UNIVERSITY OF AMSTERDAM

Prof. dr. Cees Snoek
University of Amsterdam

Head of VIS lab
Director QUVA ICAI-lab
Director Atlas ICAI-lab
Director HAVA-lab
CSO, Kepler Vision
Scientific Director Amsterdam AI

https://www.ceessnoek.info

Video Deep Learning

VIS LAB
VIDEO & IMAGE SENSE LAB

Hi I am Cees Snoek

1996-2000	**BSc/MSc Student** - UvA
2001-2005	**PhD** - UvA, CMU
2005-2011	**Postdoc** - UvA, UC Berkeley
2011-2017	**Assistant/Associate Professor** – UvA

2011-2014: Head of R&D – Euvision Technologies
2014-2017: Managing Principal Engineer - Qualcomm

2018- **Full Professor** – UvA

2018-present CSO – Kepler Vision
2018-2022: Director Master AI
2022-2024: Director ELLIS Unit Amsterdam
2022-present: Scientific Director Amsterdam AI

https://www.ceessnoek.info

Video & Image Sense Lab

We make sense of video and images with artificial and human intelligence

Prof. dr. Cees Snoek

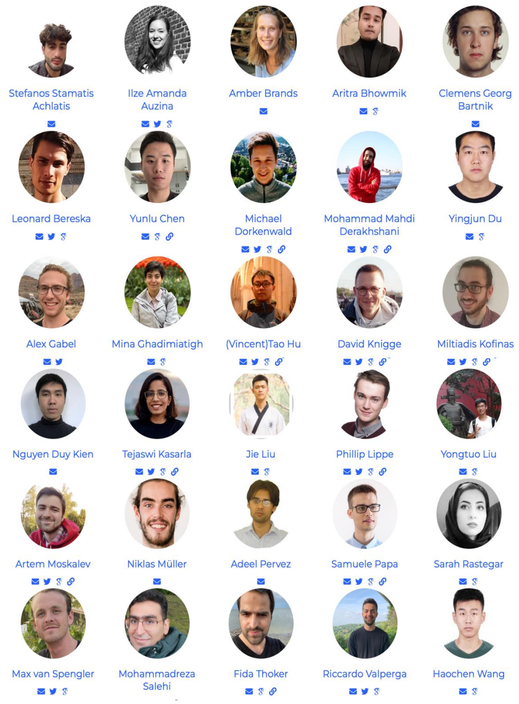
Research Staff

Dr. Efstratios Gavves

Dr. Pascal Mettes

Dr. Iris Groen

30+ PhD's & Postdocs



Lab Collaborations

Qualcomm Elekta

tomtom HAVA-Lab

2 Spin-offs

KEPLER VISION TECHNOLOGIES

Lyds.Ai

https://ivi.fnwi.uva.nl/vislab/

Video & Image Sense Lab

We make sense of video and images with artificial and human intelligence

Prof. dr. Cees Snoek

Some of the relevant AI courses we teach:

Bachelor AI

Neural Networks

Master AI

Deep Learning I
Deep Learning II
Foundation Models

Research Staff

Dr. Efstratios Gavves

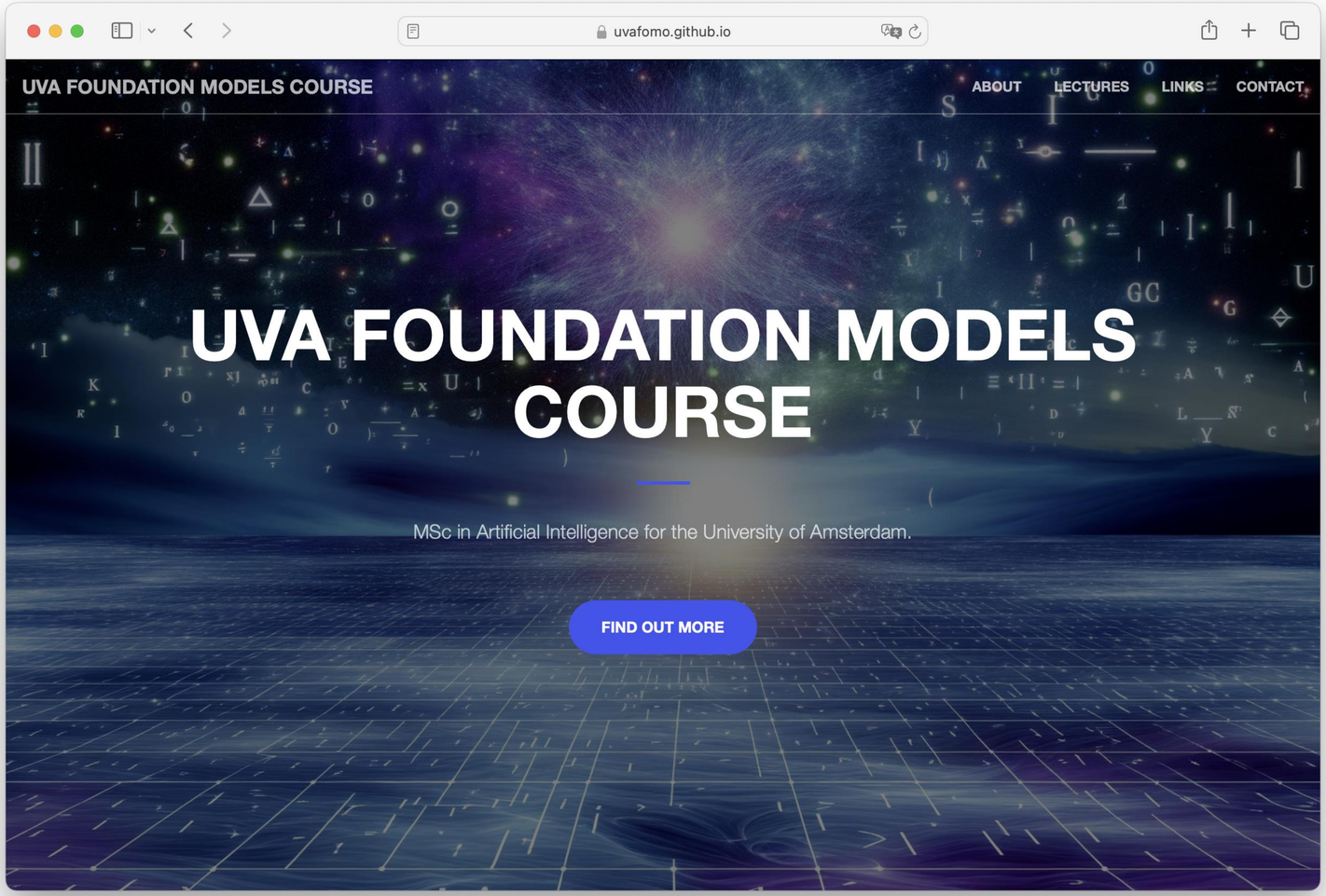
Dr. Pascal Mettes

Dr. Iris Groen

https://ivi.fnwi.uva.nl/vislab/

Block 5: Foundation Models

Prof. dr. Cees Snoek



UNIVERSITY OF AMSTERDAM

Prof. dr. Cees Snoek
University of Amsterdam

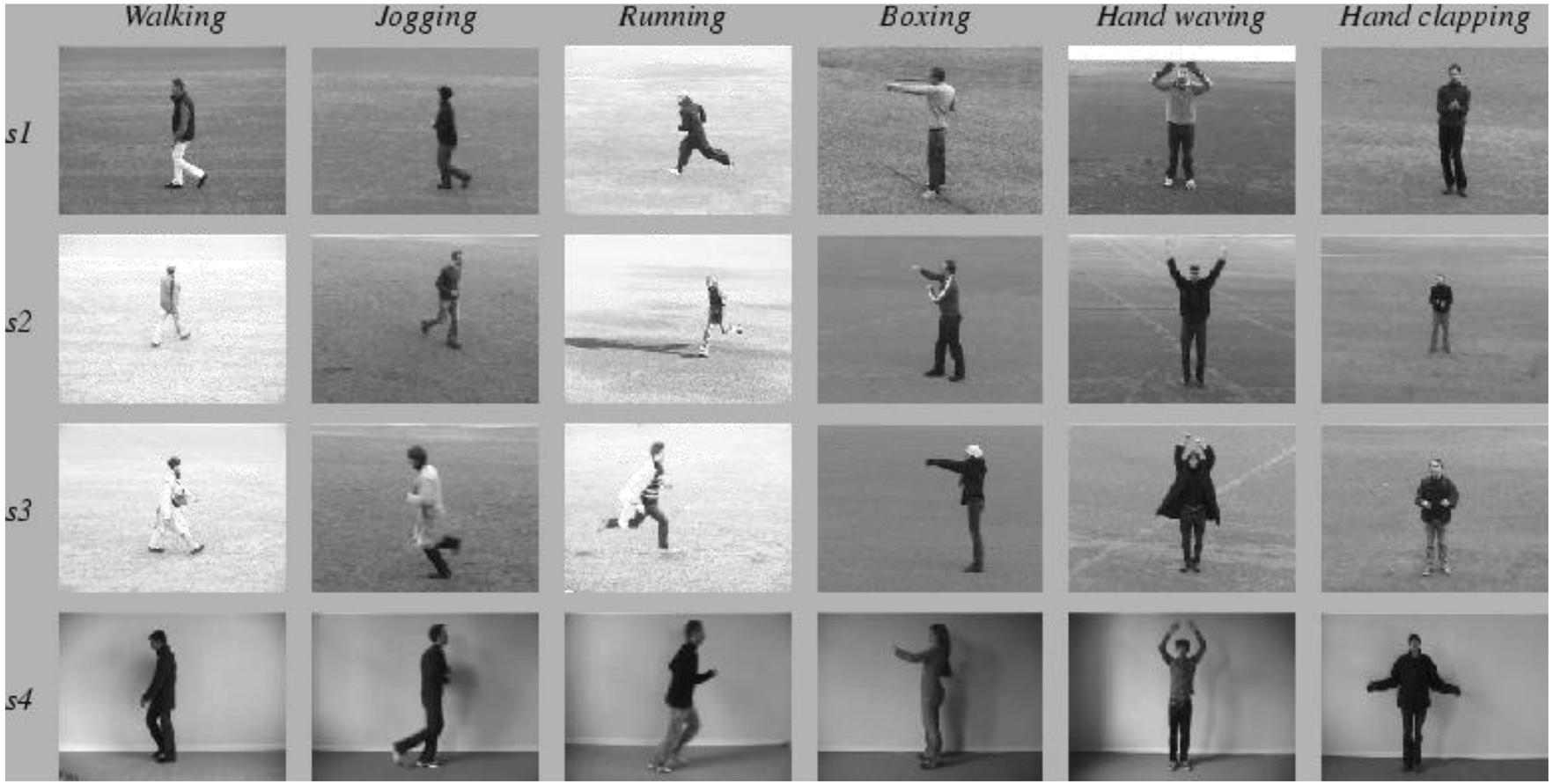
Head of VIS lab
Director QUVA ICAI-lab
Director Atlas ICAI-lab
Director HAVA-lab
CSO, Kepler Vision
Scientific Director Amsterdam AI

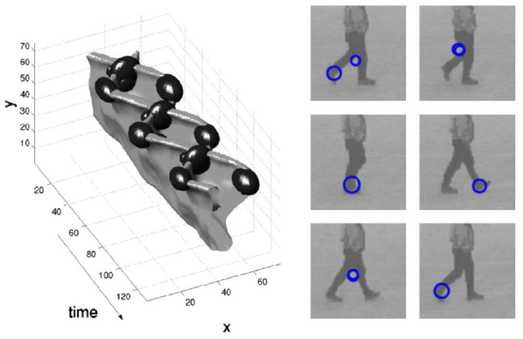
https://www.ceessnoek.info

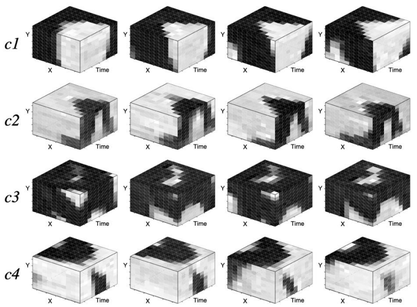
Video Deep Learning

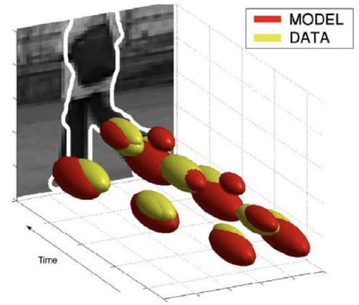
VIS LAB
VIDEO & IMAGE SENSE LAB

How it started…





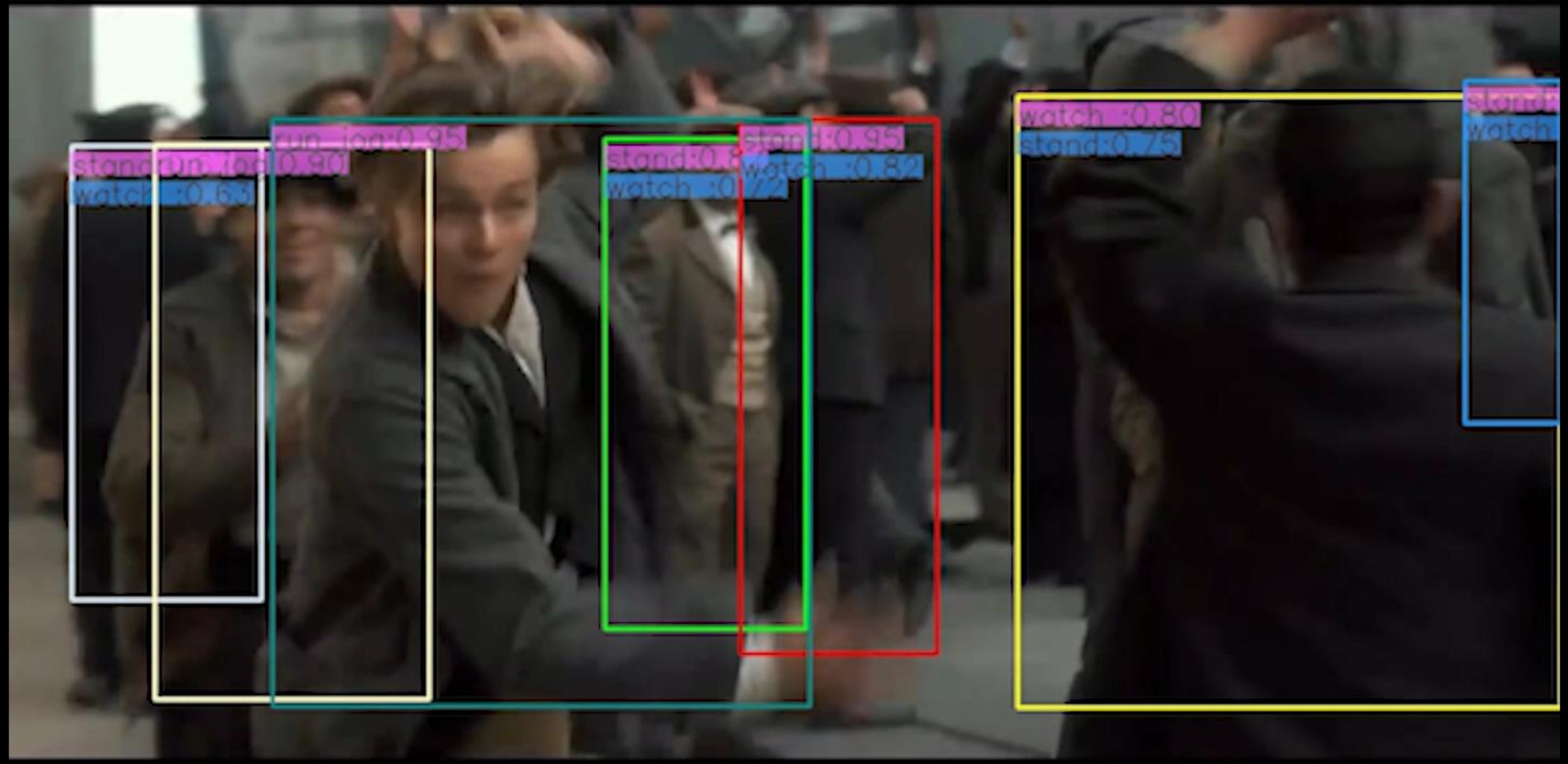




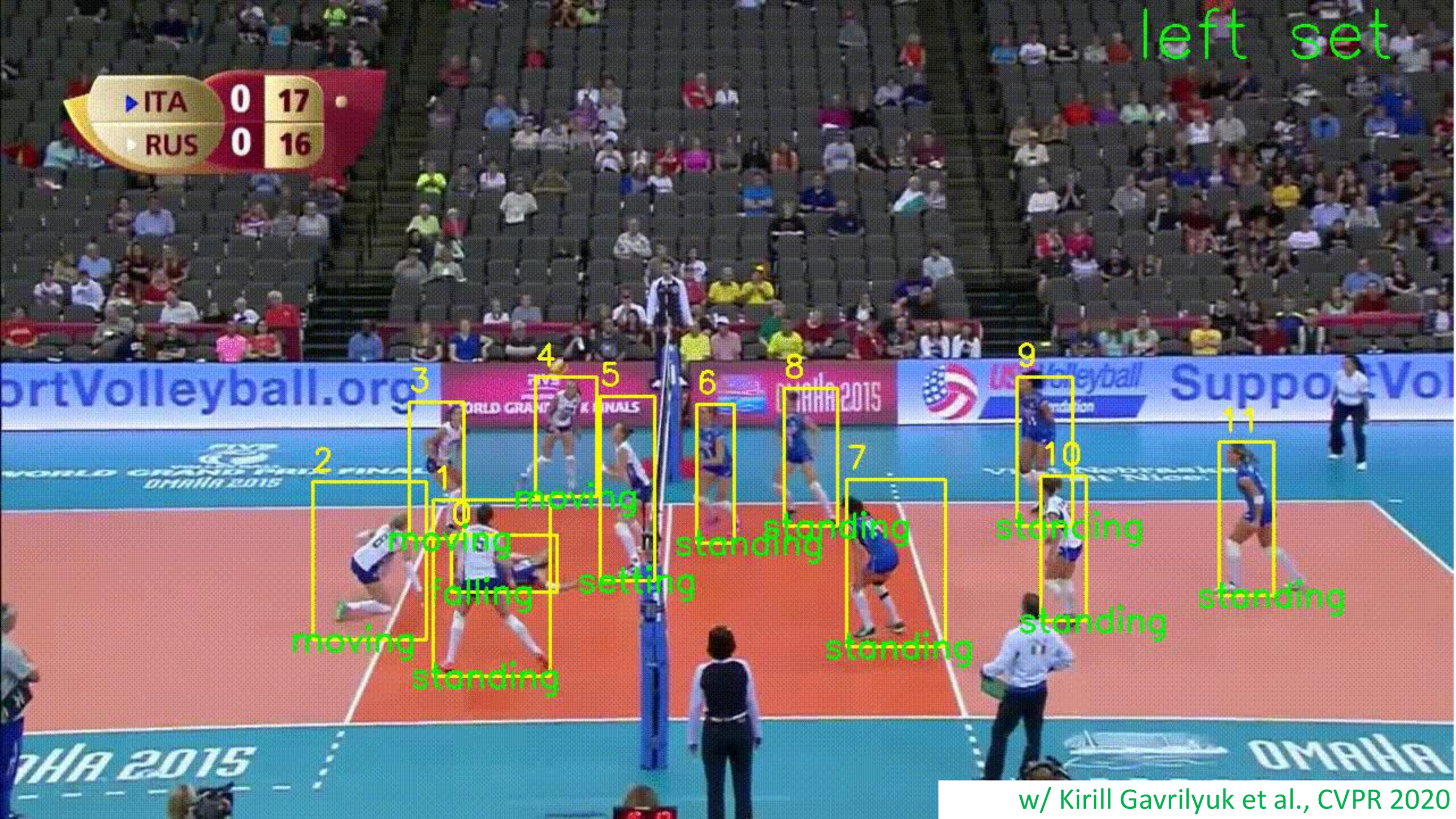
Laptev & Lindeberg, ICCV 2003

Du Tran et al., ICCV 2015

How it’s going…



w/ Jiaojiao Zhao et al., CVPR 2022



w/ Kirill Gavrilyuk et al., CVPR 2020


“gray dog running on a leash during dog show”
w/ Kirill Gavrilyuk Amir Ghodrati, & Zhenyang Li, CVPR 2019

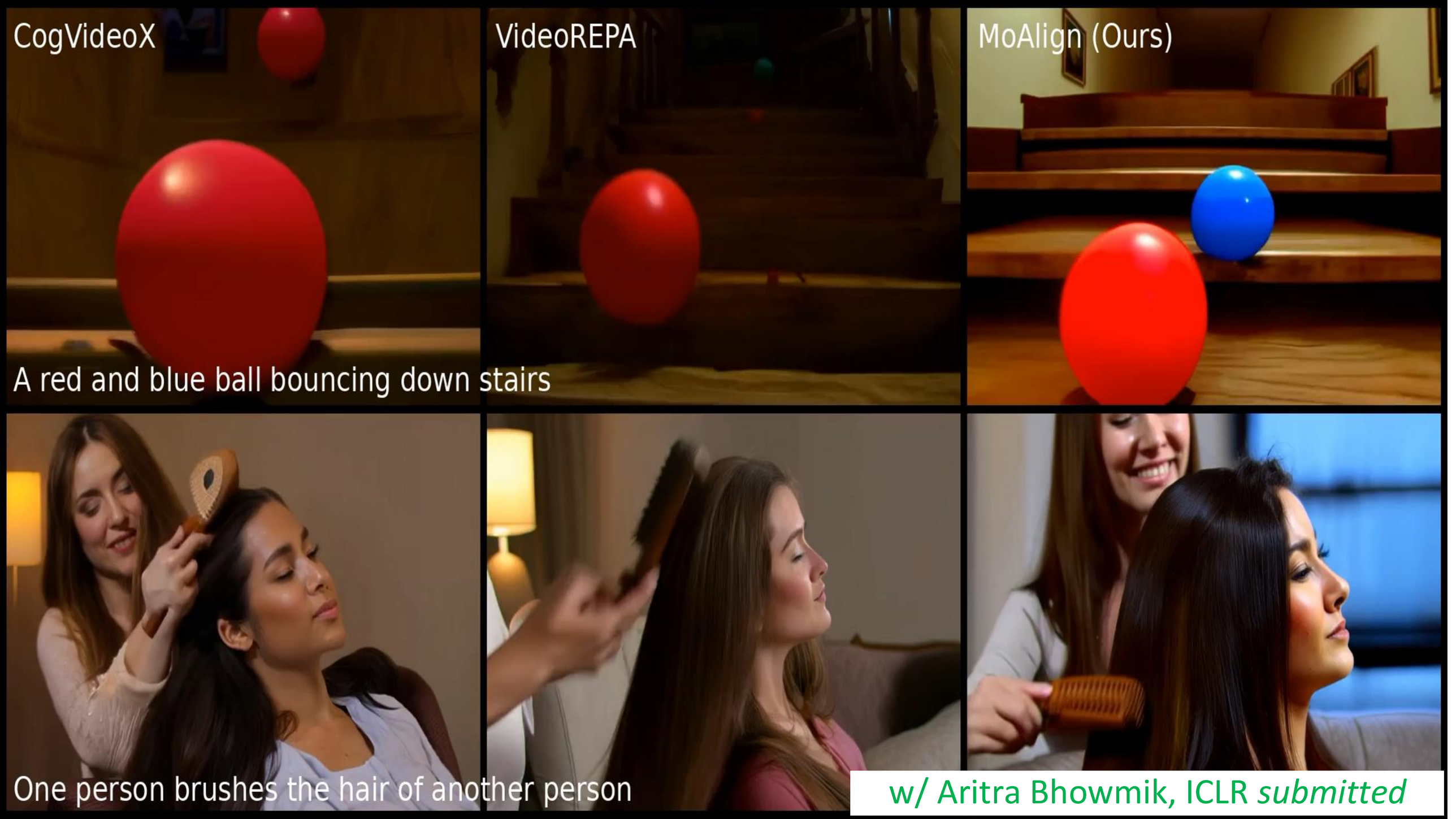
w/ Mohammadreza Salehi, Efstratios Gavves & Yuki Asano, ICCV 2023

w/ Hazel Doughty, CVPR 2022

Action: **peel**

How is the action done?

evenly, backwards, carefully, quickly, properly


CogVideoX
VideoREPA
MoAlign (Ours)
A red and blue ball bouncing down stairs
One person brushes the hair of another person
w/ Aritra Bhowmik, ICLR *submitted*

Mis-use concerns & responsibility


Heineken
PREMIUM QUALITY
CAFE MOOY
KOLKSTEEG
(CENTRUM)
Heineken Bieren Café
Café Mooy
Poffertjes
Soep p/d Dag
Tosti's
Div/ belegde
Pistolets
Hot Dog
Pizza
Ham/Kaas
Burger
Bal gehakt
Koffie/Thee
Irish Coffee
Warme Choco
met Slagroom
POOLTABLE
KARSTENGEL
LEVERWORST
KAAS


世界人民大团结万岁

Powerful yet irresponsible

- Mis-alignment with human values
- Hallucination
- Lacking adaptability to social dynamics and cultural context
- Limited transparency and explainability
- Non-inclusive and often closed access
- Unsustainable energy footprint
- Lacking robustness

University of Amsterdam
Data Science Centre

HAVA-Lab

What defines **human-aligned video-AI**, how can it be made computable, and what determines its societal acceptance?

How can we **embed laws, societal values, and ethics** in video AI's algorithm lifecycle?

Is there one solution for all, or do we need specialized **algorithms for each domain?**


Cees Snoek


Pascal Mettes


Iris Groen


Heleen Janssen


Tobias Blanke


Paula Helm


Marie Lindegaard


Erwin Berkhout


Stevan Rudinac

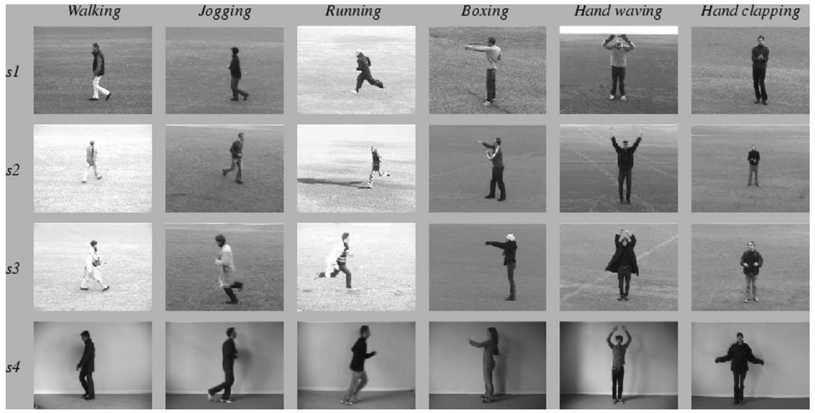

Marlies Schijven

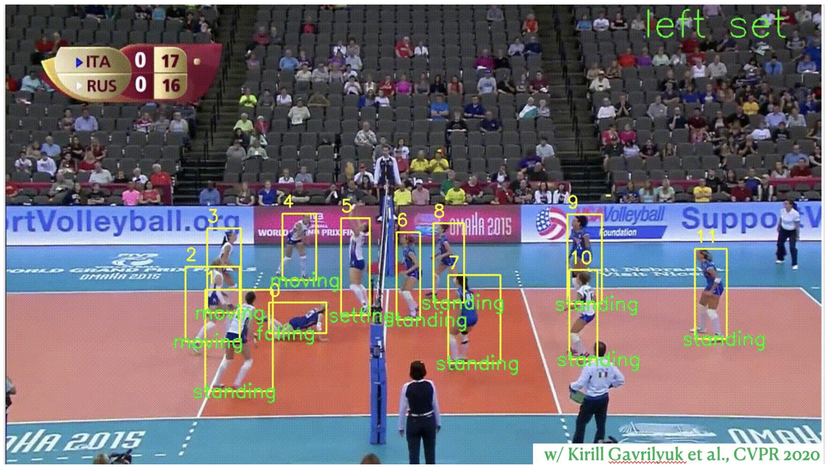

which indicate the reliability of different modalities, to learn our dual branch prediction.
Feature Projection. We use pre-trained unimodal encoders $f_m(\cdot)$ to extract the features of each modality. These features are then flattened to tokens $\mathbf{F}_m = [\mathbf{f}_1, \ldots, \mathbf{f}_{l_m}] \, \mathbf{F}_m \in \mathbb{R}^{l_m \times d_m}$ with l_m tokens and d_m feature dimension for modality m. The feature tokens of different modalities are unaligned and in different feature spaces with different lengths l_m and dimensions d_m. For example, audio tokens correspond to a specific range of time and frequency, while video tokens represent a specific spatial-temporal video segment. Existing works with modality-complete training data concatenate features of different modalities as input to a multimodal transformer. However, without modality-complete data in training, a multimodal transformer cannot learn to find cross-modal correspondences. Thus, we need a model that
sets of modality-incomplete data $X_{\{a\}}$, and $X_{\{b\}}$, and make predictions
-complete data $X_{\{a,b\}}$. Instead of concatenation we use the summation operation, which
to accumulate information from any number of available modalities, as long
are in a common feature space. However, projecting modality-
the common space as in previous works [24, 27] loses a lot of

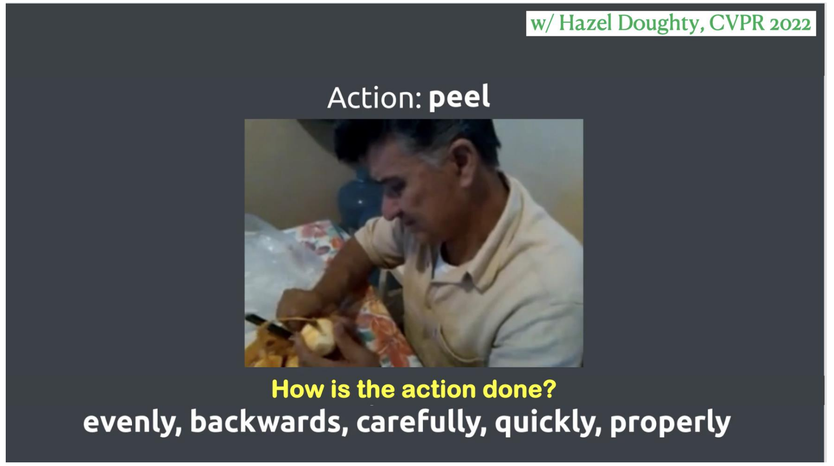
Powerful yet irresponsible

- Mis-alignment with human values
- Hallucination
- Lacking adaptability to social dynamics and cultural context
- Limited transparency and explainability
- Non-inclusive and often closed access
- Unsustainable energy footprint
- Lacking robustness

What assumption do all these works have in common at training time?







Empirical risk minimization and the i.i.d. assumption

Empirical risk minimization

Definition. *Given a set of labeled data points* $S = ((x_1, y_1), ..., (x_n, y_n))$*, the* empirical risk *of a predictor* $f : \mathcal{X} \to \mathcal{Y}$ *with respect to the sample* S *is defined as*

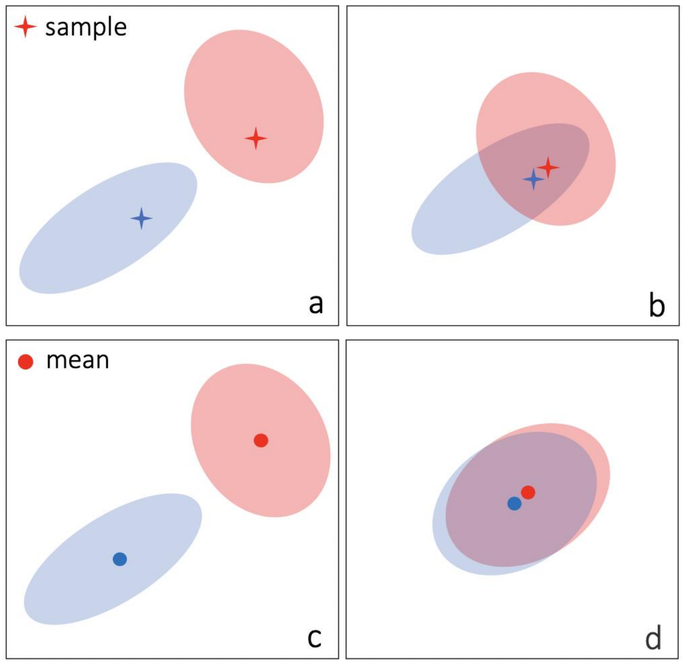
$$R_S[f] = \frac{1}{n} \sum_{i=1}^{n} loss(f(x_i), y_i) \, .$$

i.i.d. assumption

It is typically assumed that training, validation and test set are independent and identically distributed.

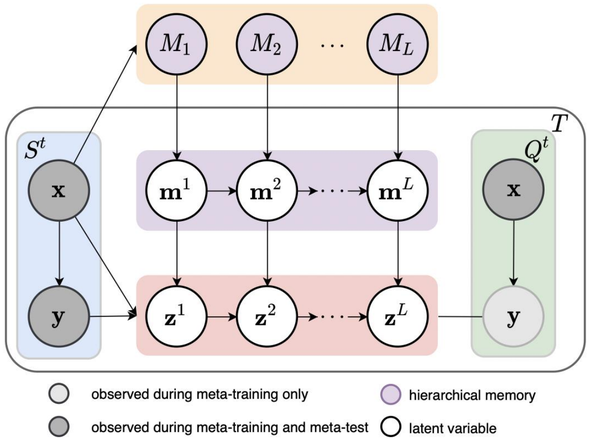
Machine learning inspiration

Domain-invariant learning



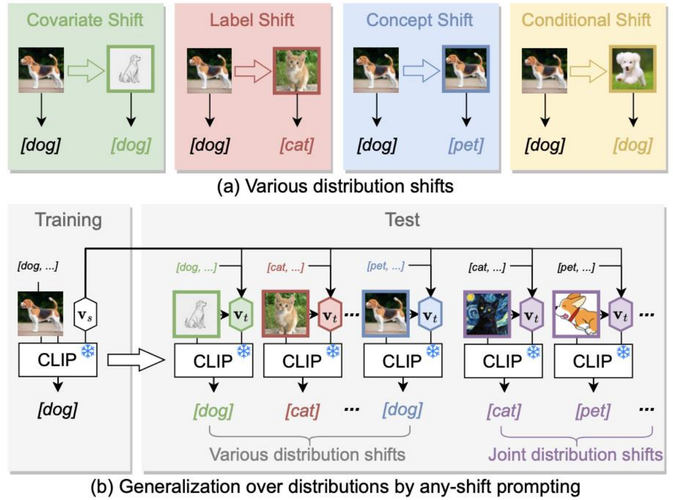
w/ Zehao Xiao et al., ICML 2021

Meta-learning



w/ Yingjun Du et al., ICLR 2022

Prompt learning



w/ Zehao Xiao et al., CVPR 2024

More is different

4 August 1972, Volume 177, Number 4047 SCIENCE

Philip Anderson crystallized the idea of emergence, arguing that "at each level of complexity entirely new properties appear" — that is, although, for example, chemistry is subject to the laws of physics, we cannot infer the field of chemistry from our knowledge of physics.

The reductionist hypothesis may still be a topic for controversy among philosophers, but among the great majority of active scientists I think it is accepted without question. The workings of our minds and bodies, and of all the animate or inanimate matter of which we have any detailed knowledge, are assumed to be controlled by the same set

planation of phenomena in terms of known fundamental laws. As always, distinctions of this kind are not unambiguous, but they are clear in most cases. Solid state physics, plasma physics, and perhaps also biology are extensive. High energy physics and a good part of nuclear physics are intensive. There is always much less intensive research going on than extensive. Once new fundamental laws are discov-

search which I think is as fundamental in its nature as any other. That is, it seems to me that one may array the sciences roughly linearly in a hierarchy, according to the idea: The elementary entities of science X obey the laws of science Y.

