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


Shashanka Venkataraman


Mamshad Rizve


João Carreira


Yuki M. Asano*


Yannis Avrithis*

Inria
amazon
Google DeepMind
University of Amsterdam
IARAI institute of advanced research in artificial intelligence

[* equal last authors. order random]

Why Self-Supervised Learning is cool!

Scale to billions of images

Avoids problems with labelling

Improved performance on downstream tasks

Do we need billions of images for pretraining?

- Face recognition and color sensitivity developed in three months.
- Depth perception takes five months.
- Visual acuity takes six months.

[de Haan *et al.,* 2001, Adams 1987, Campos *et al.,*1978, Sokol, 1978]

Do we need billions of images for pretraining?

- Face recognition and color sensitivity developed in three months.
- Depth perception takes five months.
- Visual acuity takes six months.

- Humans observe surroundings in one continuous stream, interrupted by sleep.

[de Haan *et al.,* 2001, Adams 1987, Campos *et al.,*1978, Sokol, 1978]

Videos open exciting new direction

Visual development

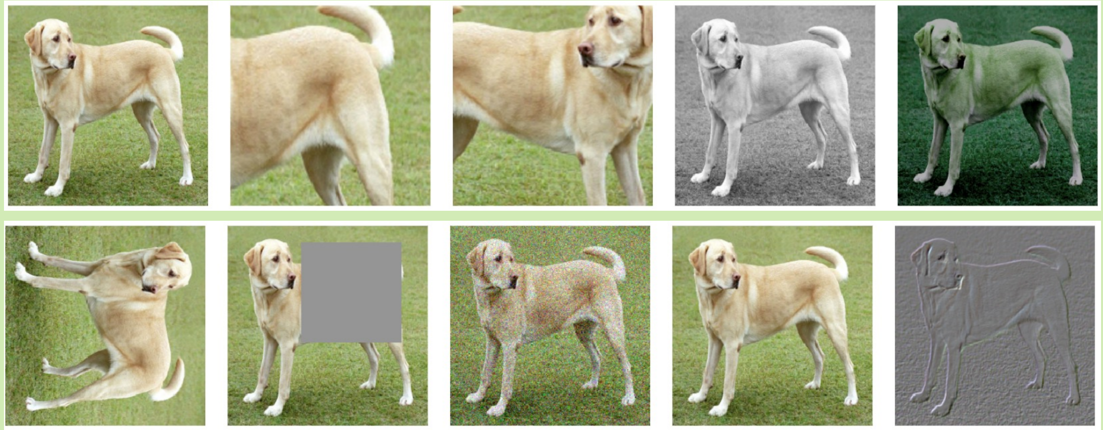
Understanding physics

Embodied AI

Platforms with insane scale

Image *vs.* Video based SSL

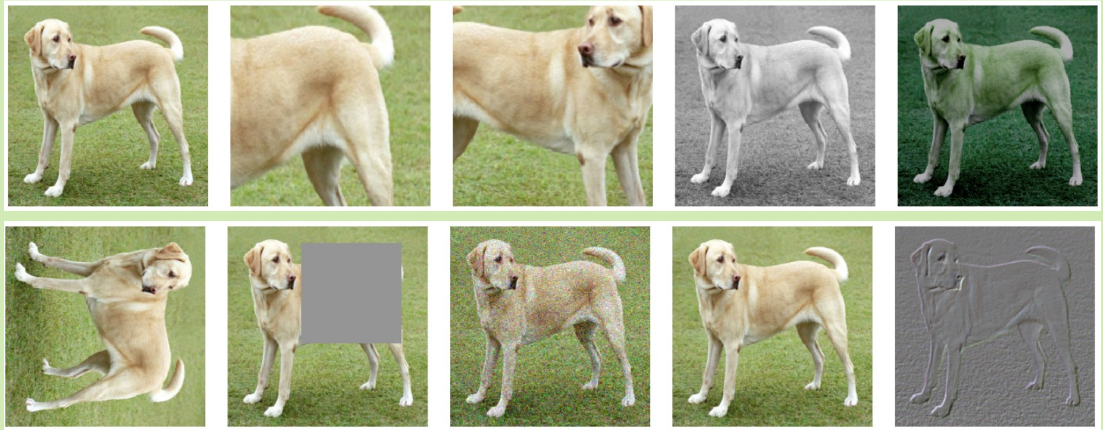
Hand-crafted data augmentations



crop, flip, blur, solarization,
random mask etc.

Image *vs.* Video based SSL

Hand-crafted data augmentations



crop, flip, blur, solarization,
random mask etc.

Natural data augmentations

Object occlusion

Perspective distortion

low-illumination

Learning Image Encoders From Video

- A new dataset of open-source first-person video for the purpose of virtual “walking tours”.

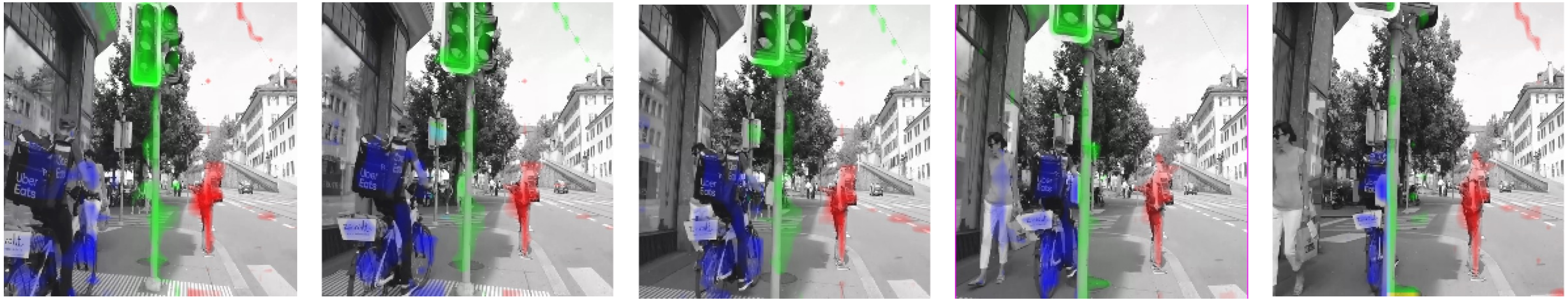




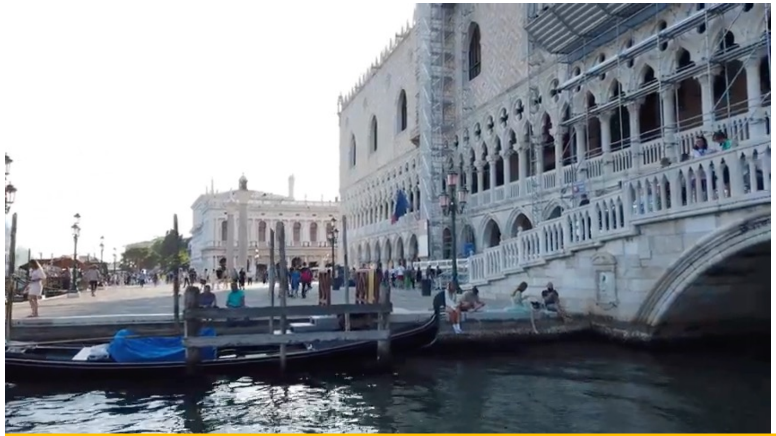
Learning Image Encoders From Video

- A new dataset of open-source first-person video for the purpose of virtual “walking tours”.

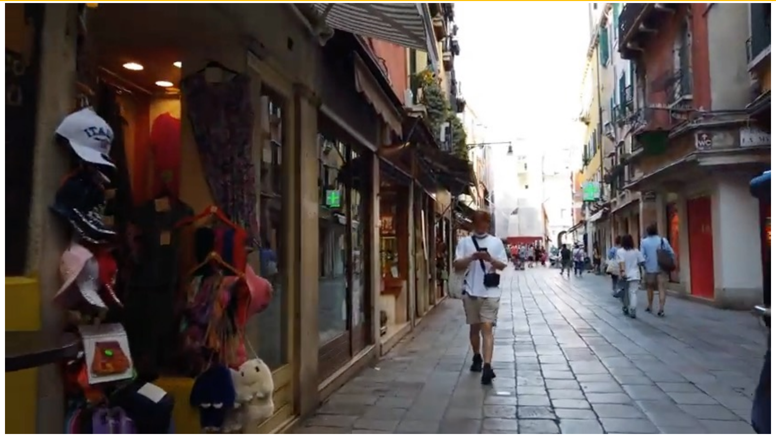
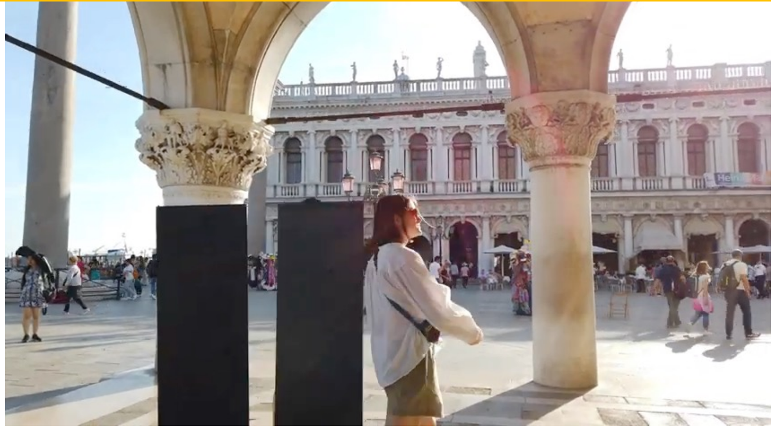
- A new SSL framework, to discover and track objects over time in an end-to-end manner, using transformer cross-attention.



