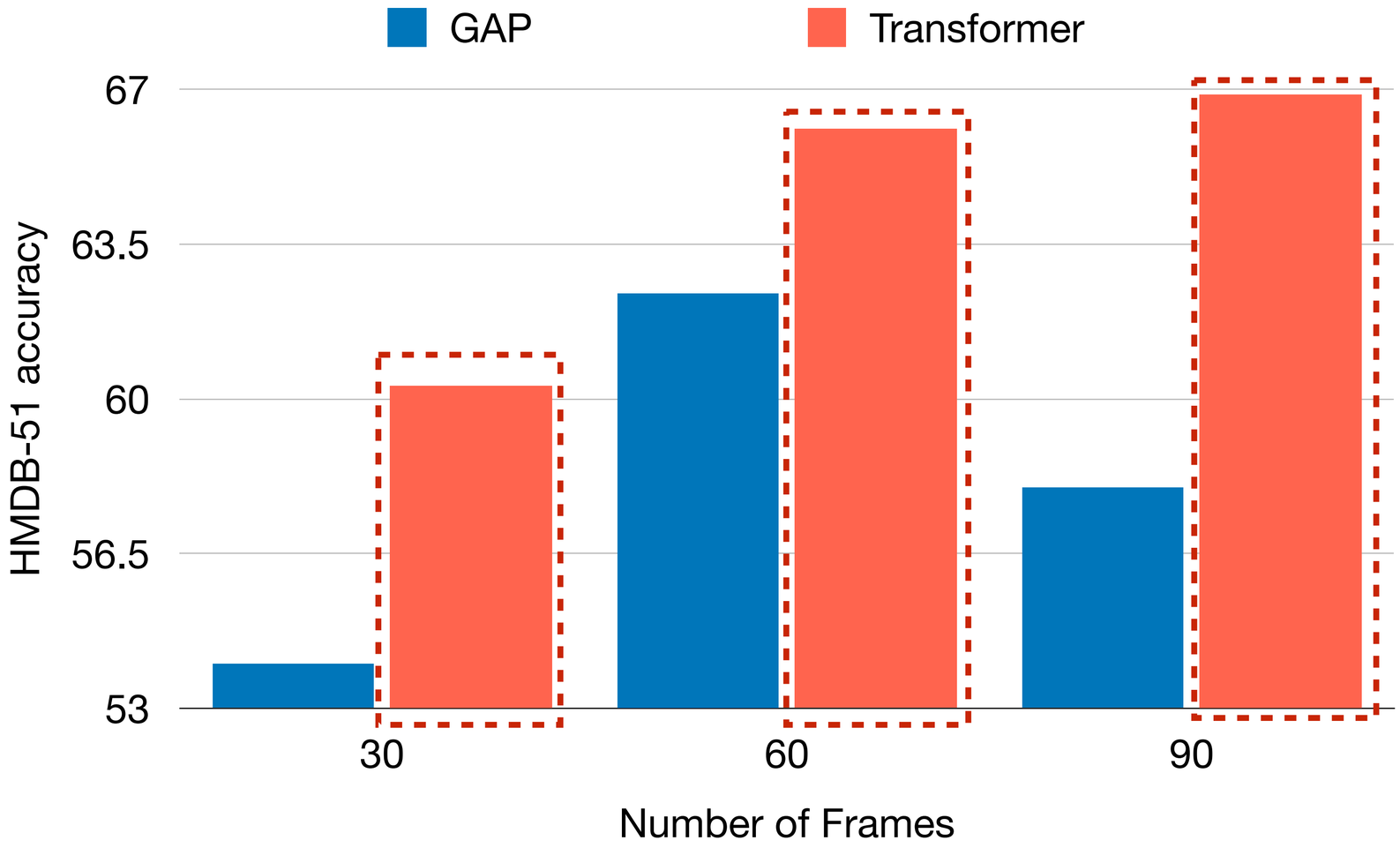


Analysis

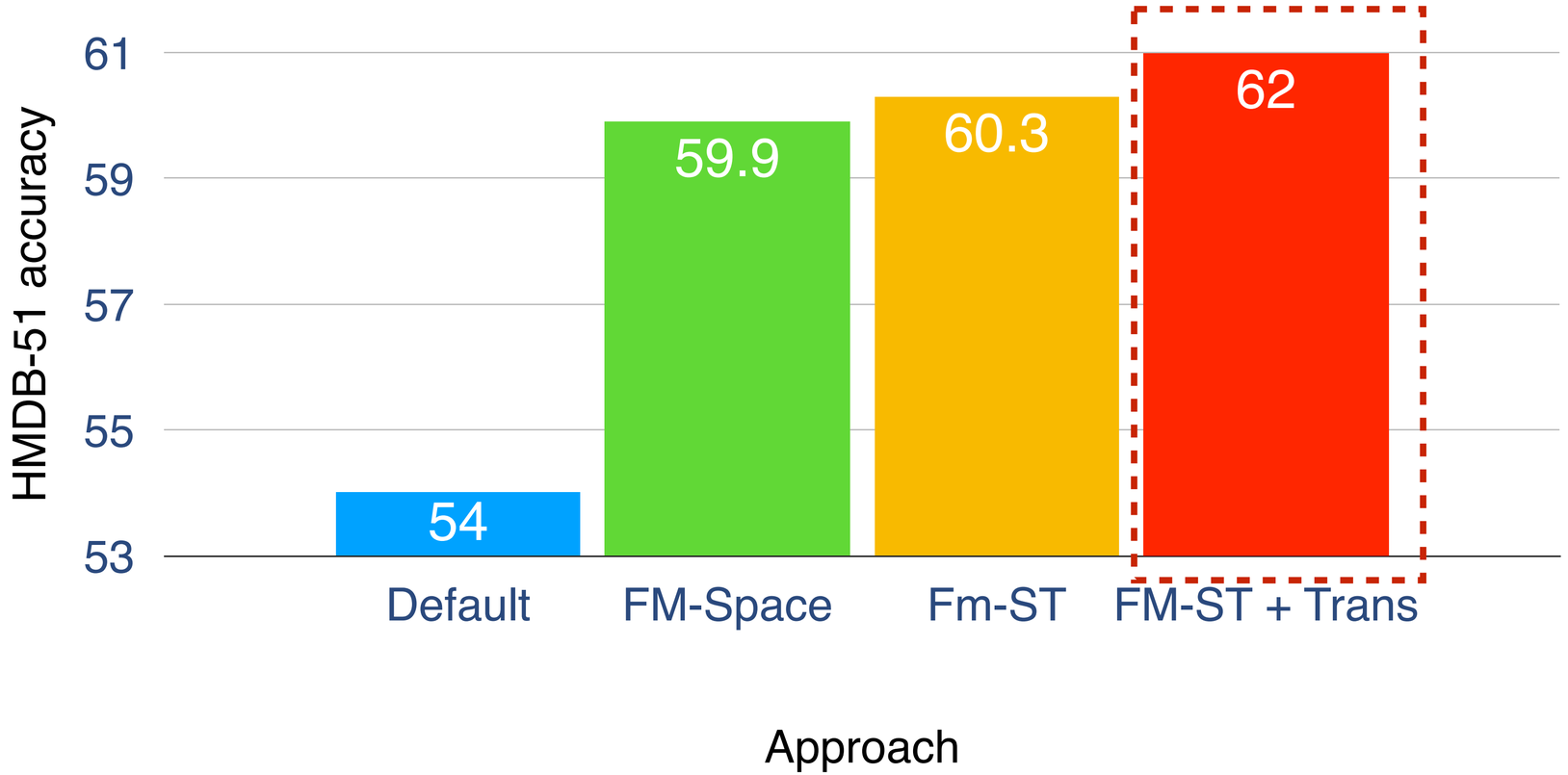
Feature Cropping Improves Video Representation Learning