X	Y
solid state or many-body physics	elementary particle physics
chemistry	many-body physics
molecular biology	chemistry

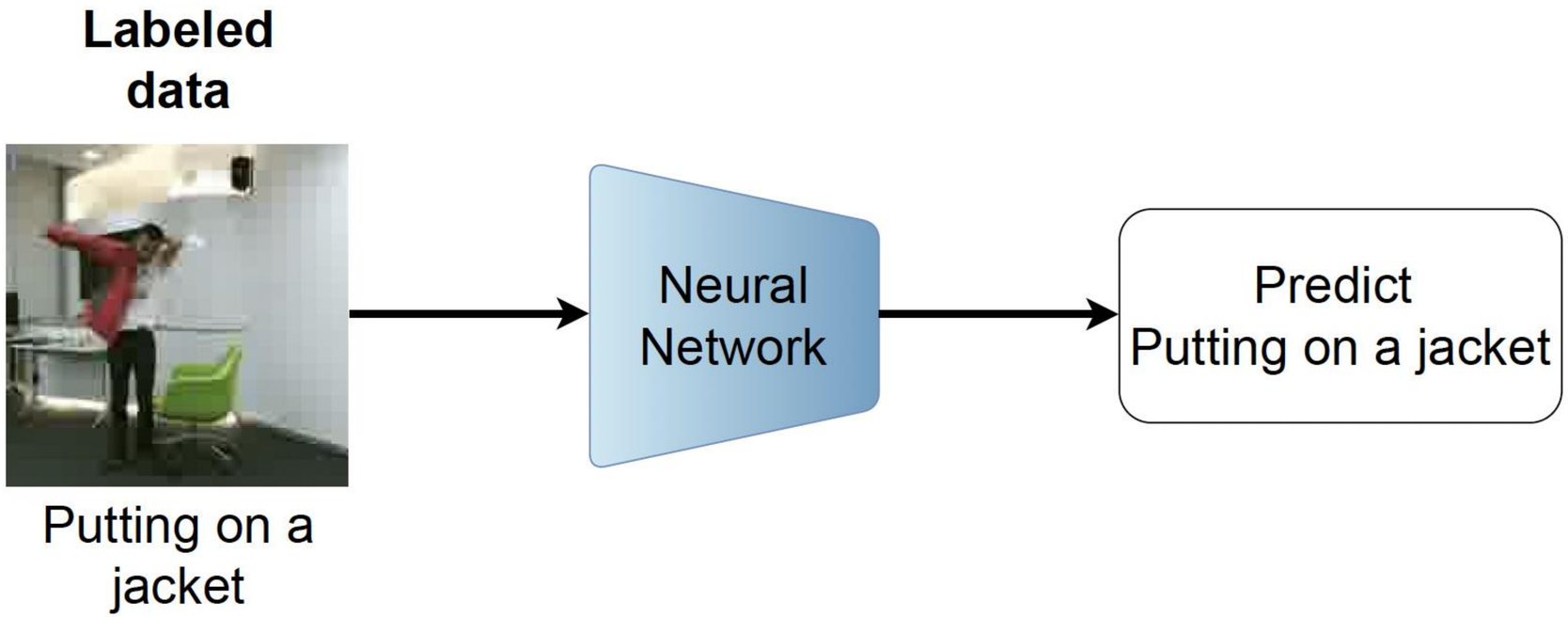
This lecture

Looks into the generalization abilities of modern video AI

1. Role of pre-training
2. Role of time
3. Role of multimodality

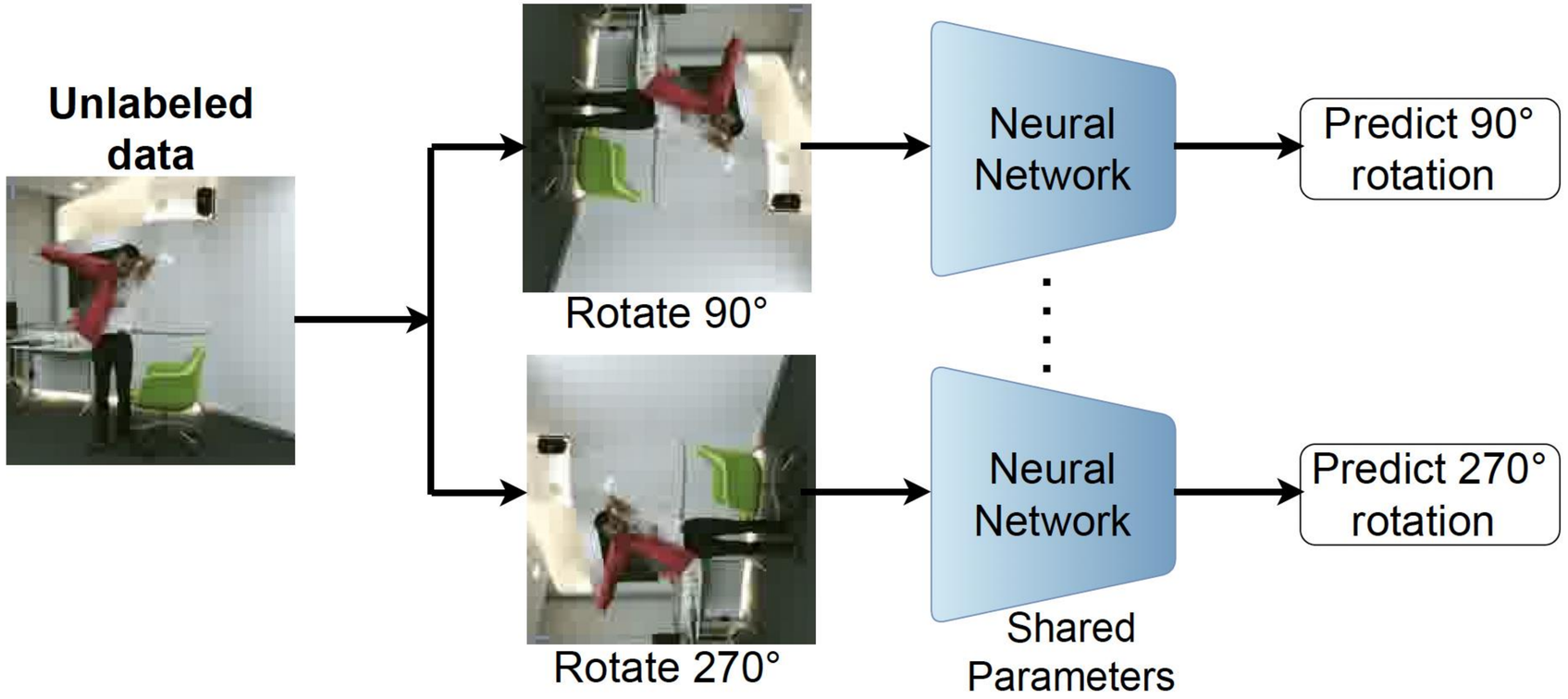
1. Role of pre-training

Supervised learning



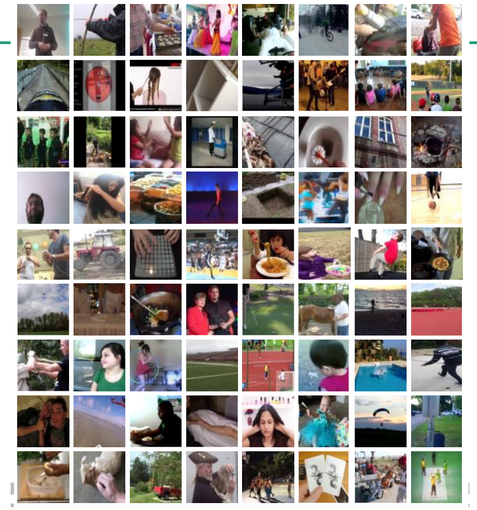
Depends on a manual labeling effort, which is costly, errorprone, and biased

Self-supervised learning using a proxy task



Self-supervised learning exploits (imposed) regularities in the data to learn from.

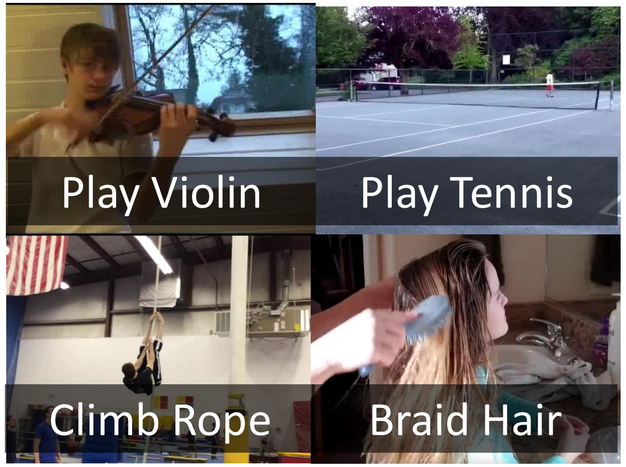
Self-Supervision



Model

Representation

Pretext Task



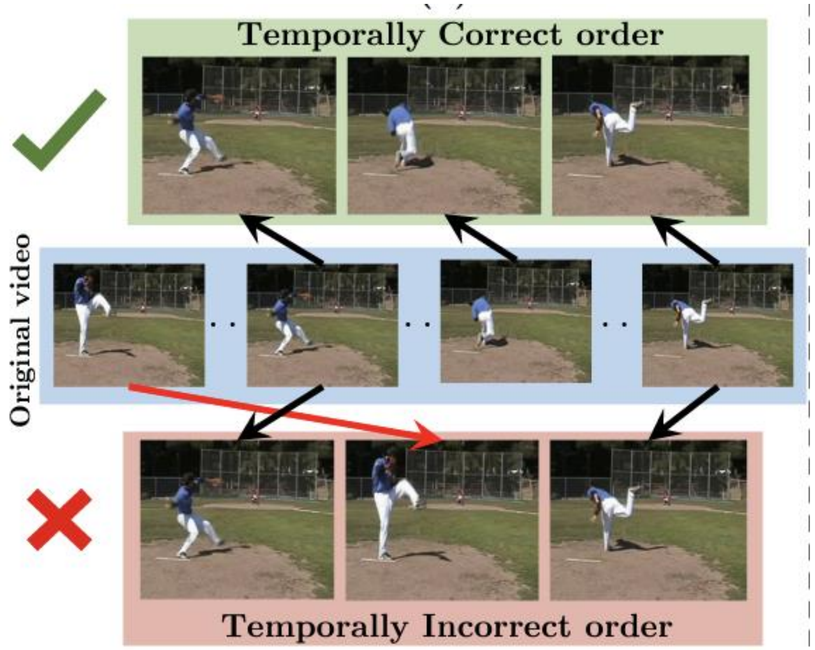
Model

Representation

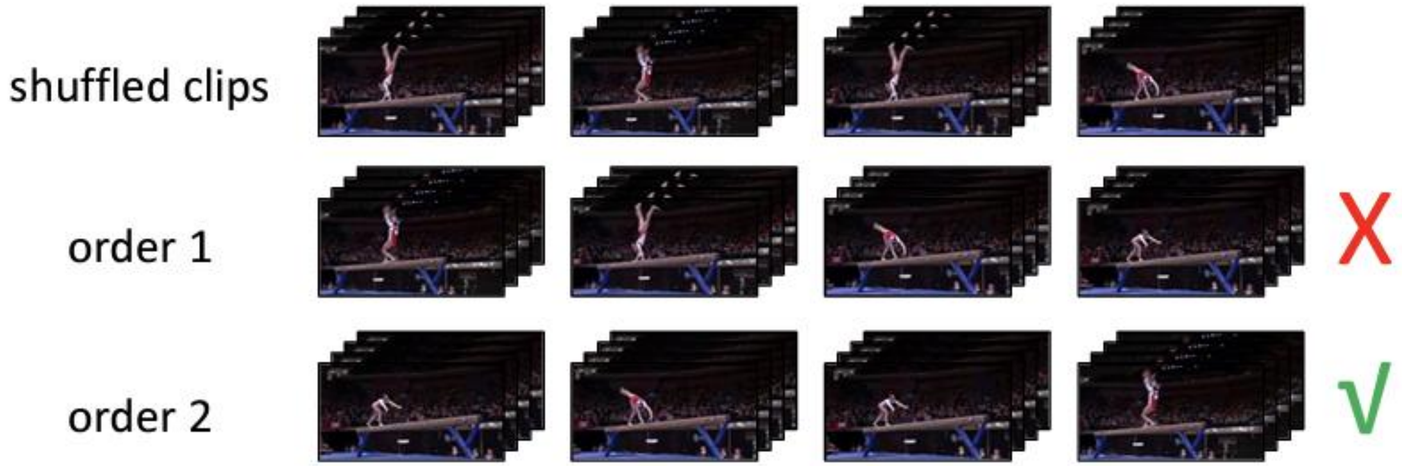
Classifier

Labels

Example proxy tasks



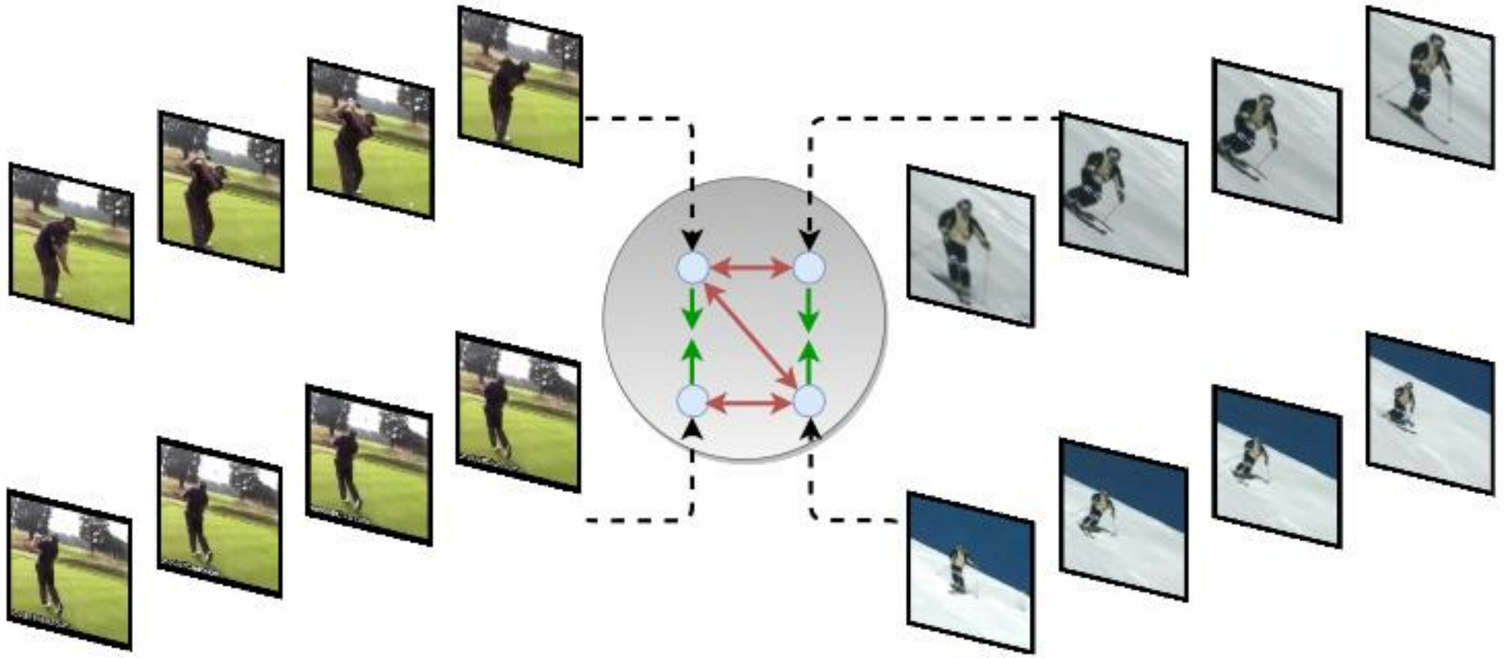
Shuffle and Learn, Mishra et. al., ECCV 2016



Video Clip Order Prediction, Xu et al., CVPR 2019

A more advanced proxy task: contrastive learning

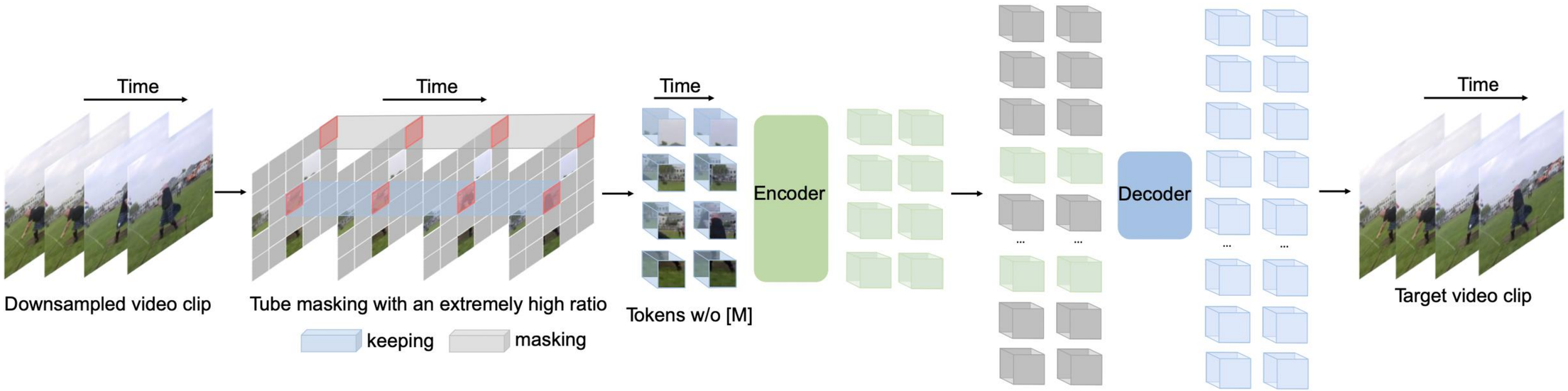
Uses Instance discrimination and enforces augmentation invariance.



Adaptation of image-based methods like MoCo, SimCLR, to video domain.

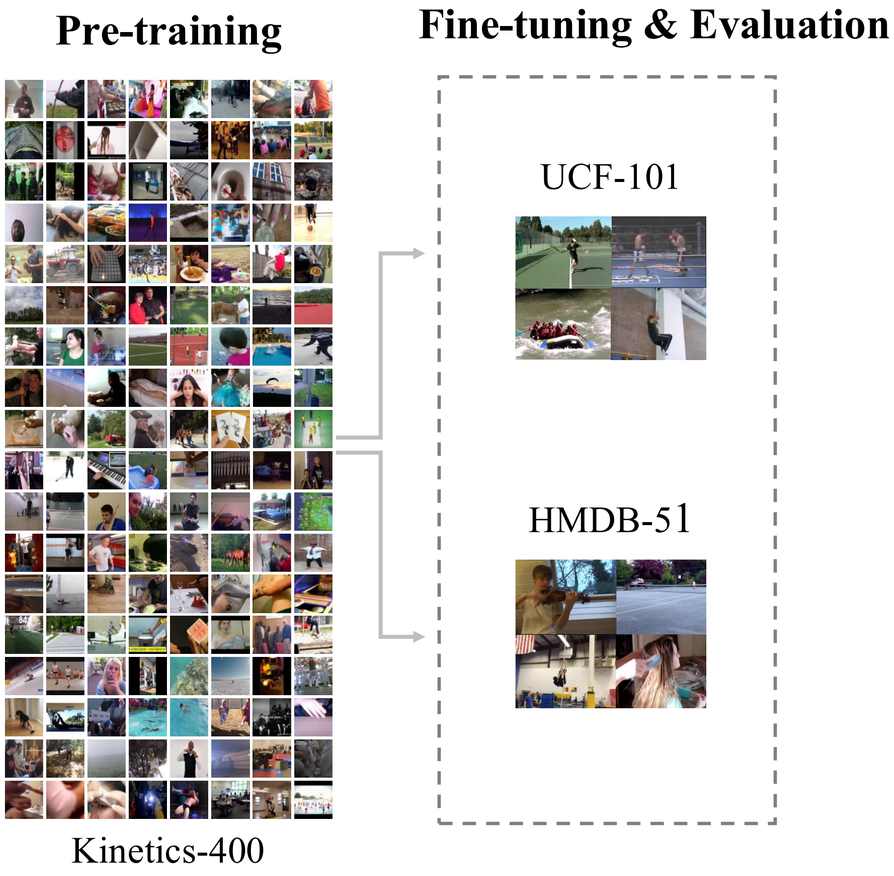
Masked auto encoding transformers

VideoMAE masks random cuboids and reconstructs the missing one

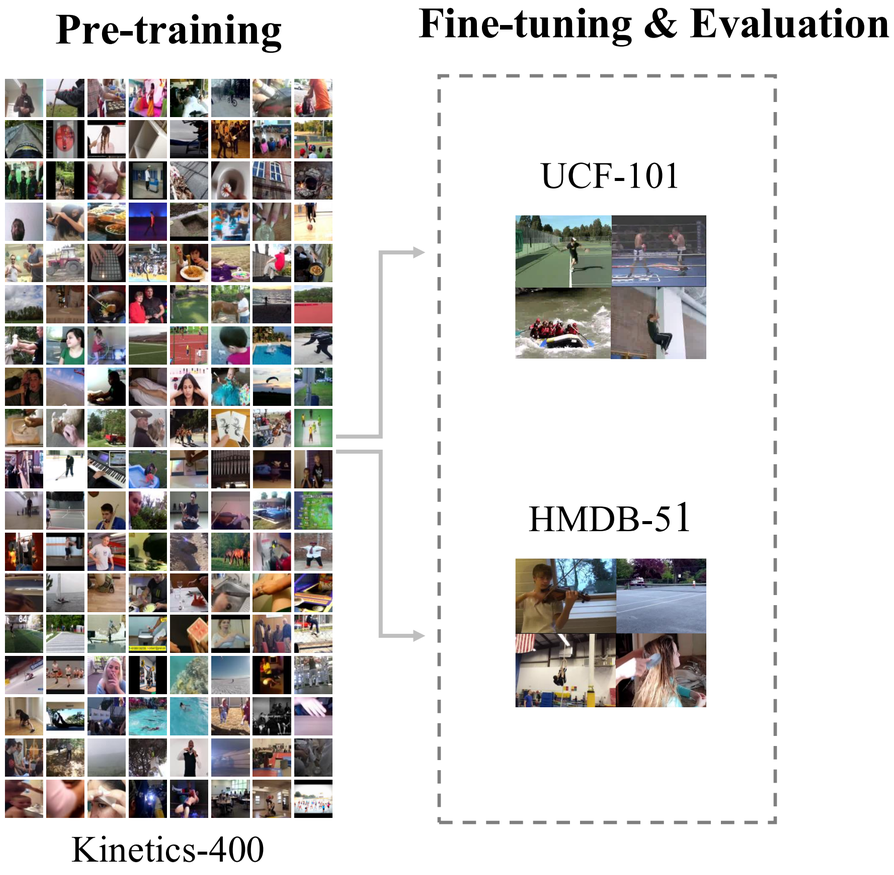


Zhan Tong, Yibing Song, Jue Wang, Limin Wang. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. In NeurIPS, 2022.

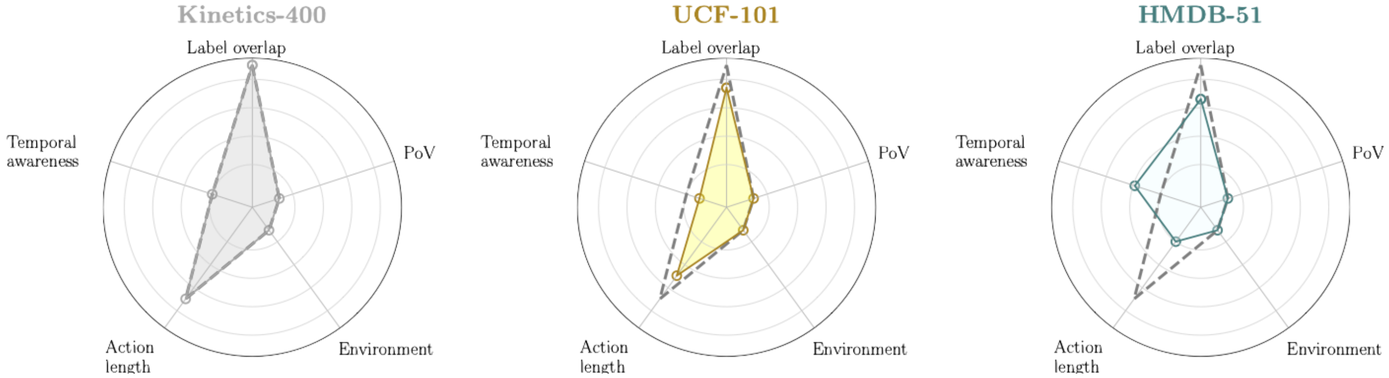
Problem: Video self-supervised learning evaluation



Problem: Video self-supervised learning evaluation



Pre-training and evaluation video too similar?



What if downstream video task is different?

Airport, shopping mall, hospital, *etc.*

1.a The problem of video evaluation

Fida Mohammad Thoker
University of Amsterdam

Hazel Doughty
University of Amsterdam

Piyush Bagad
University of Amsterdam

Cees Snoek
University of Amsterdam

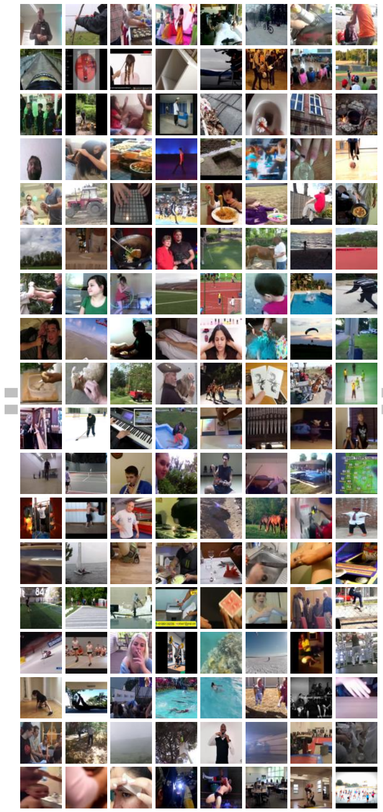
How Severe is Benchmark-Sensitivity in Video Self-Supervised Learning? In *ECCV* 2022.

Extended version in submission to the International Journal of Computer Vision, 2026.

Project Website

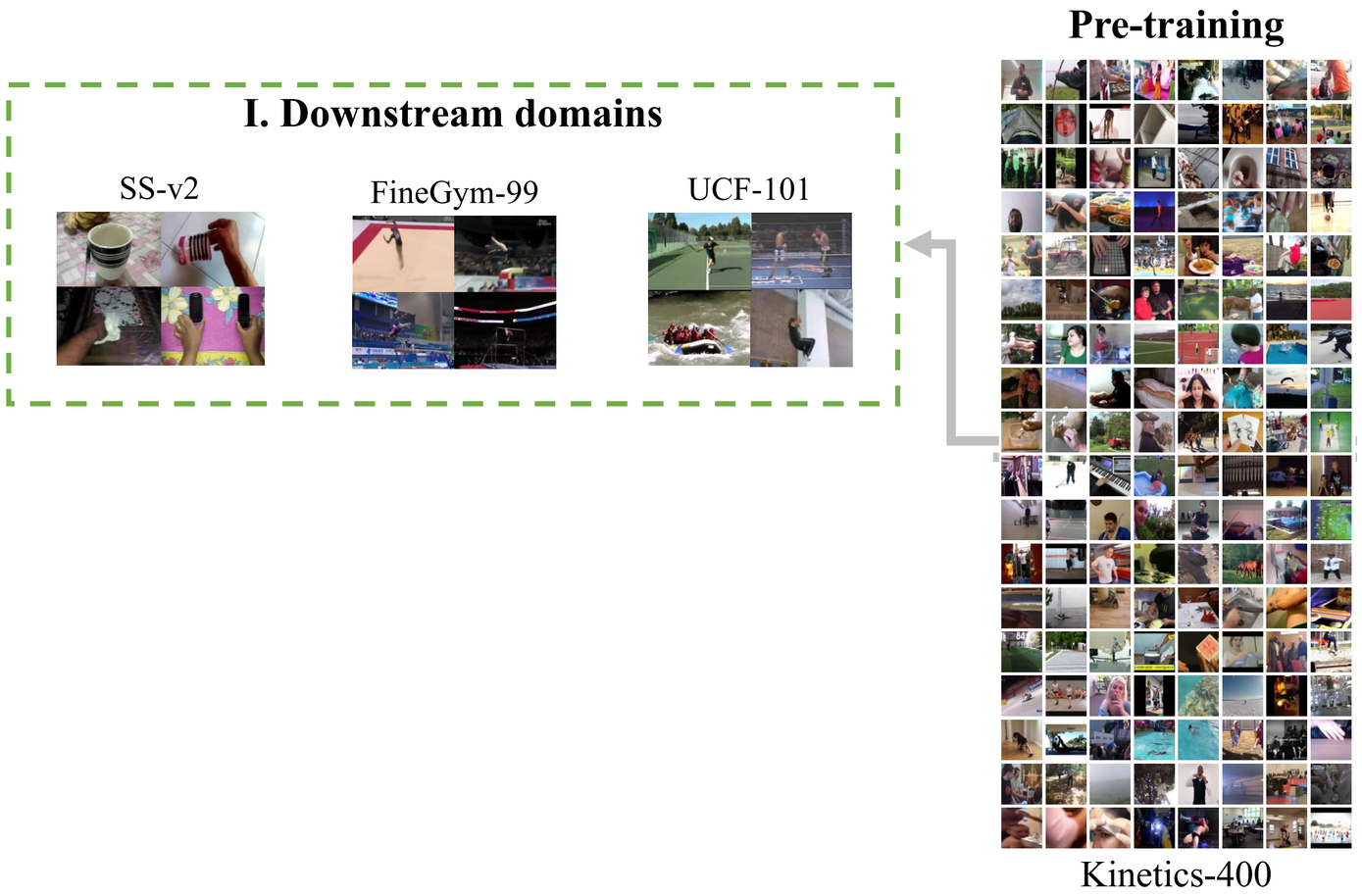
Proposed evaluation: four factors of sensitivity

Pre-training

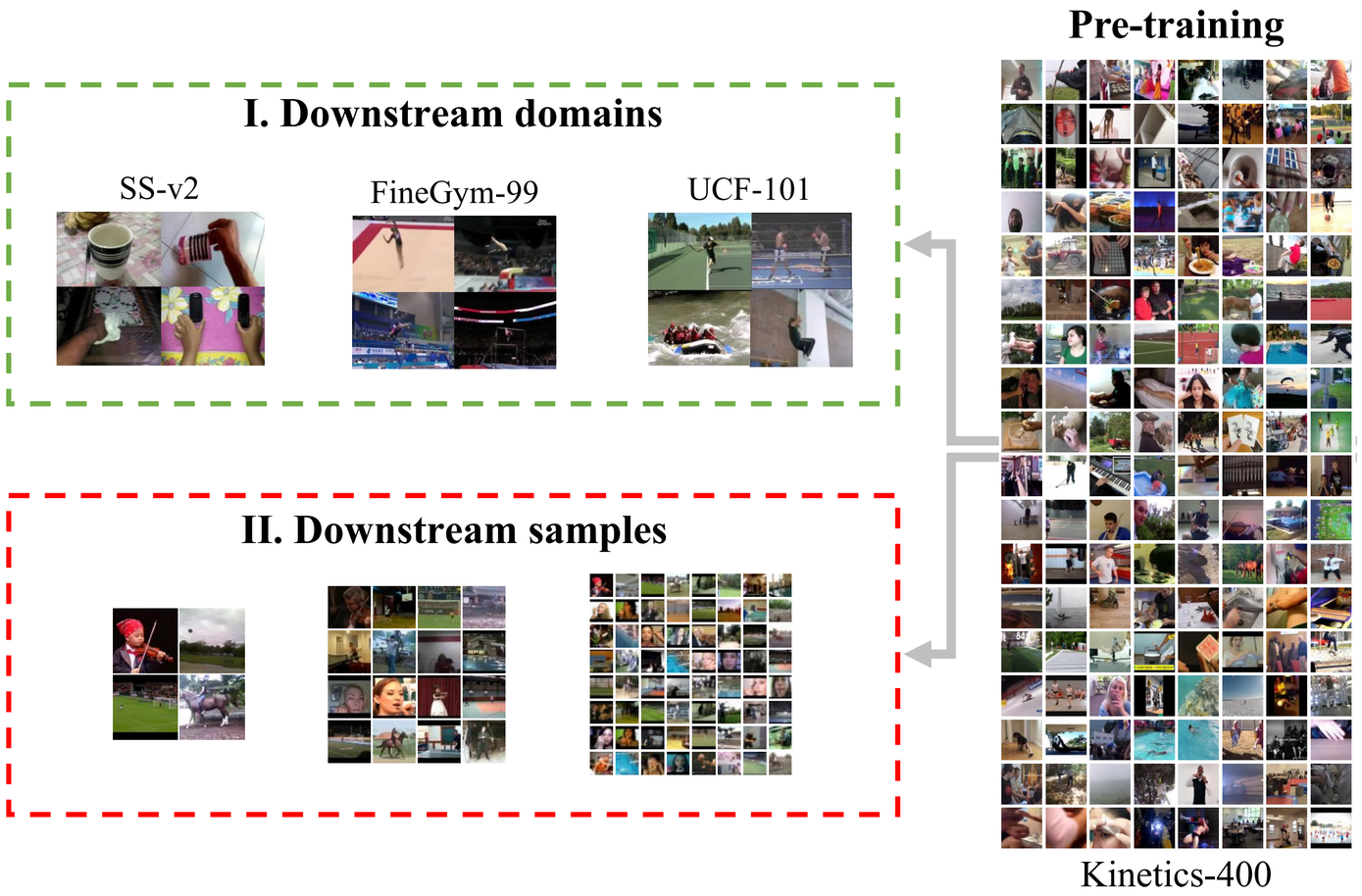


Kinetics-400

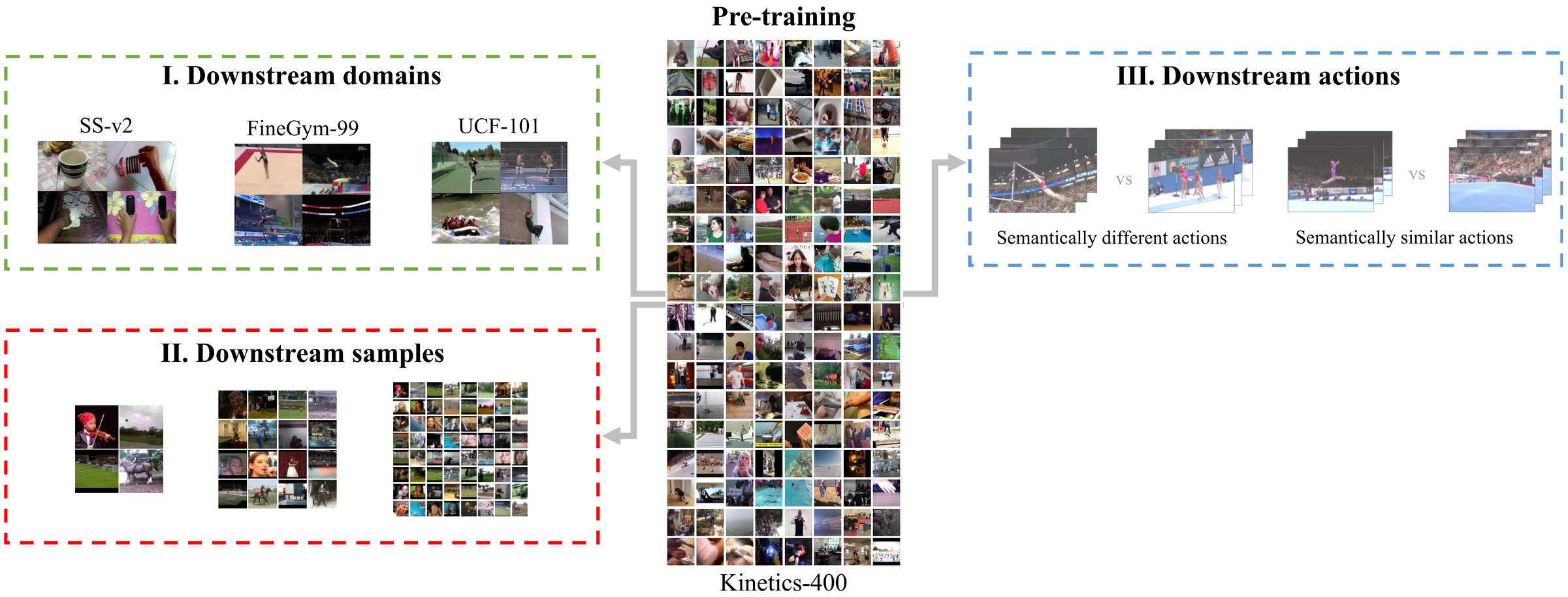
Proposed evaluation: four factors of sensitivity



