Is Imagenet Worth 1 video? Learning Strong Image Encoders From 1 Long Unlabelled Video

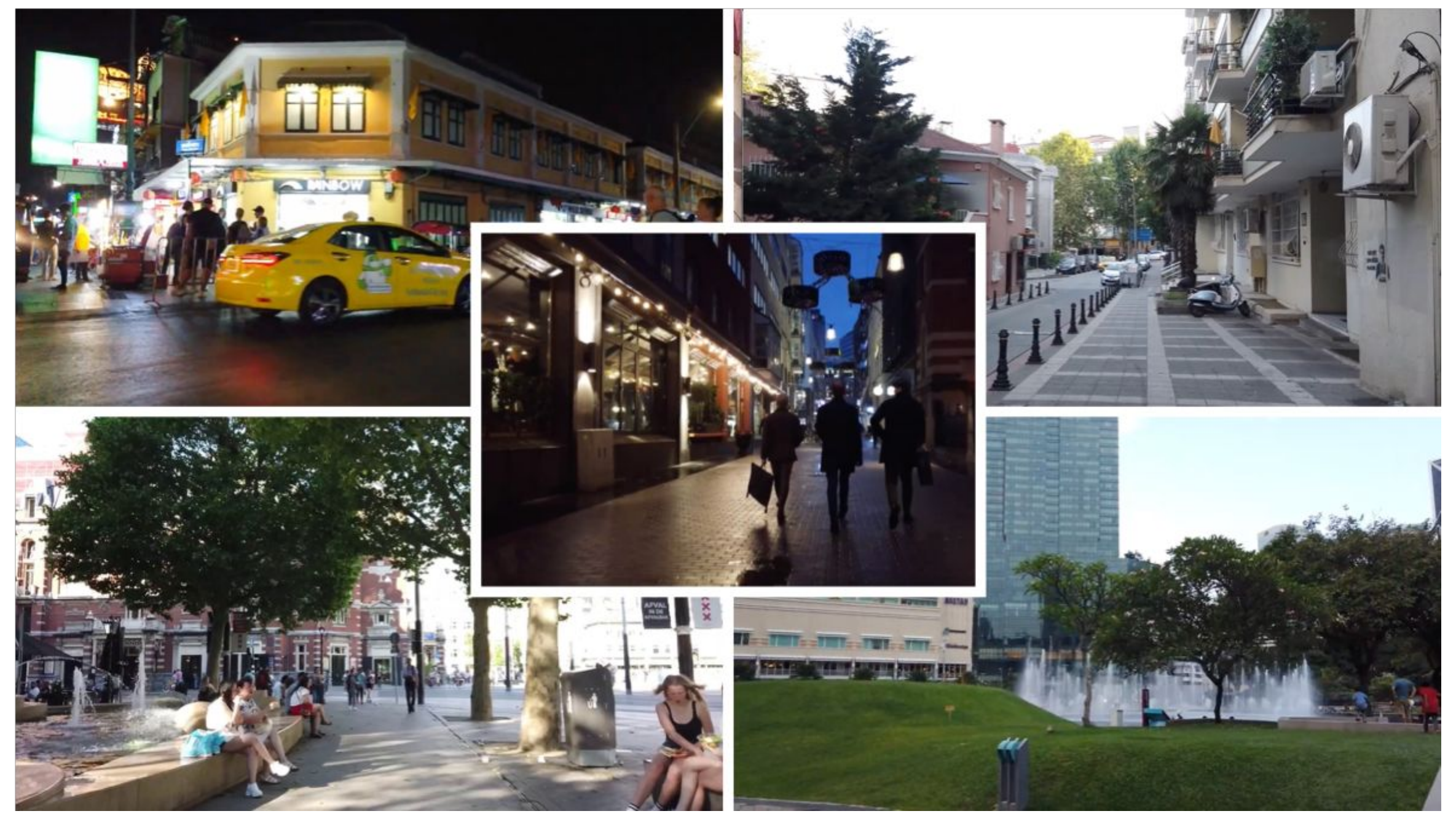
Shashanka Venkataramanan, Mamshad Rizve, João Carreira, Yuki M. Asano*, Yannis Avrithis*

We train a self-supervised ViT with **1 video** from scratch and **outperform DINO**