Walking Tour Dataset



10 x 4K videos from different cities, Avg duration – 1hr 38min, ~700 classes, License - CC-BY

Walking Tour Dataset

- Some interesting properties in Walking Tour videos

  1. Natural transition in lighting conditions.

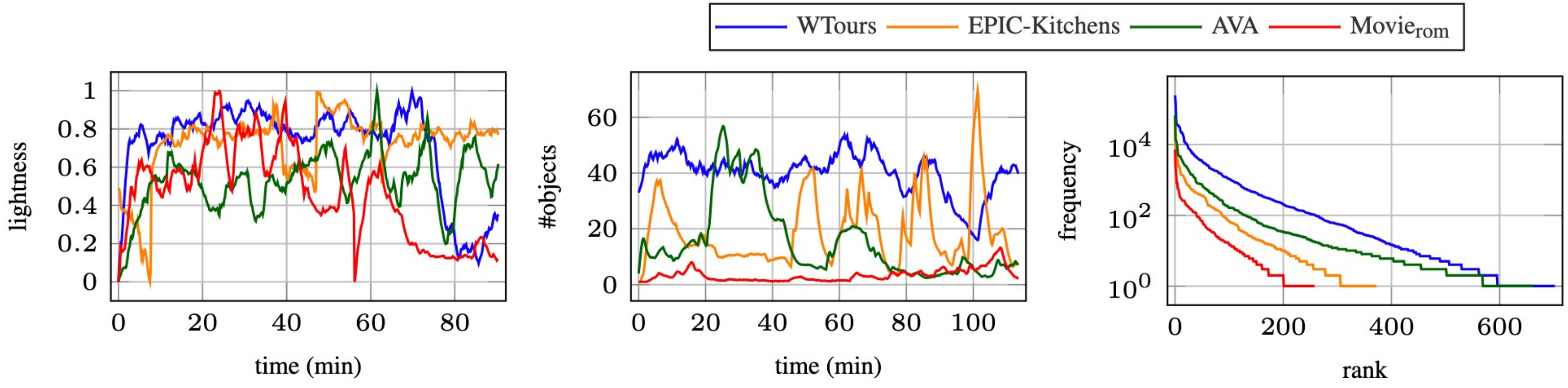
(a) Lightness

Walking Tour Dataset

- Some interesting properties in Walking Tour videos

  1. Natural transition in lighting conditions.

  2. Large number of objects and actions.

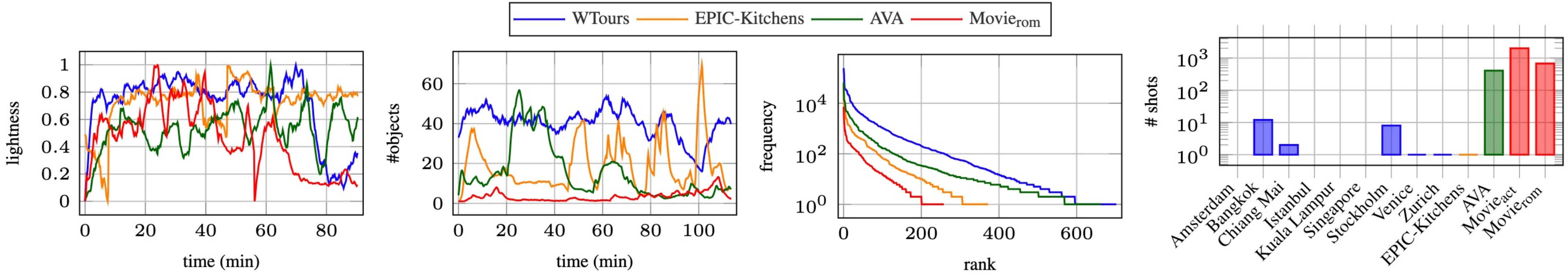


(a) Lightness (b) #Objects per frame (c) Total #classes

Walking Tour Dataset

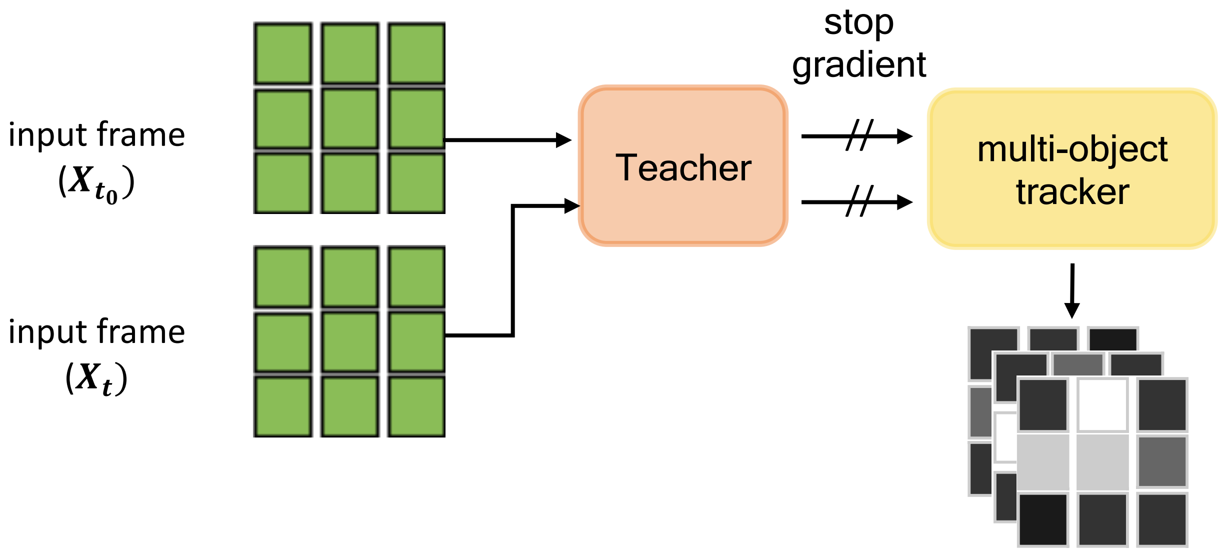
- Some interesting properties in Walking Tour videos

