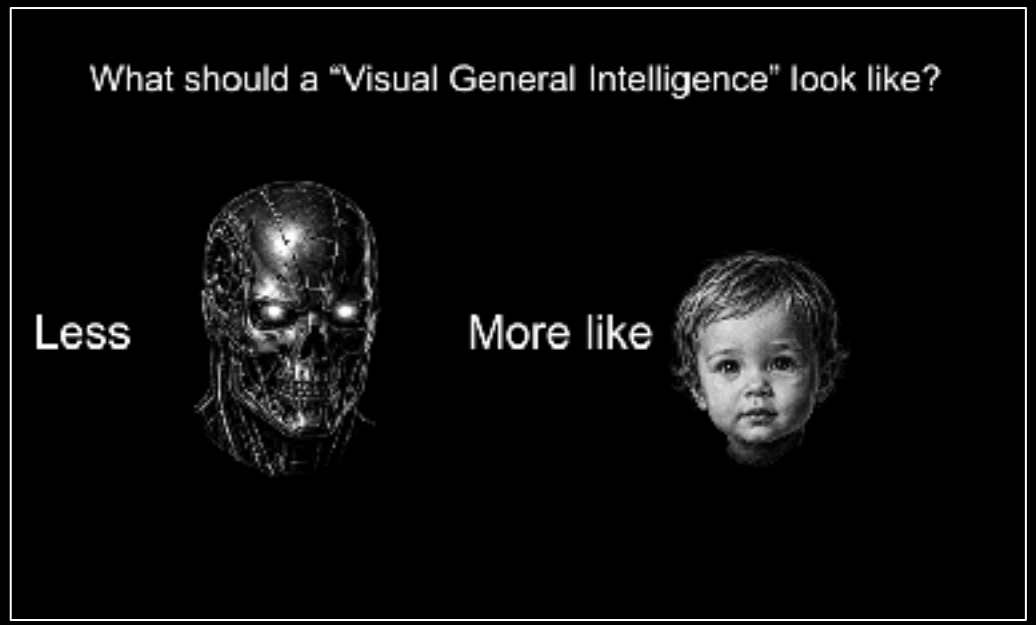
Yuki M. Asano - Fundamental AI Lab

What should a “Visual General Intelligence” look like?

Less

My grandma’s definition:
You’re intelligent if you have a good life, otherwise you’re stupid.



💡: We should view intelligence as metric of a (learning) process

Intelligence as a metric of a learning process: key factors

Performance on tasks

Amount of supervision/feedback

Ability to generalise/adapt

Intelligence as a metric of a learning process: key factors

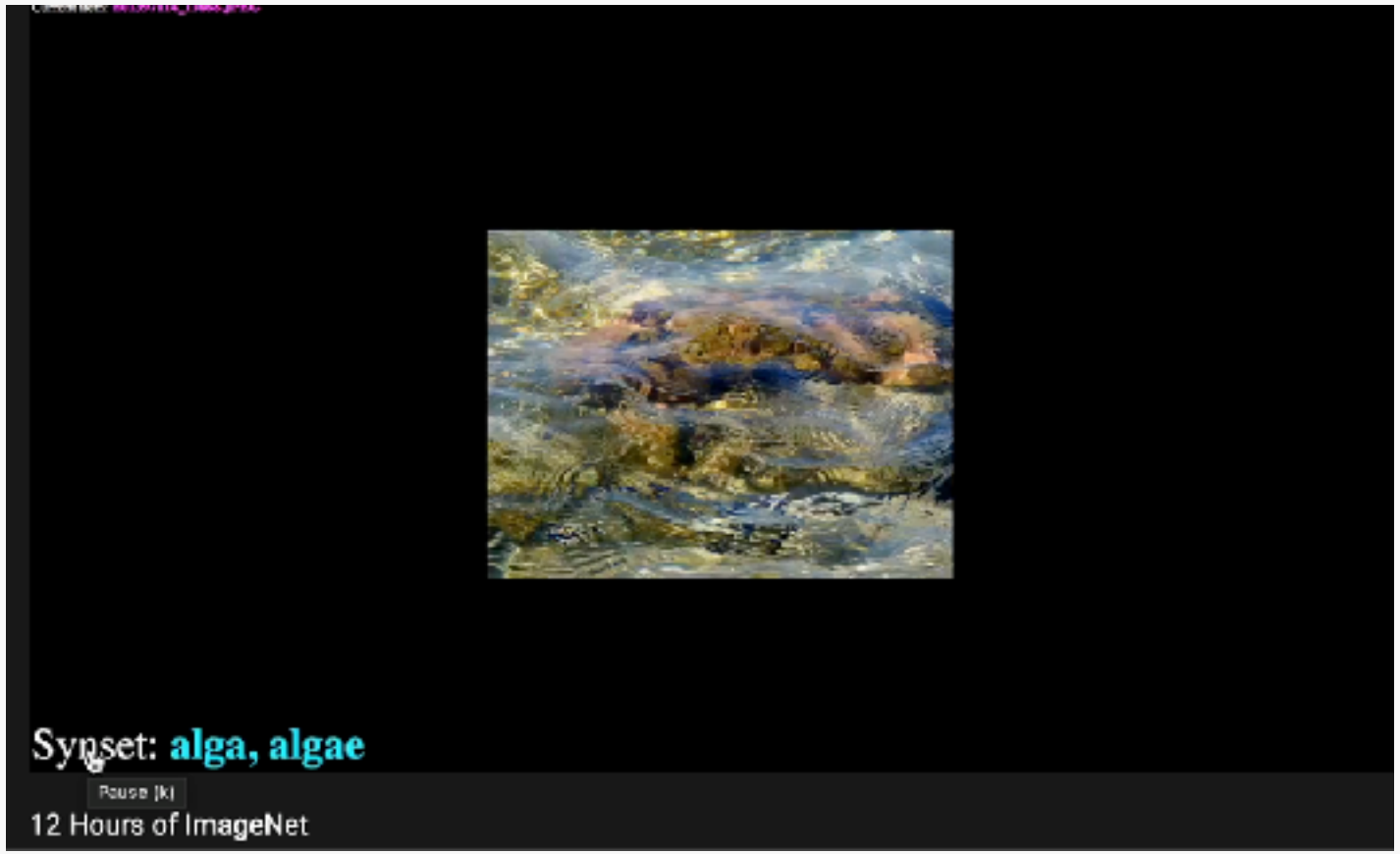
Performance on tasks

Amount of supervision/feedback

Ability to generalise/adapt

Self-Supervised Learning

Don't give this to your babies



Visual continuity linked to quality of vision [1] (invariant object representations wrt. to position)

[1] Unsupervised Natural Experience Rapidly Alters Invariant Object Representation in Visual Cortex. Li and DiCarlo. Science 2008.

Current Vision Foundation Models are trained with images. Videos can enable new directions

Visual development for AI

"Get" physics

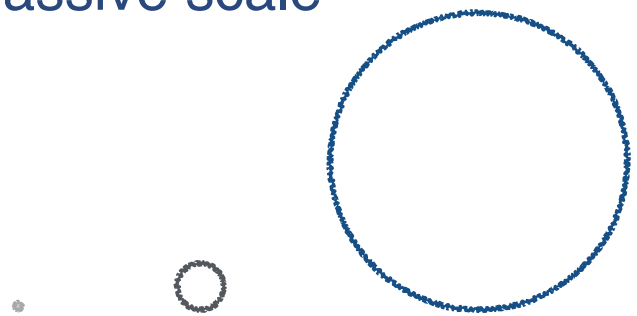
Embodied AI

+ their *insane* scale:


I work at the state…
monitoring the internet traffic.
>90% of it is video-content.

Self-supervised Learning has benefits besides scalability

Massive scale



sup. << weak sup. << raw

No cost of relabelling

No language bias

Fundamentals

So we should work on unlabelled video,

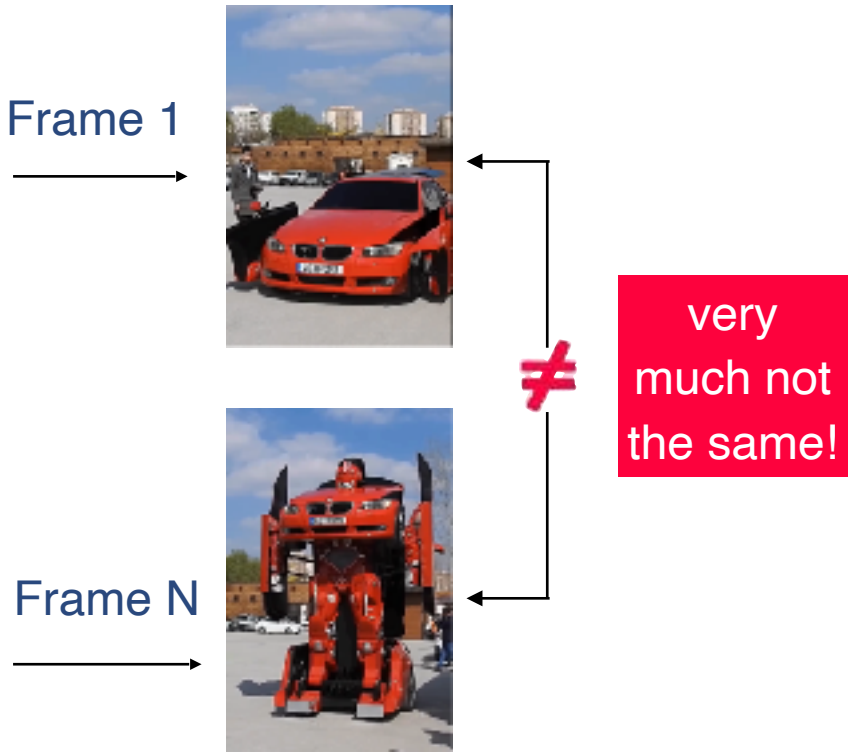
..but *how?*

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

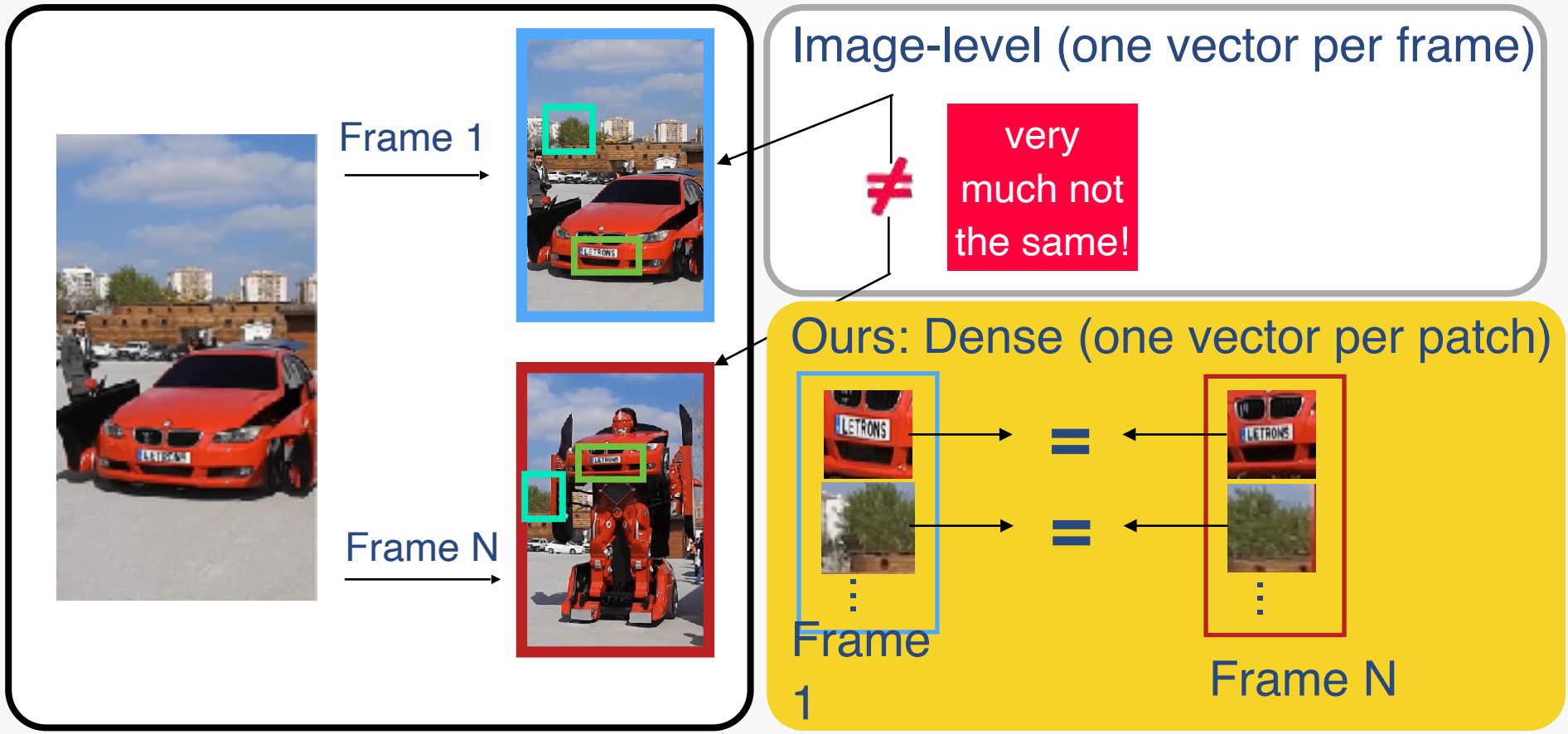
Images: SimCLR, MoCo, SwaAV et al.

key principle: view-invariance

But does this generally make sense?

