University of Amsterdam

Modern Self-supervised Image Representation Learning from Videos

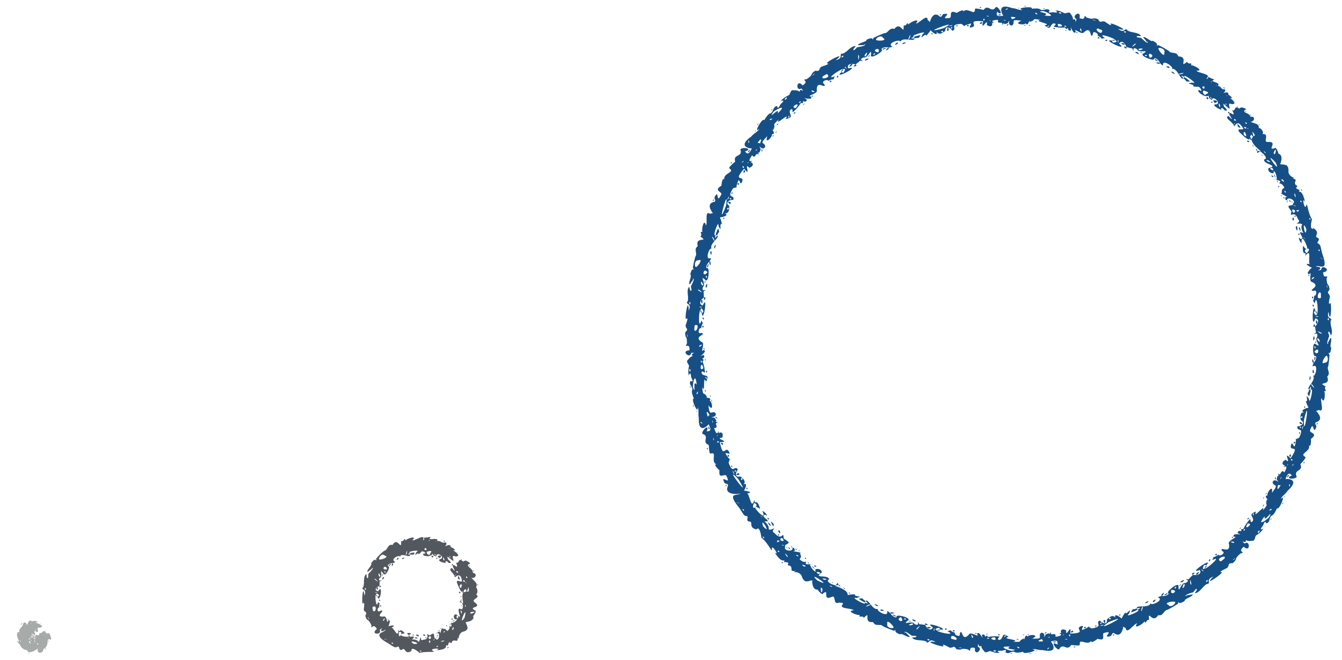
Yuki Asano

NeurIPS 2023

Van Gogh self-portraits 1888, 1889

Why do we want Self-supervised Learning in the age of CLIP *et al*?

Massive scale

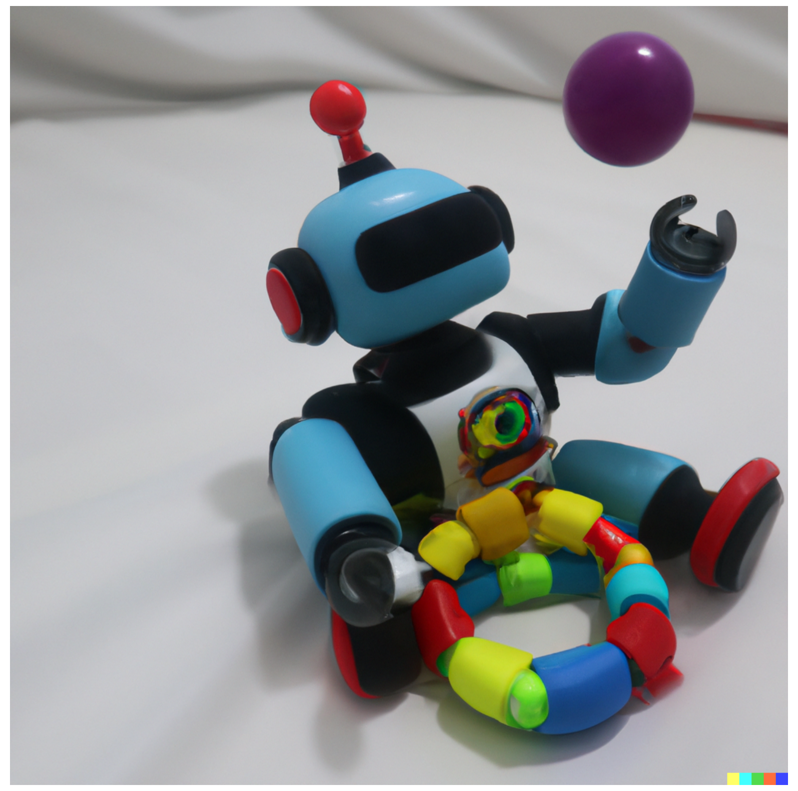


sup. << weak sup. << raw

Cost of (re)labelling

Problems of labels

Fundamentals



Especially videos open exiting new directions

Visual development for AI

"Get" physics

Embodied AI

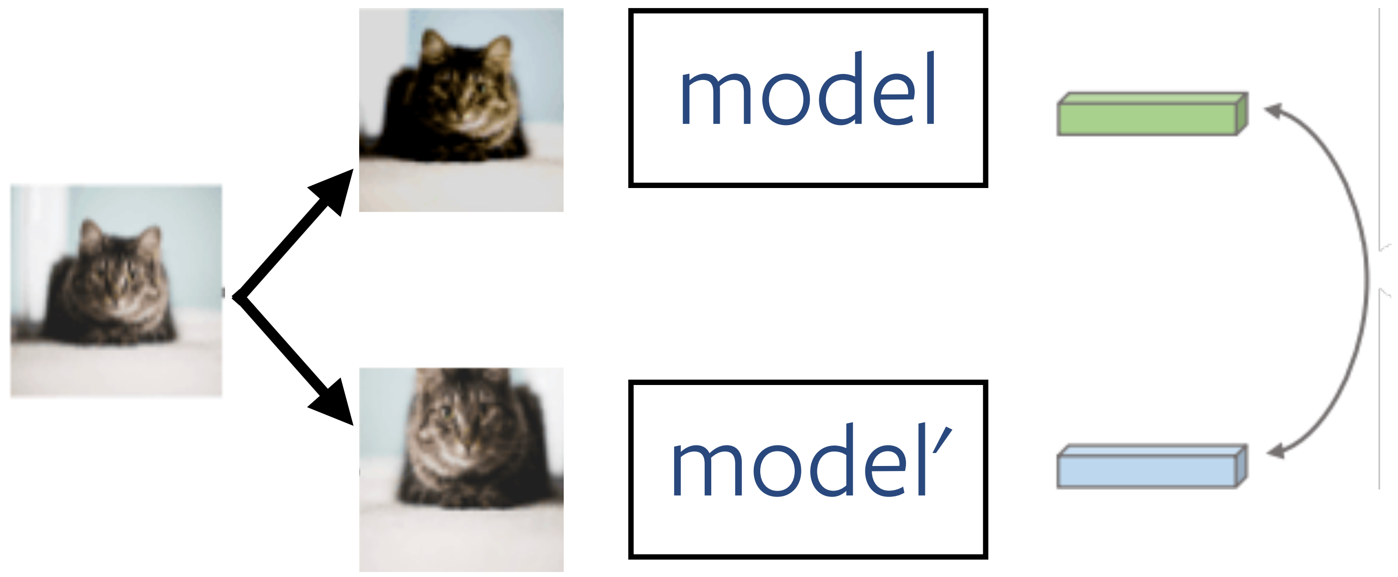
Bonus: *insane* scale:

LEARNING IMAGE ENCODERS FROM TIME

Piet Mondrian, paintings 1910, 1911, 1913, 1930

Augmentations are crucial in classic image-SSL, but forcing frames to be invariant is limiting

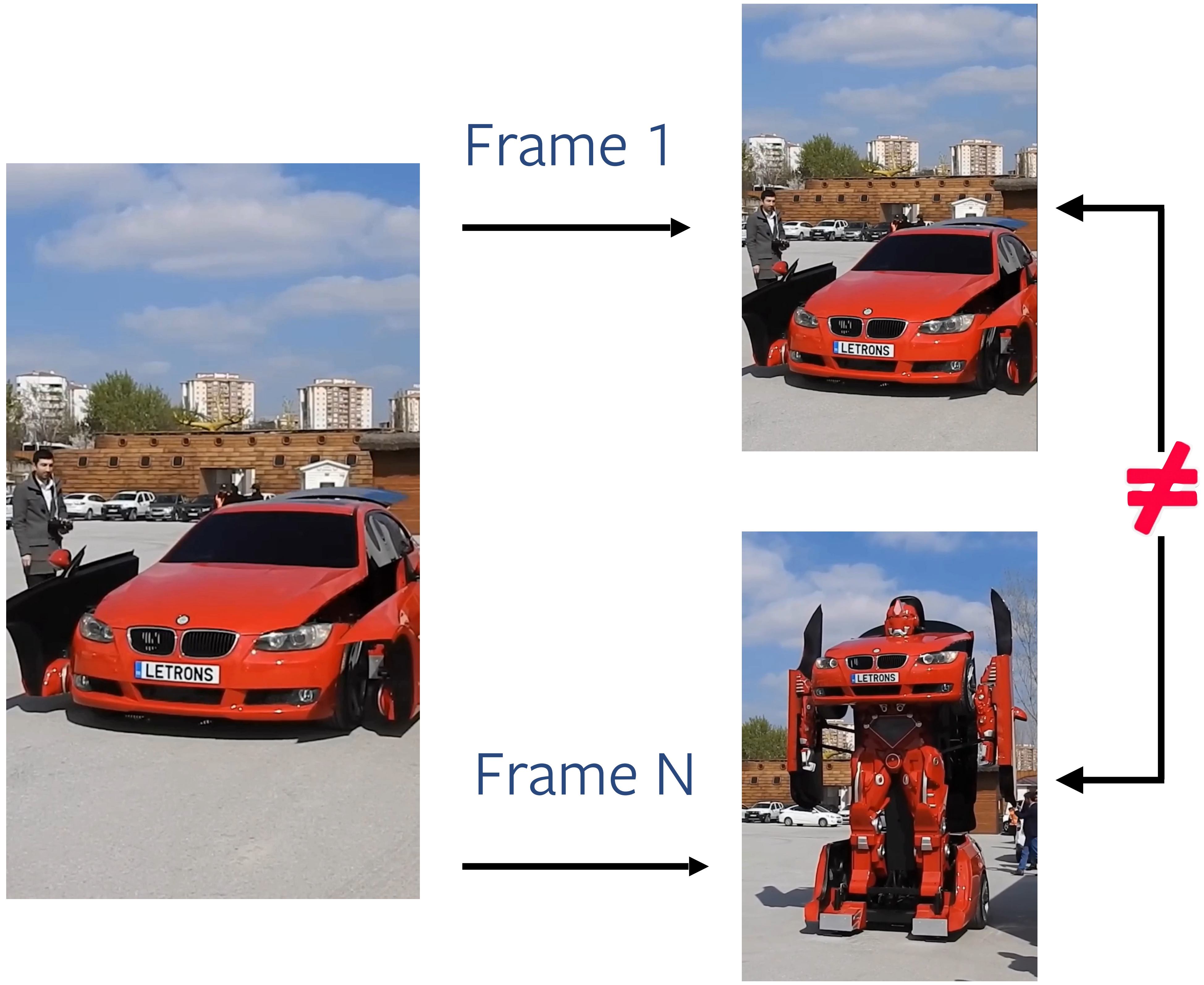
Images: SimCLR, MoCo, SwaAV et al.



key principle: view-invariance

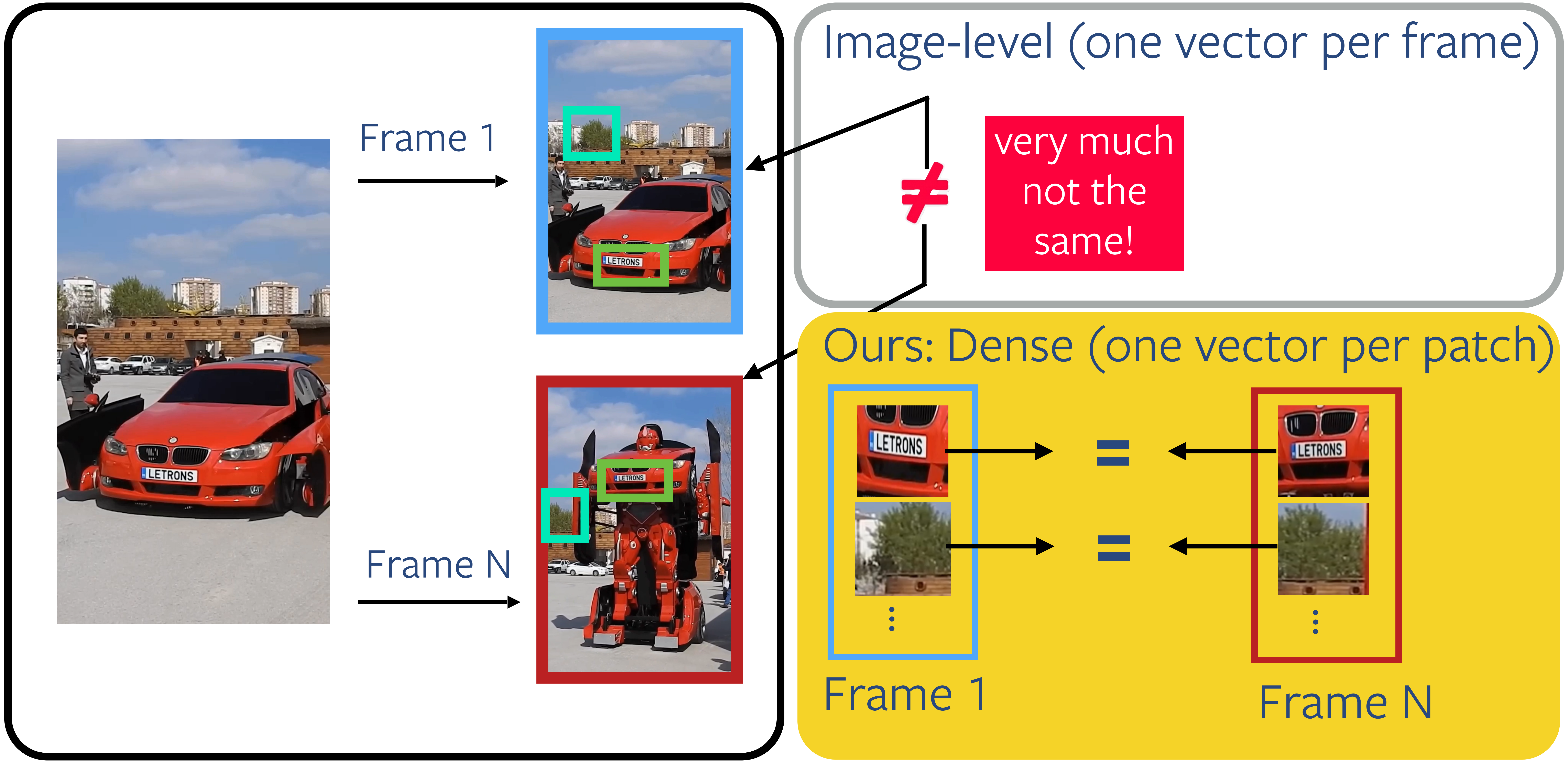
Might be ok for videos like this:

But does this generally make sense?



very much not the same!

