Unified Compositional Query Machine with Multimodal Consistency for Video-based Human Activity Recognition

Tuyen Tran, Thao Minh Le, Hung Tran, Truyen Tran
Applied Artificial Intelligence Institute (A2I2), Deakin University, Australia

MOTIVATION

▶ *Problem*

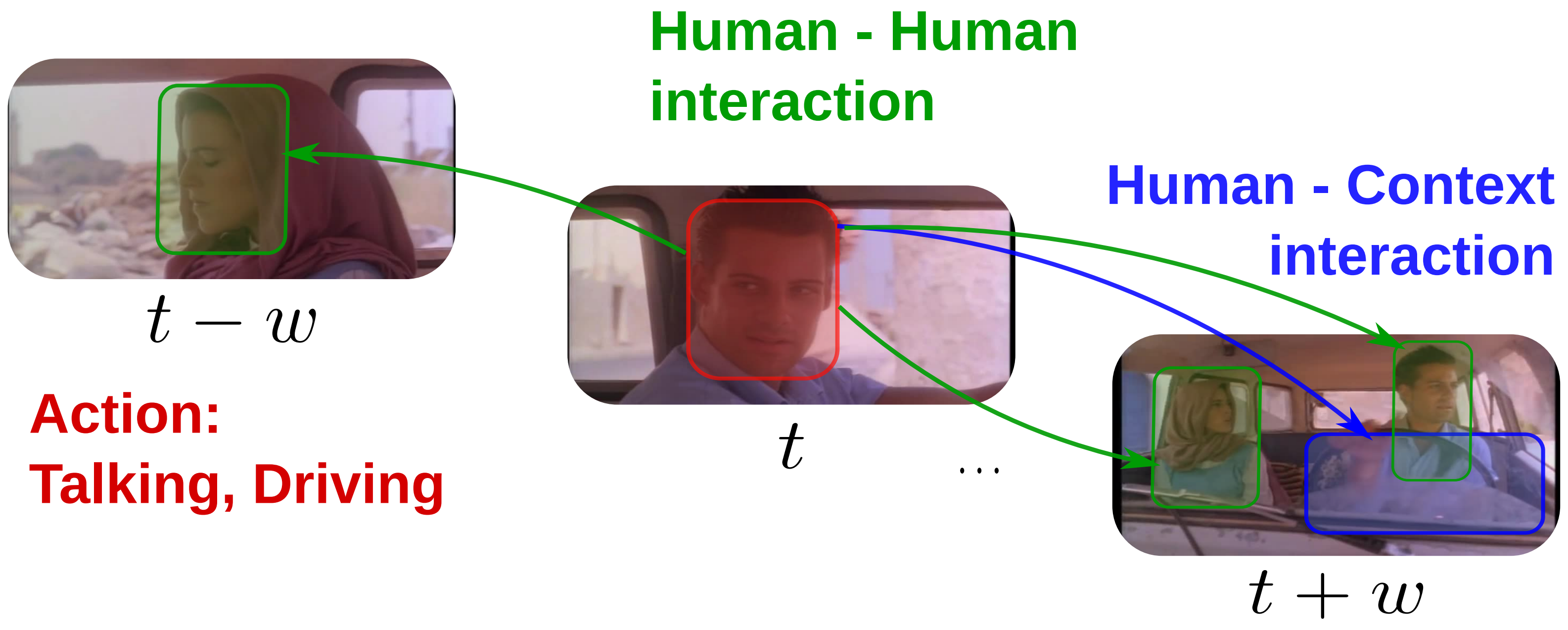
▷ Human Actions Influenced by Spatio-Temporal Interaction