Paper, Dataset & Code

Which video do we use? 🤔

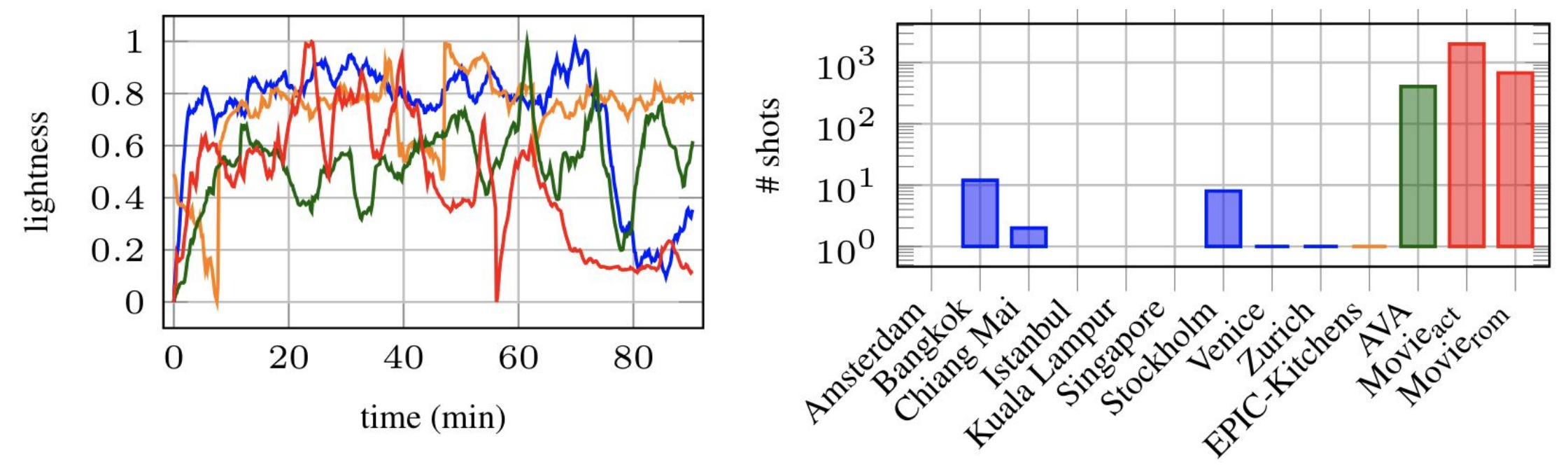
- New dataset of open-source first-person video for the purpose of virtual **Walking Tours**
- 10 x 4K videos, Avg. duration: 1 hr 38 min, **License - CC-BY**