Solution is obvious

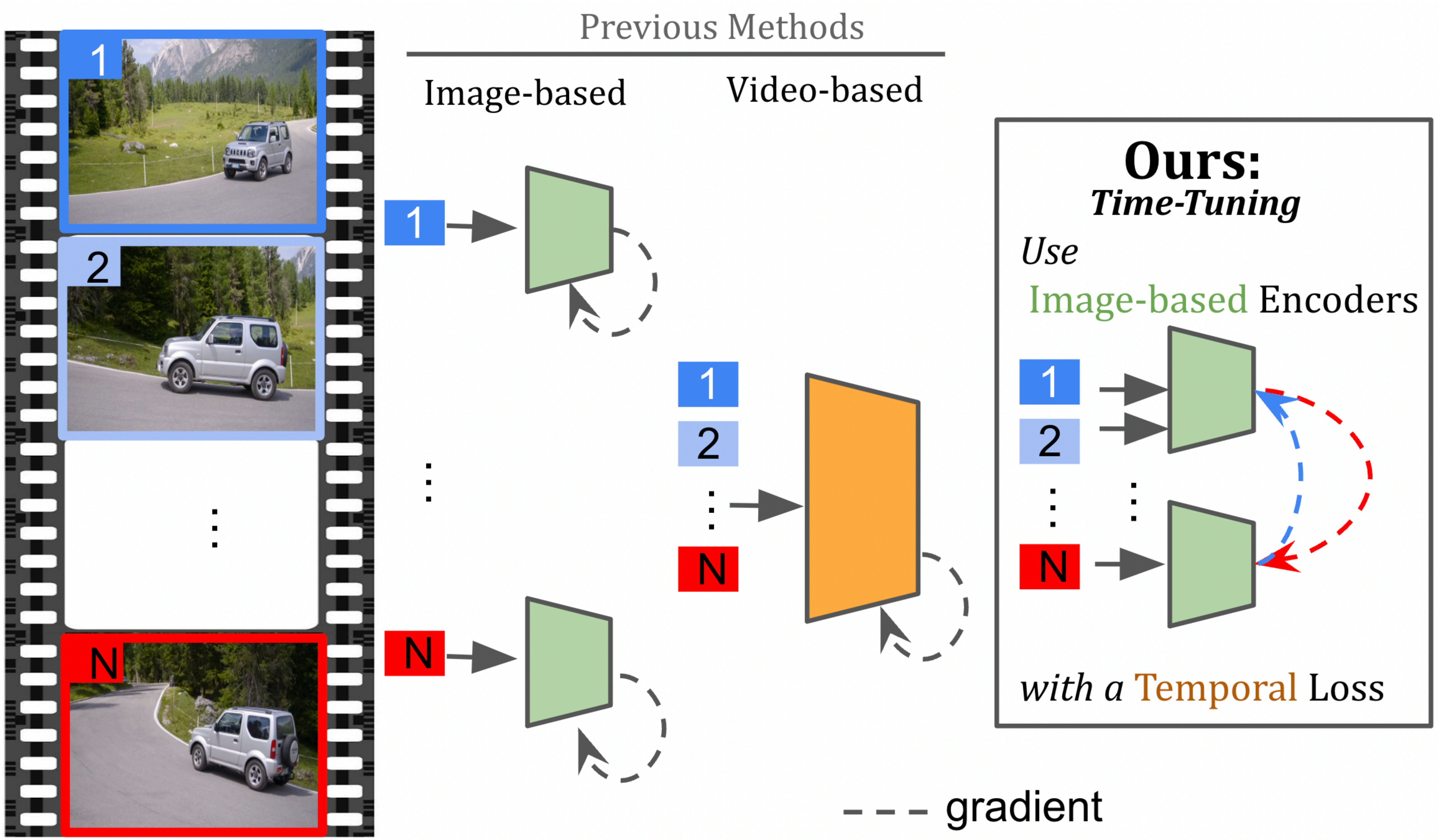


Salehi, Gavves, Snoek, Asano. *Time does tell: self-supervised time-tuning of dense image representations.* ICCV 2023

We *model* a video by tracking image patches, and aligning their clustered features

Salehi, Gavves, Snoek, Asano. *Time does tell: self-supervised time-tuning of dense image representations.* ICCV 2023

Using videos to learn self-supervised image encoders

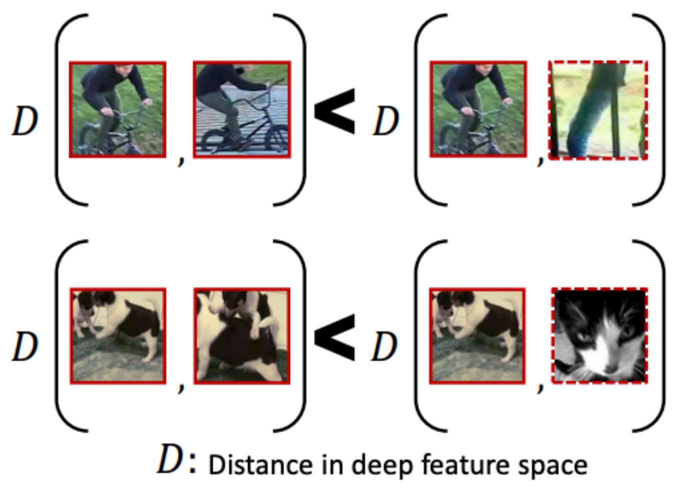


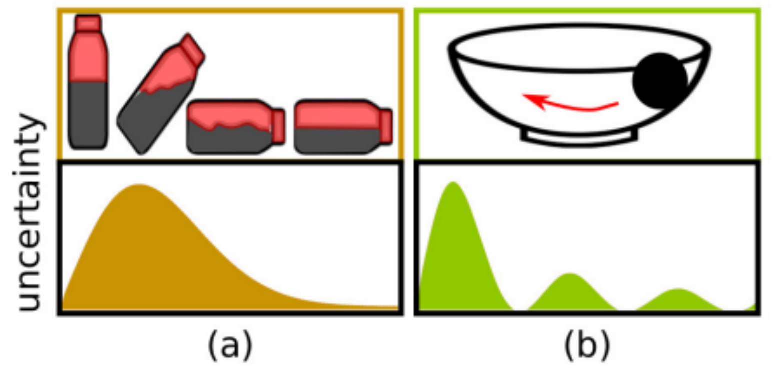
How is time utilised?	As data source	As data dimension	As self-supervised signal

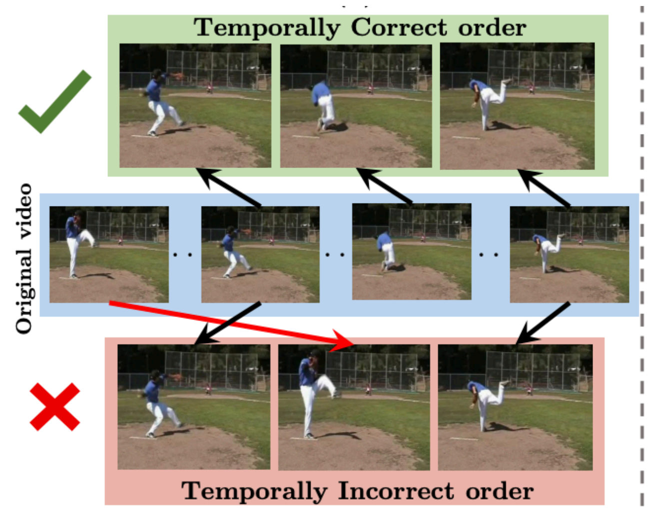
Salehi, Gavves, Snoek, Asano. *Time does tell: self-supervised time-tuning of dense image representations.* ICCV 2023

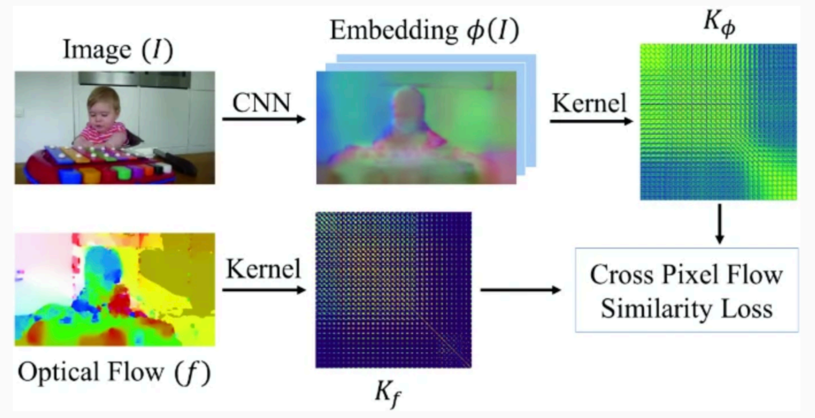
This field has a rich history. And now it's time to get back to it

2015-2019

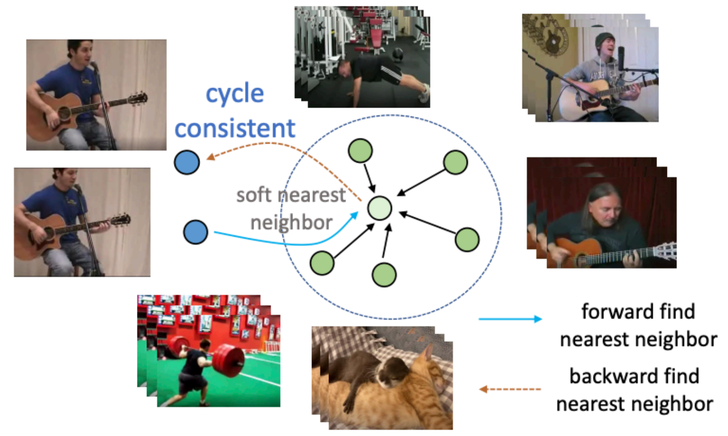



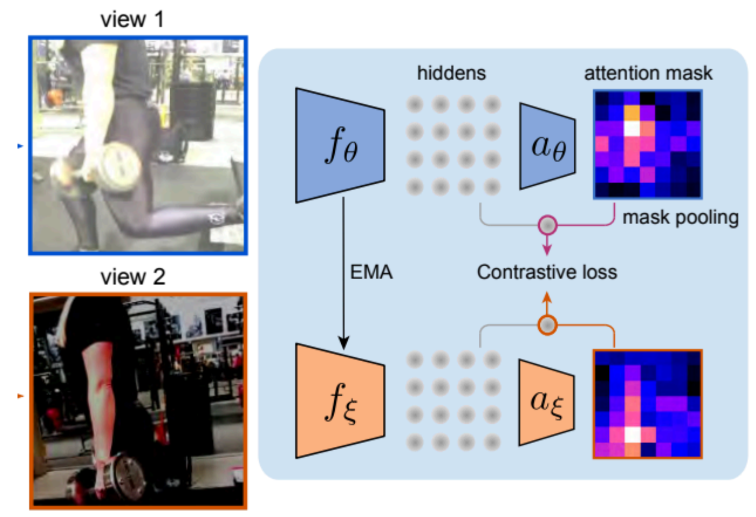


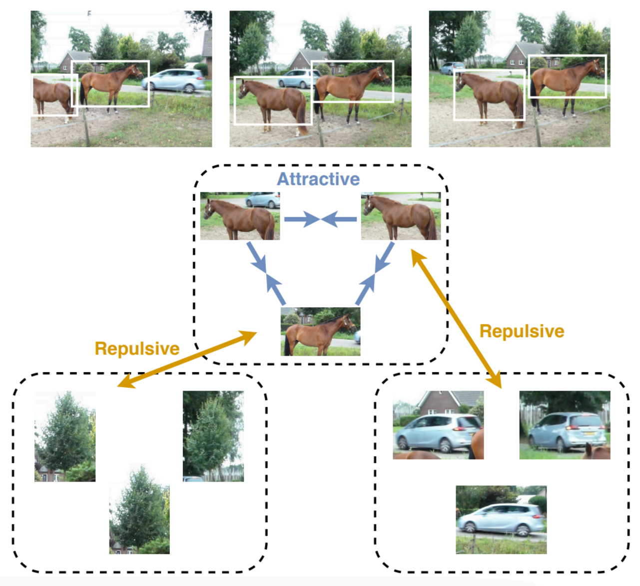


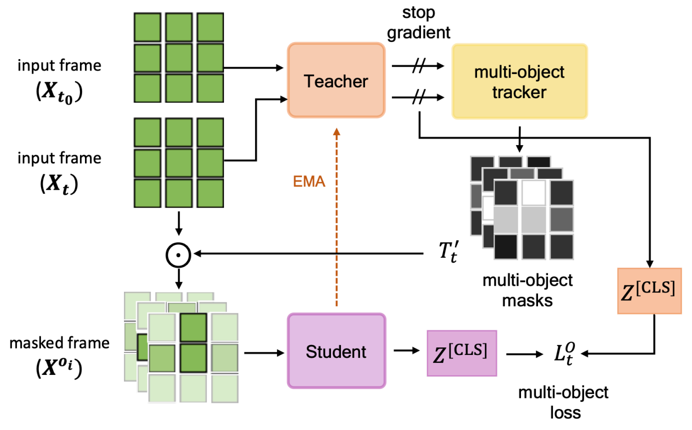


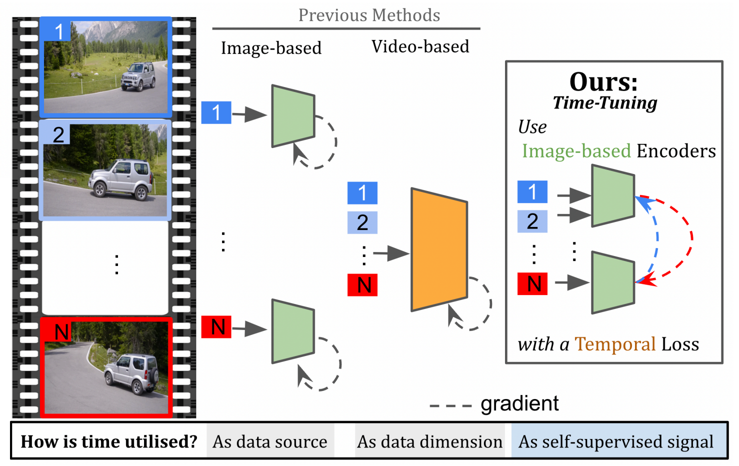
2020 onwards getting competitive to ImageNet

