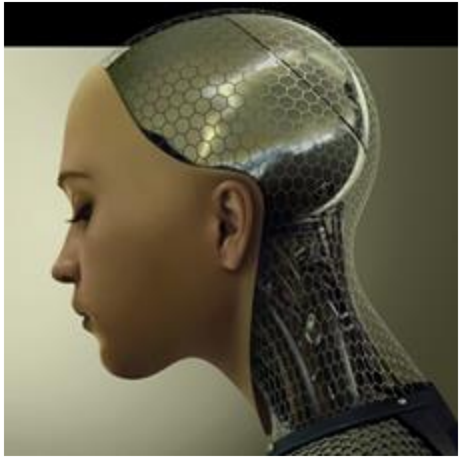
AVA: A Video Dataset of Atomic Visual Actions



CVPR 2018

Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik

Answer phone






Climb (e.g., a mountain)

Clink glass

Dig






Fall down

Give/Serve (an object) to (a person)

Hug (a person)

run/jog
walk
jump
stand
sit
lie/sleep
bend/bow
crawl
swim
dance
get up
fall down
crouch/kneel
martial art

Pose (14)

talk to
watch
listen to
sing to
kiss
hug
grab
lift
kick
give/serve … to …
take … from …
play with kids
hand shake
hand clap
hand wave
fight/hit
push

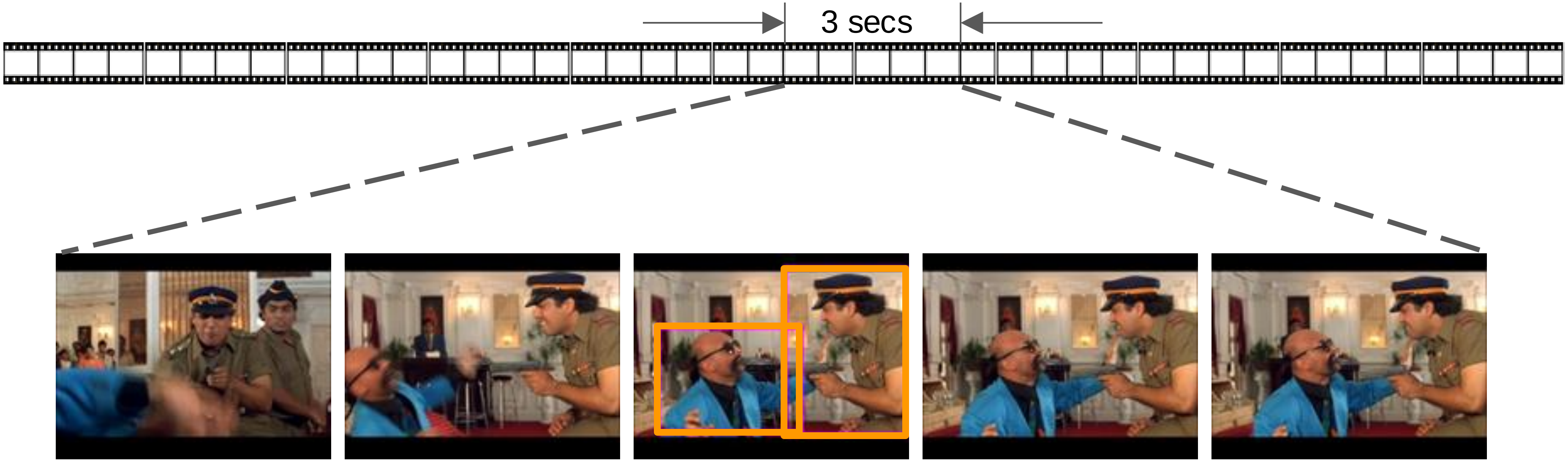
Person-person (17)

lift/pick up
put down
carry
hold
throw
catch
eat
drink
cut
hit
stir
press
extract
read
write
smoke
sail boat
row boat
fishing
touch
cook
kick
paint
dig
shovel
chop
shoot
take a photo
brush teeth
clink glass
work on a computer
answer phone
climb (e.g., mountain)
play board game
play with pets
drive (e.g., a car)
push (an object)
pull (an object)
point to (an object)
play musical instrument
text on/look at a cellphone
turn (e.g., screwdriver)
dress / put on clothing
ride (e.g., bike, car, horse)
watch (e.g., TV)
open
close
enter
exit

Person-object (49)

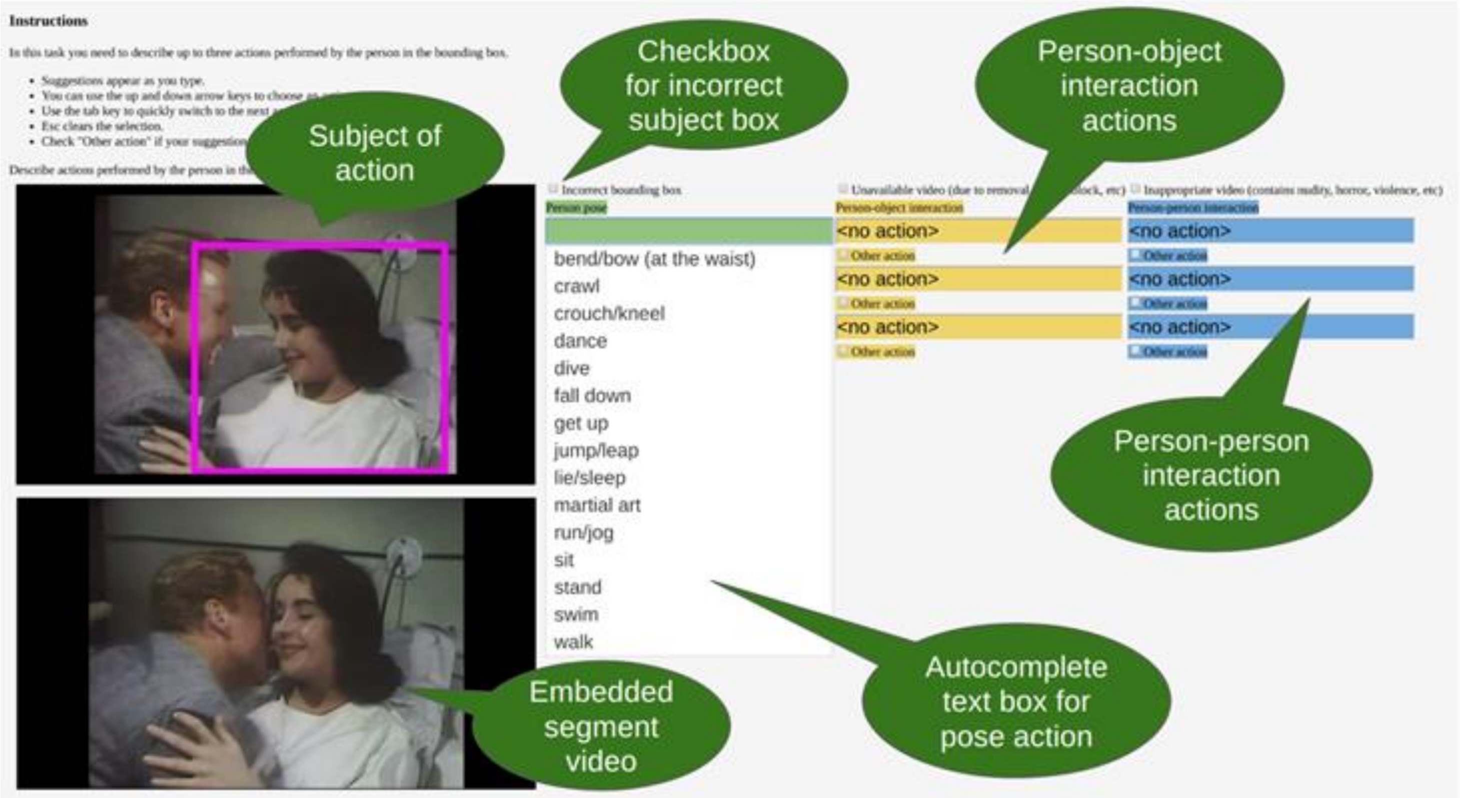
Label Annotation

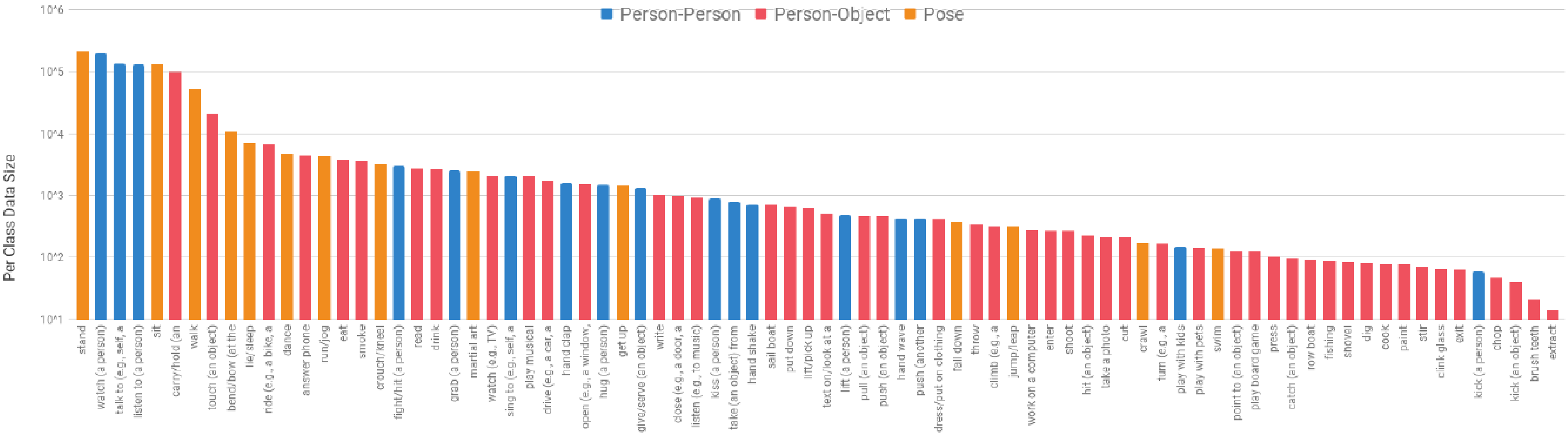
Annotation Goal



Left: Kneel, Talk to
Right: Stand, Listen, Shoot

User Interface for Action Selection





Person-Person
Person-Object
Pose
10^6
10^5
10^4
10^3
10^2
10^1
Per Class Data Size
stand
watch (a person)
talk to (e.g., self, a
listen to (a person)
sit
carry/hold (an
walk
touch (an object)
bend/bow (at the
lie/sleep
ride (e.g., a bike, a
dance
answer phone
run/jog
eat
smoke
crouch/kneel
fight/hit (a person)
read
drink
grab (a person)
martial art
watch (e.g., TV)
sing to (e.g., self, a
play musical
drive (e.g., a car, a
hand clap
open (e.g., a window,
hug (a person)
get up
give/serve (an object)
write
close (e.g., a door, a
listen (e.g., to music)
kiss (a person)
take (an object) from
hand shake
sail boat
put down
lift/pick up
text on/look at a
lift (a person)
pull (an object)
push (an object)
hand wave
push (another
dress/put on clothing
fall down
throw
climb (e.g., a
jump/leap
work on a computer
enter
shoot
hit (an object)
take a photo
cut
crawl
turn (e.g., a
play with kids
play with pets
swim
point to (an object)
play board game
press
catch (an object)
row boat
fishing
shovel
dig
cook
paint
stir
clink glass
exit
kick (a person)
chop
kick (an object)
brush teeth
extract

Around the World in 3,000 Hours of Egocentric Video

Jitendra Malik

Meta AI / UC Berkeley

This talk based on slides from Kristen Grauman and the Ego4D team

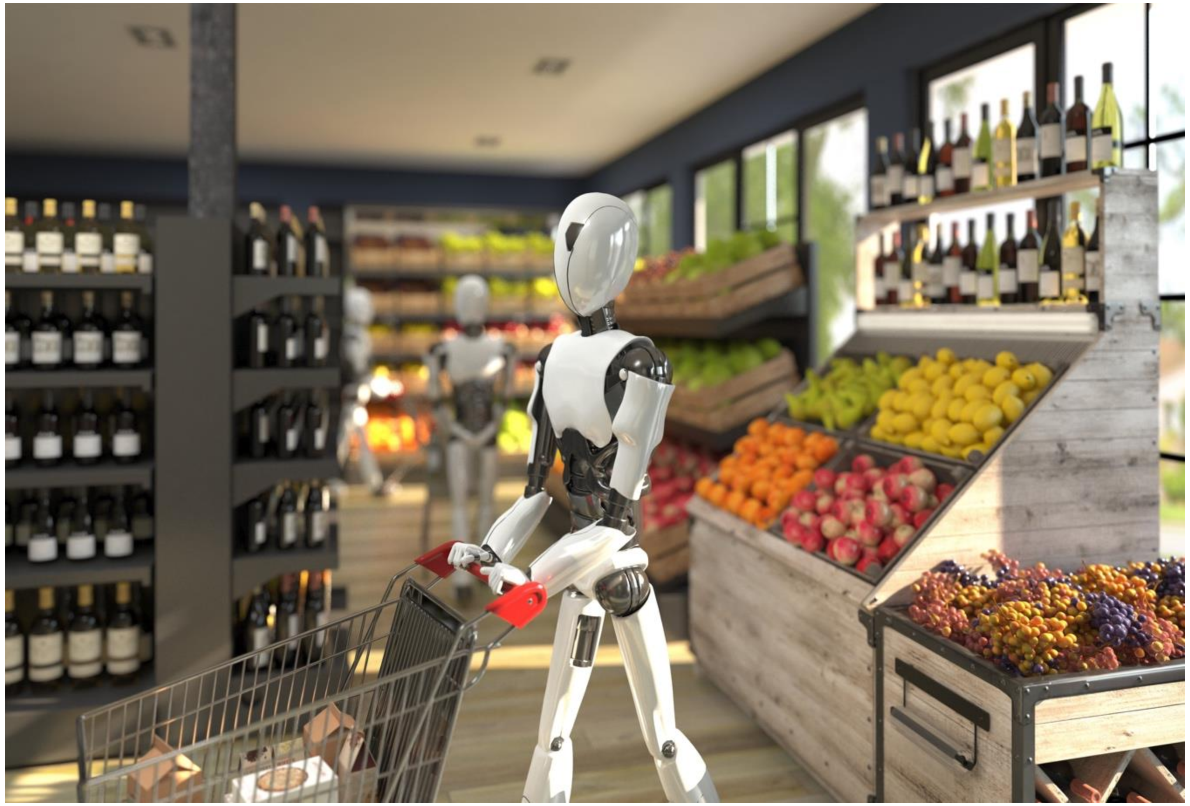
Ego4D team

Ego4D: Around the World in 3,000 Hours of Egocentric Video

Kristen Grauman[1,2], Andrew Westbury[1], Eugene Byrne[*1], Zachary Chavis[*3], Antonino Furnari[*4], Rohit Girdhar[*1], Jackson Hamburger[*1], Hao Jiang[*5], Miao Liu[*6], Xingyu Liu[*7], Miguel Martin[*1], Tushar Nagarajan[*1,2], Ilija Radosavovic[*8], Santhosh Kumar Ramakrishnan[*1,2], Fiona Ryan[*6], Jayant Sharma[*3], Michael Wray[*9], Mengmeng Xu[*10], Eric Zhongcong Xu[*11], Chen Zhao[*10], Siddhant Bansal[17], Dhruv Batra[1], Vincent Cartillier[1,6], Sean Crane[7], Tien Do[3], Morrie Doulaty[13], Akshay Erapalli[13], Christoph Feichtenhofer[1], Adriano Fragomeni[9], Qichen Fu[7], Christian Fuegen[13], Abrham Gebreselasie[12], Cristina González[14], James Hillis[5], Xuhua Huang[7], Yifei Huang[15], Wenqi Jia[6], Weslie Khoo[16], Jachym Kolar[13], Satwik Kottur[13], Anurag Kumar[5], Federico Landini[13], Chao Li[5], Yanghao Li[1], Zhenqiang Li[15], Karttikeya Mangalam[1,8], Raghava Modhugu[17], Jonathan Munro[9], Tullie Murrell[1], Takumi Nishiyasu[15], Will Price[9], Paola Ruiz Puentes[14], Merey Ramazanova[10], Leda Sari[5], Kiran Somasundaram[5], Audrey Southerland[6], Yusuke Sugano[15], Ruijie Tao[11], Minh Vo[5], Yuchen Wang[16], Xindi Wu[7], Takuma Yagi[15], Yunyi Zhu[11], Pablo Arbeláez[†14], David Crandall[†16], Dima Damen[†9], Giovanni Maria Farinella[†4], Bernard Ghanem[†10], Vamsi Krishna Ithapu[†5], C. V. Jawahar[†17], Hanbyul Joo[†1], Kris Kitani[†7], Haizhou Li[†11], Richard Newcombe[†5], Aude Oliva[†18], Hyun Soo Park[†3], James M. Rehg[†6], Yoichi Sato[†15], Jianbo Shi[†19], Mike Zheng Shou[†11], Antonio Torralba[†18], Lorenzo Torresani[†1,20], Mingfei Yan[†5], Jitendra Malik[1,8]

[1]Facebook AI Research (FAIR), [2]University of Texas at Austin, [3]University of Minnesota, [4]University of Catania, [5]Facebook Reality Labs, [6]Georgia Tech, [7]Carnegie Mellon University, [8]UC Berkeley, [9]University of Bristol, [10]King Abdullah University of Science and Technology, [11]National University of Singapore, [12]Carnegie Mellon University Africa, [13]Facebook, [14]Universidad de los Andes, [15]University of Tokyo, [16]Indiana University, [17]International Institute of Information Technology, Hyderabad, [18]MIT, [19]University of Pennsylvania, [20]Dartmouth

Why egocentric video?
Robot learning



Robots that can learn from video how to manipulate human-centric objects and navigate in human-centric spaces

EGO 4D

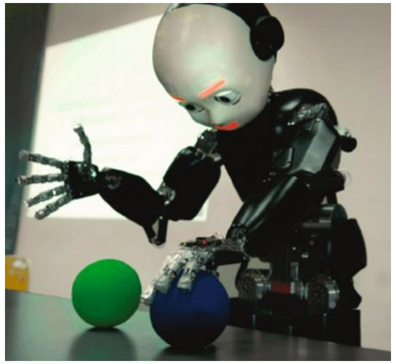
First-person perception and learning

Status quo:

Learning and inference with “disembodied” images/videos.

On the horizon:

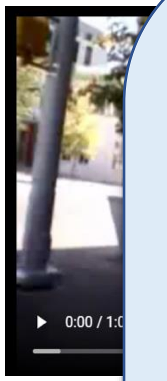
Visual learning in the context of agent goals, interaction, and multi-sensory observations.



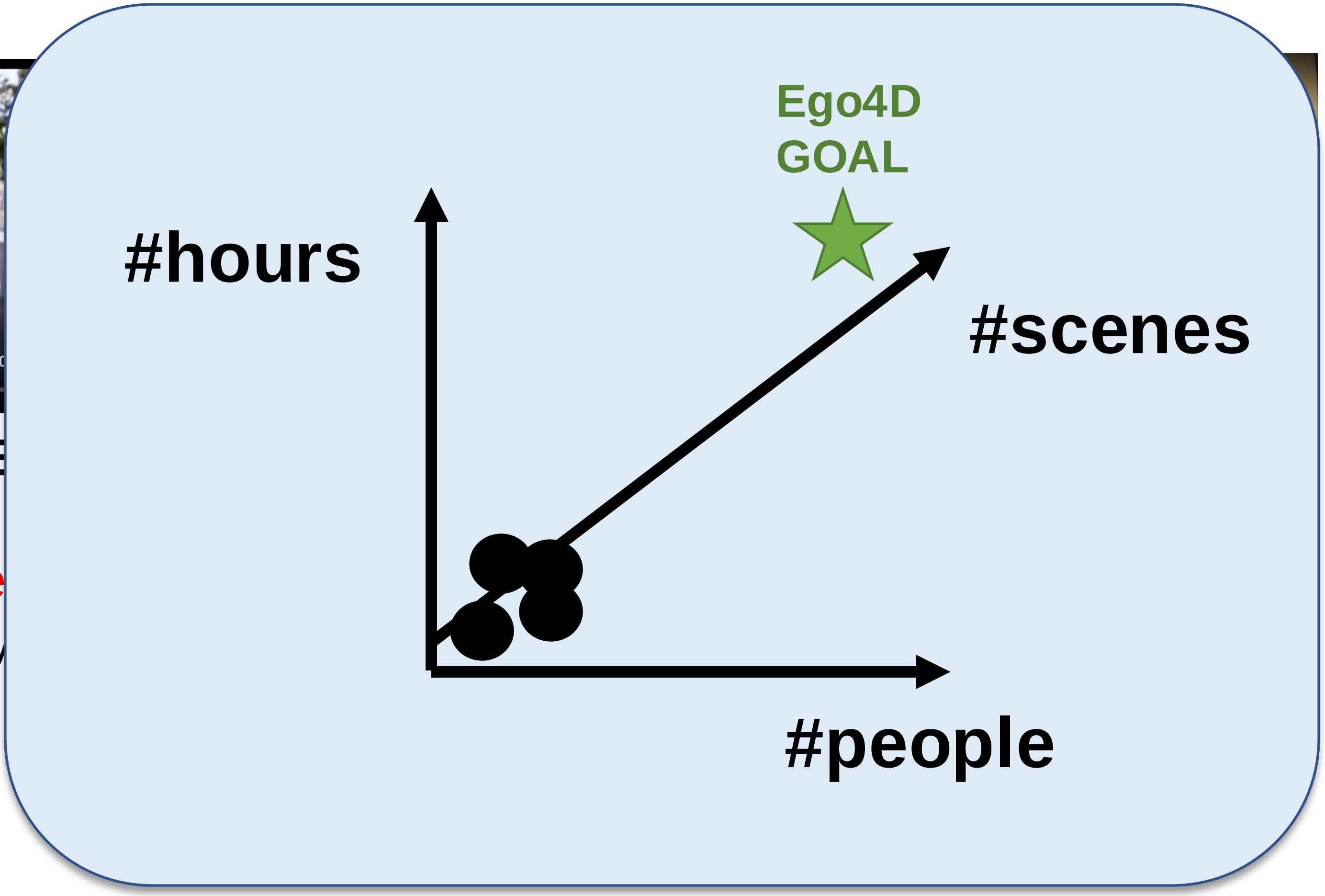
Existing first-person video datasets

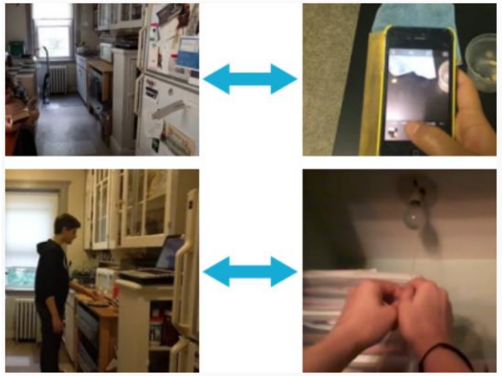
Inspire this effort, but call for greater scale, content, diversity

EPIC Kitchens
Damen et al. 2020
45 people, 100 hrs
kitchens only



UT E
Lee
4 pe
daily





Charades-Ego
Sigurdsson 2018
71 people, 34 hrs
indoor

Ego4D: a new massive egocentric video dataset

Goal: Large-scale "in the wild" first-person video dataset

Catalyze research in multimodal egocentric perception

Content:

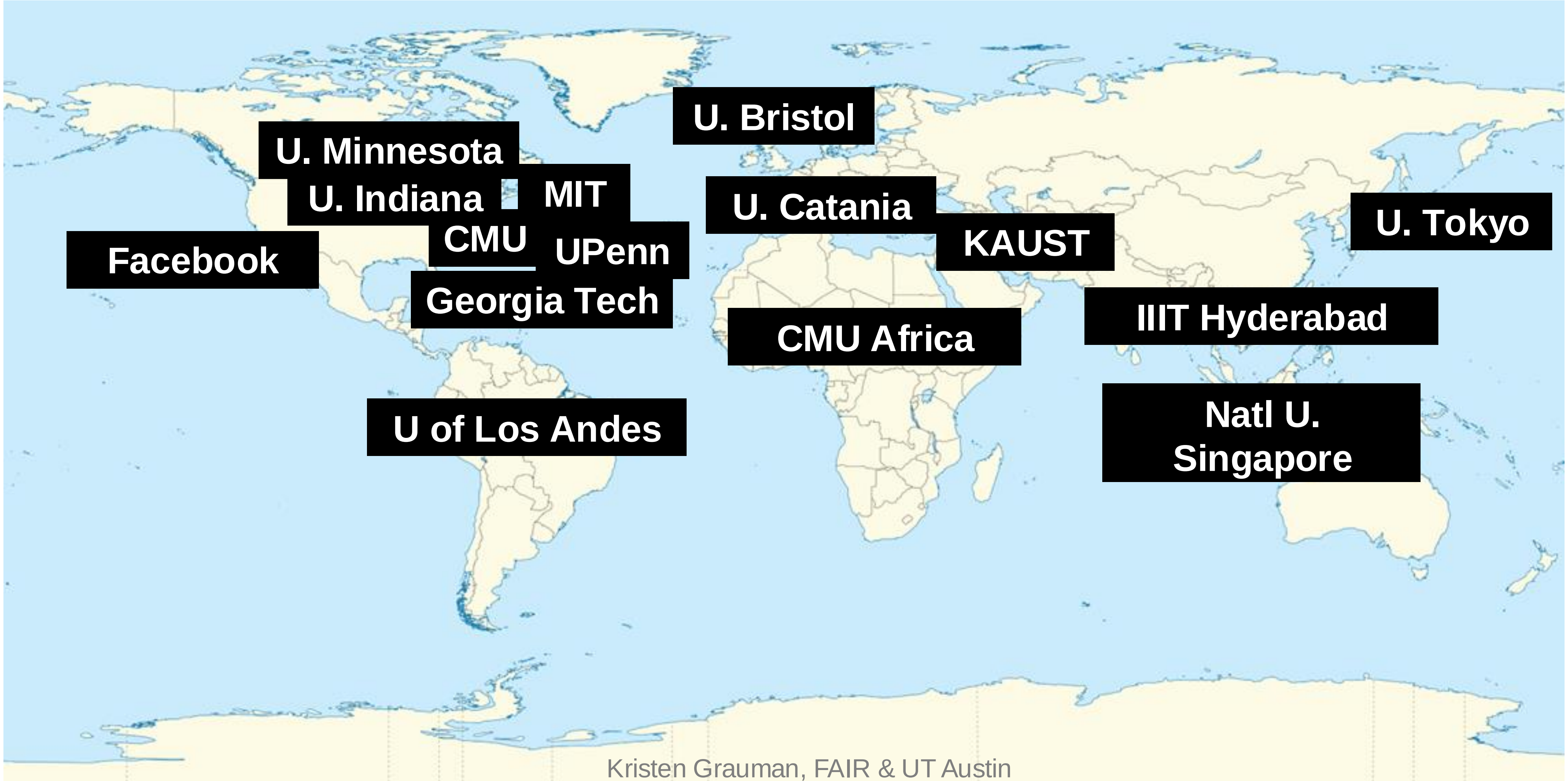
- 3,025 hours of video from 74 cities & 9 countries
- 855 unique camera wearers – not just graduate students!
- Daily life activities – work, home, shopping, commute, street
- Multi-modal sensing: audio, 3D scans, IMU, stereo, multi-camera
- Benchmark challenge for the research community

Timeline:

- Collection began early 2020
- Paper released last week, data will be released late Nov 2021