▷ Inherent disparities between modalities

▶ *Research Aim*

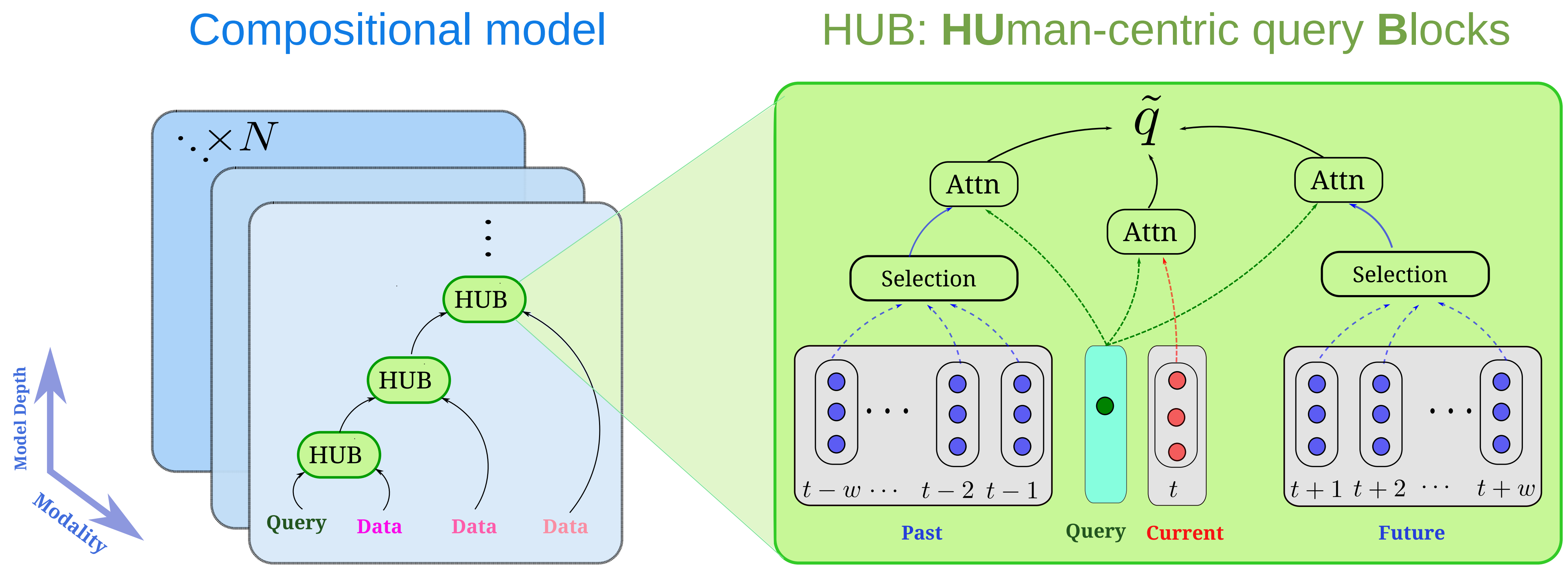
▷ Modeling the spatio-temporal complexity of human interactions

