Time is Precious: Self-Supervised Learning Beyond Images

EUROPEAN CONFERENCE ON COMPUTER VISION

MILANO
2024

ECCV
EUROPEAN CONFERENCE ON COMPUTER VISION
MILANO 2024

Welcome from the organizers

Shashanka
Venkataramanan

Mohammadreza
Salehi

Yuki
Asano

valeo.ai

UNIVERSITY OF AMSTERDAM

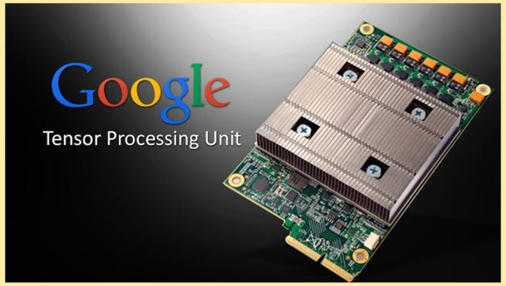
Schedule for today...

EUROPEAN CONFERENCE ON COMPUTER VISION
MILANO 2024

Schedule

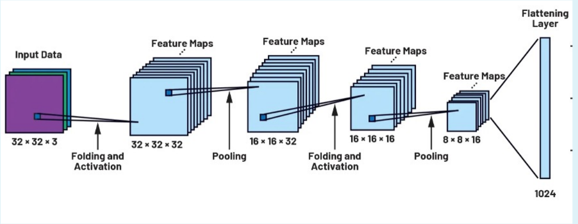
Title	Speaker	Time (CST)
Introduction	Mohammadreza	09:00 - 09:10
Part (1): Learning image encoders from videos Prior works	Shashanka	09:10 - 09:50
Part (2): New Vision Foundation Models from Video(s): 1-video pretraining, tracking image-patches	Yuki M. Asano	09:50 - 10:30
Coffee Break		10:30 - 11:00
Applications (1): Learning from one continous stream: single-stream continual learning, massively parallel video models, perceivers	João Carreira	11:00 - 11:40
Applications (2): What makes Generative video models tick? Emu Video (text-to-video), FlowVid (video-to-video), factorizing text-to-video generation, efficiency	Ishan Misra	11:40 - 12:20
Applications (3): SSL from the perspective of a developing child Audio-visual dataset, development of early word learning, learning from children	Emin Orhan	12:20 - 13:00
Conclusion, Open Problems & Final remarks	Yuki M. Asano	13:00 - 13:10

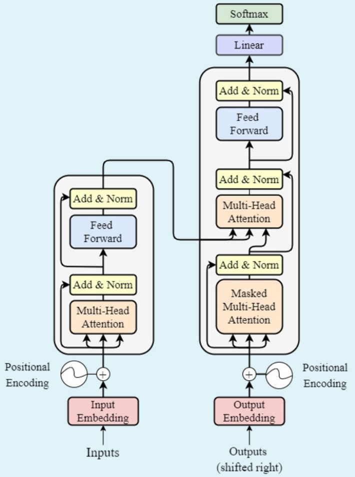
What are the main factors of AI progress?



Hardware

Interaction





Models
(MLP, CNN, Transformer)

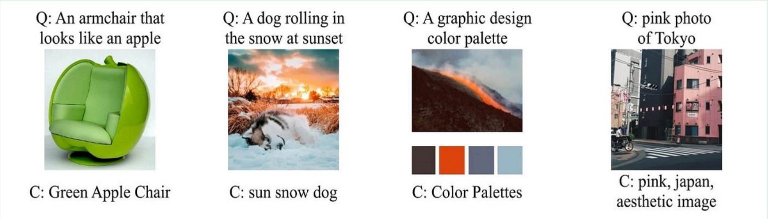


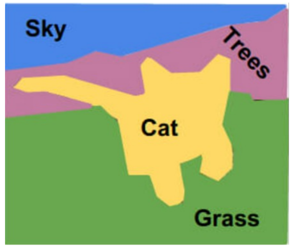
Figure 3: **LAION-5B examples.** Sample images from a nearest neighbor search in LAION-5B using CLIP embeddings. The image and caption (C) are the first results for the query (Q).