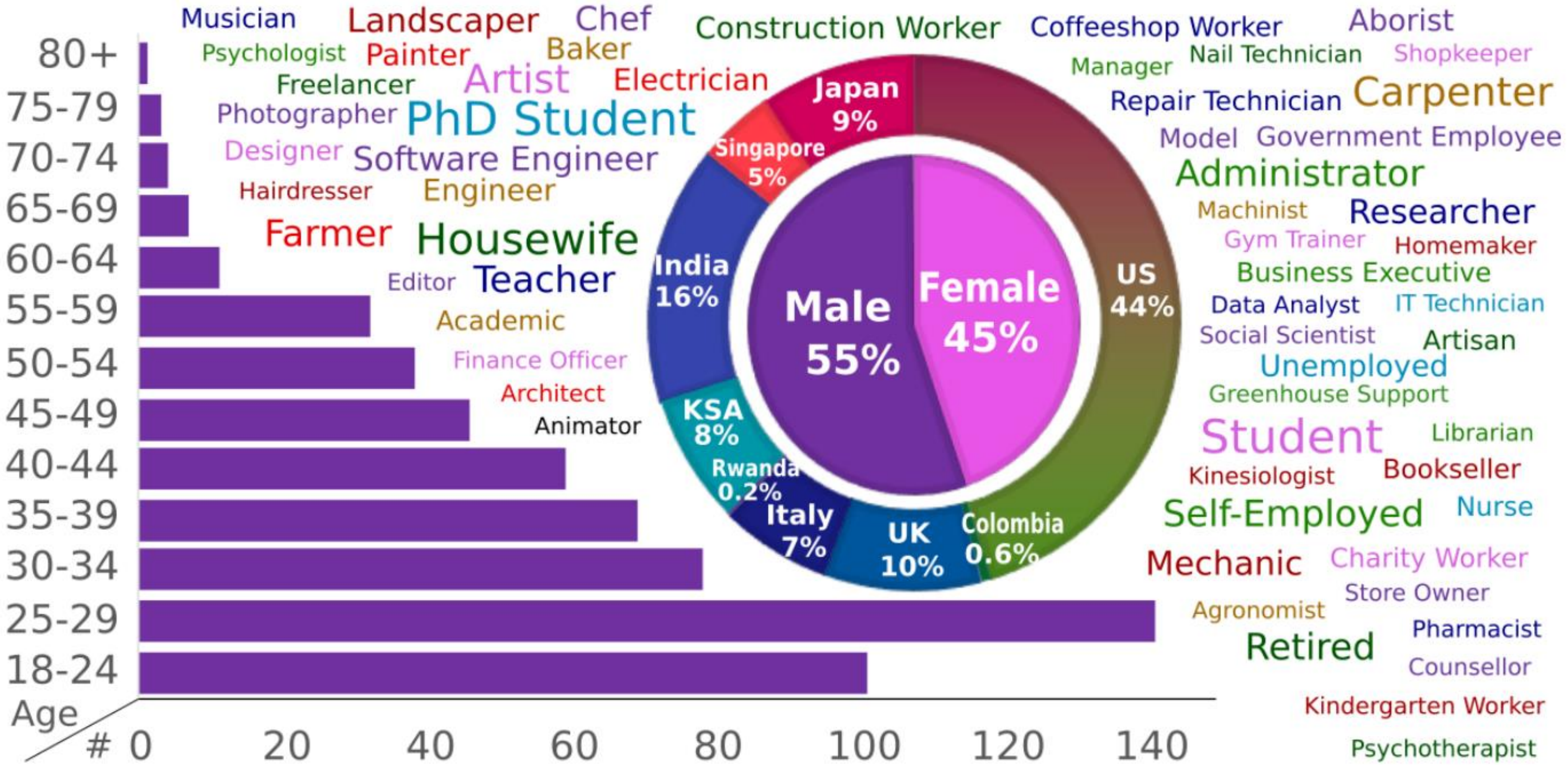
Ego4D consortium

Towards geographically diverse ego-video coverage



855 camera wearers

Self-reported demographics and countries of residence



Daily-life scenarios

How people spend their days: US Bureau of Labor Statistics

Everyday activities in the home:

- Sleeping
- Daily hygiene
- Doing hair/make-up
- Cleaning / laundry
- Cooking
- Talking with family me
- Hosting a party
- Eating
- Yardwork / shoveling
- Household managem for kids
- Fixing something in th
- Playing with pets
- Crafting/knitting/sewing/drawing/ painting/etc

Errands

- Grocery shopping
- Clothes, shopping
- Getting car fixed
- Attending a lecture/class
- Writing on whiteboard
- Video call
- Eating at the cafeteria
- Making coffee

Entertainment/Leisure

- Watching movies at cinema
- Watching tv
- Reading books
- BBQ'ing/picnics
- Going to a salon (nail, hair, spa)
- Getting a tattoo / piercing
- Volunteering
- Practicing a musical instrument
- Attending a festival or fair

Exercise:

- Going to the gym
- Yoga practice
- Swimming in a pool/ocean
- Working out at home
- Cycling / jogging
- Dancing
- Working out outside
- Walking on street
- Going to the park
- Hiking

ansportation:

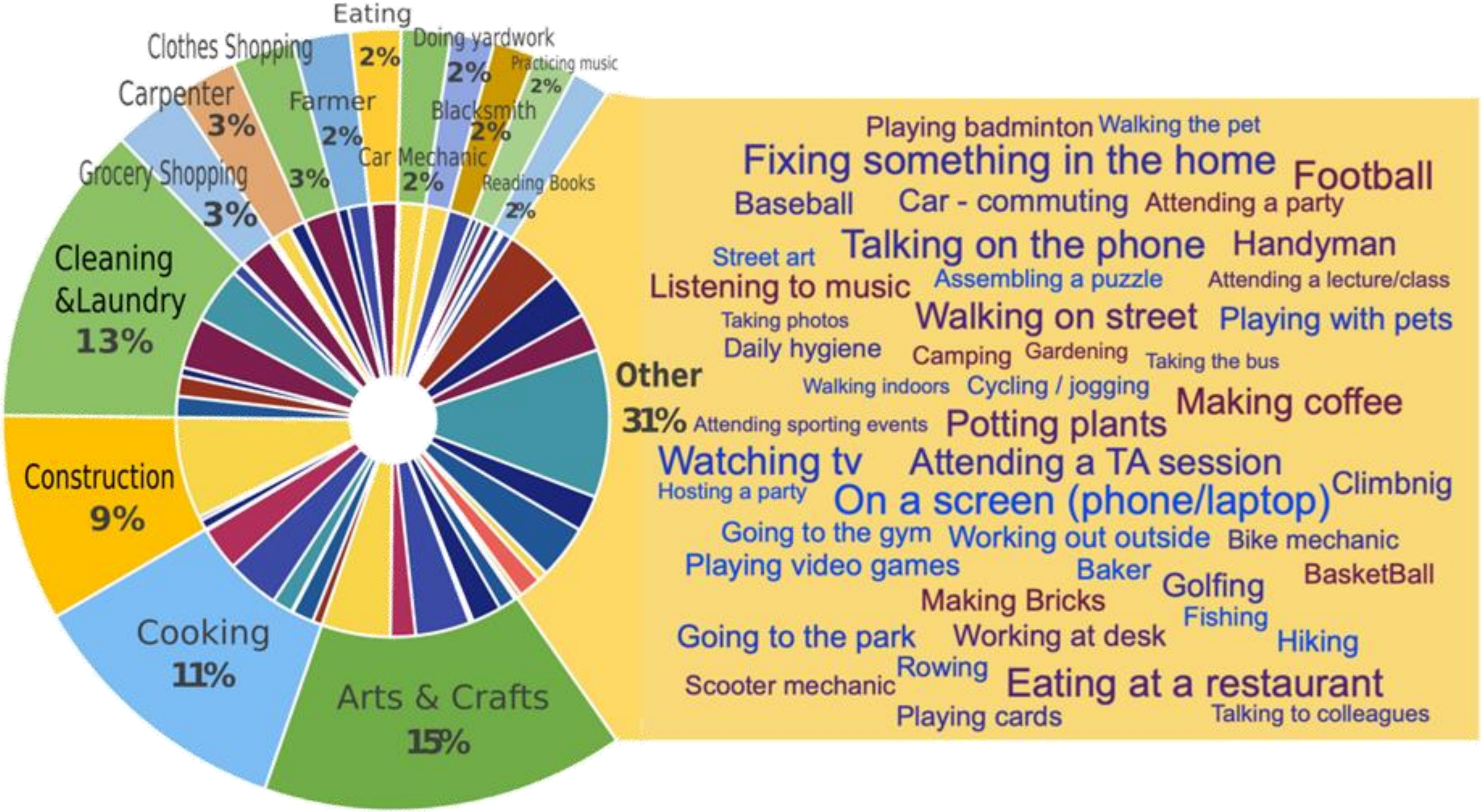
- Car - commuting, road trip
- Bus
- Train
- Airplane
- Bike
- Skateboard/scooter

https://www.bls.gov/news.release/atus.nr0.htm

Key tenet in Ego4D:
Capture unscripted, daily-life activity

Daily-life scenarios

Wide variety of activity in the home, workplace, outdoors, errands



Kristen Grauman, FAIR & UT Austin

Wearable cameras

GoPro

Vuzix Blade

Pupil Labs

ZShade

WeeView

We deploy a variety of head-mounted cameras.

Privacy and ethics

- Mobilized leading experts in first-person video capture, with expertise in privacy, de-identification, and responsible in-the-wild data collection
- Each partner underwent separate months-long IRB review process, overseeing ethical and privacy standards for data collection, management, and informed consent.
- Consent forms signed by all recorded people where relevant
- State-of-the-art de-identification processes, featuring both automated and manual reviews for faces, screens, credit cards, and other identifiers

Ego4D data: everyday activity around the world

- 3,025+ hours of video
- 855+ camera wearers
- Geographic diversity
- Occupational diversity
- Unscripted daily life activity

Ego4D data: everyday activity around the world

- 3,025+ hours of video
- 855+ camera wearers
- Geographic diversity
- Occupational diversity
- Unscripted daily life activity

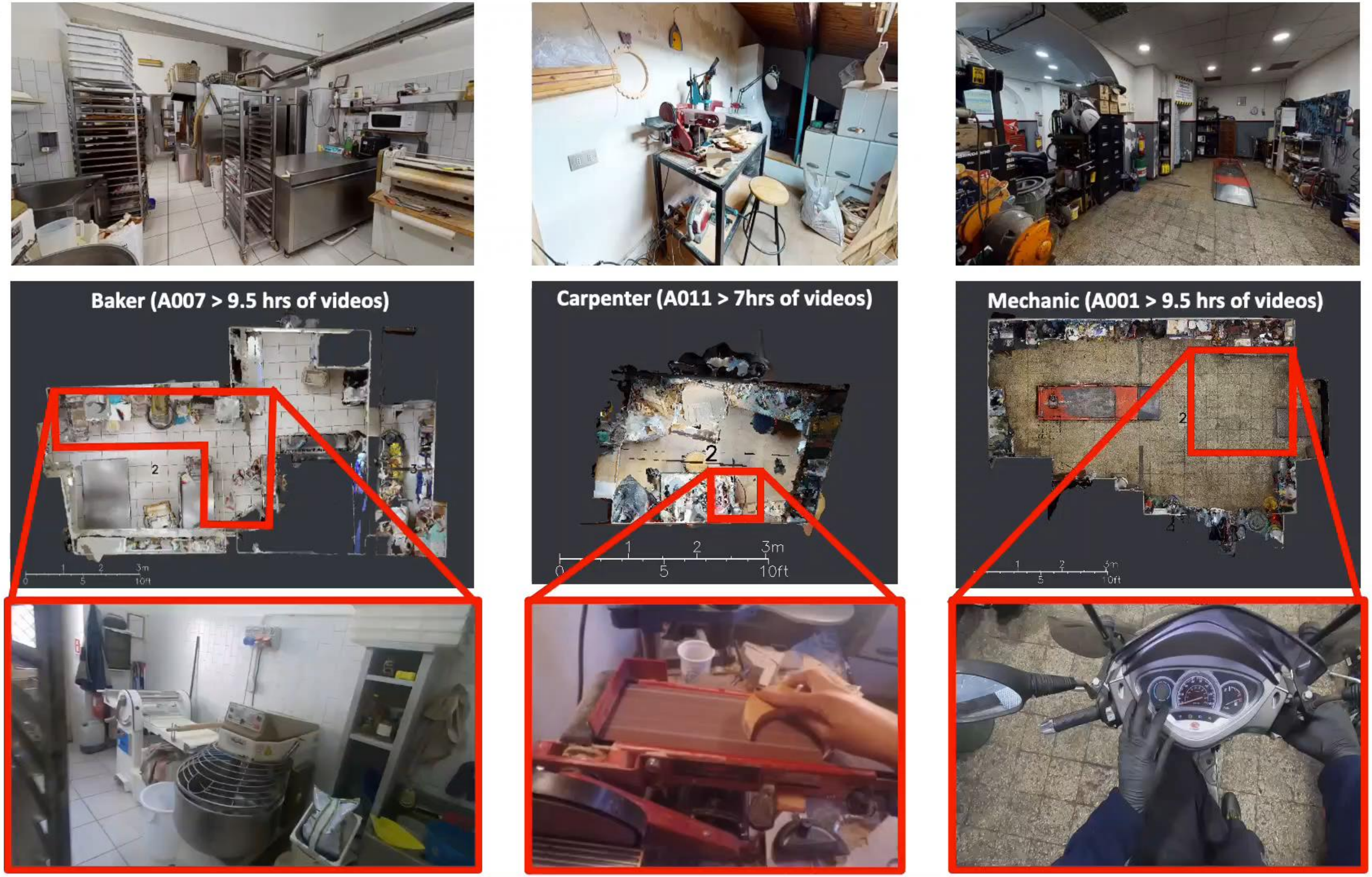
Ego4D data: 3D environment scans

EGO4D@UNICT Examples

3D

FloorPlan

EGO



Available for 491 hours of video

Ego4D data: multi-camera and eye gaze

Multiple simultaneous egocentric cameras

Eye gaze

Ego4D benchmark suite

Past

Episodic Memory

"where is my X?"

Present

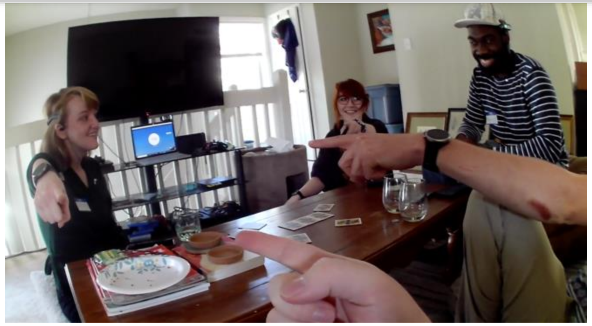
Hands & Objects

"what am I doing and how?"

Present

Audio-visual Diarization

"who said what when?"



Social Interaction

"who is attending to whom?"

Future

Forecasting

"what will I do next?"

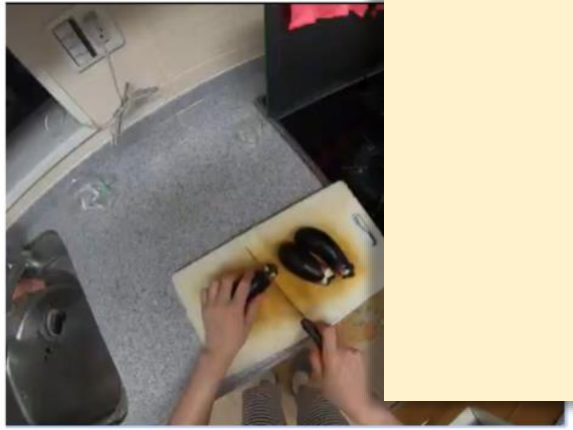
Annotation: text narrations

Dense descriptive text of each camera wearer activity + clip-level summaries

13 sentences per minute

4M sentences on 3,025 hours

Annotation: benchmarks



What did I put in the drawer?

More than 250,000 hours of annotator effort

ns (text)

emporal labels

nts

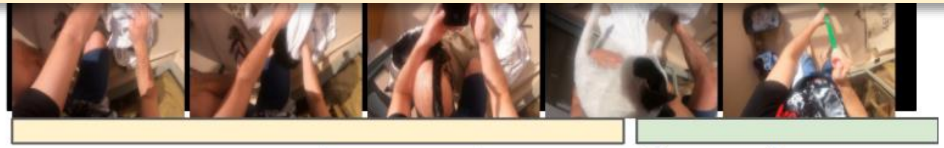
s

objects

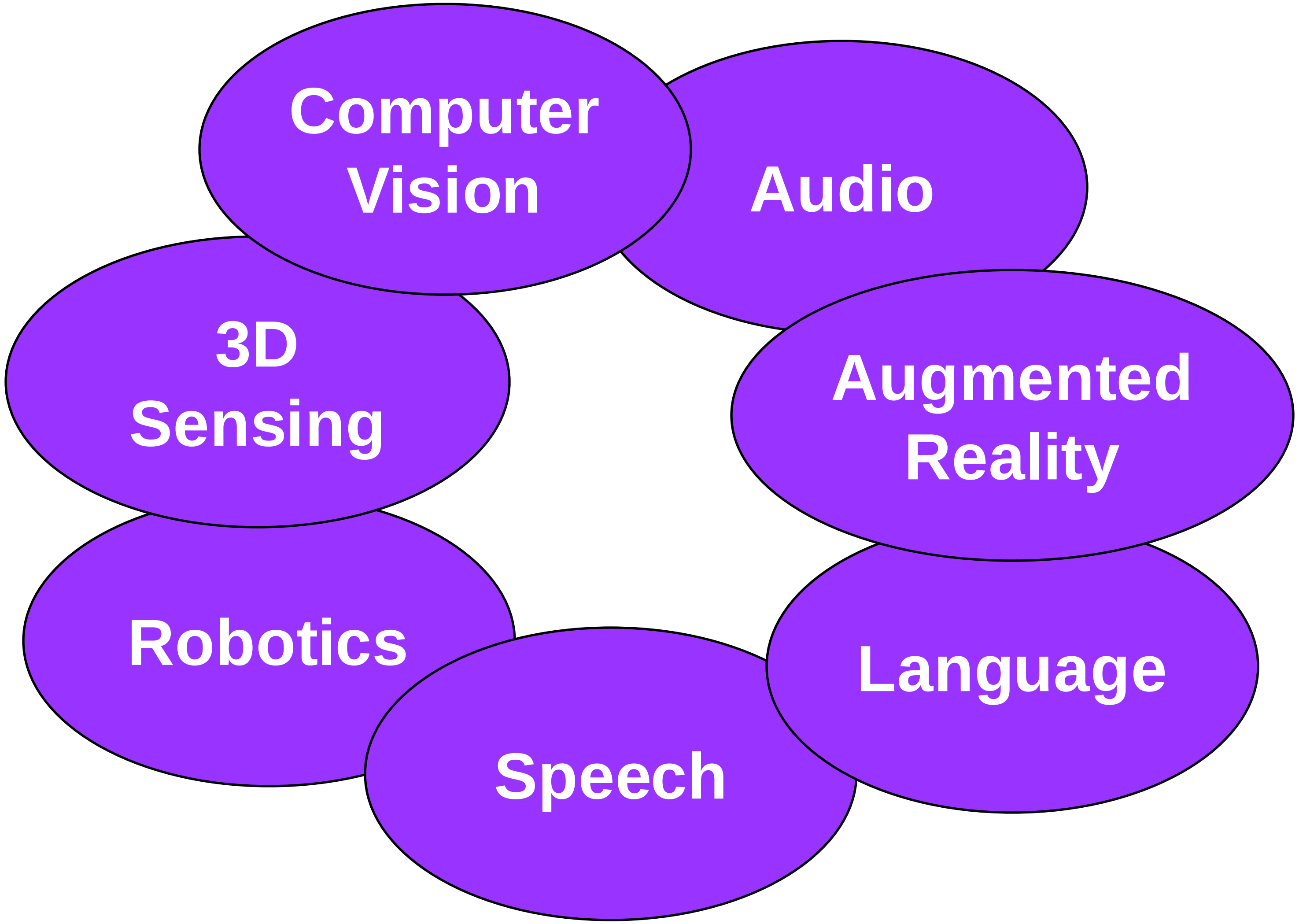
changes

odal labels

- Query+response
- Speech transcript



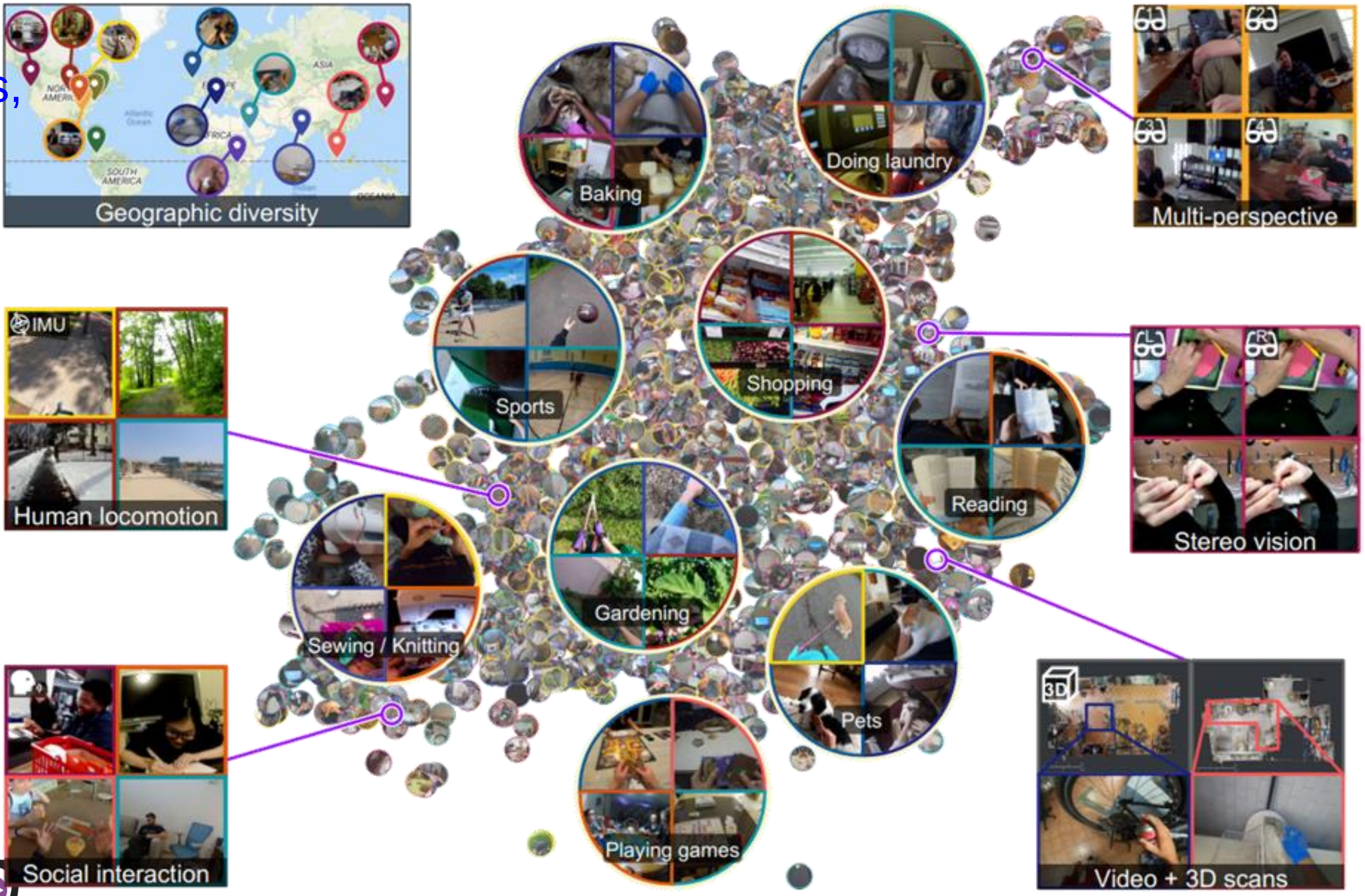
Ego4D: for vision and beyond



9 countries,
74 cities

IMU
836 hrs

Audio
2,207 hrs



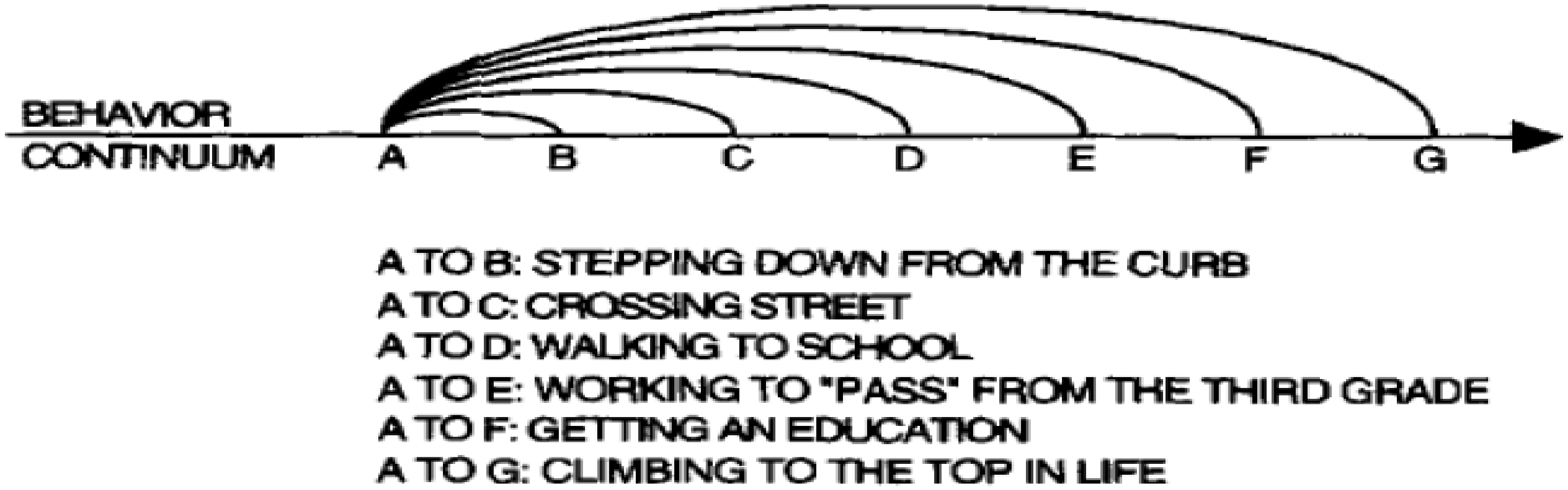
Multiple
synchronized
ego-cameras
224 hrs

Stereo
80 hrs

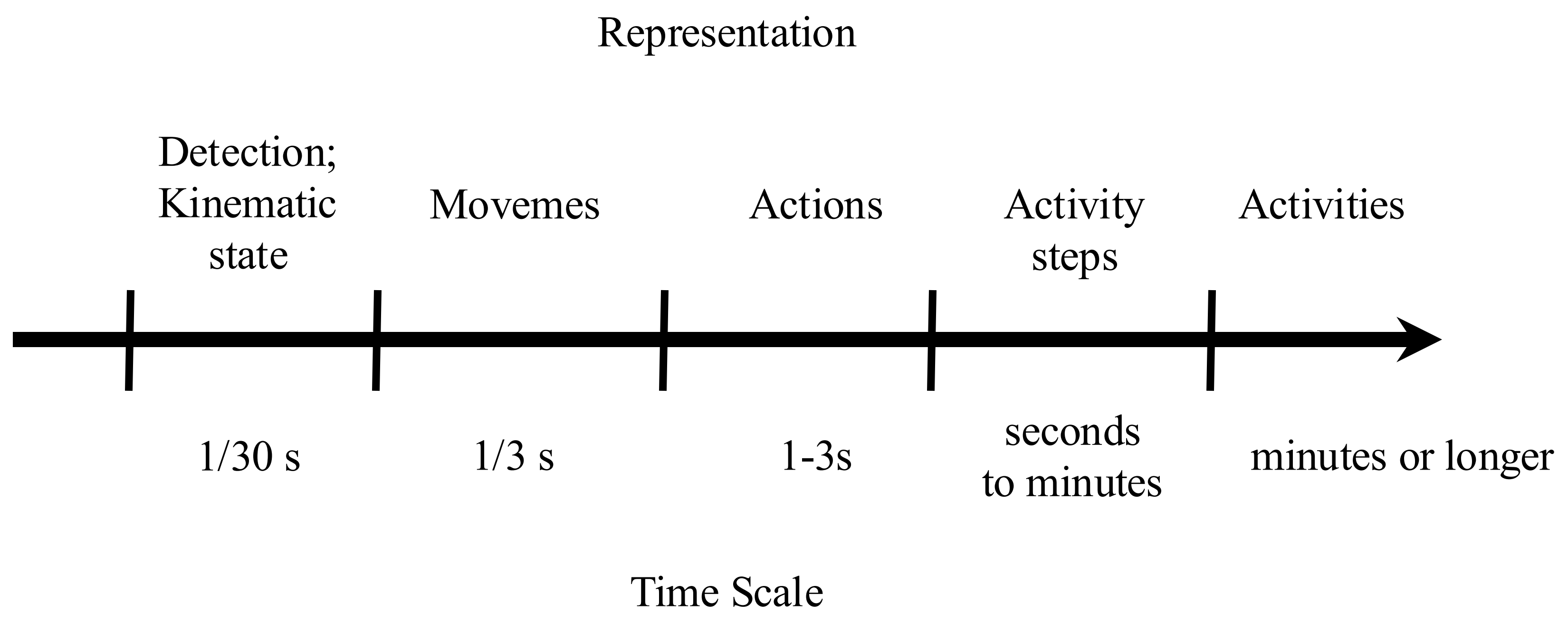
3D
environment
scans
491 hrs

What we need to understand video

- The hierarchical structure of human behavior- movement, goals, actions and events e.g. Barker and Wright (1954).