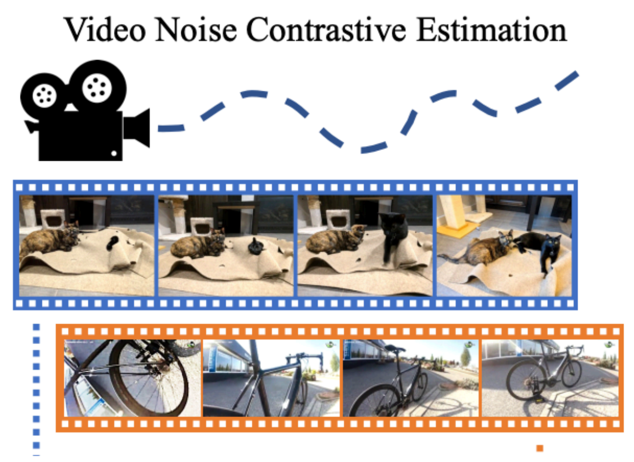


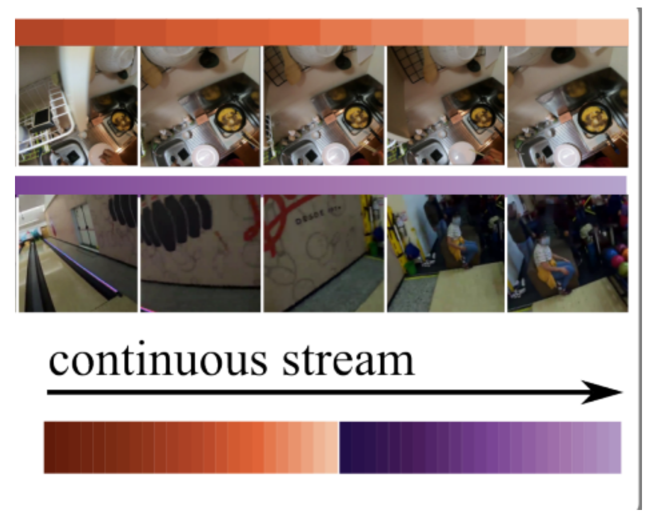








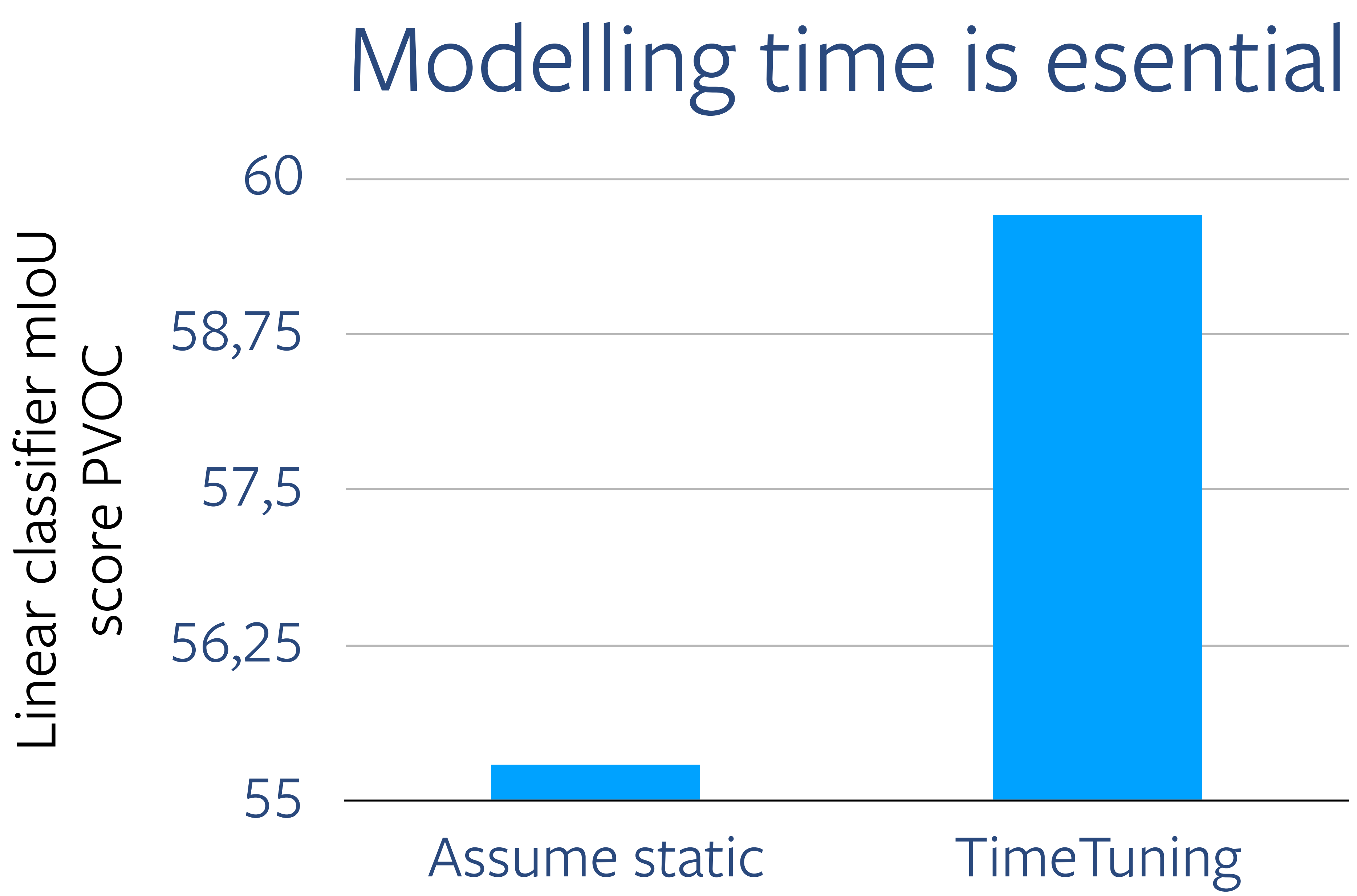


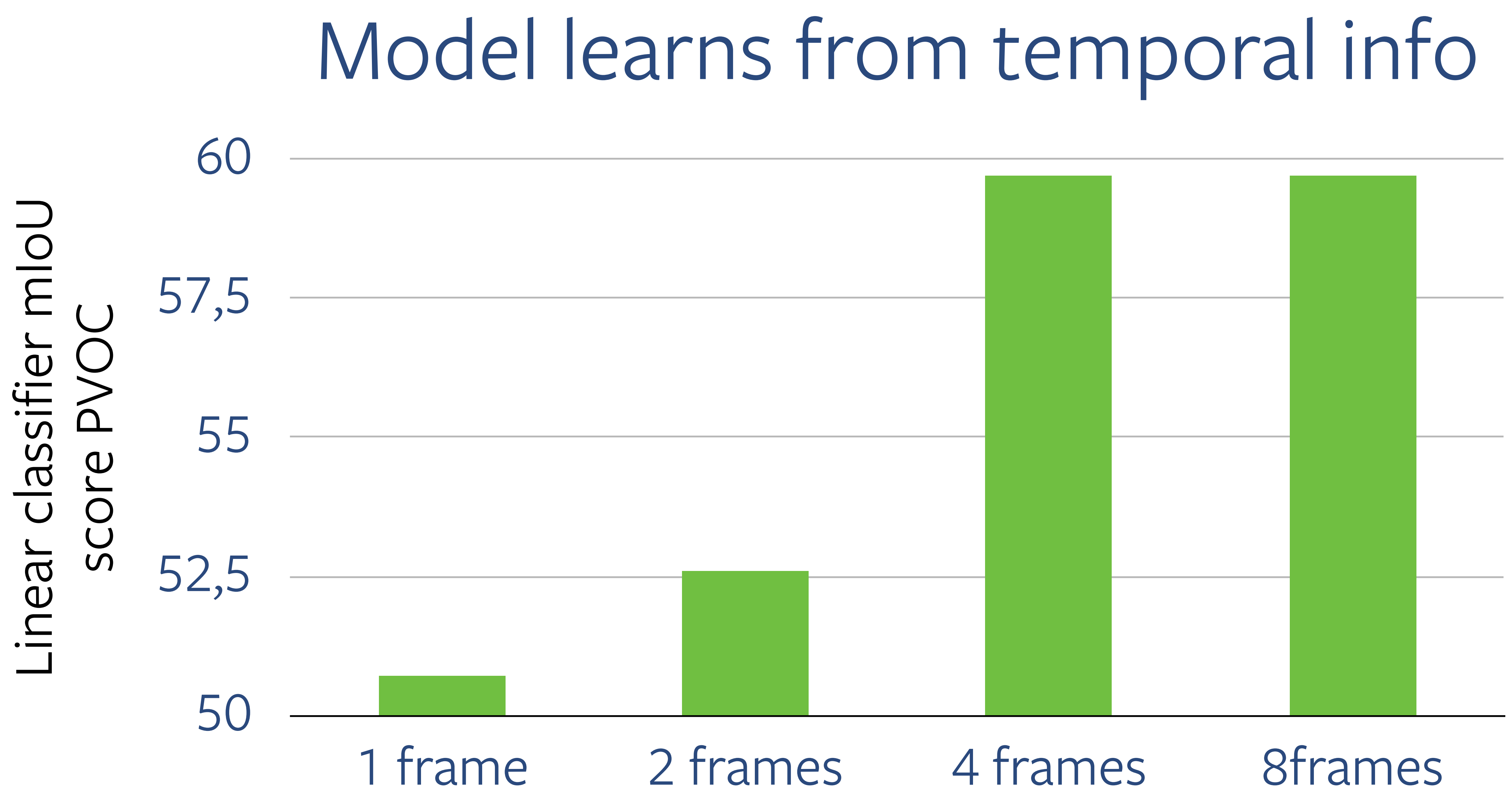


Some references:
2002 Wiskott, Sejnowski. Slow Feature Analysis: Unsupervised Learning of Invariances
2015 Agrawal, Carreira, Malik. Learning to See by Moving: predict egomotion from frames
2015 Wang, Gupta. Unsupervised Learning of Visual Representations using Videos
2015 Goroshin, Bruna, Eigen, LeCun. Unsupervised feature learning from temporal data
2015 Ramanathan, Tang, Mori, Fei-Fei. Learning Temporal Embeddings for Complex Video Analysis
2016 Gao, Jayamaran, Graumann. Object-Centric Representation Learning from Unlabeled Videos
2017 Wang, Kaiming, Gupta. Transitive Invariance for Self-supervised Visual Representation Learning
2018 Wei, Lim, Zisserman, Freeman. Learning and Using the Arrow of Time
2019 Jayamaran, Ebert, Efros, Levine: Time-Agnostic Prediction: Predicting Predictable Video Frames
2019 Mahendran, Thewlis, Vedaldi. focus on motion: cross-pixel flow

2020 Tschannen, Djolonga, Ritter, Mahendran, Zhai, Houlsby, Gelly, Lucic. Self-Supervised Learning of Video-Induced Visual Invariances.
2021 Wu, Wang. Contrastive Learning of Image Representations with Cross-Video Cycle-Consistency
2020 Gordon, Ehsani, Fox, Farhadi. Watching the World Go By: Representation Learning from Unlabeled Videos
2023 Parthasarathy, Eslami, Carreira, Henaff. Self-supervised video pretraining yieldshuman-aligned visual representations
2023 Salehi, Gavves, Snoek, Asano. Time does tell: self-supervised time-tuning of dense image representations
2023 Venkataramanan, Rizve, Carreira, Avrithis*, Asano*. Is ImageNet worth 1 video? Learning strong image encoders from 1 long unlabelled video.
2023 Carreira et al. Learning from One Continuous Video Stream

Ablations demonstrate using time helps learn better features





Unsupervised Semantic Segmentation on videos

[simply running k-means on a couple of videos' spatial features, k=10]