Datasets

How can we use the data?

Supervised

X : 

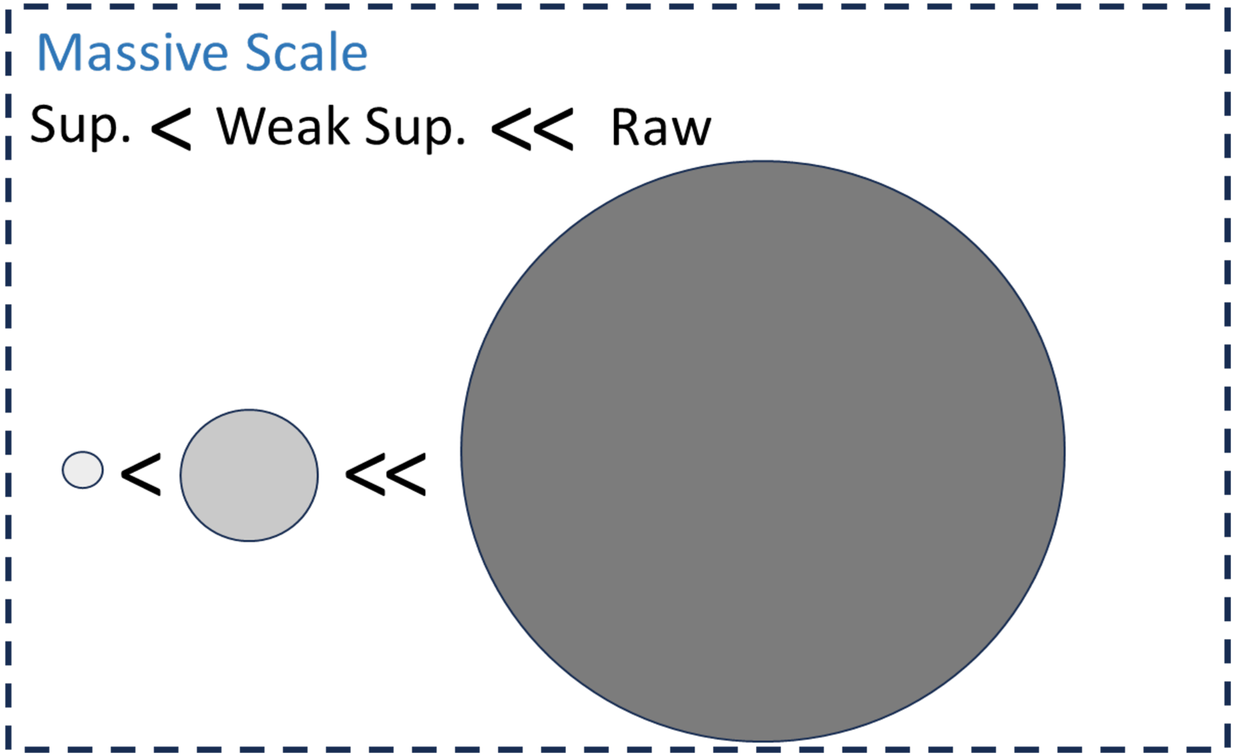
y : Cat  , 

Weakly-supervised

X : 