Temporal scales of human motion

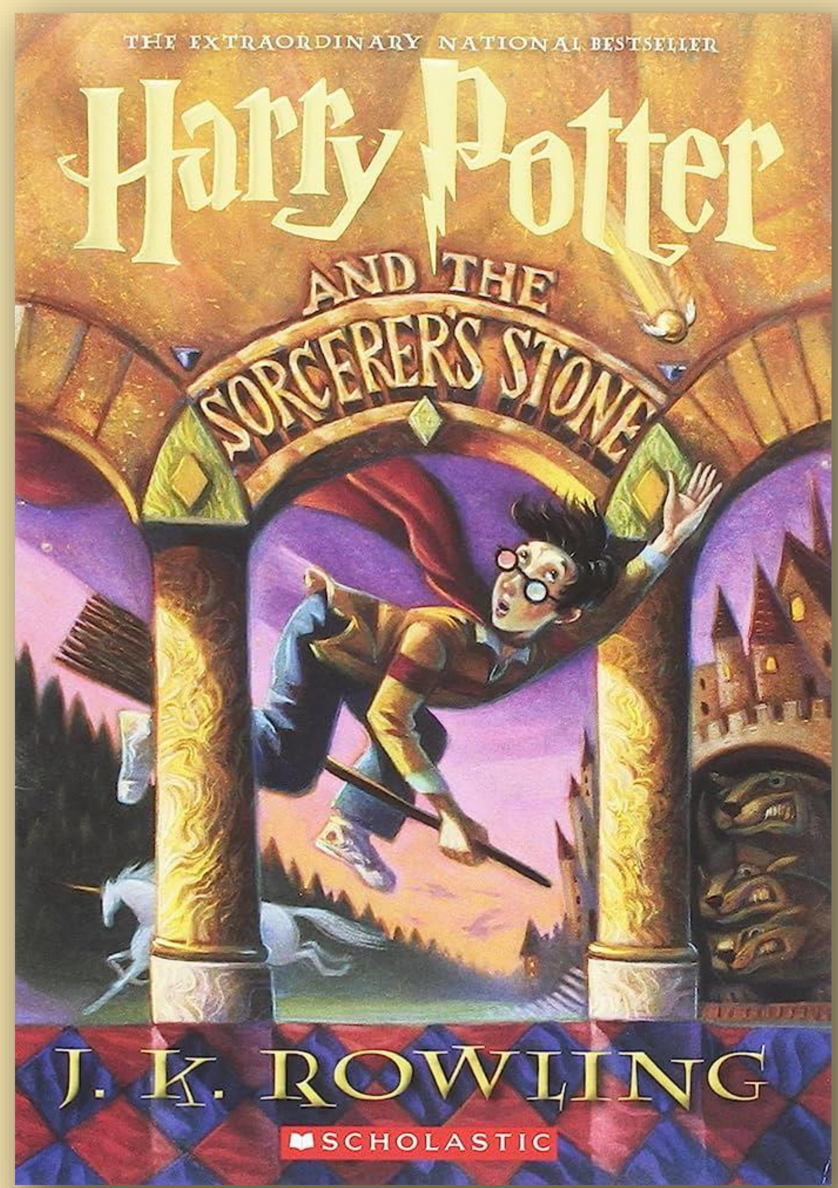


Representations with more complex semantics are at longer time scales

The multiple spatiotemporal scales

- The finest scale can be understood as physical action, but the larger scales are best understood in terms of goals and intentionality

ACTION = MOVEMENT + GOAL



What position does Harry play on the Quidditch team?

- Answer: Seeker

What does the Sorcerer's Stone do?

- Answer: It produces the Elixir of Life which makes the drinker immortal, and can also turn any metal into pure gold.

What is the name of the Mirror which had the Sorcerer’s Stone hidden within it?

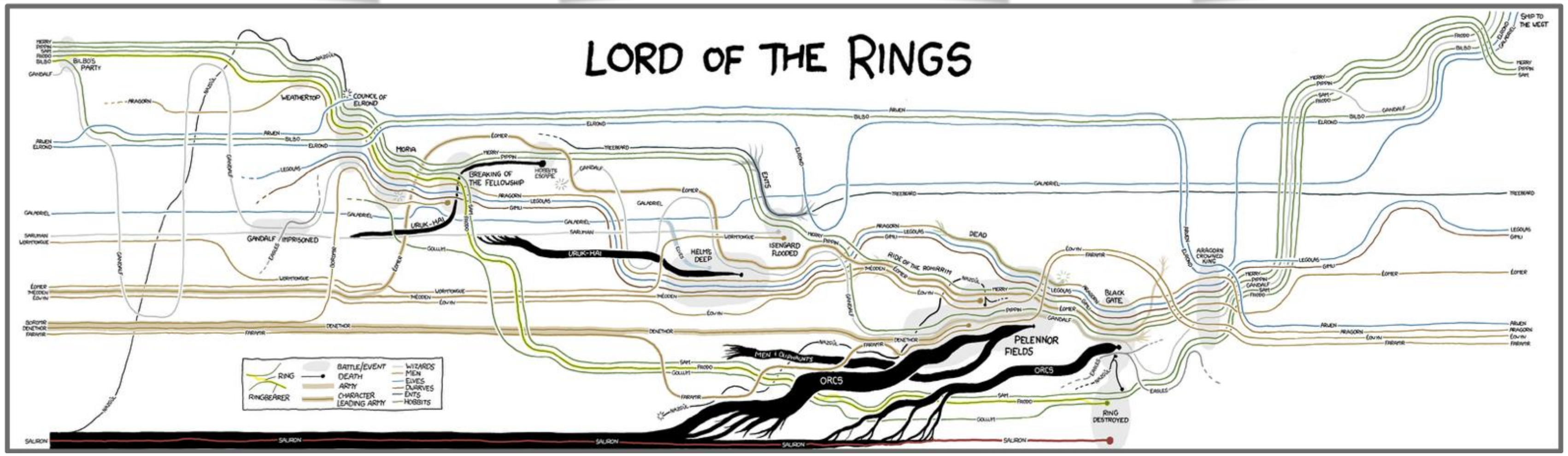
- Answer: The Mirror of Erised

or we could just watch the movie!

Visual World in Hours

...

...



Movies allow learning perception and reasoning over goals, behaviors & intents of agents in a rich dynamic contextual setup.

Two uses of video

- Exteroception
  - It teaches us about the external world. We build mental models of behavior (physical, social..) and use them to interpret, predict and control
- Proprioception
  - It tells me about my current state in the world. Helps produce an episodic memory situated in space and time, and guides action in a context-specific way

EgoSchema: An unsolved problem in long-range video understanding

NeurIPS 2023

Karttikeya Mangalam

Raiymbek Akshulakov

Jitendra Malik

UC Berkeley

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Berkeley
UNIVERSITY OF CALIFORNIA

Long Video

Q: What water activity is being Shown? (Action classification)
A: Swimming

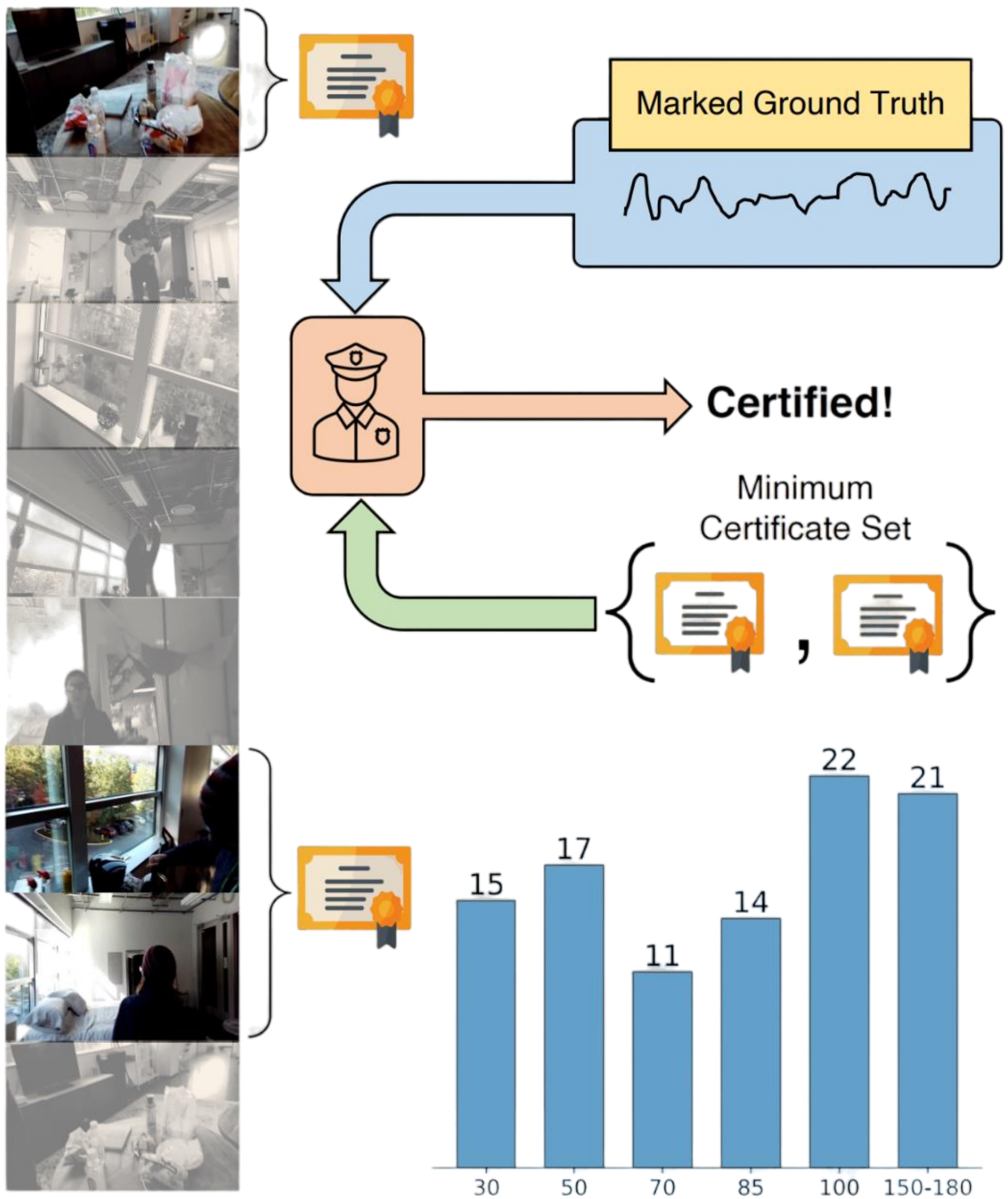
Q: What swimming stroke are They doing?
A: Butterfly stroke

Q: How did michael phelps position change over the race?
A: ...

!=

Long Temporal Understanding Task

How to Make this notion of Long Temporal task precise?



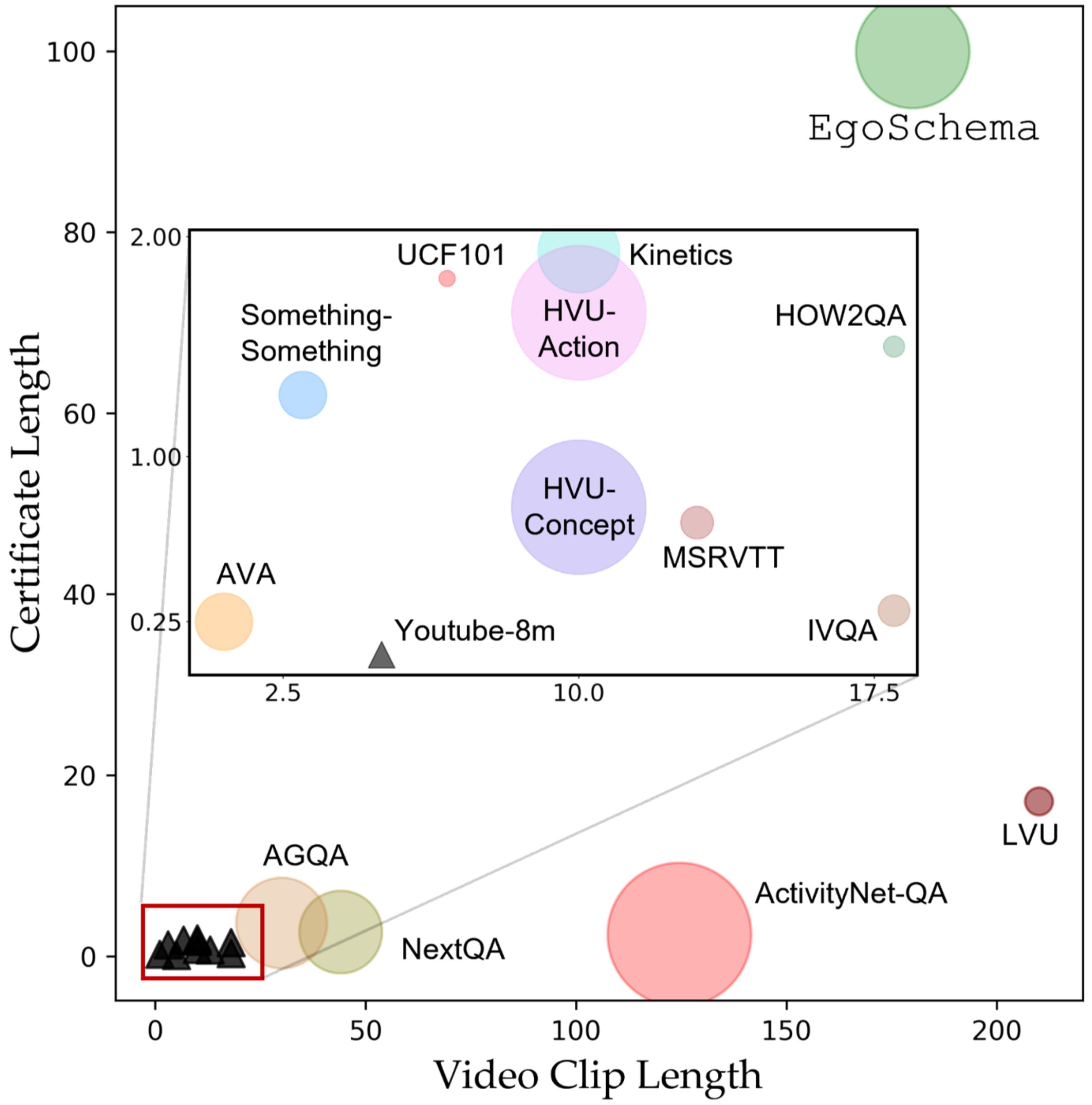
Temporal Certificates

The **smallest set** of sub-clips that are both necessary and sufficient (on its own) to convince a human verifier of the veracity of the marked annotation

Certificates for 15 video datasets & benchmarks

All existing video dataset have very small temporal certificate lengths, with most popular datasets being less than 2 seconds.

Even the longest dataset (LVU) is less than 20 seconds.



Benchmarking State-of-the-art On EgoSchema

Model	Release	Inference Params	Evaluation Setting	QA Acc
Choosing the correct $\mathcal{A}$ uniformly at random				20.0%
FrozenBiLM [57]	Oct 2022	1.2B	10 frames 90 frames	26.4% 26.9%
VIOLET [14]	Sept 2022	198M	5 frames 75 frames	19.9% 19.6%
mPLUG-Owl [59]	May 2023	7.2B	1 frame 5 frames 10 frames 15 frames 30 frames	27.0% 31.1% 29.6% 28.7% 20.0%
InternVideo [48]	Dec 2022	478M	10 frames 30 frames 90 frames	31.4% 31.8% **32.1%**

Nothing works!

7B+ parameter SOTA video & language models achieve < 33% MCQ accuracy

(random accuracy is 20%)

Benchmarking Human Performance On Egoschema

Evaluation Setting	QA Accuracy
180 frames	67.2%
In <1 min	67.0%
In <3 min	68.0%
No constraint	75.0%
Video $\rightarrow$ Text	**76.2%**

A huge gap exists between current SOTA model and human performance on EgoSchema

We believe this presents a very exciting opportunity for future research!

Is scaling token-based LLM like models the answer ?

- Doesn’t capture the essence of the 4D world
- Complexity likely to be too high
- …

Book Name	Word Count	Movie Length (min)	Number of Frames (x60x24)	Number of Tokens (x196)
"The Great Gatsby" by F. Scott Fitzgerald	47.0 K	143	205.9 K	40.4 M
"Sorcerer's Stone" by J.K. Rowling	78.0 K	152	218.9 K	42.9 M
"The Hobbit" by J.R.R. Tolkien	95.0 K	169	243.4 K	47.7 M
"Pride and Prejudice" by Jane Austen	120.0 K	129	185.8 K	36.4 M
"The Shining" by Stephen King	200.0 K	146	210.2 K	41.2 M
"To Kill a Mockingbird" by Harper Lee	100.0 K	129	185.8 K	36.4 M
"The Godfather" by Mario Puzo	144.0 K	175	252.0 K	49.4 M
"Jurassic Park" by Michael Crichton	127.0 K	127	182.9 K	35.8 M
"Gone with the Wind" by Margaret Mitchell	418.0 K	238	342.7 K	67.2 M
"Lord of the Flies" by William Golding	60.0 K	92	132.5 K	26.0 M

Book Name	Word Count	Movie Length (min)	Number of Frames (x60x24)	Number of Tokens (x196)
"The Great Gatsby" by F. Scott Fitzgerald	47.0 K	143	205.9 K	40.4 M
"Sorcerer's Stone" by J.K. Rowling	78.0 K	152	218.9 K	42.9 M
"The Hobbit" by J.R.R. Tolkien	95.0 K	169	243.4 K	47.7 M
"Pride and Prejudice" by Jane Austen	120.0 K	129	185.8 K	36.4 M
"The Shining" by Stephen King	200.0 K	146	210.2 K	41.2 M
"To Kill a Mockingbird" by Harper Lee	100.0 K	129	185.8 K	36.4 M
"The Godfather" by Mario Puzo	144.0 K	175	252.0 K	49.4 M
"Jurassic Park" by Michael Crichton	127.0 K	127	182.9 K	35.8 M
"Gone with the Wind" by Margaret Mitchell	418.0 K	238	342.7 K	67.2 M
"Lord of the Flies" by William Golding	60.0 K	92	132.5 K	26.0 M
Youtube		**9.4 B**	**13478.4 B**	**2641.8 T**

Book Name	Word Count	Movie Length (min)	Number of Frames (x60x24)	Number of Tokens (x196)
"The Great Gatsby" by F. Scott Fitzgerald	47.0 K	143	205.9 K	40.4 M
"Sorcerer's Stone" by J.K. Rowling	78.0 K	152	218.9 K	42.9 M
"The Hobbit" by J.R.R. Tolkien	95.0 K	169	243.4 K	47.7 M
"Pride and Prejudice" by Jane Austen	120.0 K	129	185.8 K	36.4 M
"The Shining" by Stephen King	200.0 K	146	210.2 K	41.2 M
"To Kill a Mockingbird" by Harper Lee	100.0 K	129	185.8 K	36.4 M
"The Godfather" by Mario Puzo	144.0 K	175	252.0 K	49.4 M
"Jurassic Park" by Michael Crichton	127.0 K	127	182.9 K	35.8 M
"Gone with the Wind" by Margaret Mitchell	418.0 K	238	342.7 K	67.2 M
"Lord of the Flies" by William Golding	60.0 K	92	132.5 K	26.0 M
Youtube		**9.4 B**	**13478.4 B**	**2641.8 T**
Llama-2				**2 T**

Multiscale Vision Transformers (MViT)

Haoqi Fan*, Bo Xiong*, Karttikeya Mangalam*, Yanghao Li*, Zhicheng Yan, Jitendra Malik, Christoph Feichtenhofer*

Facebook AI Research (FAIR)

github.com/facebookresearch/pytorchvideo

github.com/facebookresearch/SlowFast

* Equal technical contribution

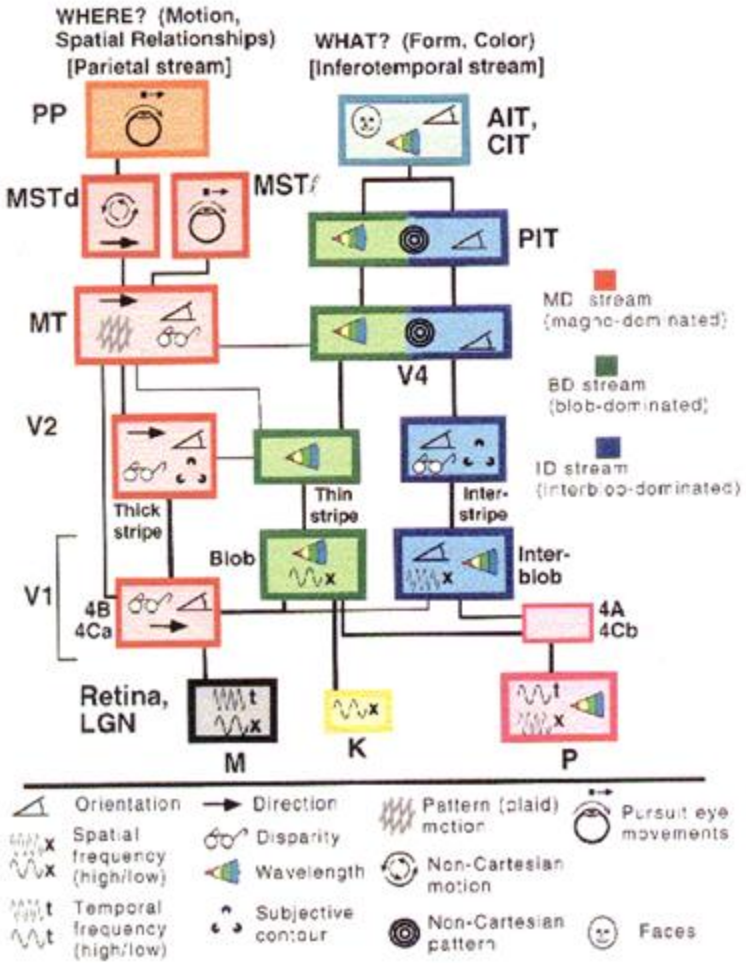
Motivation

Visual pathways are made up of hierarchical multiscale representations

RECEPTIVE FIELDS OF OPTIC NERVE FIBRES IN THE SPIDER MONKEY

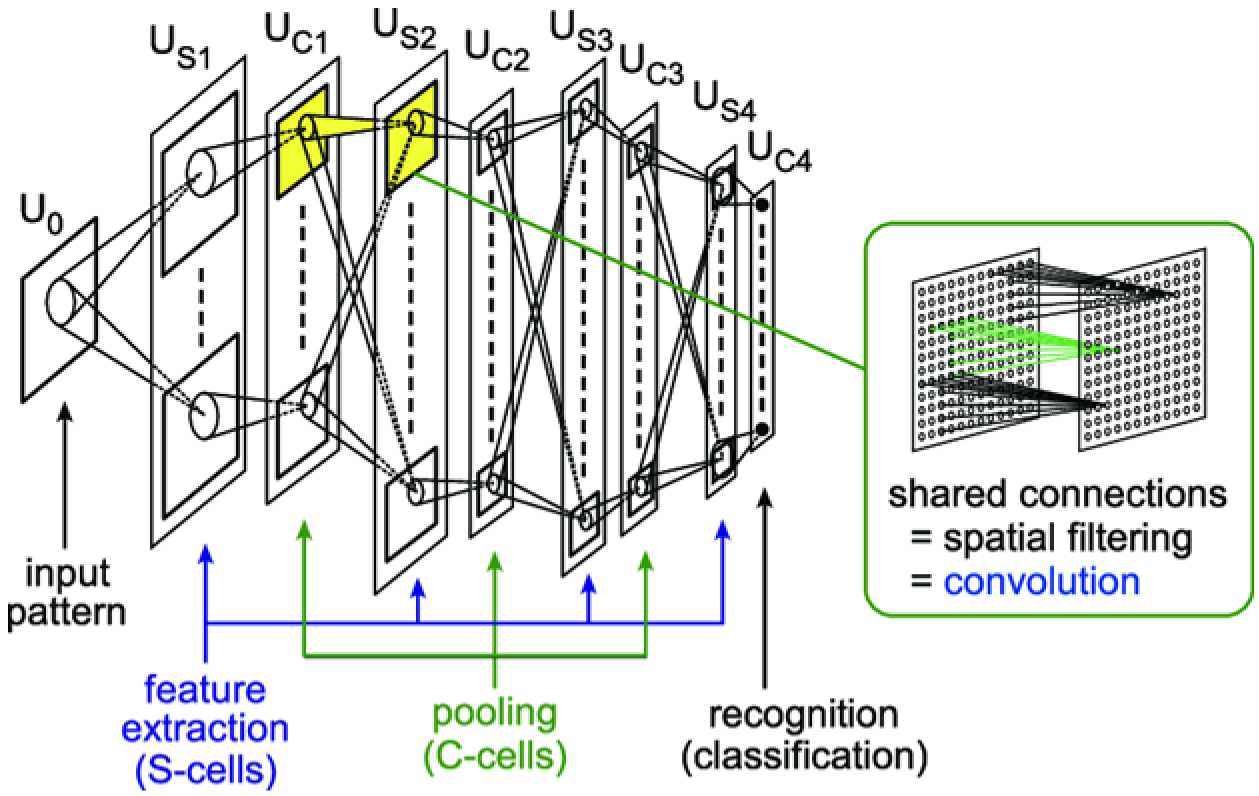
BY D. H. HUBEL AND T. N. WIESEL

From the Neurophysiology Laboratory, Department of Pharmacology, Harvard Medical School, Boston, Mass., U.S.A.

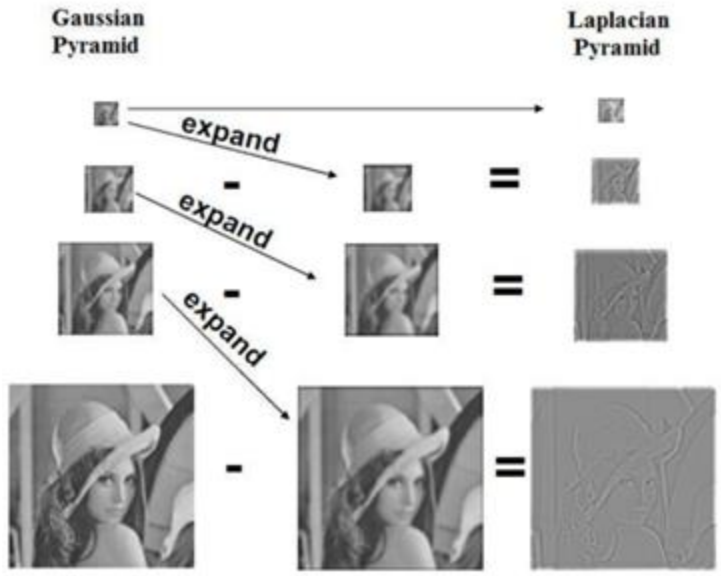


[1] Van Essen, David C., and Jack L. Gallant. "Neural mechanisms of form and motion processing in the primate visual system." *Neuron* 13.1 (1994): 1-10.

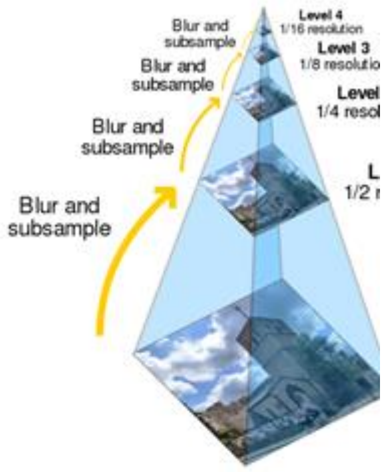
Motivation



Neocognitron



Laplacian Pyramid Codes

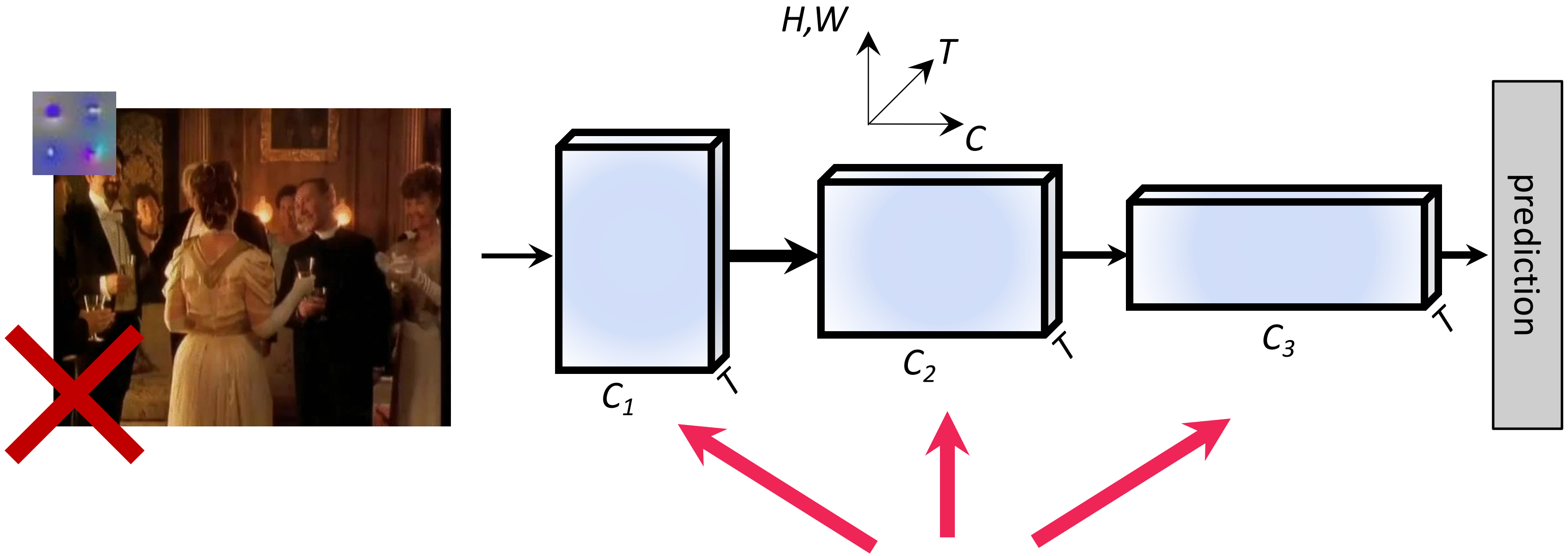


Gaussian Image Pyramids

[1] Fukushima, Kunihiko, and Sei Miyake. "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition." *Competition and cooperation in neural nets*. Springer, Berlin, Heidelberg, 1982. 267-285.
[2] Burt, Peter J., and Edward H. Adelson. "The Laplacian pyramid as a compact image code." Readings in computer vision. Morgan Kaufmann, 1987. 671-679.