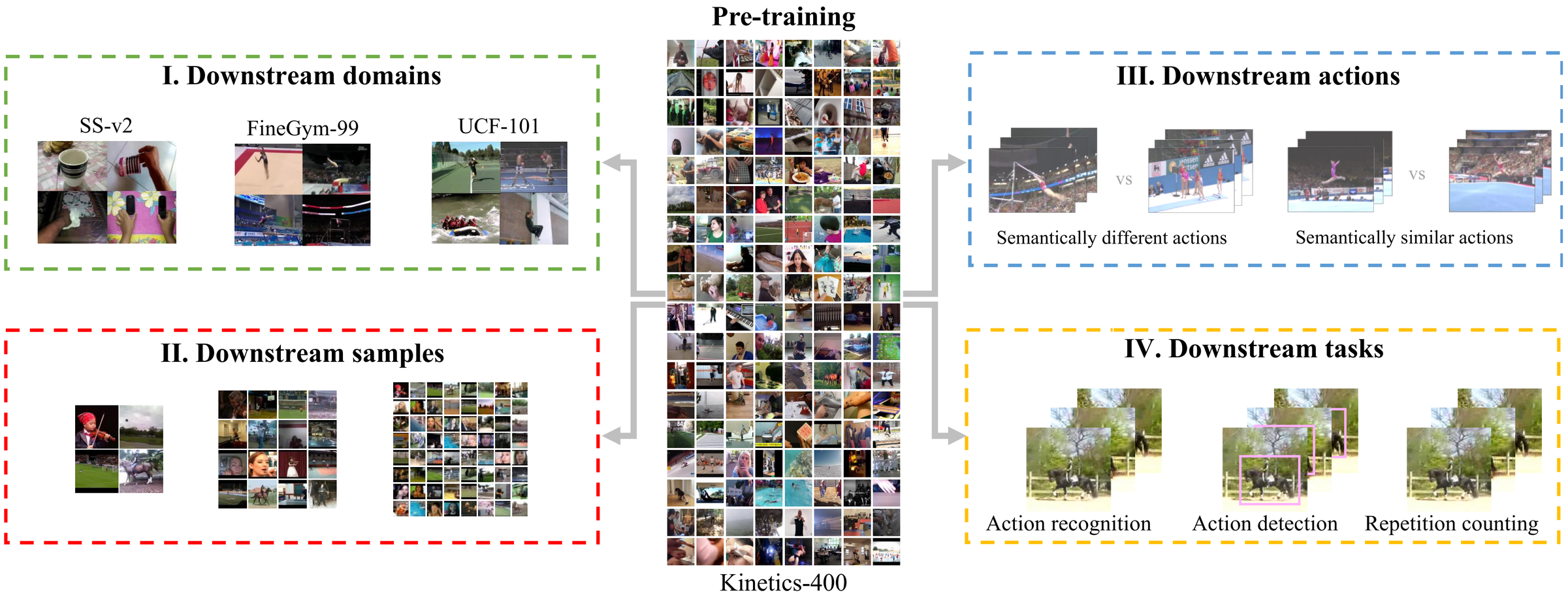
Proposed evaluation: four factors of sensitivity



Proposed evaluation: four factors of sensitivity

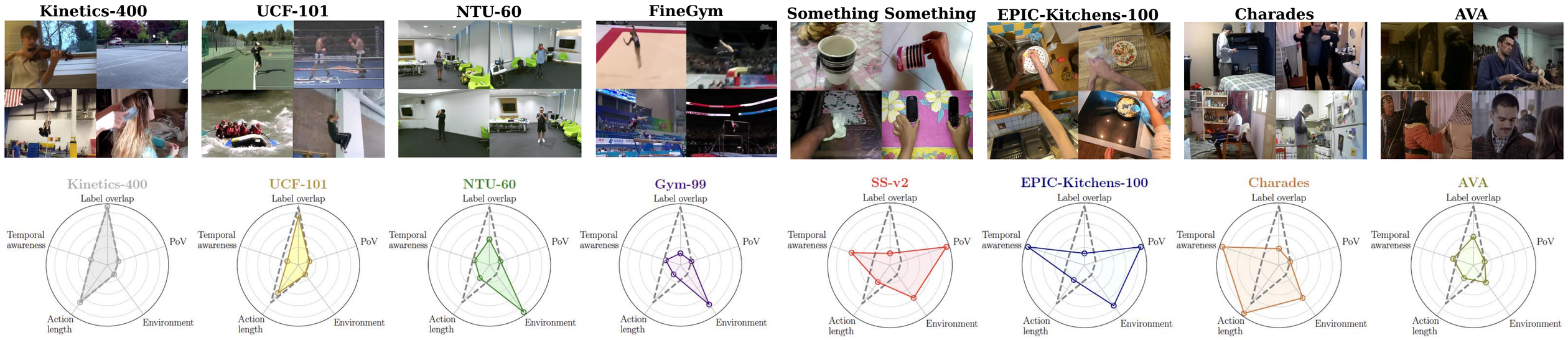


Proposed evaluation: four factors of sensitivity



7 datasets / 6 tasks / 500 experiments

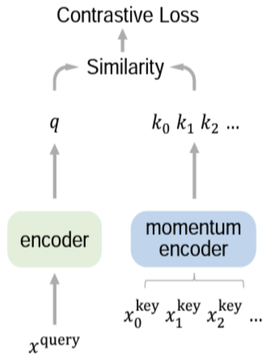
Considerable variety in video domain, the actions and tasks



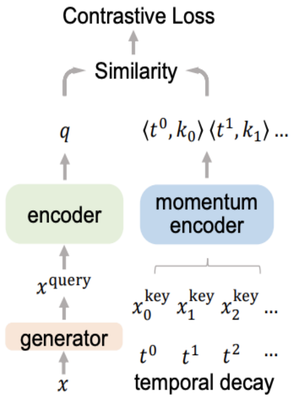
Tasks: *Action classification, Action detection, Repetition counting, Arrow of time prediction, Spatio-temporal detection, Multi-label classification*

9 video self-supervised learners

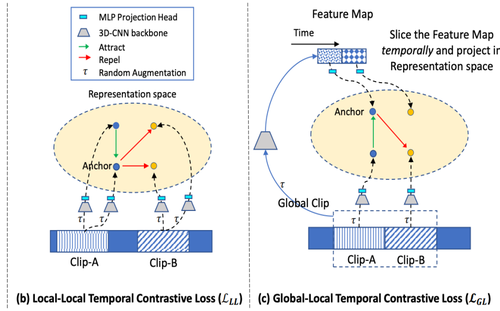
MoCo



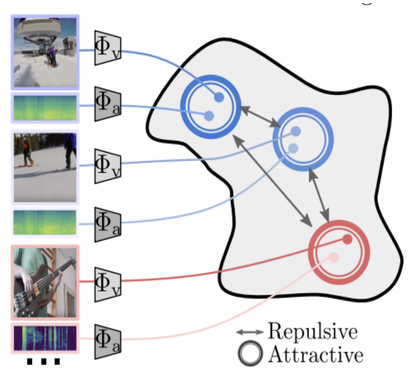
Video MoCo



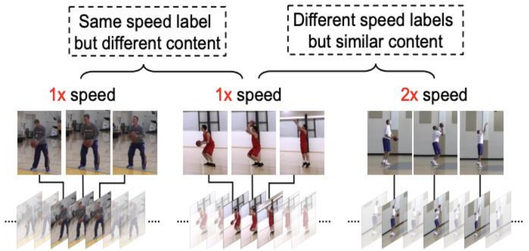
TCLR



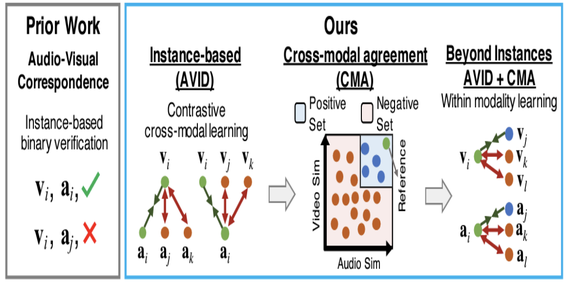
GDT



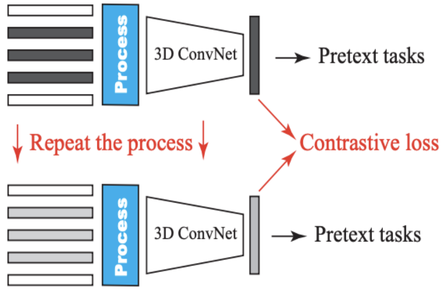
RSPnet



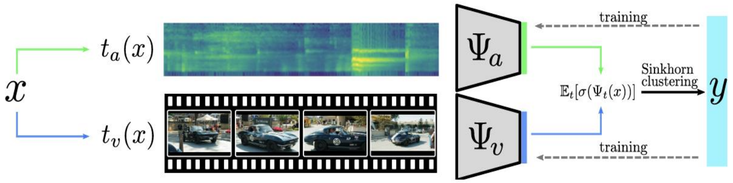
AVID-CMA



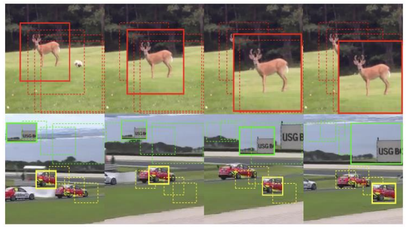
Pretext Contrast



SeLaVi

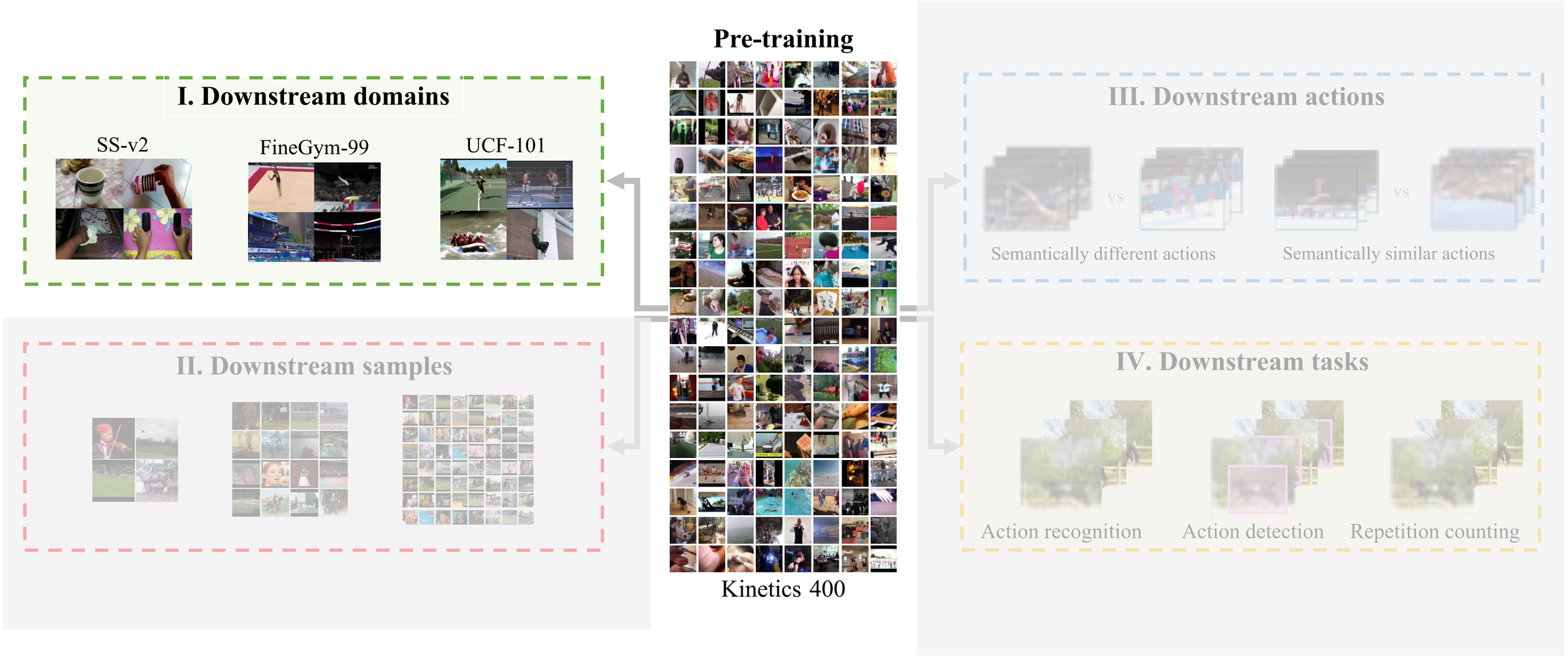


CtP



All methods come with weights for a R(2+1)D-18 network pre-trained on Kinetics-400

Sensitivity factor I: Downstream domain



Sensitivity factor I: Downstream domain

Downstream Domains

Pre-training	Finetuning				
	UCF101	NTU60	Gym99	SSv2	EK 100
None	75.4	92.9	89.4	56.8	25.7
MoCo	83.5	93.4	90.6	57.0	26.4
SeLaVi	84.9	92.8	88.9	56.4	33.8
VideoMoCo	85.8	94.1	90.5	58.8	43.6
Pretext-Contrast	86.6	93.9	90.3	57.0	34.3
RSPNet	88.5	93.9	91.3	59.4	42.7
AVID-CMA	89.3	94.0	90.6	53.8	29.9
CtP	89.8	94.3	92.2	60.2	42.8
TCLR	90.8	94.1	91.5	60.0	36.2
GDT	91.1	93.9	90.4	57.8	37.3
Supervised	94.1	93.9	91.8	61.0	47.7

Increasing domain shift →

Sensitivity factor I: Downstream domain

Downstream Domains

Pre-training	Finetuning				
	UCF101	NTU60	Gym99	SSv2	EK 100
None	75.4	92.9	89.4	56.8	25.7
MoCo	83.5	93.4	90.6	57.0	26.4
SeLaVi	84.9	92.8	88.9	56.4	33.8
VideoMoCo	85.8	94.1	90.5	58.8	43.6
Pretext-Contrast	86.6	93.9	90.3	57.0	34.3
RSPNet	88.5	93.9	91.3	59.4	42.7
AVID-CMA	89.3	94.0	90.6	53.8	29.9
CtP	89.8	94.3	92.2	60.2	42.8
TCLR	90.8	94.1	91.5	60.0	36.2
GDT	91.1	93.9	90.4	57.8	37.3
Supervised	94.1	93.9	91.8	61.0	47.7

Increasing domain shift →

Sensitivity factor I: Downstream domain

Downstream Domains

Pre-training	Finetuning				
	UCF101	NTU60	Gym99	SSv2	EK 100
None	75.4	92.9	89.4	56.8	25.7
MoCo	83.5	93.4	90.6	57.0	26.4
SeLaVi	84.9	92.8	88.9	56.4	33.8
VideoMoCo	85.8	94.1	90.5	58.8	43.6
Pretext-Contrast	86.6	93.9	90.3	57.0	34.3
RSPNet	88.5	93.9	91.3	59.4	42.7
AVID-CMA	89.3	94.0	90.6	53.8	29.9
CtP	89.8	94.3	92.2	60.2	42.8
TCLR	90.8	94.1	91.5	60.0	36.2
GDT	91.1	93.9	90.4	57.8	37.3
Supervised	94.1	93.9	91.8	61.0	47.7

Increasing domain shift →

Sensitivity factor I: Downstream domain

Downstream Domains

Pre-training	Finetuning				
	UCF101	NTU60	Gym99	SSv2	EK 100
None	75.4	92.9	89.4	56.8	25.7
MoCo	83.5	93.4	90.6	57.0	26.4
SeLaVi	84.9	92.8	88.9	56.4	33.8
VideoMoCo	85.8	94.1	90.5	58.8	43.6
Pretext-Contrast	86.6	93.9	90.3	57.0	34.3
RSPNet	88.5	93.9	91.3	59.4	42.7
AVID-CMA	89.3	94.0	90.6	53.8	29.9
CtP	89.8	94.3	92.2	60.2	42.8
TCLR	90.8	94.1	91.5	60.0	36.2
GDT	91.1	93.9	90.4	57.8	37.3
Supervised	94.1	93.9	91.8	61.0	47.7

Increasing domain shift →

Sensitivity factor I: Downstream domain

Downstream Domains

Pre-training	Finetuning				
	UCF101	NTU60	Gym99	SSv2	EK 100
None	75.4	92.9	89.4	56.8	25.7
MoCo	83.5	93.4	90.6	57.0	26.4
SeLaVi	84.9	92.8	88.9	56.4	33.8
VideoMoCo	85.8	94.1	90.5	58.8	43.6
Pretext-Contrast	86.6	93.9	90.3	57.0	34.3
RSPNet	88.5	93.9	91.3	59.4	42.7
AVID-CMA	89.3	94.0	90.6	53.8	29.9
CtP	89.8	94.3	92.2	60.2	42.8
TCLR	90.8	94.1	91.5	60.0	36.2
GDT	91.1	93.9	90.4	57.8	37.3
Supervised	94.1	93.9	91.8	61.0	47.7

Increasing domain shift →

Sensitivity factor I: Downstream domain

Downstream Domains

Pre-training	Finetuning				
	UCF101	NTU60	Gym99	SSv2	EK 100
None	75.4	92.9	89.4	56.8	25.7
MoCo	85.8	93.1	88.8	57.0	26.4
SeLaVi	84.9	92.8	88.9	56.4	33.8
VideoMoCo	85.8	94.1	90.5	58.8	43.6
PreCLR	86.5	93.5	90.3	57.6	34.8
RSPNet	88.5	93.9	91.3	59.4	42.7
AVID+CMA	89.5	94.0	90.6	55.8	29.9
CtP	89.8	94.3	92.2	60.2	42.8
TCLR	90.8	94.1	91.5	60.0	36.2
GDT	91.1	93.9	90.4	57.8	37.3
Supervised	94.1	93.9	91.8	61.0	47.7

Downstream Domains

UCF-101 finetuning performance **does not generalize** to other target domains.

Increasing domain shift

Sensitivity factor I: Downstream domain

Downstream Domains

Pre-training	Finetuning				
	UCF101	NTU60	Gym99	SSv2	EK 100
None	75.4	92.9	89.4	56.8	25.7
MoCo	83.5	93.4	90.6	57.0	26.4
SeLaVi	84.9	92.8	88.9	56.4	33.8
VideoMoCo	85.8	94.1	90.5	58.8	43.6
Pretext-Contrast	86.6	93.9	90.3	57.0	34.3
RSPNet	88.5	93.9	91.3	59.4	42.7
AVID-CMA	89.3	94.0	90.6	53.8	29.9
CtP	89.8	94.3	92.2	60.2	42.8
TCLR	90.8	94.1	91.5	60.0	36.2
GDT	91.1	93.9	90.4	57.8	37.3
Supervised	94.1	93.9	91.8	61.0	47.7

Increasing domain shift →

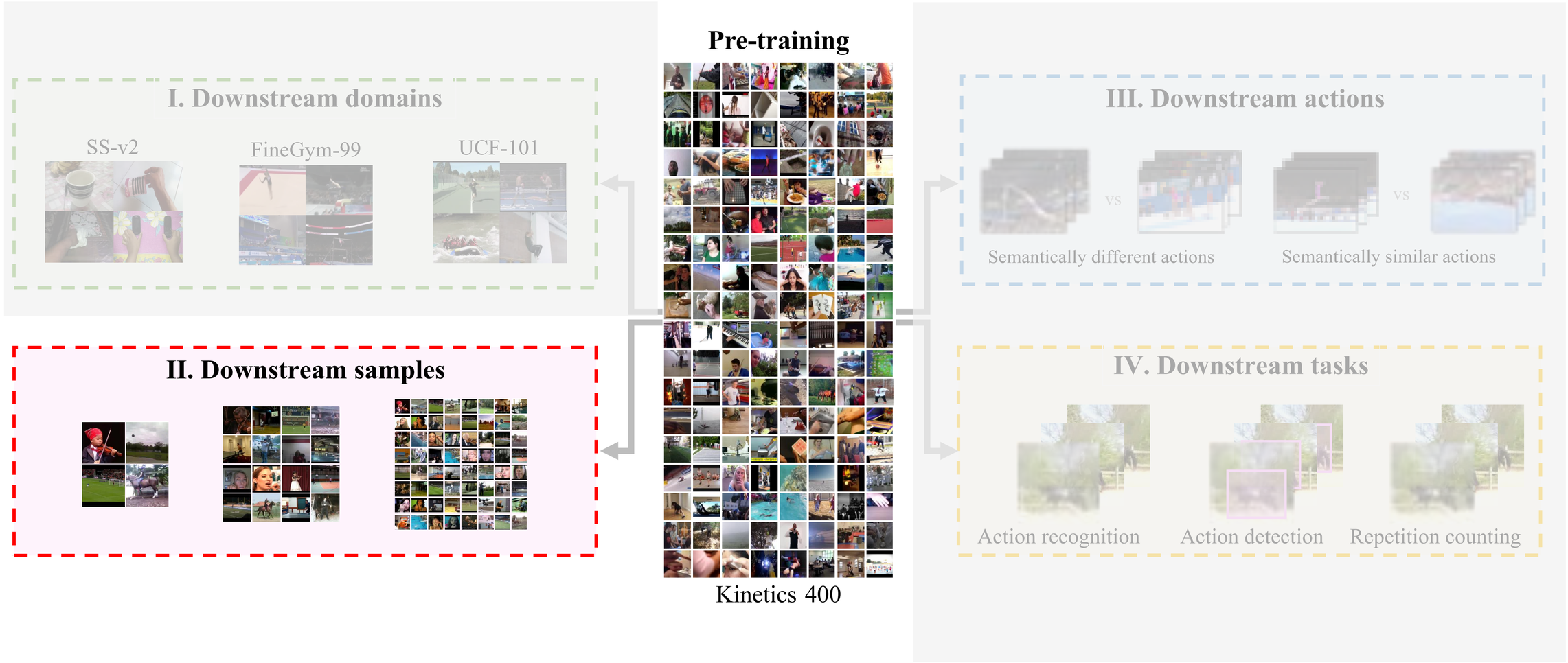
Sensitivity factor I: Downstream domain

Downstream Domains

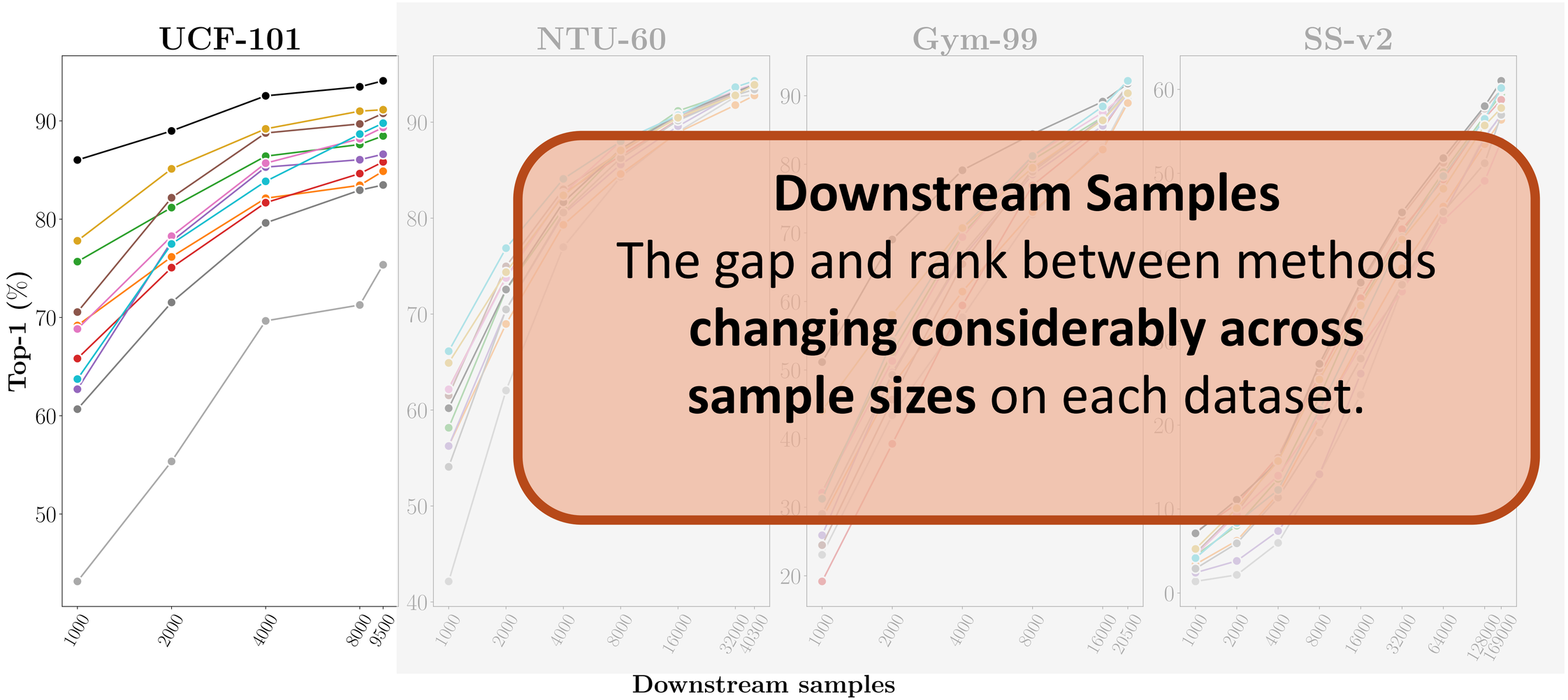
Pre-training	Finetuning				
	UCF101	NTU60	Gym99	SSv2	EK 100
None	75.4	92.9	89.4	56.8	25.7
MoCo	83.5	93.4	90.6	57.0	26.4
SeLaVi	84.9	92.8	88.9	56.4	33.8
VideoMoCo	85.8	94.1	90.5	58.8	43.6
Pretext-Contrast	86.6	93.9	90.3	57.0	34.3
RSPNet	88.5	93.9	91.3	59.4	42.7
AVID-CMA	89.3	94.0	90.6	53.8	29.9
CtP	89.8	94.3	92.2	60.2	42.8
TCLR	90.8	94.1	91.5	60.0	36.2
GDT	91.1	93.9	90.4	57.8	37.3
Supervised	94.1	93.9	91.8	61.0	47.7

Increasing domain shift →

Sensitivity factor II: Downstream samples



Sensitivity factor II: Downstream samples



Sensitivity factor III & IV: Downstream actions & tasks

Downstream Actions

Most self-supervised methods are **sensitive to action granularity** in downstream dataset.

Downstream Tasks

UCF-101 action classification performance is **mildly indicative** on other tasks.

Key takeaways

No clear winner, different methods standing out in different settings.

Supervised pre-training is dominant across all sensitivity factors.

Contrastive methods encouraging **temporal distinctiveness** transfer well.

We select a subset of experiments as the **‘SEVERE’ benchmark**

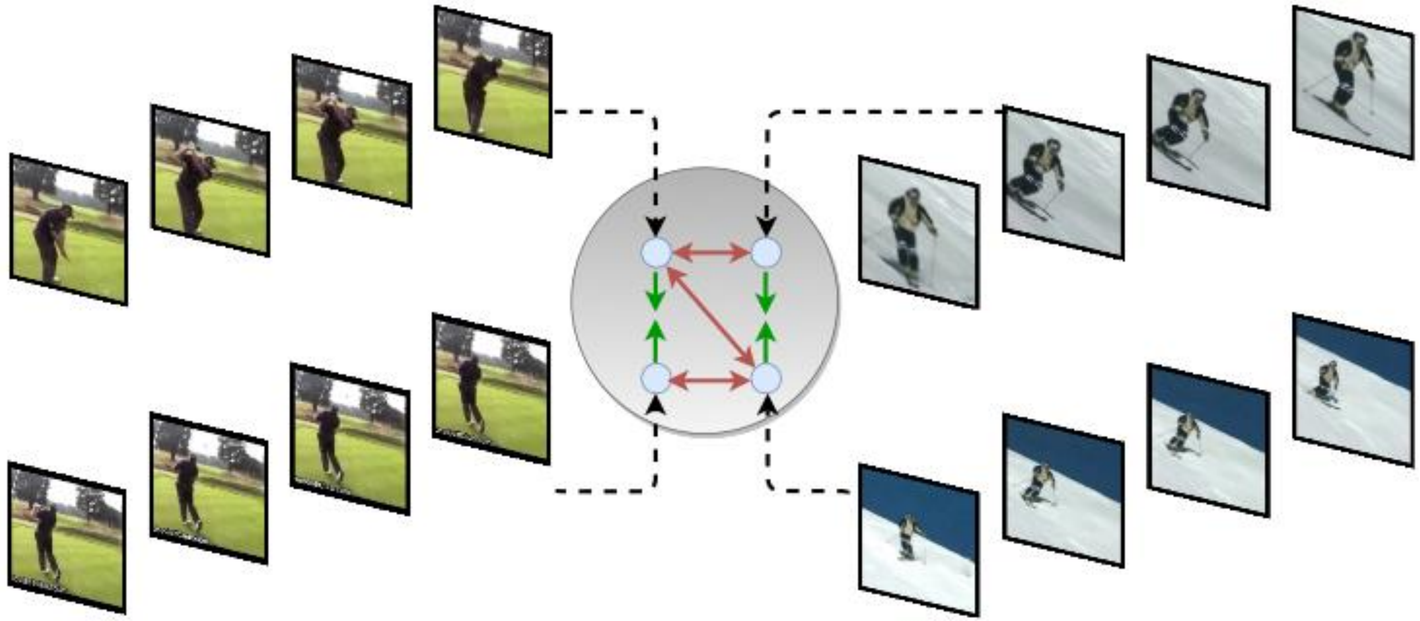
SEVERE benchmark: subset of our experiments

Pre-training	Existing	SEVERE-benchmark							
		Domains		Samples		Actions		Tasks	
	UCF101	SS-v2	Gym-99	UCF (10^3)	Gym-99 (10^3)	FX-S1	UB-S1	UCF-RC	Charades-MLC
None	75.4	56.8	89.4	43.1	23.1	45.0	84.0	0.232	7.9
MoCo	83.5	57.0	90.6	60.7	29.0	65.1	85.0	0.220	8.1
SeLaVi	84.9	56.4	88.9	69.2	28.3	50.2	81.5	0.171	8.2
VideoMoCo	85.8	58.8	90.5	65.8	19.2	60.4	82.1	0.171	10.5
Pretext-Contrast	86.6	57.0	90.3	62.7	25.9	65.8	86.2	0.168	8.9
RSPNet	88.5	59.4	91.3	75.7	32.2	63.5	85.1	0.151	9.1
AVID-CMA	89.3	53.8	90.6	68.8	32.1	67.2	88.4	0.162	8.4
CtP	89.8	60.2	92.2	63.7	31.2	79.7	88.4	0.178	9.6
TCLR	90.8	60.0	91.5	70.6	24.5	61.0	85.3	0.149	11.1
GDT	91.1	57.8	90.4	77.8	44.1	65.7	81.6	0.137	8.5
Supervised	94.1	61.0	91.8	86.0	51.2	81.0	86.9	0.137	23.6

Enables future video self-supervised methods to evaluate generalization along 4 factors.

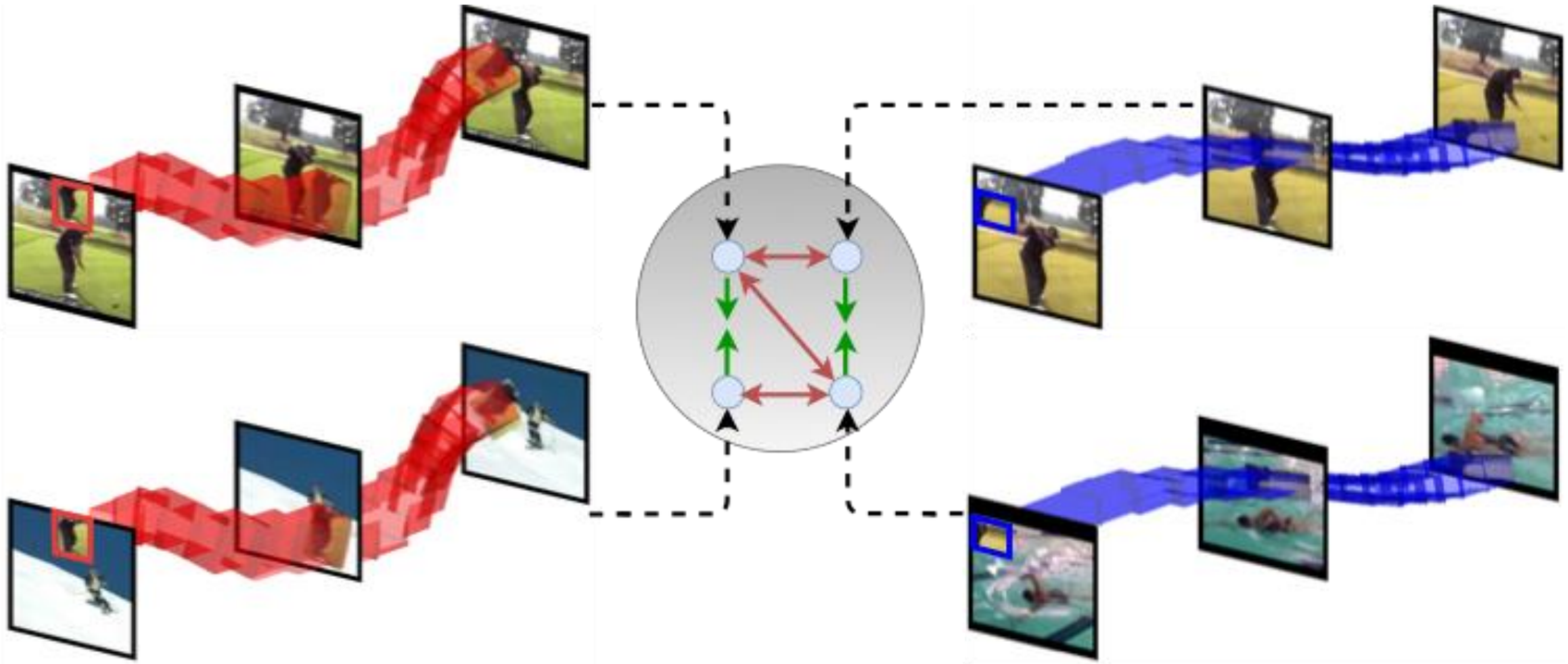
Problem of holistic contrastive learning

Uses instance discrimination and enforces augmentation invariance.



- 👎 Favours **coarse-grained** features
- 👎 Exploits background **shortcut**
- 👎 Limits **generalizability**
- 👎 Motion-variety constraints cause **data hunger**

Solution: add synthetic tubelets during pretraining



1.b The problem of video contrastive-learning

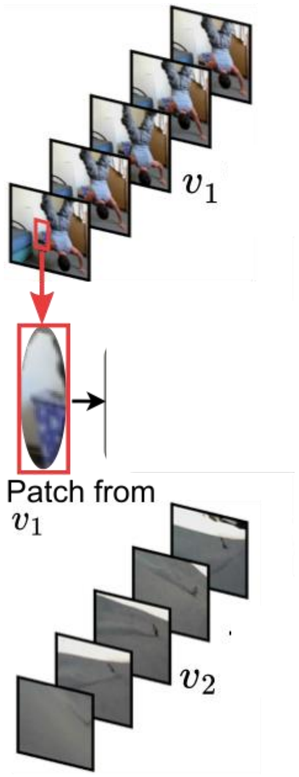
Fida Mohammad Thoker
University of Amsterdam

Hazel Doughty
University of Amsterdam

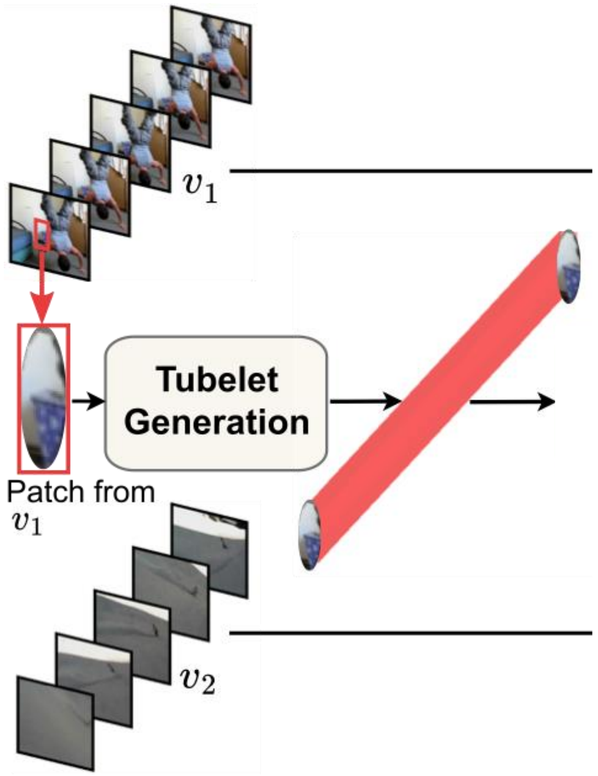
Cees Snoek
University of Amsterdam

Tubelet-Contrastive Self-Supervision for Video-Efficient Generalization. In *ICCV* 2023.