DINO

❌ part-centric maps

STEGO

❌ noisy maps

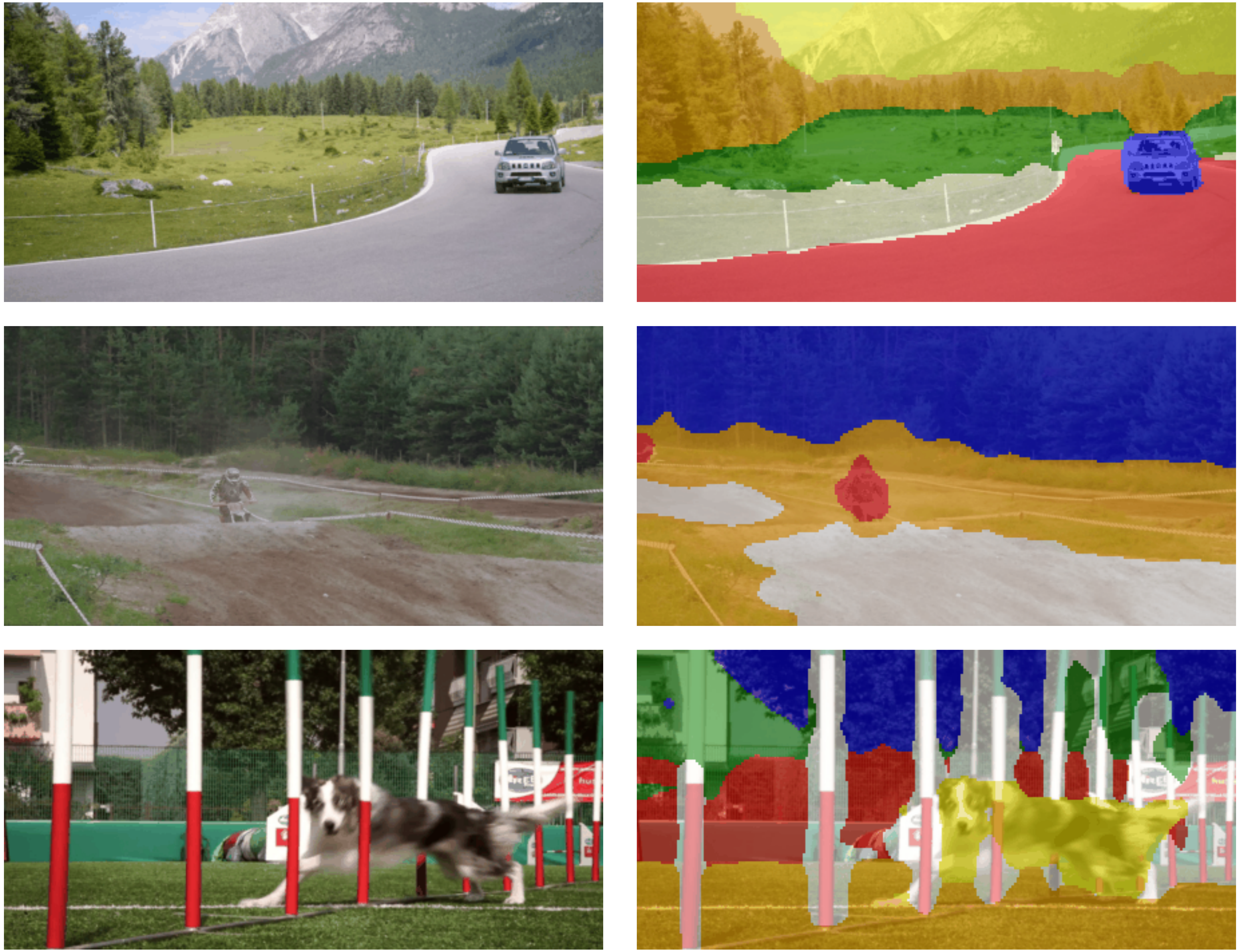
Ours

✅ crisp semantic maps

Unsupervised Semantic Segmentation on videos

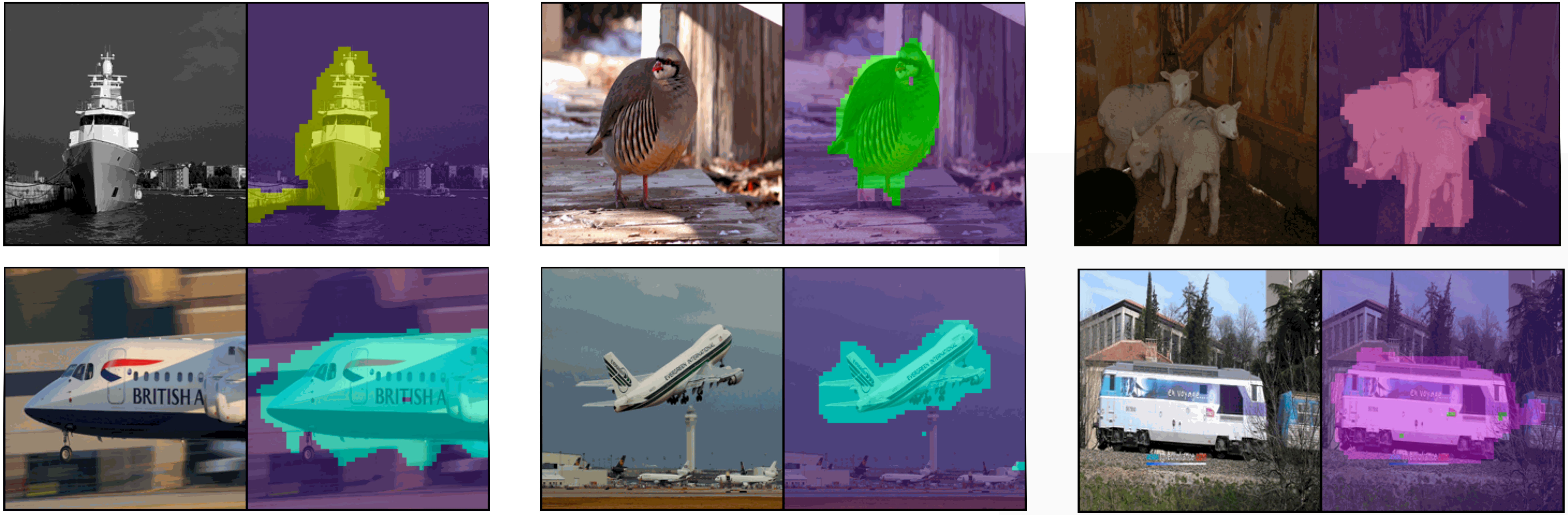
[here: running k-means on the whole video's spatial features, k=5]

More examples



UNIVERSITY OF AMSTERDAM
YUKI ASANO

Also good performance on images, despite having been tuned on videos.



TimeTuning:
DINO as init & use temporal info of videos.

How powerful is time without image-pretraining?

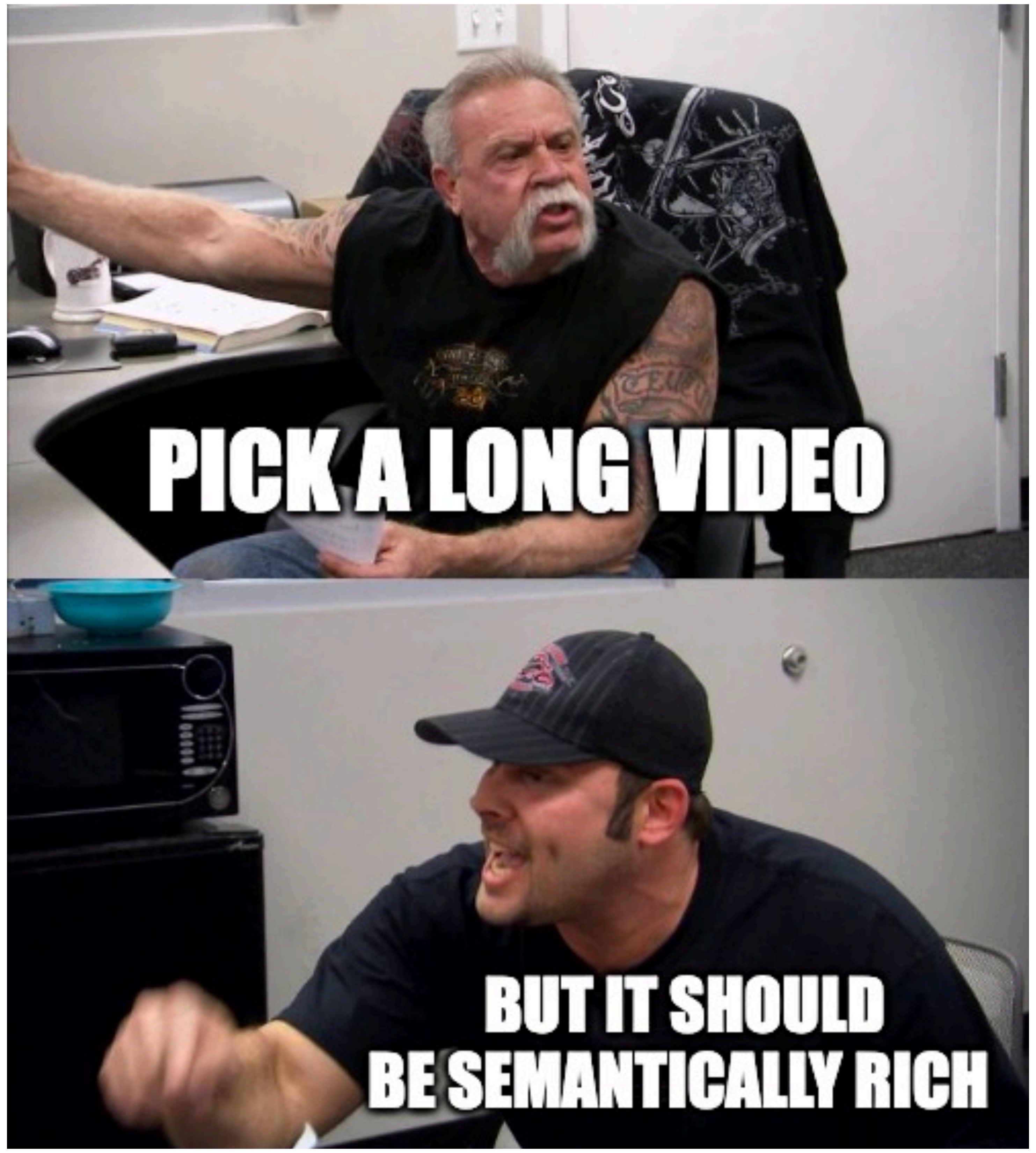
Study the extreme: try to learn from a **single video, from scratch.**

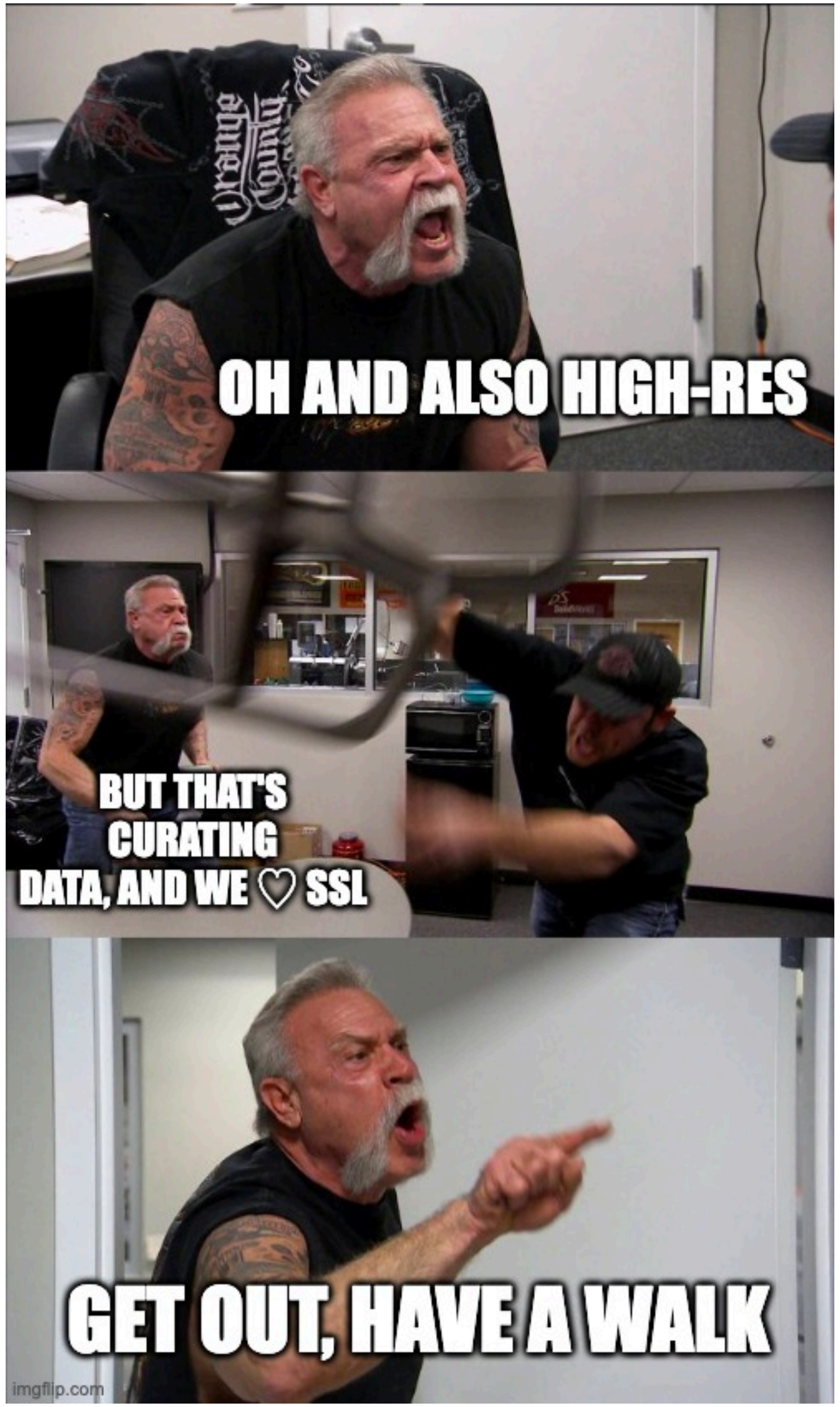
Motivated by: Asano Rupprecht, Vedaldi. *A critical analysis of self-supervision, or what we can learn from a single image*. ICLR 2020

Vermeer, The Milkmaid 1660

IS ALL WE NEED ONE LONG VIDEO WITH MANY DETAILS?

Us figuring out which video to use



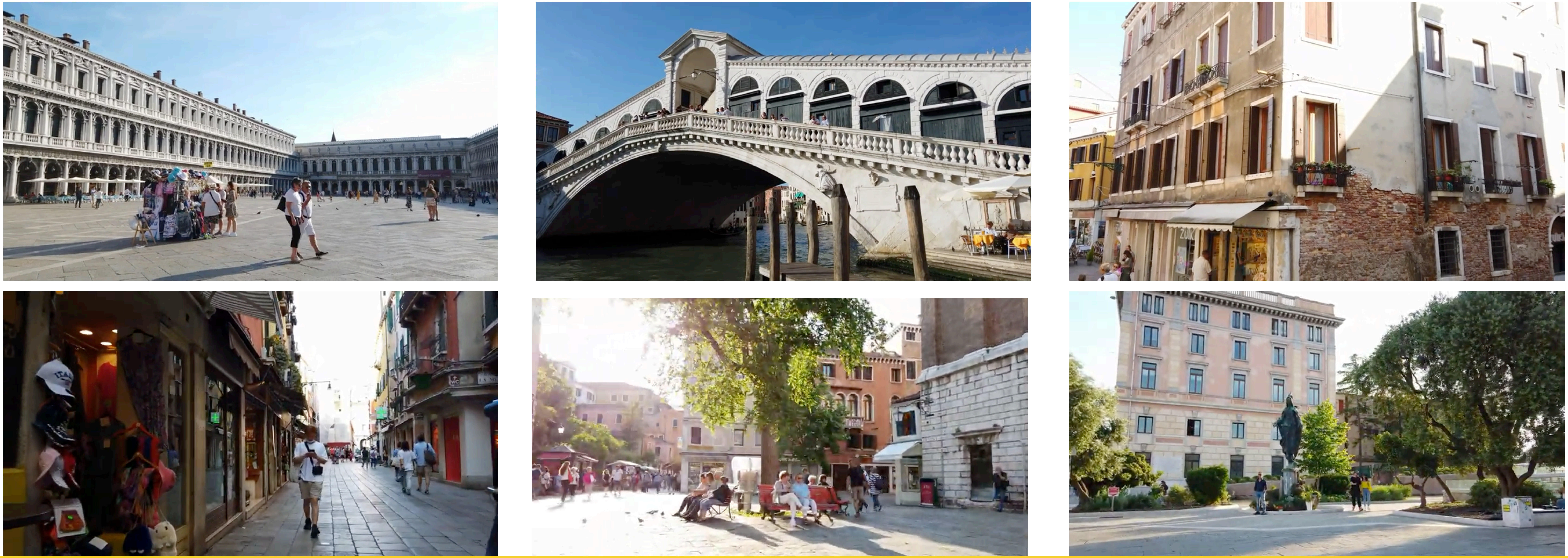


- ✓ Long
- ✓ High-res, smooth
- ✓ Semantically rich
- ✓ Scalable (we ❤️ SSL)

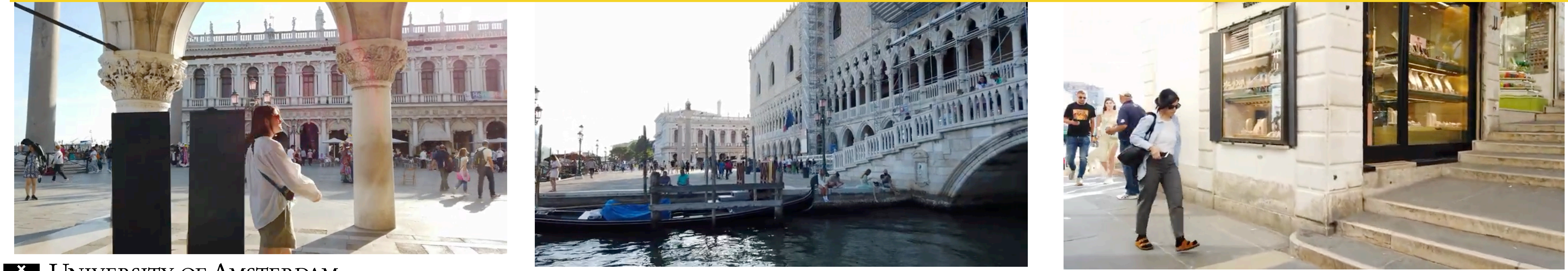
↓

Walking Tours

WTours proposed for learning video compression in ACCV 2022: Wiles et al. Compressed Vision for Efficient Video Understanding.