Status quo: Video Recognition with 3D CNNs

4D tensors of shape T x H x W x C



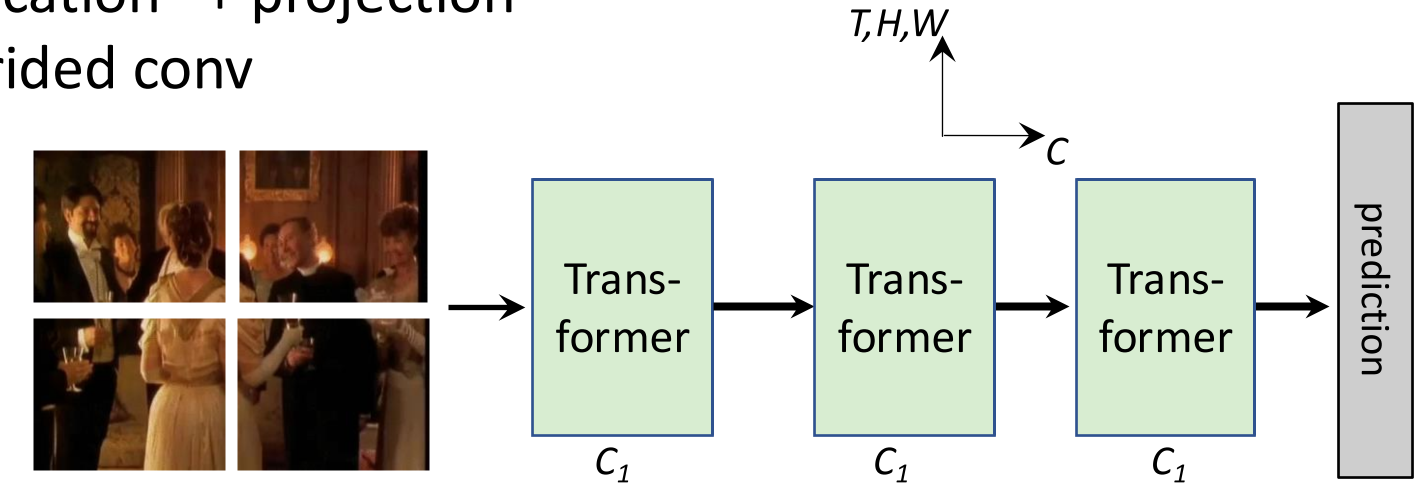
(T) x H x W shrinks, C grows.

e.g. 1x1x1 & 3x3x3 convs.

Vision Transformers (ViT)

“Patchification” + projection
a.k.a. strided conv

2D tensors of shape THW x C

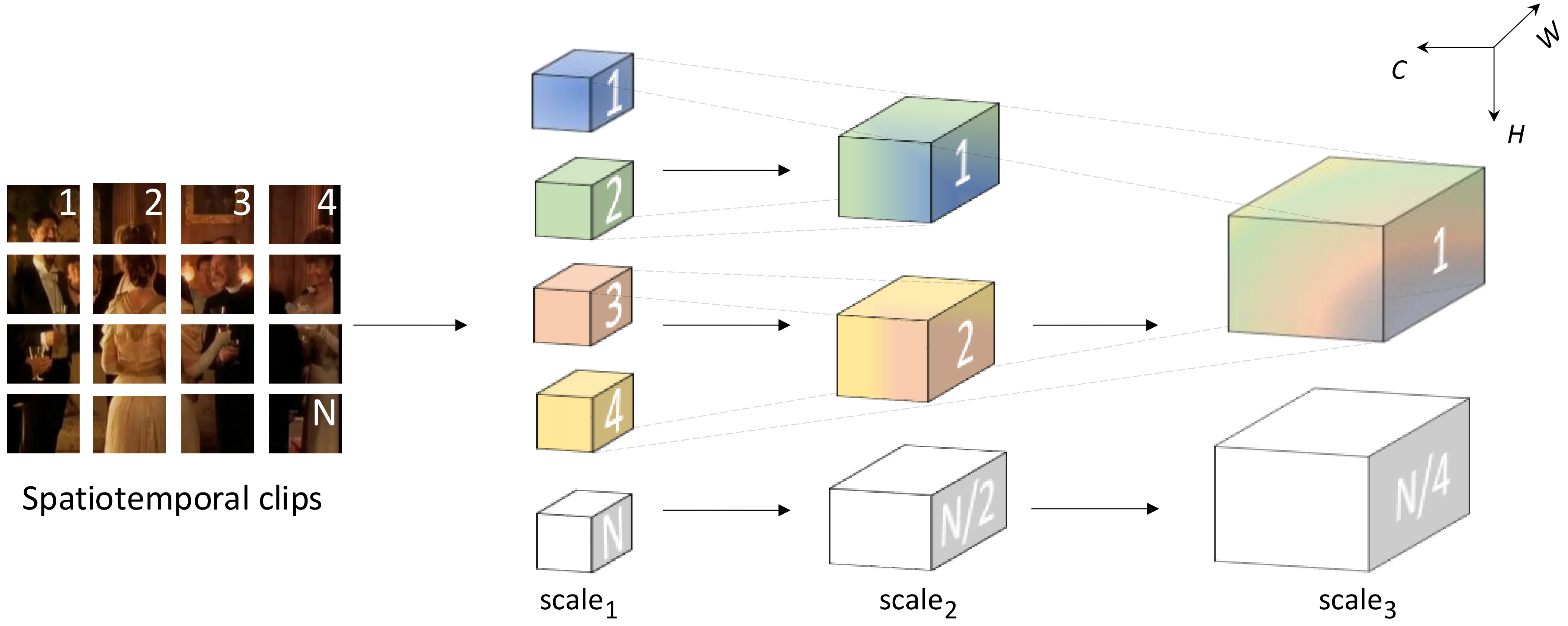


Spatiotemporal clips

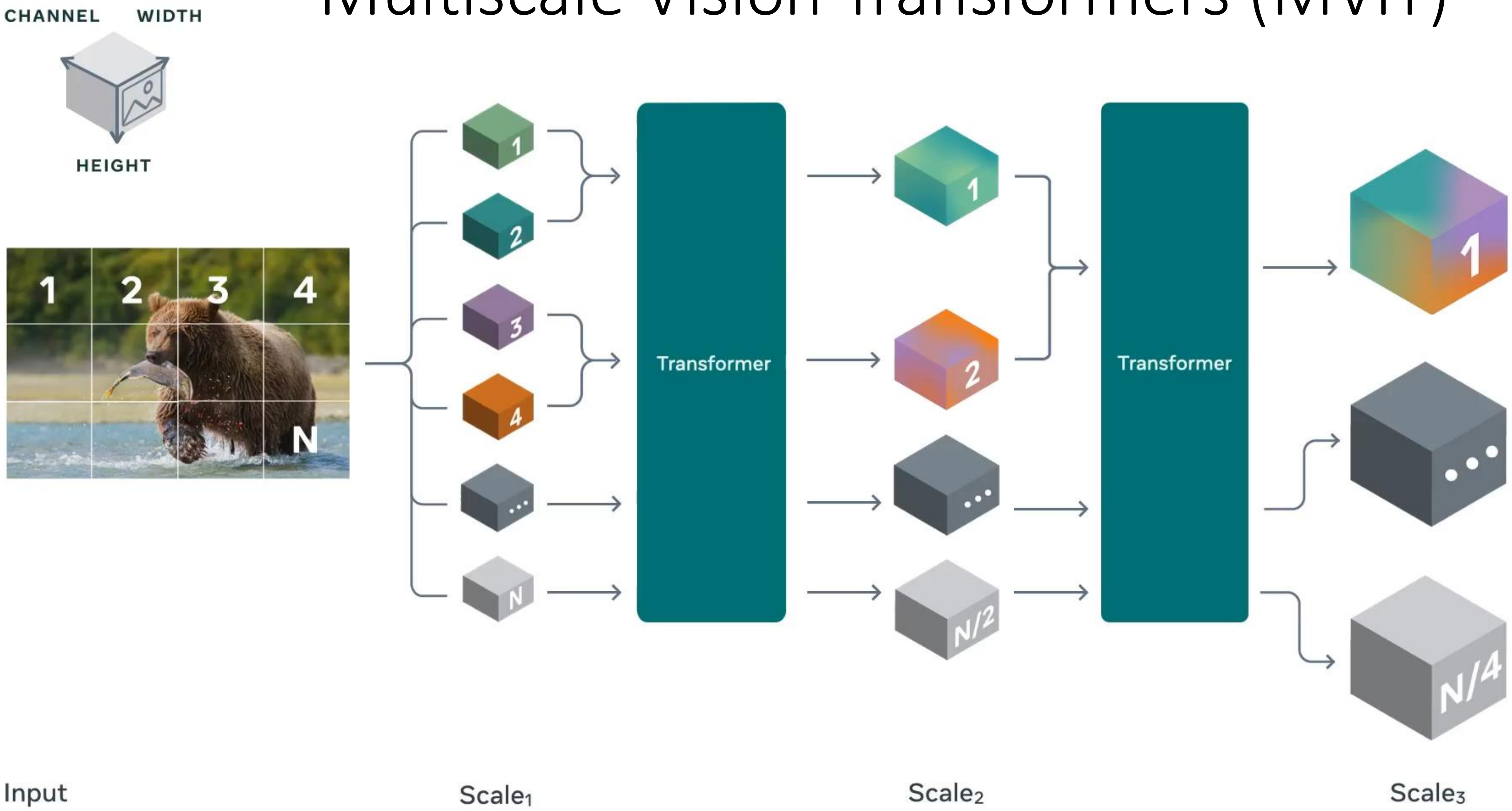
THW constant, C constant.

1x1 convs & TxHxW attention.

Multiscale Vision Transformers (MViT)



Multiscale Vision Transformers (MViT)

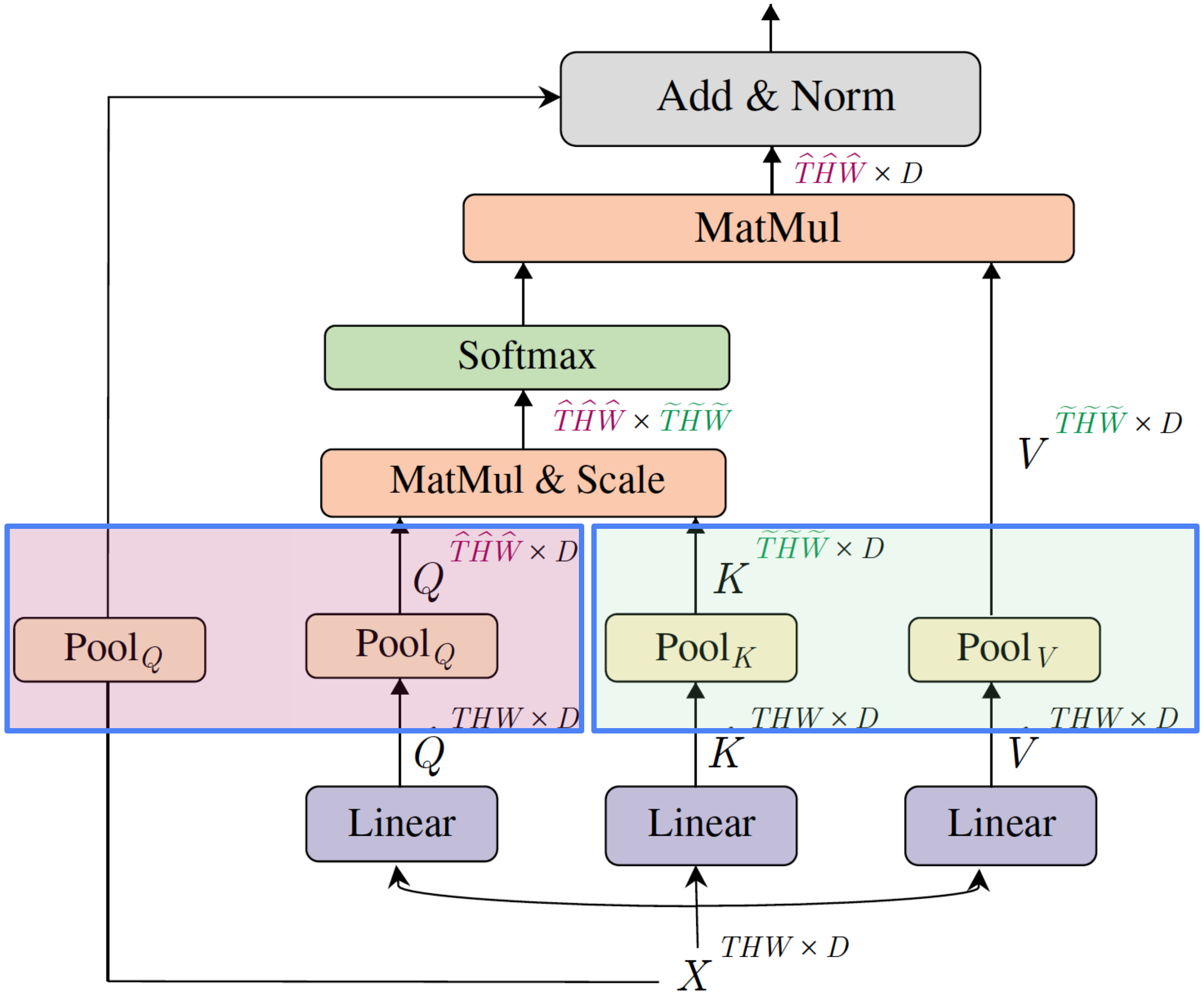


MViT:
(1) **Pooling** attention

- Multi Head Pooling Attention (***MHPA***)
- Pooling either **K**ey & **V**alue (P^K & P^V) or **Q**uery (P^Q)
- Pooling Attention (PA):

$$\mathrm{PA}(\cdot) = \mathrm{Softmax}(\mathcal{P}(Q; \boldsymbol{\Theta}_Q)\mathcal{P}(K; \boldsymbol{\Theta}_K)^T/\sqrt{d})\mathcal{P}(V; \boldsymbol{\Theta}_V)$$

- Pooling **K, V** reduces attention computation
- Pooling **Q** reduces output dimension
- Channel expansion is done with the MLP block of the previous stage



Status quo: ImageNet ViT-B

- Video deals with a T-times longer sequence length vs. image transformers; e.g. ViT-B becomes the following

stage	operators	output sizes
data layer	stride $\tau \times 1 \times 1$	$T \times H \times W$
patch_1	$1 \times 16 \times 16$, D stride $1 \times 16 \times 16$	$D \times T \times \frac{H}{16} \times \frac{W}{16}$
scale_2	$\begin{bmatrix} \text{MHA}(D) \\ \text{MLP}(4D) \end{bmatrix} \times N$	$D \times T \times \frac{H}{16} \times \frac{W}{16}$

ViT vs. MViT
(2) **Stage** design

ViT

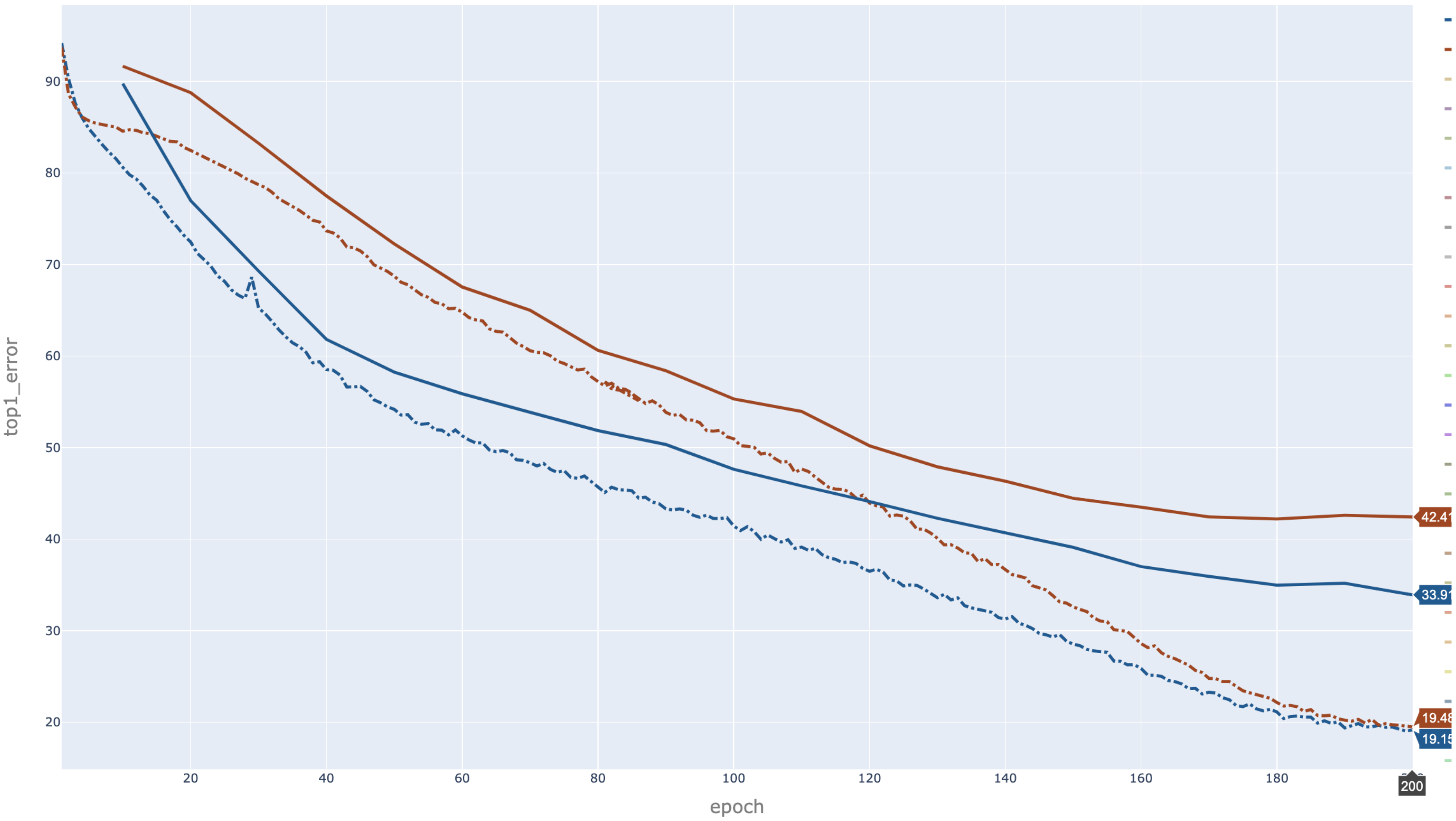
stage	operators	output sizes
data layer	stride $\tau \times 1 \times 1$	$T \times H \times W$
patch$_1$	$1 \times 16 \times 16$, D stride $1 \times 16 \times 16$	$D \times T \times \frac{H}{16} \times \frac{W}{16}$
scale$_2$	$\begin{bmatrix} \text{MHA}(D) \\ \text{MLP}(4D) \end{bmatrix} \times N$	$D \times T \times \frac{H}{16} \times \frac{W}{16}$

MViT

stages	operators	output sizes
data layer	stride $\tau \times 1 \times 1$	$D \times T \times H \times W$
cube$_1$	$c_T \times c_H \times c_W$, D stride $s_T \times 4 \times 4$	$D \times \frac{T}{s_T} \times \frac{H}{4} \times \frac{W}{4}$
scale$_2$	$\begin{bmatrix} \text{MHPA}(D) \\ \text{MLP}(4D) \end{bmatrix} \times N_2$	$D \times \frac{T}{s_T} \times \frac{H}{4} \times \frac{W}{4}$
scale$_3$	$\begin{bmatrix} \text{MHPA}(2D) \\ \text{MLP}(8D) \end{bmatrix} \times N_3$	$2D \times \frac{T}{s_T} \times \frac{H}{8} \times \frac{W}{8}$
scale$_4$	$\begin{bmatrix} \text{MHPA}(4D) \\ \text{MLP}(16D) \end{bmatrix} \times N_4$	$4D \times \frac{T}{s_T} \times \frac{H}{16} \times \frac{W}{16}$
scale$_5$	$\begin{bmatrix} \text{MHPA}(8D) \\ \text{MLP}(32D) \end{bmatrix} \times N_5$	$8D \times \frac{T}{s_T} \times \frac{H}{32} \times \frac{W}{32}$

MHPA = Multihead Pooling Attention

ViT-B (red) vs. MViT-B (blue)

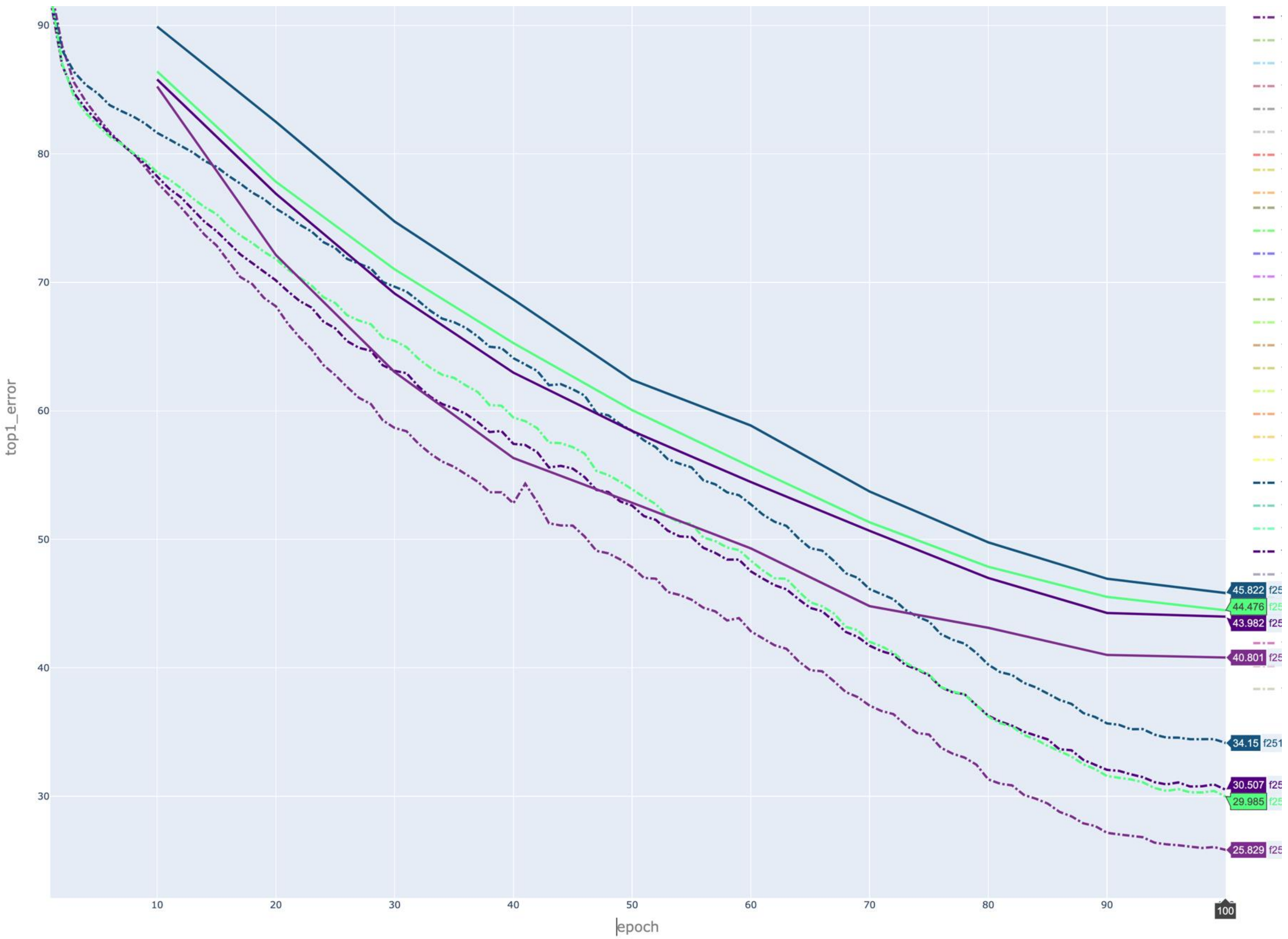


Ablations: Two scales in ViT

variant	$[\mathrm{N}_1, \mathrm{N}_2]$	FLOPs (G)	Mem (G)	Acc
ViT-B	[12, 0]	179.6	16.8	68.5
2-scale ViT-B, Q pool	[6, 6]	**111.1** **(−68.5)**	9.8 **(−7.0)**	**71.0** **(+1.5)**
ViT-B, K, V pool	[12, 0]	148.4 **(−31.2)**	**8.9** **(−7.9)**	69.1 **(+0.6)**

Table 8. **Query (scale stage) and Key-Value pooling on ViT-B.** Introducing a *single* extra resolution stage into ViT-B boosts accuracy by +1.5%. Pooling K, V provides +0.6% accuracy. Both techniques allow dramatic FLOPs/memory savings.

Pooling Q / [K, V] / both+ stages



No pooling, – 56G flops (DeiT-S)

P_Q – 46G flops (DeiT-S)

$P_{K,V}$ – 48G flops (DeiT-S)

$P_{Q,K,V}$ + grow width – 35G flops

ViT vs. MViT: Instantiations

ViT-B

stage	operators	output sizes
data	stride $8\times1\times1$	$8\times224\times224$
patch_1	$1\times16\times16$, 768 stride $1\times16\times16$	$768\times8\times14\times14$
scale_2	$\begin{bmatrix}\text{MHA}(768)\\ \text{MLP}(3072)\end{bmatrix}\times12$	$768\times8\times14\times14$

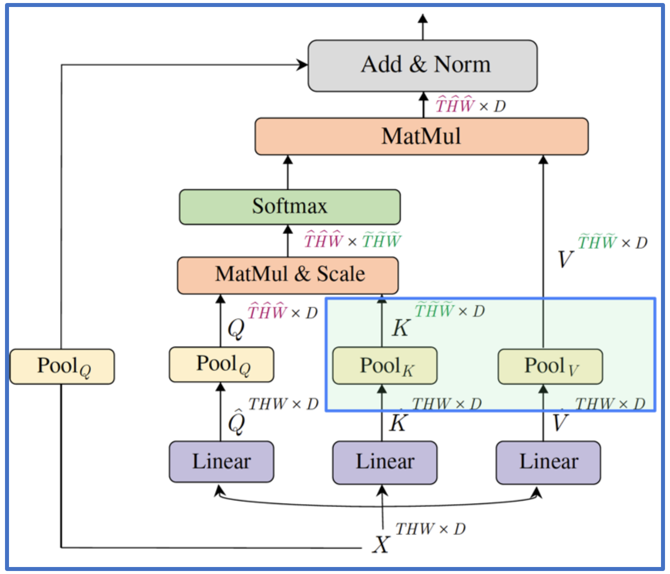
(a) **ViT**-B with **179.6**G FLOPs, **87.2**M param, **16.8**G memory, and **68.5**% top-1 accuracy.