1. Natural transition in lighting conditions.
2. Large number of objects and actions.
3. Natural transition in scenes.



(a) Lightness (b) #Objects per frame (c) Total #classes (d) Shots

DoRA: Discover and tRAck

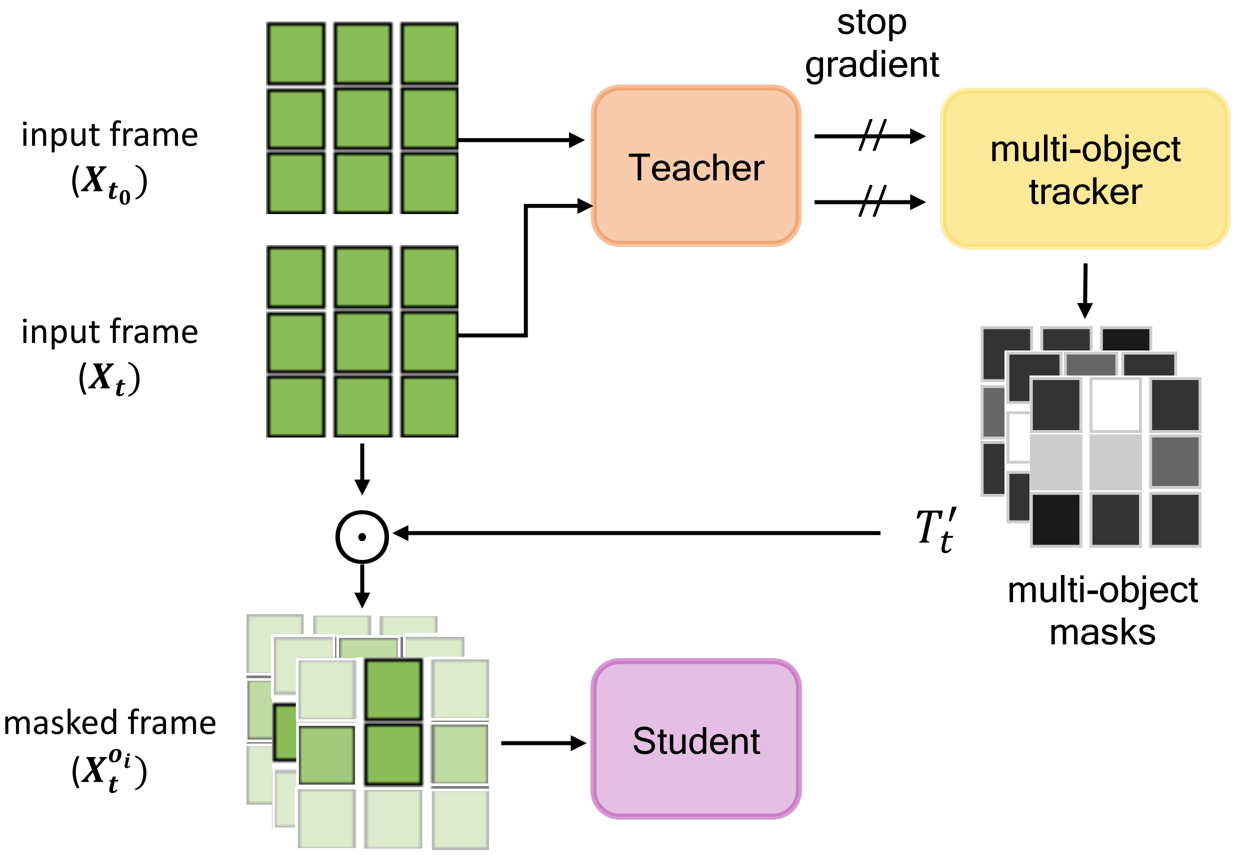


High-level idea

1. Use attention from [cls] token to detect and track multiple objects.

2. Enforce invariance of features over time.

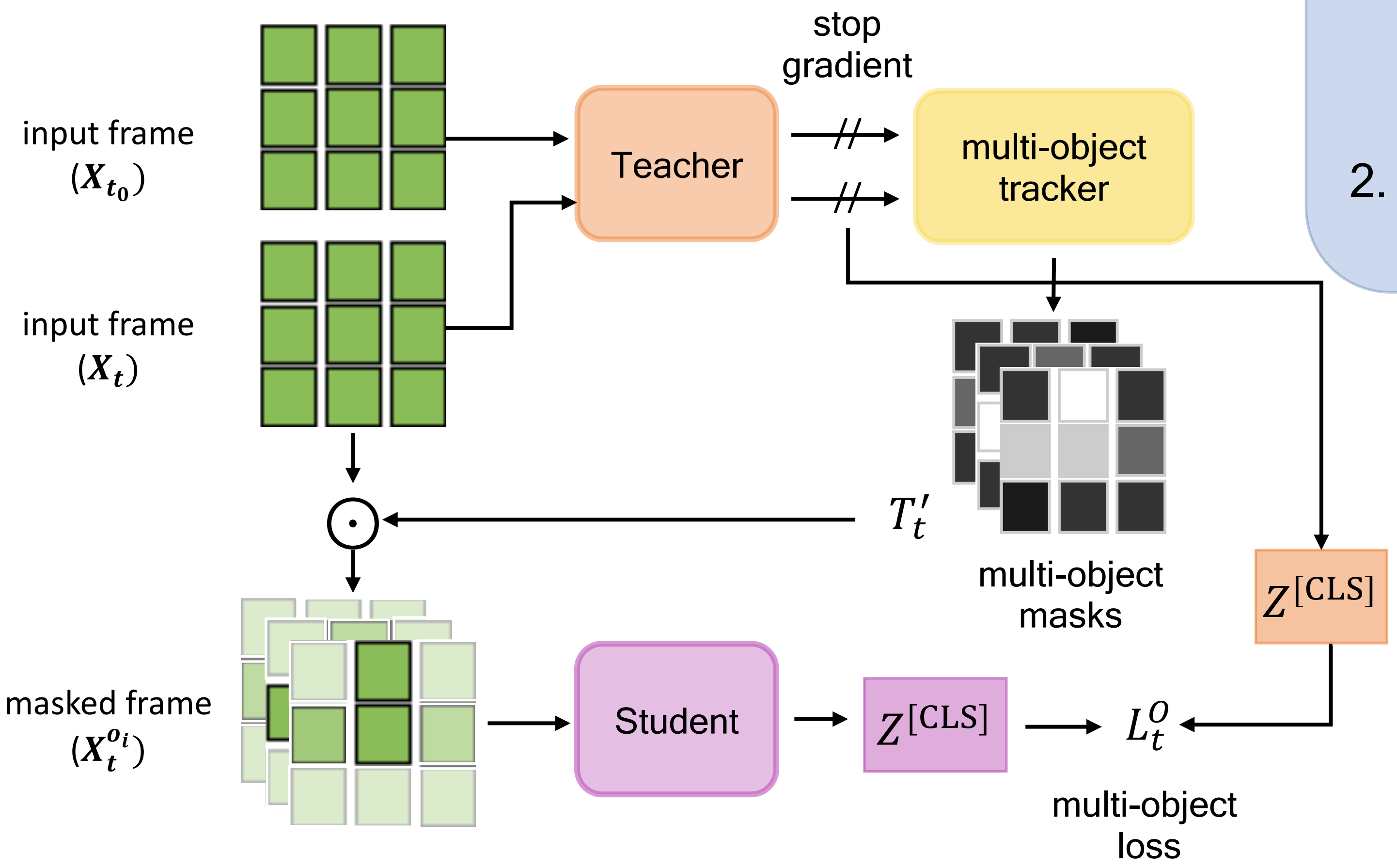
DoRA: Discover and tRAck



High-level idea

1. Use attention from [cls] token to detect and track multiple objects.