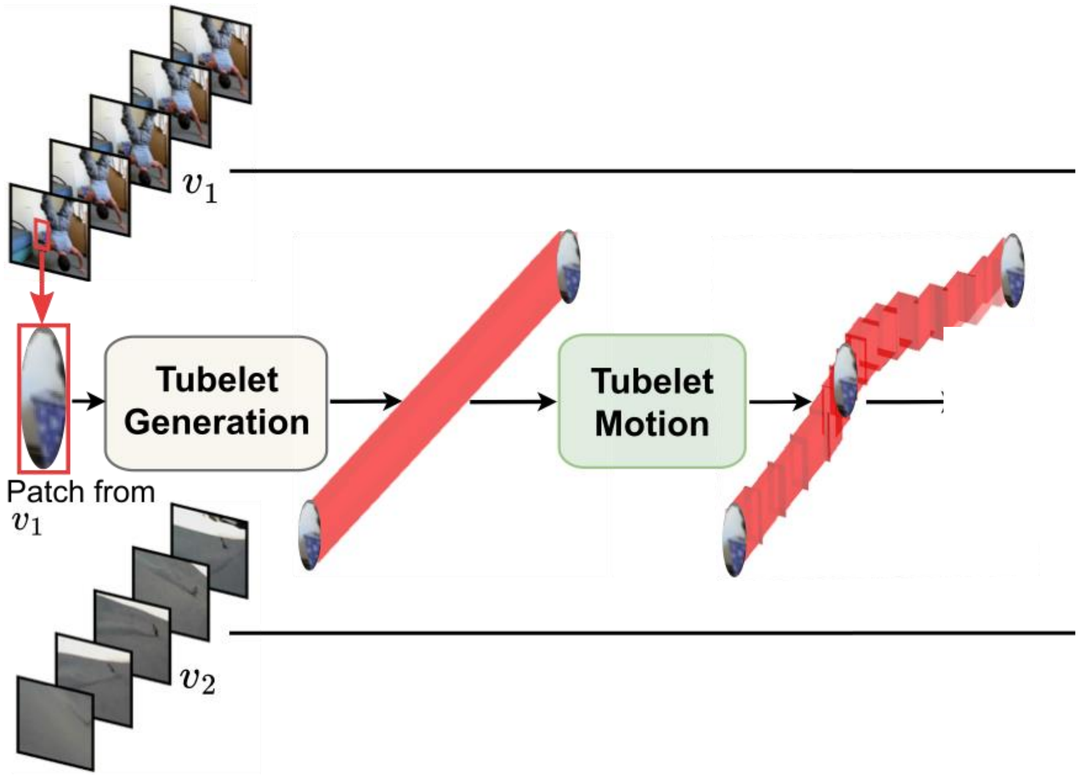
Step 0: Crop a random patch from one clip

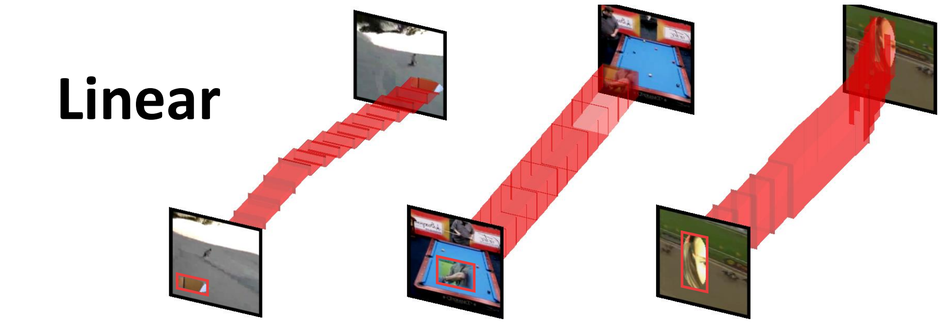


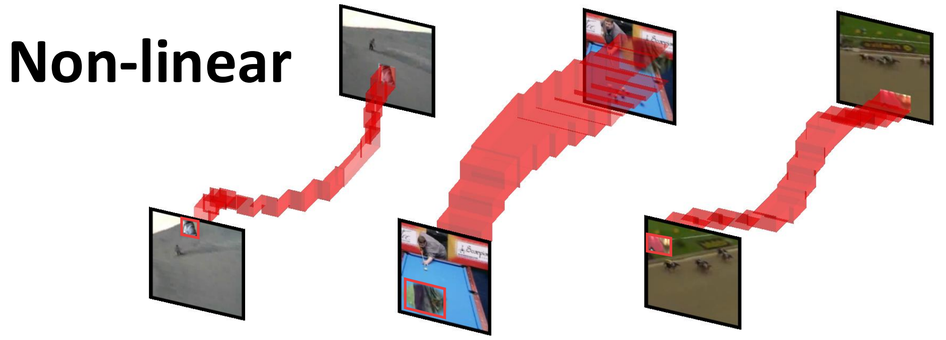
Step 1: Generate a tubelet



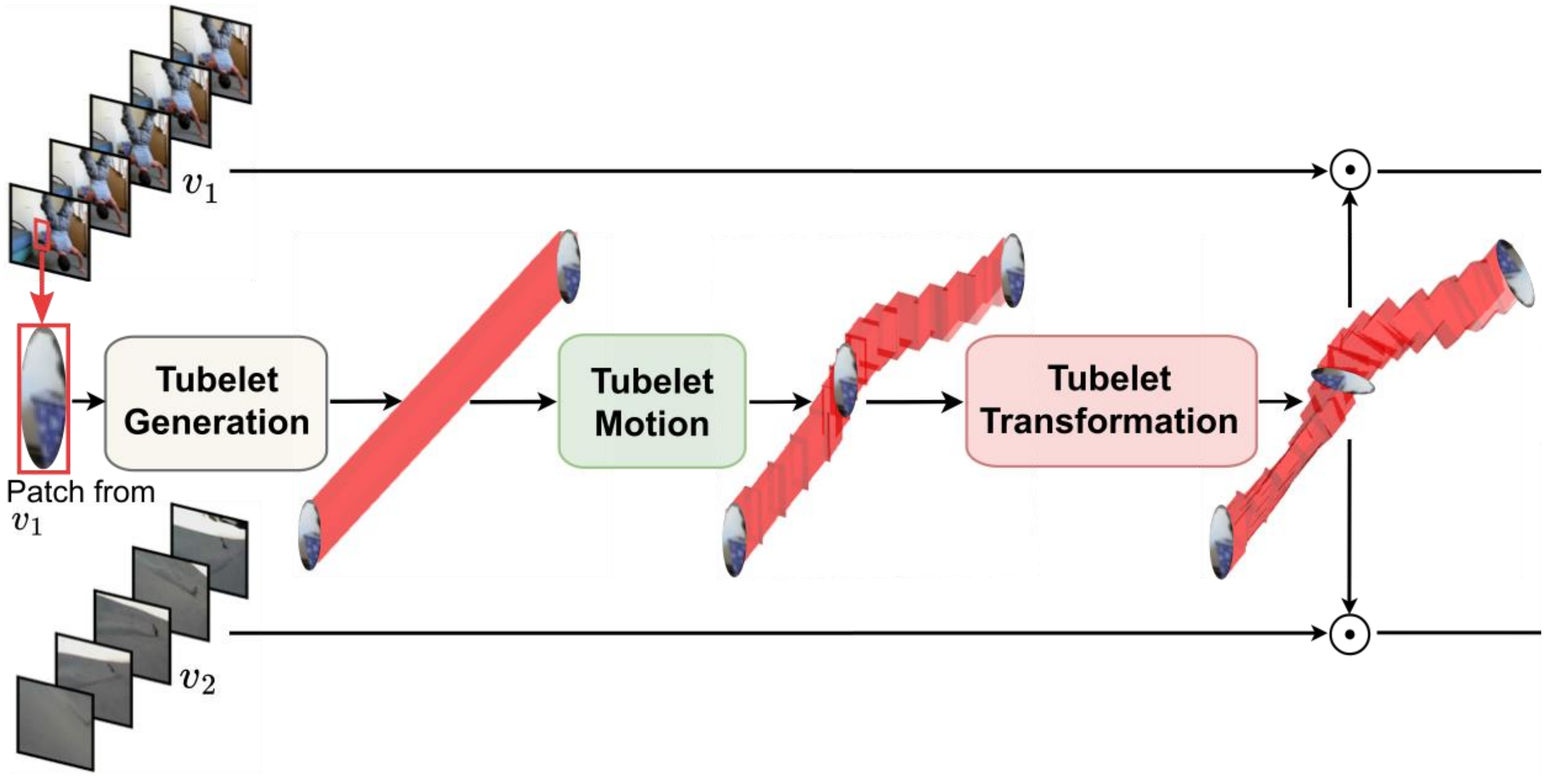
Step 2: Add motion to the patch



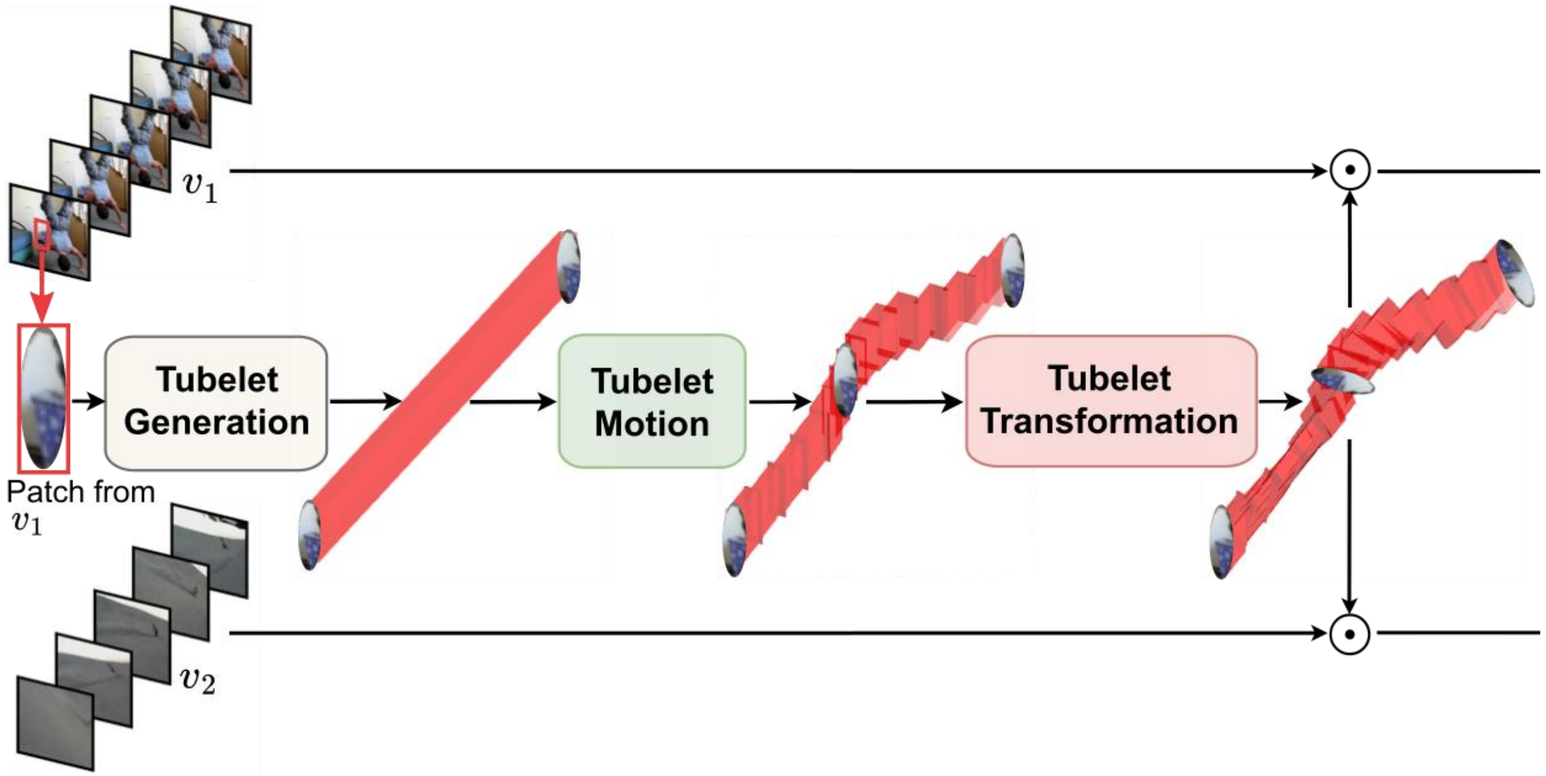


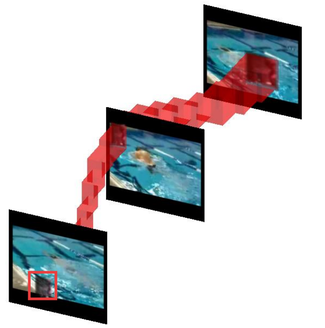


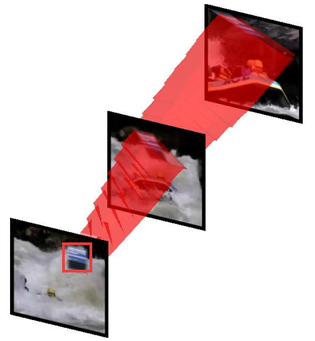
Step 3: Add motion complexity by transformations



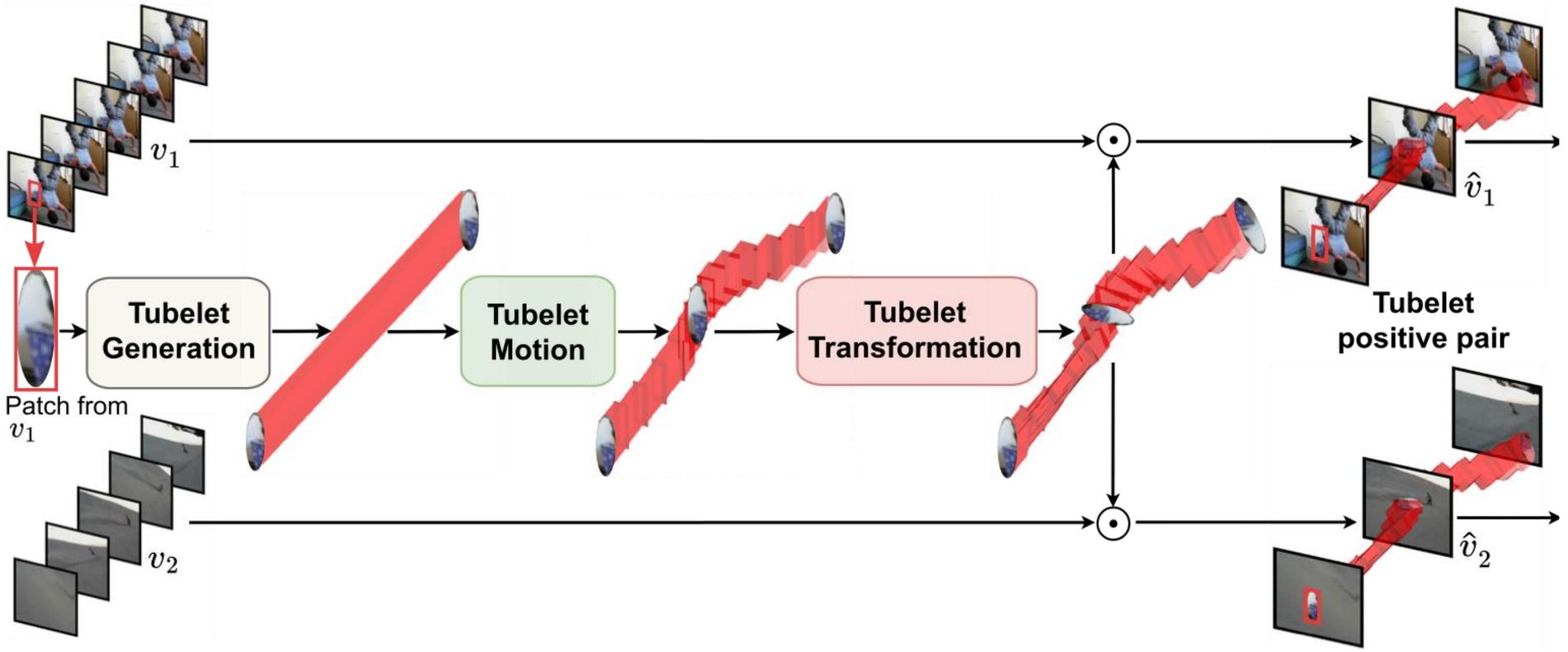
Step 3: Add motion complexity by transformations



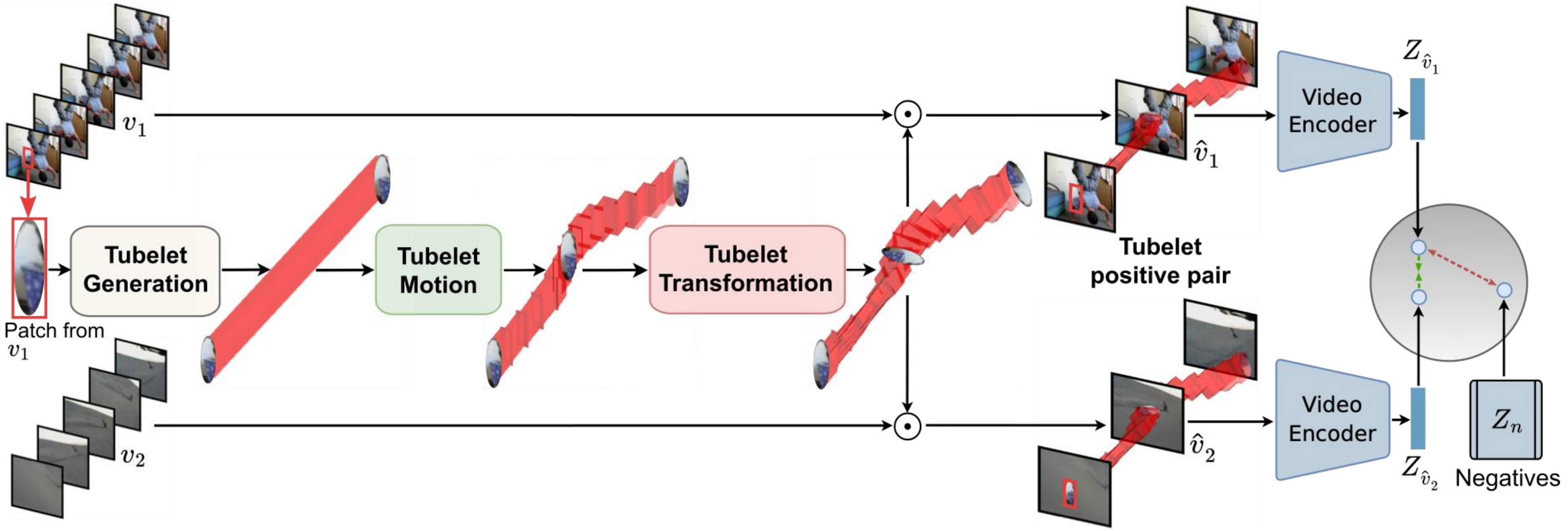




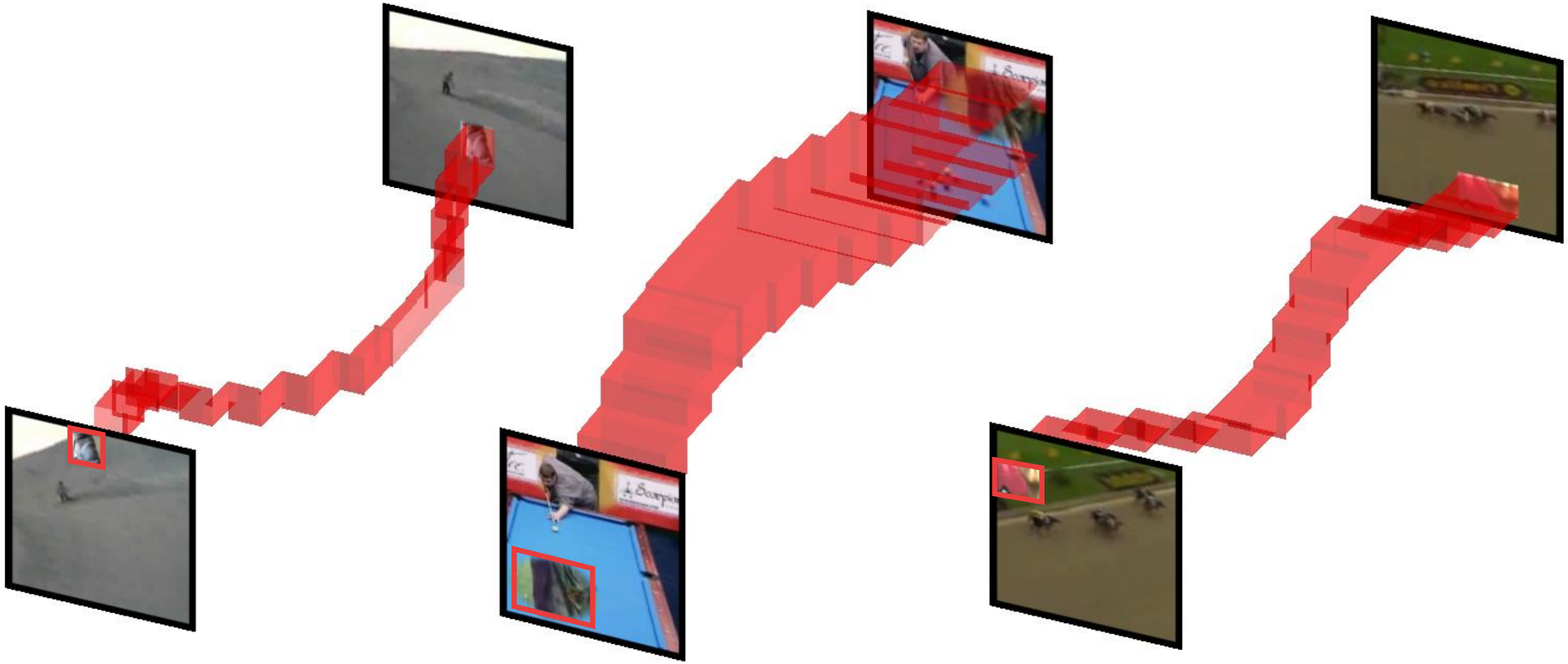
Step 4: Overlay identical tubelet on two clips



Step 5: Tubelet-contrastive learning



Examples of synthetically added tubelets



Ablations

	UCF (10^3)	Gym (10^3)	SSv2-Sub	UB-S1
Video Contrast				
Baseline	57.5	29.5	44.2	84.8
Tubelet Contrast				
Tubelet Generation	48.2	28.2	40.1	84.1
Tubelet Motion	63.0	45.6	47.5	90.3
Tubelet Transformation	65.5	48.0	47.9	90.9

Table 2: **Tubelet-Contrastive Learning** considerably outperforms video contrast on multiple downstream settings. Tubelet motion and transformations are key.

Tubelet Motion	UCF (10^3)	Gym (10^3)	SSv2-Sub	UB-S1
No motion	48.2	28.2	40.1	84.1
Linear	55.2	34.6	45.3	88.5
Non-Linear	63.0	45.6	47.5	90.3

Table 3: **Tubelet Motions.** Learning from tubelets with non-linear motion benefits multiple downstream settings.

Transformation	UCF (10^3)	Gym (10^3)	SSv2-Sub	UB-S1
None	63.0	45.6	47.5	90.5
Scale	65.1	46.5	47.0	90.5
Shear	65.2	47.5	47.3	90.9
Rotation	65.5	48.0	47.9	90.9

Table 4: **Tubelet Transformation.** Adding motion patterns to tubelet-contrastive learning through transformations improves downstream performance. Best results for rotation.

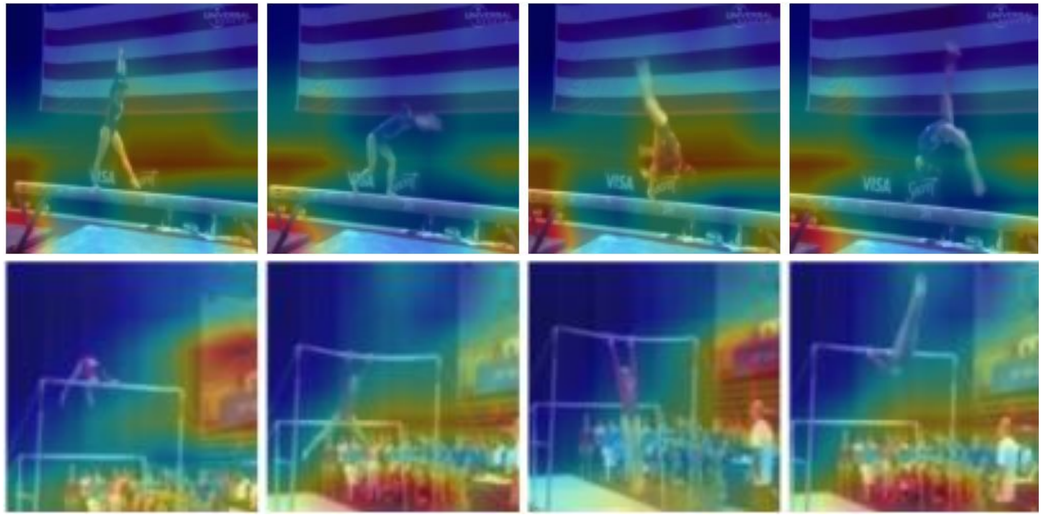
#Tubelets	UCF (10^3)	Gym (10^3)	SSv2-Sub	UB-S1
1	62.0	39.5	47.1	89.5
2	65.5	48.0	47.9	90.9
3	66.5	46.0	47.5	90.9

Table 5: **Number of Tubelets.** Overlaying two tubelets in positive pairs improves downstream performance.

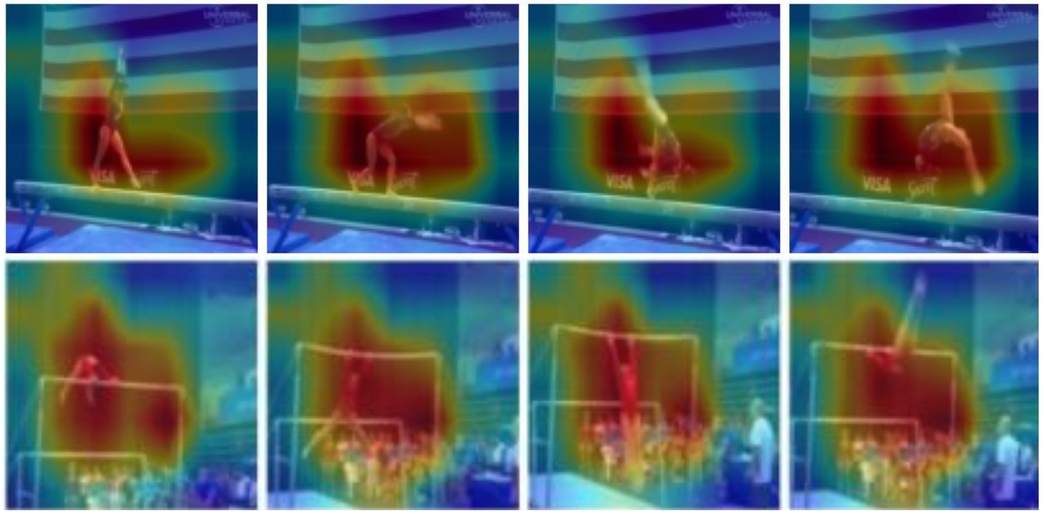
Ablations confirm intuitions

What does the model learn?

Video-contrastive learning

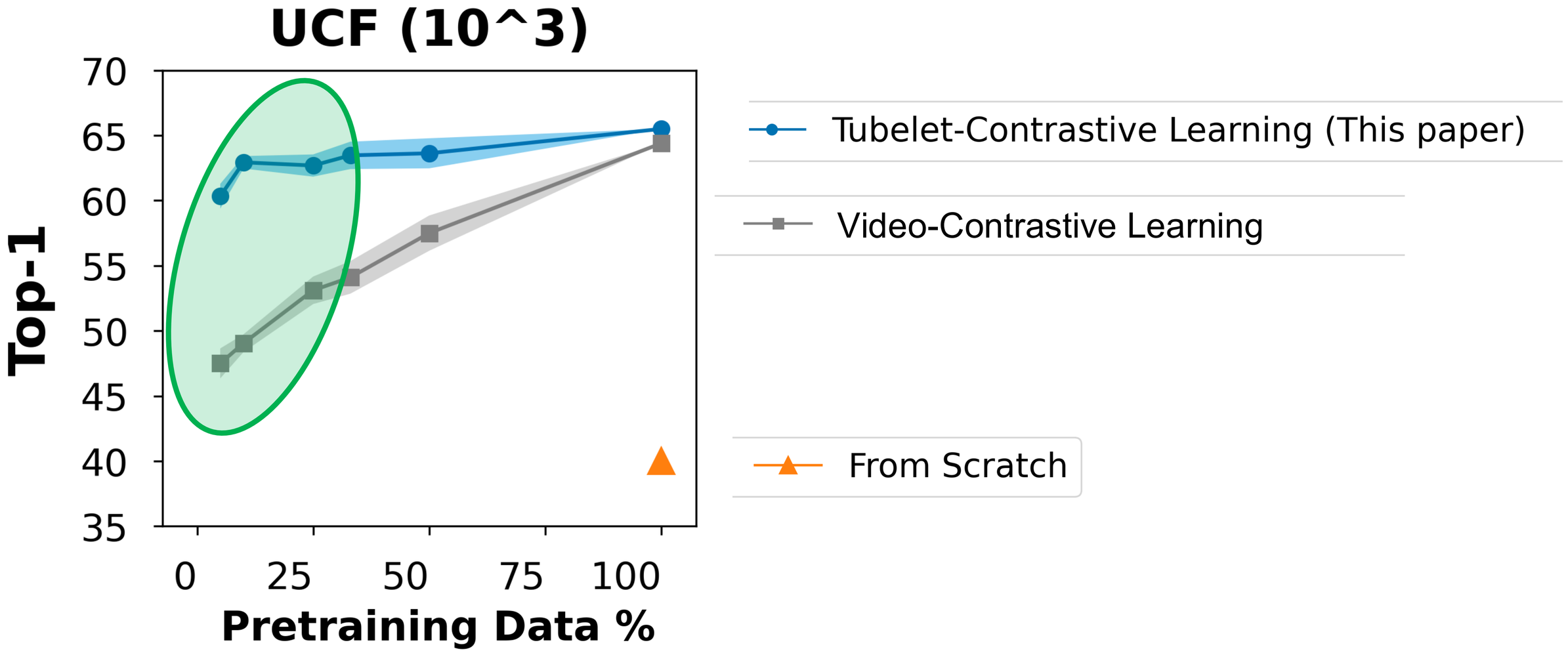


Proposed tubelet-contrastive learning

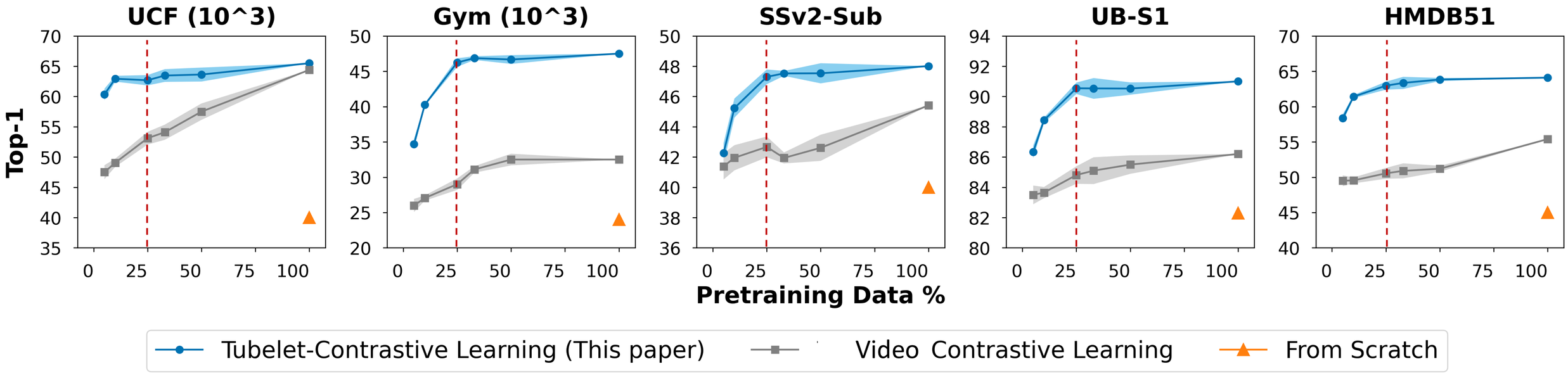


Without seeing any FineGym videos during training, our approach attends to motion

Adding synthetic motion improves data efficiency



Key benefit: we need 4x less video data

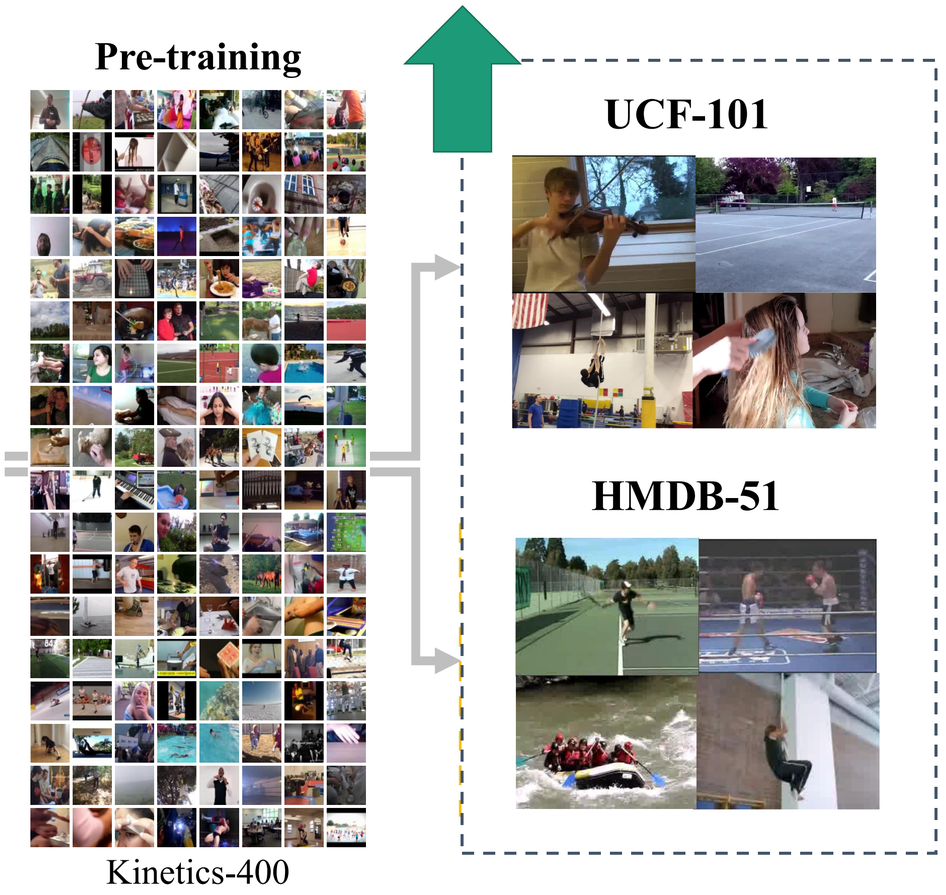


Tubelets simulate a richer variety of fine-grained motion than present in the original video