MViT-B

stage	operators	output sizes
data	stride $4\times1\times1$	$16\times224\times224$
cube_1	$3\times7\times7$, 96 stride $2\times4\times4$	$96\times8\times56\times56$
scale_2	$\begin{bmatrix}\text{MHPA}(96)\\ \text{MLP}(384)\end{bmatrix}\times1$	$96\times8\times56\times56$
scale_3	$\begin{bmatrix}\text{MHPA}(192)\\ \text{MLP}(768)\end{bmatrix}\times2$	$192\times8\times28\times28$
scale_4	$\begin{bmatrix}\text{MHPA}(384)\\ \text{MLP}(1536)\end{bmatrix}\times11$	$384\times8\times14\times14$
scale_5	$\begin{bmatrix}\text{MHPA}(768)\\ \text{MLP}(3072)\end{bmatrix}\times2$	$768\times8\times7\times7$

(b) **MViT**-B with **70.3**G FLOPs, **36.5**M param, **6.8**G memory, and **77.2**% top-1 accuracy.

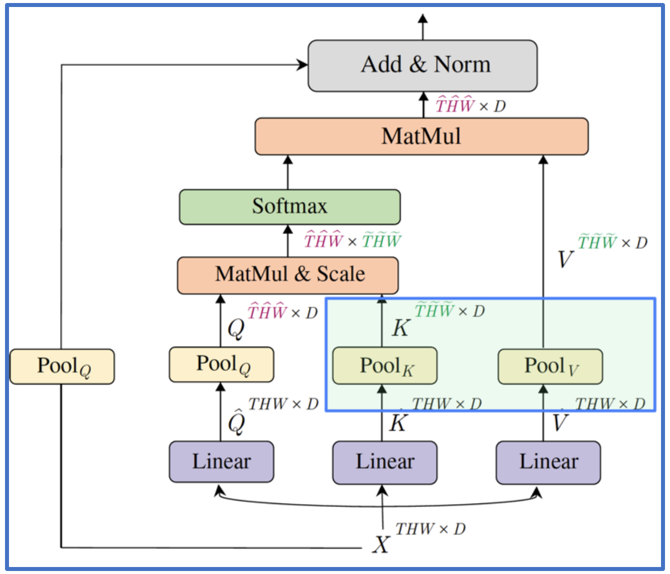
Ablations: **K,V** pooling



stride s	adaptive	FLOPs	Mem	Acc
none	n/a	130.8	16.3	**77.6**
1×4×4		71.4	8.2	75.9
2×4×4		64.3	6.6	74.8
2×4×4	✓	83.6	9.1	77.1
1×8×8	✓	70.3	6.8	77.2
2×8×8	✓	**63.7**	**6.3**	75.8

Table 12. **Key-Value pooling**: Vary stride $\mathbf{s} = s_T \times s_H \times s_W$, for pooling K and V. "adaptive" reduces stride w.r.t. stage resolution.

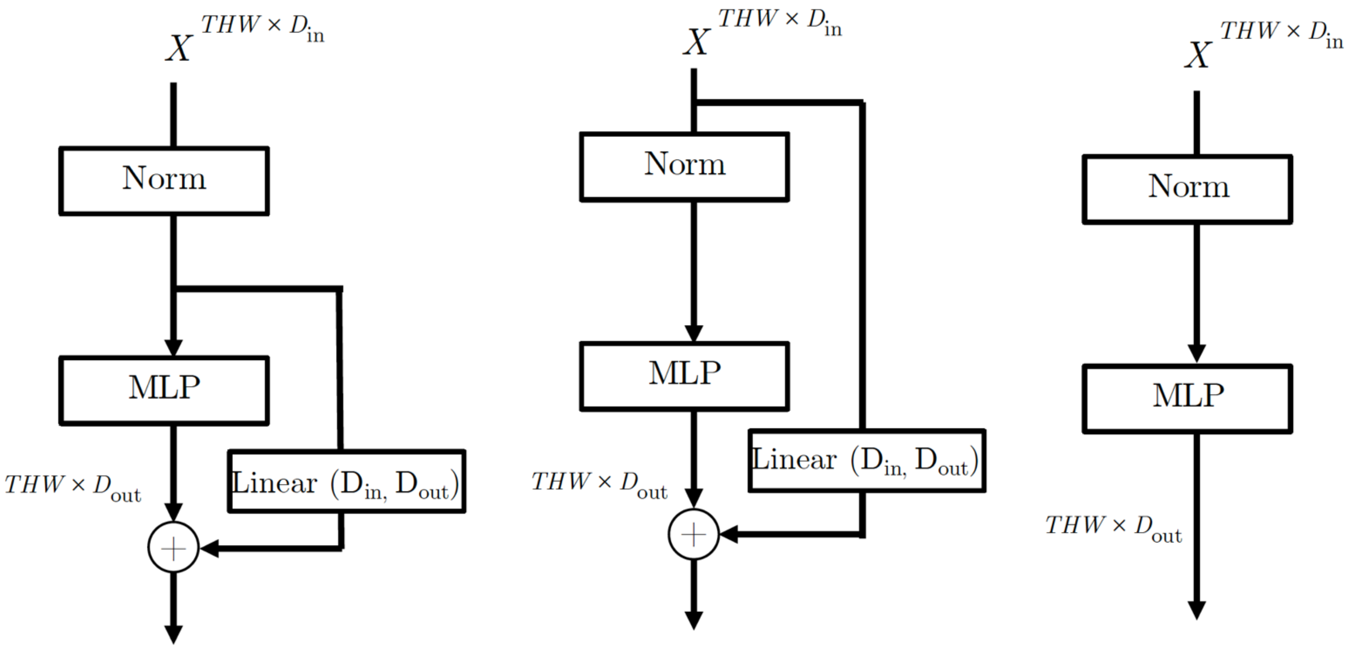
Ablations: Pooling kernel & function



kernel $\mathbf{k}$	pooling func	Param	Acc
$\mathbf{s}$	max	36.5	76.1
$2\mathbf{s}+1$	max	36.5	75.5
$\mathbf{s}+1$	max	36.5	77.2
$\mathbf{s}+1$	average	36.5	75.4
$\mathbf{s}+1$	conv	36.6	78.3
$3\times3\times3$	conv	36.6	**78.4**

Table 13. **Pooling function**: Varying the kernel $\mathbf{k}$ as a function of stride $\mathbf{s}$. Functions are average or max pooling and conv which is a learnable, channel-wise convolution.

Ablations: Skip-connections at stage-transitions



(a) **normalized** skip-connection

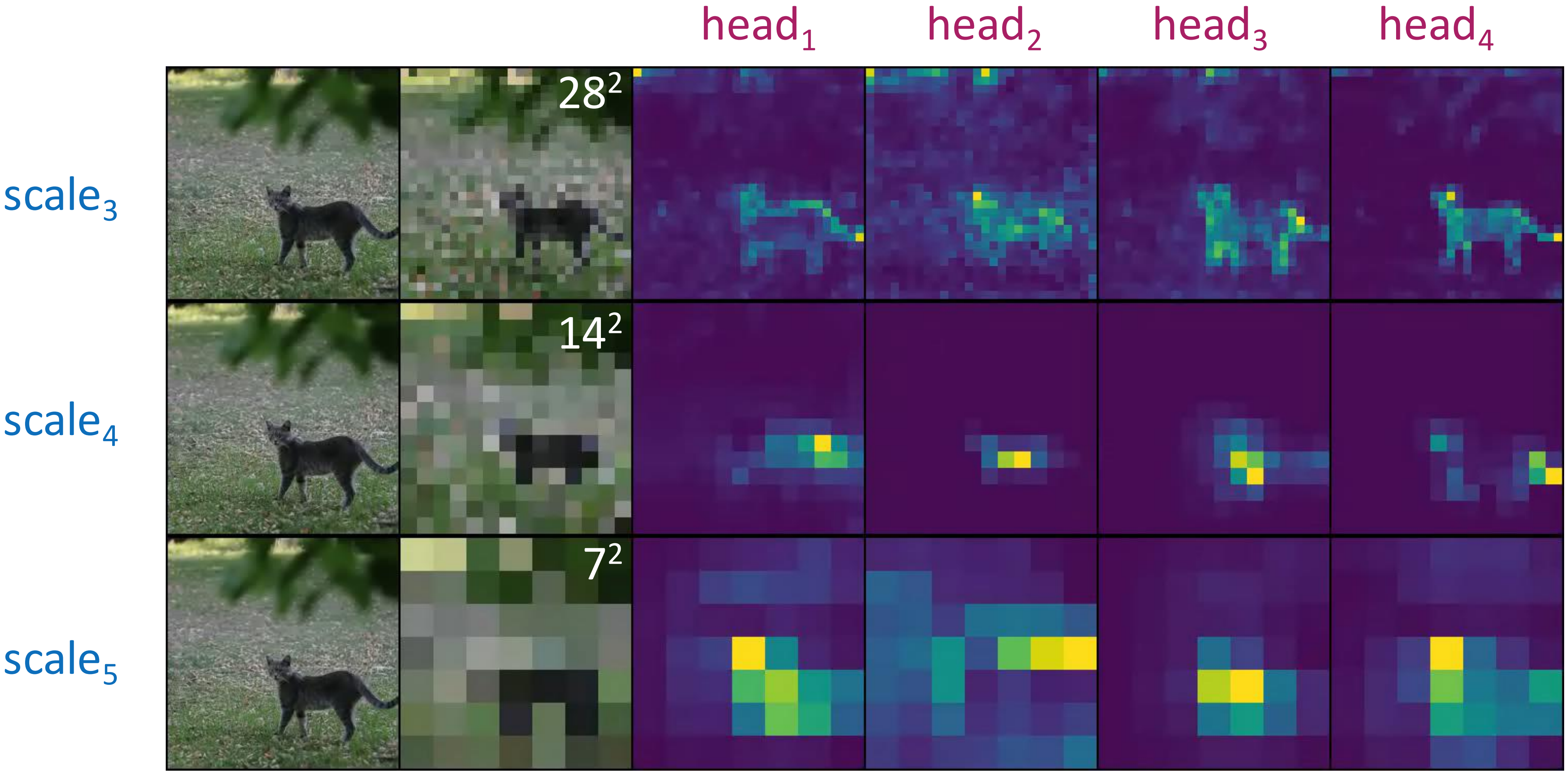
(b) unnormalized skip-connection

(c) no skip-connection

	method	top-1	top-5
(a)	**normalized** skip-connection	**77.2**	**93.1**
(b)	unnormalized skip-connection	74.6	91.3
(c)	no skip-connection	74.7	91.8

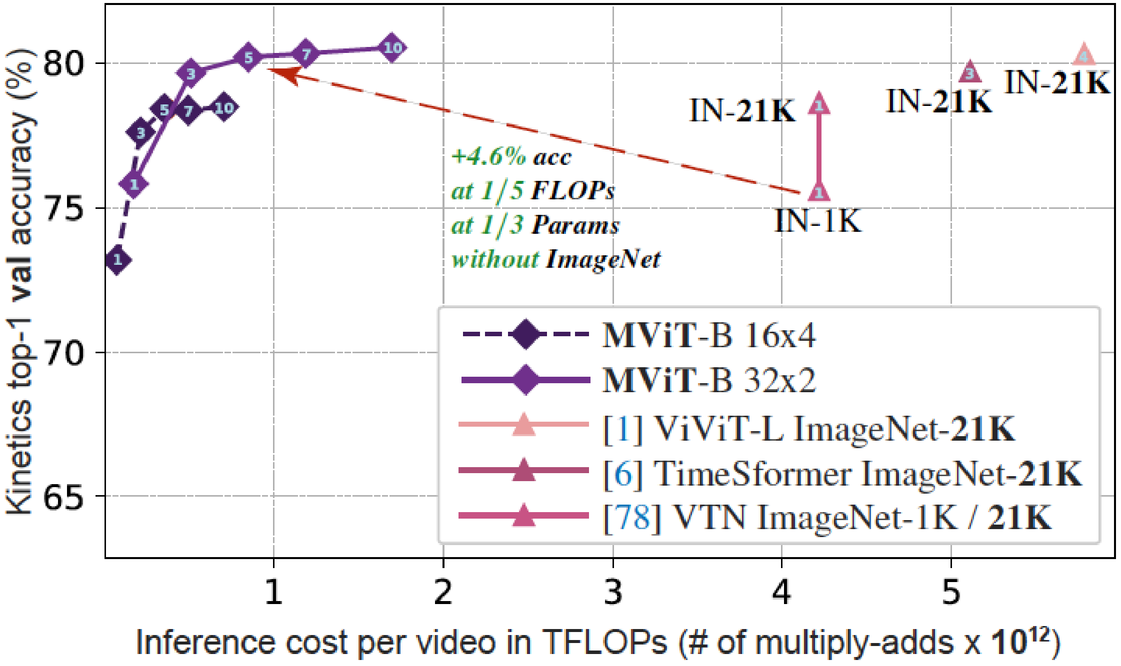
Table A.4. **Skip-connections at stage-transitions on K400.** We use our base model, MViT-B 16×4. Normalizing the skip-connection at channel expansion is essential for good performance.

Multiscale attention heads



Comparison to concurrent work on Kinetics

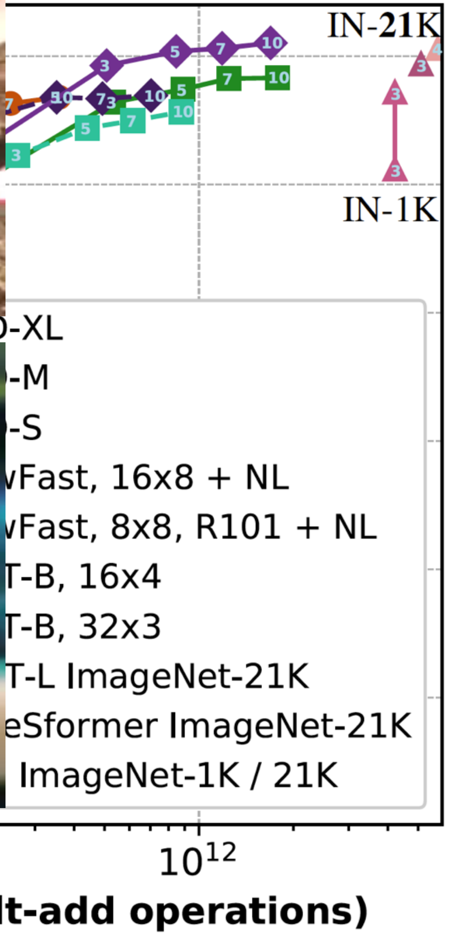
- Accuracy/computation trade-off
- 3 concurrent video transformers
  - require up to 10x more computation
  - rely on ImageNet-21K to be competitive in accuracy
    - IN-21K has **60x** more labels than Kinetics-400 (and some classes overlap)



[78] Neimark, Bar, Zohar, Asselmann (arXiv 2021). Video Transformer Network
[6] Bertasius, Wang, Torresani (arXiv 2021). Is Space-Time Attention All You Need for Video Understanding?
[1] Arnab, Dehghani, Heigold, Sun, Lučić, Schmid (arXiv 2021). ViViT: A Video Vision Transformer

The questionable (1)

Are T… s?



- At larg…
- Smalle…
- Setting has to be fair
- e.g., IN-21K has **60x** more labels than Kinetics-400
    - (and some classes overlap)
    - IN-21K: 100+ "snake" classes
    - IN-1K: ~15 "snake" classes
    - K400: "holding snake" class

[1] Neimark, Bar, Zohar, Asselmann (arXiv 2021). Video Transformer Network
[2] Bertasius, Wang, Torresani (arXiv 2021). Is Space-Time Attention All You Need for Video Understanding?
[3] Arnab, Dehghani, Heigold, Sun, Lučić, Schmid (arXiv 2021). ViViT: A Video Vision Transformer

The questionable (1)

Are Transformers **really** better than CNNs?

- Transformers are faster on GPUs

model	clips/sec	Acc	FLOPs×views	Param
X3D-M [27]	7.9	74.1	4.7×1×5	3.8
SlowFast R50 [28]	5.2	75.7	65.7×1×5	34.6
SlowFast R101 [28]	3.2	77.6	125.9×1×5	62.8
ViT-B [24]	3.6	68.5	179.6×1×5	87.2
MViT-S, max-pool	**12.3**	74.3	32.9×1×5	26.1
MViT-B, max-pool	6.3	77.2	70.3×1×5	36.5

+ 5.8% top-1

3.4x faster

Ablation: Training throughput measured in clips/s

The questionable (2)

MViT vs. ViT: Shuffling input frames

model	shuffling	FLOPs (G)	Param (M)	Acc
MViT-B		70.3	36.5	**77.2**
MViT-B	✓			70.1 **(–7.1)**
ViT-B		179.6	87.2	68.5
ViT-B	✓			68.4 **(–0.1)**

Table 7. **Shuffling frames in inference. MViT**-B severely drops (–7.1%) for shuffled temporal input, but ViT-B models appear to *ignore* temporal information as accuracy remains similar (–0.1%).

The questionable (2)

Vanilla positional embedding is not very effective

	positional embedding	Param (M)	Acc
(i)	none	36.2	75.8
(ii)	space-only	36.5	76.7
(iii)	joint space-time	38.6	76.5
(iv)	**separate** in space & time	36.5	**77.2**

Table 9. **Effect of separate space-time positional embedding.** Backbone: **MViT**-B, 16×4. FLOPs are 70.3G for all variants.

Comparison to prior work

model	pre-train	top-1	top-5	FLOPs×views	Param
Two-Stream I3D [11]	-	71.6	90.0	216 × NA	25.0
ip-CSN-152 [94]	-	77.8	92.8	109×3×10	32.8
SlowFast 8×8 +NL [29]	-	78.7	93.5	116×3×10	59.9
SlowFast 16×8 +NL [29]	-	79.8	93.9	234×3×10	59.9
X3D-M [28]	-	76.0	92.3	6.2×3×10	3.8
X3D-XL [28]	-	79.1	93.9	48.4×3×10	11.0
ViT-B-VTN [76]	ImageNet-1K	75.6	92.4	4218×1×1	114.0
ViT-B-VTN [76]	ImageNet-**21K**	78.6	93.7	4218×1×1	114.0
ViT-B-TimeSformer [6]	ImageNet-**21K**	80.7	94.7	2380×3×1	121.4
ViT-L-ViViT [1]	ImageNet-**21K**	81.3	94.7	3992×3×4	307.0
ViT-B (our baseline)	ImageNet-**21K**	79.3	93.9	180×1×5	87.2
ViT-B (our baseline)	-	68.5	86.9	180×1×5	87.2

Table 4. **Comparison with previous work on Kinetics-400**.

12x fe

FLO

model	pretrain	top-1	top-5	GFLOPs×views	Param
SlowFast 16×8 +NL [34]	-	81.8	95.1	234×3×10	59.9
X3D-M	-	78.8	94.5	6.2×3×10	3.8
X3D-XL	-	81.9	95.5	48.4×3×10	11.0
ViT-B-TimeSformer [8]	IN-**21K**	82.4	96.0	1703×3×1	121.4
ViT-L-ViViT [1]	IN-**21K**	83.0	95.7	3992×3×4	310.8
MViT-B, 16×4	-	82.1	95.7	70.5×1×5	36.8
MViT-B, 32×3	-	83.4	96.3	170×1×5	36.8
MViT-B-24, 32×3	-	**84.1**	**96.5**	236×1×5	52.9

Table 5. **Comparison with previous work on Kinetics-600**.

model	pretrain	top-1	top-5	FLOPs×views	Param
TSM-RGB [76]	IN-1K+K400	63.3	88.2	62.4×3×2	42.9
MSNet [68]	IN-1K	64.7	89.4	67×1×1	24.6
TEA [73]	IN-1K	65.1	89.9	70×3×10	-
ViT-B-TimeSformer [8]	IN-**21K**	62.5	-	1703×3×1	121.4
ViT-B (our baseline)	IN-**21K**	63.5	88.3	180×3×1	87.2
SlowFast R50, 8×8 [34]	K400	61.9	87.0	65.7×3×1	34.1
SlowFast R101, 8×8 [34]	K400	63.1	87.6	106×3×1	53.3
MViT-B, 16×4	K400	64.7	89.2	70.5×3×1	36.6
MViT-B, 32×3	K400	67.1	90.8	170×3×1	36.6
MViT-B, 64×3	K400	**67.7**	**90.9**	455×3×1	36.6
MViT-B, 16×4	K600	66.2	90.2	70.5×3×1	36.6
MViT-B, 32×3	K600	67.8	91.3	170×3×1	36.6
MViT-B-24, 32×3	K600	**68.7**	**91.5**	236×3×1	53.2

Table 6. **Comparison with previous work on SSv2**.

Applying the architecture to image classification

- Multiscale idea is space/time agnostic
- MViT on ImageNet: We **simply** remove temporal dimension
- Training deep MViT networks works out of the box
  - +2% over DeiT-B at lower FLOPs
  - +0.7% better than concurrent Swin-B at lower FLOPs/Params
- Multiscale Vision Transformers perform state-of-the-art on ImageNet, even though it was designed for video classification
- Code github.com/facebookresearch/pytorchvideo
  github.com/facebookresearch/SlowFast

ImageNet top-1 accuracy

model	pretrain	Acc	FLOPs (G)	Param (M)
RegNetZ-4GF [27]		83.1	4.0	28.1
RegNetZ-16GF [27]		84.1	15.9	95.3
EfficientNet-B7 [99]		84.3	37.0	66.0
DeiT-S [101]		79.8	4.6	22.1
DeiT-B [101]		81.8	17.6	86.6
DeiT-B $\uparrow 384^2$ [101]		83.1	55.5	87.0
Swin-B (concurrent) [79]		83.3	15.4	88.0
Swin-B $\uparrow 384^2$ (concurrent) [79]		84.2	47.0	88.0
MViT-B-16, max-pool		82.5	7.8	37.0
MViT-B-16		83.0	7.8	37.0
MViT-B-24		84.0	14.7	72.7
MViT-B-24-320^2		84.8	32.7	72.9
MViT-L-48 $\uparrow 384^2$		**86.0**	140.7	218.5

Conclusion

- Multiscale Vision Transformers connect multiscale feature hierarchies with transformers
- MViT hierarchically *expands feature* complexity while *reducing* visual *resolution*
- In empirical evaluation, it shows advantage over single-scale vision transformers
- MViT achieves substantial gains with respect to concurrent transformers (ViT) across major video benchmarks at a fraction of the inference compute
- ***Without*** using any ***external data***: Other concurrent works report a failure to train video transformer models without ImageNet pre-training
- For ***image recognition***: MViT outperforms concurrent work on ImageNet and even though it has been designed for video

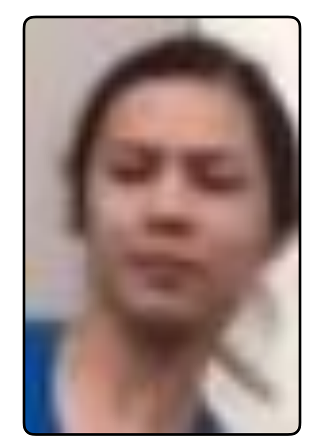
github.com/facebookresearch/pytorchvideo
github.com/facebookresearch/SlowFast

Learning Individual Styles of Conversational Gesture


Shiry Ginosar*


Amir Bar*

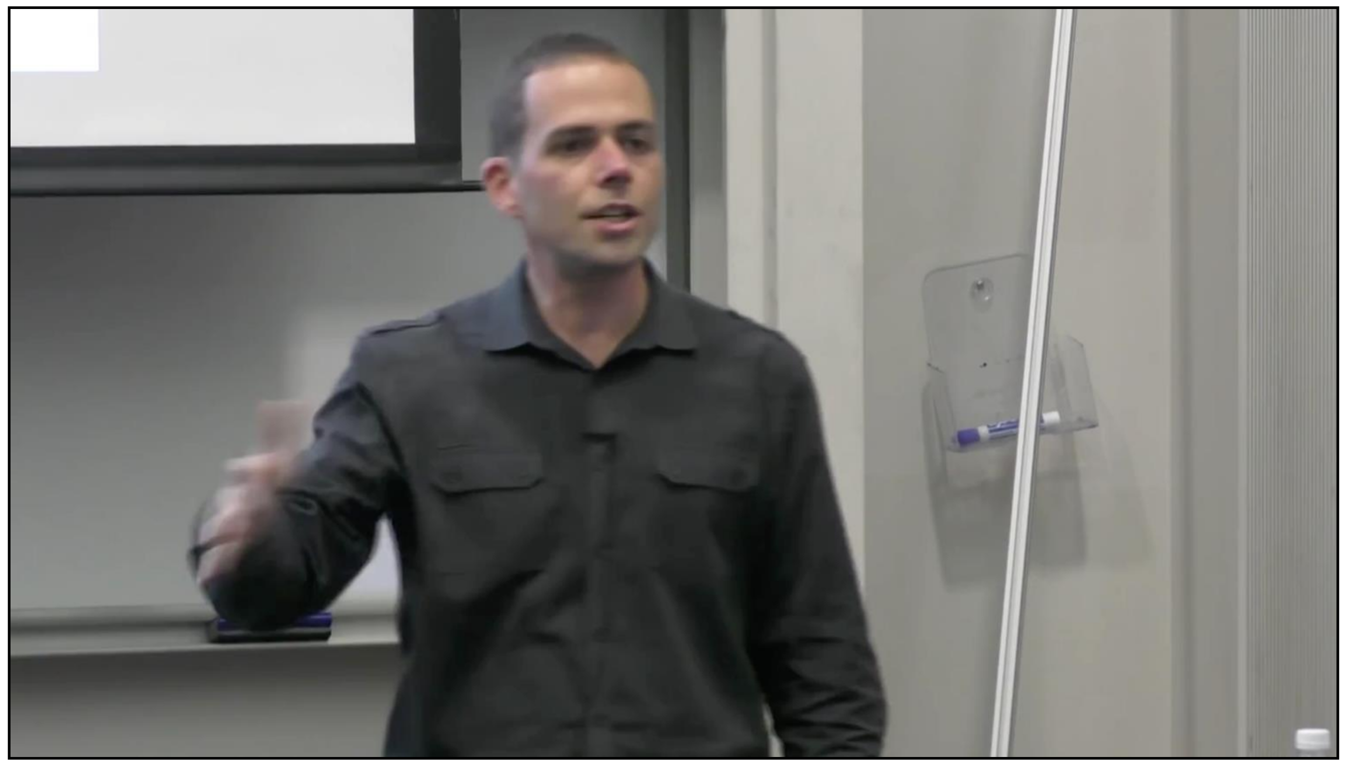
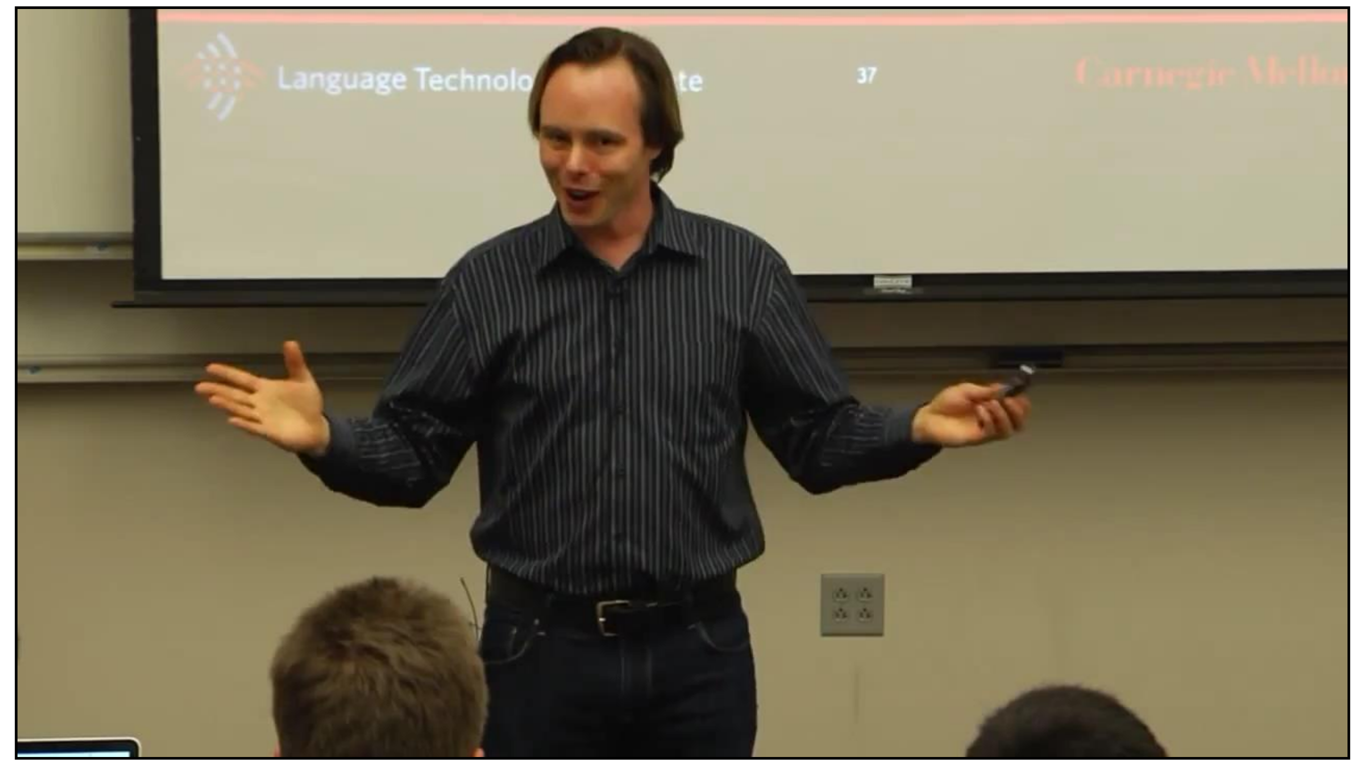