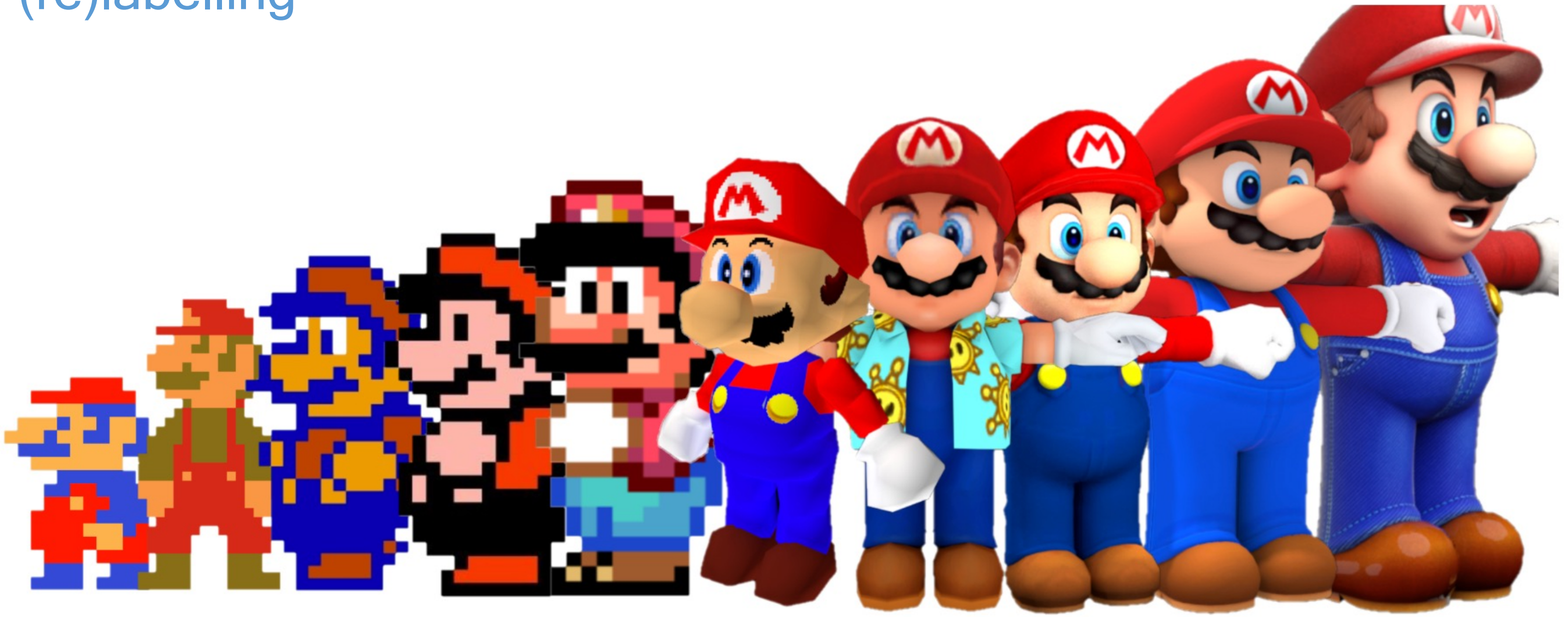
y : A playful kitten walking through a grassy field on a bright, sunny day.

Challenges of having labels



Challenges of having labels

Cost of (re)labelling



Super Mario from 1981 to 2017

Image source: Sypros Gidaris and Andrei Bursuc, Advances in self-supervised learning: Introduction

Challenges of having labels

Problem of labels

Problem of captions

Example from Flickr30k

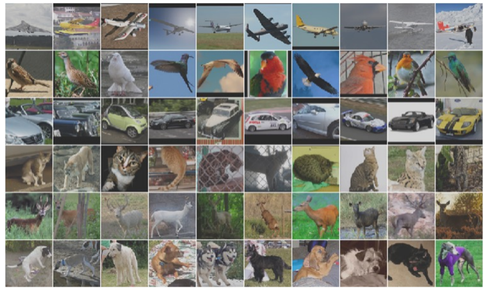
A hot, blond girl getting criticized by her boss.

Labels or captions can ignore the context

Self-supervised Learning as a solution

- Designing $f(X)$ to create y: Extracting Free Supervisory Signals from Data

Pretraining



Unlabeled data

Transformations

Rotate(x, degree)

DNN

Proxy task

predict the rotation degree

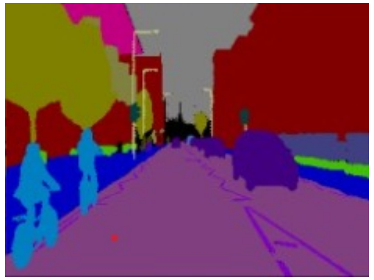
Downstream tasks

Labeled data

DNN

Target task

less labeled



What Makes Self-Supervised Learning Effective?