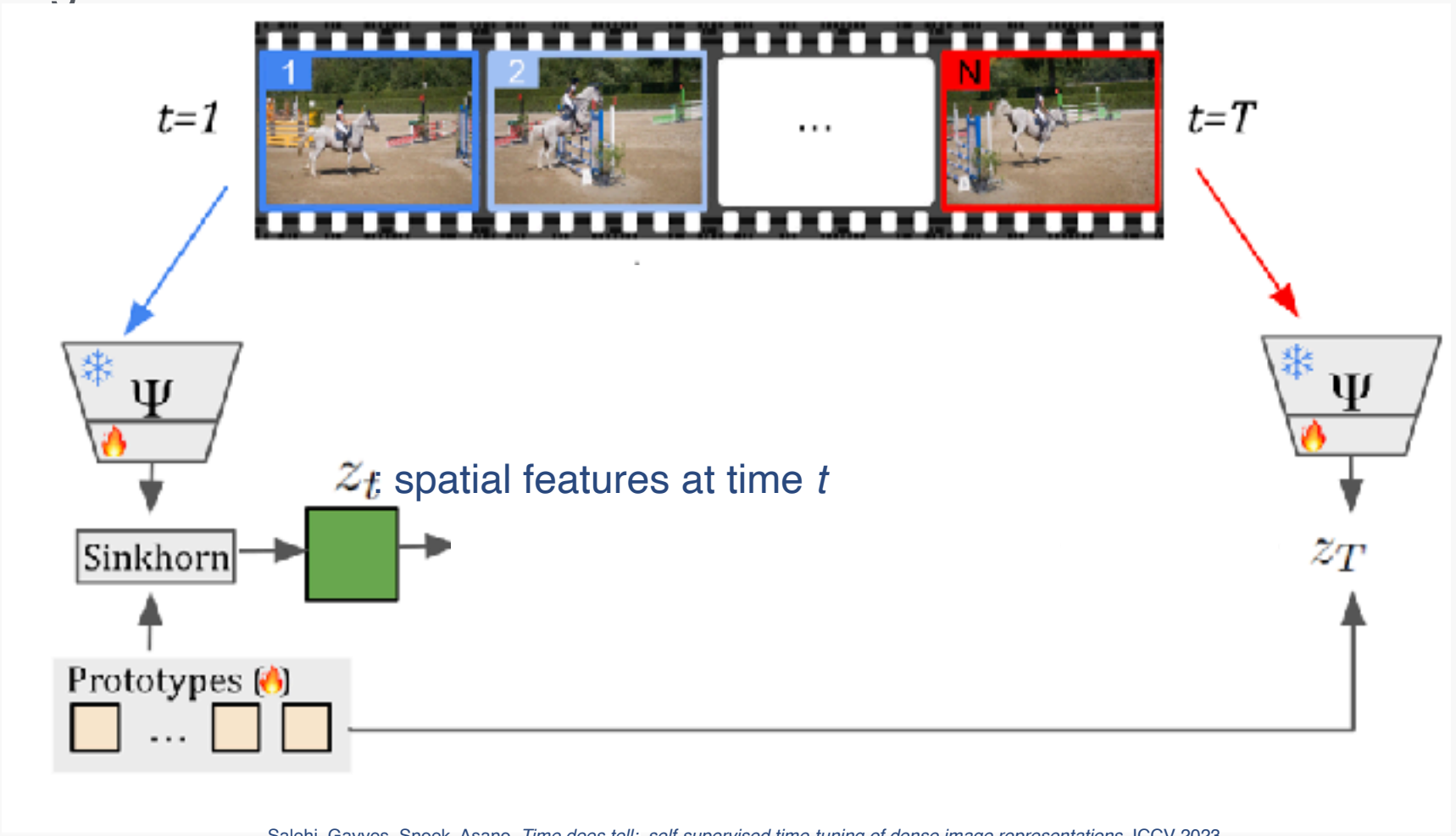


Solution is obvious



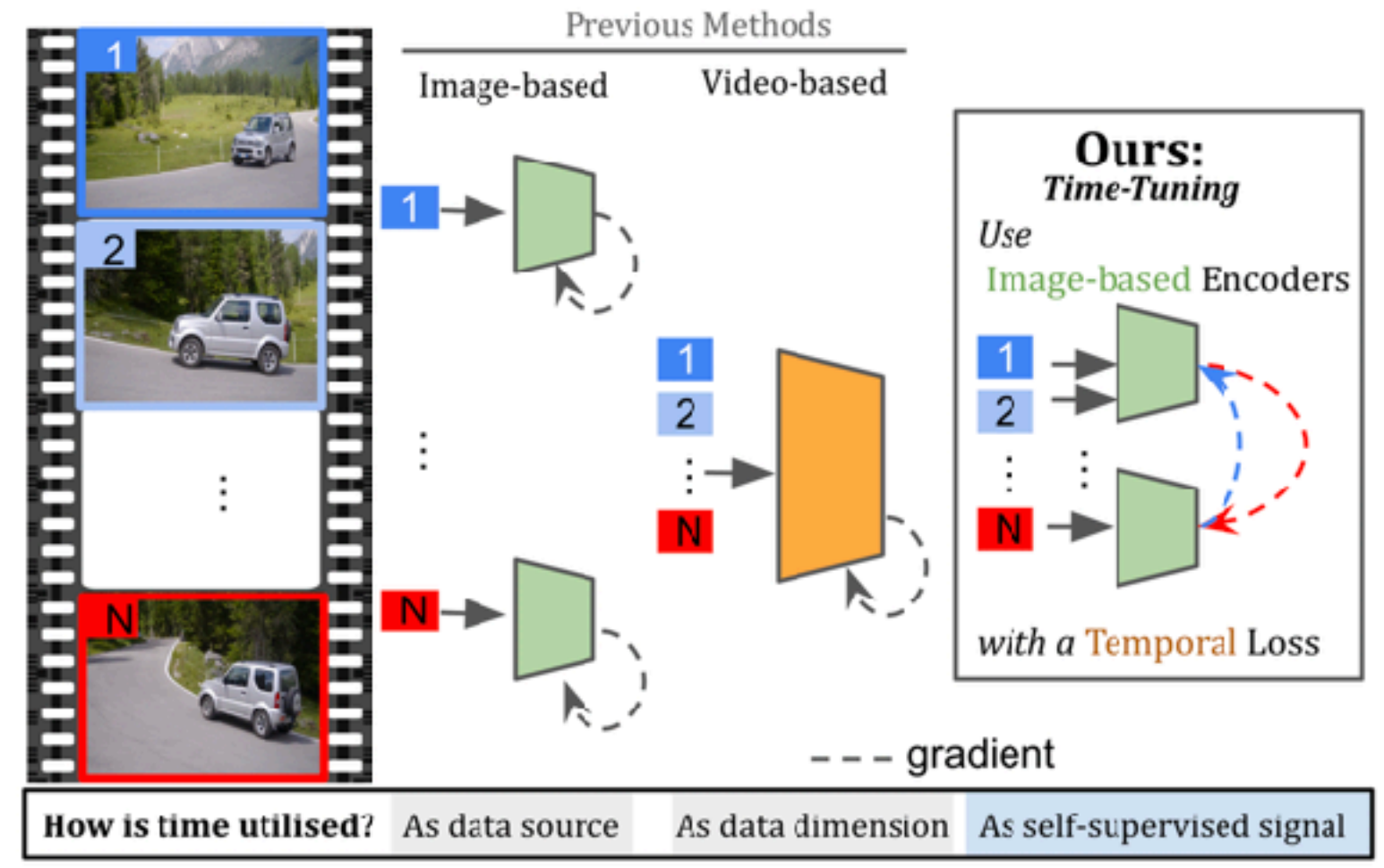
Salehi, Gavves, Snoek, Asano. *Time does tell: self-supervised time-tuning of dense image representations.* ICCV 2023

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

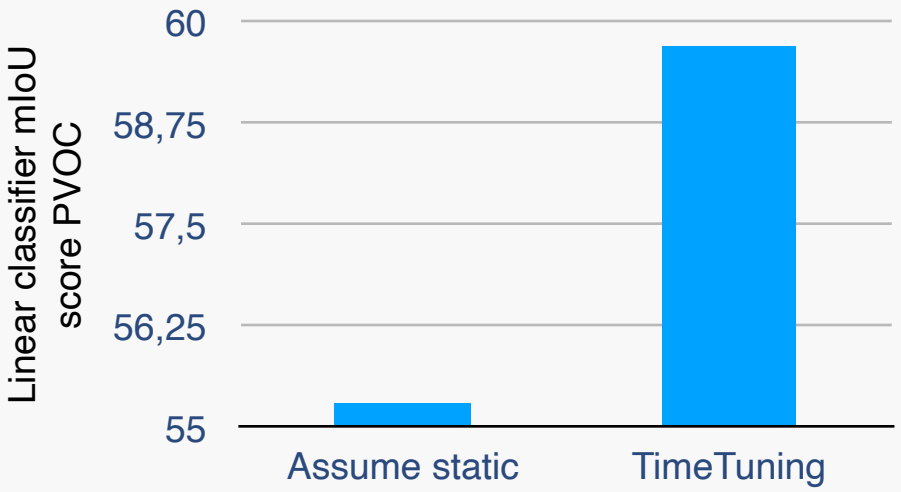
Using videos to learn self-supervised image encoders



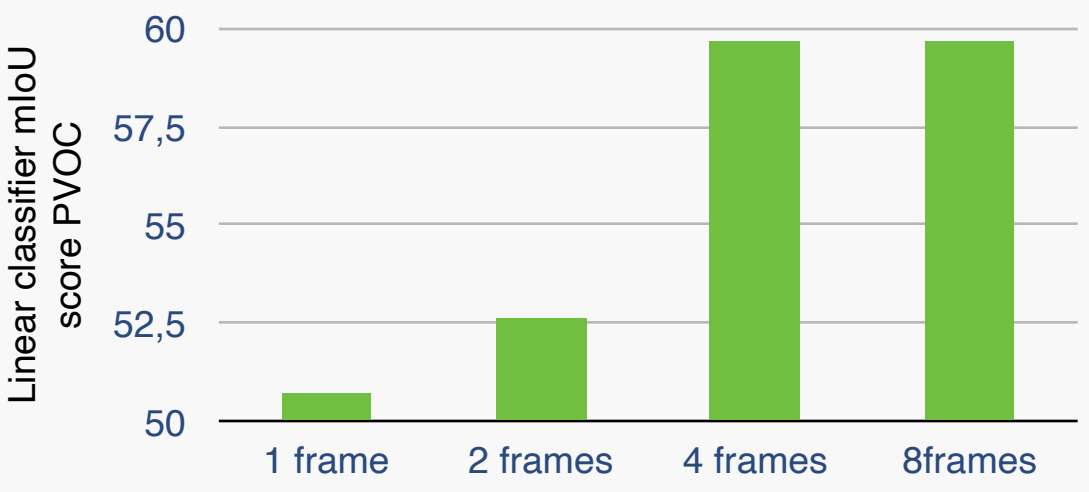
Salehi, Gavves, Snoek, Asano. *Time does tell: self-supervised time-tuning of dense image representations.* ICCV 2023

Ablations demonstrate using time helps learn better features

Modelling time is esential



Model learns from temporal info



Results

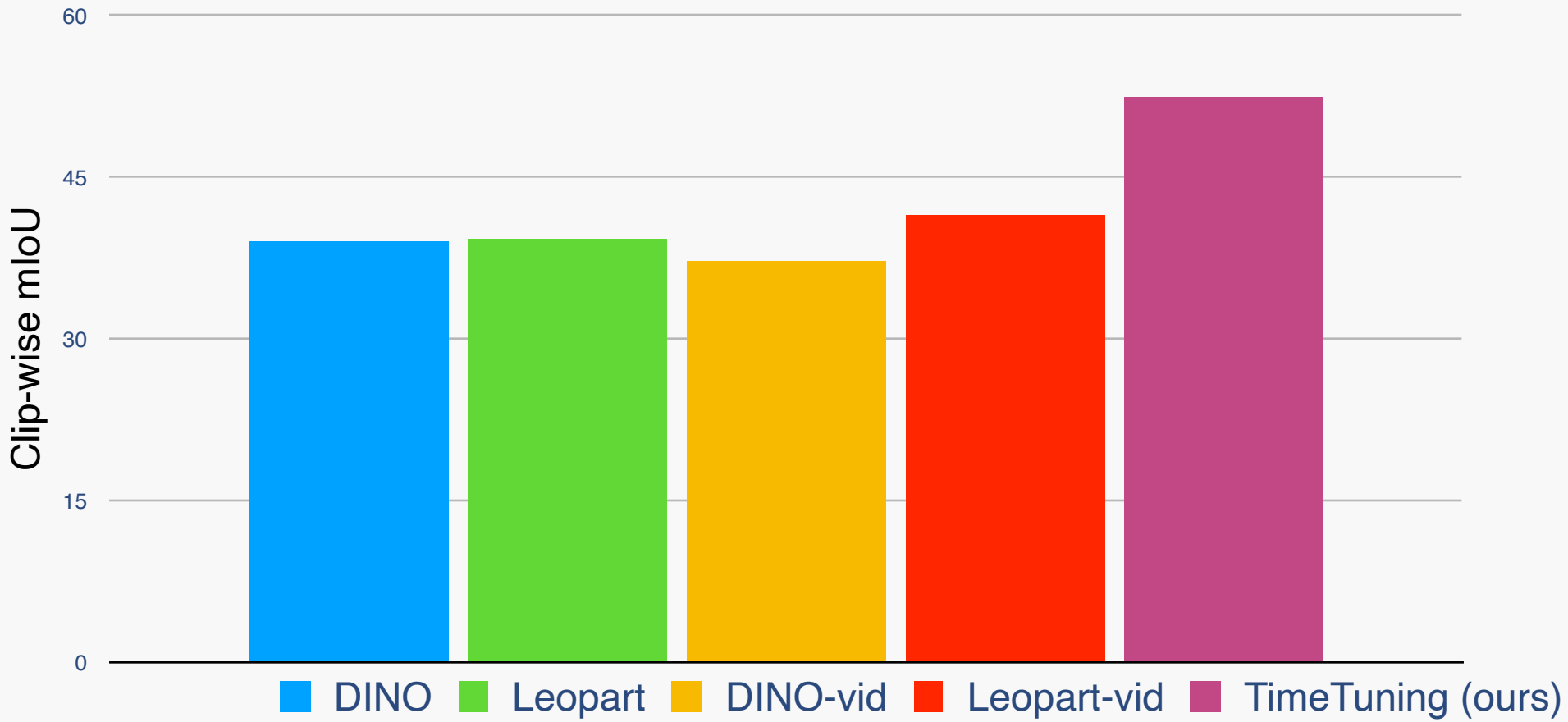
SoTA on unsupervised video segmentation

	Clustering					
	YTVOS			DAVIS		
	F	*C*	*D*	*F*	*C*	*D*
Trained on Images						
Resnet50	44.0	43.4	1.7	39.3	37.4	4.2
SwAV [8]	39.5	38.2	3.2	32.0	29.6	7.3
DINO [9]	39.1	37.9	1.9	30.2	31.0	1.6
Leopart [74]	39.2	37.9	11.7	30.3	30.2	16.5
Trained on Videos						
STEGO*	41.5	40.3	2.0	31.9	31.0	3.2
DINO*	37.2	36.1	1.2	29.3	29.2	2.4
Leopart*	41.5	40.5	7.7	37.5	36.5	12.6
TIMET(ours)	**52.5**	**51.3**	**13.3**	**53.7**	**53.0**	**20.5**

SoTA on unsupervised image segmentation

	Pascal VOC			
	K=21	K=500	LC	FCN
Trained on Images				
ResNet-50	4.5	36.5	53.8	-
DINO [9]	5.5	17.4	50.6	60.6
SwAV [8]	11.6	35.7	50.7	-
MaskContrast [57]	35.0	45.4	49.2	-
DenseCL [61]	-	43.6	49.0	69.4
STEGO [21]	7.0	19.5	59.1	63.5
CrOC [52]	20.6	-	61.6	-
Leopart [74]	**36.6**	50.5	**68.0**	70.1
Trained on Videos				
STEGO*	4.0	15.5	51.1	55.5
Leopart*	14.9	21.2	53.2	63.2
Flowdino† [70]	-	-	59.4	-
TIMET (ours)	34.5	**53.2**	**68.0**	**70.6**

Results on unsupervised video semantic segmentation



Unsupervised Semantic Segmentation on videos

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

Fundamental AI Lab

UTN

DINO

❌ part-centric maps

STEGO

❌ noisy maps

Ours

✅ crisp semantic maps

Key take-aways

- **Videos** provide rich supervision signal
- Don't use frame-wise invariance across time, but instead patch-level invariance
- Start with strong model, further **improve it**