2. Enforce invariance of features over time.

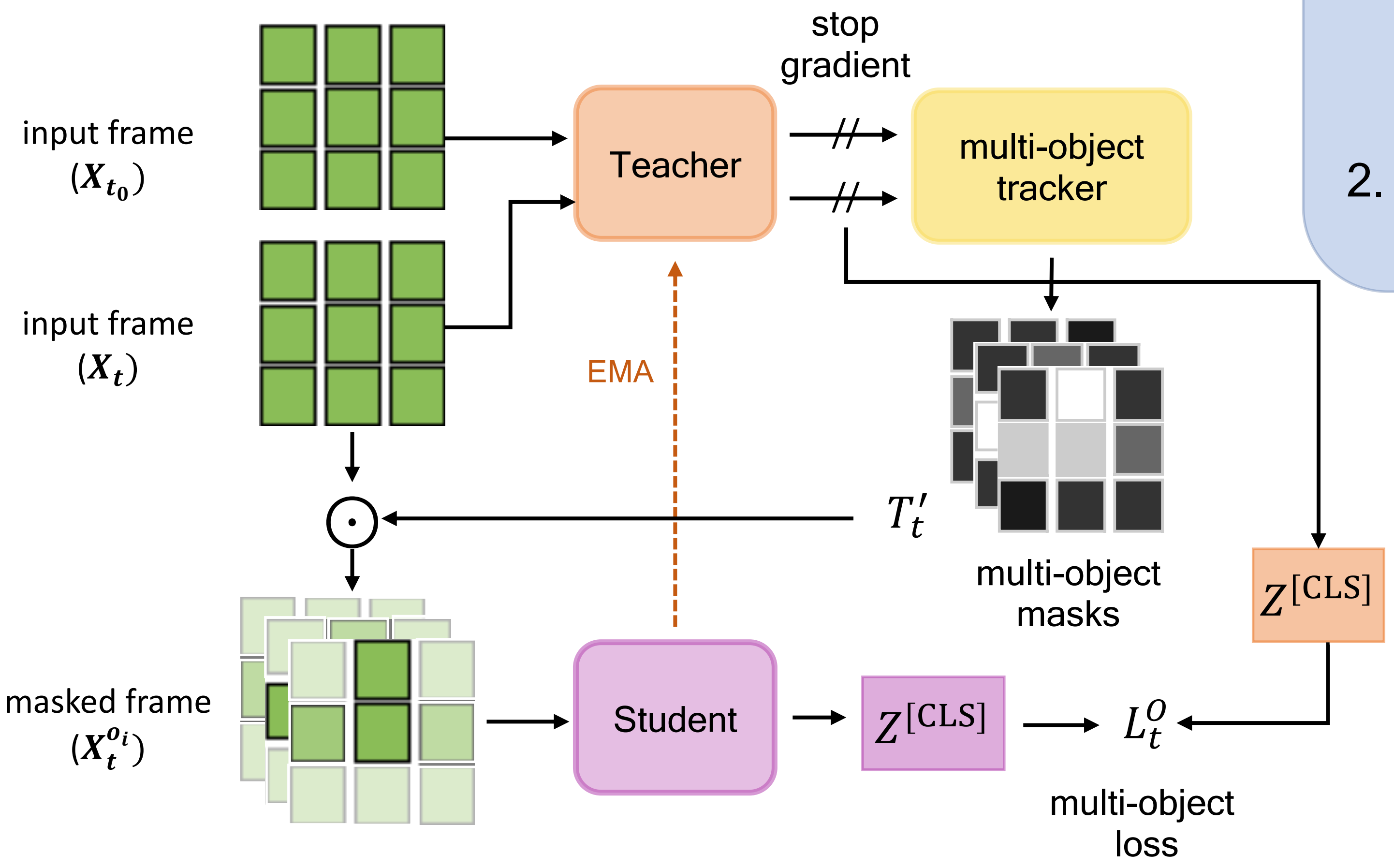
DoRA: Discover and tRAck



High-level idea

1. Use attention from [cls] token to detect and track multiple objects.
2. Enforce invariance of features over time.

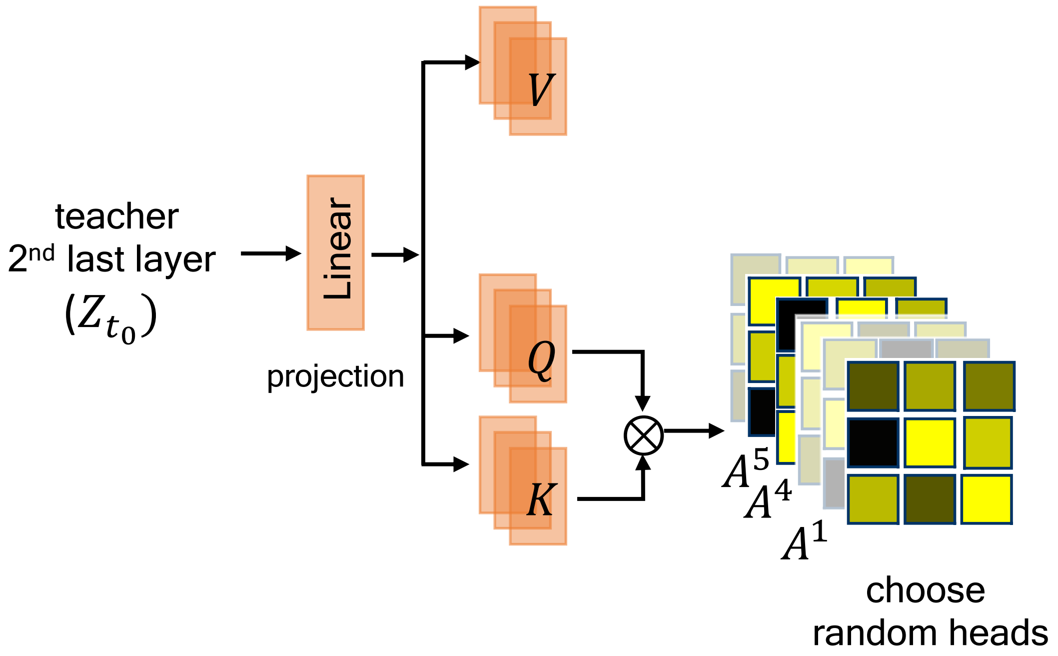
DoRA: Discover and tRAck



High-level idea

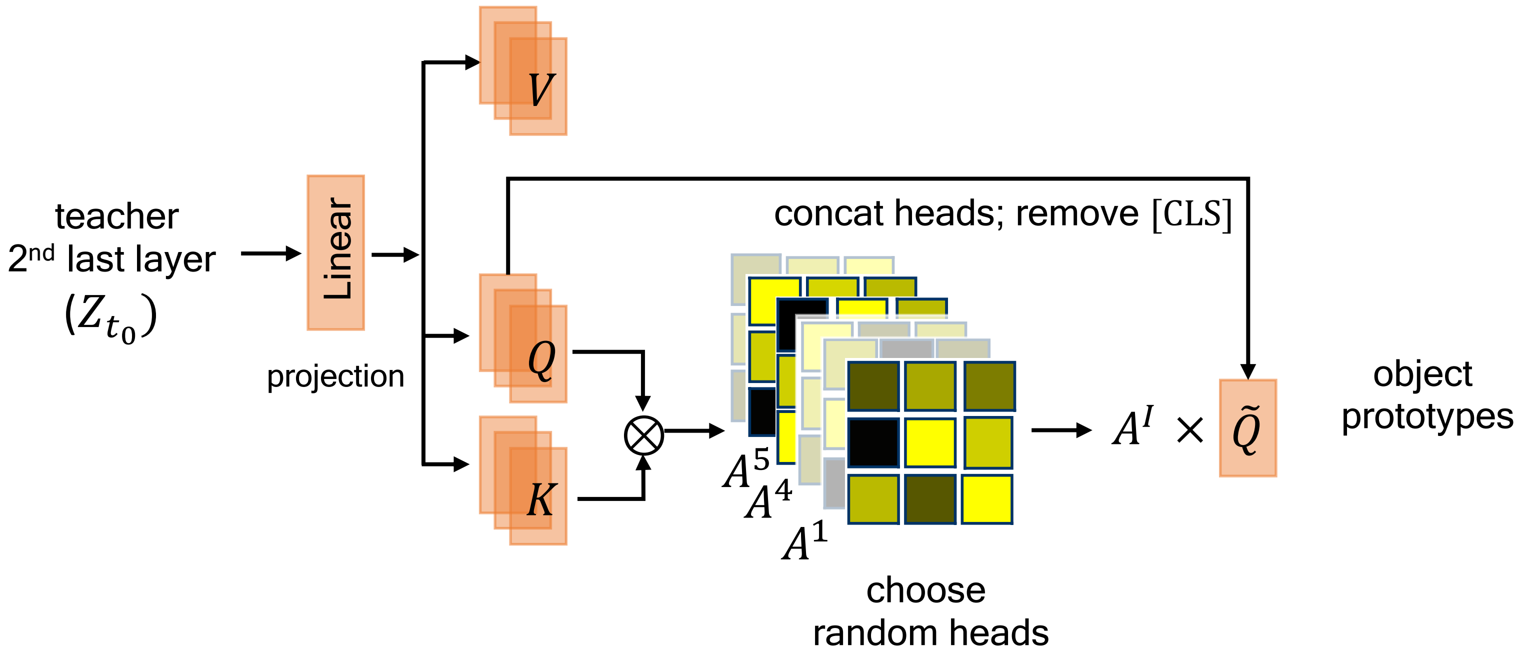
1. Use attention from [cls] token to detect and track multiple objects.
2. Enforce invariance of features over time.