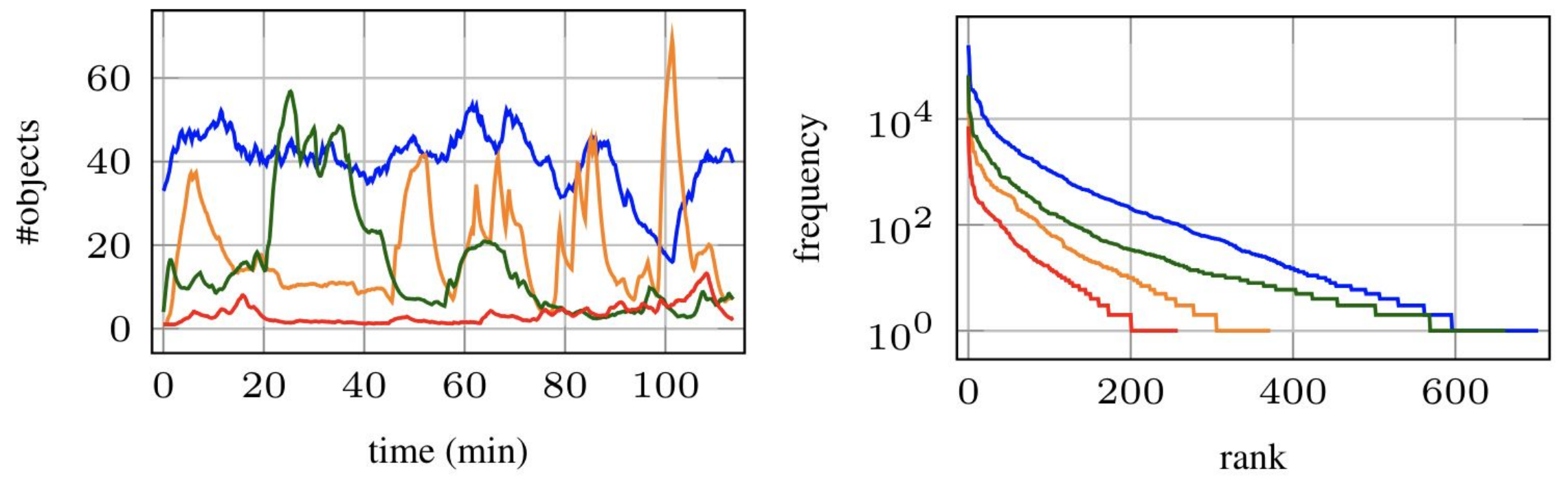
Interesting properties of WTours

Lightness measure

Scene changes

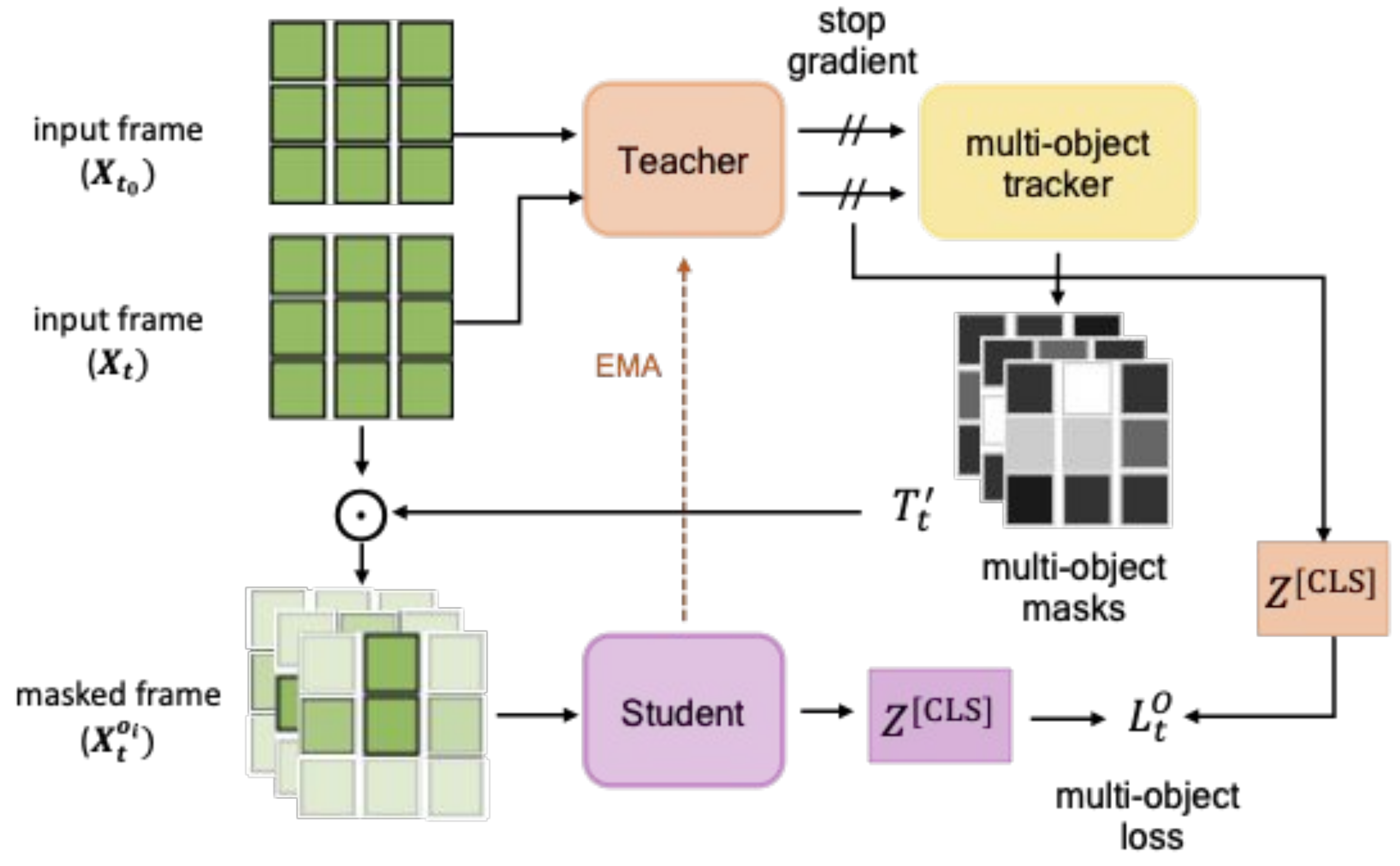


Large # objects and classes