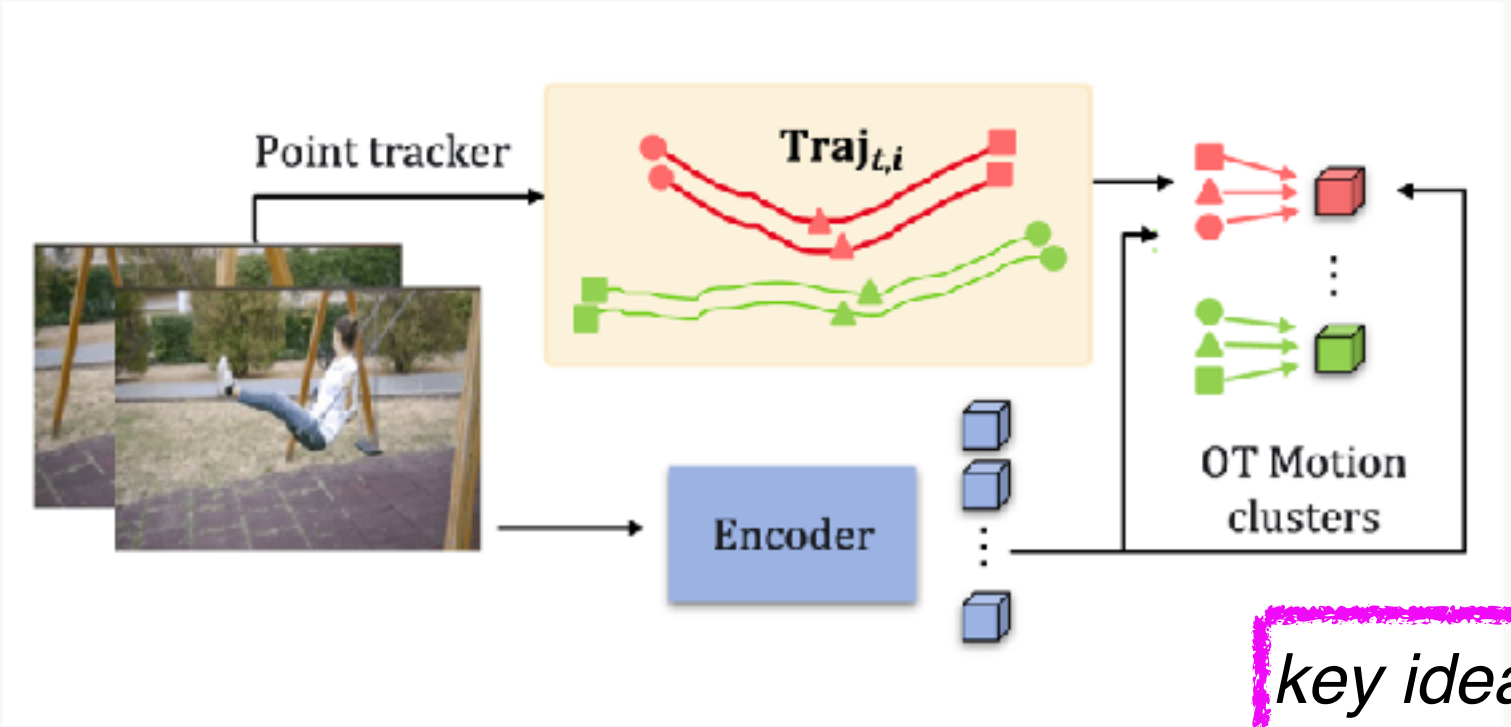
Also: this guy!

Advanced Machine Intelligence

REAL WORLD. REAL INTELLIGENCE.

Upgrades (ICCV 2025 & ICML 2026)

MoSiC: Optimal-Transport Motion Trajectory for Dense Self-Supervised Learning



key idea:
Densely cluster point-trajectories with Optimal Transport (OT)

https://arxiv.org/abs/2506.08694

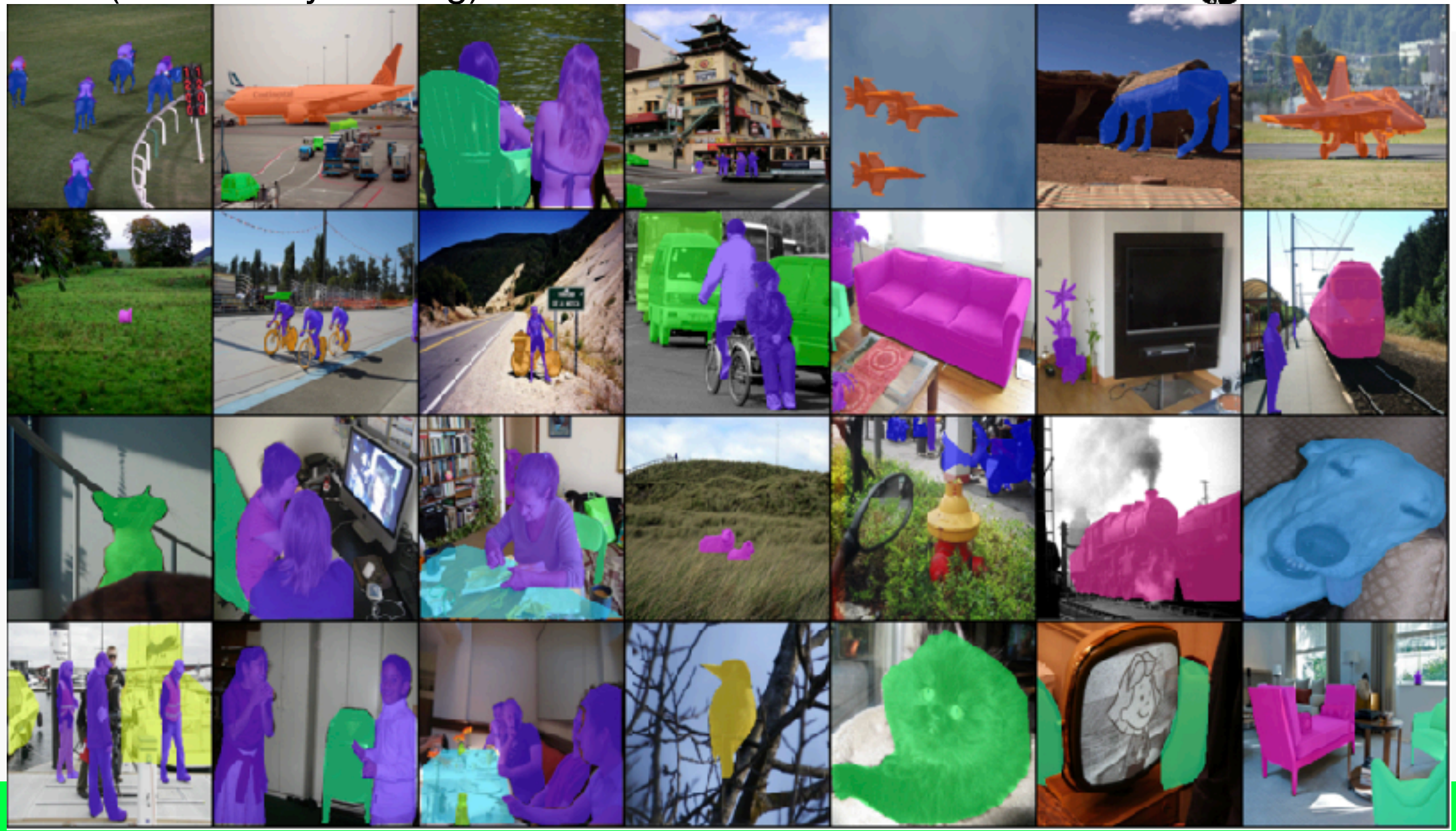
Frozen In-Context Semantic Segmentation Evaluation

Large gains even compared to TimeT

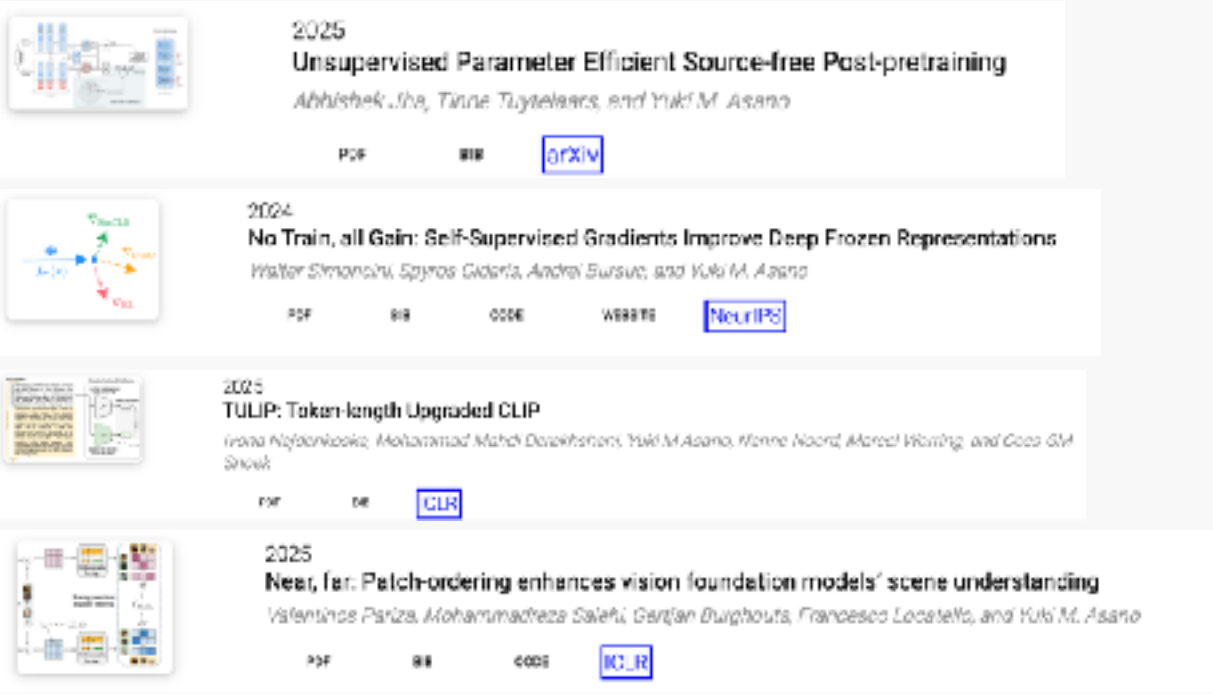
Method	Backbone	Params	ADE20K 1/128	ADE20K 1/64	ADE20K 1/8	ADE20K 1/1	Pascal VOC 1/128	Pascal VOC 1/64	Pascal VOC 1/8	Pascal VOC 1/1
Trained on Images										
DINO [10]	ViT-S/16	21M	9.5	11.0	15.0	17.9	26.4	30.5	41.3	48.7
CrOC [54]	ViT-S/16	21M	8.7	10.8	15.2	17.3	34.0	41.8	53.8	60.5
SelfPatch [66]	ViT-S/16	21M	10.0	10.9	14.7	17.7	28.4	32.6	43.2	50.8
Leopart [71]	ViT-S/16	21M	12.9	14.8	19.6	23.9	44.6	49.7	58.4	64.5
CrIBo [36]	ViT-S/16	21M	14.6	17.3	22.7	26.6	53.9	59.9	66.9	72.4
DINOv2 [42]	ViT-S/14	21M	22.8	26.4	33.5	38.8	56.0	62.4	72.3	77.0
Finetuned on Videos										
TimeT [50]	ViT-S/16	21M	12.1	14.1	18.9	23.2	38.1	43.8	55.2	62.3
MoSiC	ViT-S/14	21M	**23.8**	**27.4**	**35.7**	**40.7**	**62.5**	**66.6**	**74.7**	**78.2**
Trained on Images										
MAE [23]	ViT-B/16	85M	10.0	11.3	15.4	18.6	3.5	4.1	5.6	7.0
DINO [10]	ViT-B/16	85M	11.5	13.5	18.2	21.5	33.1	37.7	49.8	57.3
Leopart [71]	ViT-B/16	85M	14.6	16.8	21.8	26.7	50.1	54.7	63.1	69.5
Hummingbird [5]	ViT-B/16	85M	11.7	15.1	22.3	29.6	50.5	57.2	64.3	71.8
CrIBo [36]	ViT-B/16	85M	15.9	18.4	24.4	28.4	55.9	61.8	69.2	74.2
DINOv2 [42]	ViT-B/14	85M	24.2	27.6	34.7	39.9	55.7	61.8	72.4	77.1
Finetuned on Videos										
MoSiC	ViT-B/14	85M	**25.4**	**29.3**	**37.3**	**42.6**	**65.5**	**69.8**	**76.9**	**80.5**

(without any training) -- PVOC ICL eval

(without any training) -- PVOC ICL eval

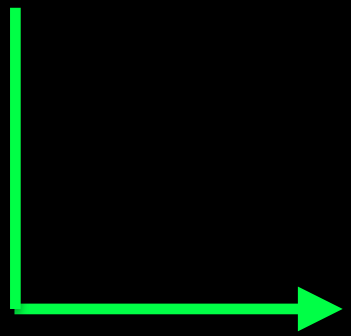


More post-training works



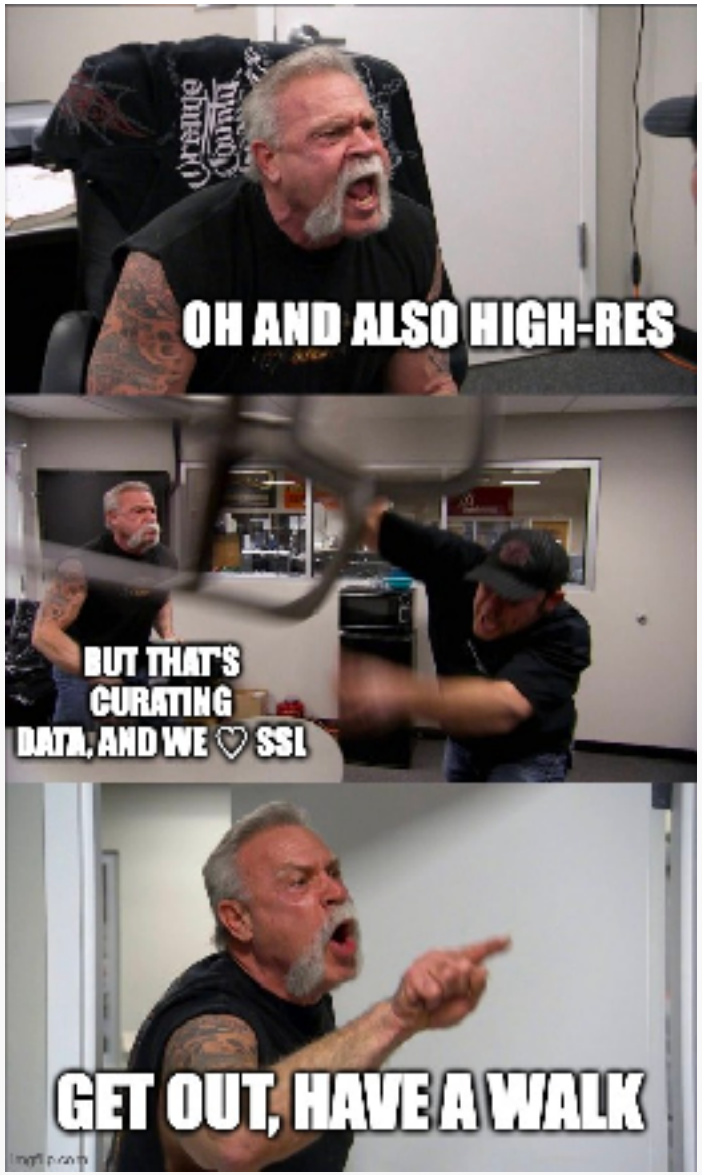
So videos can help post-train models.