DoRA: Multi-Object Tracker



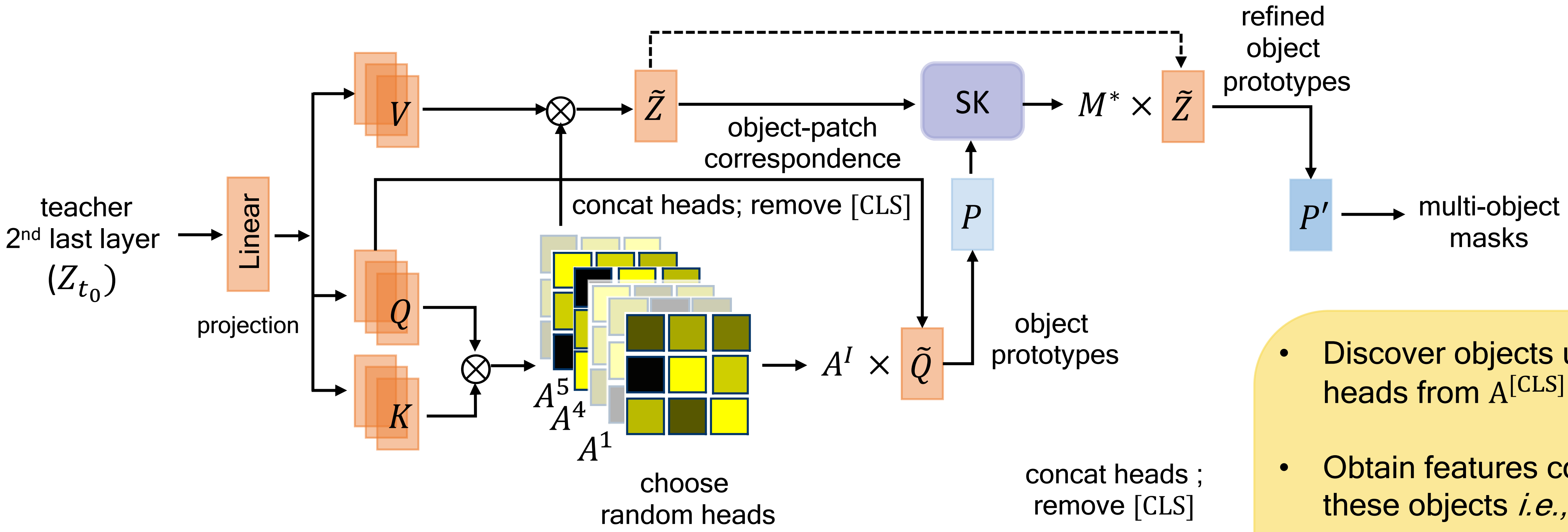
- Discover objects using three random heads from $A^{[CLS]}$

DoRA: Multi-Object Tracker



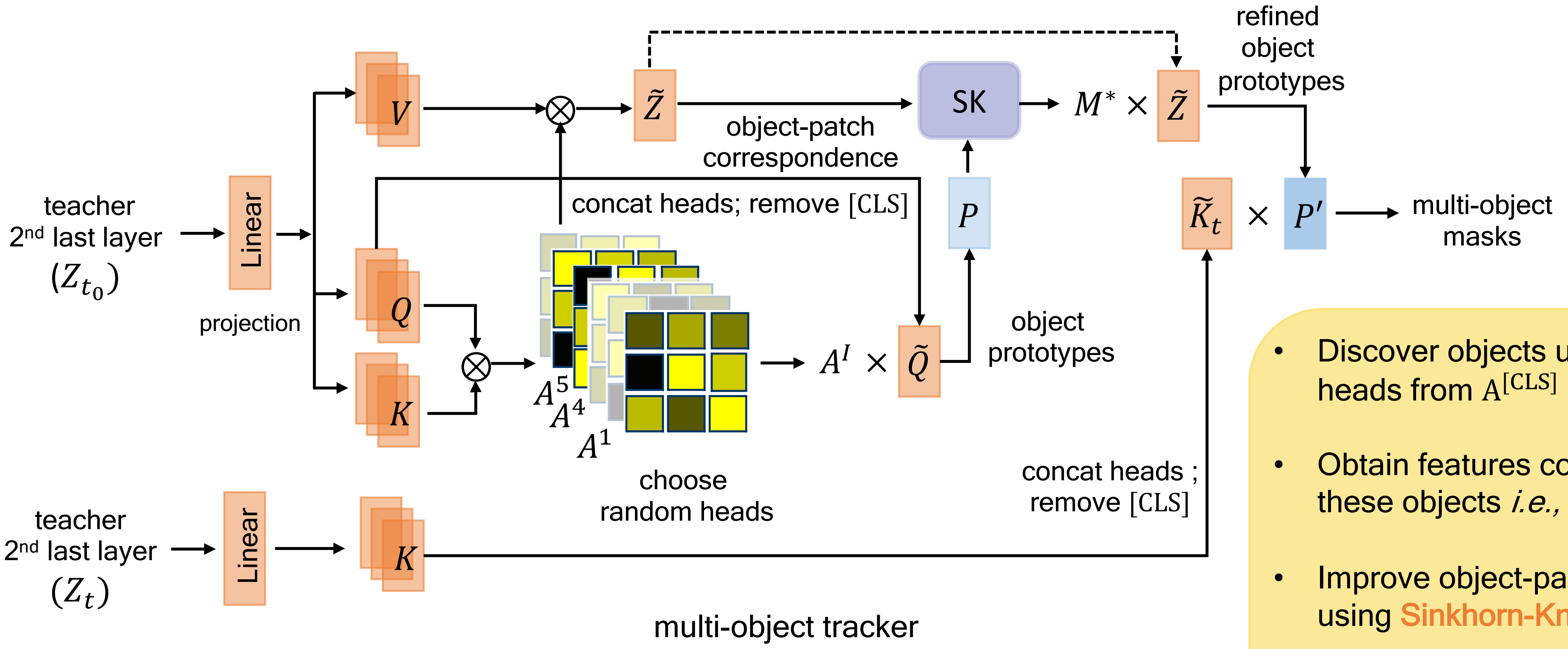
- Discover objects using three random heads from $A^{[CLS]}$
- Obtain features corresponding to these objects *i.e.,* object prototypes.

DoRA: Multi-Object Tracker



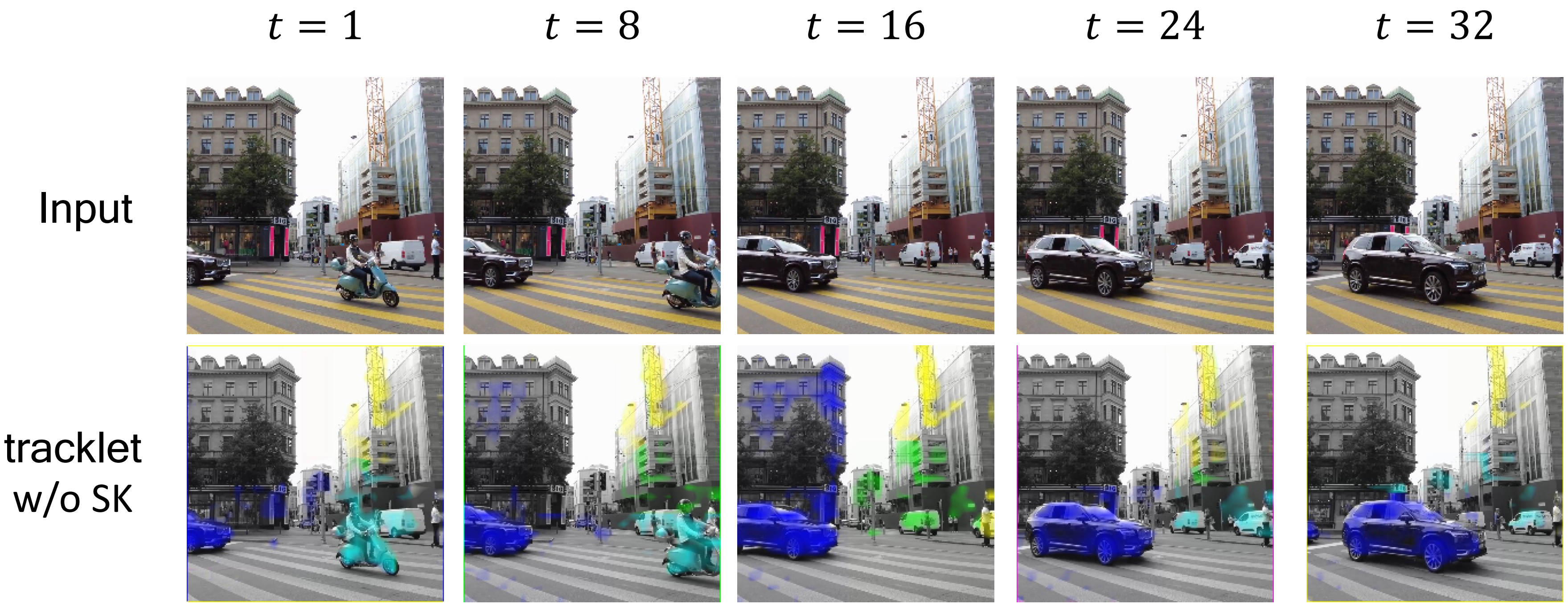
- Discover objects using three random heads from $A^{[CLS]}$
- Obtain features corresponding to these objects *i.e.*, object prototypes.
- Improve object-patch correspondence using Sinkhorn-Knopp.