The dataset consists of 10x 4K videos of different cities' Walking Tours.



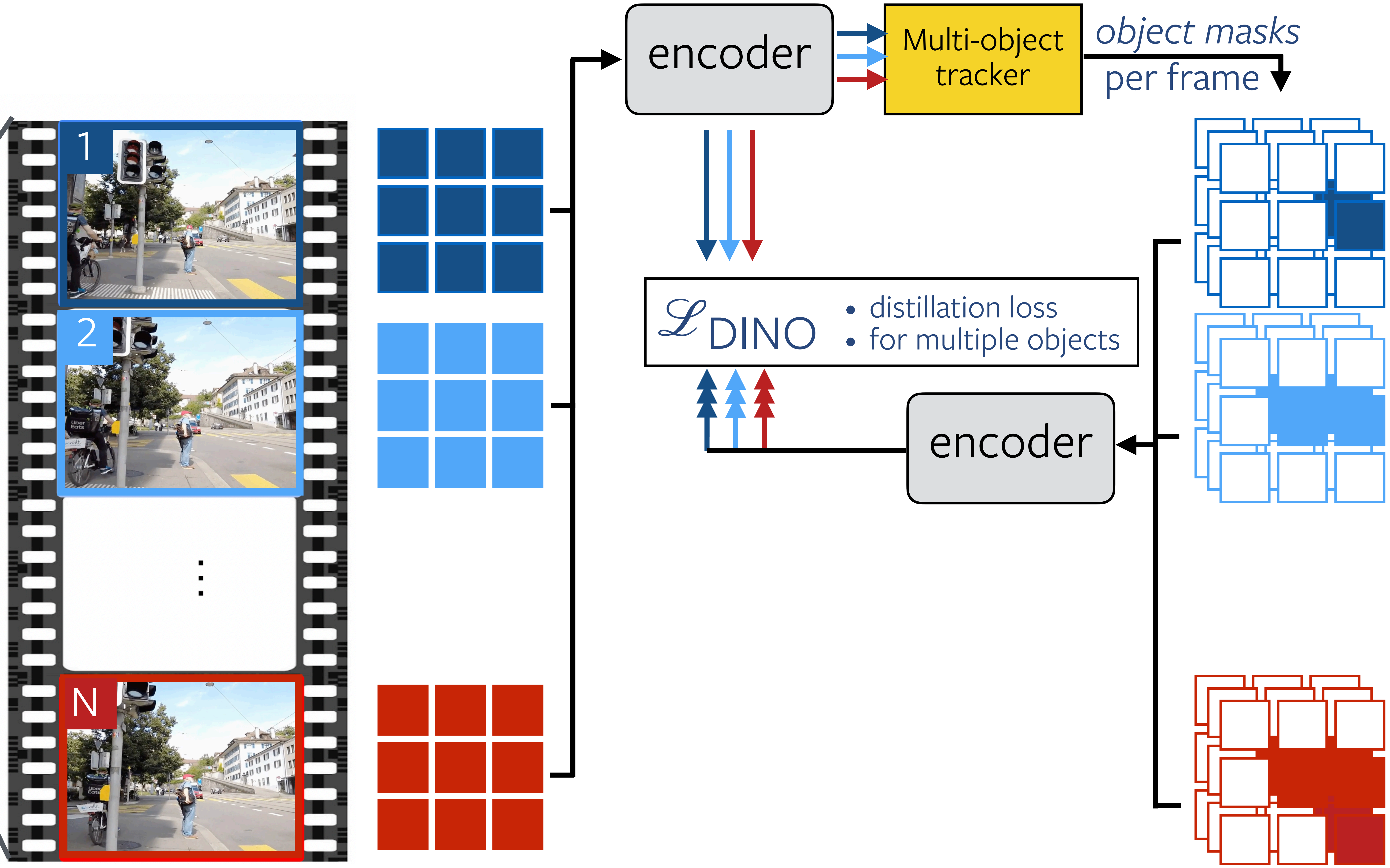
WT Venice: https://www.youtube.com/watch?v=fGX0Te6pFvk. CC-BY Poptravel.

Dora: **D**isc**o**ver and T**ra**ck

High-level idea:

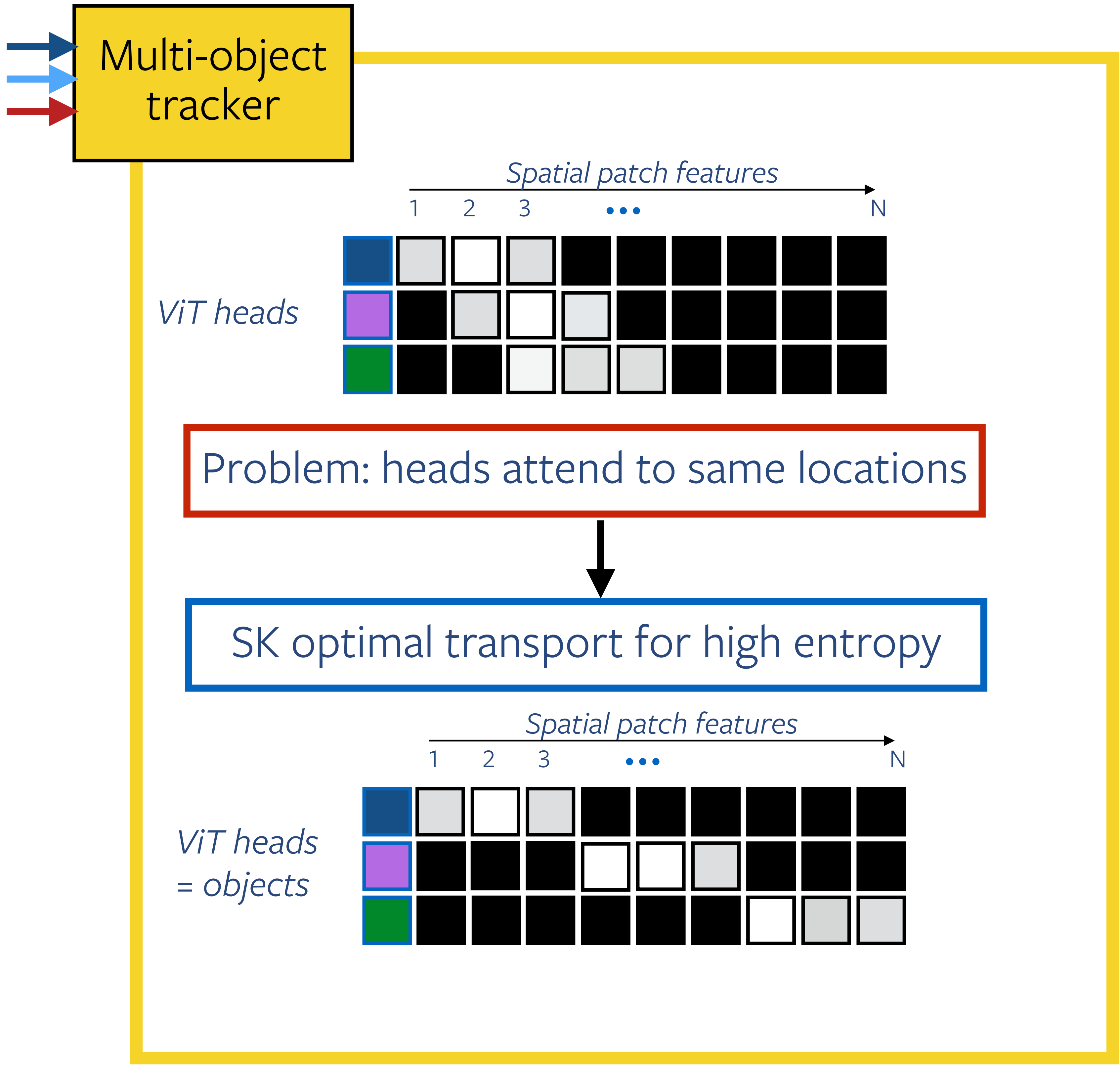
1) track multiple objects across time
2) enforce invariance of features across time

Much like Dora, we walk around and learn from what we see.

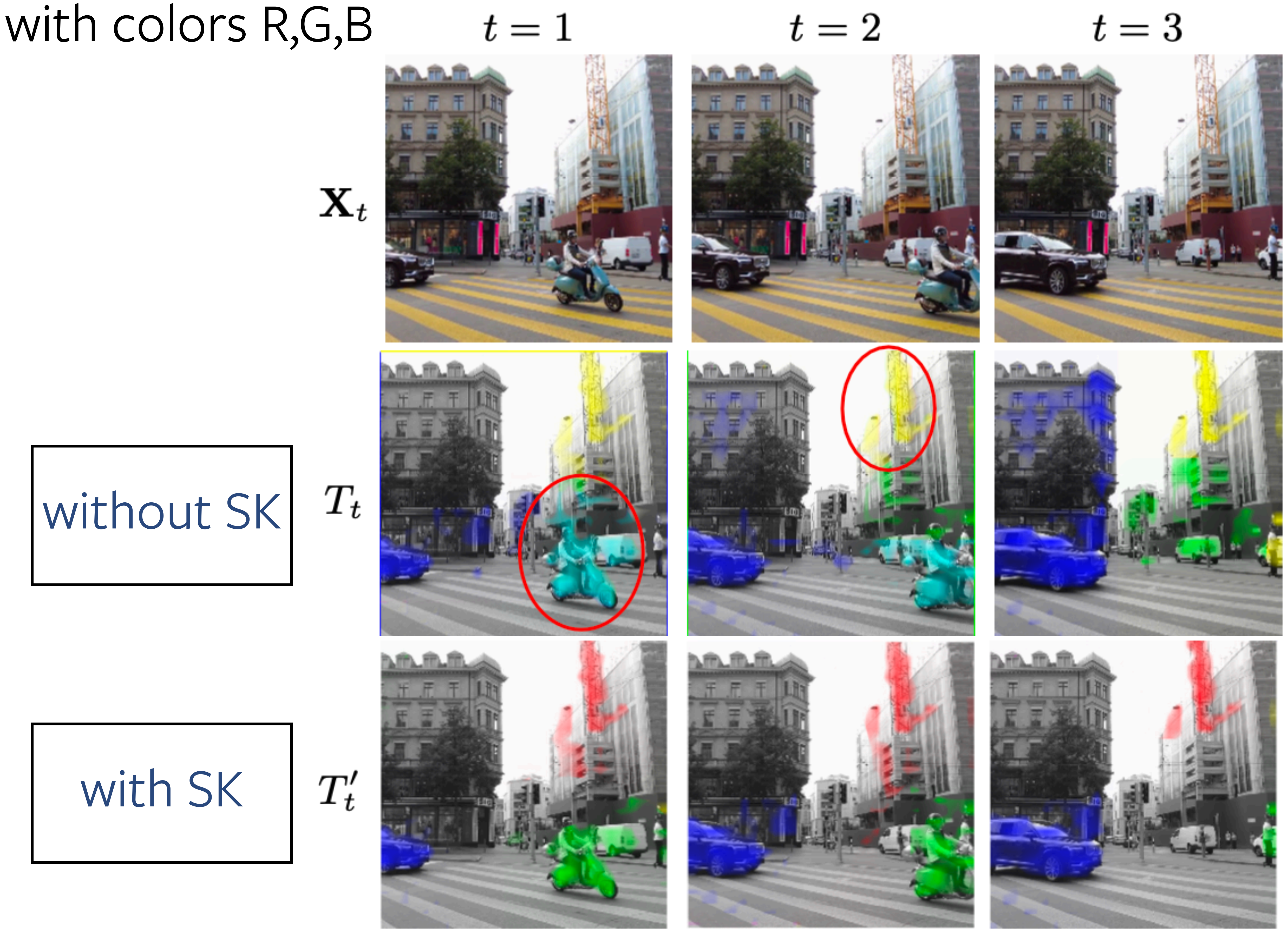


Venkataramanan, Rizve, Carreira, Asano*, Avrithis*. *Is ImageNet worth 1 video? Learning strong image encoders from 1 long unlabelled video.* ArXiv 2023

Spreading attention with Sinkhorn-Knopp



Visualise attention of
3 heads with colors R,G,B

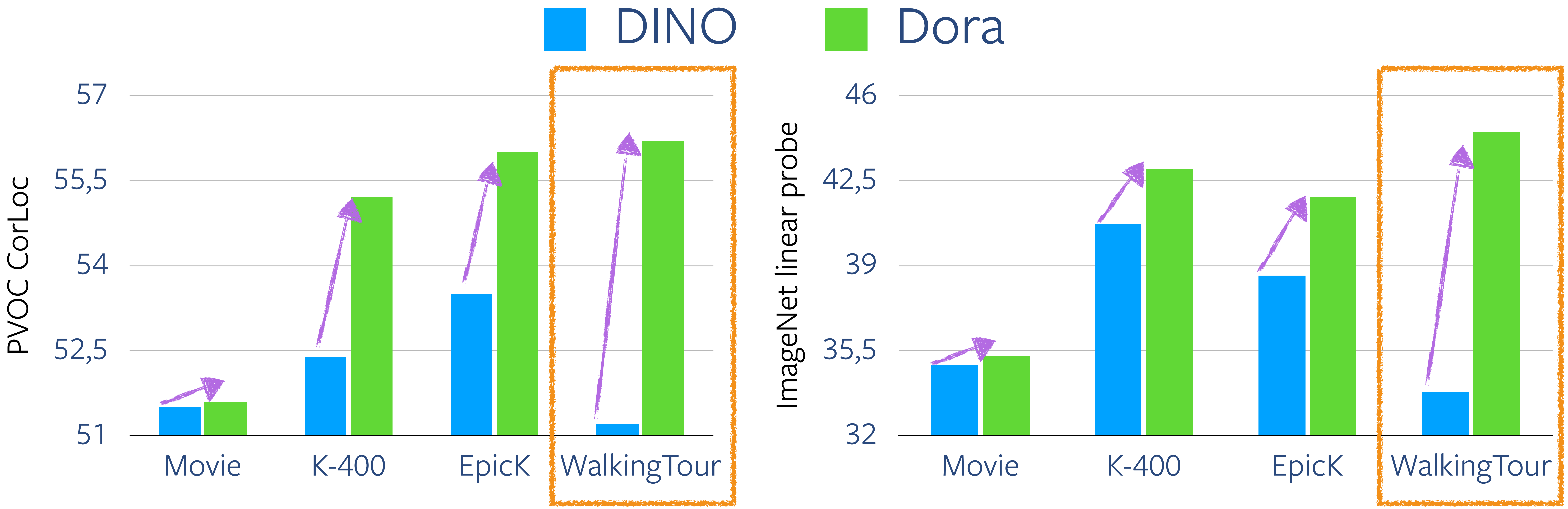


Venkataraman, Rizve, Carreira, Asano*, Avrithis*. *Is ImageNet worth 1 video? Learning strong image encoders from 1 long unlabelled video.* ArXiv 2023