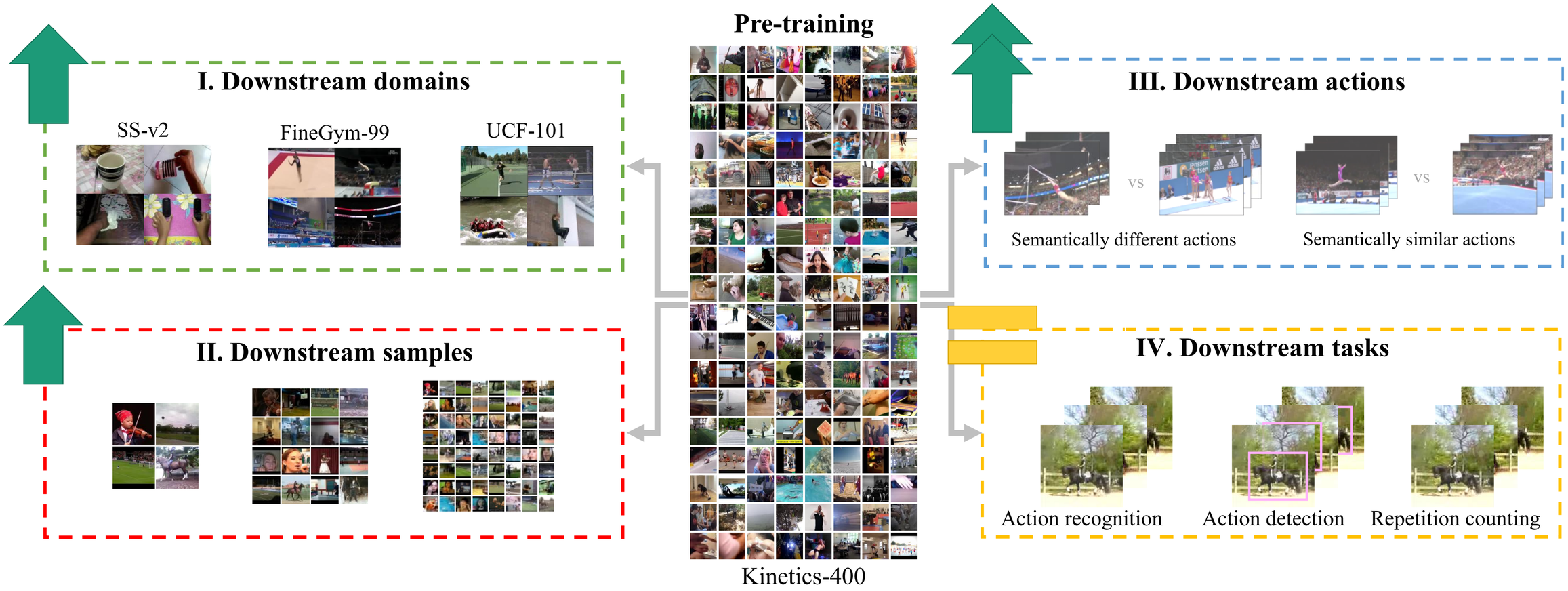
Solid accuracy gain on UCF-101 and HMDB-51

R(2+1)D Backbone pretrained on Kinetics-400

Method	Modality	UCF101	HMDB51
Pace Prediction [76]	RGB	77.1	36.6
VideoMoCo [56]	RGB	78.7	49.2
RSPNet [58]	RGB	81.1	44.6
SRTC [46]	RGB	82.0	51.2
FAME [10]	RGB	84.8	53.5
MCN [45]	RGB	84.8	54.5
AVID-CMA [52]	RGB+Audio	87.5	60.8
TCLR [9]	RGB	88.2	60.0
TE [31]	RGB	88.2	62.2
CtP [74]	RGB	88.4	61.7
MotionFit [20]	RGB+Flow	88.9	61.4
GDT [57]	RGB+Audio	89.3	60.0
Ours w/ mini-Kinetics	RGB	**90.7**	**65.0**
Ours w/ Kinetics	RGB	**91.0**	**64.1**



Generalization on SEVERE-benchmark



Generalization on SEVERE-benchmark

		Domains		Samples		Actions		Tasks			
	Backbone	SSv2	Gym99	UCF (10^3)	Gym (10^3)	FX-S1	UB-S1	UCF-RC↓	Charades	**Mean**	**Rank↓**
SVT [61]	ViT-B	59.2	62.3	83.9	18.5	35.4	55.1	0.421	35.5	51.0	8.9
VideoMAE [71]	ViT-B	69.7	85.1	77.2	27.5	37.0	78.5	0.172	12.6	58.1	8.3
Supervised [72]	R(2+1)D-18	60.8	92.1	86.6	51.3	79.0	87.1	0.132	23.5	70.9	3.9
None	R(2+1)D-18	57.1	89.8	38.3	22.7	46.6	82.3	0.217	7.9	52.9	11.6
SeLaVi [2]	R(2+1)D-18	56.2	88.9	69.0	30.2	51.3	80.9	0.162	8.4	58.6	11.0
MoCo [23]	R(2+1)D-18	57.1	90.7	60.4	30.9	65.0	84.5	0.208	8.3	59.5	9.1
VideoMoCo [56]	R(2+1)D-18	59.0	90.3	65.4	20.6	57.3	83.9	0.185	10.5	58.6	9.1
Pre-Contrast [69]	R(2+1)D-18	56.9	90.5	64.6	27.5	66.1	86.1	0.164	8.9	60.5	9.0
AVID-CMA [51]	R(2+1)D-18	52.0	90.4	68.2	33.4	68.0	87.3	0.148	8.2	61.6	9.0
GDT [57]	R(2+1)D-18	58.0	90.5	**78.4**	45.6	66.0	83.4	**0.123**	8.5	64.8	8.6
RSPNet [58]	R(2+1)D-18	59.0	91.1	74.7	32.2	65.4	83.6	0.145	9.0	62.6	8.0
TCLR [8]	R(2+1)D-18	59.8	91.6	72.6	26.3	60.7	84.7	0.142	**12.2**	61.7	7.6
CtP [74]	R(2+1)D-18	59.6	92.0	61.0	32.9	79.1	88.8	0.178	9.6	63.2	5.6
Ours w/ mini-Kinetics	R(2+1)D-18	59.4	92.2	65.5	**48.0**	78.3	90.9	0.150	9.0	66.0	5.4
Ours w/ Kinetics	R(2+1)D-18	**60.2**	**92.8**	65.7	47.0	**80.1**	**91.0**	0.150	10.3	**66.5**	**4.1**

Better generalization, even when using the 3x smaller Mini-Kinetics for pretraining.

Key takeaways

Contrastive learning with **synthetic tubelets** provides:

Simple and effective self-supervised video representation learning.

Data-efficient pretraining with less unlabelled video data.

Better generalization to diverse video domains and fine-grained tasks.

1c. The problem of video masked auto encoding

Mohammadreza Salehi
University of Amsterdam

Michael Dorkenwald
University of Amsterdam

Fida Mohammad Thoker
KAUST

Efstratios Gavves
University of Amsterdam

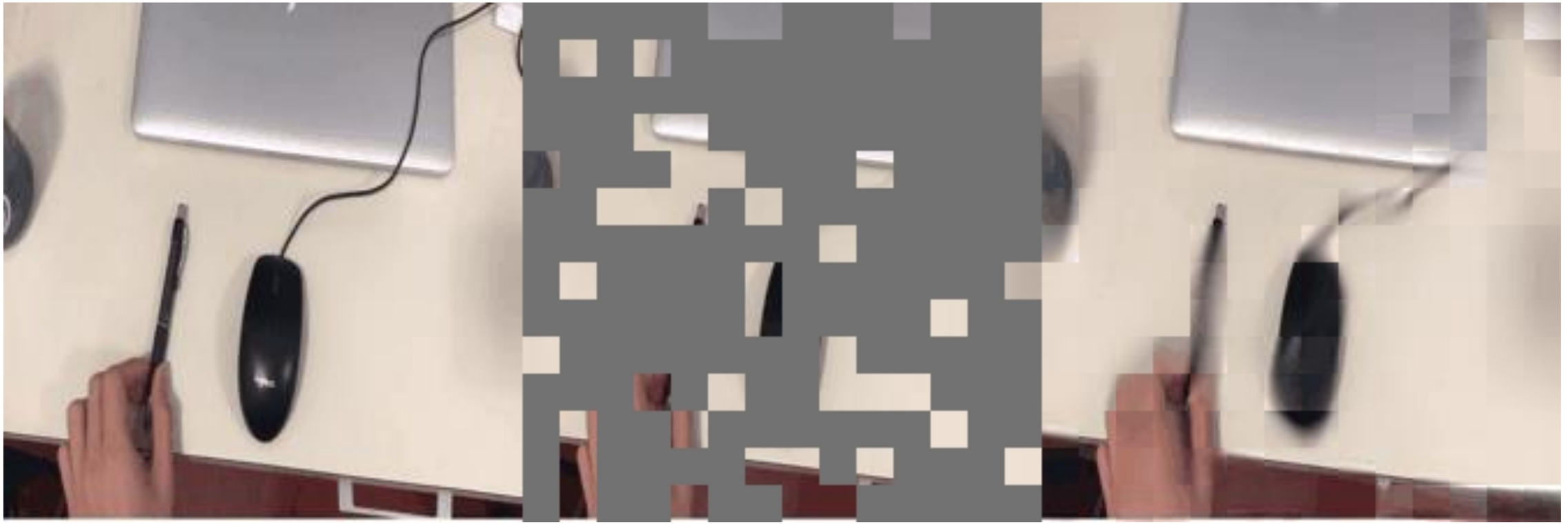
Cees Snoek
University of Amsterdam

Yuki Asano
University of Amsterdam
Currently: UoT Neuremberg

SIGMA: Sinkhorn-Guided Masked Video Modeling. In *ECCV* 2024.

QUVA Deep Vision Lab

Video MAE



Input video | Masked input (80%) | Reconstructed output video

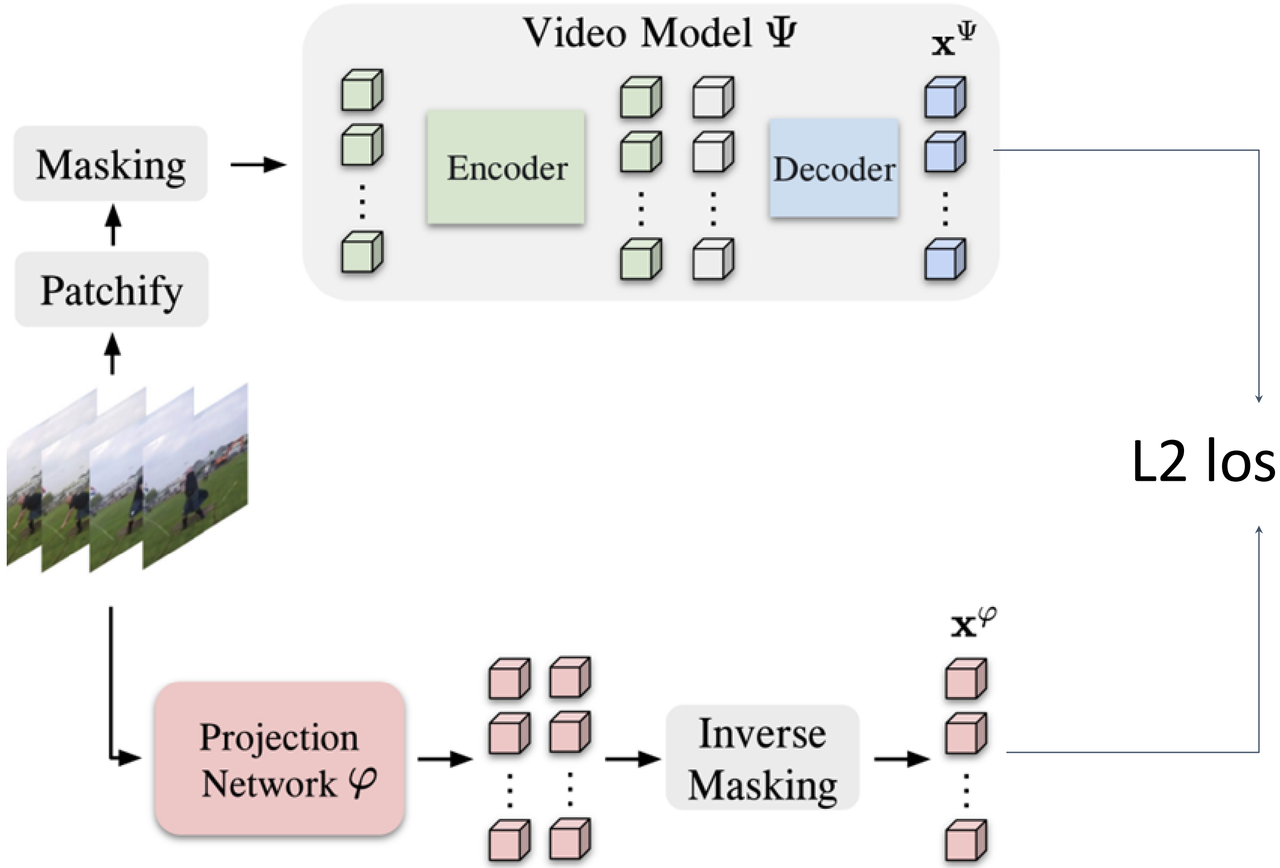
Video MAE Challenge: Poor motion modeling

Input video

Masked input

Reconstructed video

From pixel to feature reconstruction

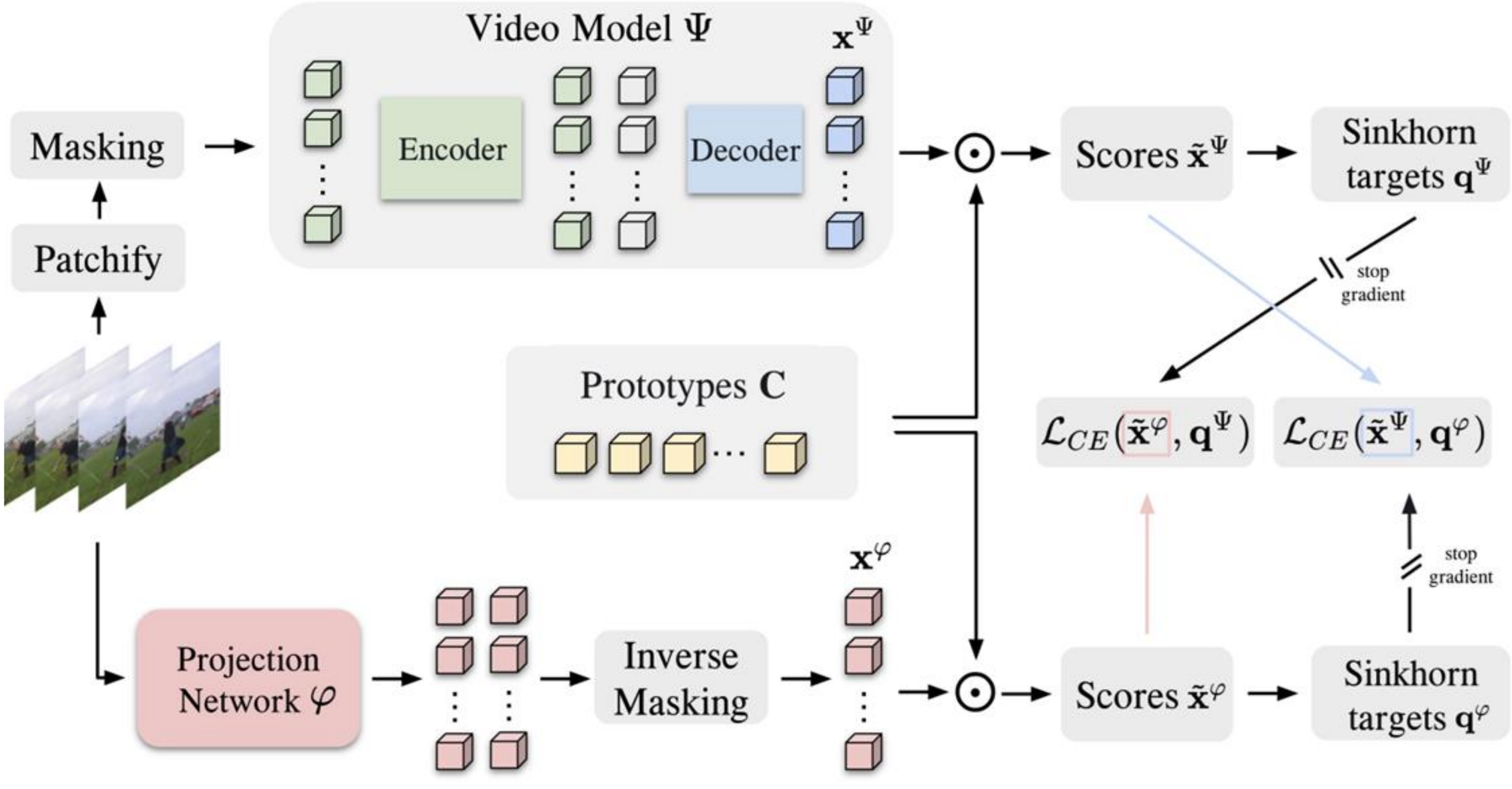


L2 loss

Trivial Solution! Mode Collapse.

Online clustering methods can help

SIGMA: Sinkhorn-Guided Masked Video Modeling

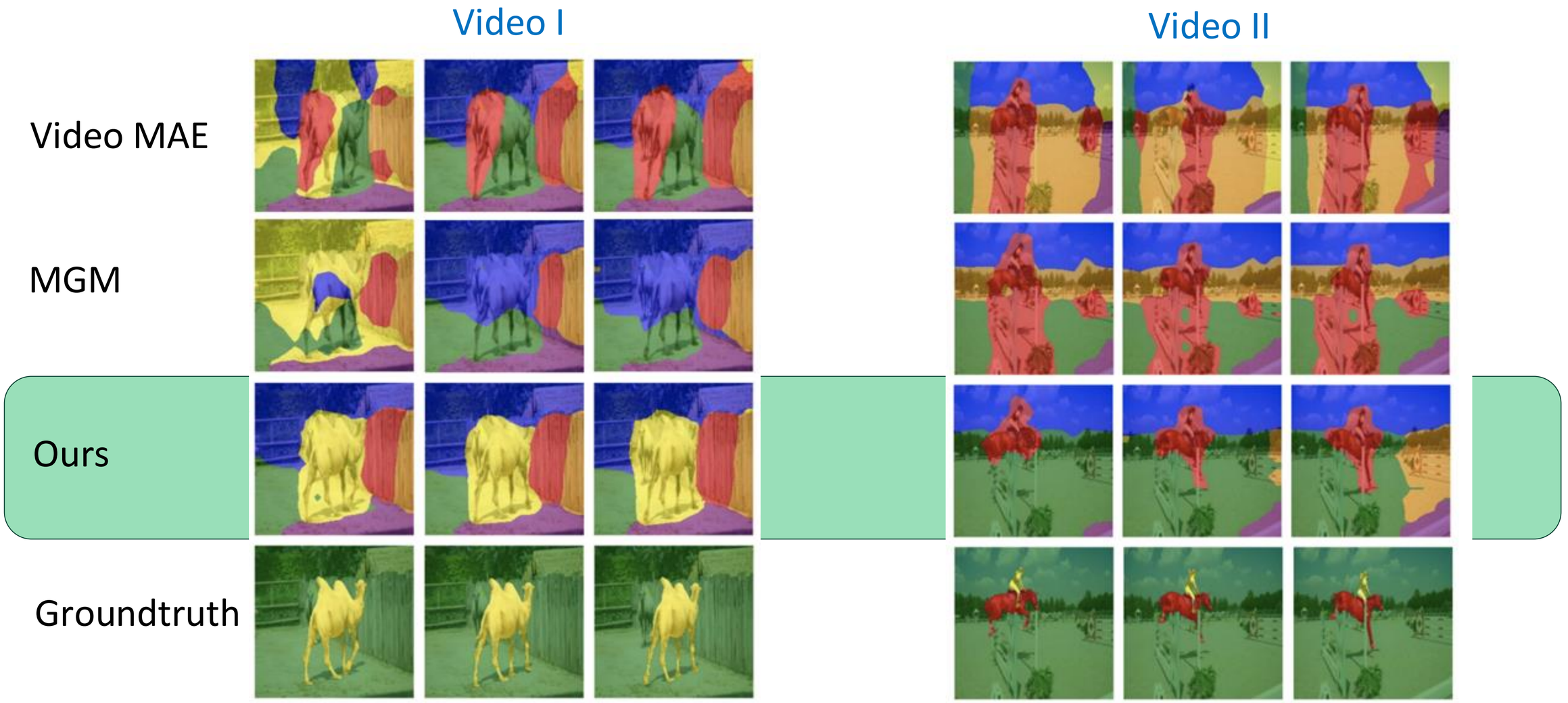


Generalization on SEVERE-benchmark

	Domains		Samples (10^3)		Actions		Tasks		Mean
	SSv2	Gym99	UCF	Gym	FX-S1	UB-S1	UCF-RC↓	Charades	
SVT [61]	59.5	62.3	<u>83.9</u>	18.5	35.4	55.1	0.421	**35.5**	51.0
MVD [83]	<u>70.0</u>	82.5	66.7	17.5	31.3	50.5	0.184	16.1	52.1
VideoMAE [76]	68.6	86.6	74.6	25.9	42.8	65.3	<u>0.172</u>	12.6	57.4
MGMAE [38]	68.9	87.2	77.2	24.0	33.7	79.5	0.181	17.9	58.8
SIGMA-MLP (ours)	69.8	<u>87.4</u>	80.2	<u>26.8</u>	<u>46.0</u>	<u>79.7</u>	0.178	20.1	<u>61.5</u>
SIGMA-DINO (ours)	**70.9**	**89.7**	**84.1**	**28.0**	**55.1**	**79.9**	**0.169**	<u>23.3</u>	**64.3**

SIGMA achieves strong generalization, outperforming prior works across most configurations.

Unsupervised video object segmentation on DAVIS



Key takeaways

Sinkhorn-clustering leads to more abstract mask reconstruction

Alleviates **training collapse,** profits from **pretrained image models**

Better generalization to video domains, samples and fine-grained actions.

2. Role of time

Piyush Bagad

Makarand Tapaswi

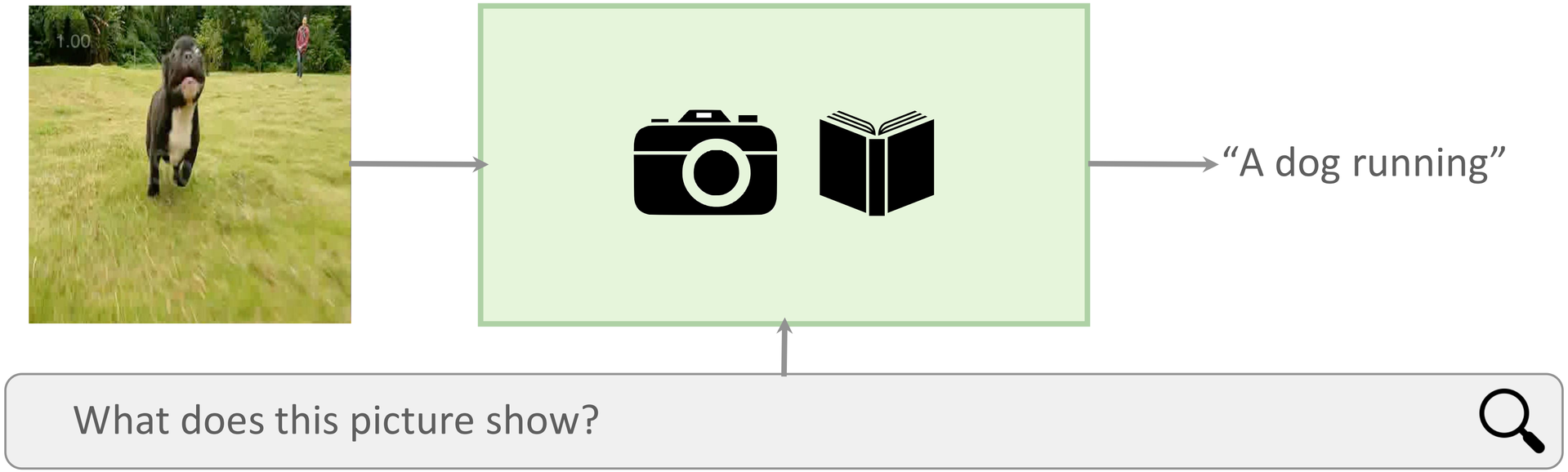
Piyush Bagad, Makarand Tapaswi, Cees G M Snoek: **Test of Time: Instilling Video-Language Models with a Sense of Time.** In: CVPR, 2023.

2a. The problem of sensing video time

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

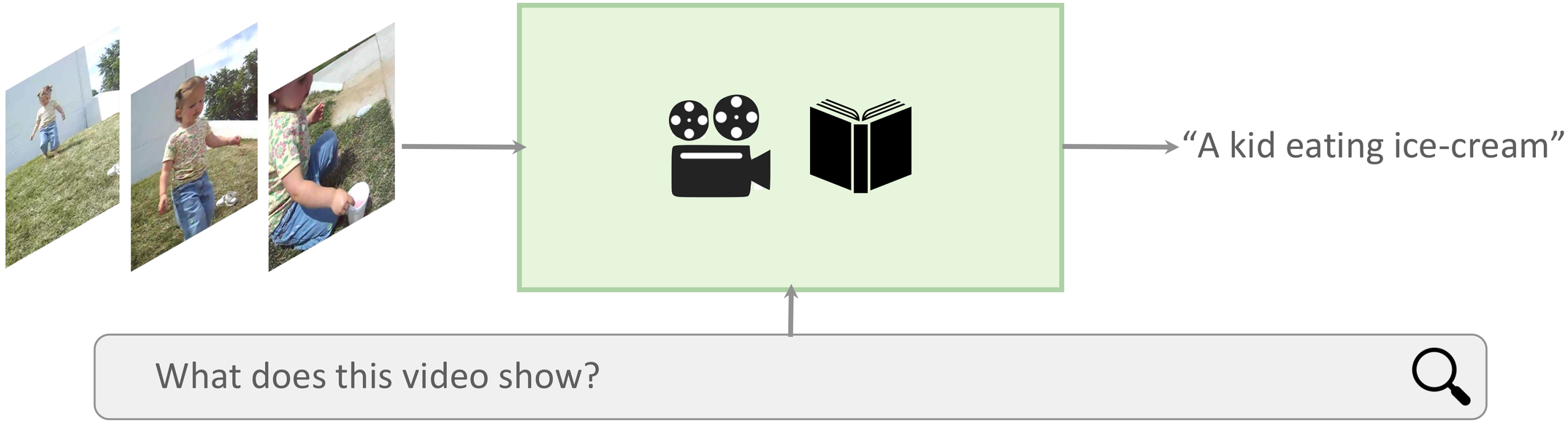
The problem

- Language interface + a few (or no) image samples



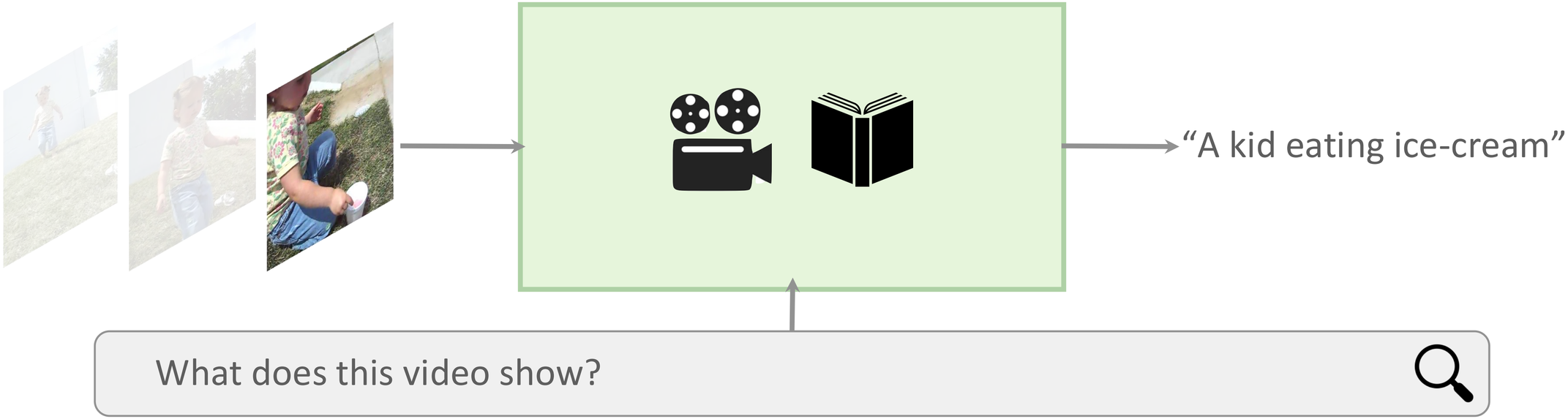
The problem

- Language interface + a few (or no) image samples
- Particularly attractive for videos given high cost



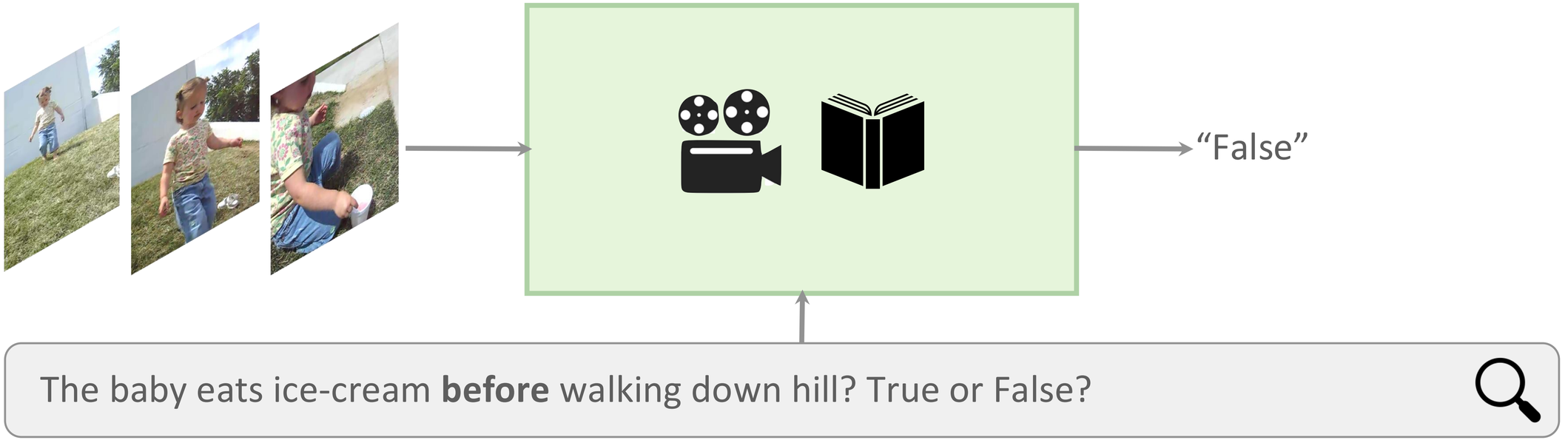
The problem

- Do video-language models truly understand time?



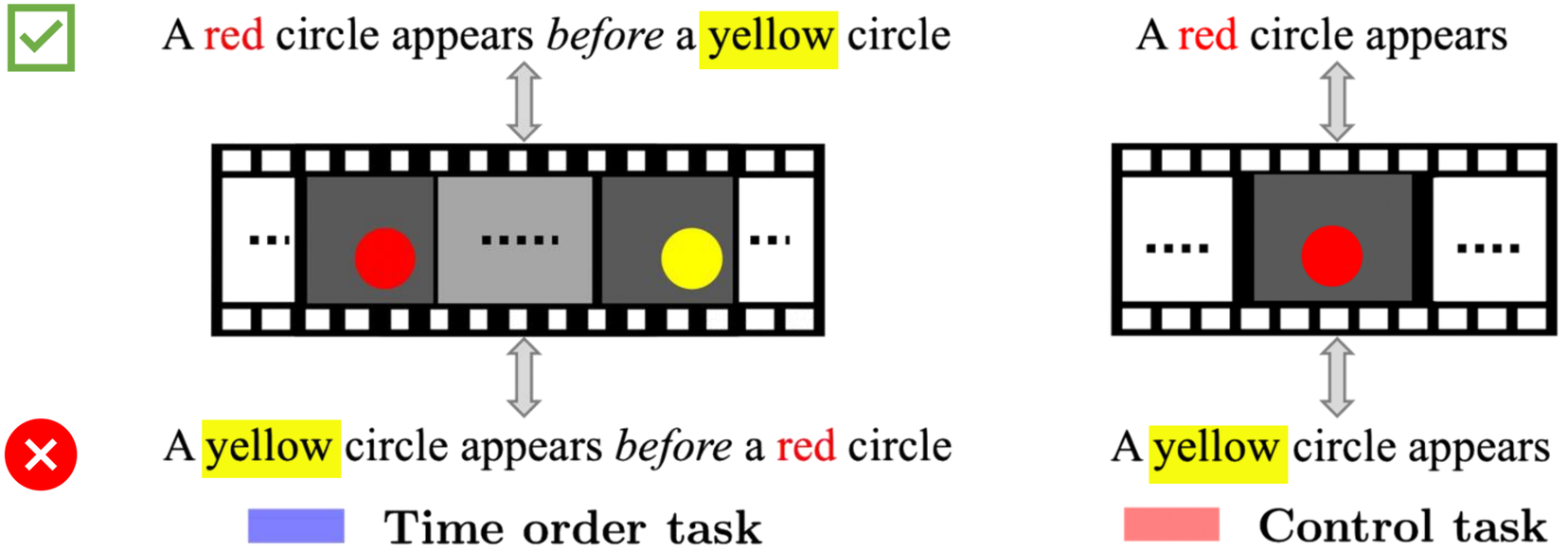
The problem

- Do video-language models truly understand time?
- Our idea for a “test of time”: ask questions that have temporal relations



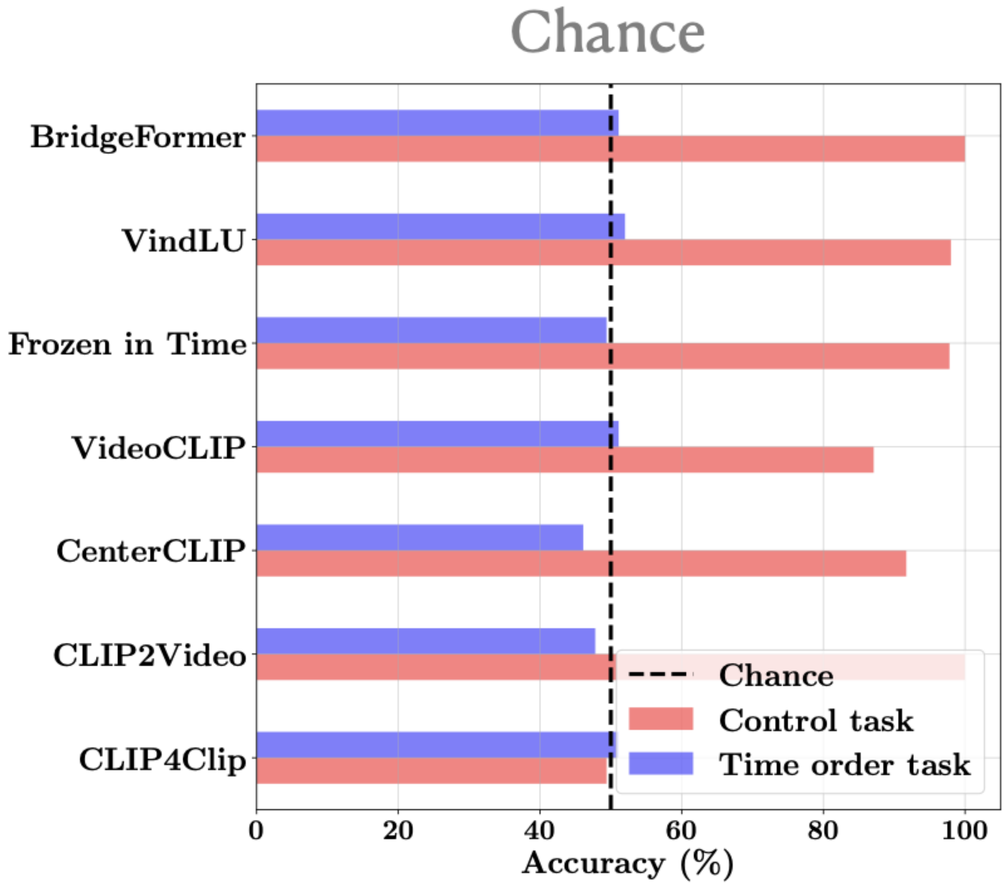
The test of time

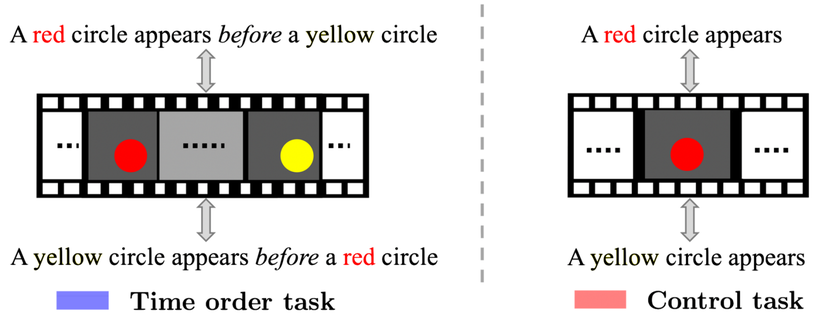
- Synthetic benchmark
- Simple ‘true’ or ‘false’ predictions



Existing models fail this test of time

- We pick a suite of seven openly available video-language models
- While excelling at the control task, they all fail at the time-order task







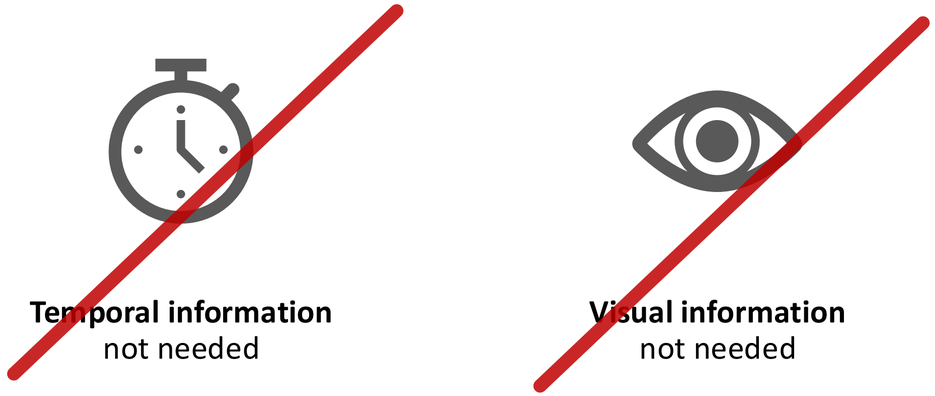

Daniel Cores Michael Dorkenwald Manuel Mucientes Yuki Asano

Daniel Cores*, Michael Dorkenwald*, Manuel Mucientes, Cees G. M. Snoek, Yuki M. Asano: **Lost in Time: A New Temporal Benchmark for VideoLLMs**. In: BMVC, 2025.