DoRA: Multi-Object Tracker

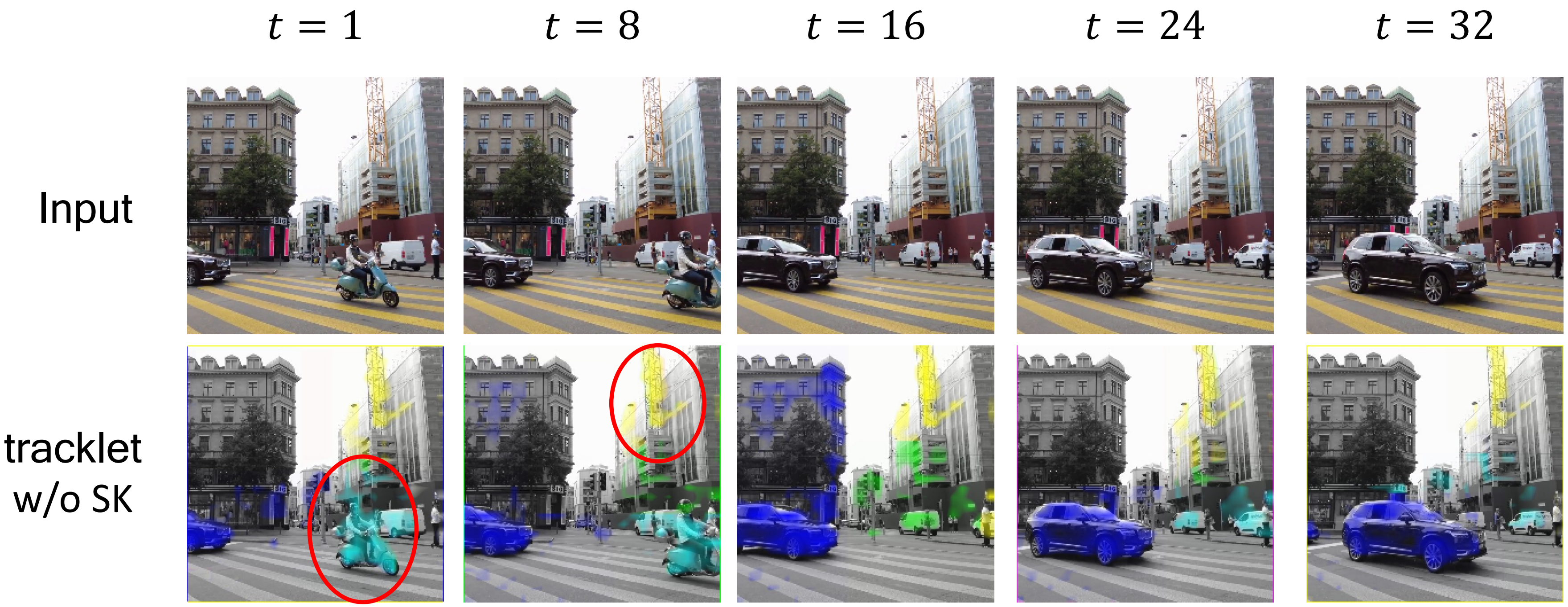


- Discover objects using three random heads from $A^{[CLS]}$
- Obtain features corresponding to these objects *i.e.*, object prototypes.
- Improve object-patch correspondence using Sinkhorn-Knopp.
- Obtain multi-object masks using cross-attention.