Gefen Kohavi

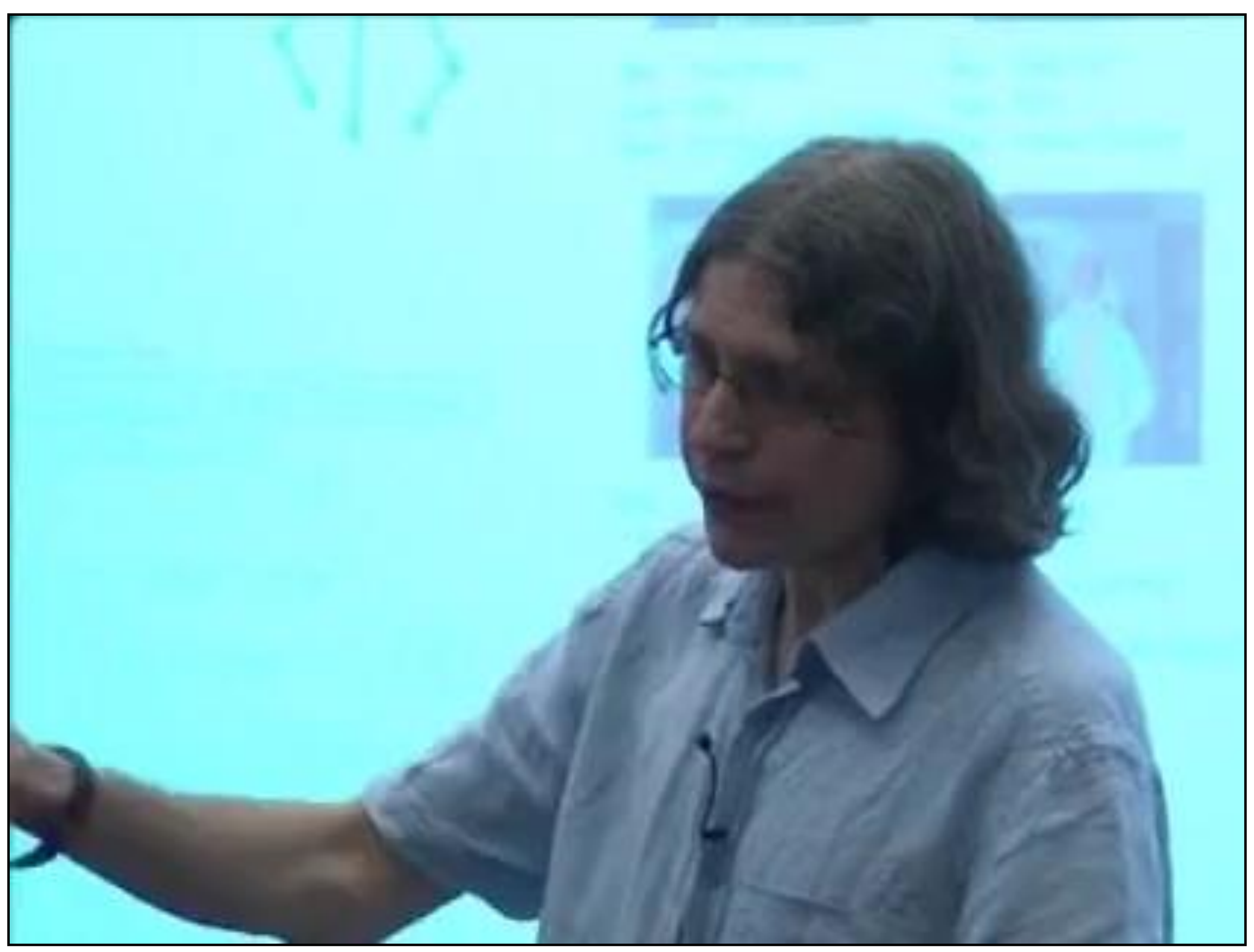
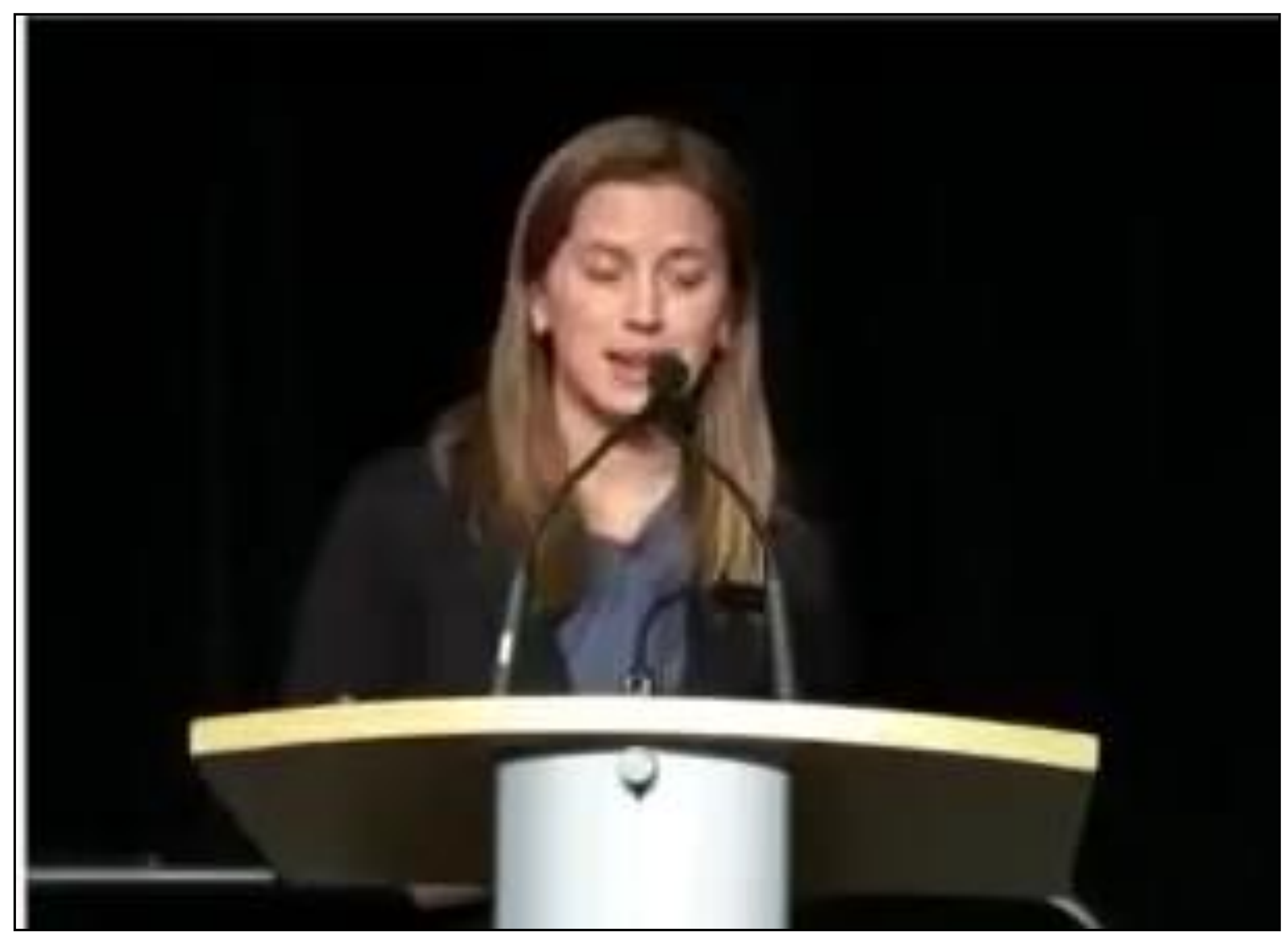

Caroline Chan


Andrew Owens


Jitendra Malik

CVPR 2019


Conversational gestures

Types of gestures:

- Iconic
- Metaphoric
- Beat
- Deictic
- Cohesive
- Emblem

Characteristic of an ***individual*** speaker.

Kendon’s continuum

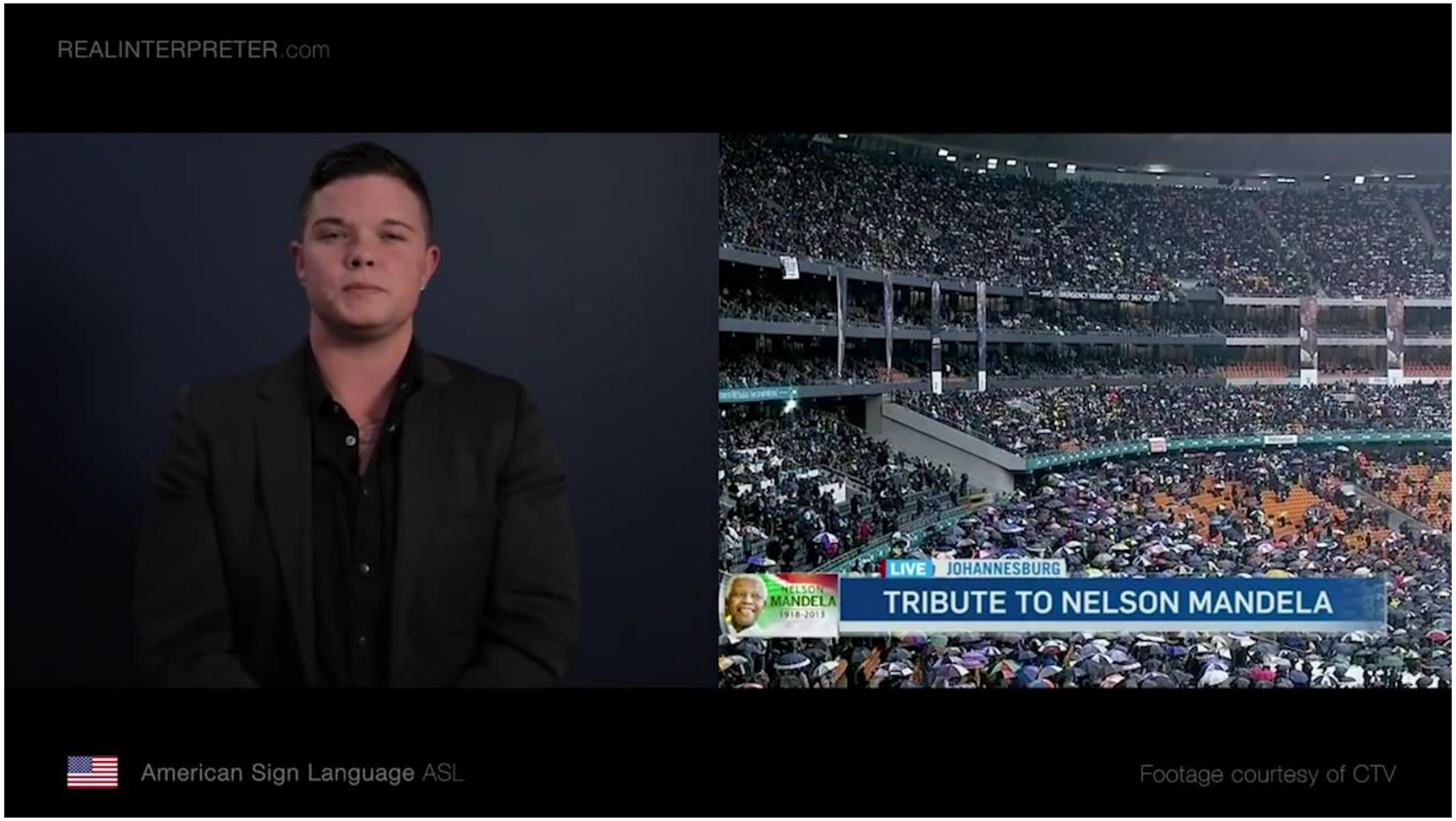
Spontaneous gesticulation | Language-like gestures | Pantomime | Emblems | Sign language

More speech,
idiosyncratic gesture

Less speech,
Socially-regulated signs

[Kendon, 2004]

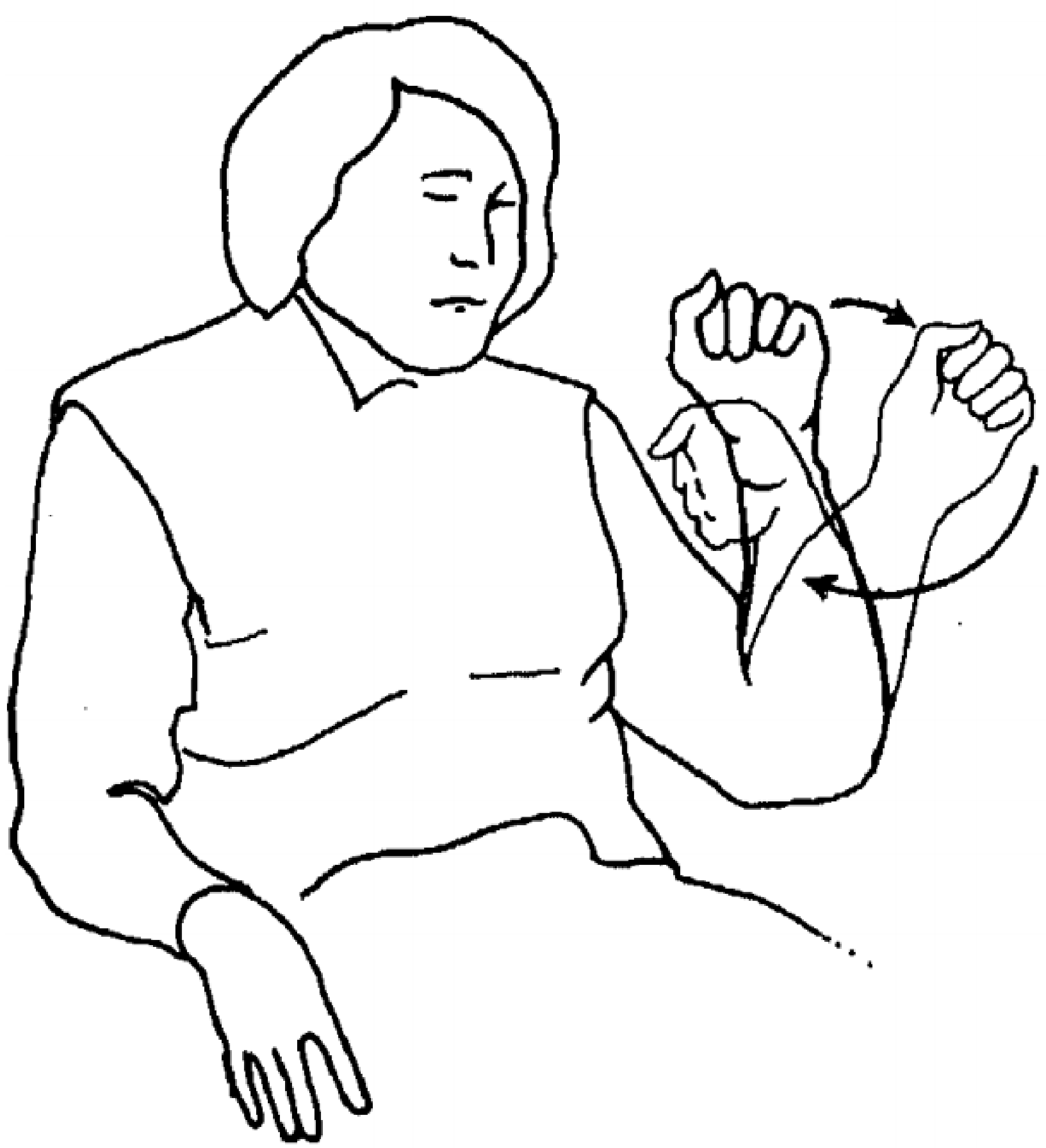
Sign language



Italianate

Related work

Psychology



McNeil 1994, Kendon 2004

Conversational agents

Cassell et al. 1994, Morency 2007, Levine et al. 2010

Sound-to-video

Suwajanakorn 2017, Shlizerman 2017, Chung 2019

Task: predict gesture from speech

Gestures dataset

10 subjects.

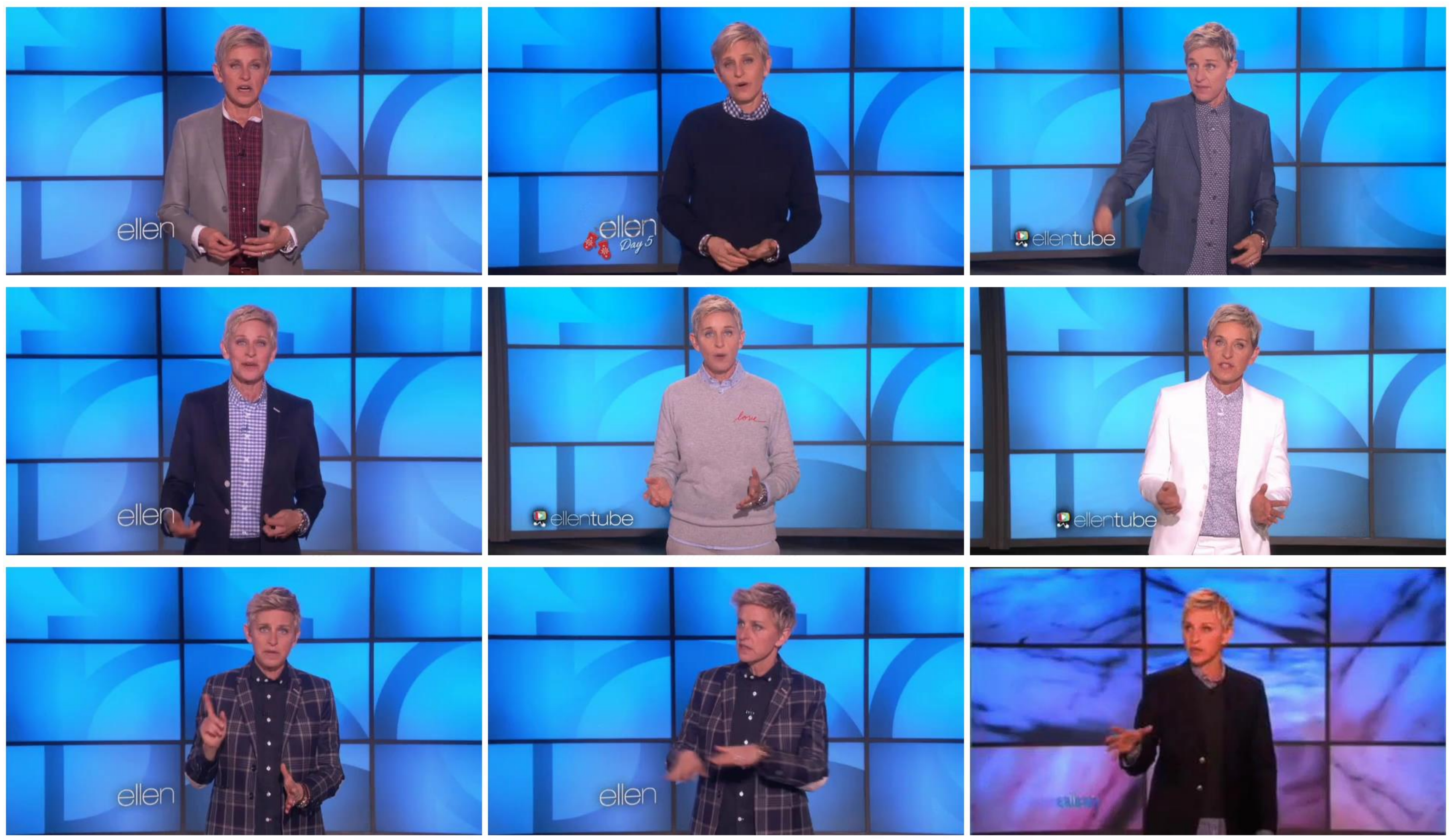
128 hours total.

Clean video intervals with frontal single speaker.

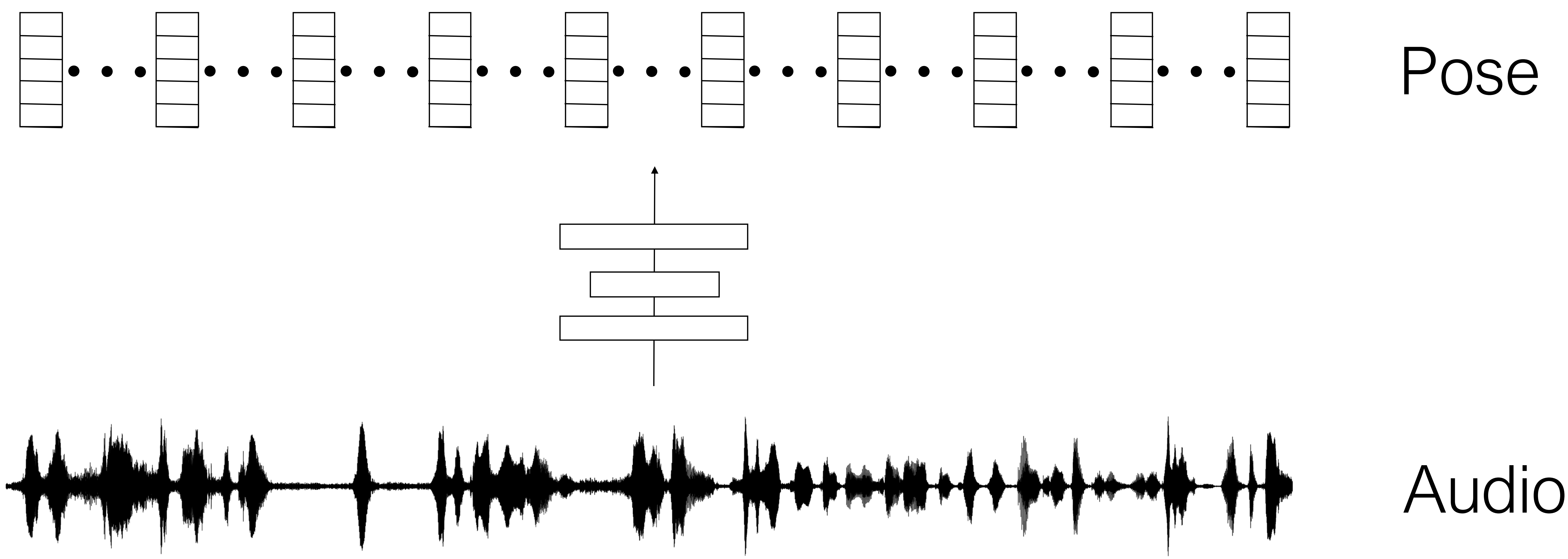
OpenPose detections for every frame.


ellentube

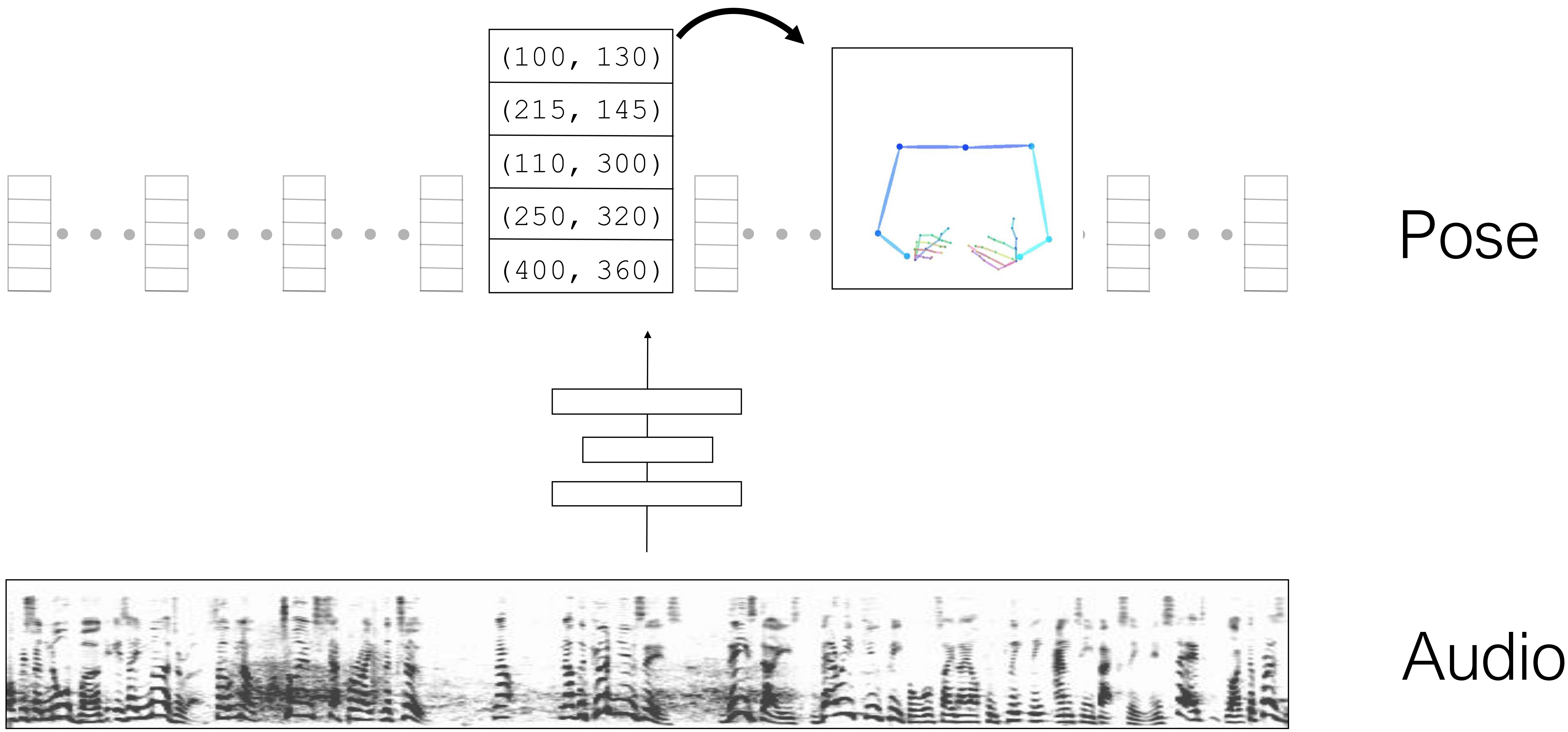
How do gesture styles differ across speakers?