2b. The problem of measuring video time

BMVC 2025

Our findings on existing VideoLLM benchmarks



We propose a new video temporal benchmark TVBench

Does time matter?

What pattern does the video depict?

- ✔ A black hexagon displayed various colored hands.
- ✘ A red circle showed multiple green feet.
- ✘ A blue pentagon portrayed different colored heads.
- ✘ A yellow square showed various purple arms

Example from MVBench

Does time matter?

Which one of these descriptions correctly matches the actions in the video?

✔ skiing	✘ competing
✘ sliding	✘ balancing

Example from MVBench

Does time matter?

Which one of these descriptions correctly matches the actions in the video?

✓ skiing

✗ commuting

✗ sliding

✗ balancing

Many questions can be answered without temporal information 😔

Example from MVBench

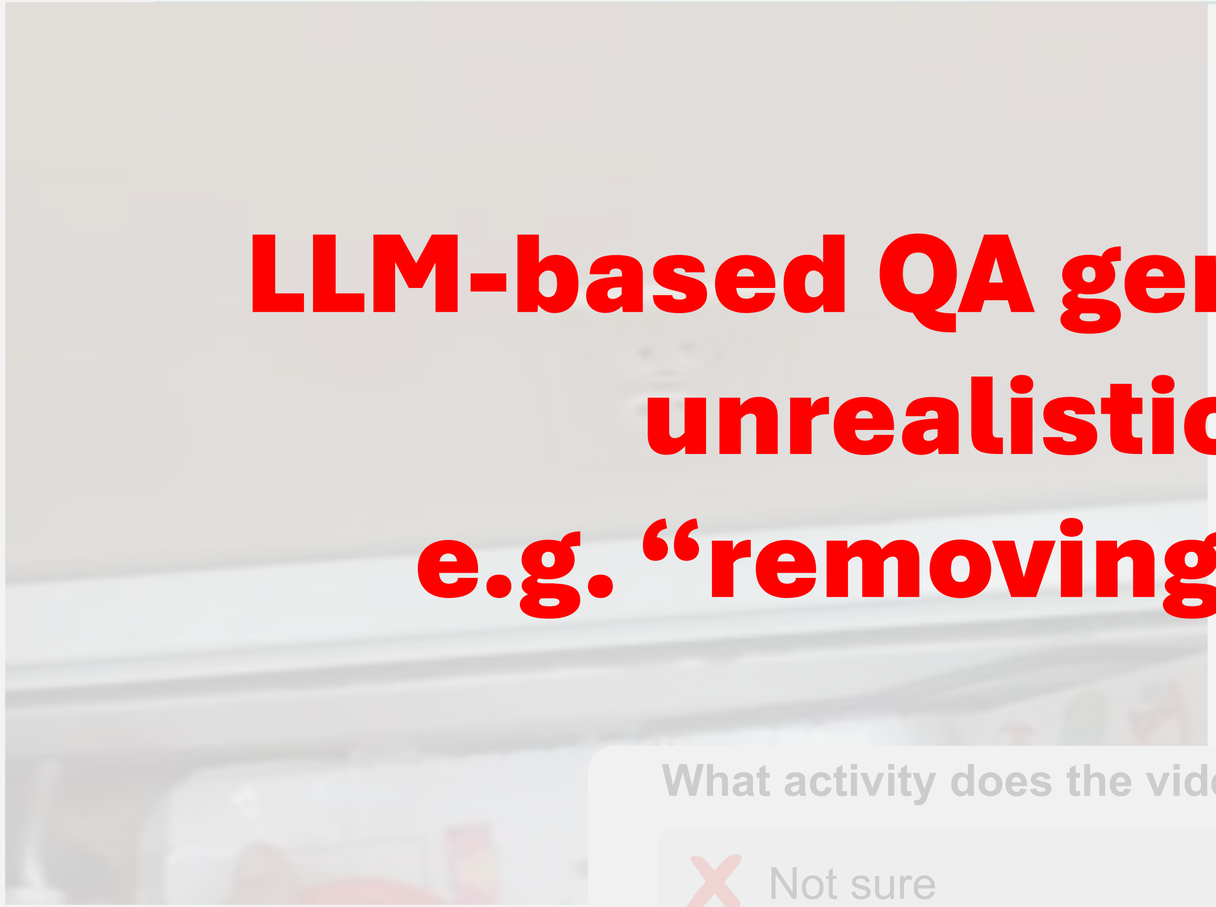
Can questions be answered solely by text?

What activity does the video depict?

- ✗ Not sure
- ✓ Plugging something into something.
- ✗ Removing something into something.

Example from MVBench

Can questions be answered solely by text?



LLM-based QA generation introduces unrealistic candidates

e.g. “removing sth into sth” 😔

What activity does the video depict?

- ✗ Not sure
- ✓ Plugging something into something.
- ✗ Removing something into something.

Example from MVBench

World knowledge sufficient?

What did Mrs. Koothrappali say after Raj told her that all the other guys going to the north pole?

- ✗ Have fun
- ✗ Well then go
- ✓ I don’t care what the other guys are doing
- ✗ What other guys?
- ✗ They not my son

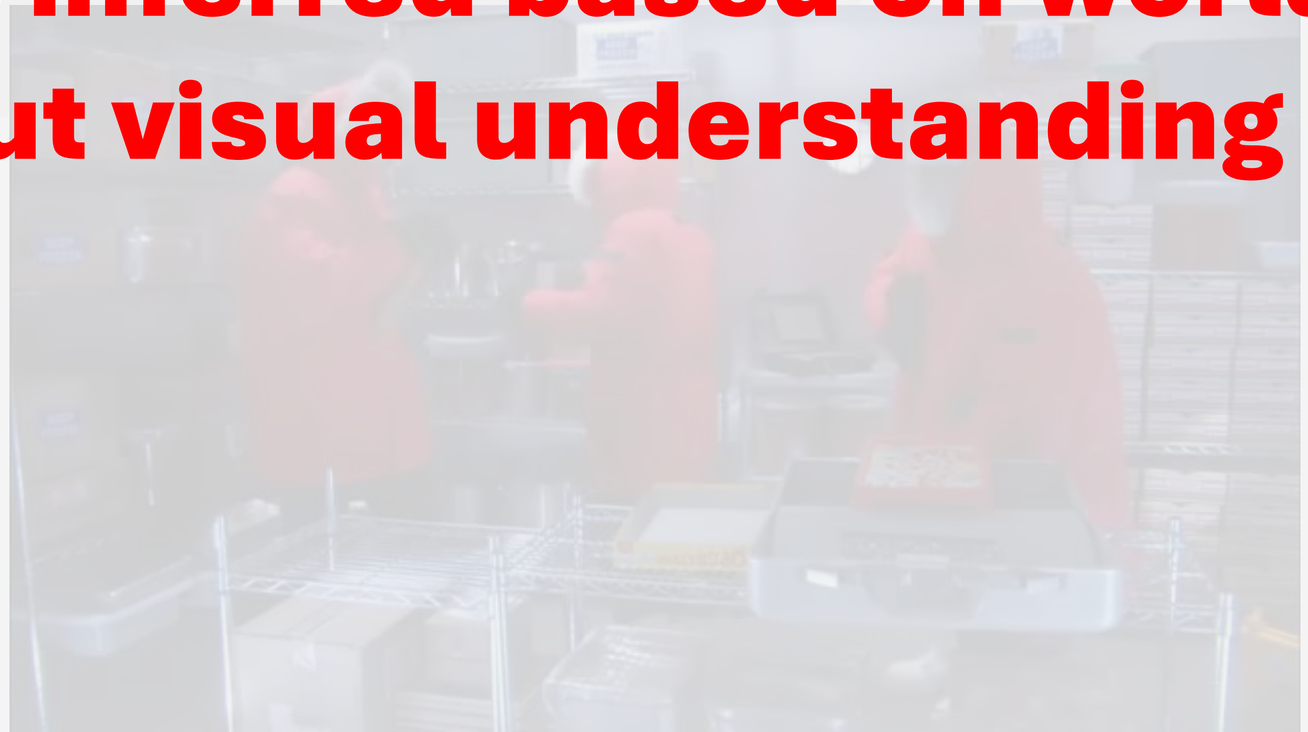
Example from MVBench

World knowledge sufficient?

What did Mrs. Koothrappali say after Raj told her that all the other guys going to the north pole?

- ✗ Have fun
- ✗ Well then go
- ✓ I don't care what the other guys are doing
- ✗ What other guys?
- ✗ They not my son

Answers can be inferred based on world knowledge without visual understanding 😔



Example from MVBench

TVBench: Temporal Video-Language Benchmark

Temporal information is crucial

Visual information is crucial

TVBench

Temporal information is **crucial** to discard incorrect candidates

Can you choose the option that matches how scenes change in the video?

✗ From the bus to prison	✓ From prison to the bus

TVBench

Temporal information is **crucial** to discard incorrect candidates

Locate the amusing part of the video?

- ✗ In the middle of the video
- ✗ At the beginning of the video
- ✗ Throughout the entire video
- ✓ At the end of the video

TVBench

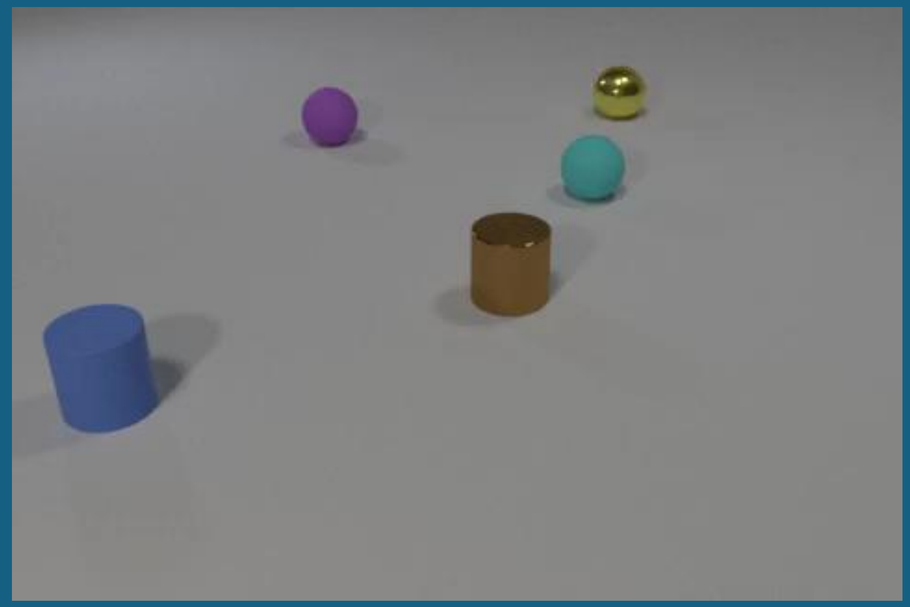
Visual information is **crucial** to discard incorrect candidates

Can you identify when the action ‘person holding a pillow’ happens in the video?

✗ Throughout the entire video	✗ In the middle of the video
✗ At the end of the video	✓ At the beginning of the video

TVBench

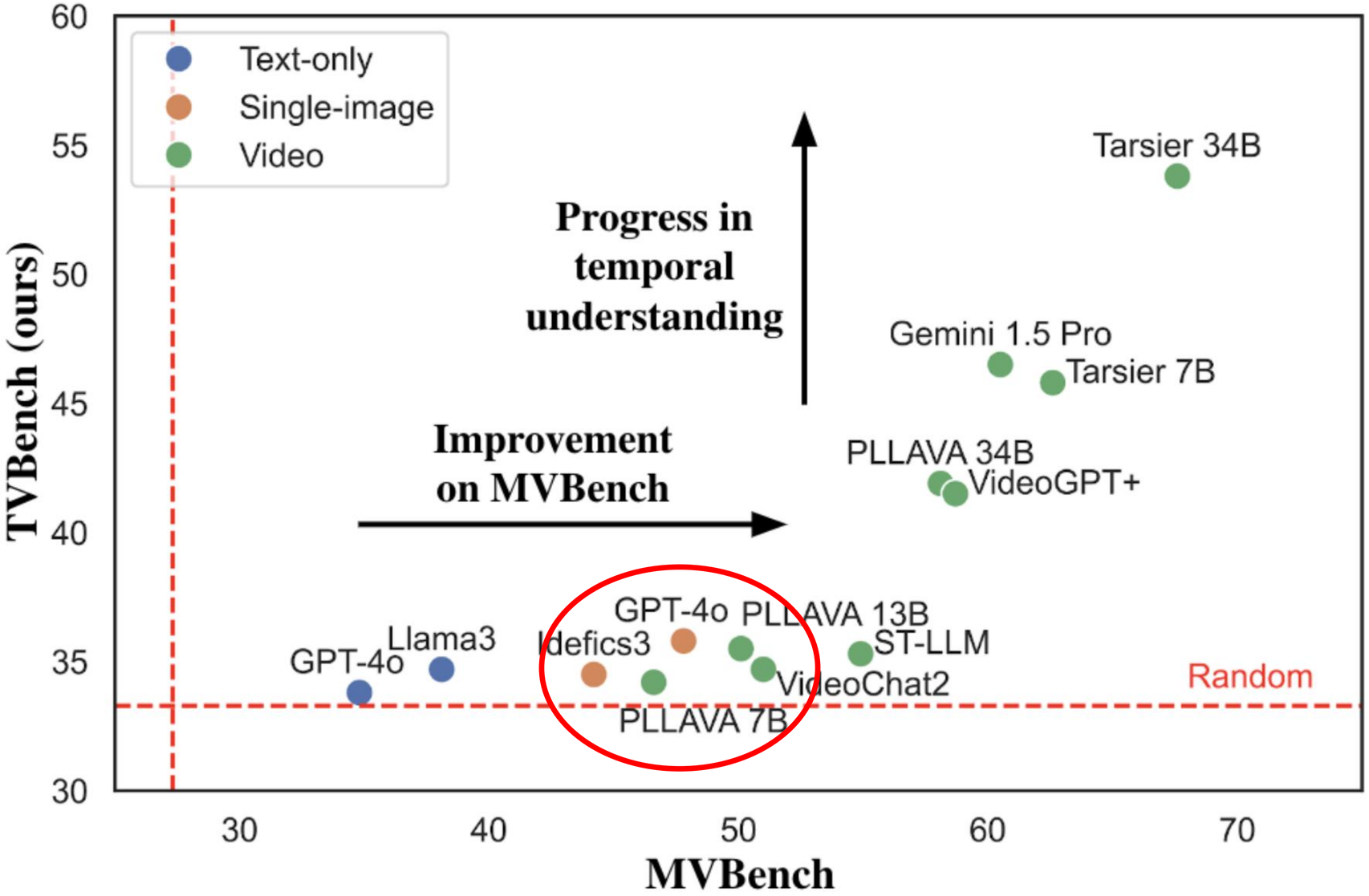
Visual information is **crucial** to discard incorrect candidates



Which direction does the yellow sphere move in the video?

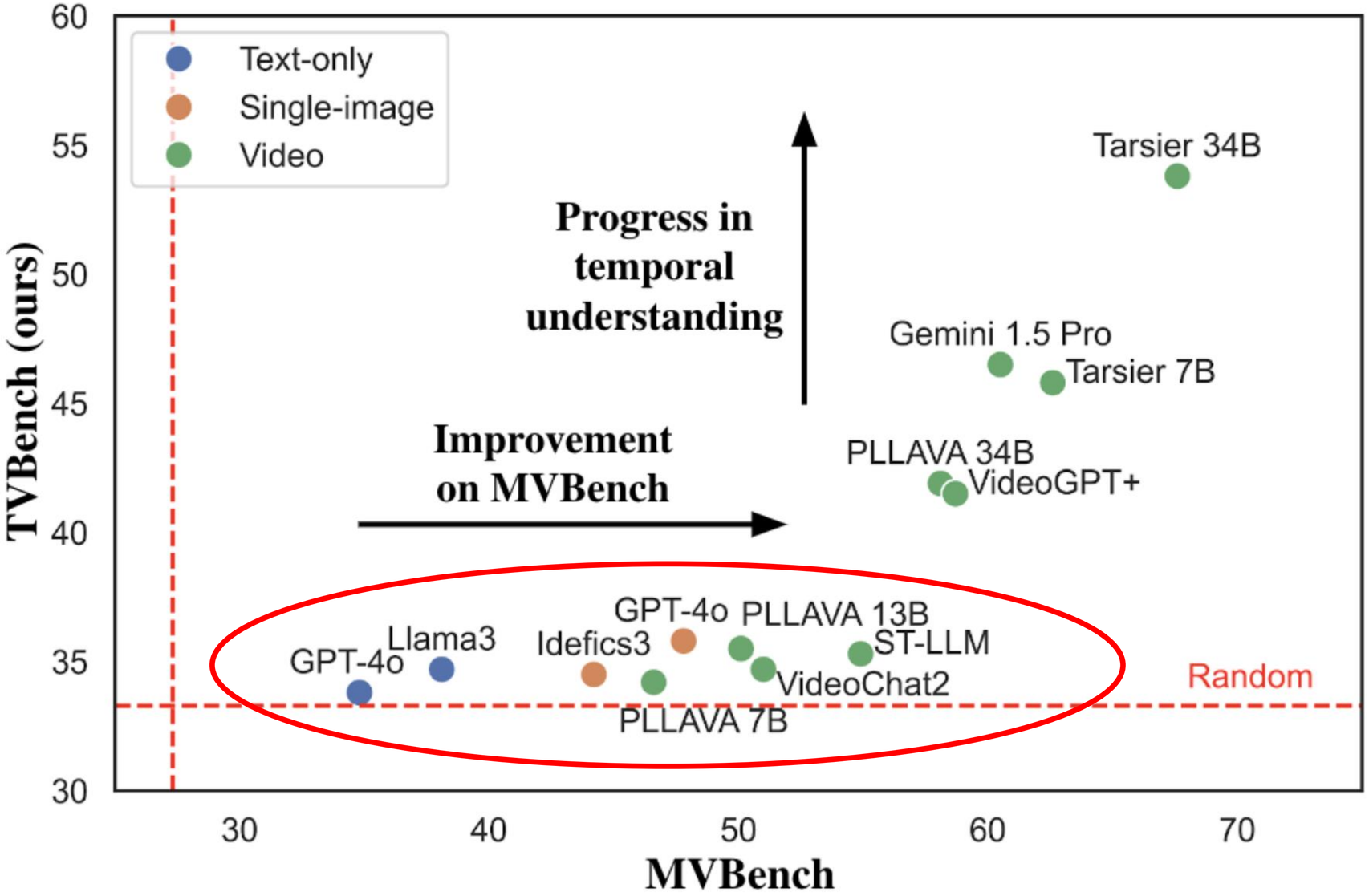
✗ Up and to the right	✗ Up and to the left
✓ Down and to the left	✗ Down and to the right

MVBench vs **TVBench**



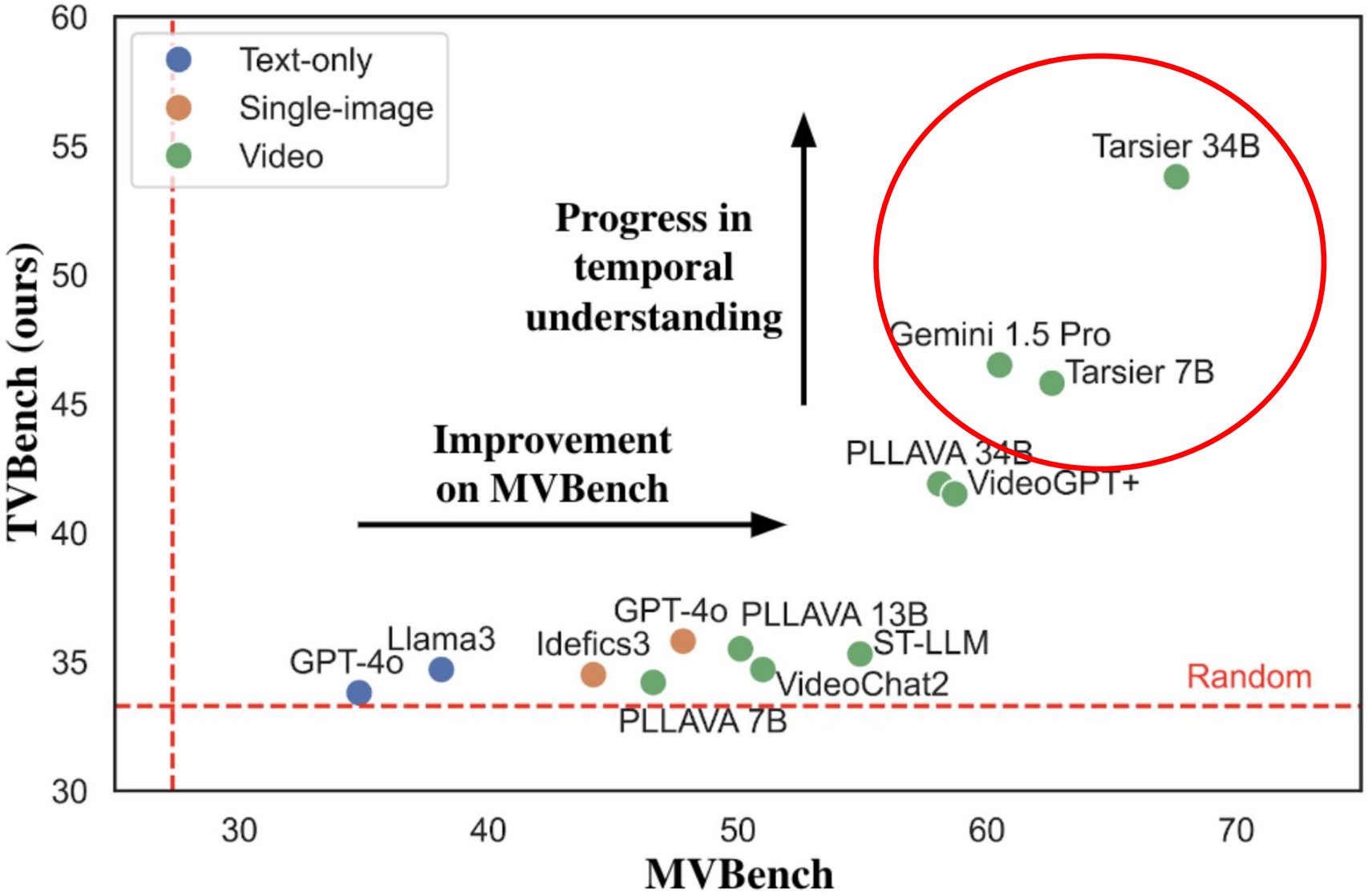
Single-image models achieve competitive performance on MVBench

MVBench vs **TVBench**



Text-only, image and many video models get random performance on TVBench

MVBench vs **TVBench**



Only temporal video models significantly outperform random baseline on TVBench

TVBench: Temporal Video-Language Benchmark

Temporal information is crucial

Visual information is crucial

Hugging Face

https://huggingface.co/datasets/FunAILab/TVBench

arXiv

https://arxiv.org/abs/2410.07752

Conclusions

Many VideoLLMs are **poor temporal performers**

Many VideoLLM benchmarks are **poor temporal evaluators**

TVBench assesses temporal understanding of VideoLLM’s explicitly

3. Role of multimodality

3.a Adaptation by video-audio

Yunhua Zhang
University of Amsterdam

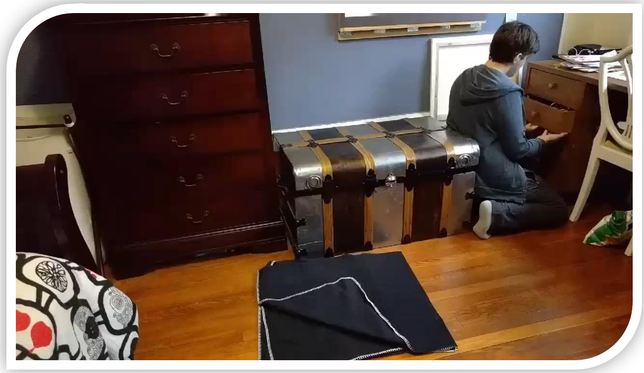
Hazel Doughty
University of Amsterdam

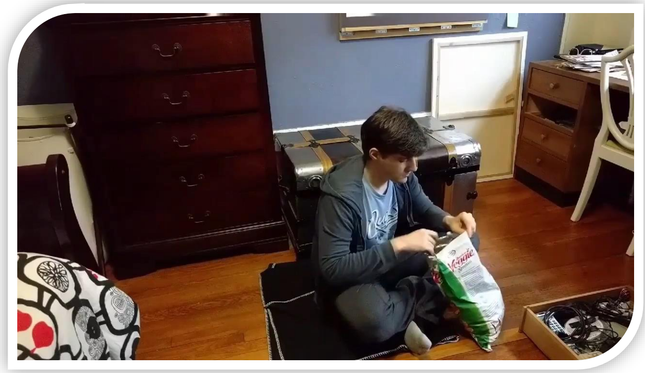
Ling Shao
Inception Institute of AI

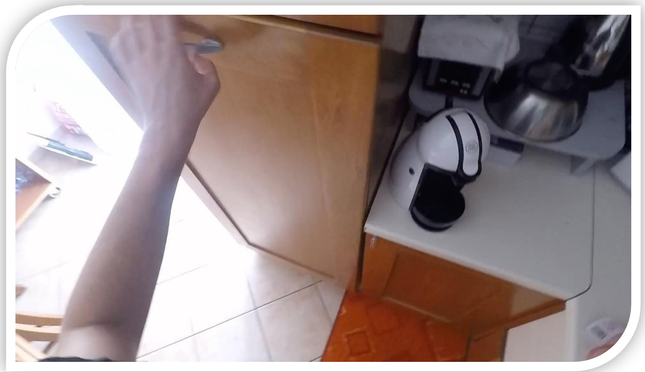
Cees Snoek
University of Amsterdam

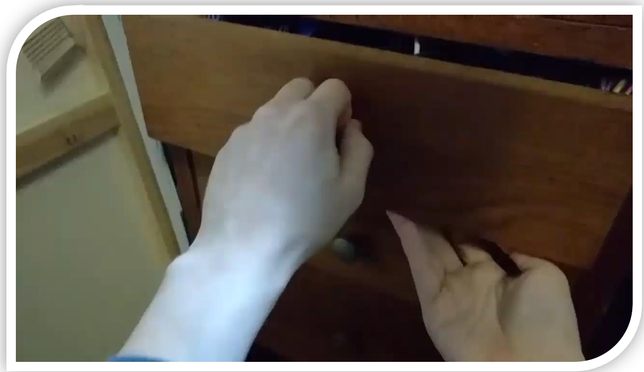
Audio-Adaptive Activity Recognition Across Video Domains. In *CVPR* 2022.

Activity recognition under domain shift









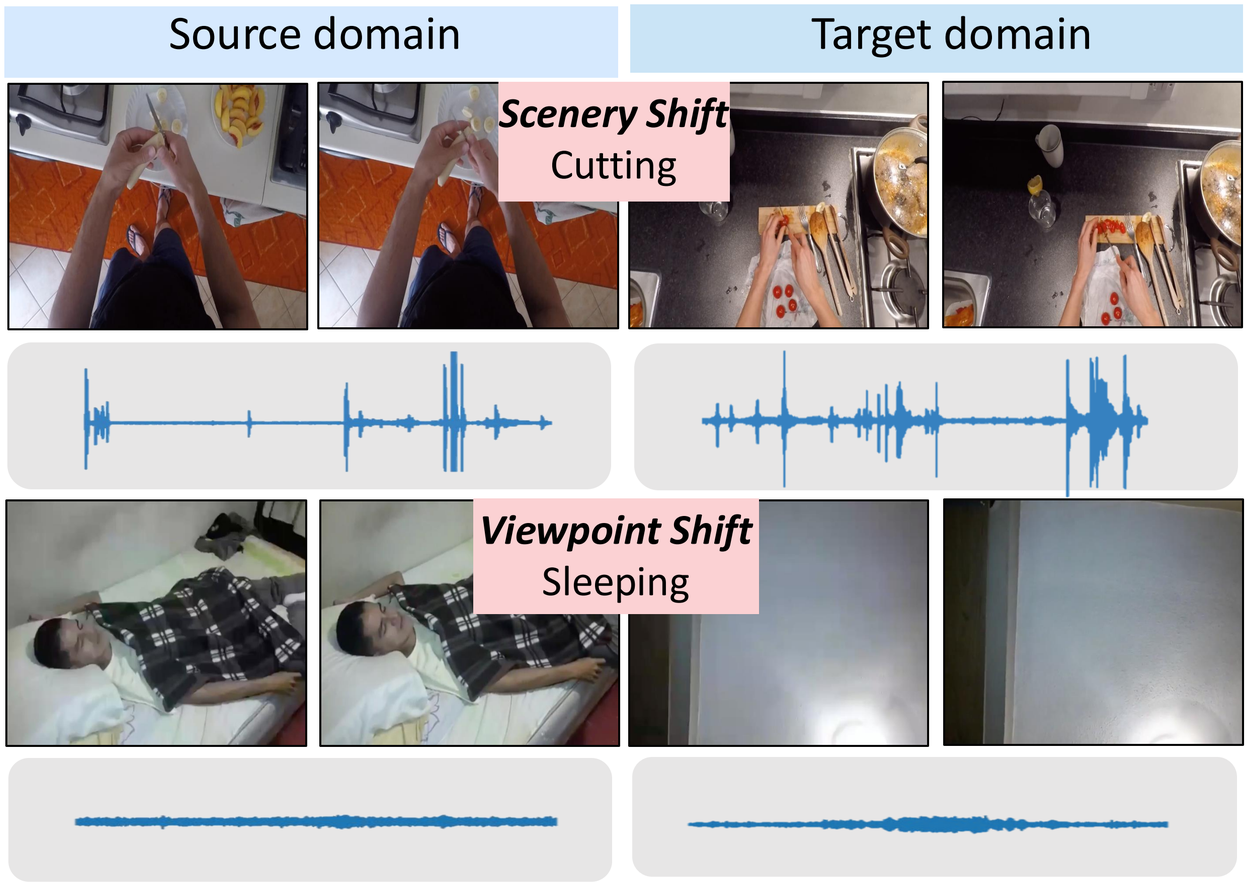
Scenery shift

Camera viewpoint shift

Actor shift

Proposed solution

*We deal with the vision distribution shift with the aid of **activity sounds**.*



⇨ Characteristic sound signals of audible activities

(Playing piano, playing guitar, …)

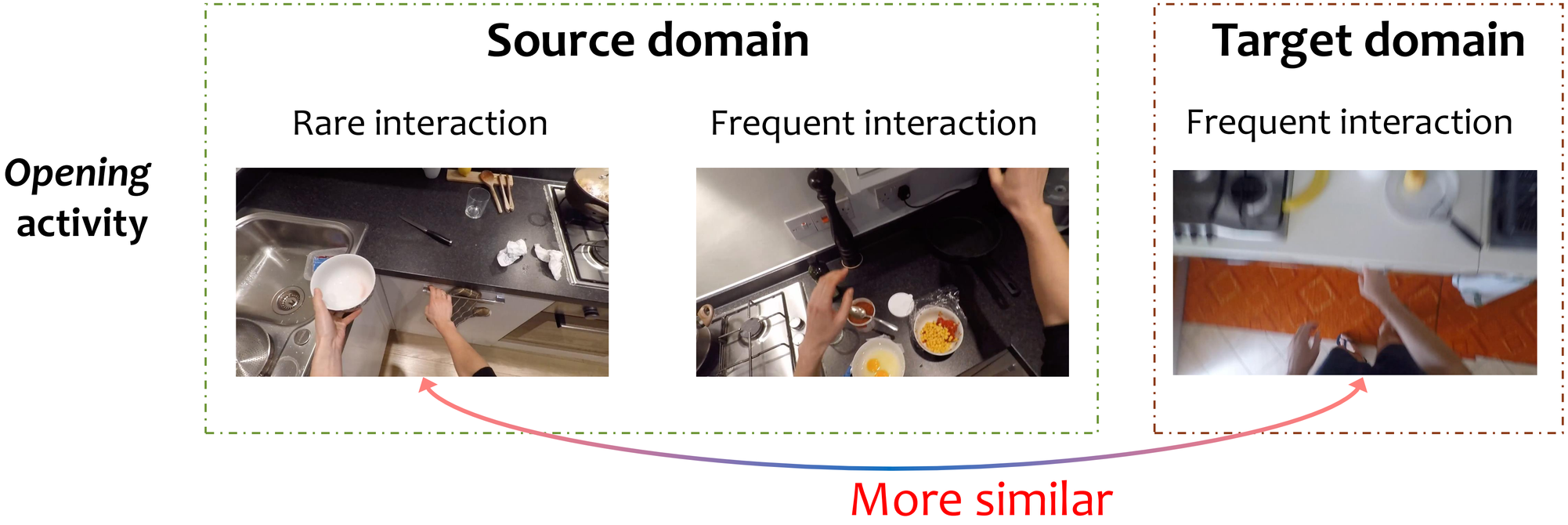
⇨ Environmental sounds of silent activities

Situp (Sounds in the gym), Camping (Outdoor sounds)

Audio-balanced learning

Motivation: videos from **different domains** often have **different label distributions**, not only in terms of activity classes but also their interactions with objects or the environment.