It needs no supervision

→ Massive scale

→ Learning general priors

→ Capturing key data features

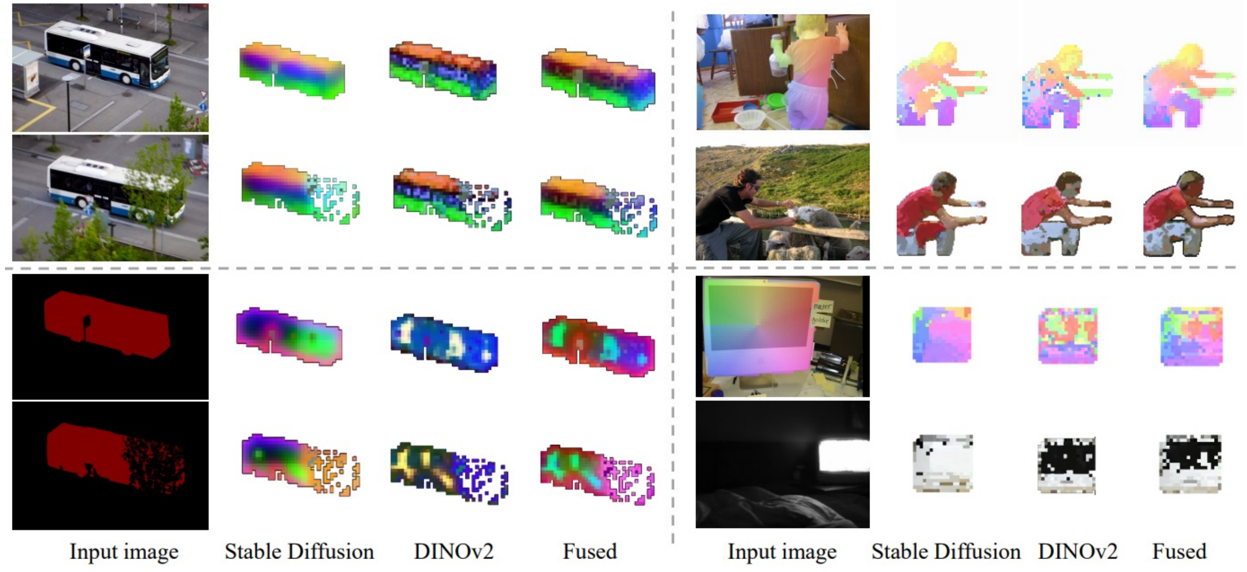
→ Transferring better to other domains

GPT, DINO, MAE, DINOv2

Applying learned priors across frameworks

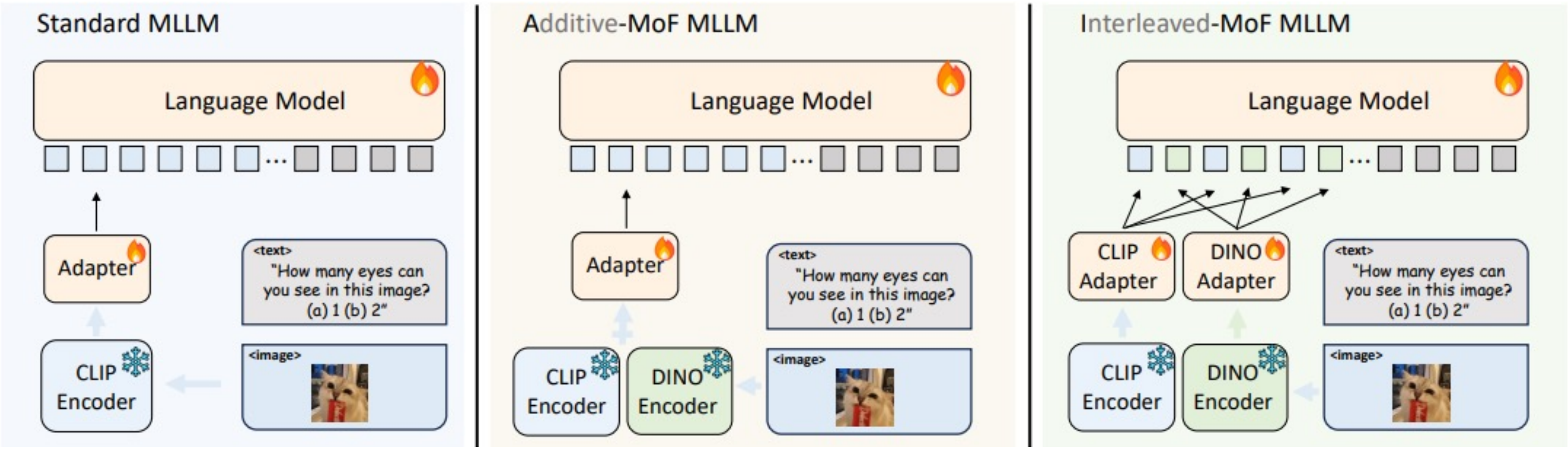
Table 3: **Evaluation on SPair-71k.** Per-class and average PCK@0.10 on test split. The methods are categorized into four types: strong supervised (S), GAN supervised (G), unsupervised with task-specific design (U^T), and unsupervised with only nearest neighboring (U^N). $*$: fine-tuned backbone. †: a trained bottleneck layer is applied on top of the features. We report *per image* PCK result for the (S) methods and *per point* result for other methods. The highest PCK among *supervised methods* and *all other methods* are highlighted in **bold**, while the second highest are underlined. Our NN-based method surpasses all previous unsupervised methods significantly.

	Method	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Dog	Horse	Motor	Person	Plant	Sheep	Train	TV	**All**
S	SCOT [34]	34.9	20.7	63.8	21.1	43.5	27.3	21.3	63.1	20.0	42.9	42.5	31.1	29.8	35.0	27.7	24.4	48.4	40.8	35.6
	CATs* [9]	52.0	34.7	72.2	34.3	49.9	57.5	43.6	66.5	24.4	63.2	56.5	52.0	42.6	41.7	43.0	33.6	72.6	58.0	49.9
	PMNC* [30]	54.1	35.9	74.9	36.5	42.1	48.8	40.0	72.6	21.1	67.6	58.1	50.5	40.1	54.1	43.3	35.7	74.5	59.9	50.4
	SCorrSAN* [24]	57.1	40.3	78.3	38.1	51.8	57.8	47.1	67.9	25.2	71.3	63.9	49.3	45.3	49.8	48.8	40.3	77.7	69.7	55.3