But how powerful is “time” really?



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

us figuring out
which video to use

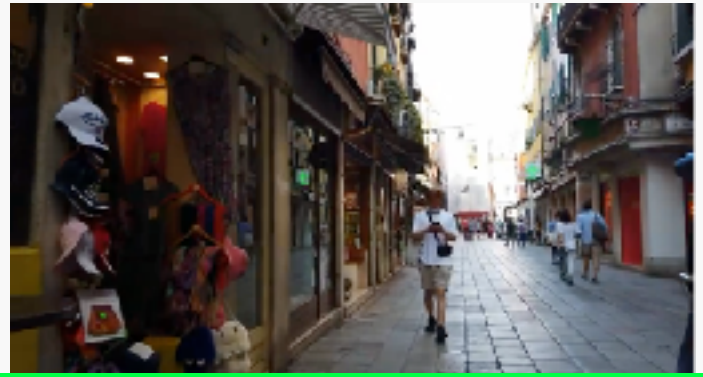


- ✓ Long
- ✓ High-res, smooth
- ✓ Semantically rich
- ✓ Scalable (we ❤️ SSL)

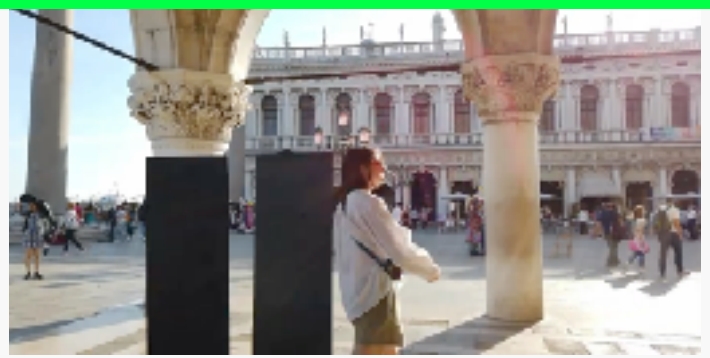
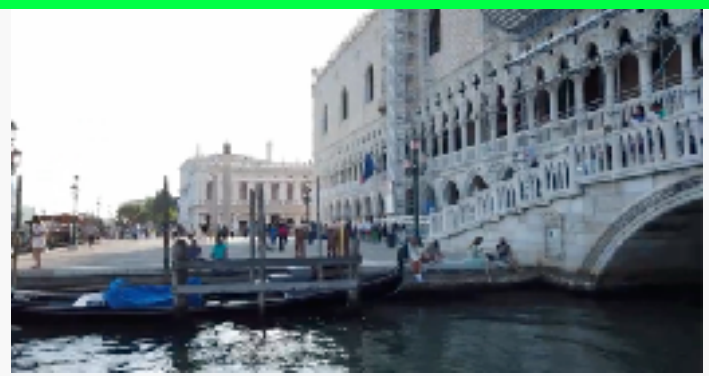
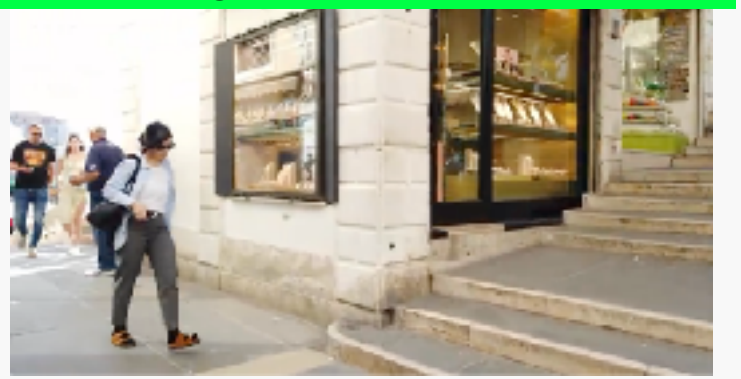
↓

Walking Tours



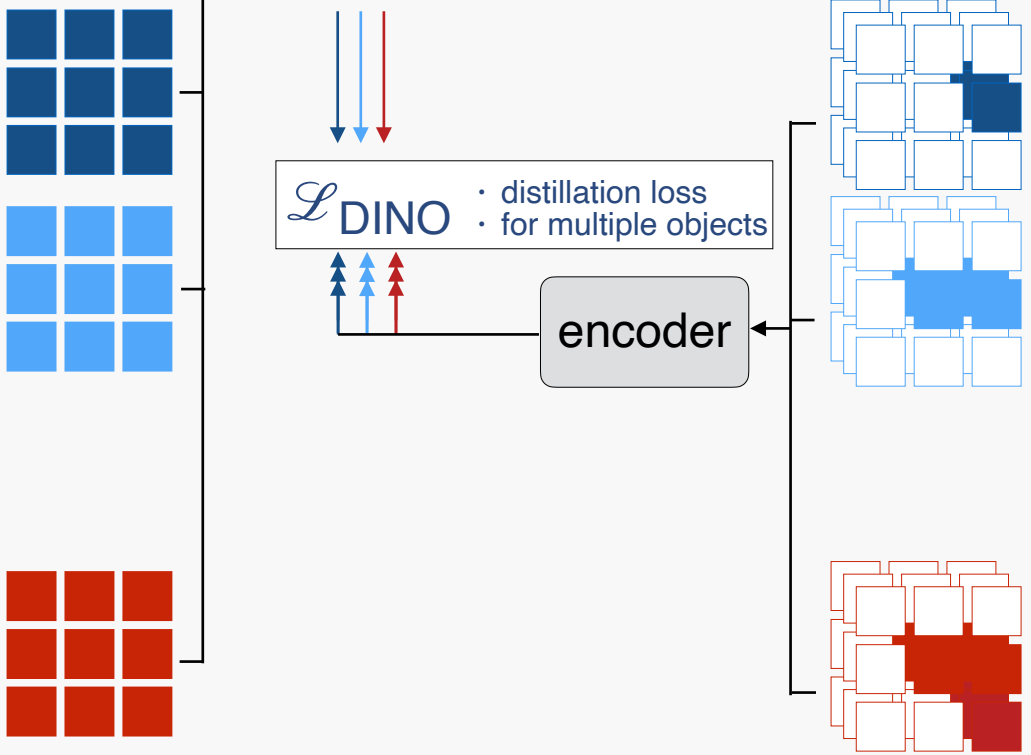
The dataset consists of 10x 4K videos of different cities' Walking Tours.

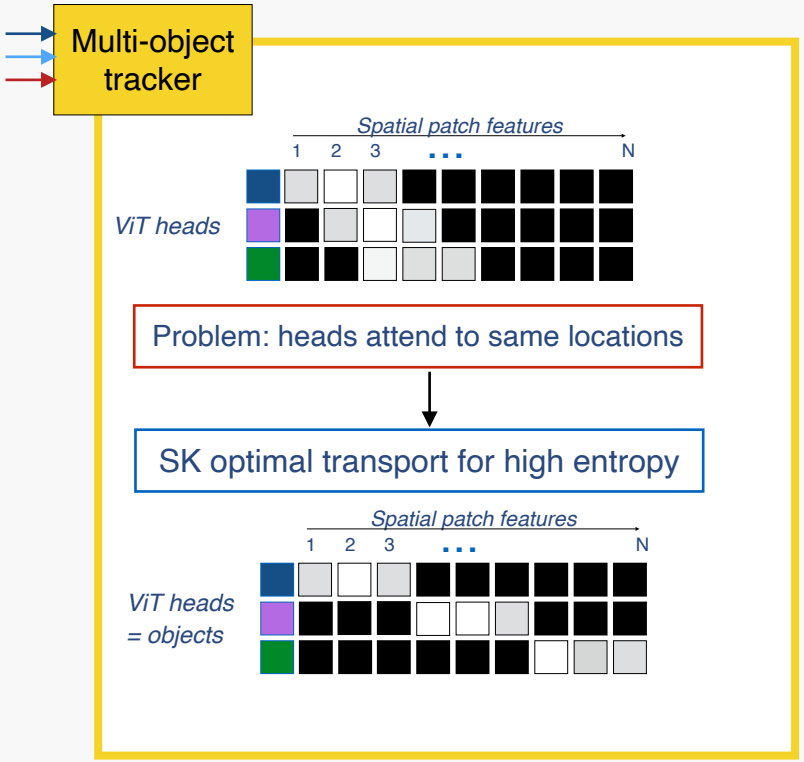
High-level idea:
1) track multiple objects across time
2) enforce invariance of features across time

Dora: **Disc**o**v**er and T**ra**ck

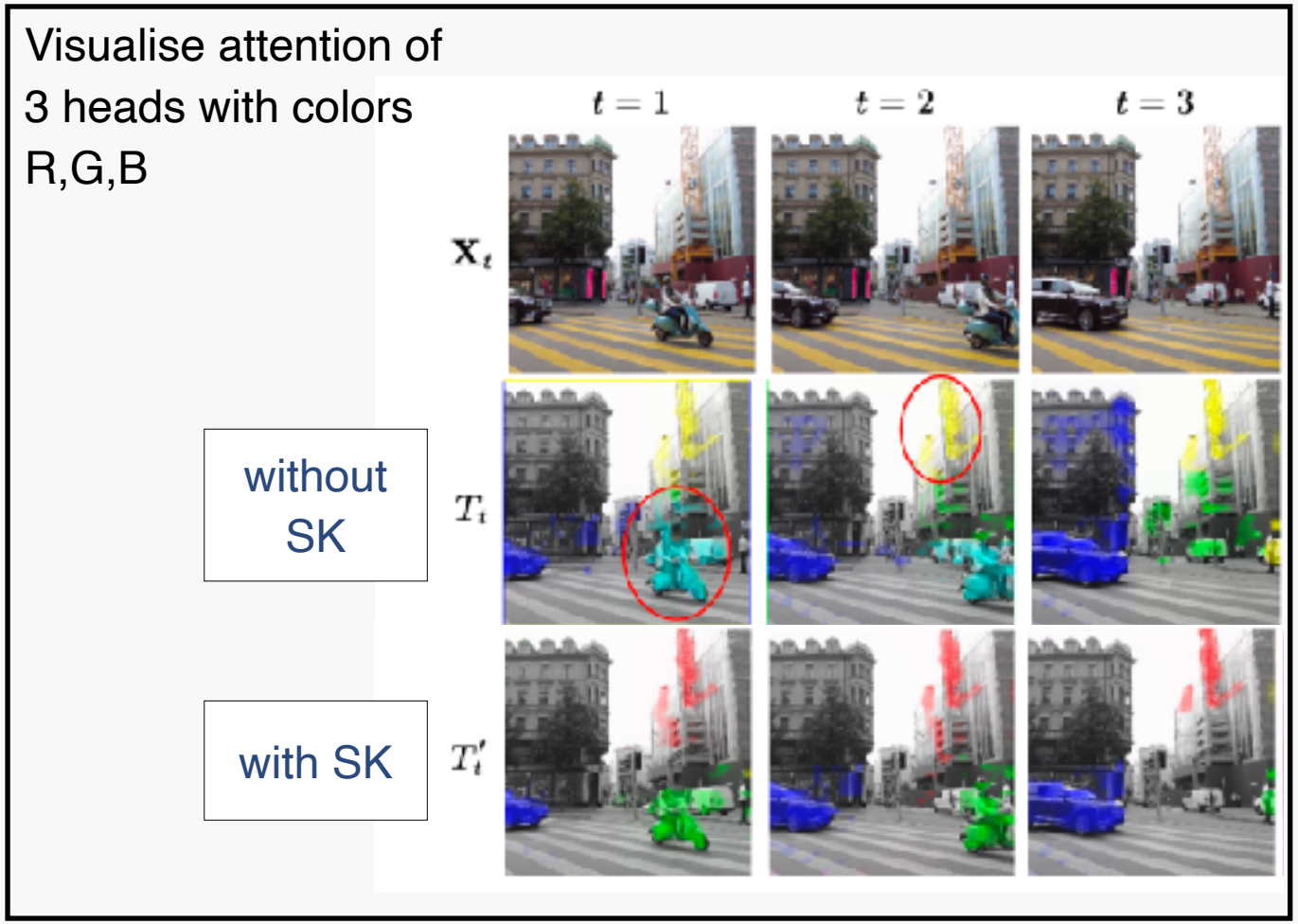
Much like Dora, we walk around and learn from what we see.



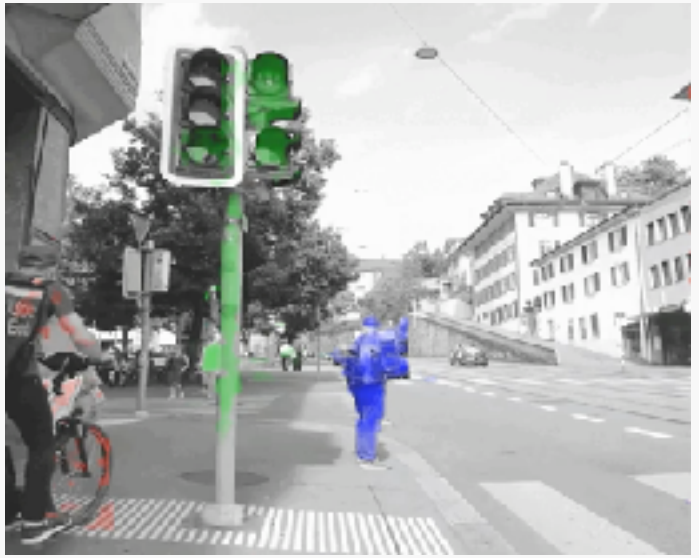
Spreading attention with Sinkhorn-Knopp



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

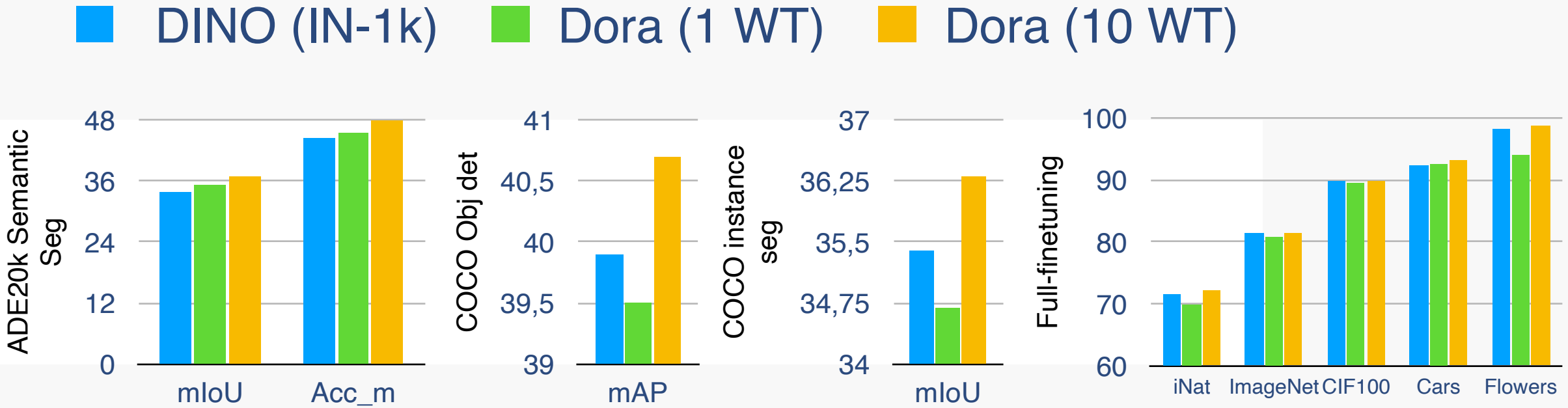


More examples: multi-object tracking in a ViT *emerges*



Venkataramanan, Rizve, Carreira, Asano*, Avrithis*. *Is ImageNet worth 1 video? Learning strong image encoders from 1 long unlabelled video.* ICLR 2024

But how does it compare against ImageNet pretraining?