More examples: multi-object tracking in a ViT *emerges*

Venkataramanan, Rizve, Carreira, Asano*, Avrithis*. *Is ImageNet worth 1 video? Learning strong image encoders from 1 long unlabelled video.* ArXiv 2023

Dora better than DINO
WT+ Dora: great match



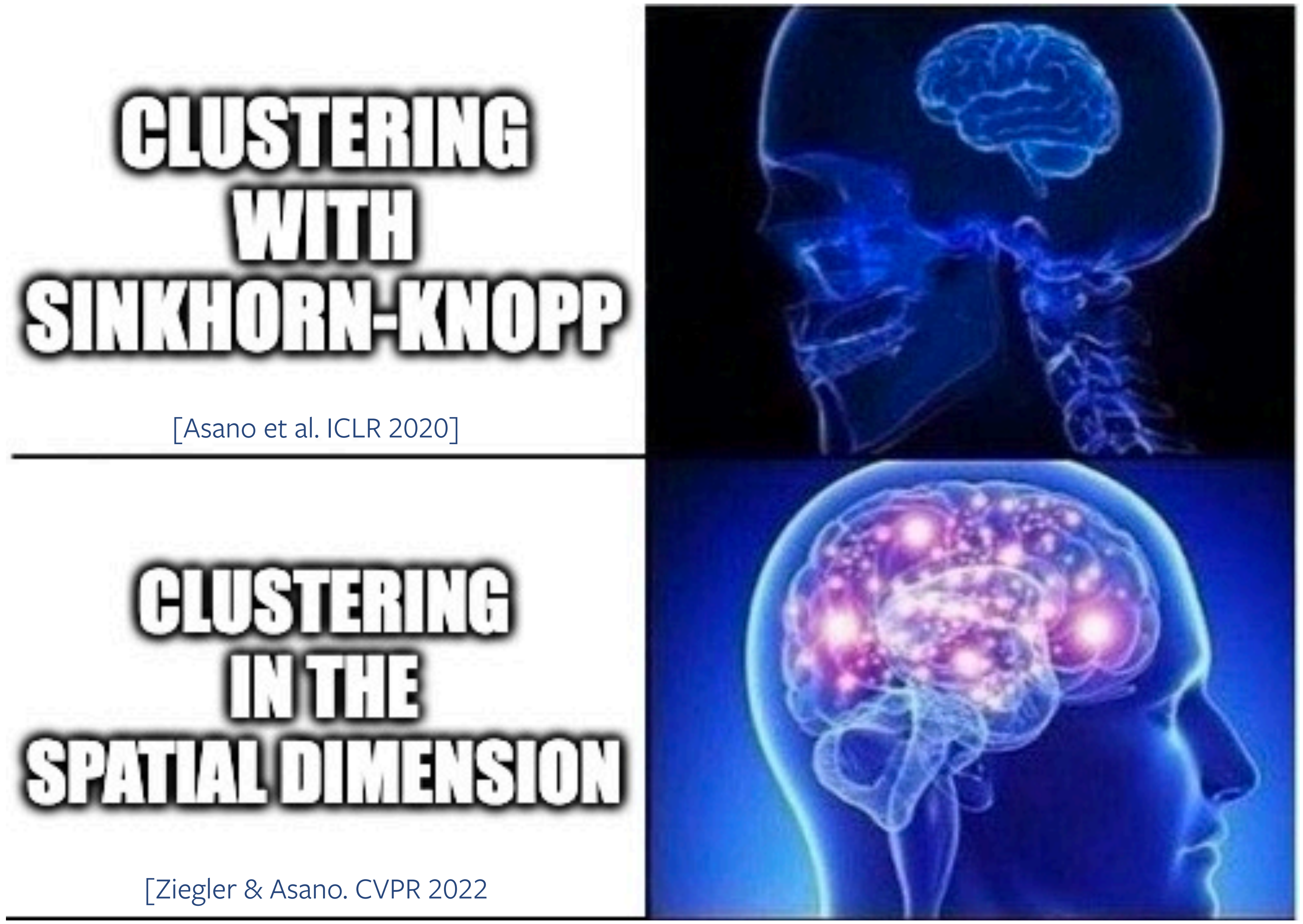
Venkataramanan, Rizve, Carreira, Asano*, Avrithis*. *Is ImageNet worth 1 video? Learning strong image encoders from 1 long unlabelled video.* ArXiv 2023

But how does it compare against ImageNet pretraining?

Dora (1WT) ~ on par with DINO (IN-1k)
Dora (10WT) > DINO (IN-1k) everywhere

Venkataraman, Rizve, Carreira, Asano*, Avrithis*. *Is ImageNet worth 1 video? Learning strong image encoders from 1 long unlabelled video.* ArXiv 2023

Summary



[Asano et al. ICLR 2020]

[Ziegler & Asano. CVPR 2022

[Salehi et al.. ICCV 2023]

[Vekataramaran et al. Arxiv 2023]

1 TimeTuning:
DINO as init & use temporal info of videos.

How powerful is time without image-pretraining?

2 Study the extreme: try to learn from a **single video**, **from scratch**.

✅ *Videos allow for strong self-supervised learning*

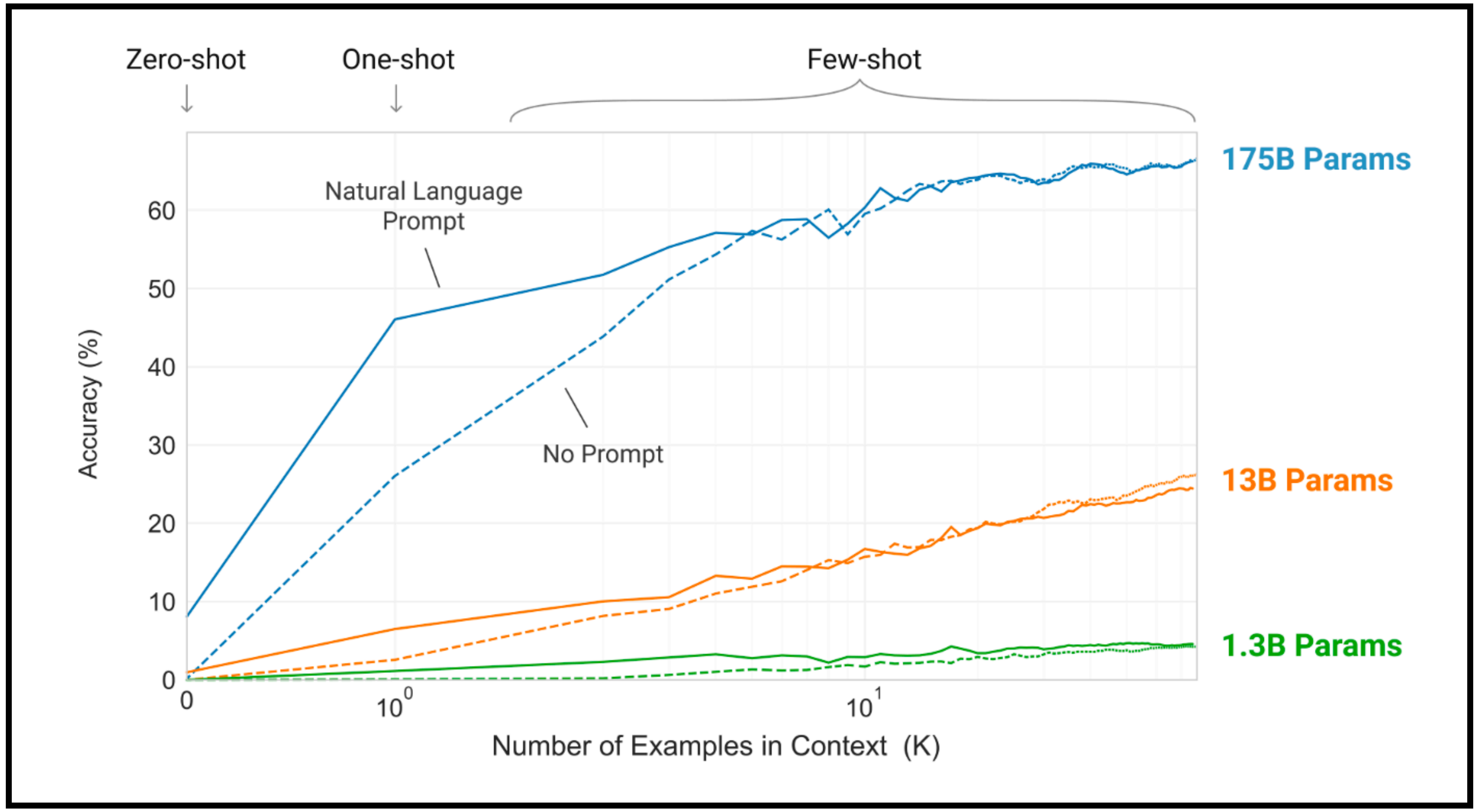
VLMs?
Can we reduce the need for paired data?

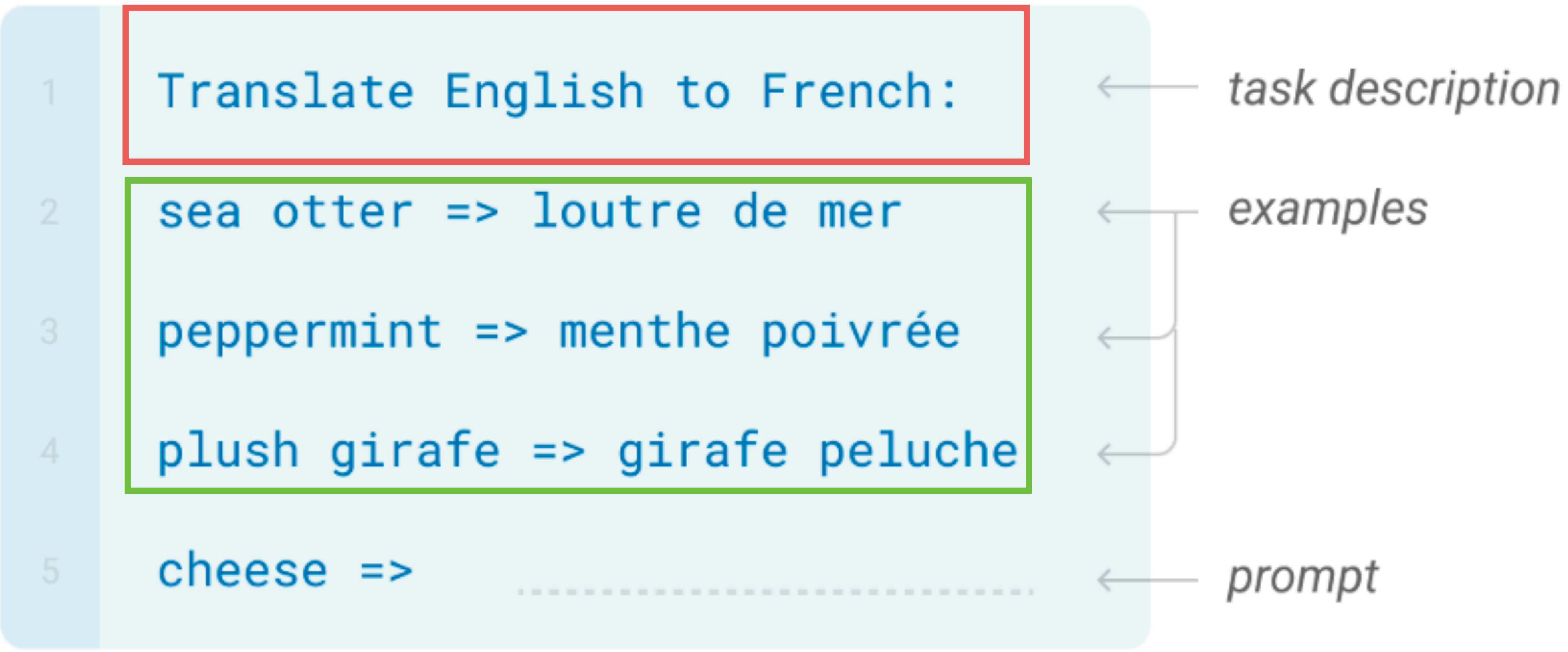
3 Use self-supervised features to create noisily paired data.

Dirk Jacobsz, Painting a Portrait of His Wife,1550

REDUCING THE NEED FOR PAIRED TEXT-IMAGE DATA

Why are LLMs so sexy? a) scaling behavior, b) In-context Learning!





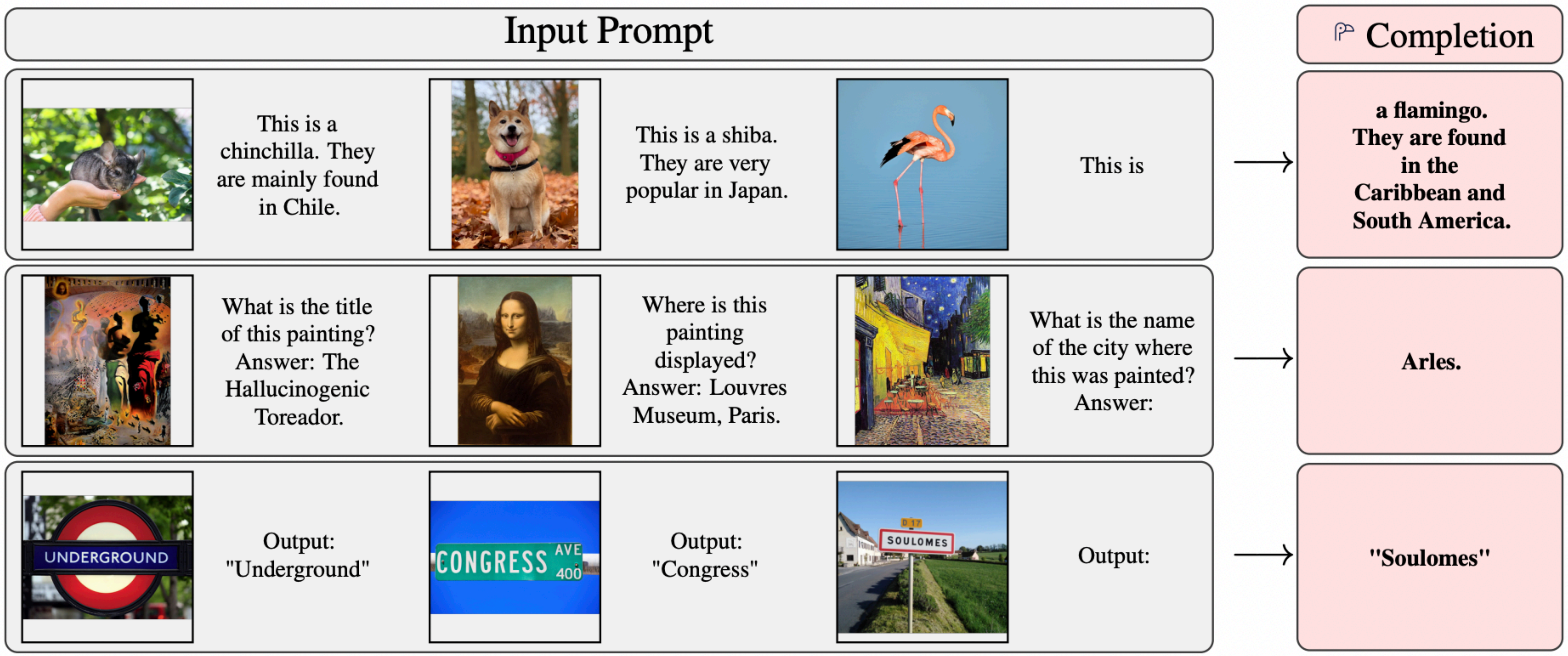
The possibility to

- *define a task* and the
- *learning-like* behaviour via few-shot examples
- with a *frozen model*