	CATs++* [10]	60.6	46.9	82.5	41.6	56.8	64.9	50.4	72.8	29.2	75.8	65.4	62.5	50.9	56.1	54.8	48.2	80.9	**74.9**	59.9
	DINOv2-ViT-B/14†	80.4	60.2	88.1	59.5	54.9	82.0	73.5	89.1	53.3	85.5	73.6	73.8	65.2	72.3	43.6	65.6	91.4	60.3	69.9
	Stable Diffusion† (**Ours**)	75.6	60.3	87.3	41.5	50.8	68.4	77.2	81.4	44.3	79.4	62.8	67.7	64.9	71.6	57.8	53.3	89.2	65.1	66.3
	Fuse-ViT-B/14† (**Ours**)	**81.2**	**66.9**	**91.6**	**61.4**	**57.4**	**85.3**	**83.1**	**90.8**	**54.5**	**88.5**	**75.1**	**80.2**	**71.9**	**77.9**	**60.7**	**68.9**	**92.4**	65.8	**74.6**
G	GANgealing [42]	-	37.5	-	-	-	-	-	67.0	-	-	23.1	-	-	-	-	-	-	57.9	-
U^T	VGG+MLS [1]	29.5	22.7	61.9	26.5	20.6	25.4	14.1	23.7	14.2	27.6	30.0	29.1	24.7	27.4	19.1	19.3	24.4	22.6	27.4
	DINO+MLS [1, 5]	49.7	20.9	63.9	19.1	32.5	27.6	22.4	48.9	14.0	36.9	39.0	30.1	21.7	41.1	17.1	18.1	35.9	21.4	31.1
	NeuCongeal [39]	-	29.1	-	-	-	-	-	53.3	-	-	35.2	-	-	-	-	-	-	-	-
	ASIC [18]	57.9	25.2	68.1	24.7	35.4	28.4	30.9	54.8	21.6	45.0	47.2	39.9	26.2	48.8	14.5	24.5	49.0	24.6	36.9
U^N	DINOv1-ViT-S/8 [2]	57.2	24.1	67.4	24.5	26.8	29.0	27.1	52.1	15.7	42.4	43.3	30.1	23.2	40.7	16.6	24.1	31.0	24.9	33.3
	DINOv2-ViT-B/14	72.7	62.0	85.2	**41.3**	40.4	52.3	51.5	71.1	36.2	67.1	64.6	67.6	61.0	68.2	30.7	62.0	54.3	24.2	55.6
	Stable Diffusion (**Ours**)	63.1	55.6	80.2	33.8	44.9	49.3	47.8	74.4	38.4	70.8	53.7	61.1	54.4	55.0	54.8	53.5	65.0	53.3	57.2
	Fuse-ViT-B/14 (**Ours**)	**73.0**	**64.1**	**86.4**	40.7	**52.9**	**55.0**	**53.8**	**78.6**	**45.5**	**77.3**	**64.7**	**69.7**	**63.3**	**69.2**	**58.4**	**67.6**	**66.2**	**53.5**	**64.0**