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

Dora (10WT) > DINO (IN-1k)

Key takeaways

- Training strong encoders **from scratch** with 1 video is possible
- Models match DINO (trained on ImageNet) in terms of performance
- The training loss is **spatially dense** and leverages **time**
- **Multi-object tracking** emerges
- **Walking videos** are great for training vision models

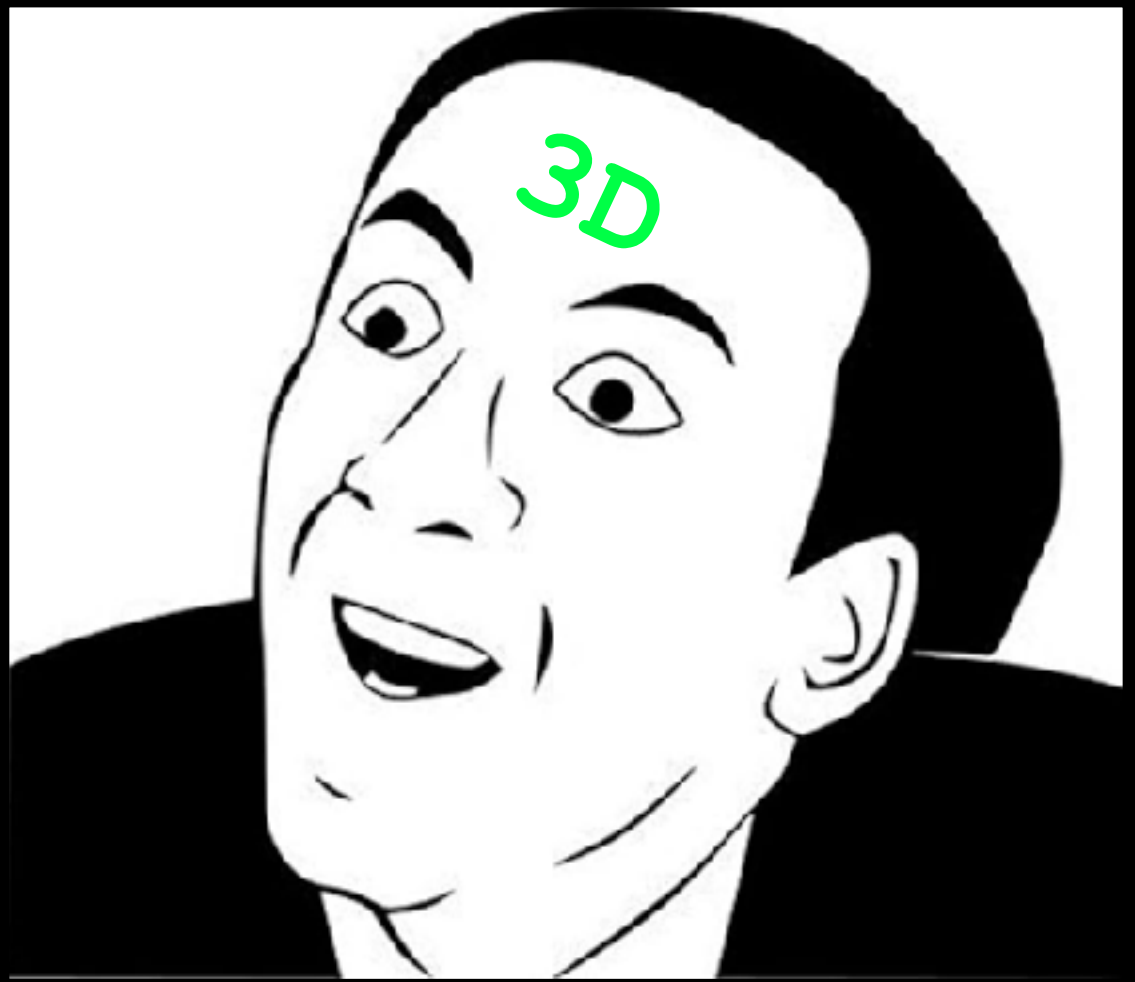
“We always have lots of raw data….”

Dataset	Source	Train	Val	Test	Total
ScanNet [23]	real	1,201	312	100	1,613
ScanNet++ [101]	real	712	178	126	1,016
S3DIS [1]	real	204	68	0	272
ArkitScenes [5]	real	4,498	549	0	5,047
HM3D [68]	real	8,881	1,119	0	10,000
Structured3D [109]	sim.	18,348	1,776	1,697	21,821
ASE [3]	sim.	90,000	10,000	0	100,000
Sonata (ours)	mixed	123,844	14,002	1,923	**139,769**

Table 1. **Data source collection.**

SOTA 3D rep. learning paper:

merely ~ 15k real scenes

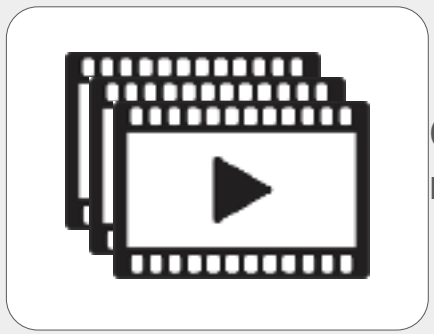


Sonata: Self-Supervised Learning of Reliable Point Representations. Wu et al. CVPR’25

We actually don’t have abundant 3D data.

How to scale 3D without 3D scans?

Our proposed solution



Web-scale
ego videos

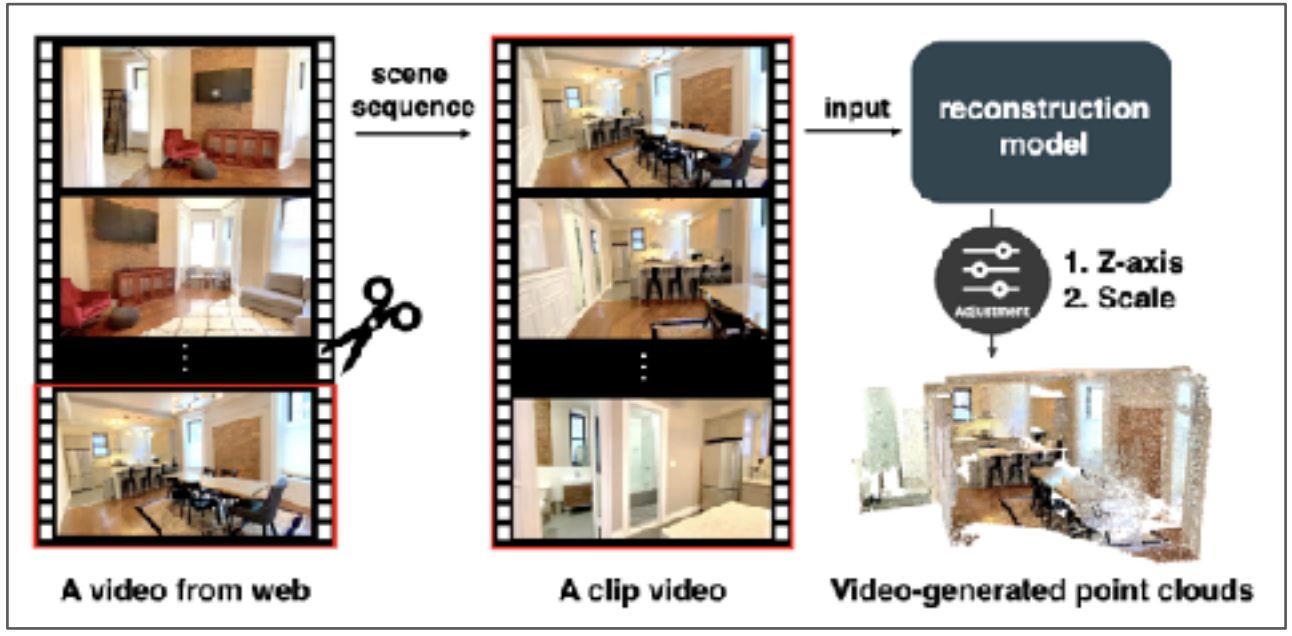
Geometry
models

Video-generated
point cloud data
(VGPC)

Noise-aware SSL
pretraining method

Yamada et al. 3D sans 3D Scans: Scalable Pre-training from Video-Generated Point Clouds. CVPR 2026

Video-generated point clouds (VGPC)

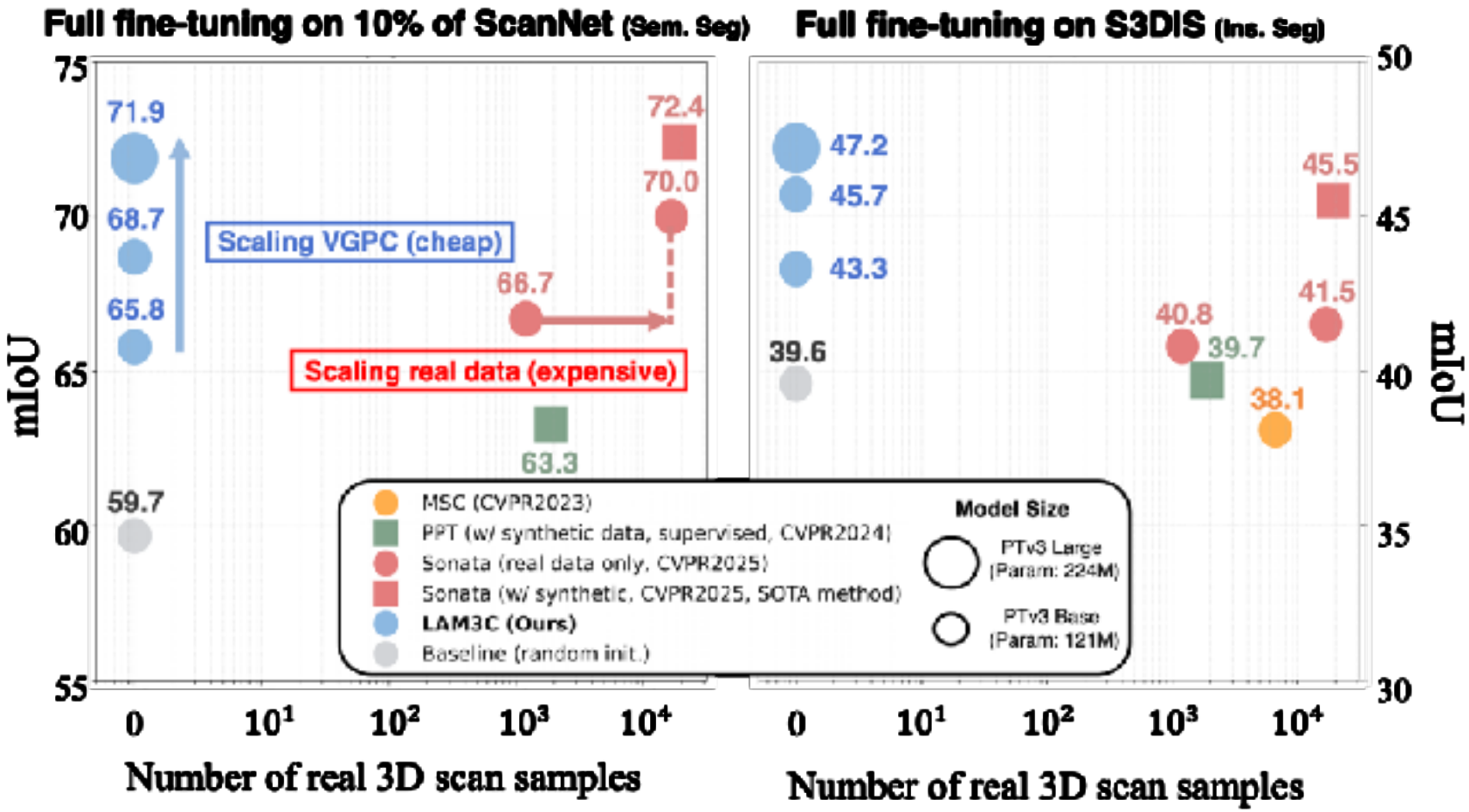


→ Scalable synthetic point cloud data pipeline

Yamada et al. 3D sans 3D Scans: Scalable Pre-training from Video-Generated Point Clouds. CVPR 2026

Super realistic point clouds from filtered web data

State-of-the-art performance with 0 real scans seen



Yamada et al. 3D sans 3D Scans: Scalable Pre-training from Video-Generated Point Clouds. CVPR 2026

Summary

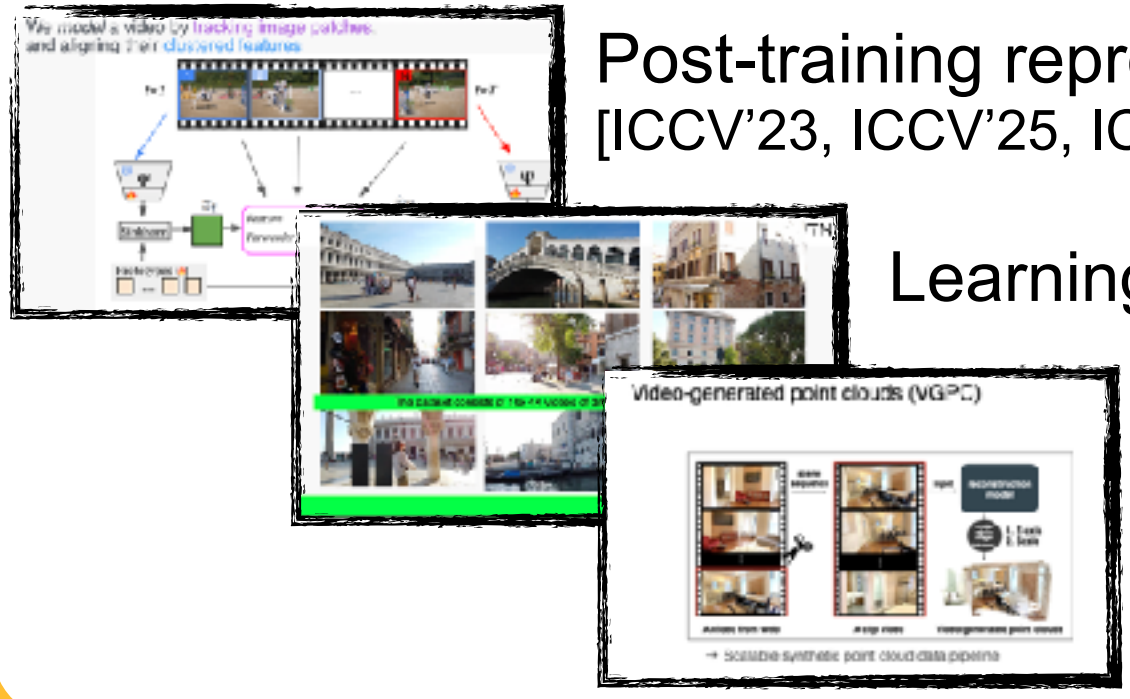
- VGPC as viable alternative to real 3D scans
- +SSL yields scalable 3D pretraining pipeline
- This makes video as key data modality for 3D

Yamada et al. 3D sans 3D Scans: Scalable Pre-training from Video-Generated Point Clouds. CVPR 2026

Put differently:

Vision from Moving

- Pretrain Foundation Models on billions of hours
- Robots collect video, sensor, and action data (*from-robots-for-robots*)



Post-training representations from video [ICCV'23, ICCV'25, ICML'26]

Learning from a single video [ICLR'24]

Video to scalable 3D [CVPR'26]

Intelligence as a metric of a learning process: key factors

Performance on tasks

Amount of supervision/feedback

Ability to generalise/adapt

Now, let’s talk about generalisation.