Solution: learn each class and each type of interaction equally



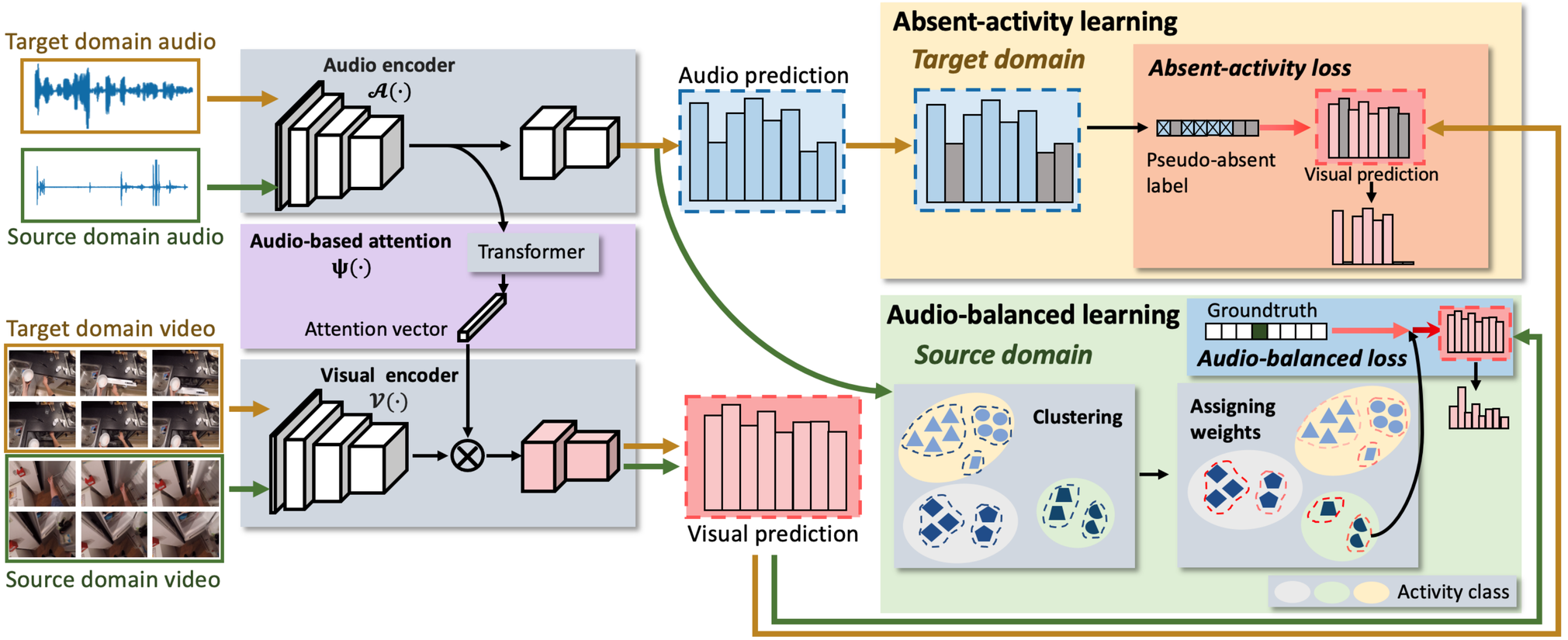
Absent-activity learning

Groundtruth activity:
pour

Absent activities predicted by audio:
wash
close
open

Audio-adaptive approach

Supervised by **audio-balanced learning** and **absent-activity learning**



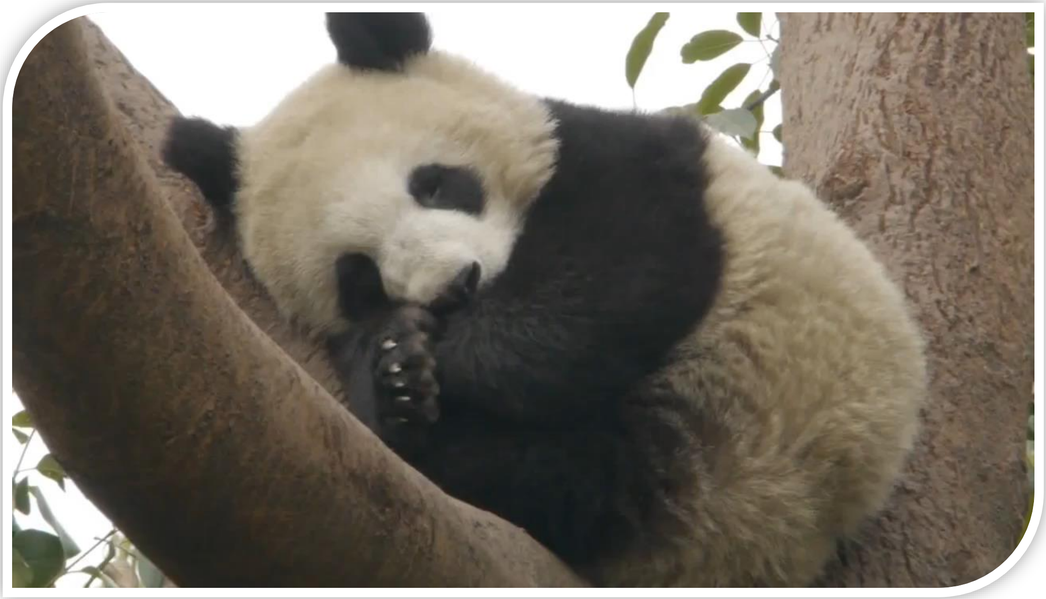
Results

	Scenery shift	Viewpoint shift
Model	**EPIC-Kitchens** Top-1 (%)	**CharadesEgo** mAP (%)
Visual-only	48.0	23.1
Ours (no audio in testing)	50.7	24.5
Ours	59.2	26.3

Actor-shift: success case

Source domain

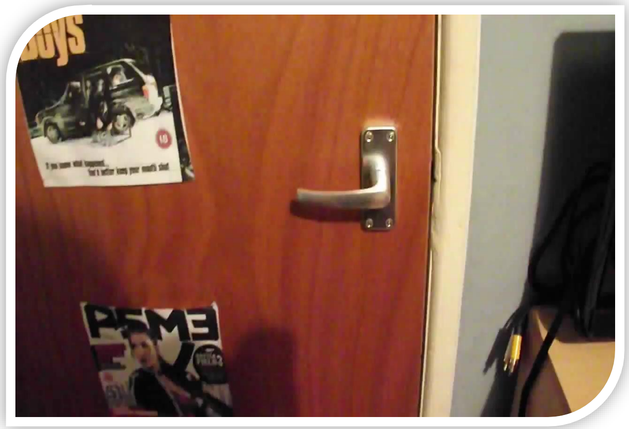
Target domain



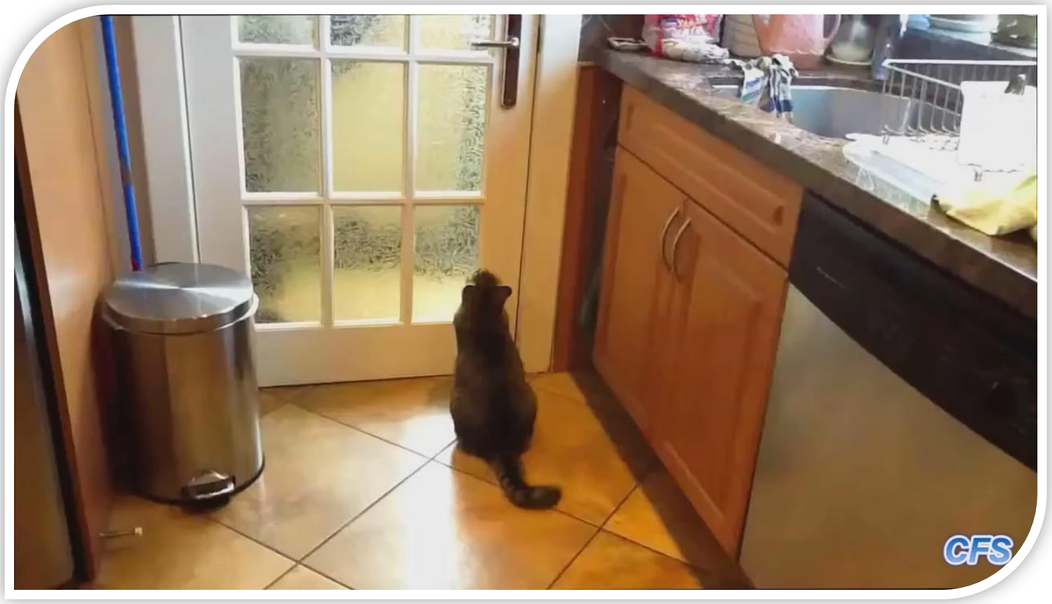
Groundtruth: *sleeping*
Prediction: *sleeping*
Confidence: 0.76

Actor-shift: success case

Source domain



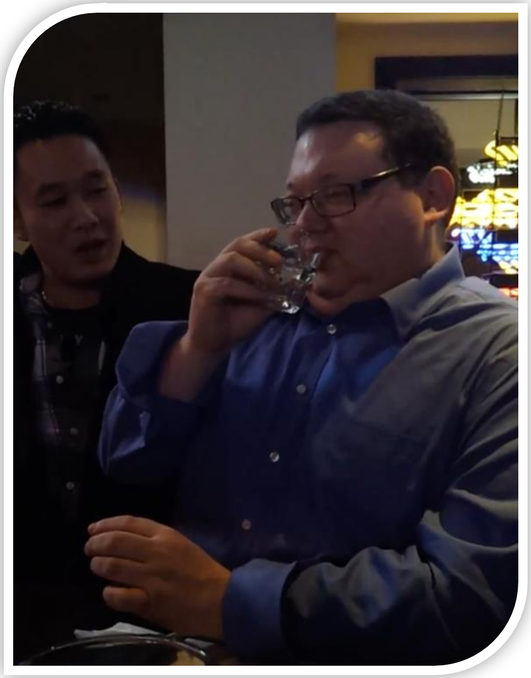
Target domain



Groundtruth: *opening door*
Prediction: *opening door*
Confidence: 0.85

Actor-shift: failure case

Source domain

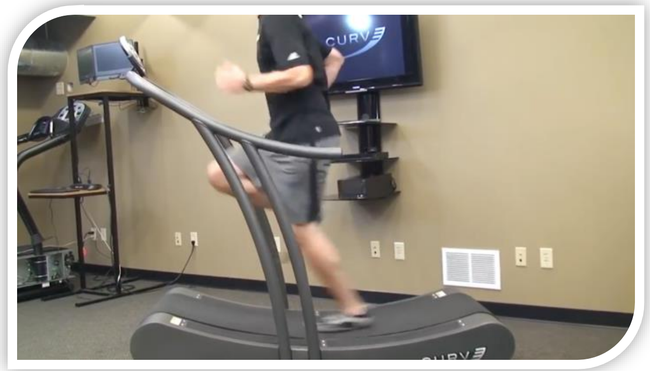


Target domain

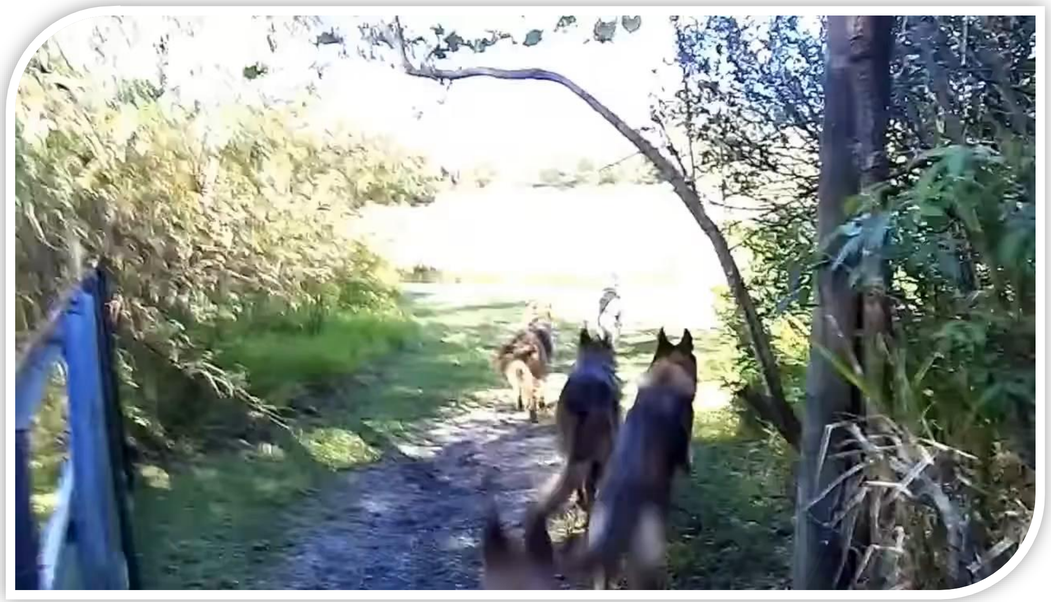
Groundtruth: *drinking*
Prediction: *eating*
Confidence: 0.35

Actor-shift: failure case

Source domain



Target domain



Groundtruth: *running*
Prediction: *swimming*
Confidence: 0.48

Key takeaways

Showed invariant properties of **sound to reduce visual domain gap**.

Better adaptation ability than visual-only solutions

Benefits from audio more than alternative audiovisual fusion methods

Generalize models to new **environments, viewpoints and actors**

3.b Adaptation at night

Yunhua Zhang
University of Amsterdam

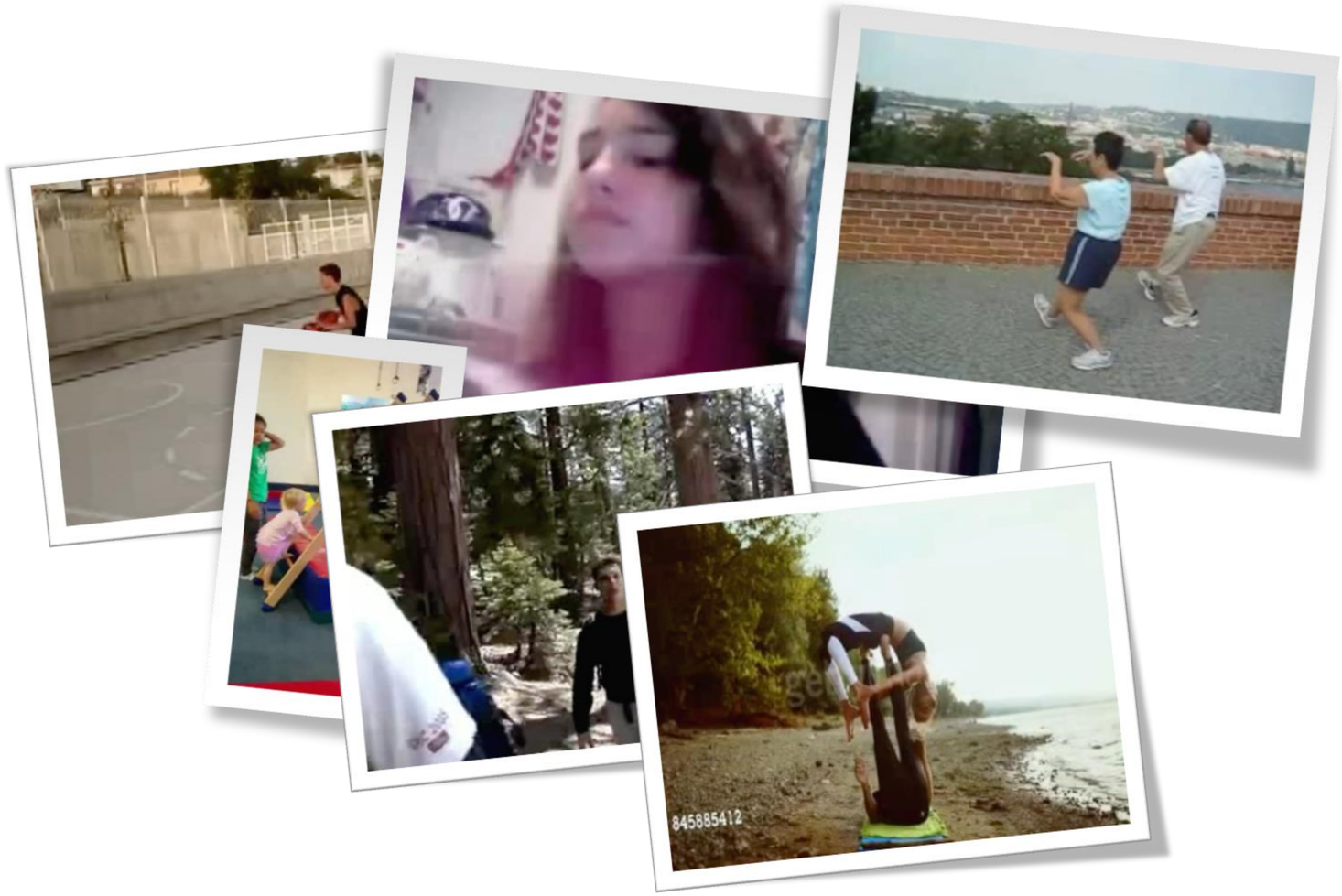
Hazel Doughty
University of Amsterdam

Cees Snoek
University of Amsterdam

Day2Dark: Pseudo-Supervised Activity Recognition beyond Silent Daylight. IJCV 2025.

Video datasets are biased to daylight conditions

Video dataset
EPIC-Kitchens
ActivityNet
Charades
Kinetics-400
Moments-in-Time
Kinetics-Sound

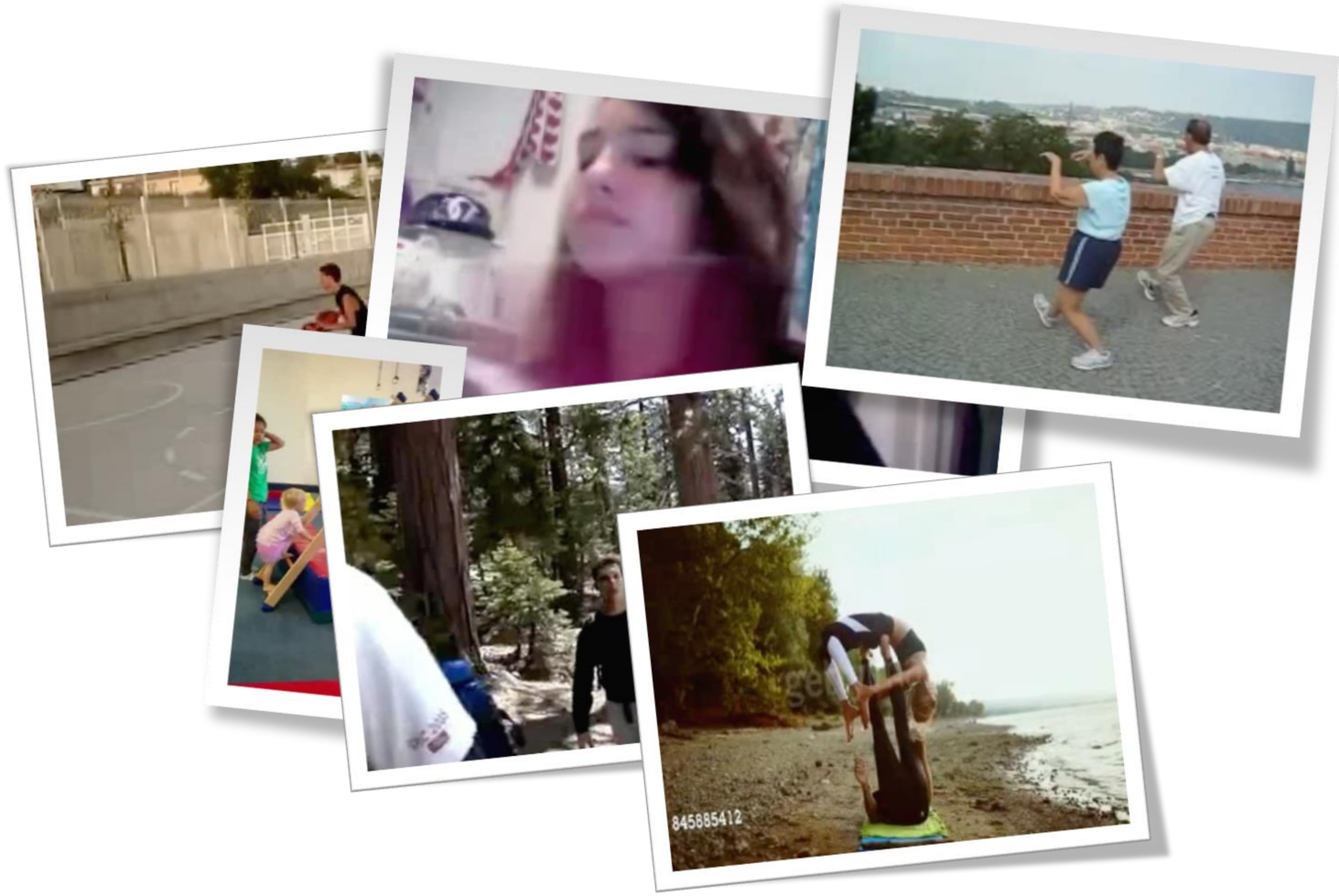


$$Y = \frac{\sum_{j=1}^{H_v \times W_v} (0.299 R_j + 0.587 G_j + 0.144 B_j)}{H_v \times W_v}.$$

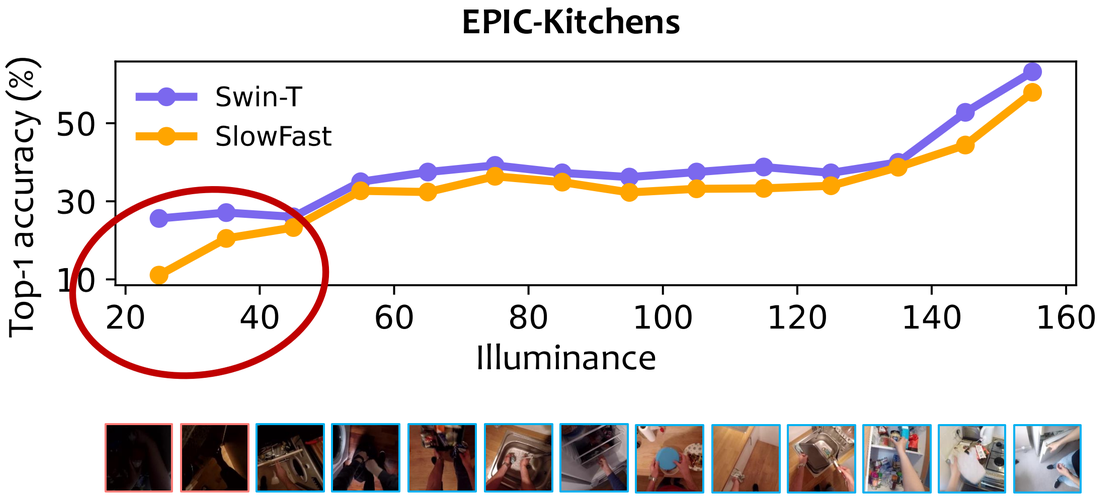
Video datasets are biased to daylight conditions

Video dataset	Dark videos (Y<=40)
EPIC-Kitchens	1.9%
ActivityNet	3.2%
Charades	3.6%
Kinetics-400	4.4%
Moments-in-Time	4.9%
Kinetics-Sound	8.3%

$$Y = \frac{\sum_{j=1}^{H_v \times W_v} (0.299R_j + 0.587G_j + 0.144B_j)}{H_v \times W_v}.$$

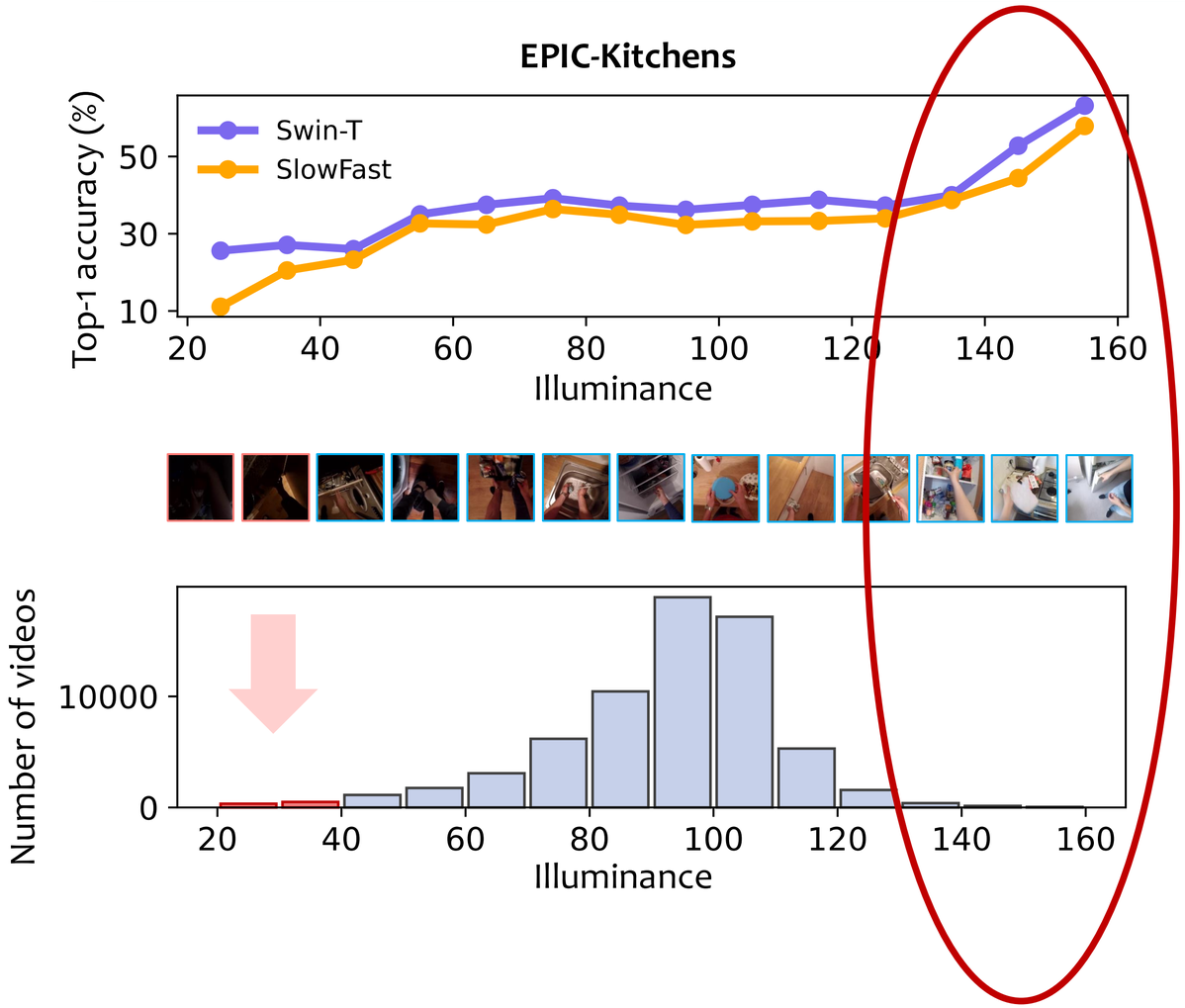


Problem statement: Day2Dark gap



Activity recognition models suffer from **performance drops** in low-illumination.

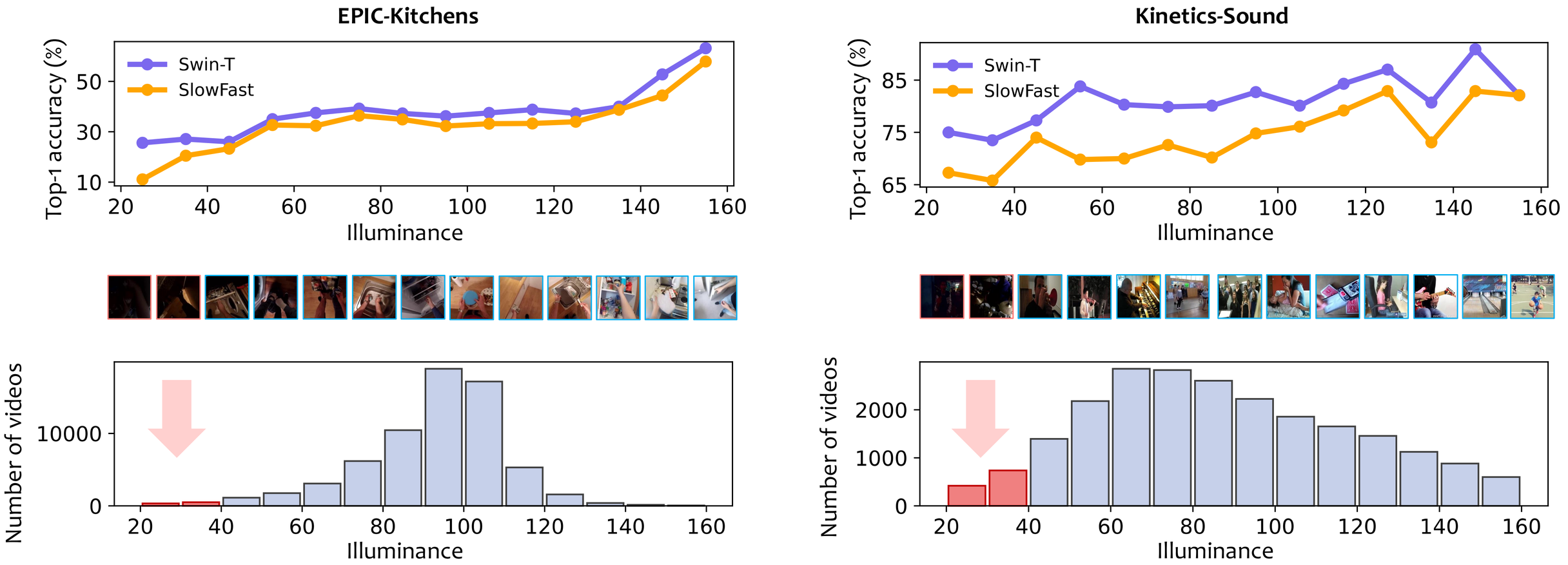
Problem statement: Day2Dark gap



Activity recognition models suffer from **performance drops** in low-illumination.

Caused by lack of **training data** and **distribution shift** by lower color contrast

Problem statement: Day2Dark gap



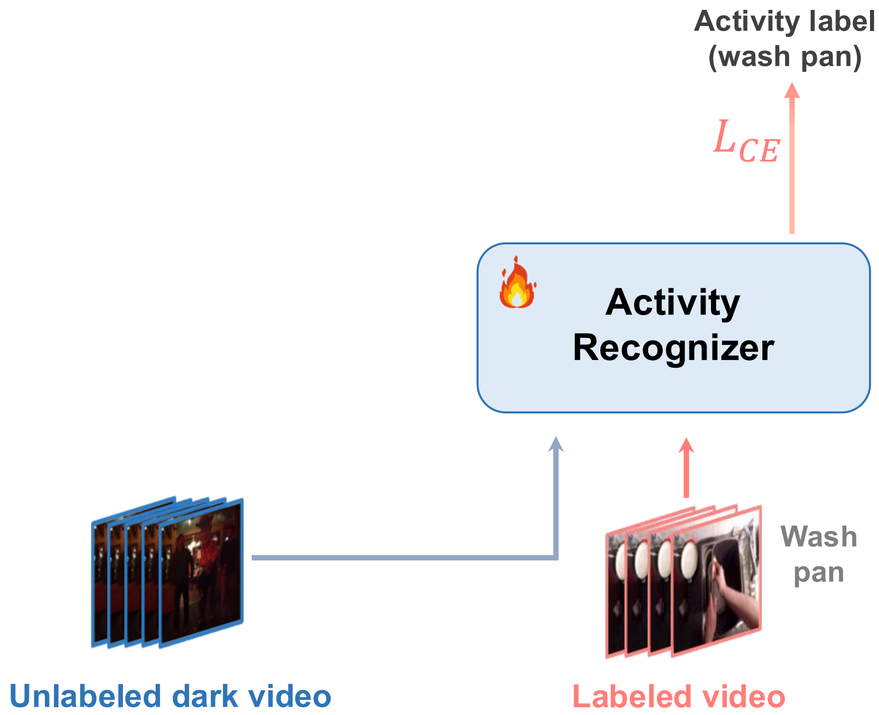
Technical contributions

I. A **pseudo-supervised learning** strategy that utilizes **unlabeled dark videos**, which do not contain target activities.

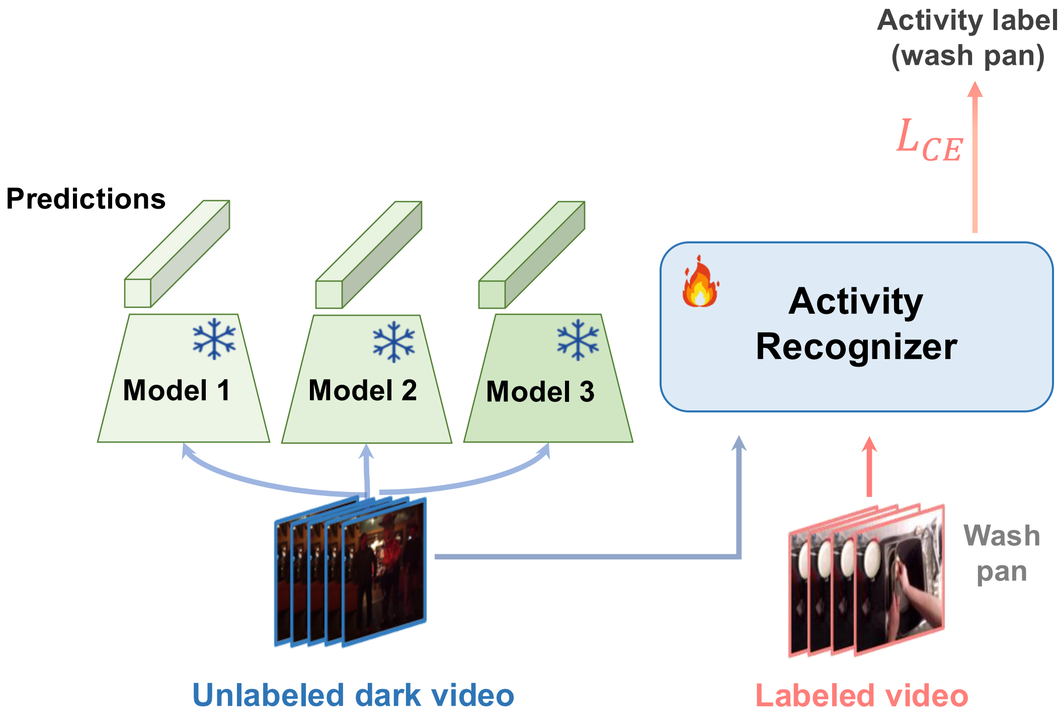
II. **Darkness-aware** audio-visual recognition to **reduce the distribution** shift and find better **cross-modal correspondences** in the dark.