allows big scaling

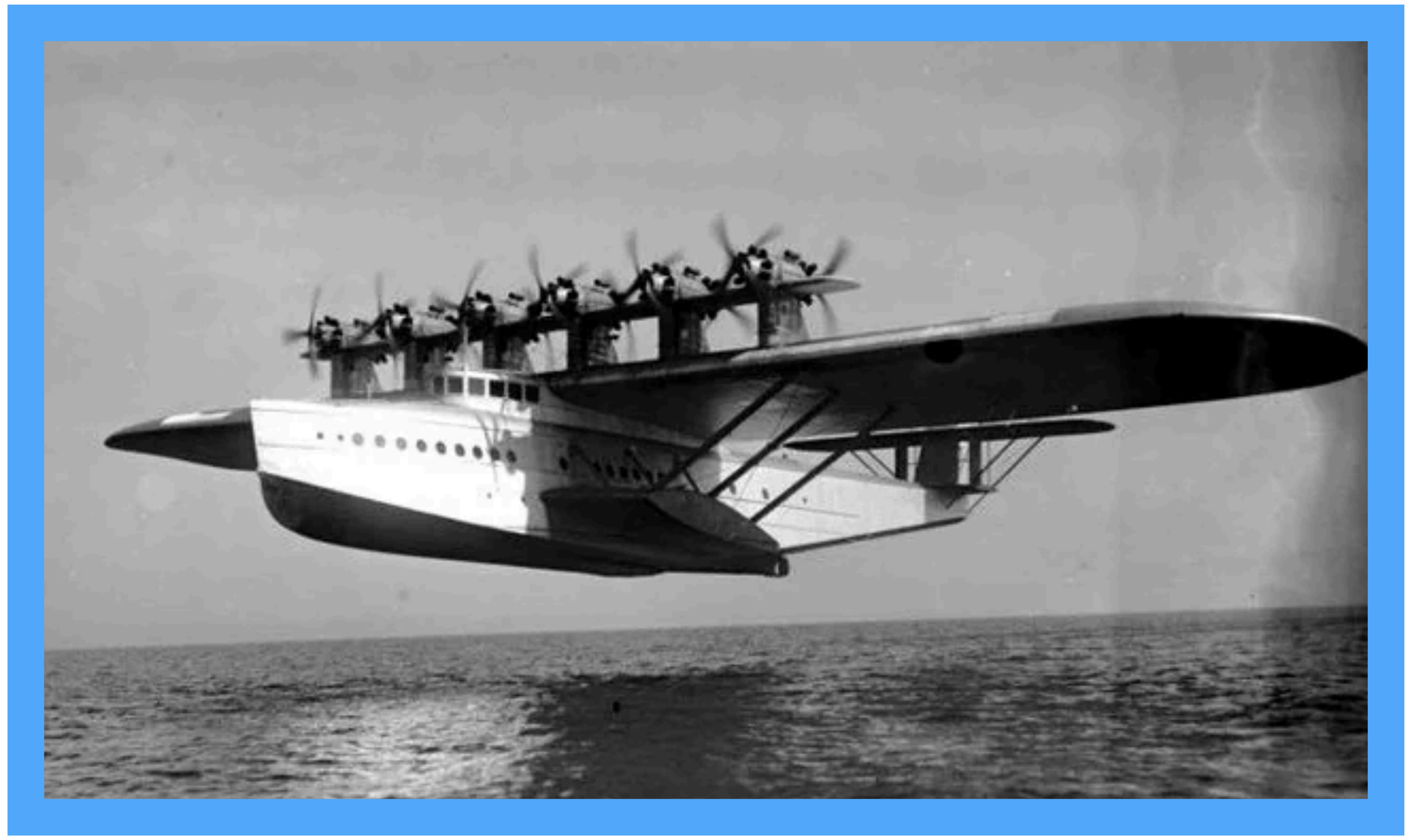
Brown et al. *Language models are few-shot learners.* NeurIPS 2020

In-context Learning emerges also for Visual Language Models

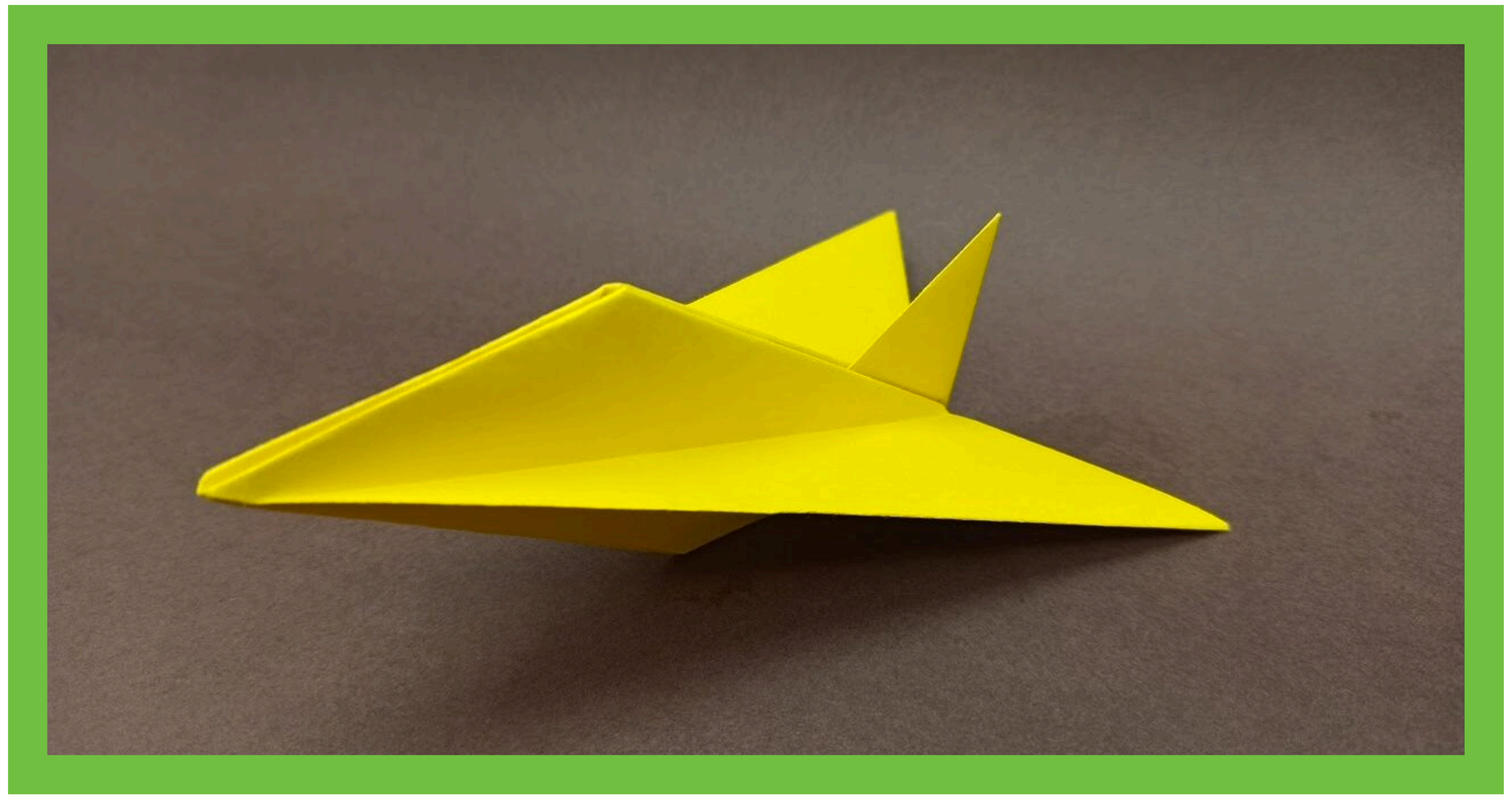


Tsimpoukelli et al. *Multimodal Few-Shot Learning with Frozen Language Models.* NeurIPS 2021
Alayrac et al. *Flamingo: a Visual Language Model for Few-Shot Learning.* NeurIPS 2022

But just because it emerges for 6B+ sized models, does it mean we cannot do this with more light-weight ones?

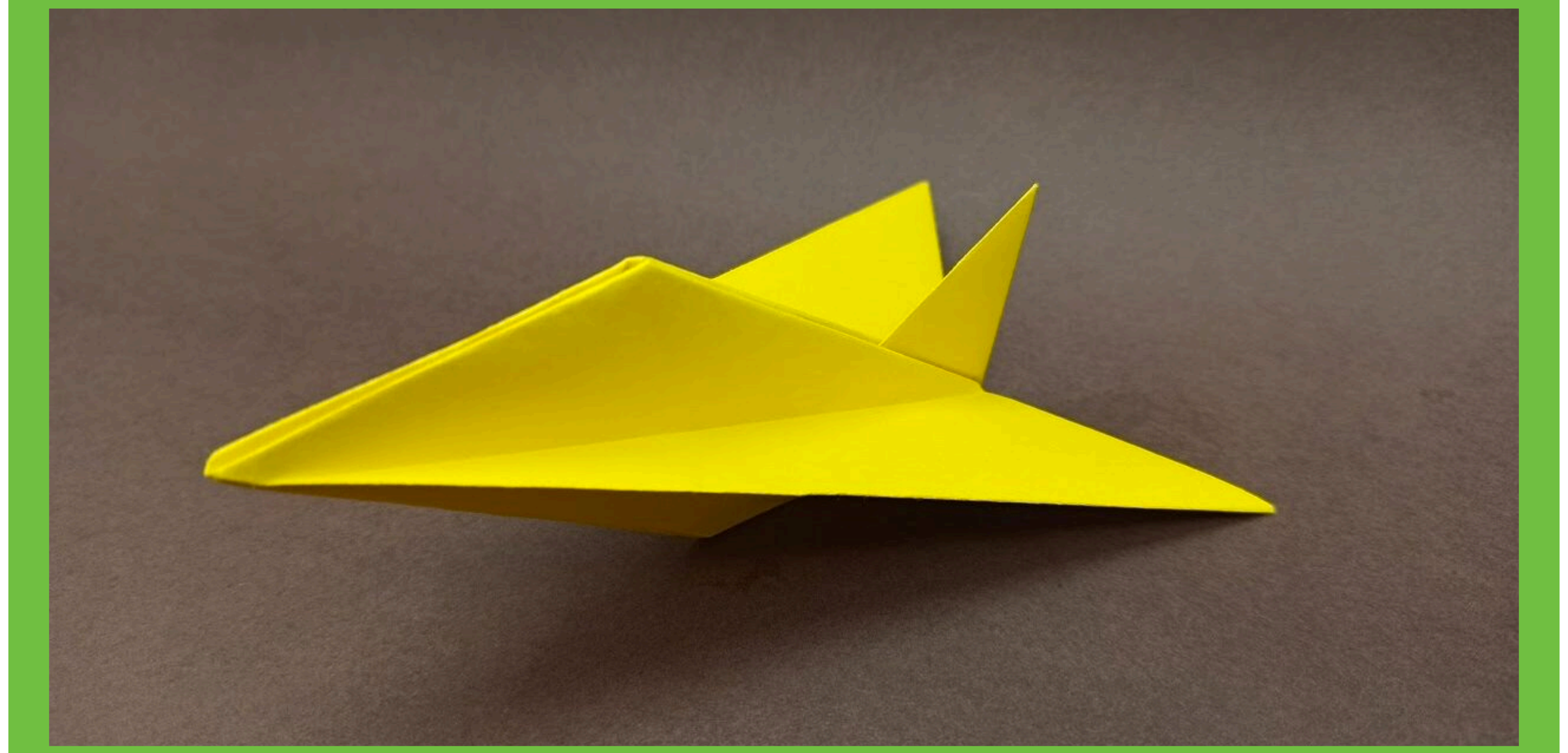


?



Our goal

open-ended multi-modal ICL
with small VLMs
<u>without using supervised data.</u>

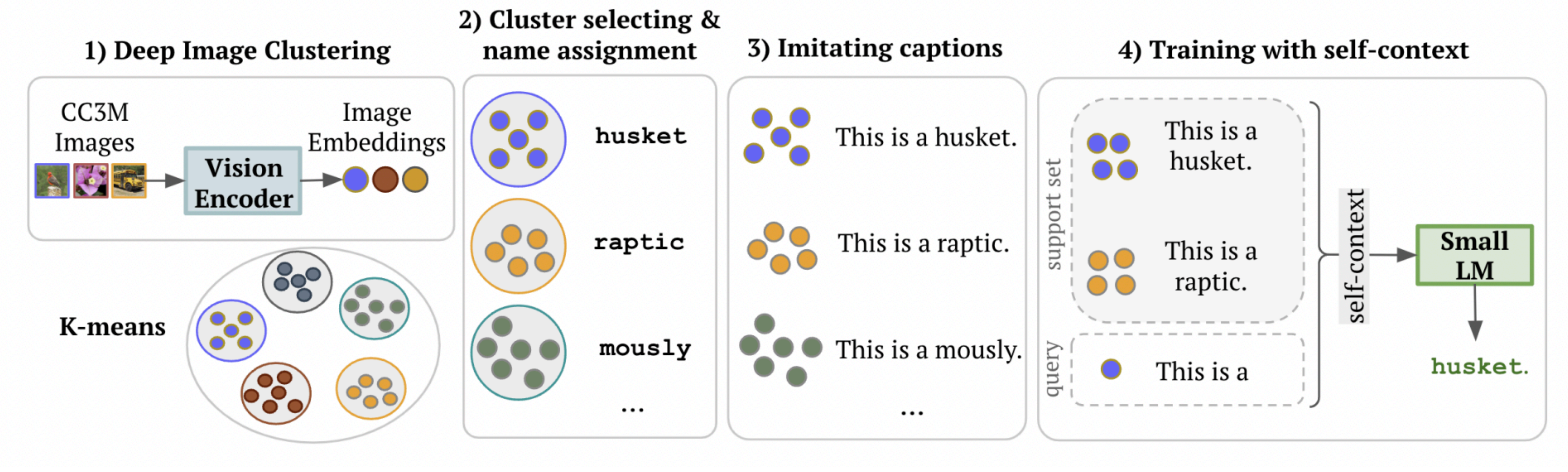


Why?