Currently, in (M)LLM space:

We believe language should help generalise:

+

Horse

Slim body, hooves, mane, tail, long face, no horn.

Rhino

Large sturdy body, thick skin, four legs, prominent single horn on the nose.

?

Unicorn

Horse-like body + mane and tail, combined with a single rhino-like horn.

From the MetaCLIP paper, appendix p.14, Table 11:

	ImageNet	Food-101	CIFAR10	CIFAR100	CUB	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	MNIST	FER-2013	STL-10	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCAM	UCF101	Kinetics700	CLEVR	HatefulMemes	SST2
MetaCLIP (400M) ViT-L	76.2	90.7	95.5	77.4	75.9	70.5	84.7	40.4	62.0	93.7	94.4	76.4	61.7	46.5	99.3	59.7	71.9	47.5	29.9	30.9	70.1	75.5	57.1	35.1	56.6	65.6
# of cls. w/ non-zero counts	703/998	52/101	10/10	93/100	1/200	193/397	0/196	8/100	40/47	15/37	86/102	61/102	10/10	12/12	10/10	2/10	32/45	1/43	0/4	190/211	1/2	5/101	122/700	8/8	1/2	2/2

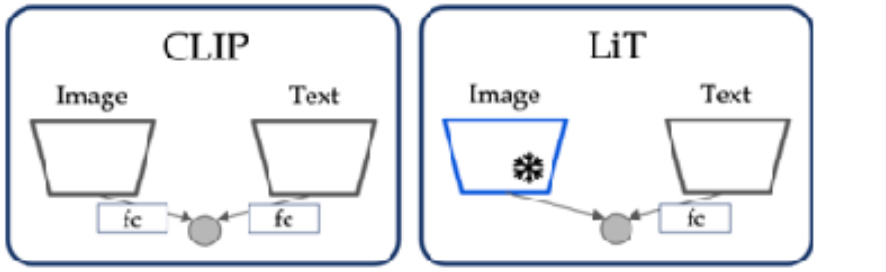
Table 11: Measuring task-alignment. First row: MetaCLIP (400M) ViT-L/14 accuracy, second row: number of classes matched in metadata

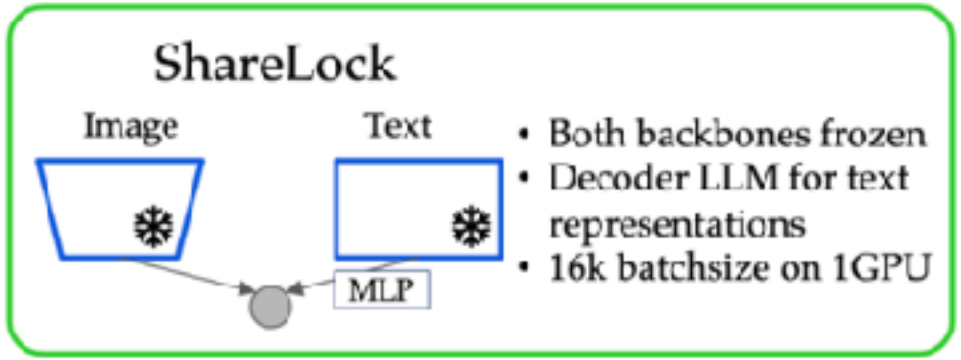
"Interestingly, there seems to be a correlation with the accuracy and the number of classes matched in the metadata."

CLIP, for the most part, is evaluated within-domain (it's just a big domain)

But surely language features, e.g. from pretrained models should help generalise?

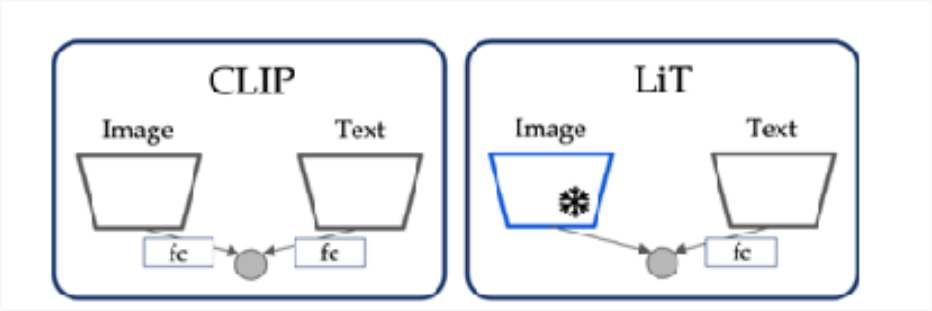
New method: **Share**d Vision-Language-**Lock**ed Tuning

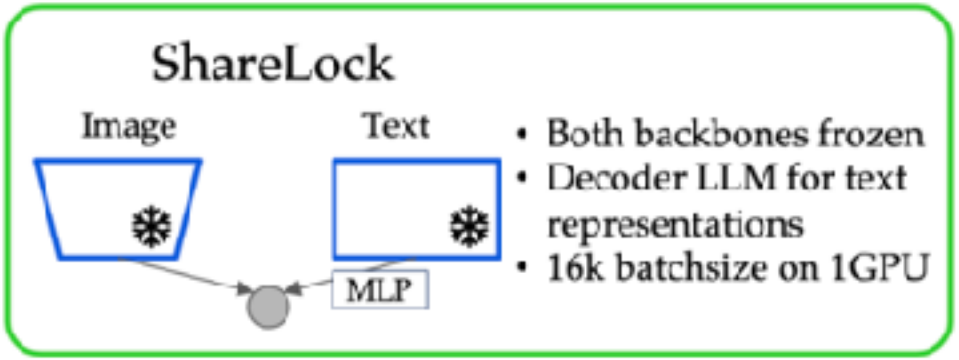




Result: CLIP-style model with that only mostly takes frozen representations

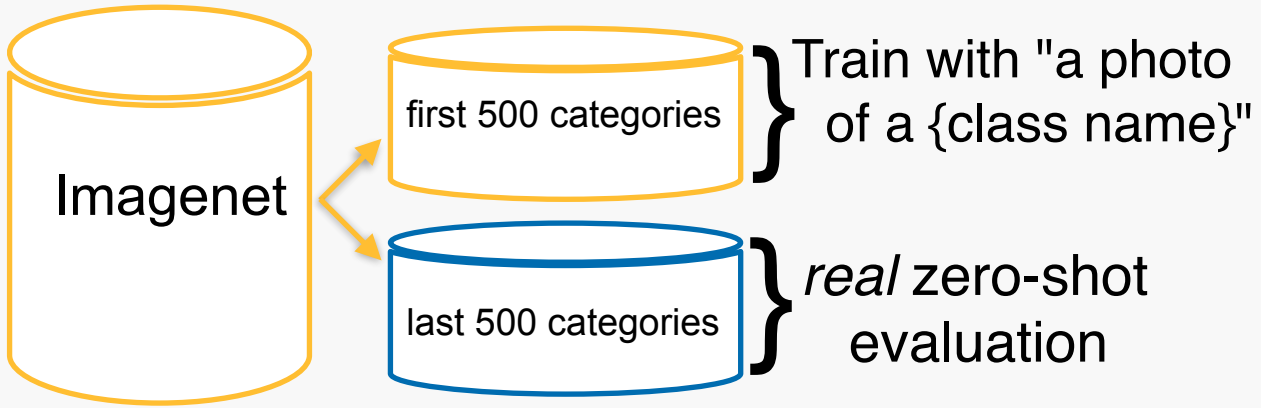
New method: **Share**d Vision-Language-**Lock**ed Tuning





Result: CLIP-style model with that only mostly takes frozen representations

New evaluation: Mutually exclusive vision-language dataset splits



Result: Clean measurement of *generalisation ability* from LLM

Decoder representations are actually really good.

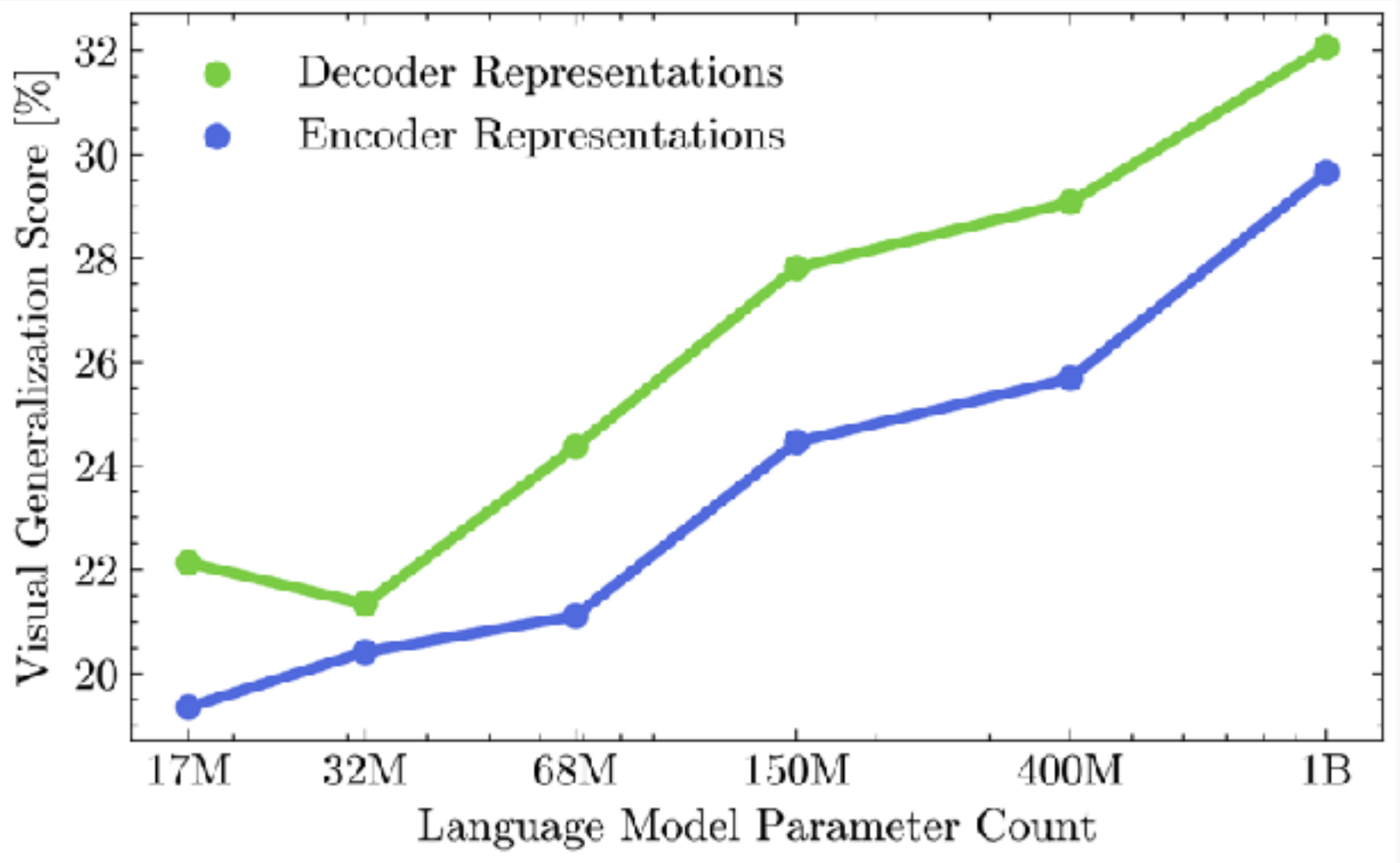
Type	Language Model	Class Names
Enc.	BERT-Large [9]	18.3
	T5-XL [47]	33.6
	Flan-UL2 [55]	37.0
	SentenceT5-XXL [39]	39.5
Dec.	Gemma 7B [16]	39.7
	Llama-3 8B [11]	<u>40.2</u>
	NV-Embed [31]	**40.5**

What people previously used

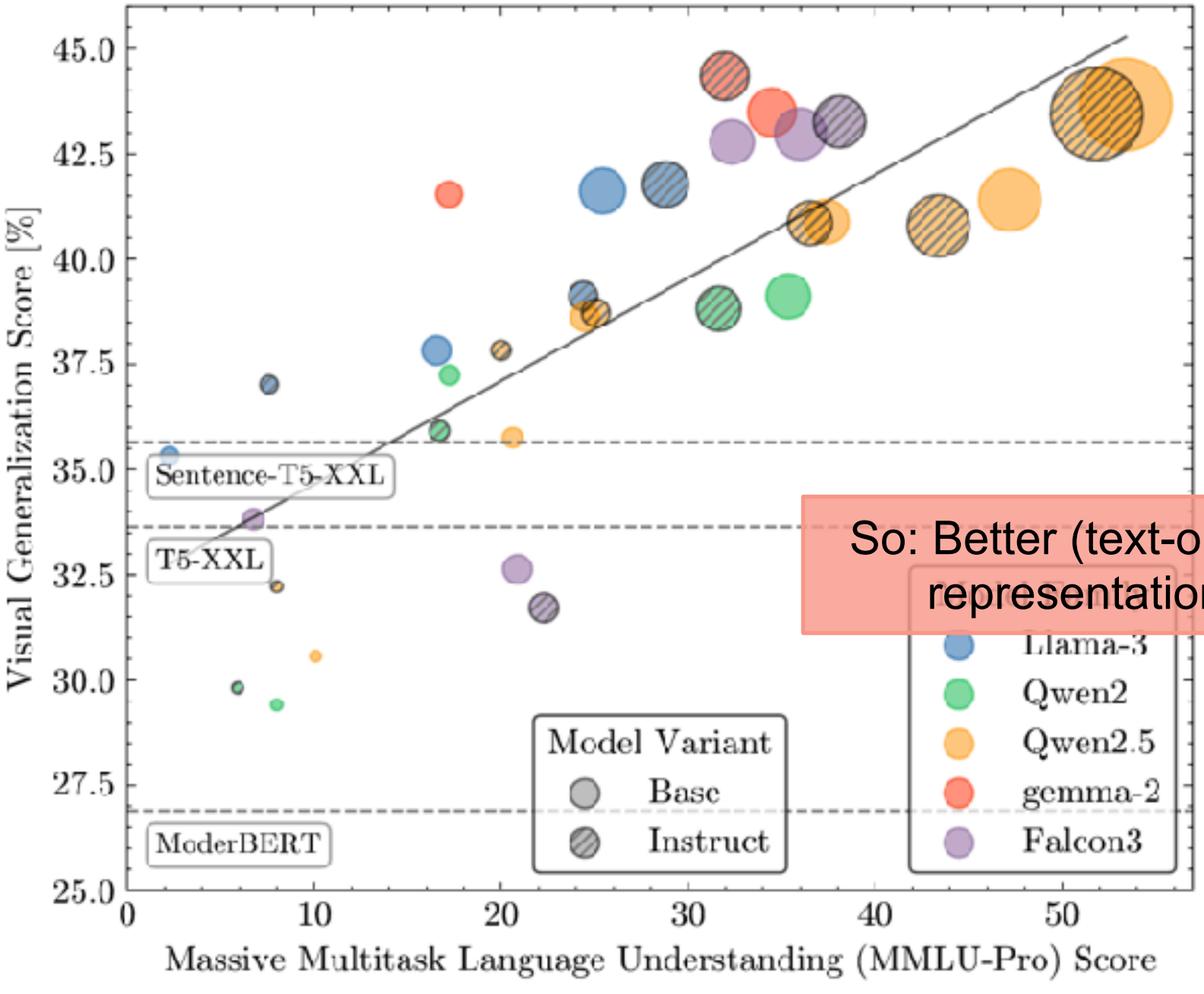
billion-scale LLMs

LLMs contain knowledge that helps visual zero-shot classification

Which representations are better?
(When using the same data)