Transformer works well for late temporal modeling

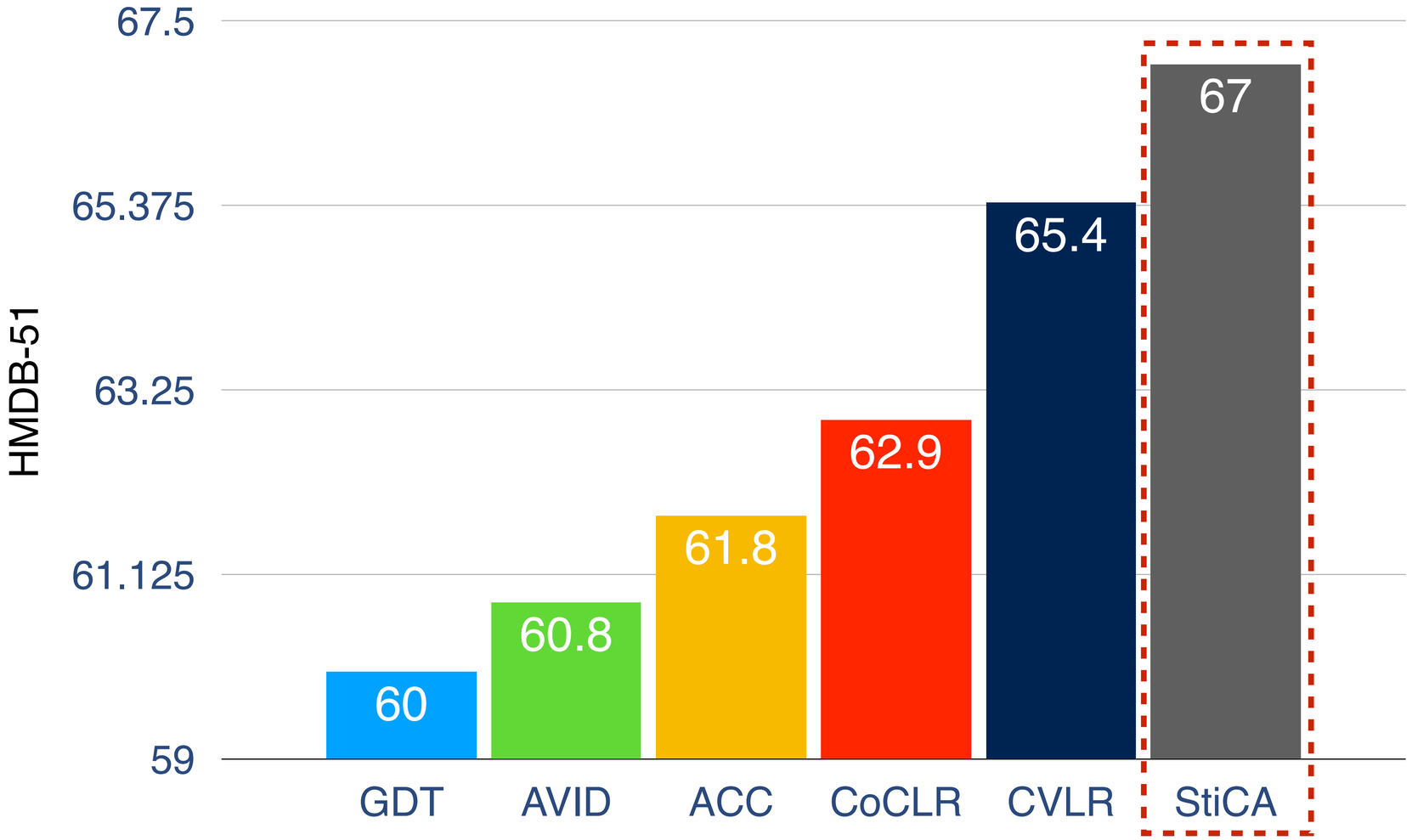


Gains are complementary



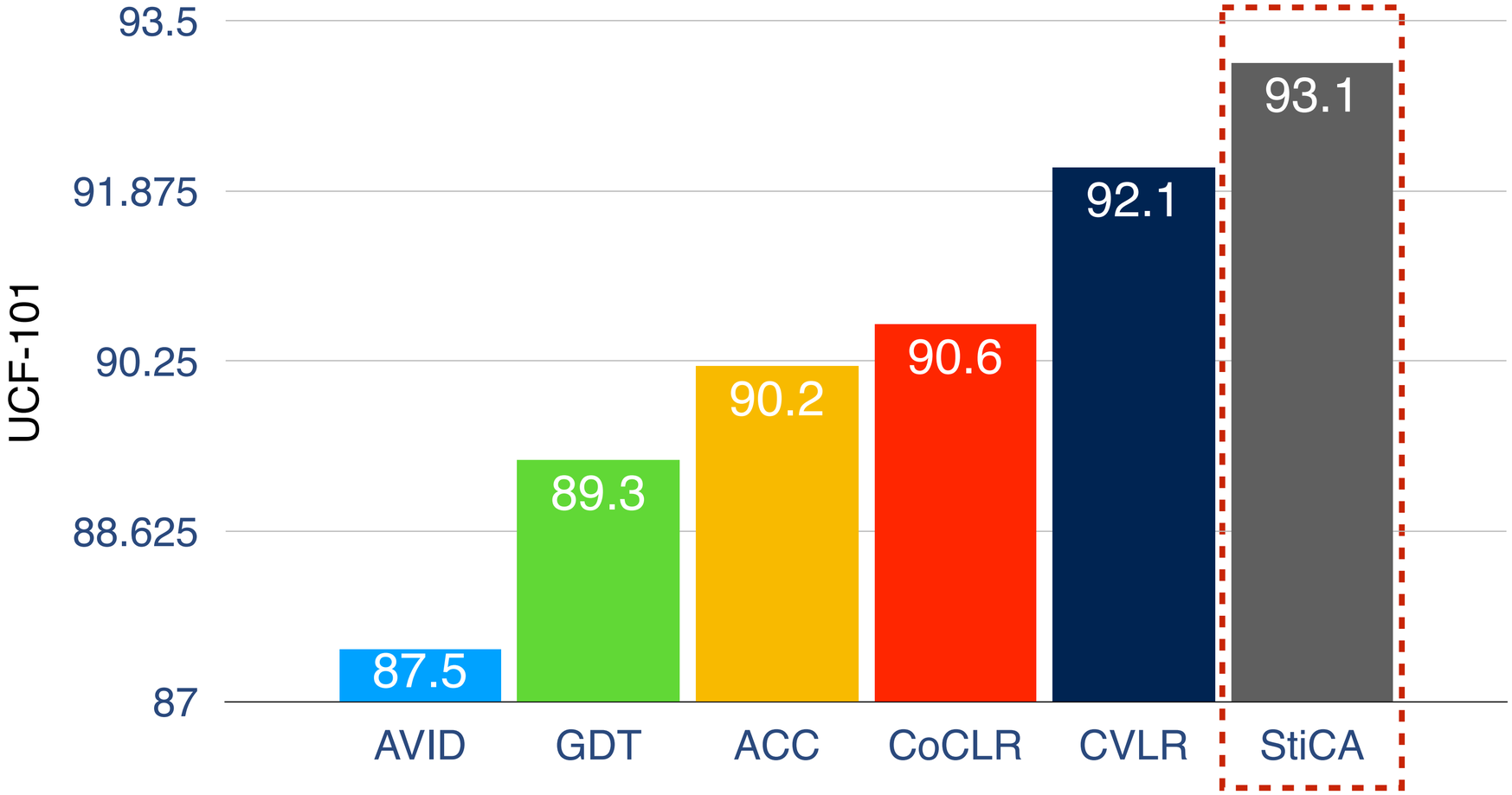
Comparison to State-of-the-Art

SOTA finetuning video-action recognition results: HDMB-51



 [Morgado et al. CPVR 2021; Patrick et al. ICCV 2021; Ma et al. ICLR 2021; Han et al. NeurIPS 2020; Qian et al. CVPR 2021]

SOTA finetuning video-action recognition results: UCF-101



 [Morgado et al. CPVR 2021; Patrick et al. ICCV 2021; Ma et al. ICLR 2021; Han et al. NeurIPS 2020; Qian et al. CVPR 2021]