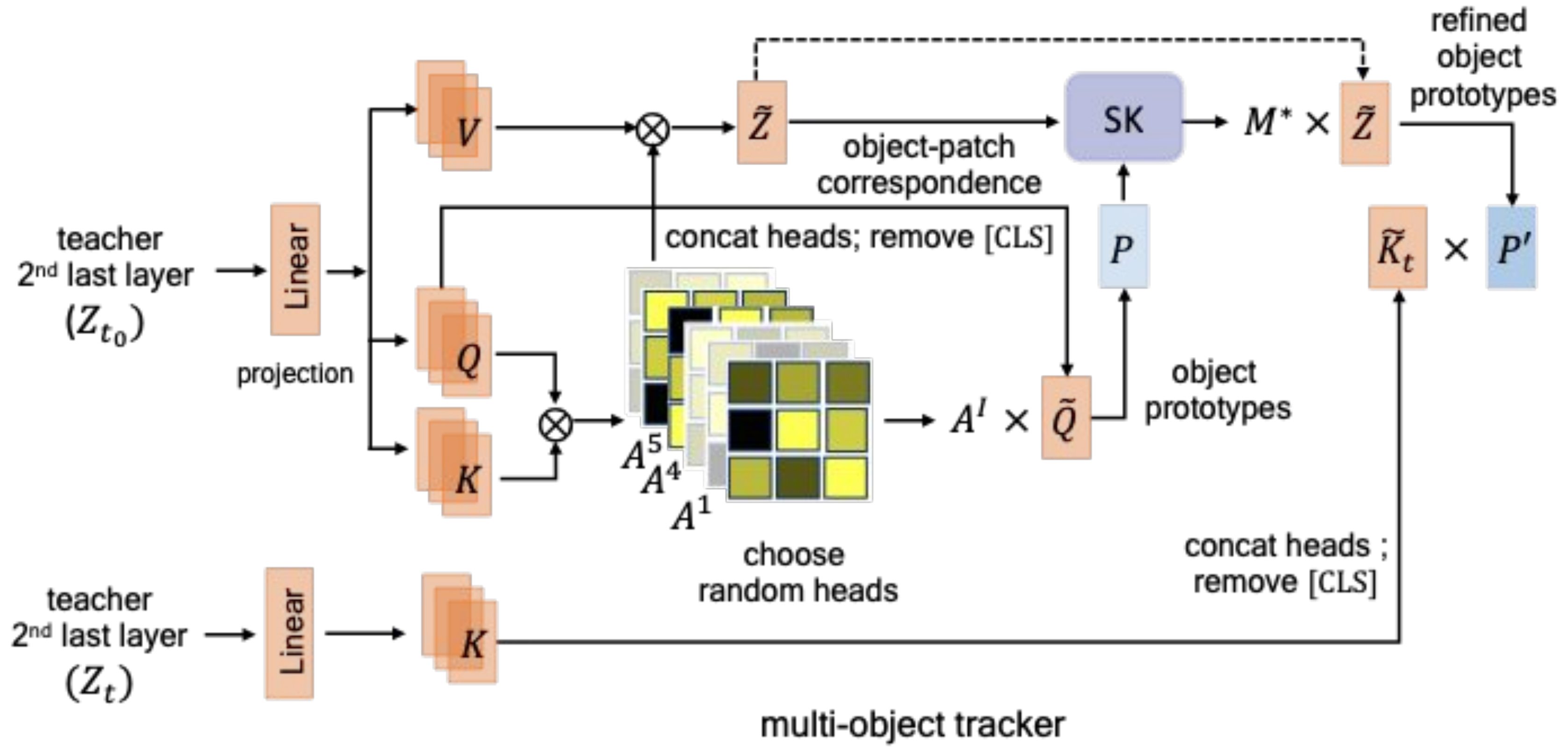


DoRA: Discover and tRAck

- We introduce **DoRA**, a new SSL image-pretraining method, aimed at learning from video frames.