A Tale of Two Features: Stable Diffusion Complements DINO for Zero-Shot Semantic Correspondence

Applying learned priors across frameworks



Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs

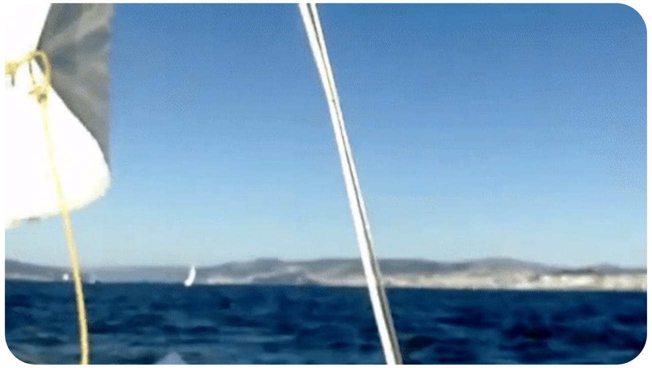
method	res	#tokens	MMVP	LLaVA	POPE
LLaVA	224^2	256	5.5	81.8	50.0
LLaVA	336^2	576	6.0	81.4	50.1
LLaVA + I-MoF	224^2	512	16.7 (+10.7)	82.8	51.0
LLaVA$^{1.5}$	336^2	576	24.7	84.7	85.9
LLaVA$^{1.5}$ + I-MoF	224^2	512	28.0 (+3.3)	82.7	86.3

Table 3. **Empirical Results of Interleaved MoF.** Interleaved MoF improves visual grounding while maintaining same level of instruction following ability.

How can we learn more real-world priors ?