Predicting gestures



Predicting gestures



Predicting gesture

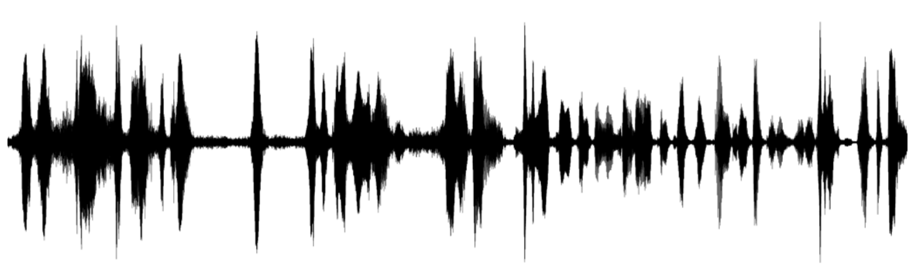
Audio input

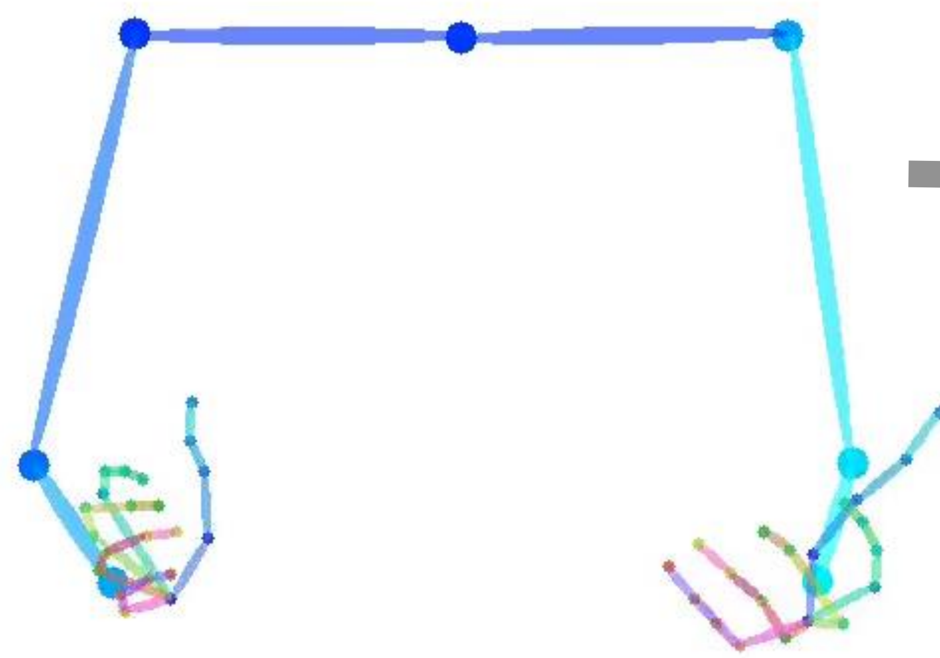
Predicting gesture

Audio input

Synthesizing a video

Ground truth face



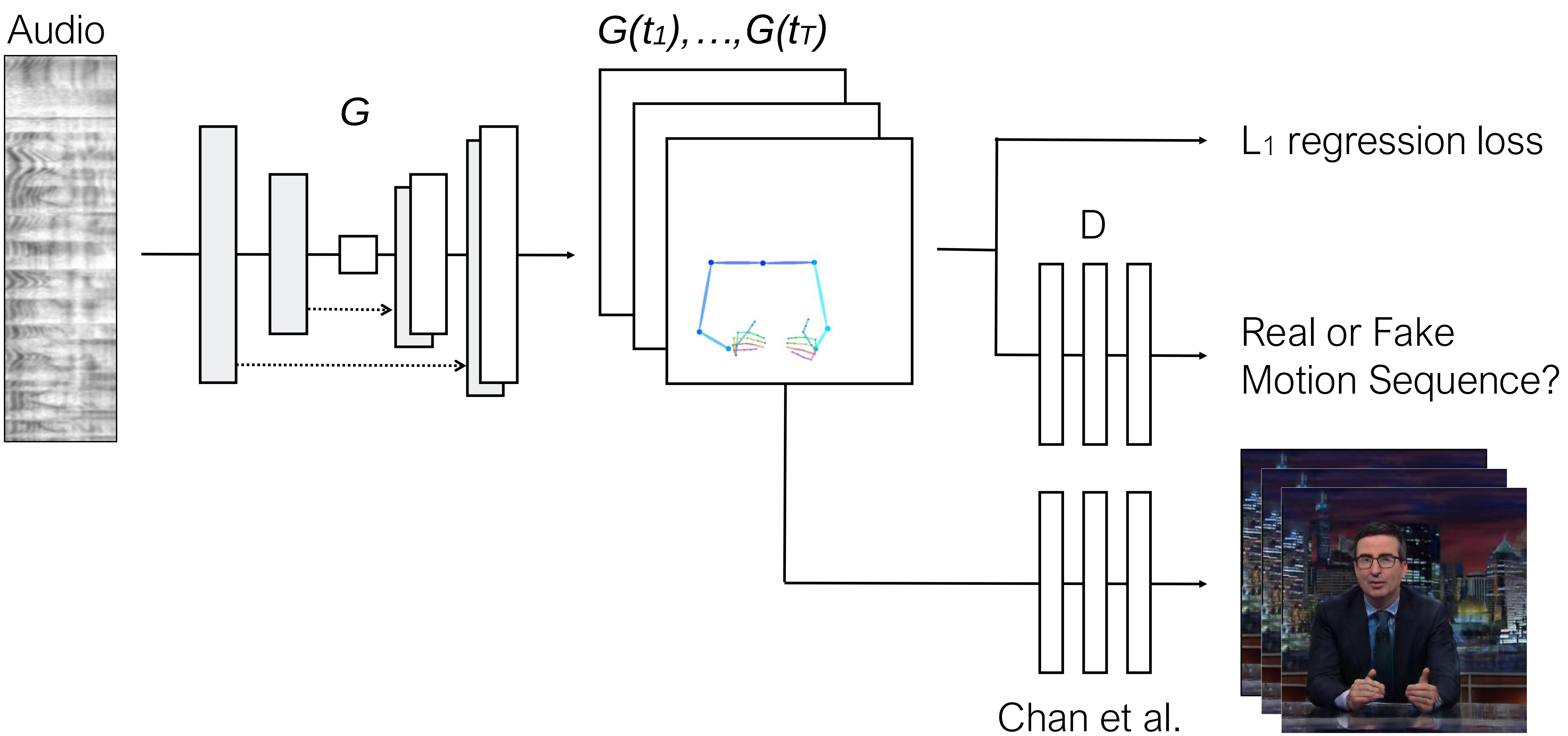


Audio input

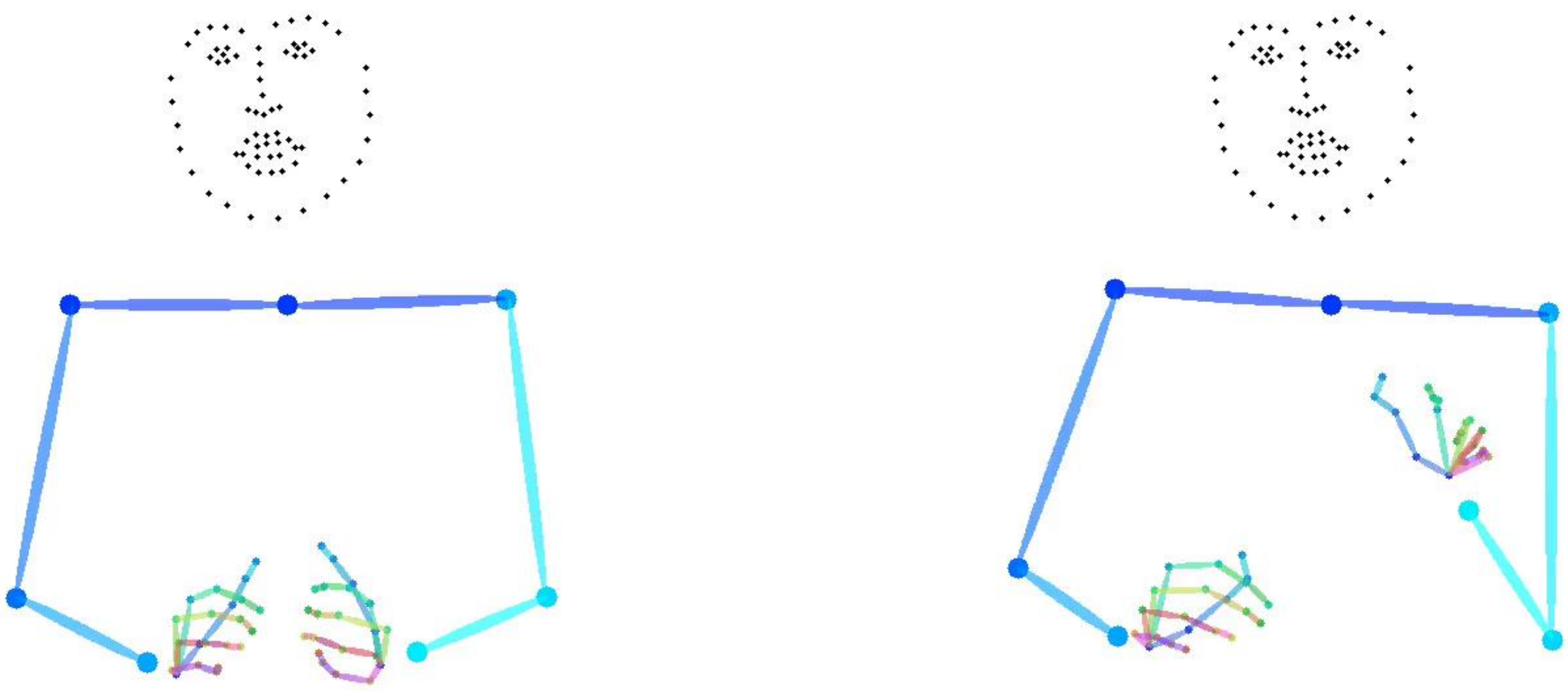
Predicted gestures

Face is ground truth

Method details

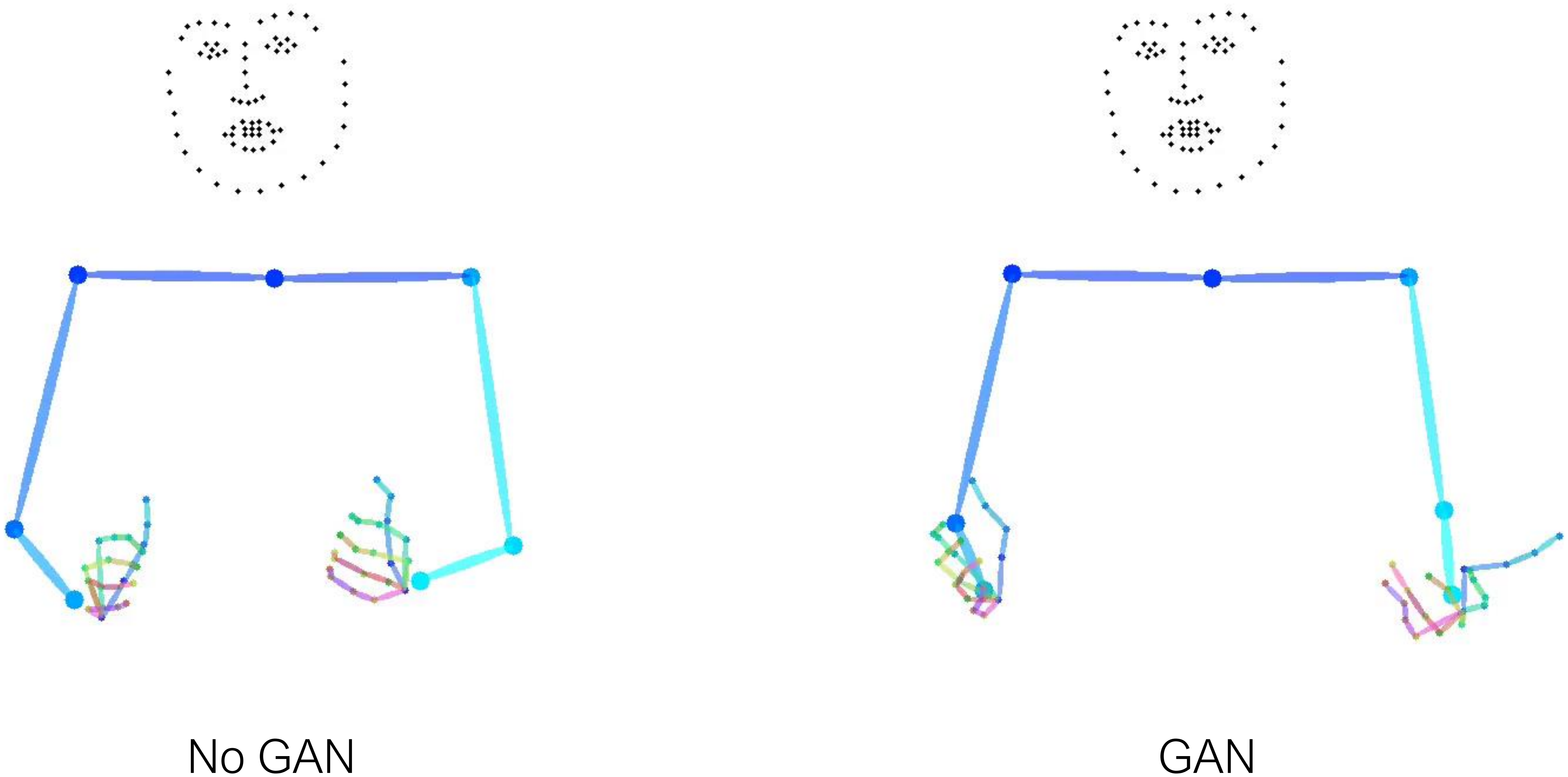


GAN loss snaps to individual styles of motion

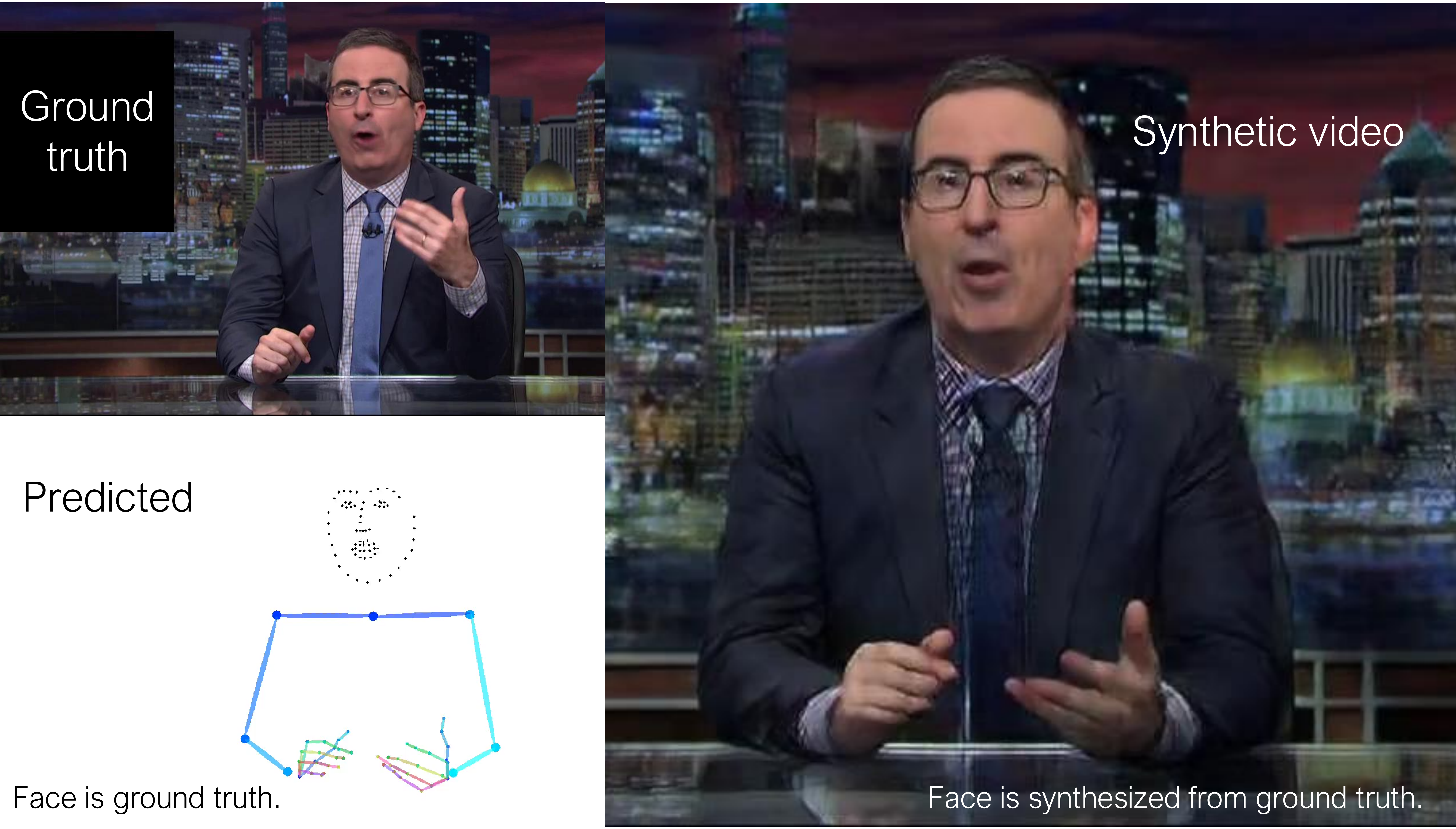


No GAN GAN

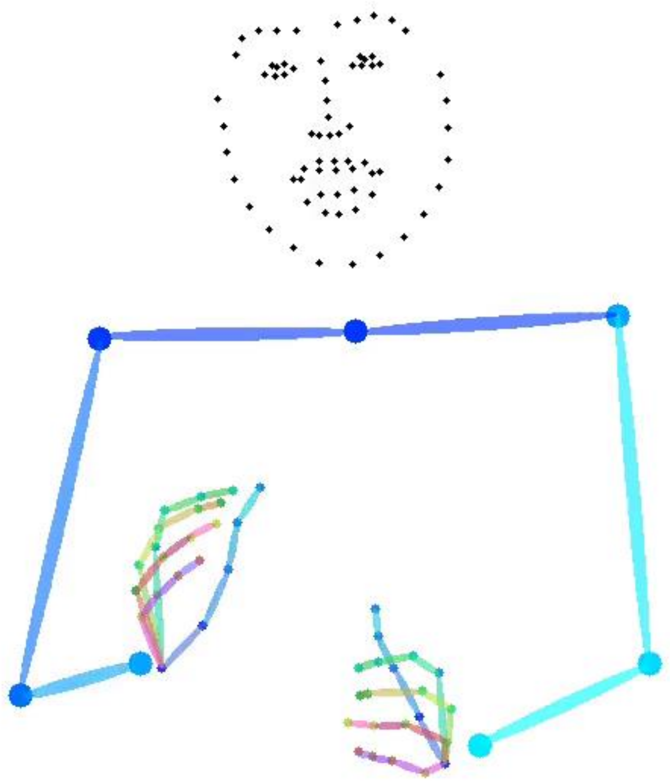
GAN loss snaps to individual styles of motion



Examples with ground truth footage

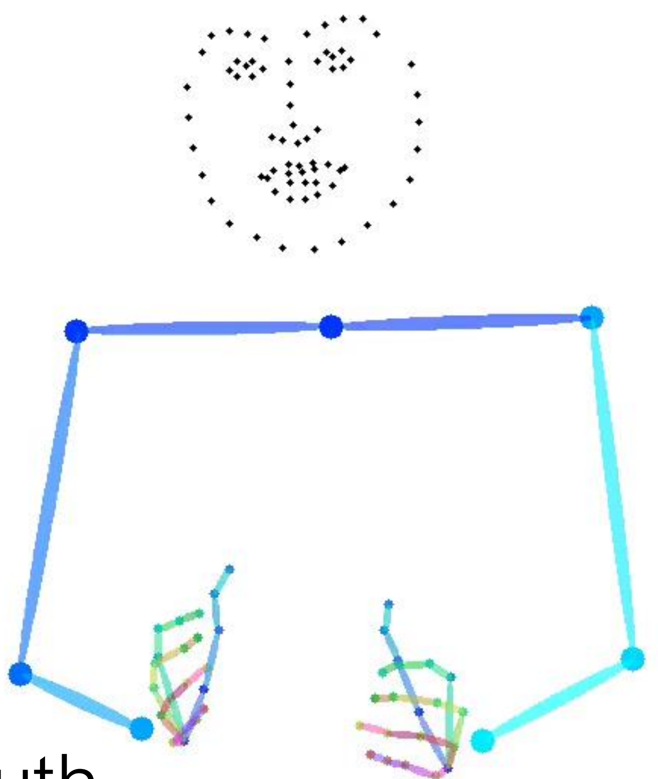

Ground truth
Predicted
Face is ground truth.
Synthetic video
Face is synthesized from ground truth.

Predicted



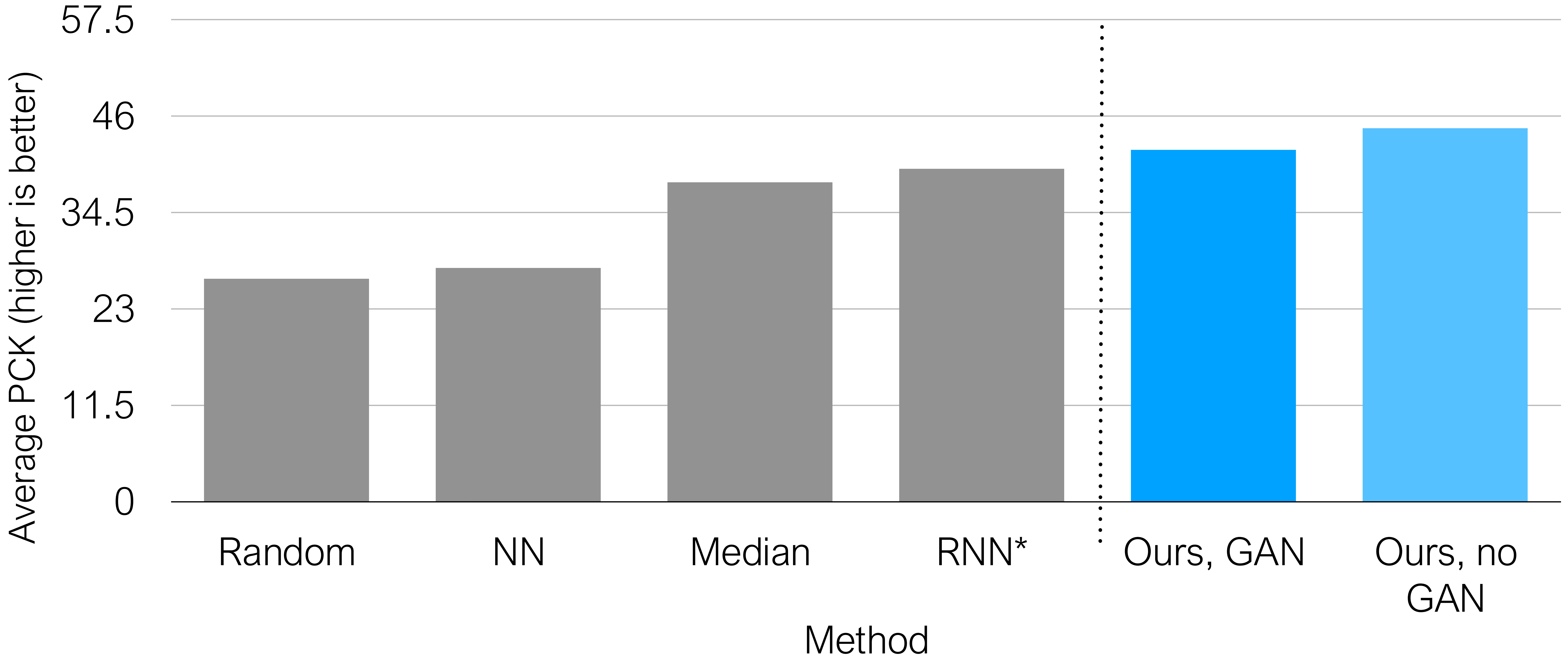
Face is ground truth.

Predicted



Face is ground truth.

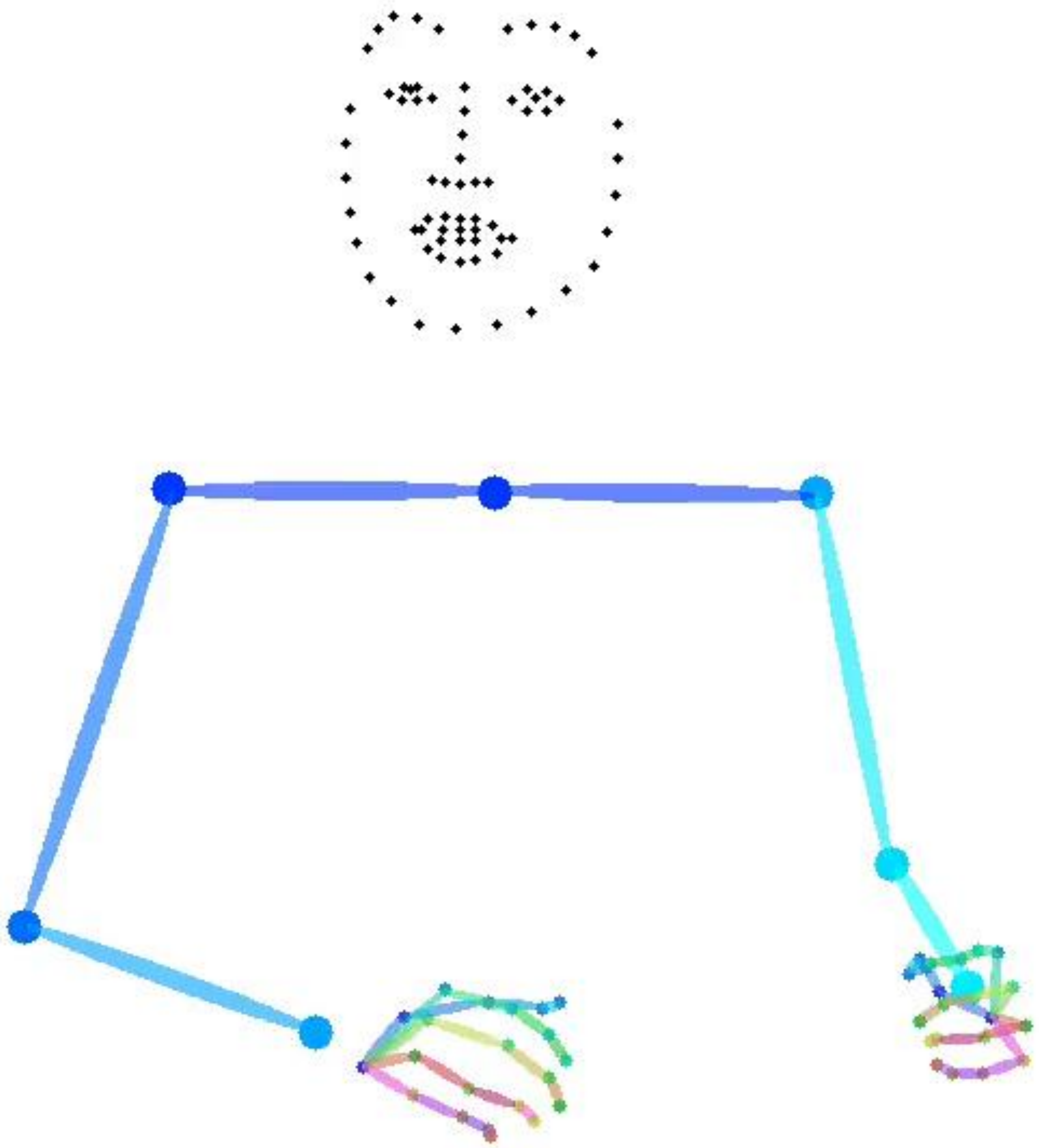
Quantitative evaluation



*[E. Shlizerman et al. Audio to body dynamics. 2018]

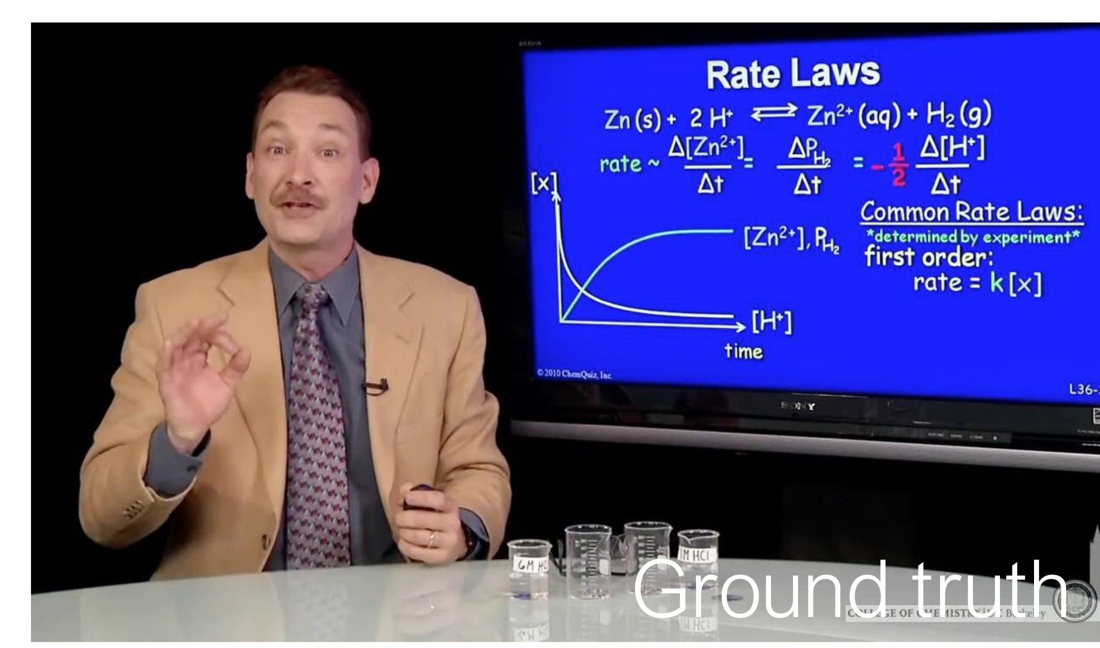
Prediction examples for other speakers
(with ground truth footage)