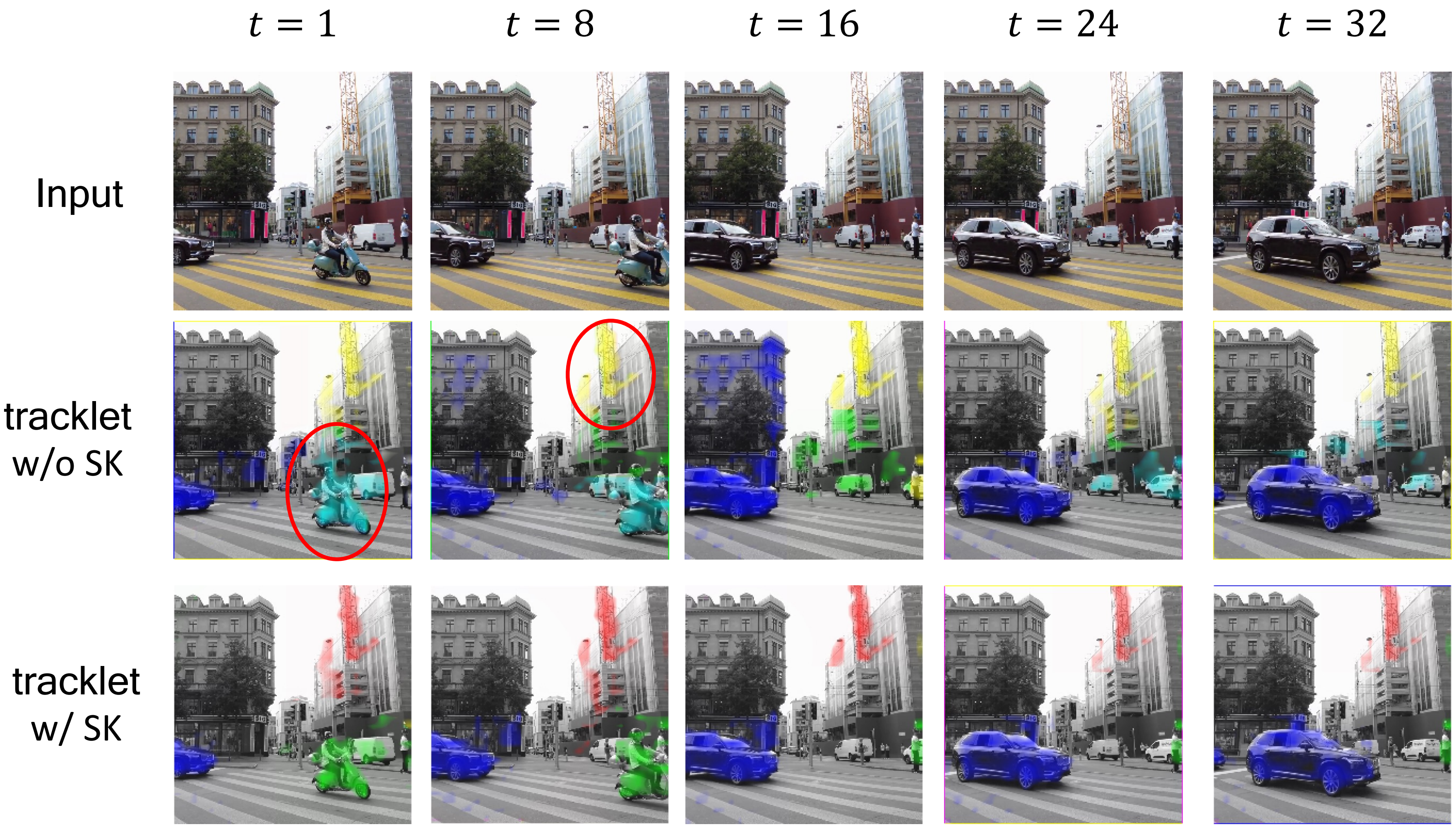
DoRA: Visualizing Tracking



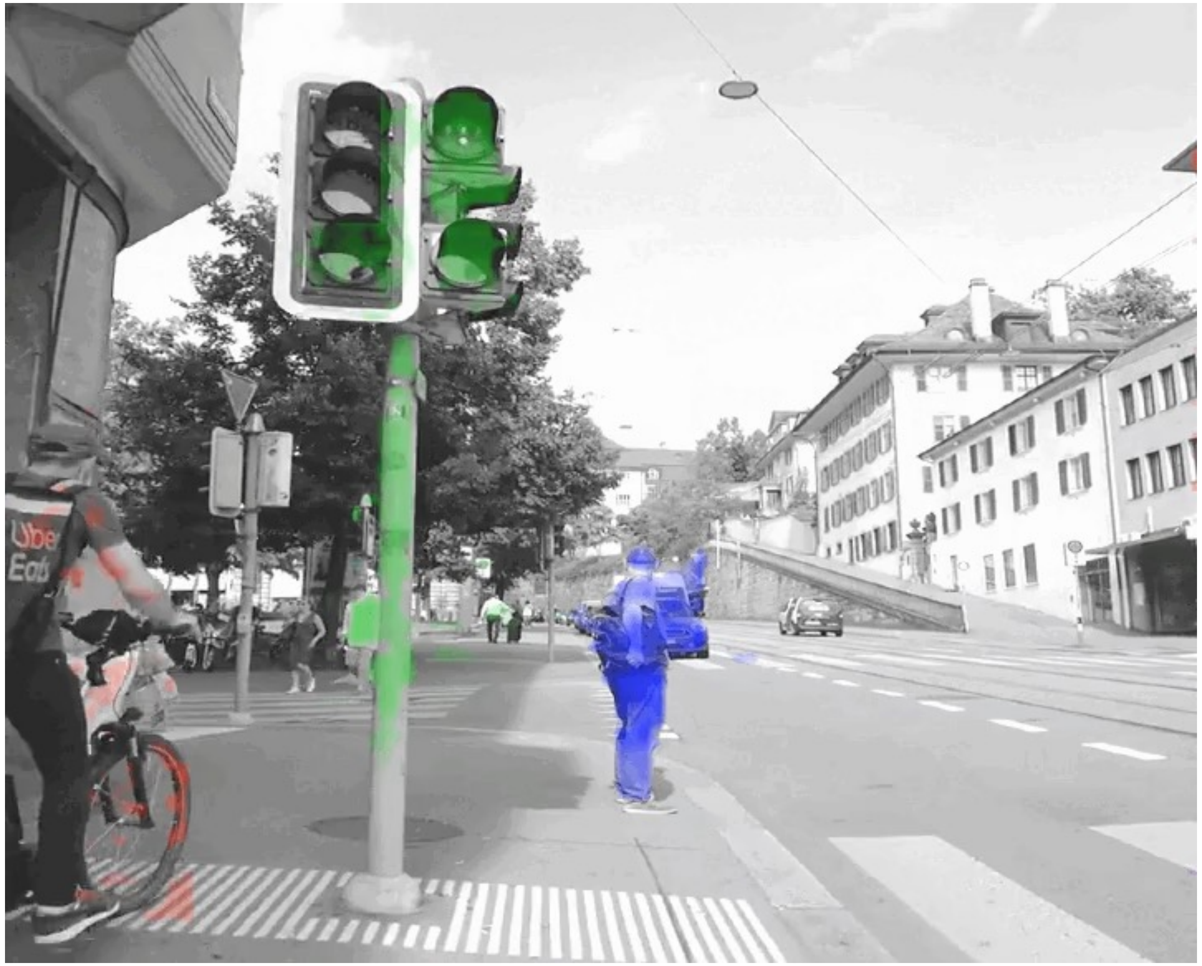
DoRA: Visualizing Tracking



DoRA: Visualizing Tracking



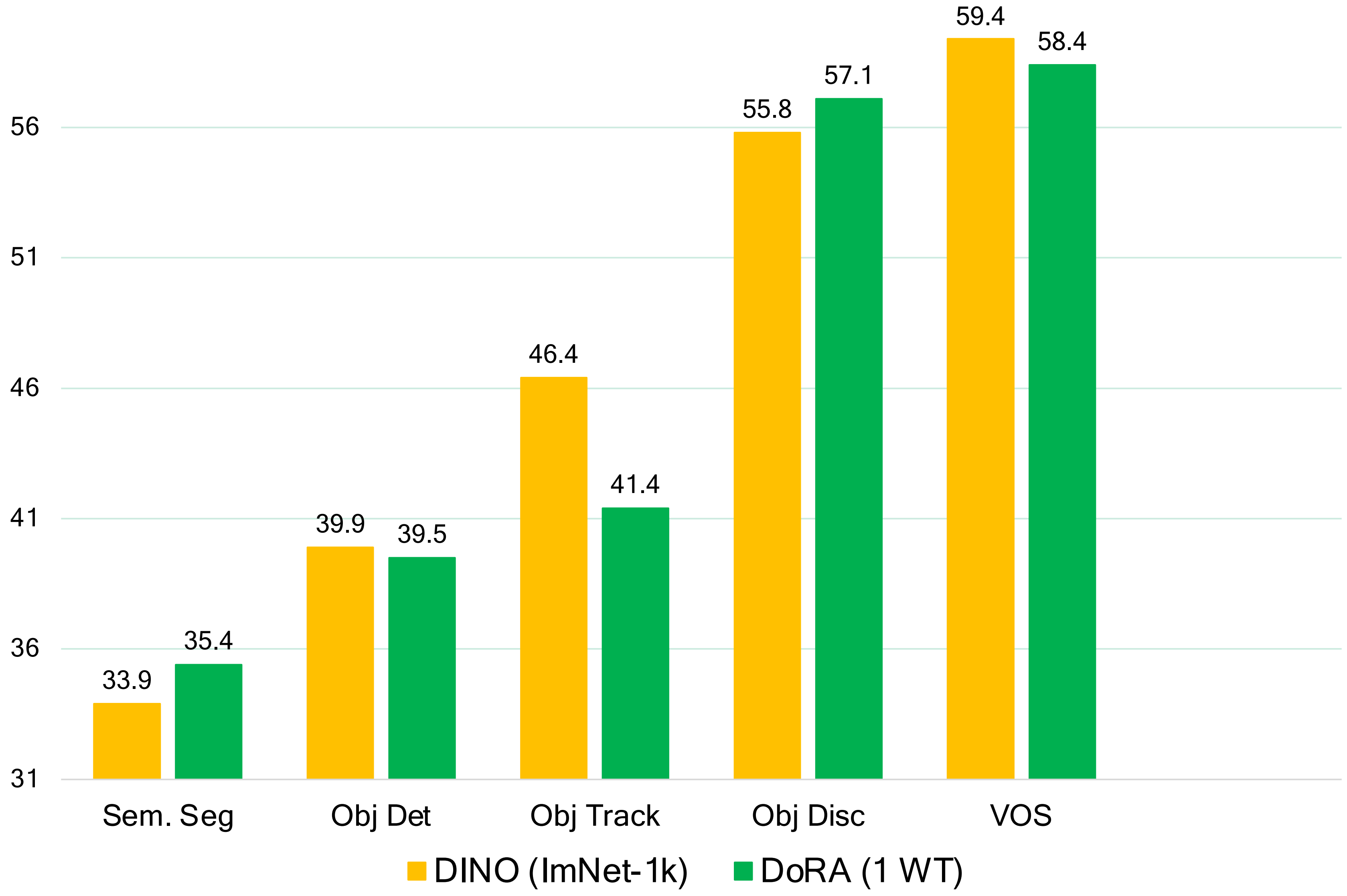
DoRA: Visualizing Tracking



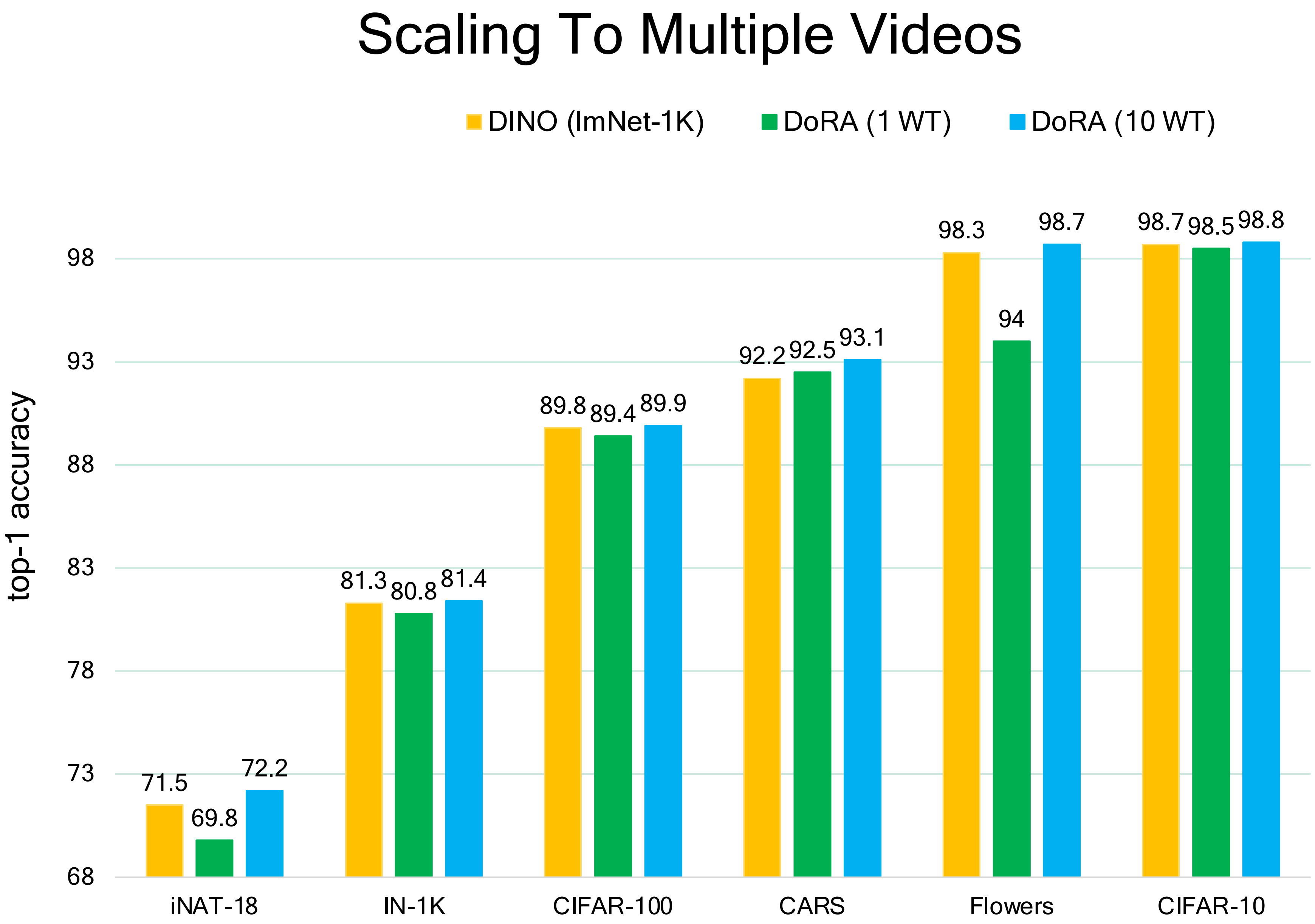
DoRA: Visualizing Tracking

Is ImageNet worth one video?