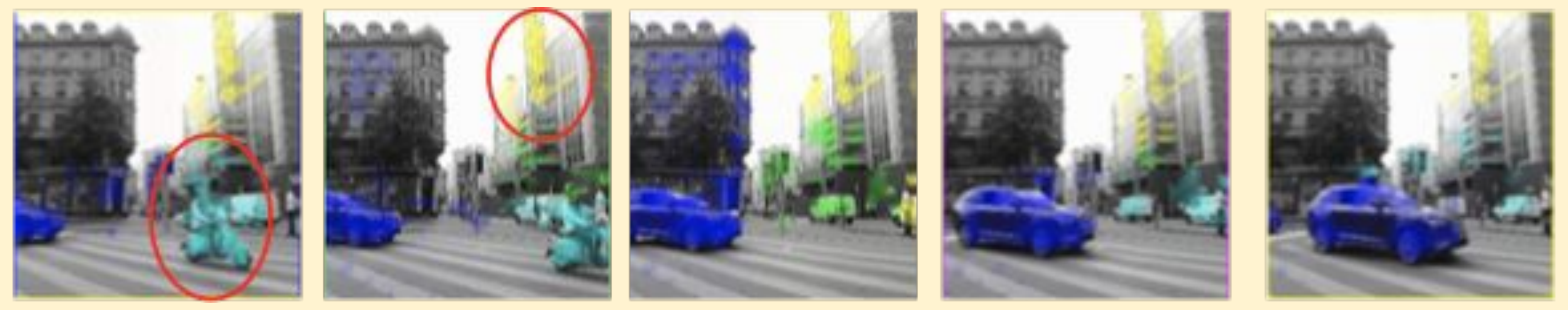


- Our SSL framework **discovers** and **tracks** objects over time in an end-to-end manner, using transformer **cross-attention**.

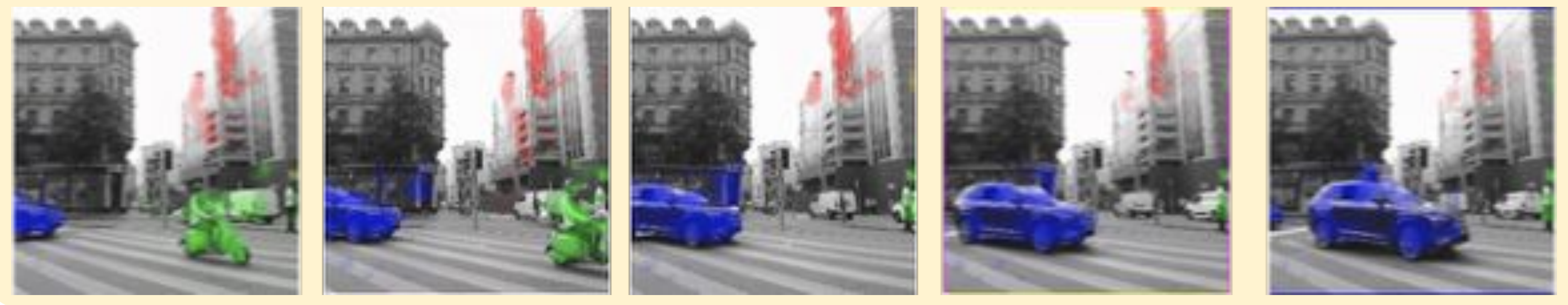


DoRA leads to emergent **tracking of objects** as ViT attention masks, even through **occlusions**