- Ultimately, paired data is rare
- ICL as algorithm: symbols replaceable
- As an existence-proof

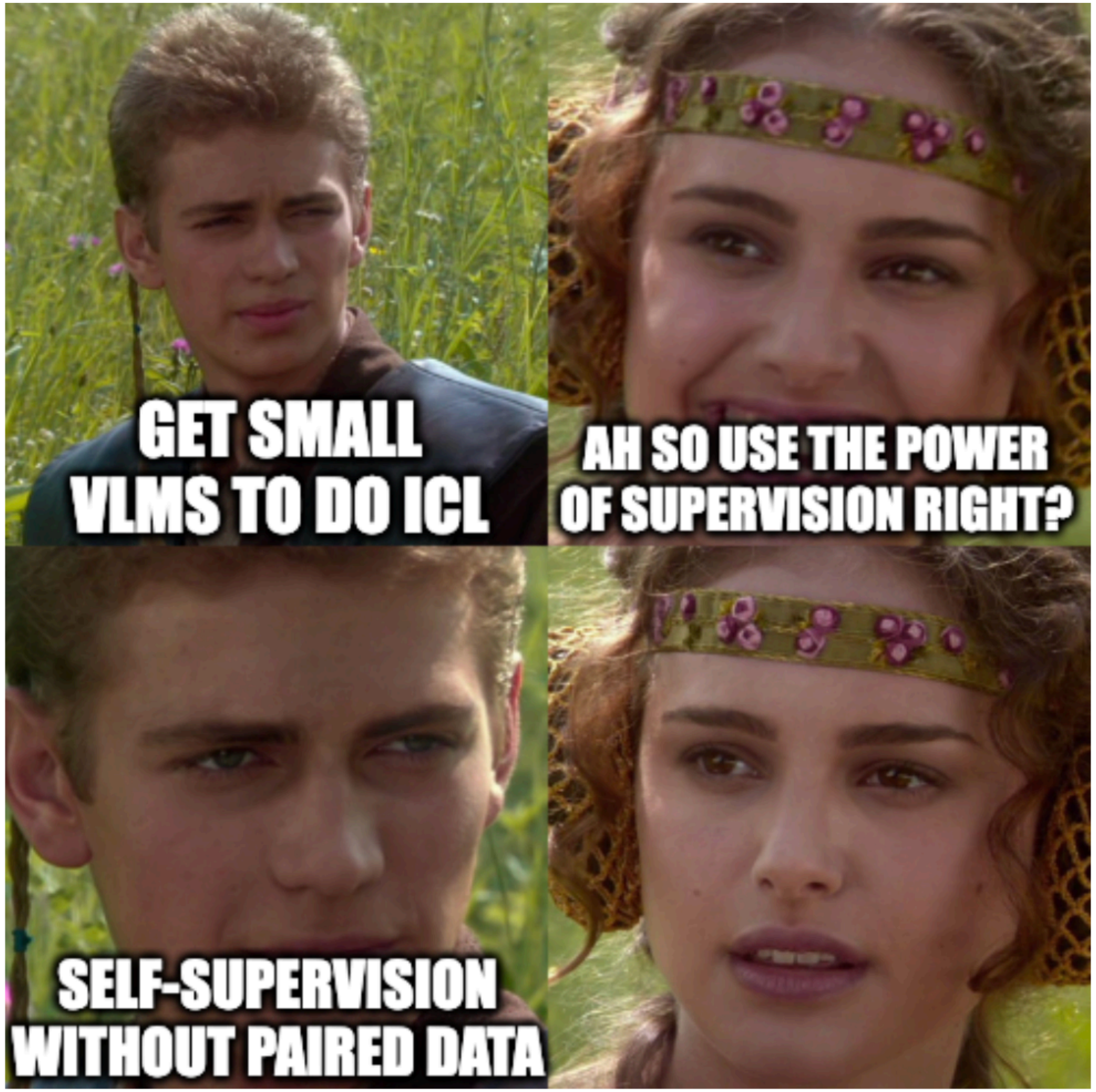
Derakhshani, Najdenkoska, Snoek, Worring, Asano. *Self-Supervised Open-Ended Classification with Small Visual Language Models.* ArXiv 2023.

How? Our method simply *mimics* supervised data by using SSL

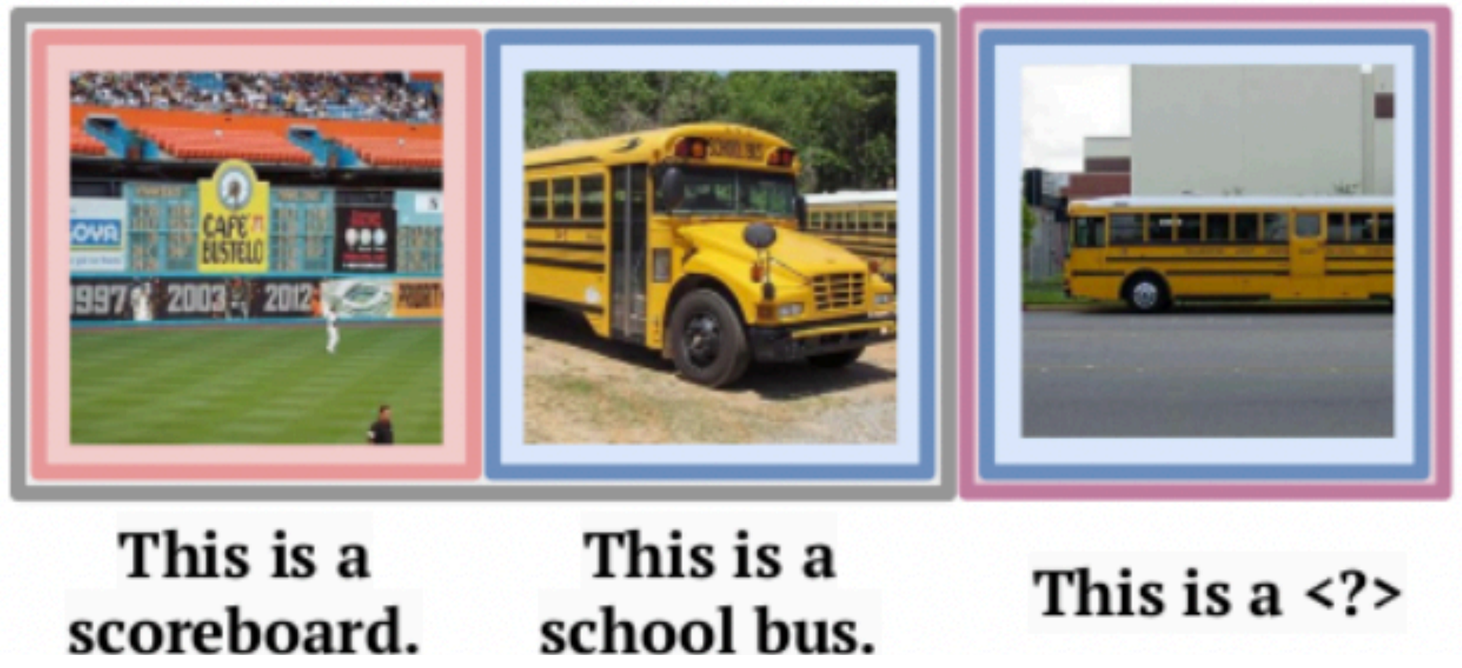


Derakhshani, Najdenkoska, Snoek, Worring, Asano. *Self-Supervised Open-Ended Classification with Small Visual Language Models.* ArXiv 2023.

Where are we?


GET SMALL VLMS TO DO ICL
AH SO USE THE POWER OF SUPERVISION RIGHT?
SELF-SUPERVISION WITHOUT PAIRED DATA

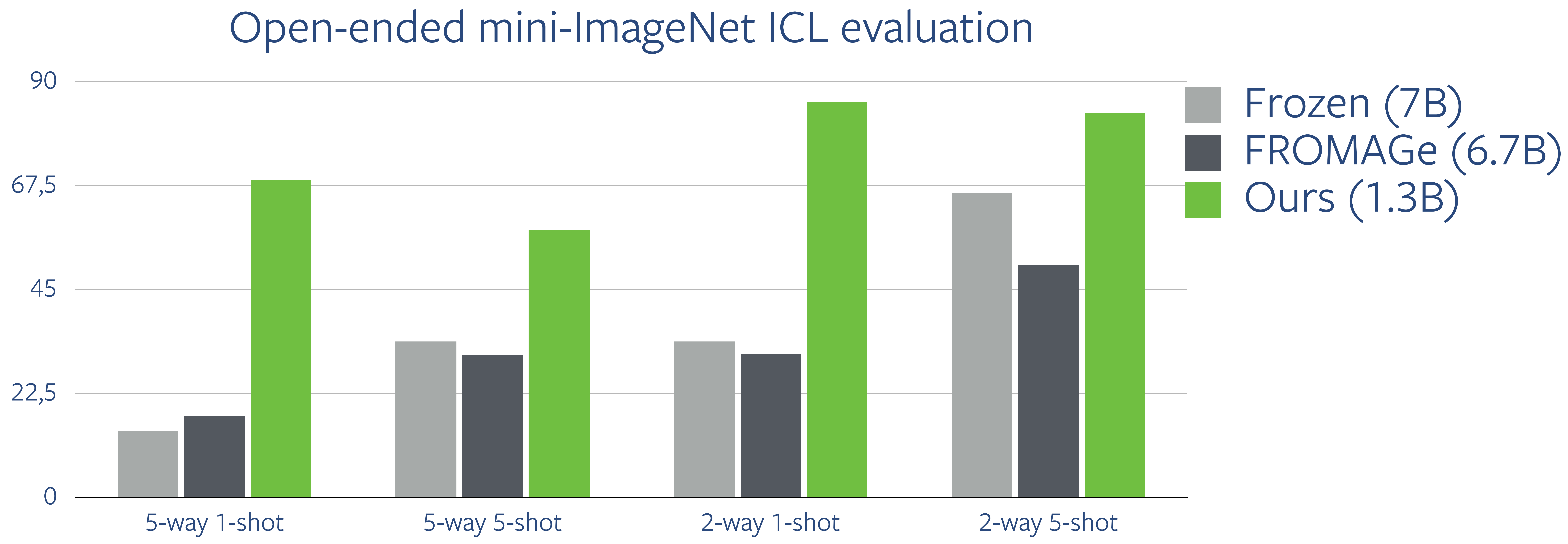
We train with fake names, but evaluation works just fine!



❌ ***ClipCap***: distinguished from all other by its long slender torso.
❌ ***FROMAGe***: school bus that is parked in the school yard.
✅ ***SeCAt (Ours)*: school bus.**

Derakhshani, Najdenkoska, Snoek, Worring, Asano. *Self-Supervised Open-Ended Classification with Small Visual Language Models.* ArXiv 2023.

Simple-as-that; beats 6B-sized models on open-ended multi-modal classif.



Derakhshani, Najdenkoska, Snoek, Worring, Asano. *Self-Supervised Open-Ended Classification with Small Visual Language Models.* ArXiv 2023.

Team for the works presented

TimeTuning


Mohammadreza Salehi


Efstratios Gavves


Cees G. M. Snoek


Yuki M. Asano

WTour Dora


Shashanka Venkataramanan


Mamshad N Rizve


Joao Carreira


Yuki M. Asano*


Yannis Avrithis*

SeCAt


Mohammad M Derakhshani[1]

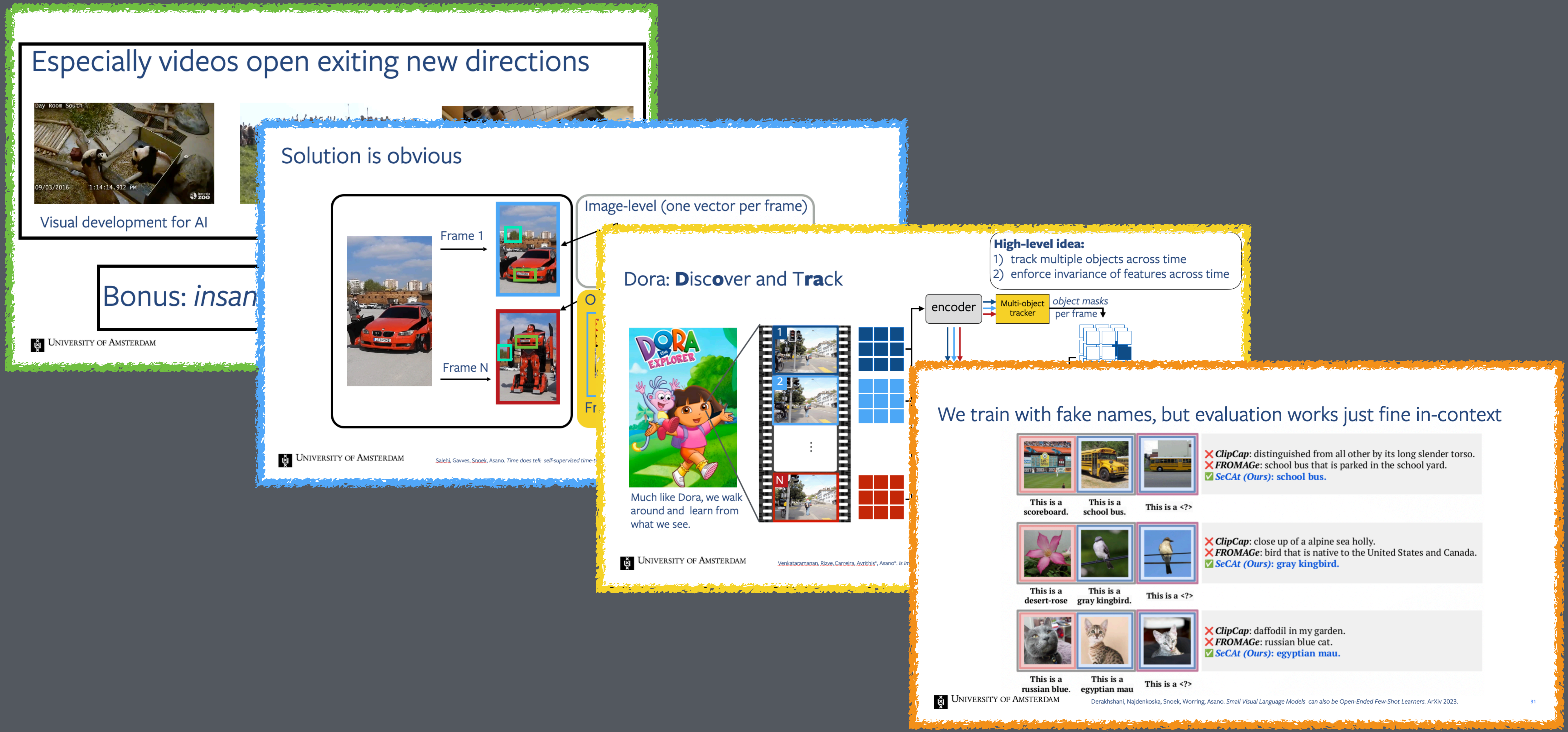

Ivona Najdenkoska[1]


Cees G. M. Snoek*


Marcel Worring*


Yuki M. Asano*

Salehi, Gavves, Snoek, Asano. *Time does tell: self-supervised time-tuning of dense image representations.* ICCV 2023
Venkataramanan, Rizve, Carreira, Avrithis, Asano. *Is ImageNet worth 1 video? Learning strong image encoders from 1 long unlabelled video.* ArXiv 2023
Derakhshani, Najdenkoska, Snoek, Worring, Asano. *Self-Supervised Open-Ended Classification with Small Visual Language Models.* ArXiv 2023.
[1]: co-first authors; *: co-last authors



Future Foundation Models will be massively pretrained with videos.
Current multi-modal training will become only the cherry on top.