Paired Ettin models [ICLR'26]

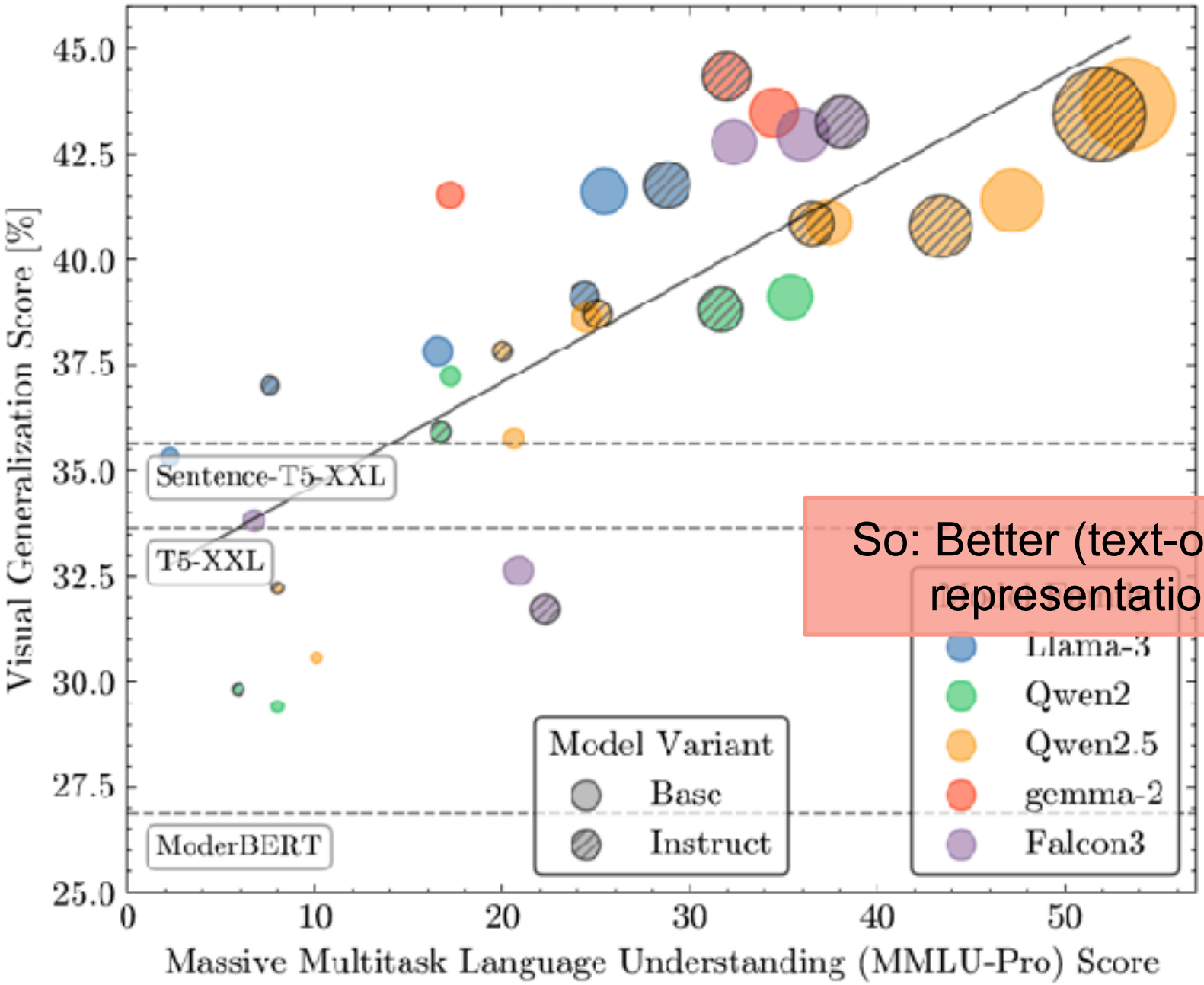


LLM's ShareLock performance correlates with (text-only) MMLU evaluation!

So: Better (text-only!) LLMs have a better representation of the **visual** world

Fundamental AI Lab

UTN



LLM's ShareLock performance correlates with (text-only) MMLU evaluation!

So: Better (text-only!) LLMs have a better representation of the **visual** world

And what if we train with actual image-caption datasets?

Fundamental AI Lab

UTN

Strong SotA for datasets 100k-12M

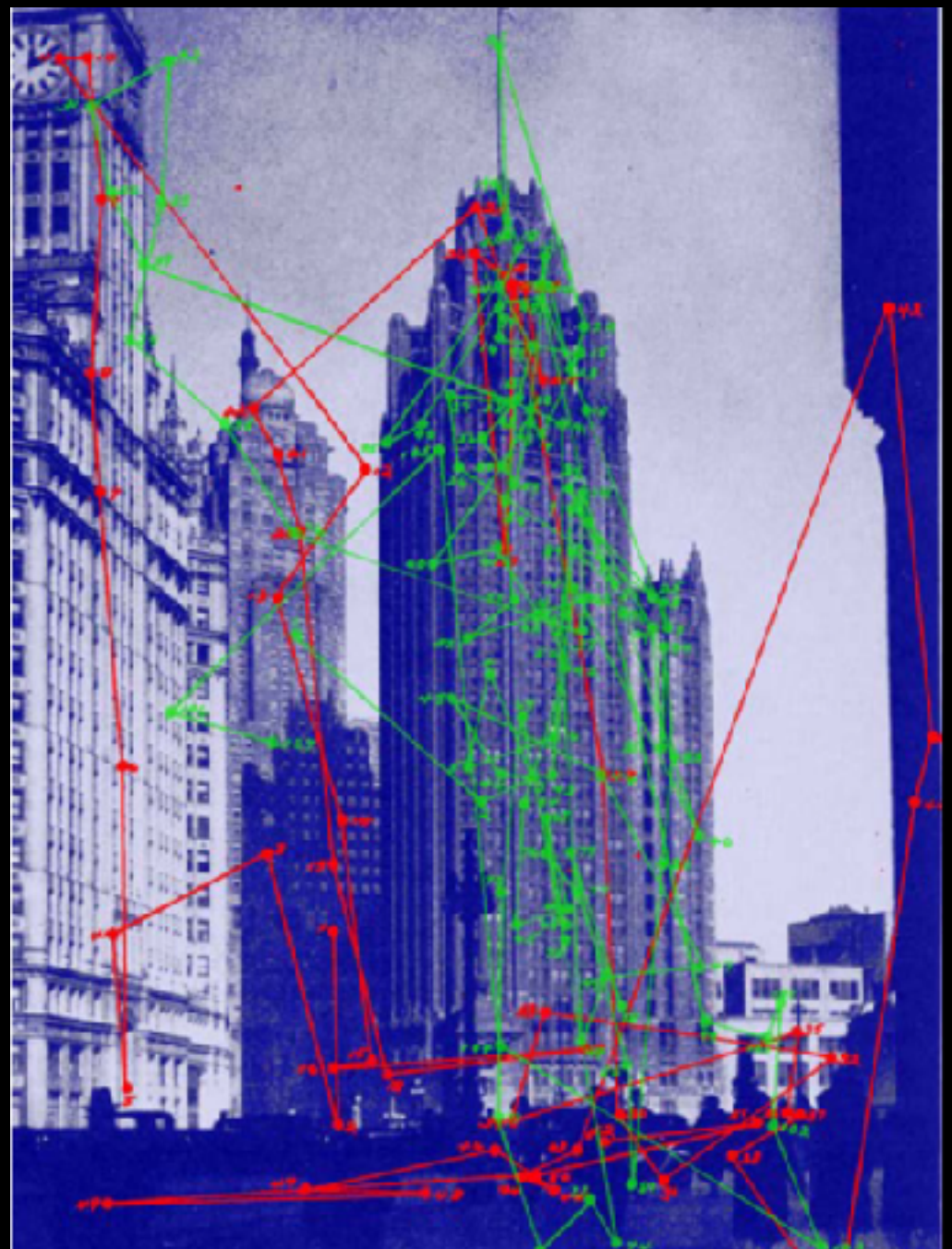
Model	Dataset	[Size]	IN-1k	IN-V2	IN-R	IN-A	IN Sketch	ObjectNet	Avg
LiT	COCO	83k	23.3	20.8	34.4	21.1	18.4	29.2	24.5
ASIF	COCO	83k	9.4	8.7	14.4	8.8	6.9	16.1	10.7
ShareLock	COCO	83k	**32.2**	**28.6**	**36.6**	**22.8**	**22.4**	**30.4**	**28.8**
LiT	CC3M Subset	563k	41.7	37.5	59.2	44.4	32.4	40.7	42.6
ASIF	CC3M Subset	563k	21.6	20.5	27.7	24.4	14.9	21.5	21.8
ShareLock	CC3M Subset	563k	**50.5**	**45.8**	**60.5**	**47.0**	**36.9**	**41.1**	**47.0**
CLIP [12]	CC3M	2.8M	16.0	13.2	17.6	3.6	6.4	8.2	10.8
SLIP [38]	CC3M	2.8M	23.5	20.2	26.8	6.8	12.1	14.3	17.3
LaCLIP [12]	CC3M	2.8M	21.3	18.6	23.5	5.0	10.6	10.2	14.9
LiT	CC3M	2.8M	44.1	39.3	62.7	45.6	34.8	**43.3**	45.0
ShareLock	CC3M	2.8M	**52.1**	**47.1**	**64.1**	**50.9**	**39.0**	43.1	**49.4**
DataComp [14]	CPool-S	3.84M	3.0	2.7	4.4	1.5	1.3	3.7	2.8
CLIP [12]	CC12M	12M	41.6	35.4	52.6	10.7	28.8	24.0	32.2
SLIP [38]	CC12M	12M	41.7	35.9	55.2	13.8	30.7	29.3	34.4
LaCLIP [12]	CC12M	12M	49.0	43.3	63.8	14.7	39.4	28.1	39.7
LiT	CC12M	8.5M	56.2	49.9	**70.3**	52.8	43.9	**47.8**	53.5
ShareLock	CC12M	8.5M	**59.1**	**53.2**	68.8	**53.4**	**44.5**	46.7	**54.3**
DataComp [14]	CPool-M	38.4M	23.0	18.9	28.0	4.3	15.1	17.7	17.8
DataComp [14]	CPool-L	384M	55.3	47.9	65.0	20.2	43.2	46.5	46.3
CLIP [46]	Proprietary	400M	68.4	61.8	77.6	50.1	48.2	55.4	60.2

TLDR:

ShareLock is an ultra-lightweight vision-language model that

shows better LLMs have better visual representations which actually generalise.

+ strong IN-1k “zero-shot” performance of 51% in <**15 GPU hours**.



Eye movement patterns when

observers are instructed to locate a person looking out of a window in the tower

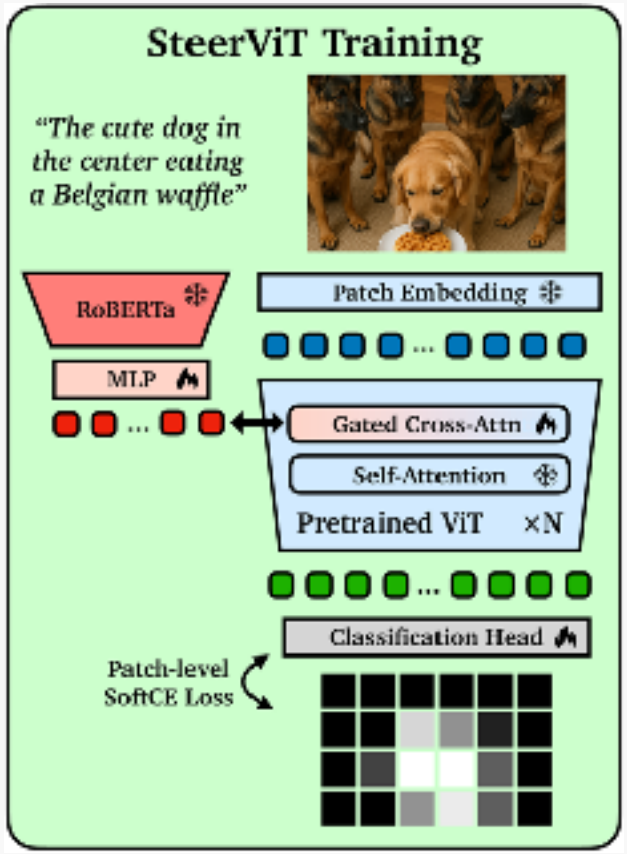
versus

when no contextual instruction is provided.

Adapted from Buswell (1935)

Can we have instruction-following visual encoding?