w/o Sinkhorn-Knopp

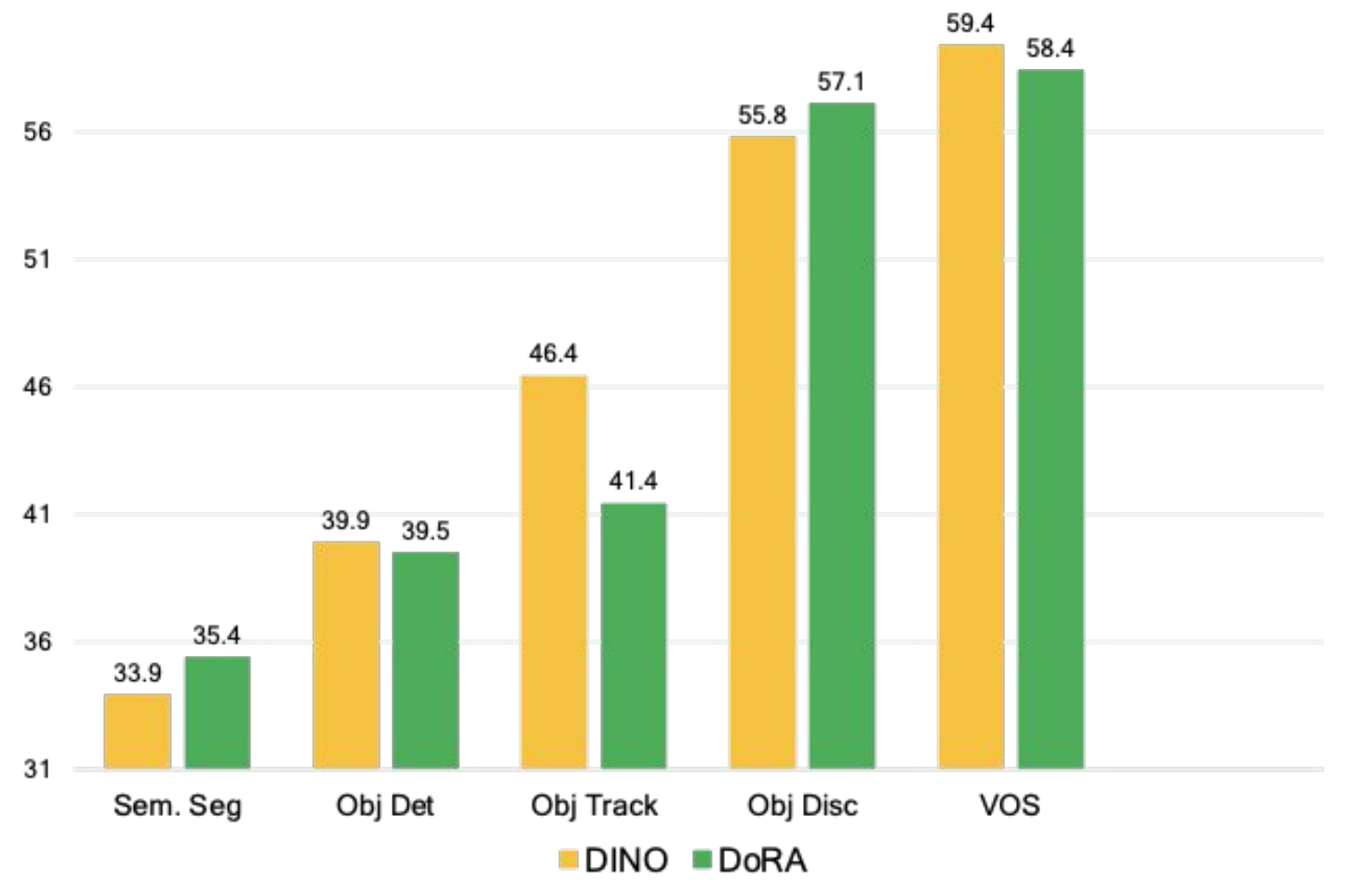


w/ Sinkhorn-Knopp

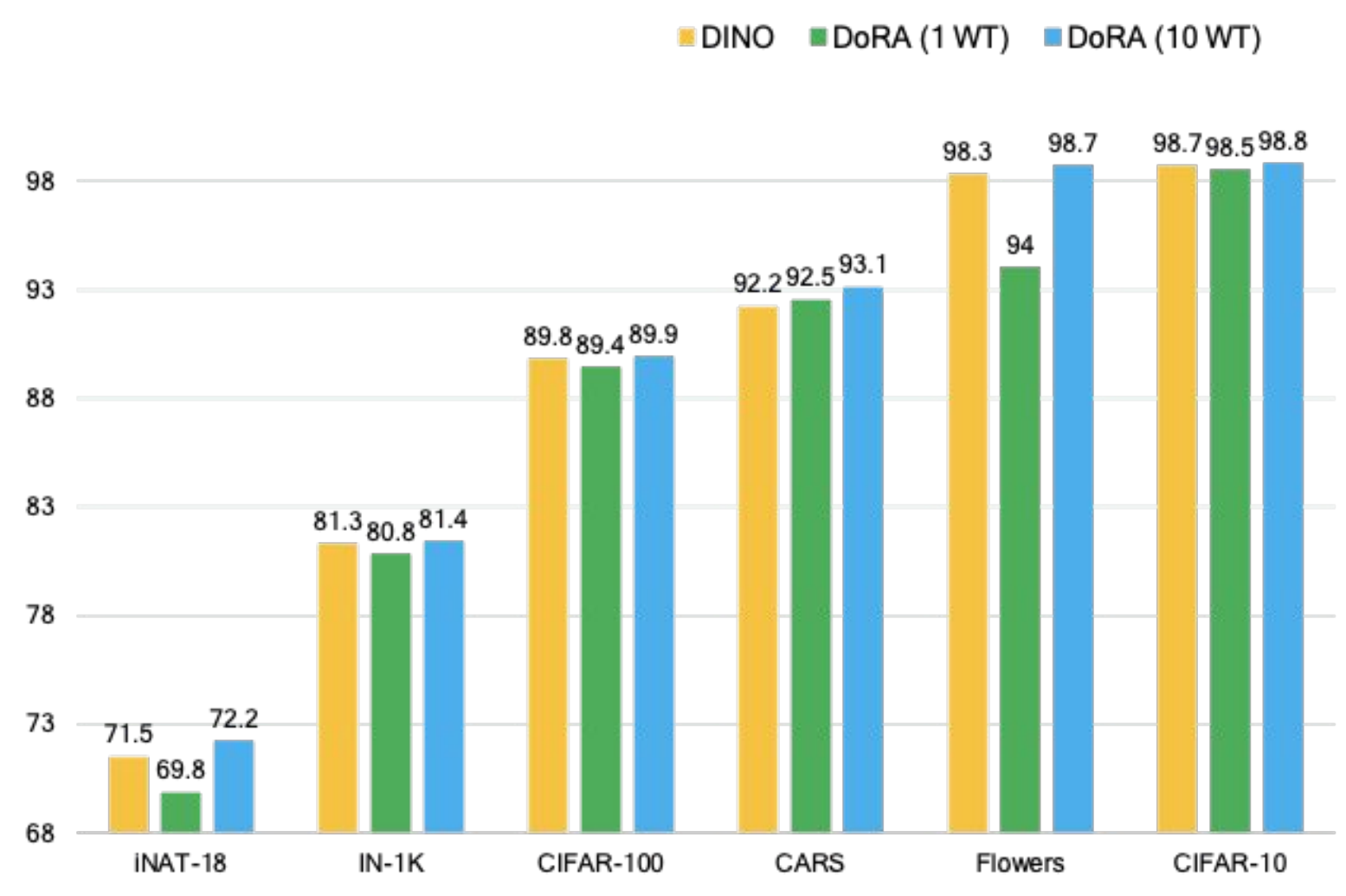