▷ Enforce cross-modal consistency through self-supervision



METHOD

▶ *Modeling the spatio-temporal complexity of human interactions*



▷ Operate at actor level features

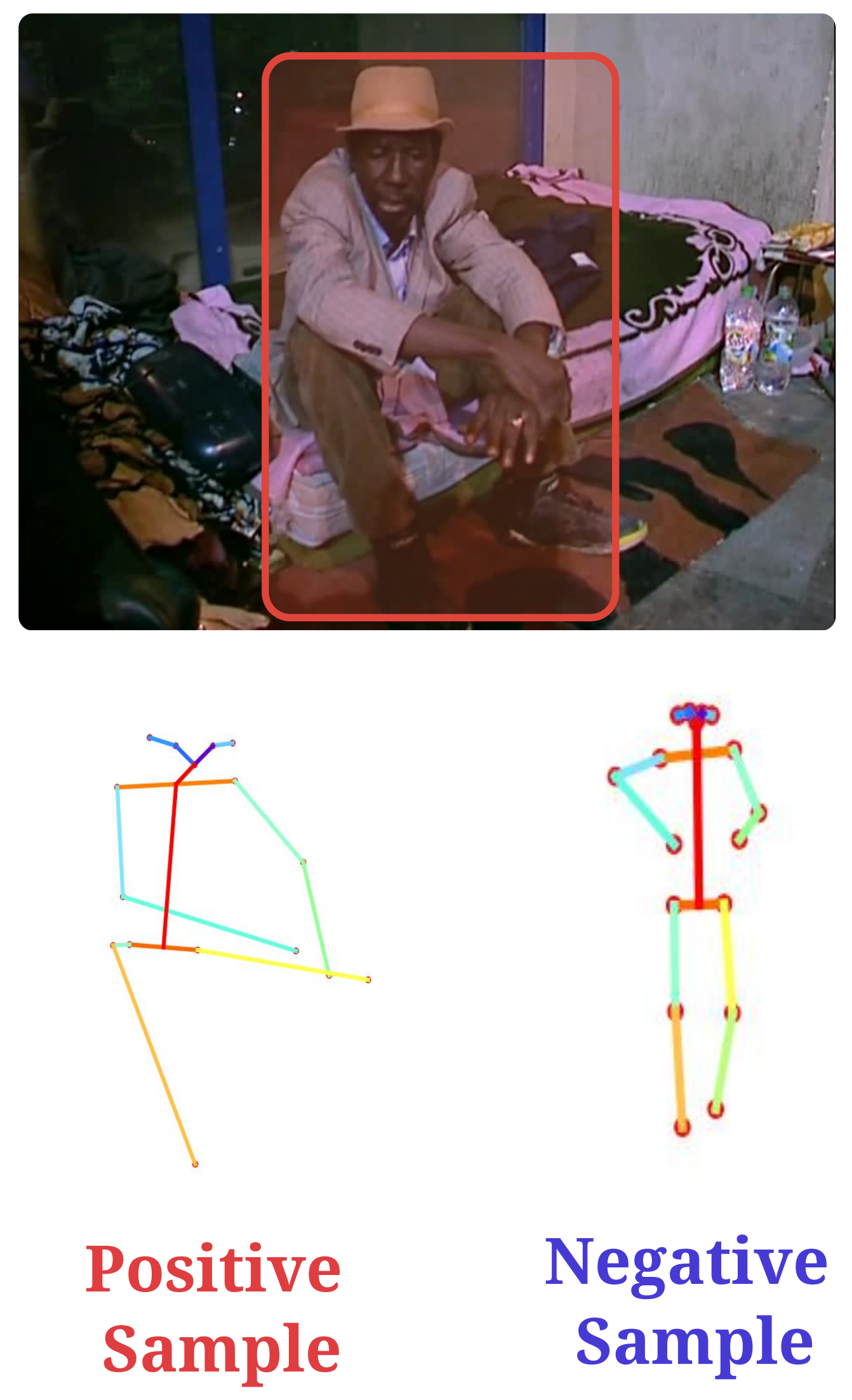
▷ Handling different kind of interaction using single computational unit