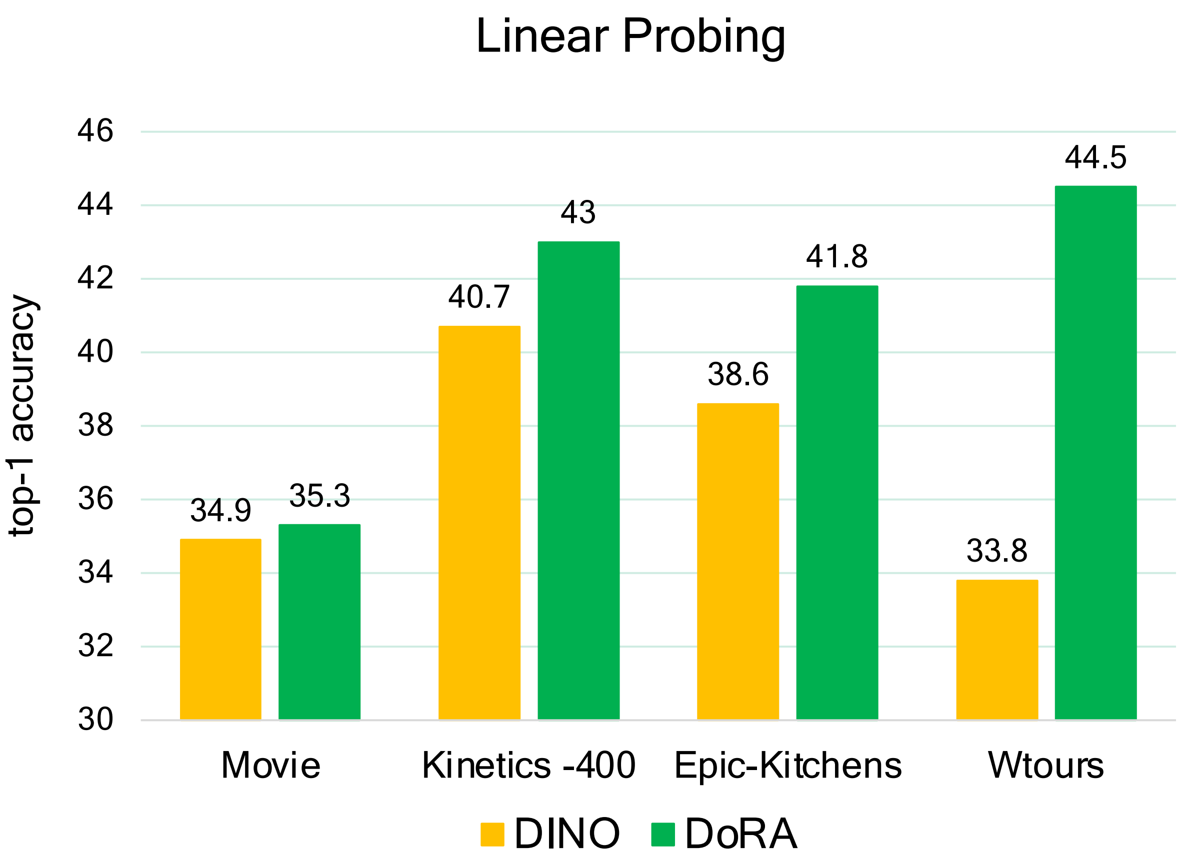
1 Video Better Than ImageNet Pretraining

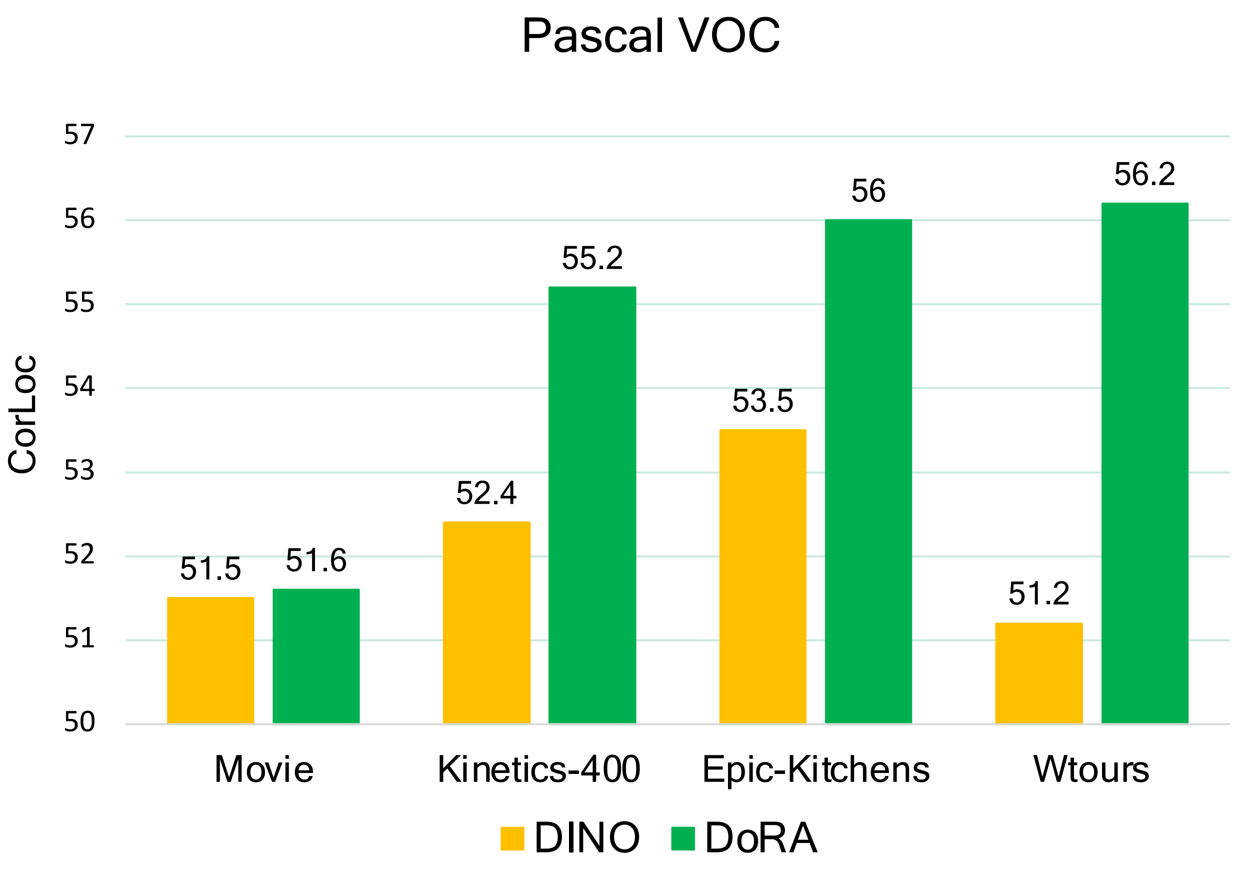


DoRA outperforms DINO on Semantic Segmentation and Object Discovery


Scaling To Multiple Videos
DINO (ImNet-1K)
DoRA (1 WT)
DoRA (10 WT)
top-1 accuracy
98
93
88
83
78
73
68
71.5
69.8
72.2
81.3
80.8
81.4
89.8
89.4
89.9
92.2
92.5
93.1
98.3
94
98.7
98.7
98.5
98.8
iNAT-18
IN-1K
CIFAR-100
CARS
Flowers
CIFAR-10

Pretraining On Different Videos





Thank you

Project Page