1 video better than ImageNet pretraining

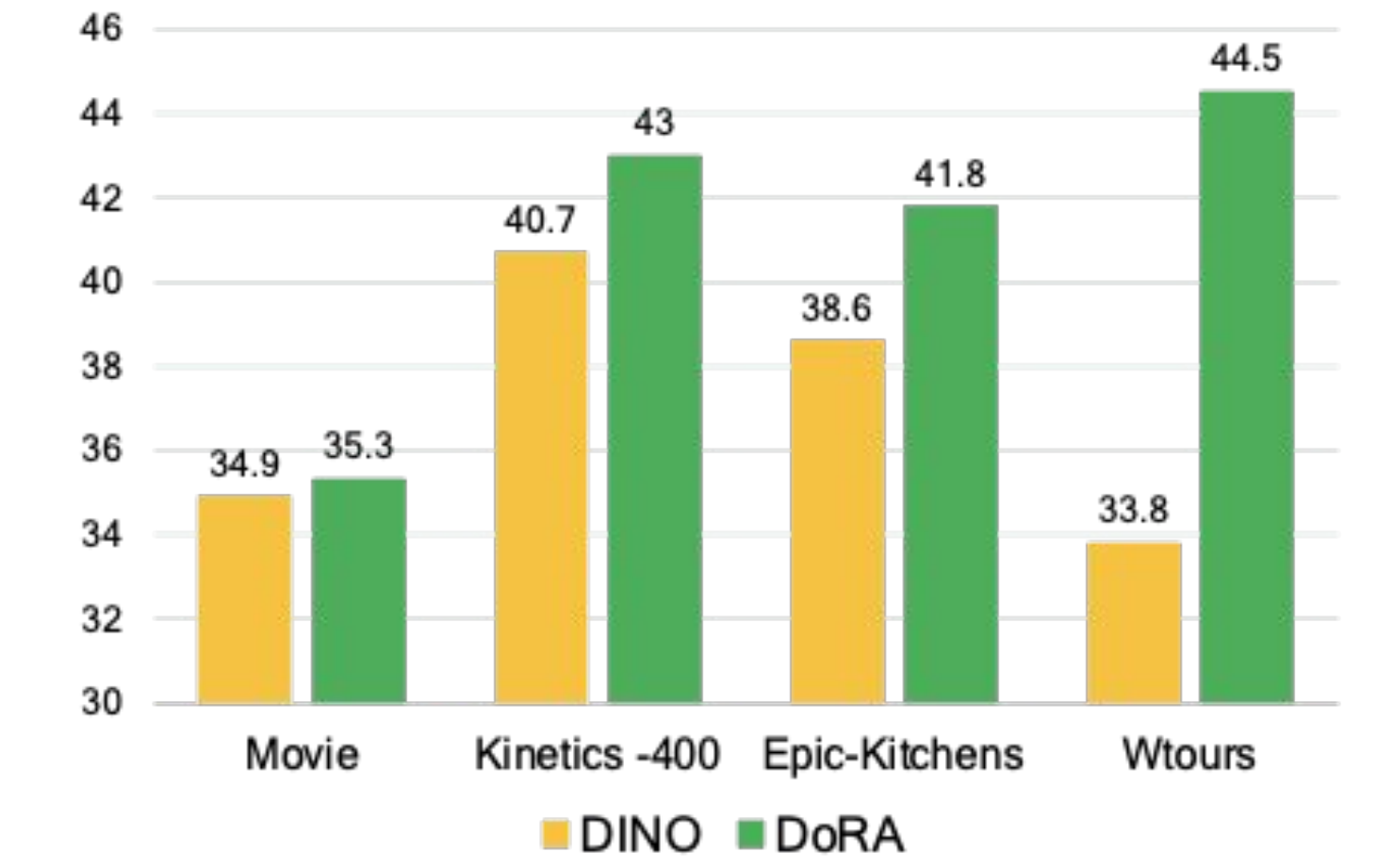
Image and video tasks

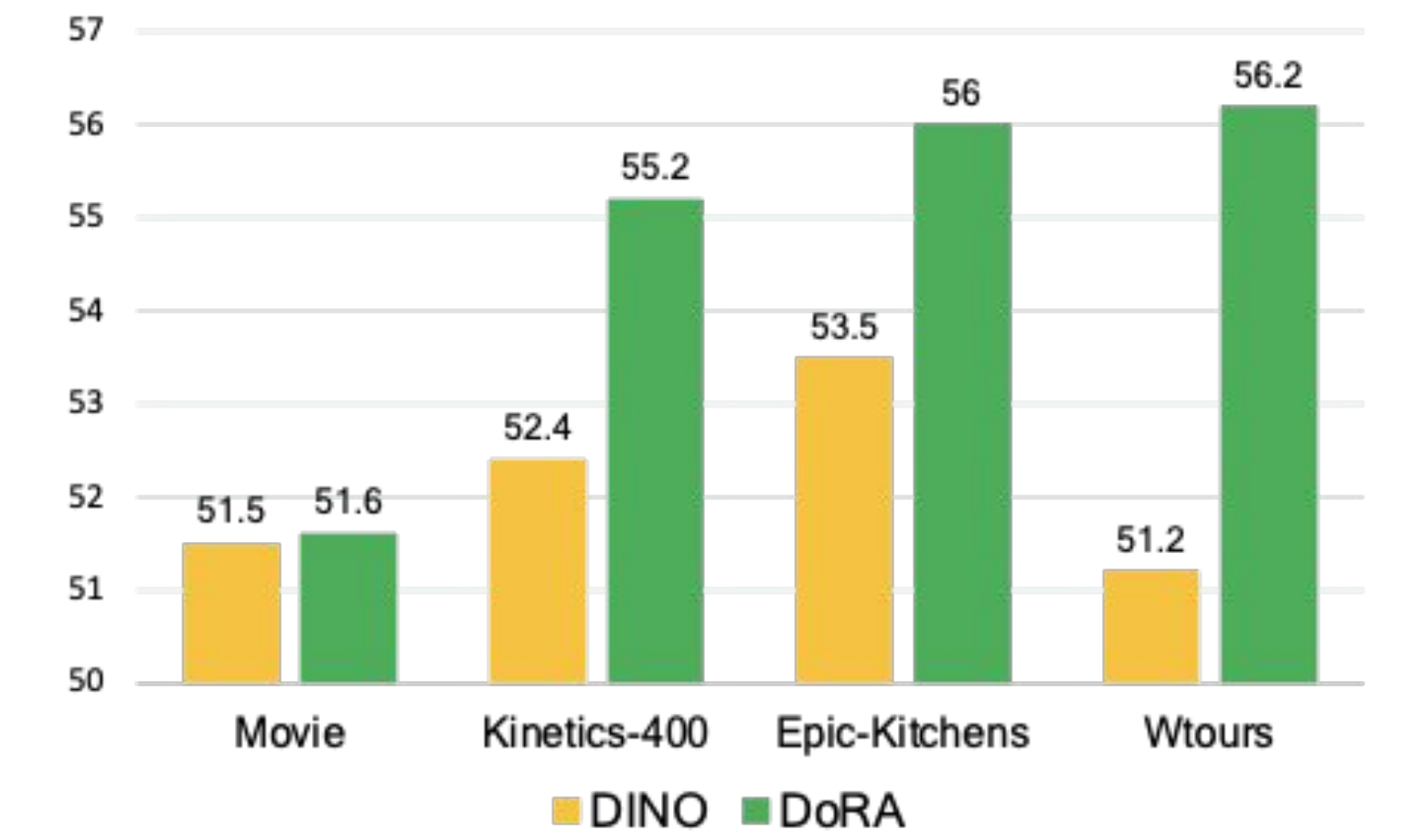


Finetuning



Different videos


Movie
Kinetics -400
Epic-Kitchens
Wtours
DINO
DoRA


Movie
Kinetics-400
Epic-Kitchens
Wtours
DINO
DoRA