▷ Extenable in modalities and model depth

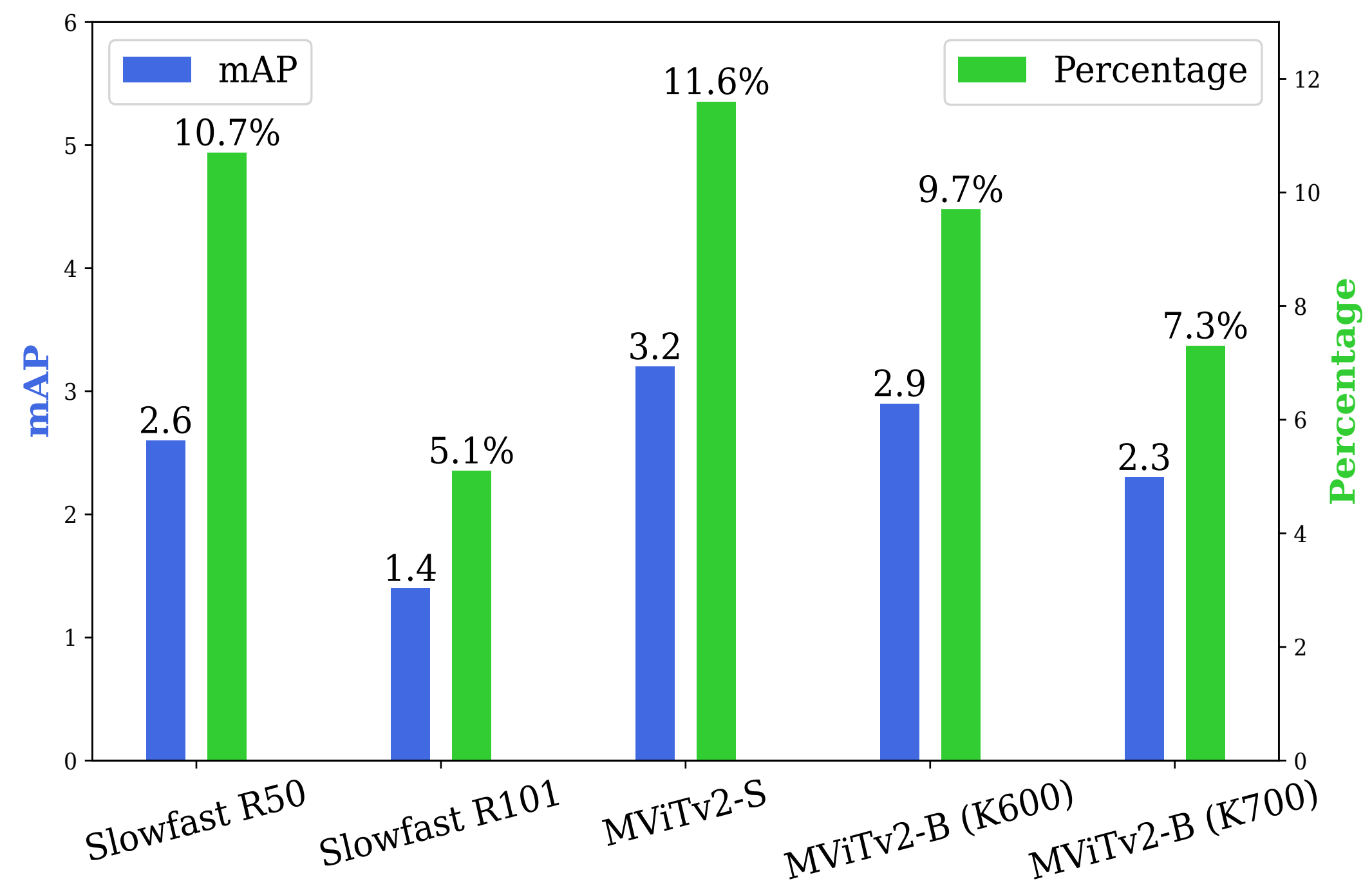
▶ *Enforce cross-modal consistency through self-supervision*

▷ Auxiliary self-supervised loss:

$$\mathcal{L}_{\mathrm{CC}} = -\log \frac{\exp\left(\mathrm{sim}\left(\hat{q}^{\mathrm{vis}}_{i,t}, \hat{q}^{\mathrm{key}}_{i,t}\right)\right)}{\sum_{k=1}^{B} \mathbb{I}_{[k\neq i]} \exp\left(\mathrm{sim}\left(\hat{q}^{\mathrm{vis}}_{i,t}, \hat{q}^{\mathrm{key}}_{k,t}\right)\right)}$$



RESULTS



Improvement over baseline: blue and green bars show point and percentage gains

Method	mAP
SlowFast [ICCV19]	23.8
ORViT [CVPR22]	26.6
MemViT [CVPR22]	29.3
MViTv1-B [ICCV19]	27.3
MViTv2-S [CVPR22]	27.6
MViTv2-B [CVPR22]	29.0
COMPUTER	**30.8**

Comparison against other methods on AVA dataset.

Paper **Code**

If you have any question, feel free to contact ***t.tran@deakin.edu.au***