Unlabeled dark video examples

I. Supervision beyond daylight



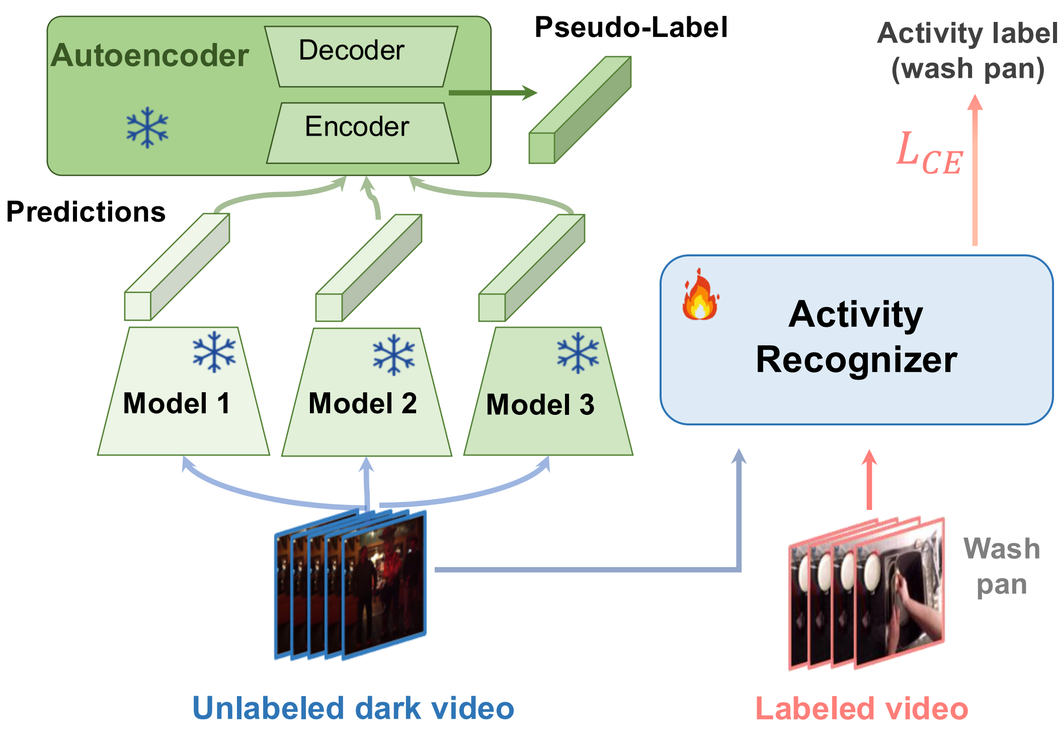
I. Supervision beyond daylight



Generate **pseudo-labels** by auxiliary models

e.g. Video-text retrieval, sound source localization *etc.*

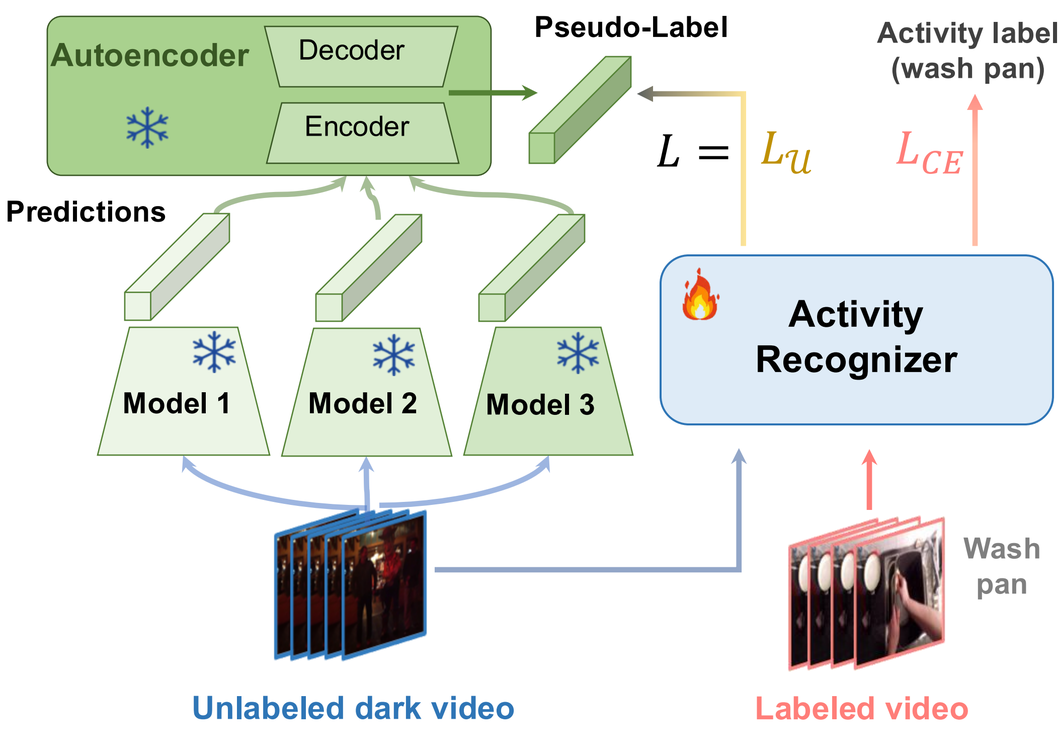
I. Supervision beyond daylight



Generate **pseudo-labels** by auxiliary models
e.g. Video-text retrieval, sound source localization *etc.*

Autoencode predictions into **latent** pseudo-label

I. Supervision beyond daylight



Generate **pseudo-labels** by auxiliary models
e.g. Video-text retrieval, sound source localization *etc.*

Autoencode predictions into **latent** pseudo-label

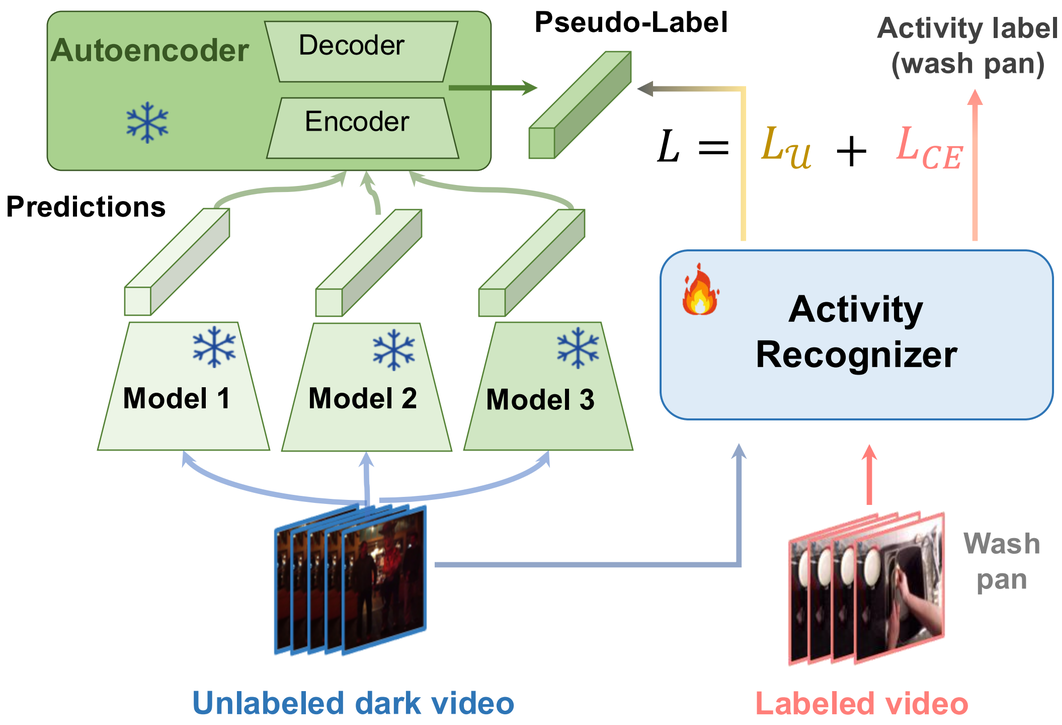
Single **distance** function as the loss

$$L_{\mathcal{U}} = \sum_{j=1}^{U} dist(\hat{q}^j, q^j)$$

Model output

Pseudo-label

I. Supervision beyond daylight



Generate **pseudo-labels** by auxiliary models

e.g. Video-text retrieval, sound source localization *etc.*

Autoencode predictions into **latent** pseudo-label

Single **distance** function as the loss

$$L_{\mathcal{U}} = \sum_{j=1}^{U} dist(\hat{q}^j, q^j)$$

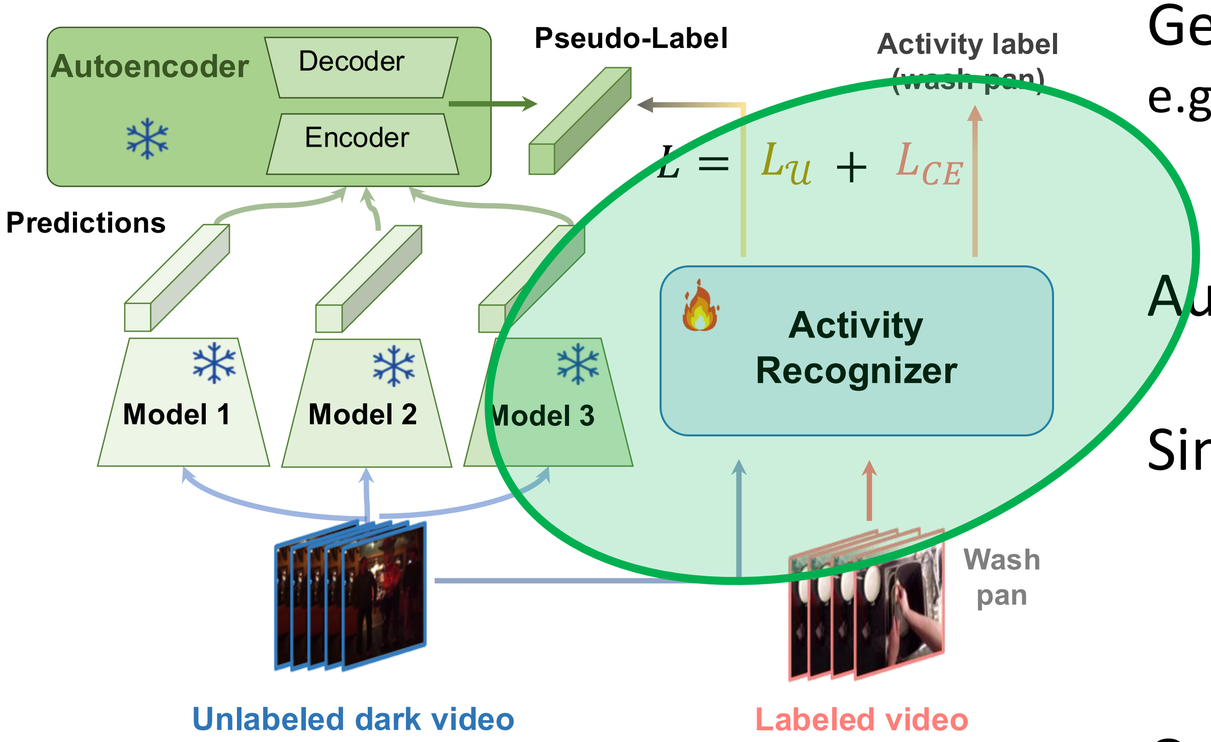
Model output

Pseudo-label

Overall training **objective:**

$$L = L_{CE} + \lambda L_U$$

I. Supervision beyond daylight



Generate **pseudo-labels** by auxiliary models
e.g. Video-text retrieval, sound source localization *etc.*

Autoencode predictions into **latent** pseudo-label

Single **distance** function as the loss

$$L_U = \sum_{j=1}^{U} dist(\hat{q}^j, q^j)$$

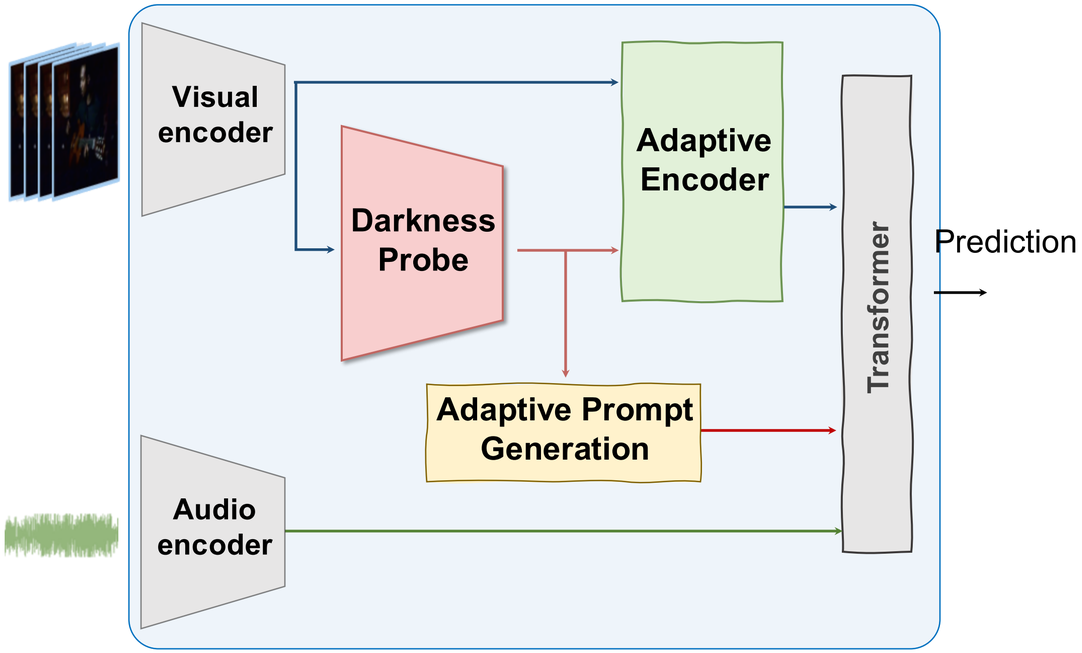
Model output

Pseudo-label

Overall training **objective:**

$$L = L_{CE} + \lambda L_U$$

II. Darkness-aware audio-visual recognition



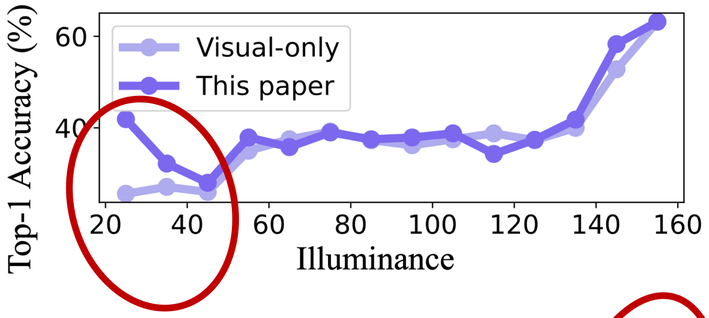
Darkness probe produces an *n*-way branch attention to adapt to the current light condition

Adaptive encoder encodes the visual features according to perceived darkness

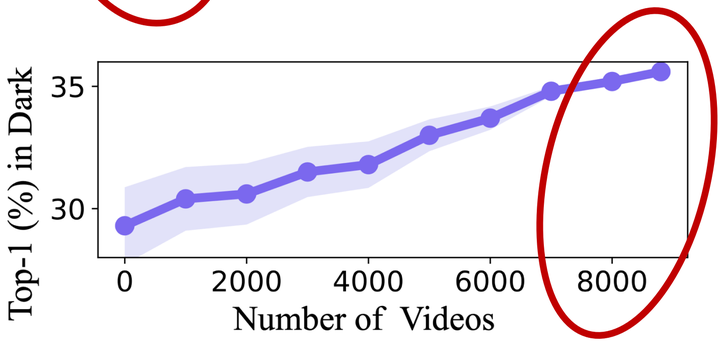
Adaptive prompt generation treats different light conditions as different tasks

Transformer fuses adapted visual features, prompts and audio features

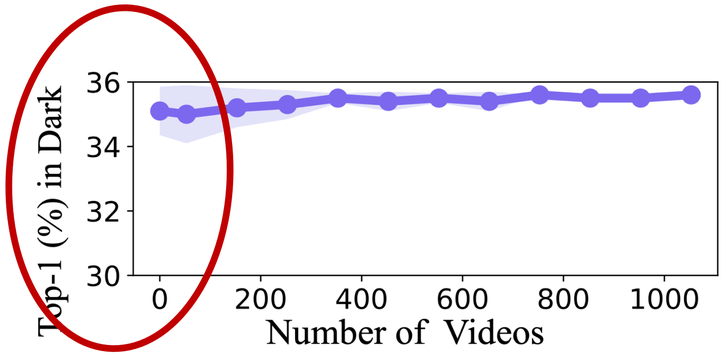
Properties of our proposal



Largest improvement for lowest illuminance.



The more unlabeled dark videos the better.



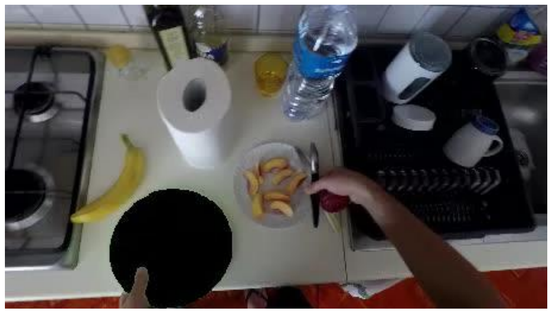
Even successful without labeled dark videos.

Bonus: also effective for occlusions

Tested on 182 EPIC-Kitchens videos with segmentation masks from Darkhalil et al.

We simulate occlusions by setting the pixel intensity of object regions to zero.





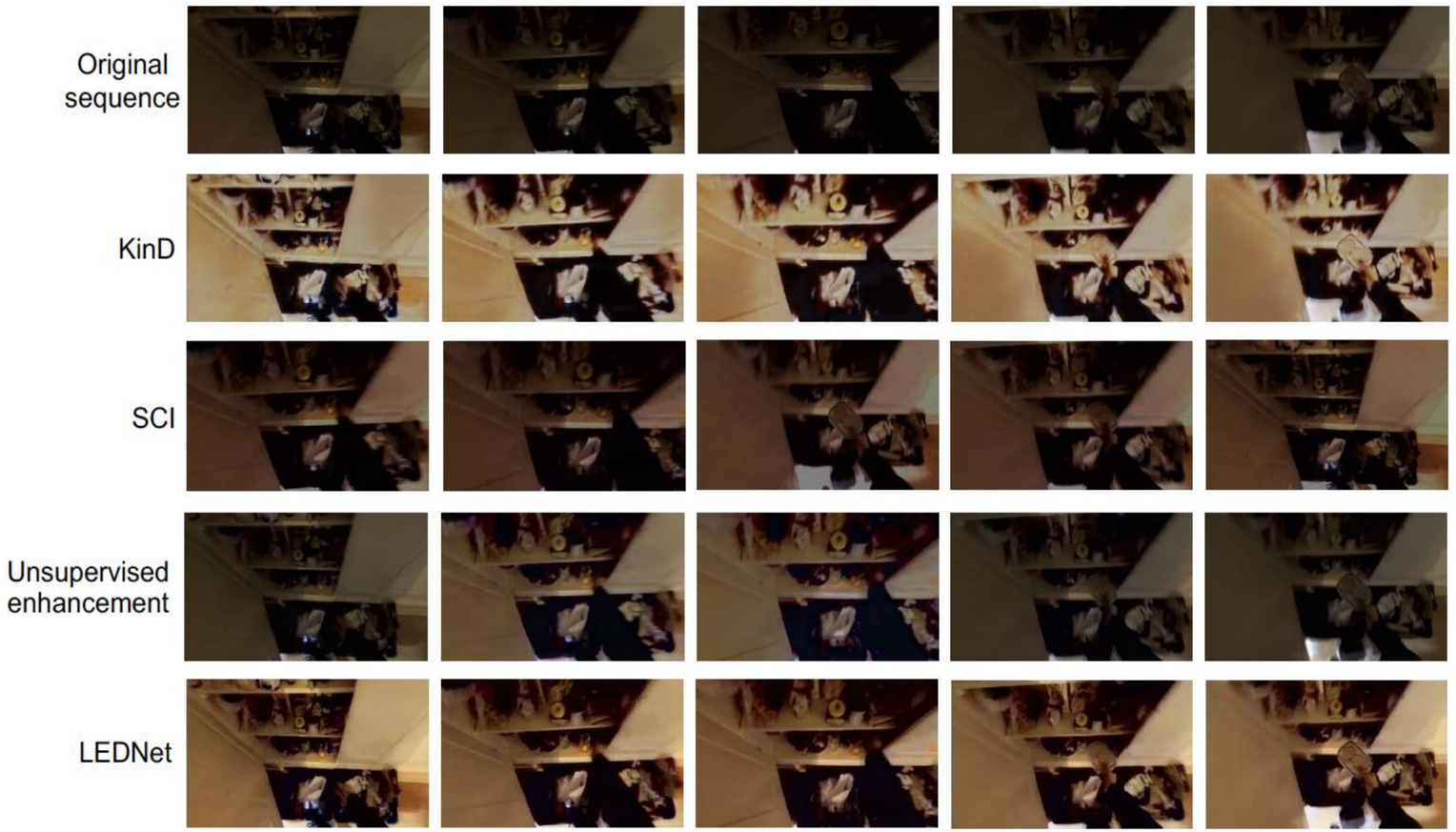

Visual encoder	26.4%
Vanilla multi-modal transformer	27.7%
This paper	***29.8%***

Comparison with image enhancement

Model	Venues	EPIC-Kitchens	
		Dark↑	GFLOPs↓
Vanilla multi-modal transformer		29.8	1.4
KinD	MM 2019	20.3	932.2
SCI	CVPR 2022	24.1	3.4
Unsupervised enhancement	ECCV 2022	26.4	108.8
LEDNet	ECCV 2022	27.8	312.0
Retinexformer	ICCV 2023	28.2	15.6
This paper		**35.6**	**1.6**

We are superior to image enhancement for both accuracy and computation time.

Qualitative result for ‘take box’



Illumination for dark frames improve, but color distortions harm activity recognition.

Failure case for ‘pick up knife’

Vanilla audio-visual transformer

Verb prediction : ***put*** ✗
Noun prediction: ***leek*** ✗

This paper

Verb prediction: ***put*** ✗
Noun prediction: ***leek*** ✗

The right hand draws more attention than the left hand

Failure case: ‘slapping’

Vanilla audio-visual transformer

prediction: ***laughing*** ✗
Confidence: 1.0

This paper

Prediction; ***laughing*** ✗
Confidence: 0.86

The environmental sound distracts the model

Key takeaways

Day2dark gap is wide-spread for multiple action recognition datasets and backbones.

Unlabeled dark videos and adaptively including **sound reduces the gap**.

Proposed model **outperforms image enhancement and alternative fusion** approaches.

3.c Generalize over unseen modality combo's

Yunhua Zhang
University of Amsterdam

Hazel Doughty
University of Amsterdam

Cees Snoek
University of Amsterdam

Learning Unseen Modality Interaction. In *NeurIPS* 2023.

Problem: Modality-complete assumption

Training

Device 1
RGB & Audio

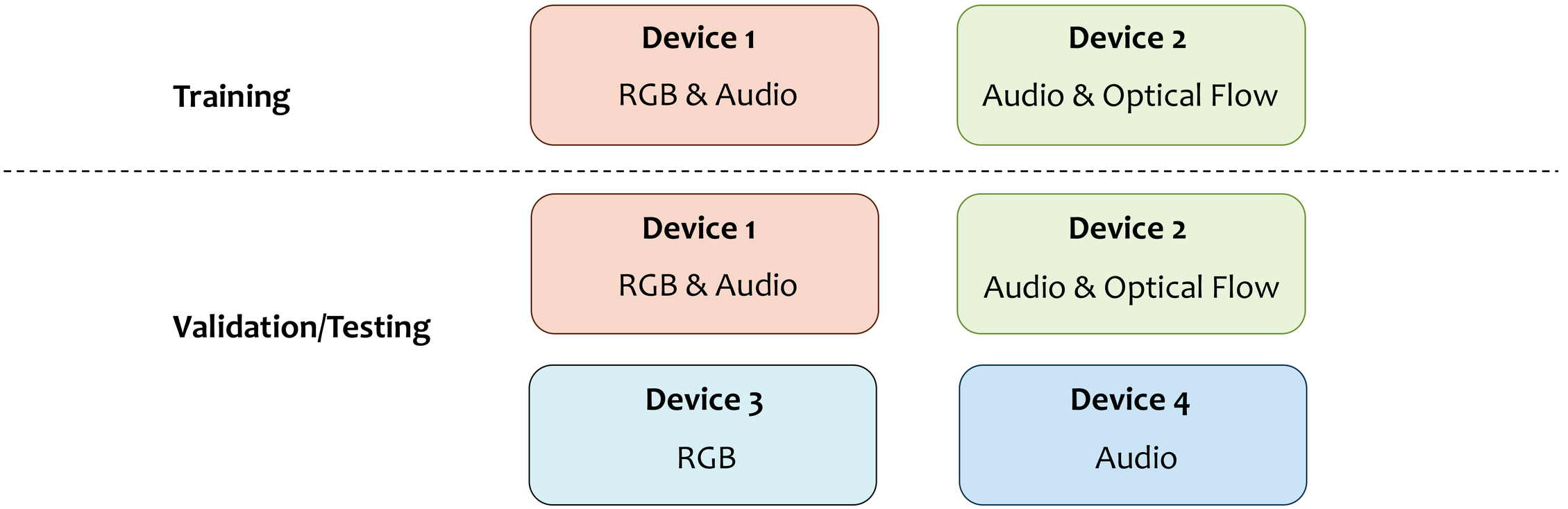
Device 2
Audio & Optical Flow

Validation/Testing

Device 1
RGB & Audio

Device 2
Audio & Optical Flow

Others: Robustness for modality-incomplete data



One or more modalities could be missing during inference

Antoine Miech, et al. "Learning a text-video embedding from incomplete and heterogeneous data." In arXiv preprint 2018.
Mengmeng Ma, et al. "Smil: Multimodal learning with severely missing modality." In AAAI 2021.
Nina Shvetsova, et al. "Everything at once-multi-modal fusion transformer for video retrieval." In CVPR 2022.

Our goal: Recognize unseen modality-interactions

Training

Device 1
RGB & Audio

Device 2
Audio & Optical Flow

Validation/Testing

Device 1
RGB & Audio

Device 2
Audio & Optical Flow

Device 3
RGB & Optical Flow

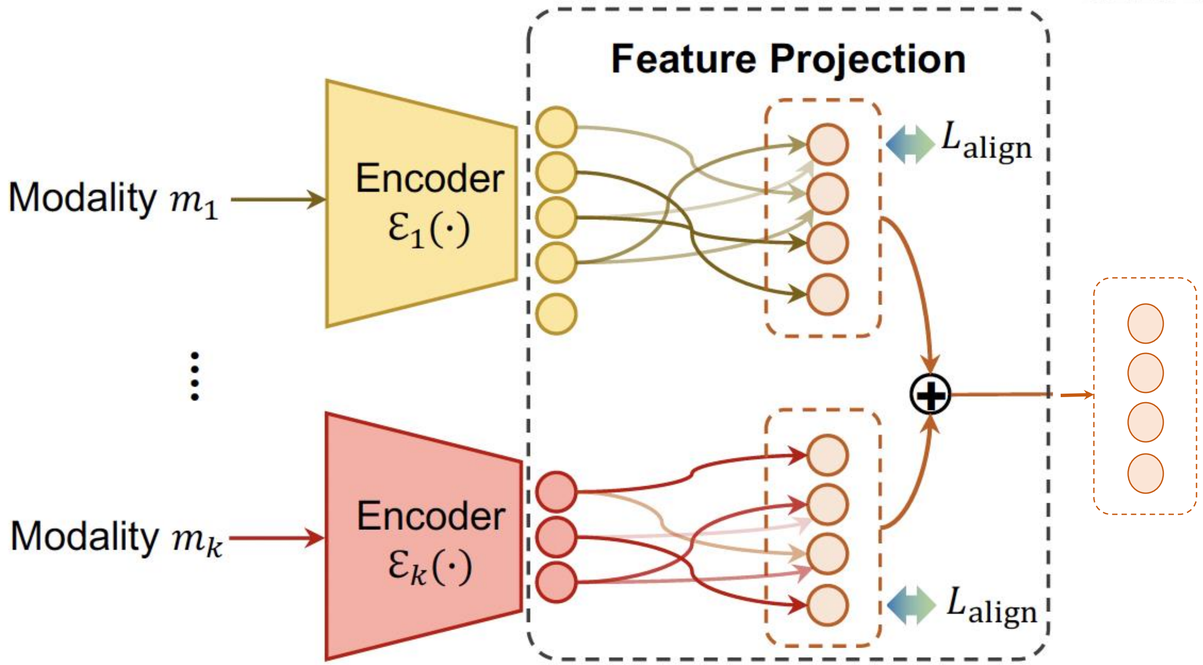
Challenges

Simple concatenation of unimodal features cannot learn cross-modal correspondences when modality-complete data unavailable.

The accumulation should be agnostic for the order of modalities, and also allow for any modality combinations

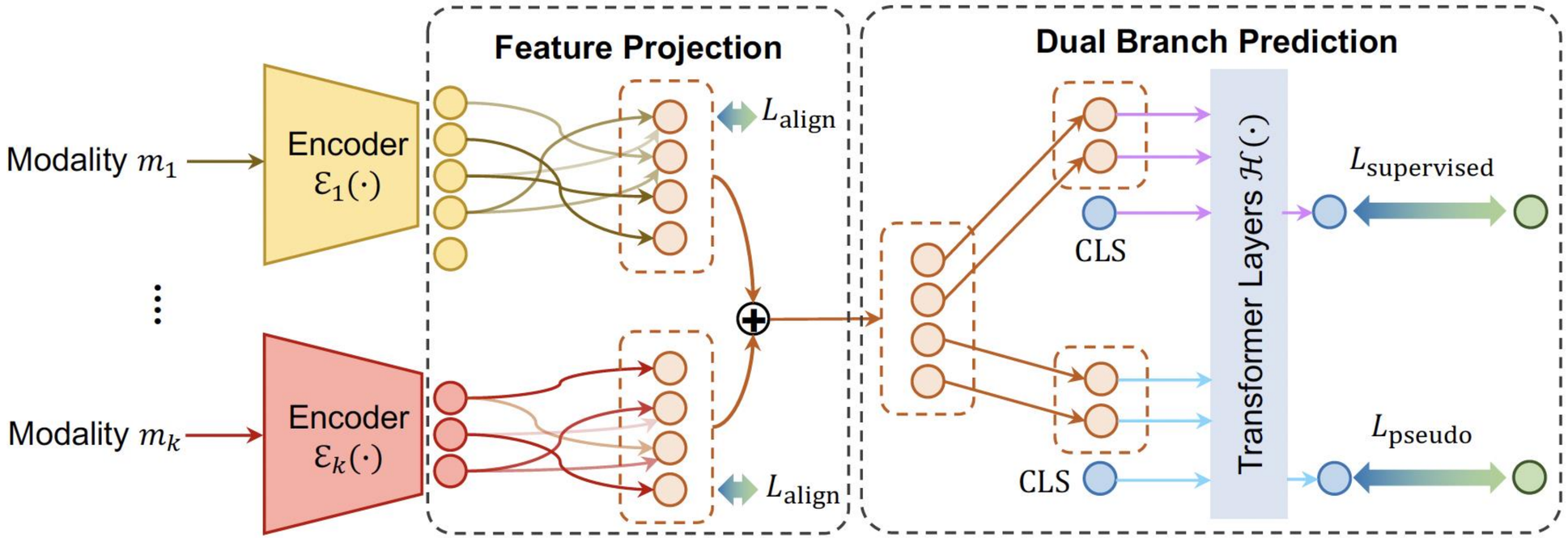
Simple addition of unimodal features is hard as modalities come in different feature spaces and dimensionalities

Approach



Project modality-specific features into a common space while maintaining differentiating information.

Approach

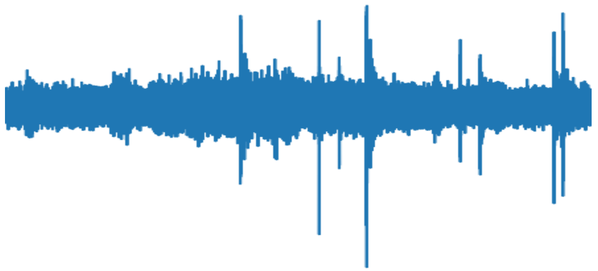


Some modalities are less discriminative than others and cause overfitting.
To mitigate overfitting, we introduce an additional pseudo-supervised branch.

Evaluation

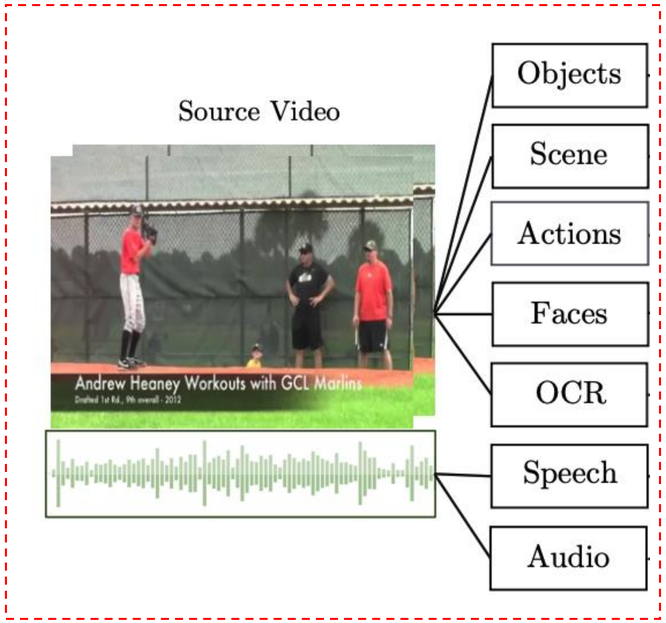
Video Classification

Audio



EPIC Kitchens with 3 provided modalities

Video Retrieval



Matching

Andrew Heaney plays baseball in a field with some other players.

MSR-VTT with 7 provided modalities

Note we define new splits to assure unseen modality interactions at test time

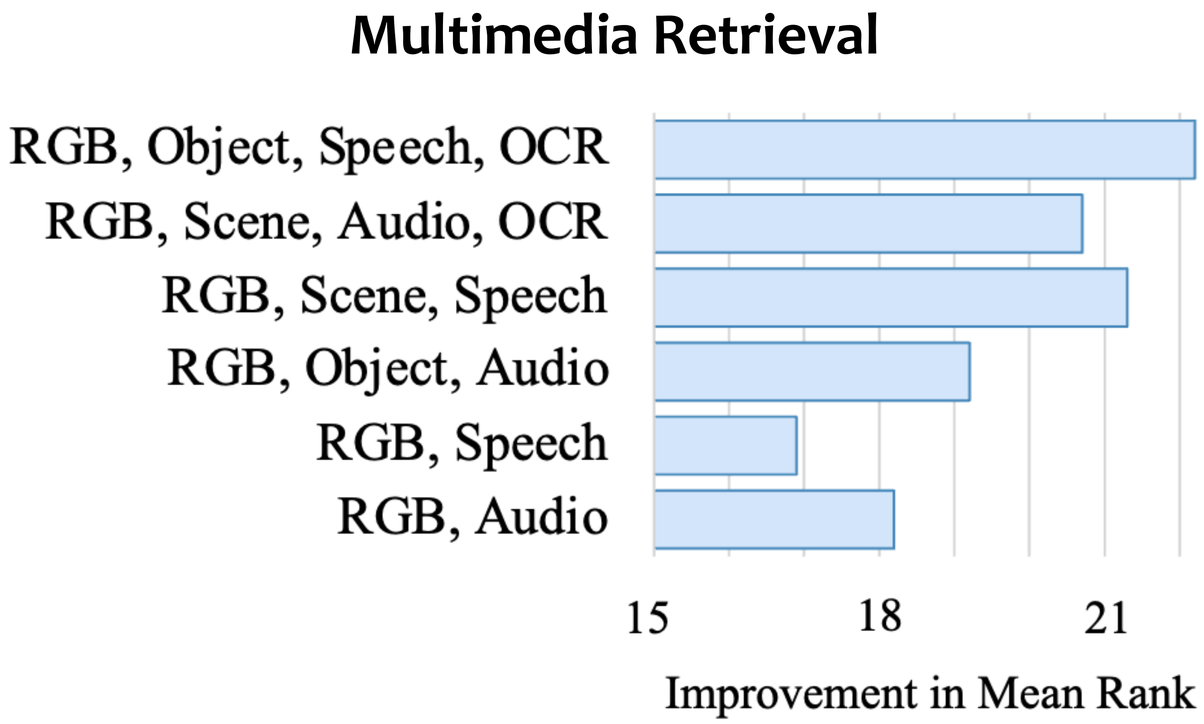
Results

	Video Classification	Video Retrieval
	Top-1 (%) ↑	MnR ↓
Late fusion	18.1	72.3
Modality Complete (Nagrani et al.)	17.5	86.2
Modality Incomplete (Recasens et al.)	18.5	72.2
Ours: *unseen modality interaction*	**23.7**	**66.2**

Without the need for modality-complete data, our method learns a more effective cross-modal fusion for unseen modality combinations

Robustness for modality incomplete at test-time

We show the improvement over a vanilla multimodal transformer.



Our model can handle any input modality

Improves robustness for all unseen combo's

Our model most effective for more modalities

Key takeaways

We can effectively make **predictions for unseen modality interactions** by feature projections and pseudo-supervision

Our approach is **suitable for classification, regression and retrieval**, and can handle a wide variety of modality combinations

Concluding encouragement

Learning to generalize in video space and time, and across modalities and tasks, is an **open research challenge**.

First ideas have started to appear, **much more research is needed**.

VIS LAB
VIDEO & IMAGE SENSE LAB

Prof. dr. Cees Snoek

https://ivi.fnwi.uva.nl/vislab/

@cgmsnoek {x, ellis.social}

Key references

- Fida Mohammad Thoker, Hazel Doughty, Piyush Bagad, Cees G M Snoek: **How Severe is Benchmark-Sensitivity in Video Self-Supervised Learning?.** In: ECCV, 2022.
- Fida Mohammad Thoker, Hazel Doughty, Cees G M Snoek: **Tubelet-Contrastive Self-Supervision for Video-Efficient Generalization.** In: ICCV, 2023.
- Mohammadreza Salehi, Michael Dorkenwald, Fida Mohammad Thoker, Efstratios Gavves, Cees G M Snoek, Yuki M Asano: **SIGMA: Sinkhorn-Guided Masked Video Modeling**. In: ECCV, 2024.
- Piyush Bagad, Makarand Tapaswi, Cees G M Snoek: **Test of Time: Instilling Video-Language Models with a Sense of Time**. In: CVPR, 2023.
- Yunhua Zhang, Hazel Doughty, Ling Shao, Cees G M Snoek: **Audio-Adaptive Activity Recognition Across Video Domains**. In: CVPR, 2022.
- Yunhua Zhang, Hazel Doughty, Cees G M Snoek: **Day2Dark: Pseudo-Supervised Activity Recognition beyond Silent Daylight.** IJCV 2025.
- Yunhua Zhang, Hazel Doughty, Cees G M Snoek: **Learning Unseen Modality Interaction**. In: NeurIPS 2023.
- Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G M Snoek, Yuki M Asano: **Lost in Time: A New Temporal Benchmark for VideoLLMs.** In: BMVC, 2025.
- Aritra Bhowmik, Denis Korzhenkov, Cees G M Snoek, Amirhossein Habibian, Mohsen Ghafoorian: **MoAlign: Motion-Centric Representation Alignment for Video Diffusion Models**. arXiv:2510.19022, 2025.