Method



- **Paradigm:** inverse-MLLM; early-fusion
- **Backbone:** any pretrained ViT
- **Approach:** gated cross-attention
- **Lightweight:** only 21M trainable parameters
- **Proxy task:** referential grounding
- **Goal:** instill language understanding into ViT

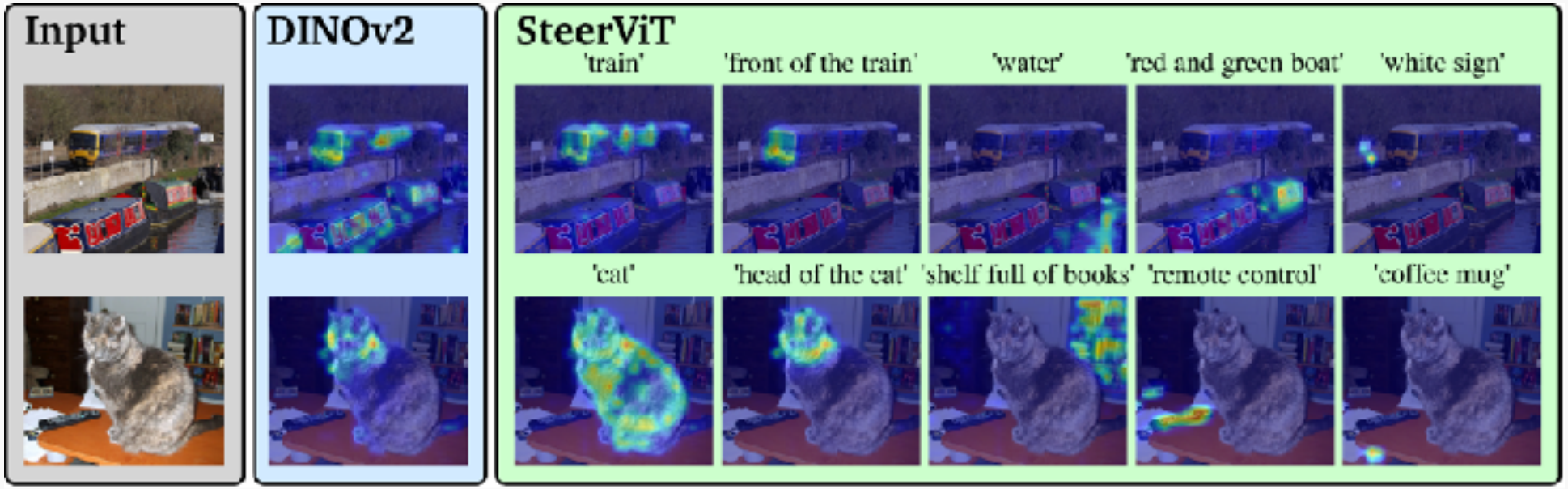
Measuring Steerability

CORE: object-conditional retrieval of images in cluttered scenes

Text changes what features encodes: 96.0 r@1 for SteerViT vs. 43.7 for DINOv2

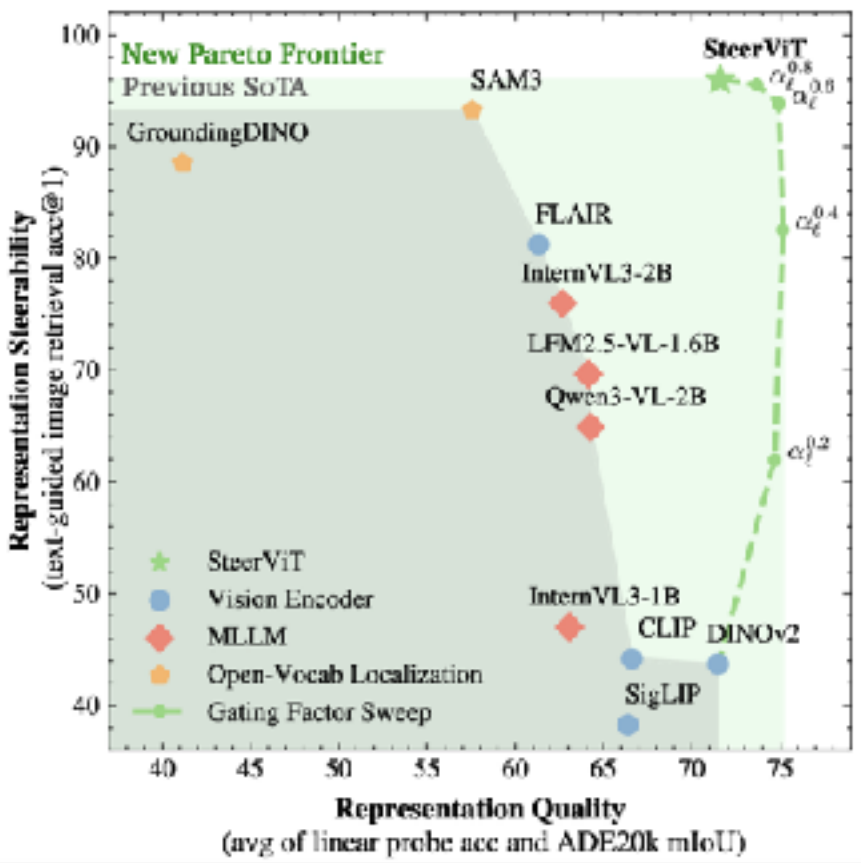
Global Aggregation

CLS Attention Maps:



Text directs local attention to queried object / region

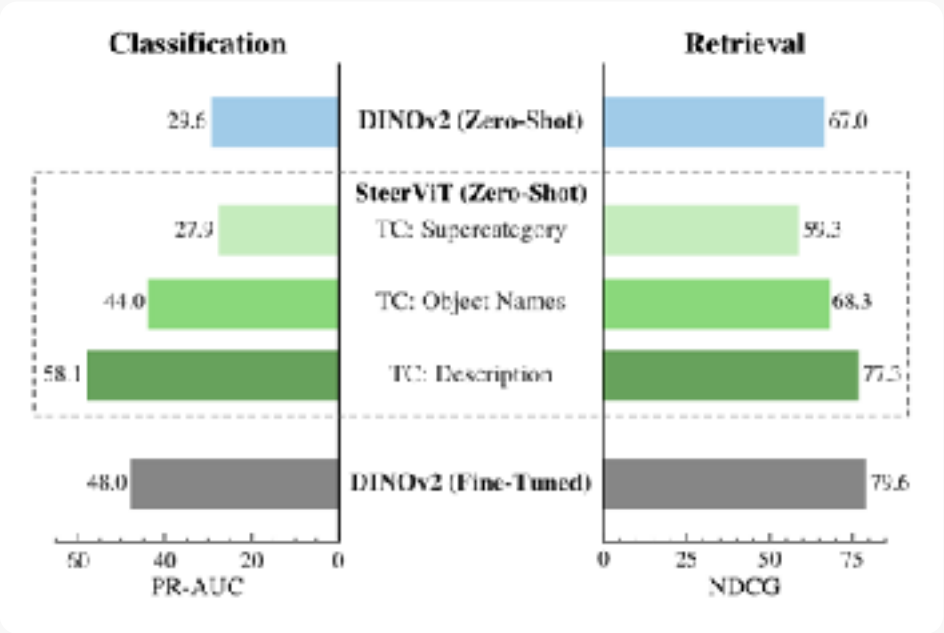
Steerability↔Quality Tradeoff



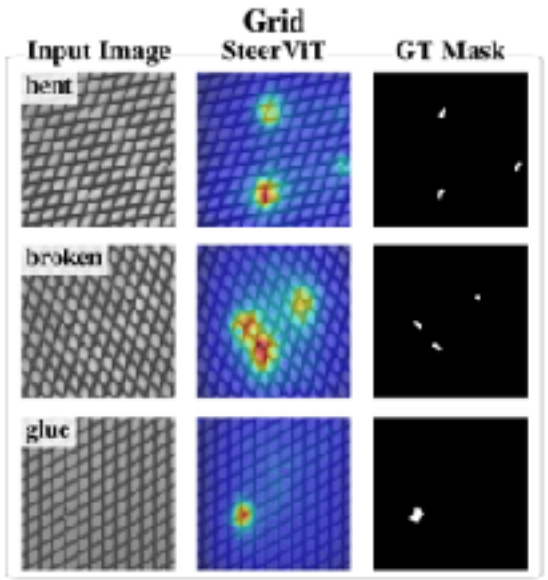
- **OV-Loc**: good steerability but low visual fidelity
- **MLLMs:** some steerability, but language-centric
- **ViTs:** strong features but no steering
- **SteerViT**: new Pareto optimal
  - Steerable representations
  - No loss in downstream performance

Applications

Personalized Object Discrimination



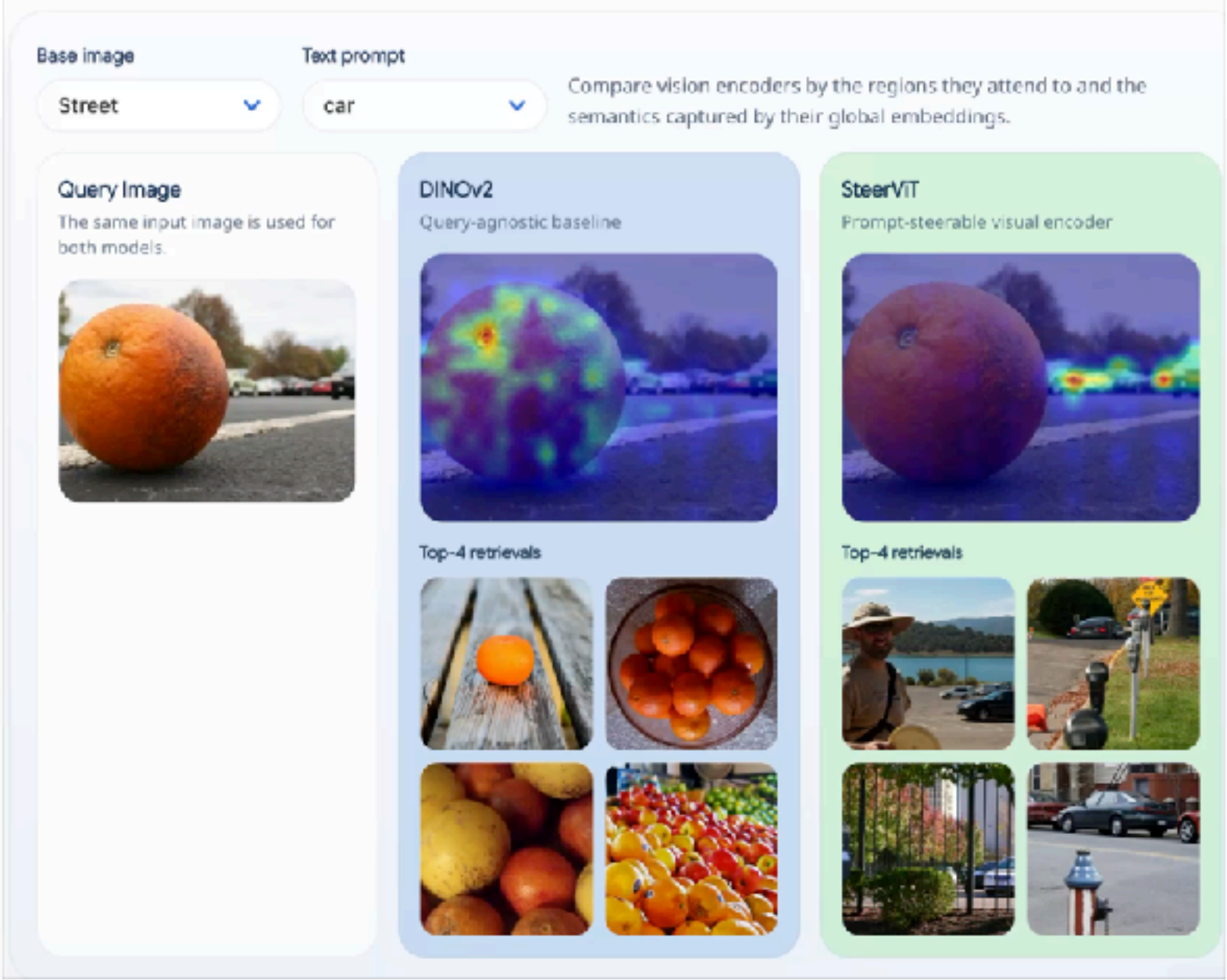
Industrial Anomaly Segmentation



Method	PRO
MaskCLIP [42]	40.5
CLIPseg [25]	34.6
SAM3 [6]	54.5
WinCLIP [14]	64.6
DIVAD [12]	73.3
FADE [21]	84.5
SteerViT	82.1

SteerViT matches or outperforms specialized methods in OOD tasks

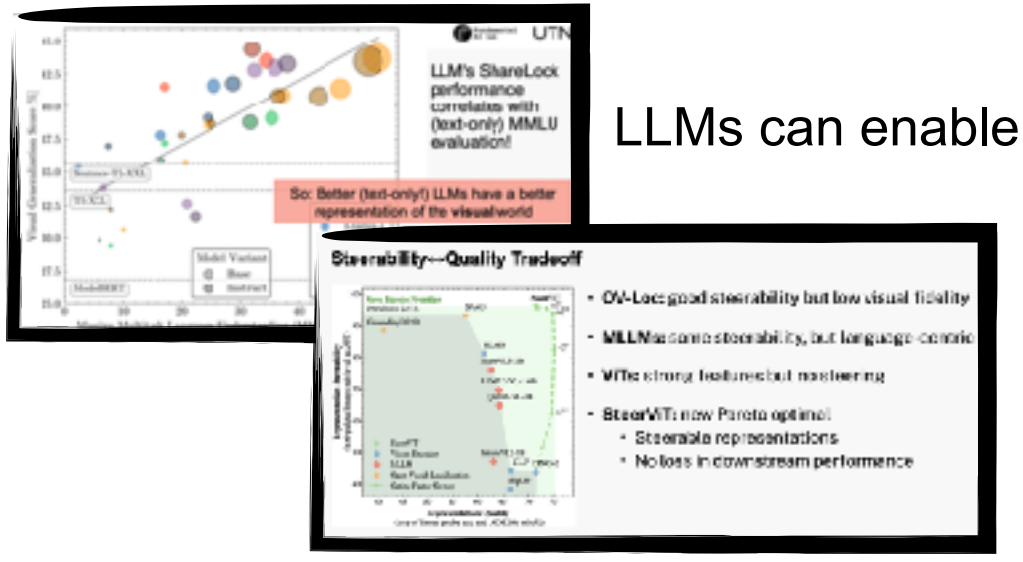
Demo


Base image
Street
Text prompt
car
Compare vision encoders by the regions they attend to and the semantics captured by their global embeddings.
Query Image
The same input image is used for both models.
DINOv2
Query-agnostic baseline
Top-4 retrievals
SteerViT
Prompt-steerable visual encoder
Top-4 retrievals

Takeaways

- **SSL-pretrained ViT**
  + lightweight supervised CA adapters
- in a “**Reverse-MLLM**” architecture
- can unlock some new **vision steering abilities**

And is worth a thought!



LLMs can enable generalisation to unseen concepts [TMLR'26]

Text-conditioned vision enables new tasks [arxiv'26]

Generalisation from Language

- Generalisation + world knowledge from LLMs
- Cross-modal learning allows new capabilities for vision

Towards VGI…

Openings for **female post-docs,** who previously did not live in Germany!

Vision from Moving

- Pretrain Foundation Models on billions of hours
- Robots collect video, sensor, and action data (*from-robots-for-robots*)

Generalisation from Language

- Generalisation + world knowledge from LLMs
- Summarize, infer semantics, generate curricula and goals
- Produce instructions, reward signals, choose exploration
- Guide how the system keeps learning over time

Hi, I'm Yuki

- Full Professor and head of Fundamental AI Lab at UTN
  - Self-supervised Learning
  - Multimodal Learning
  - Large Model Adaptation
- Previously: Oxford, Amsterdam; Meta, Qualcomm AI
  - More info: https://fundamentalailab.github.io/,
  - yuki.asano@utn.de