Videos open exciting new directions

Visual Development



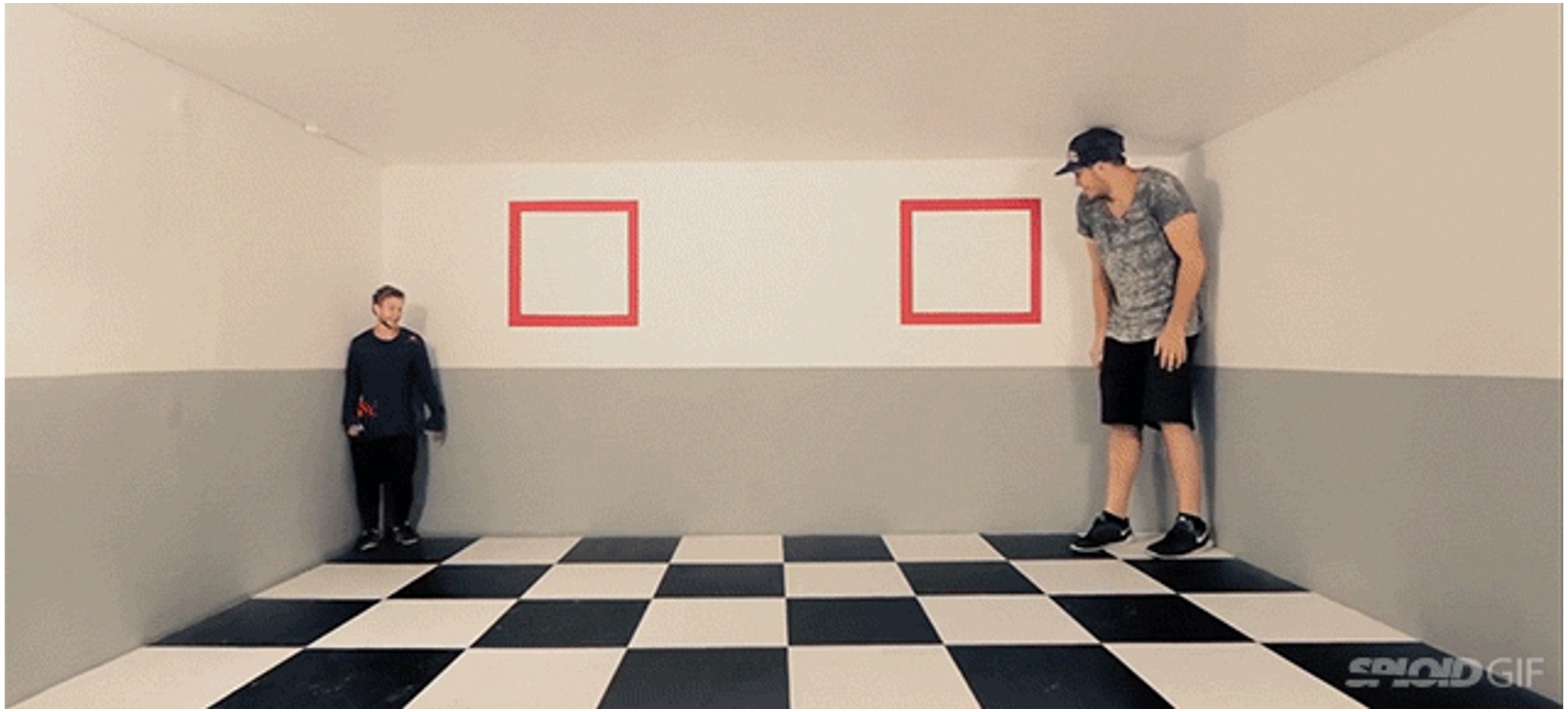
Understanding physics

Embodied AI

I've made that point before:
- LLM: 1E13 tokens x 0.75 word/token x 2 bytes/token = 1E13 bytes.
- 4 year old child: 16k wake hours x 3600 s/hour x 1E6 optical nerve fibers x 2 eyes x 10 bytes/s = 1E15 bytes.

In 4 years, a child has seen 50 times more data than the biggest LLMs.

Seeing is believing, but watching is understanding.



Ames room illusion

Seeing is believing, but watching is understanding.

Ames room illusion

Seeing is believing, but watching is understanding.

Checker shadow illusion