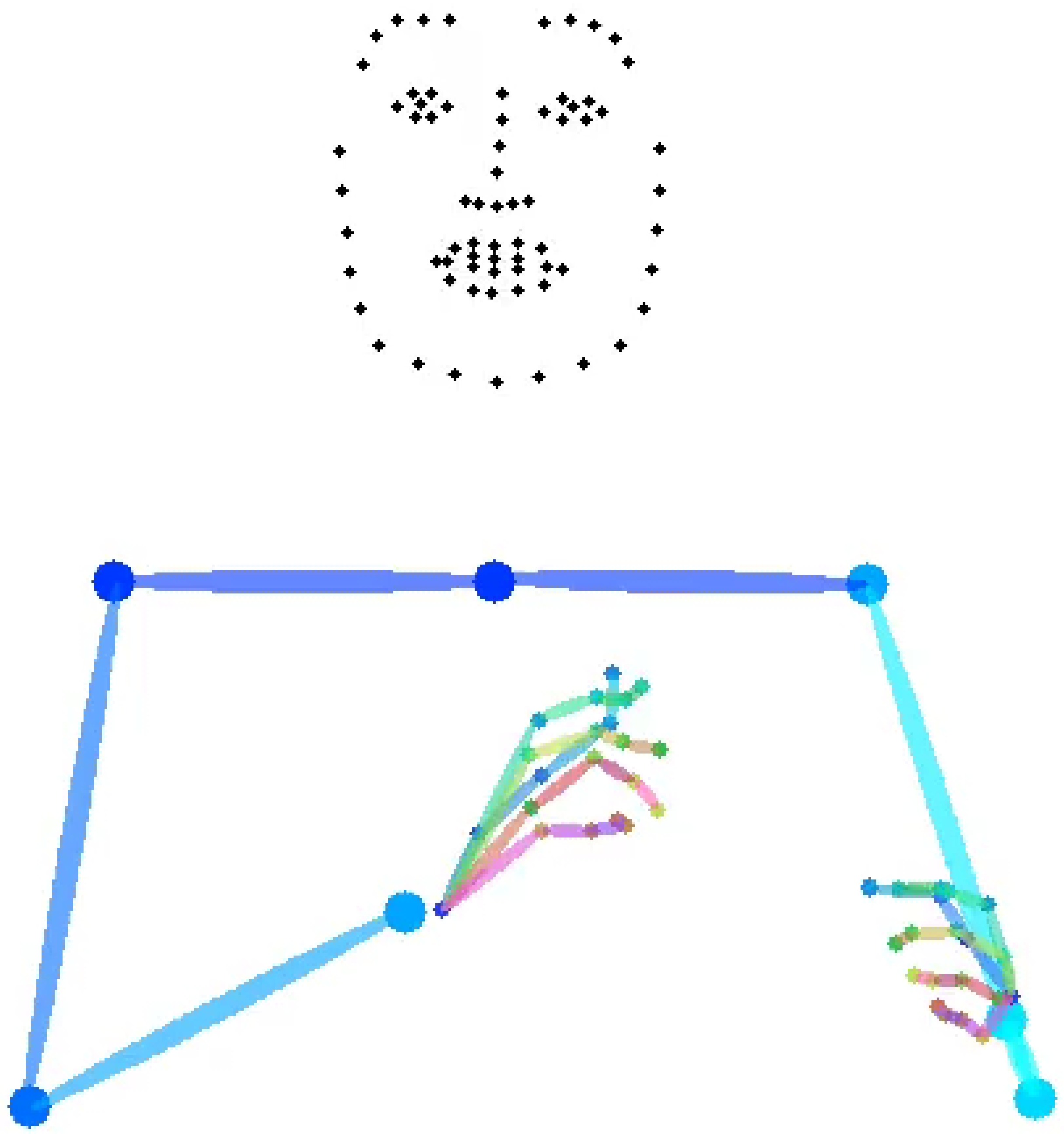
Mark Kubinec
(chemistry lecture)



Predicted from audio

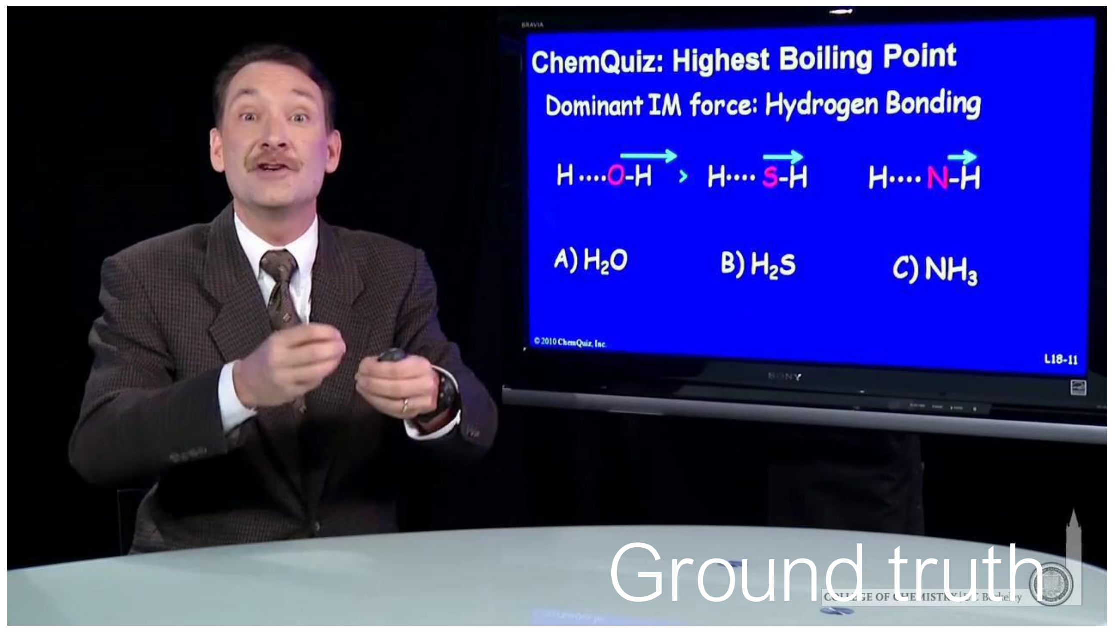
Face is ground truth





Predicted from audio

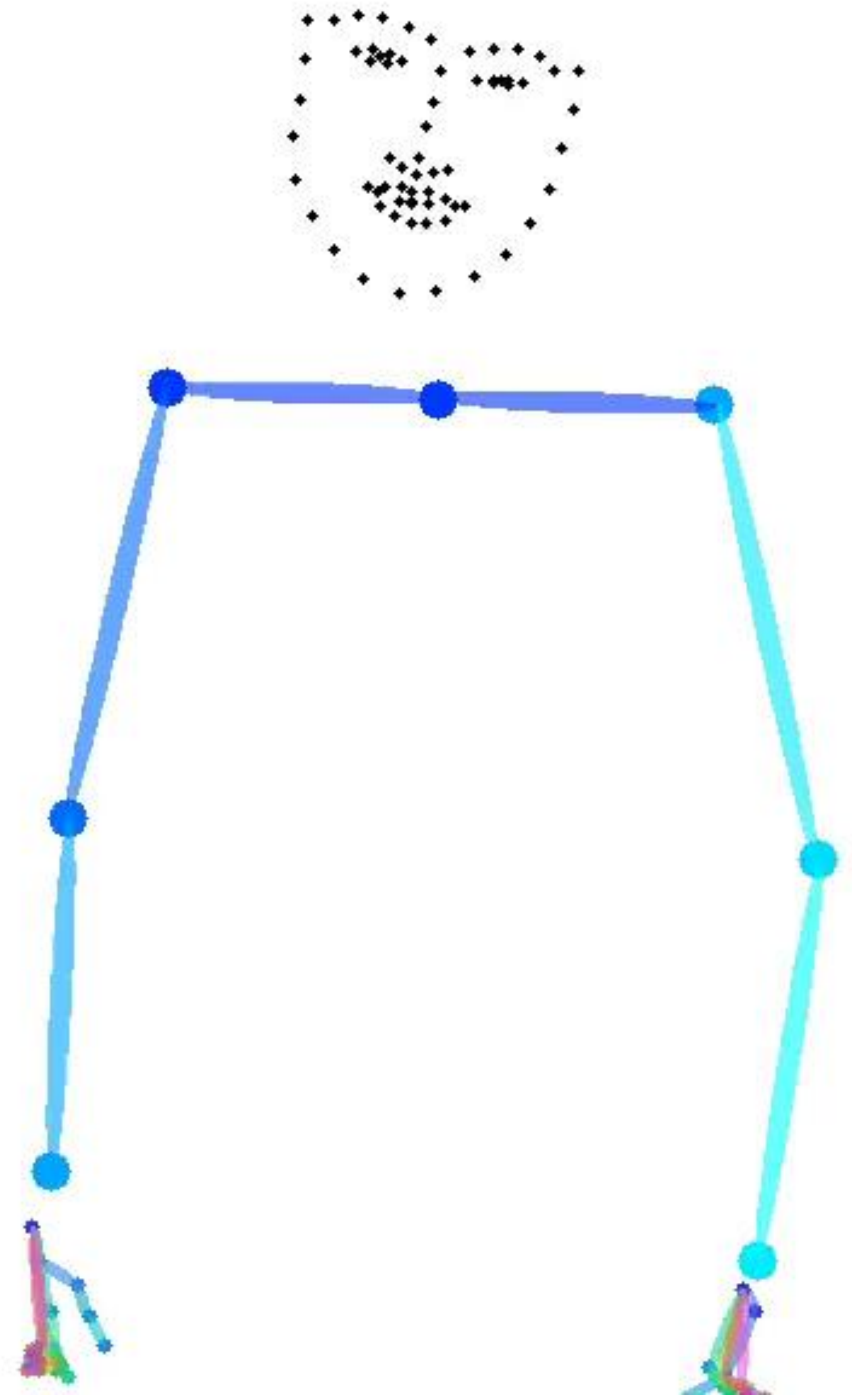
Face is ground truth

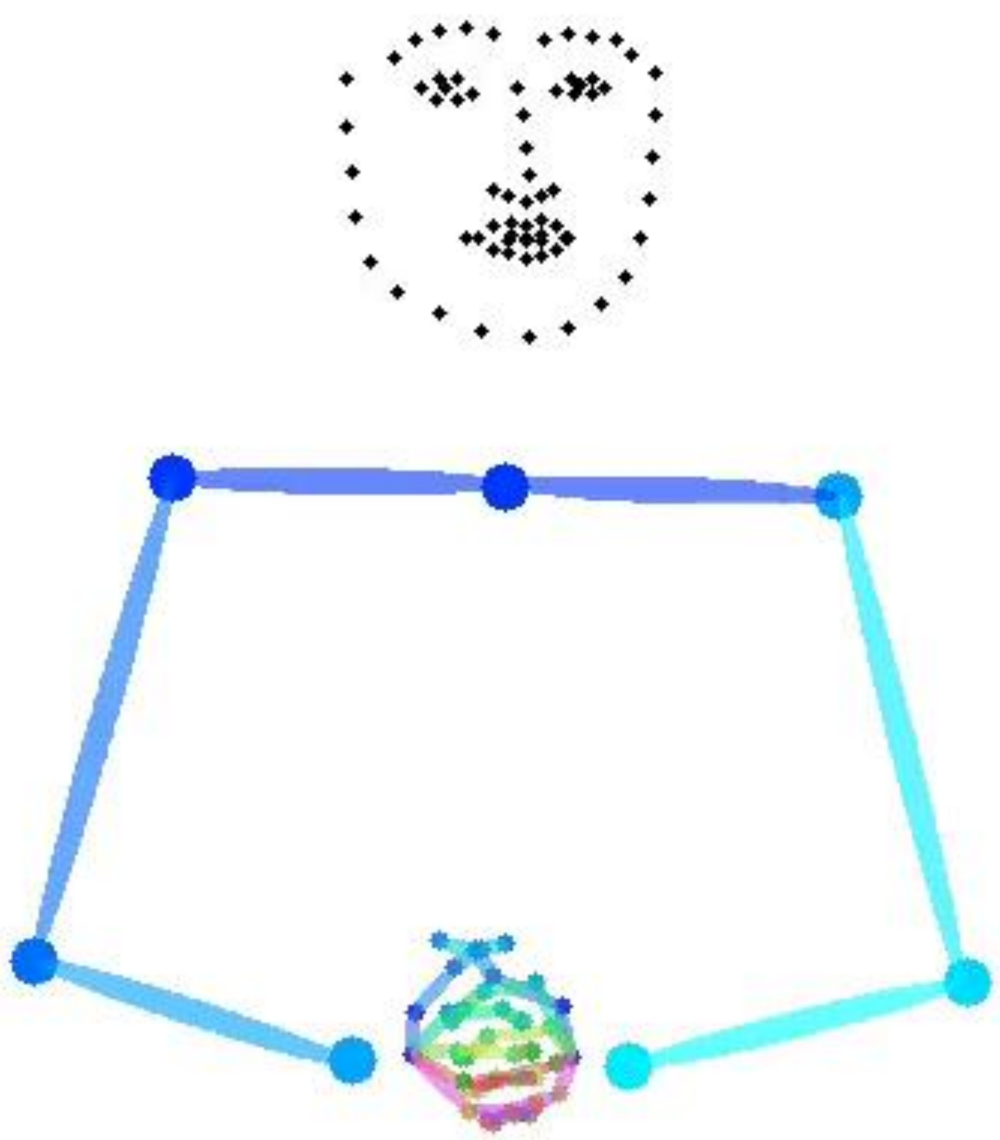


Conan O’Brien

Predicted from audio

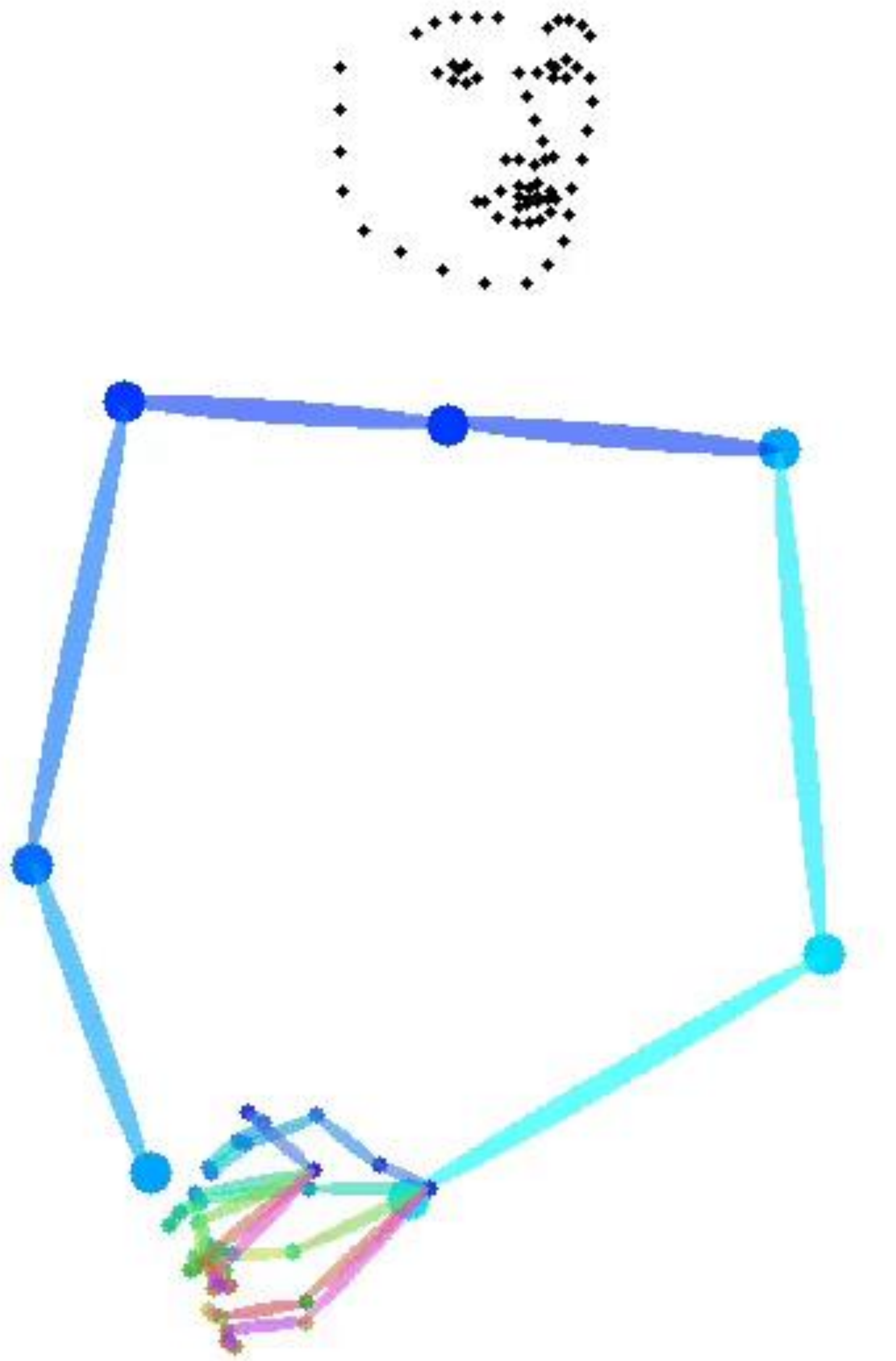
Face is ground truth





Predicted from audio

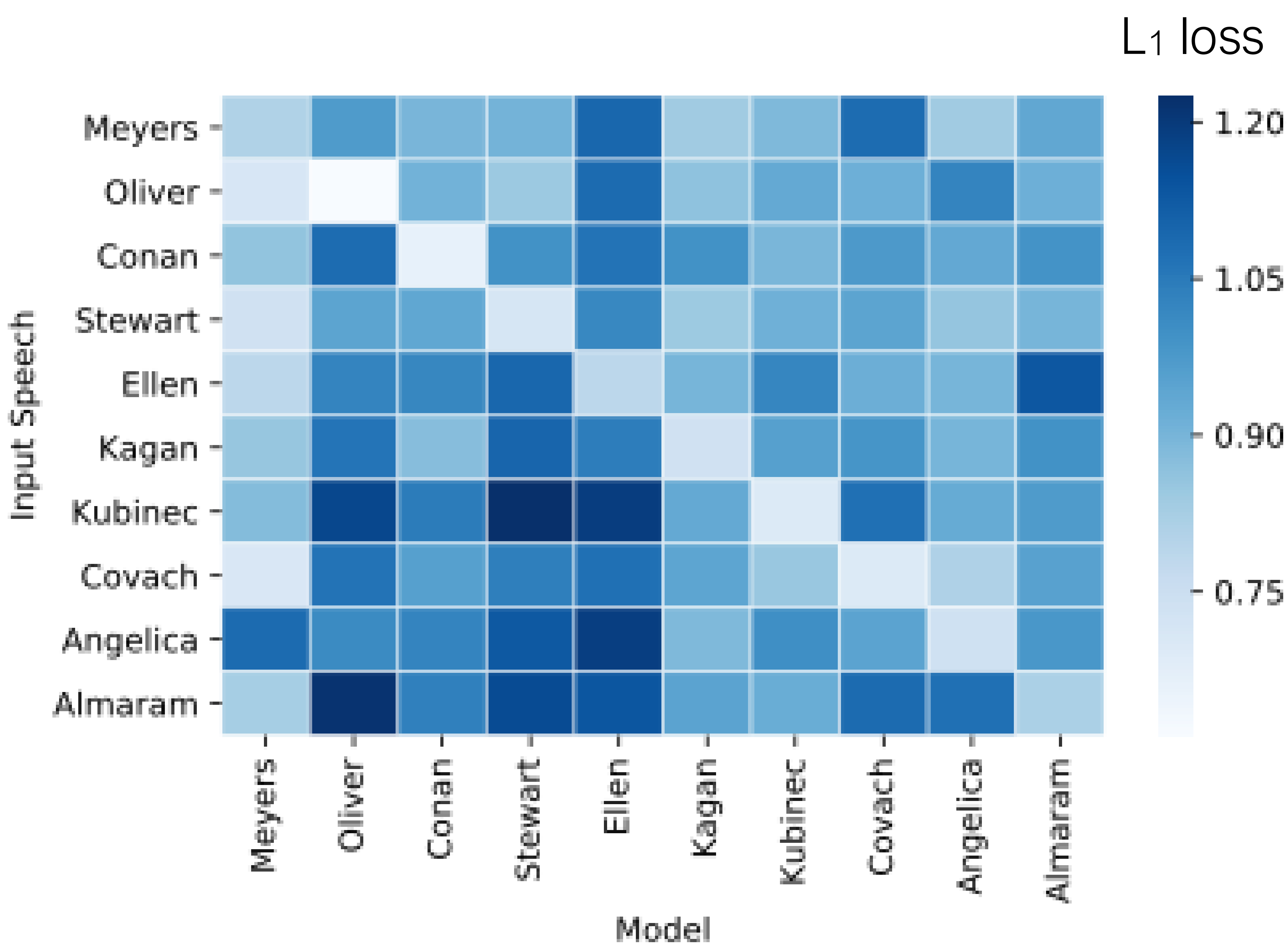
Face is ground truth



Predicted from audio

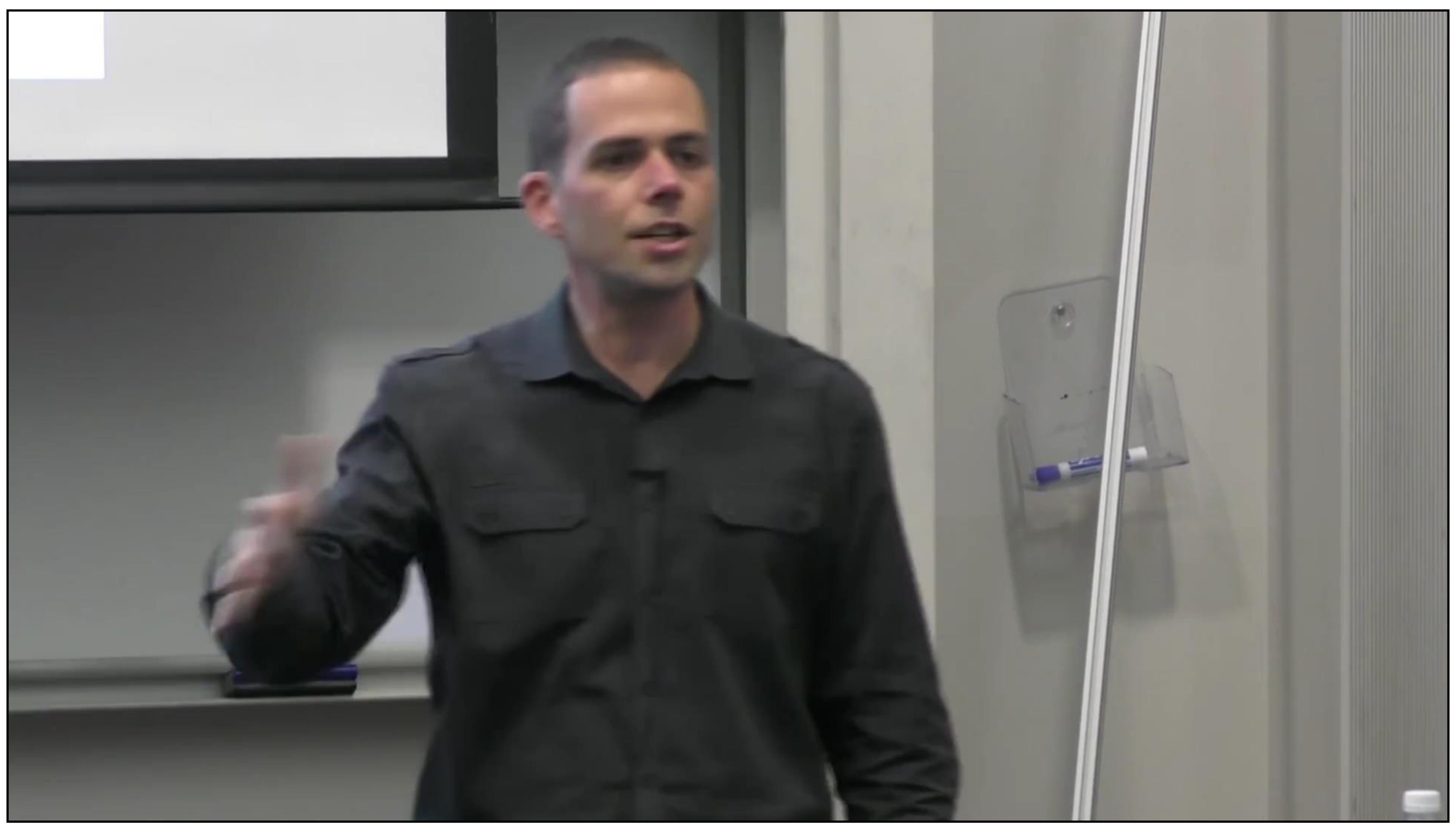
Face is ground truth

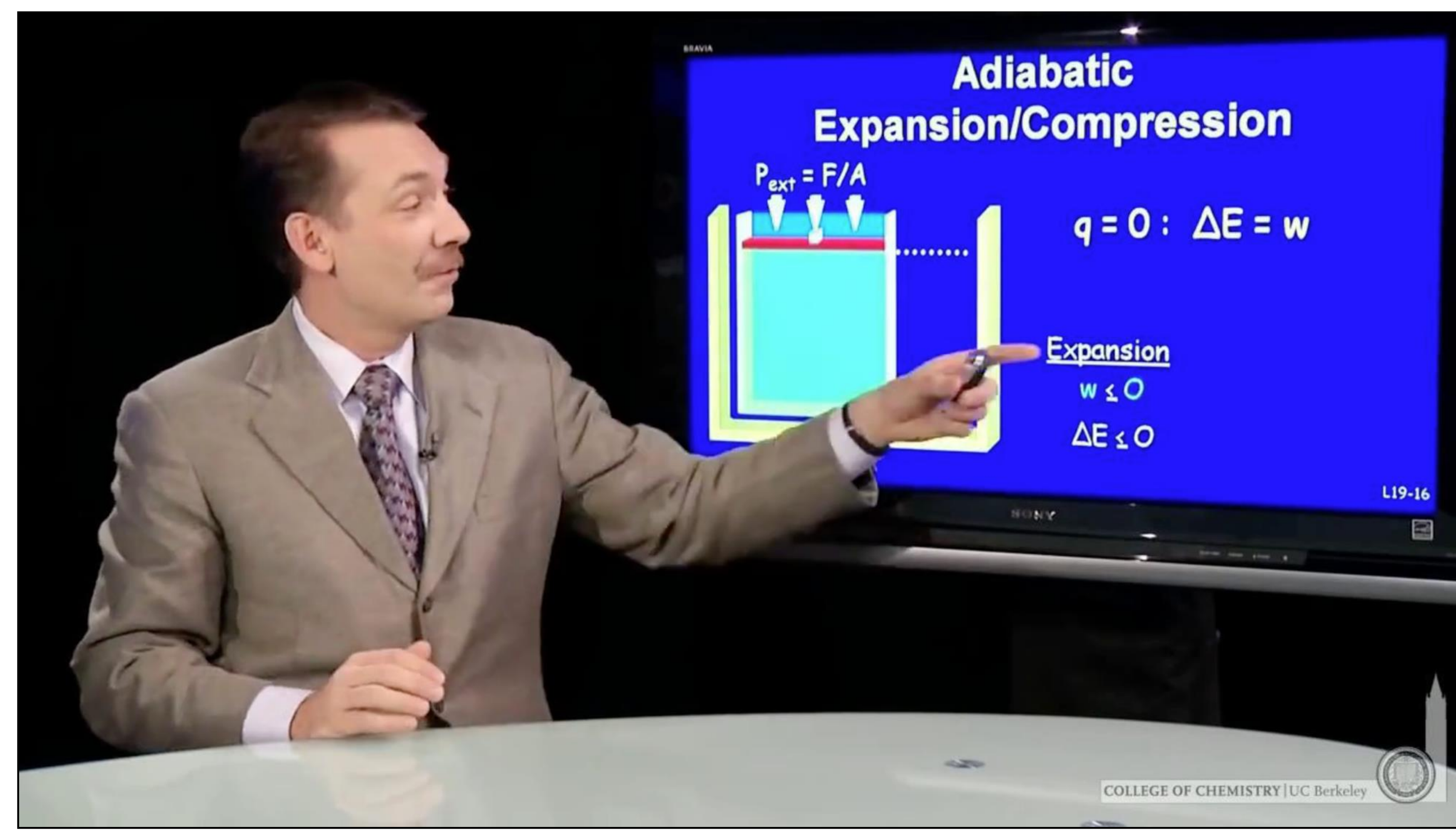
Gestures are person specific.



For every speaker audio input (row) we apply all other individually trained speaker models (columns).

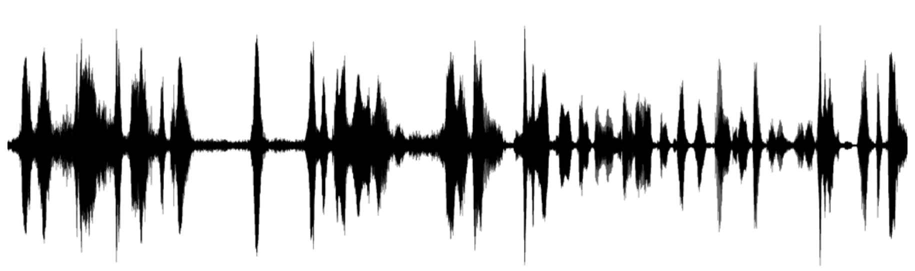
But we can still try to transfer!

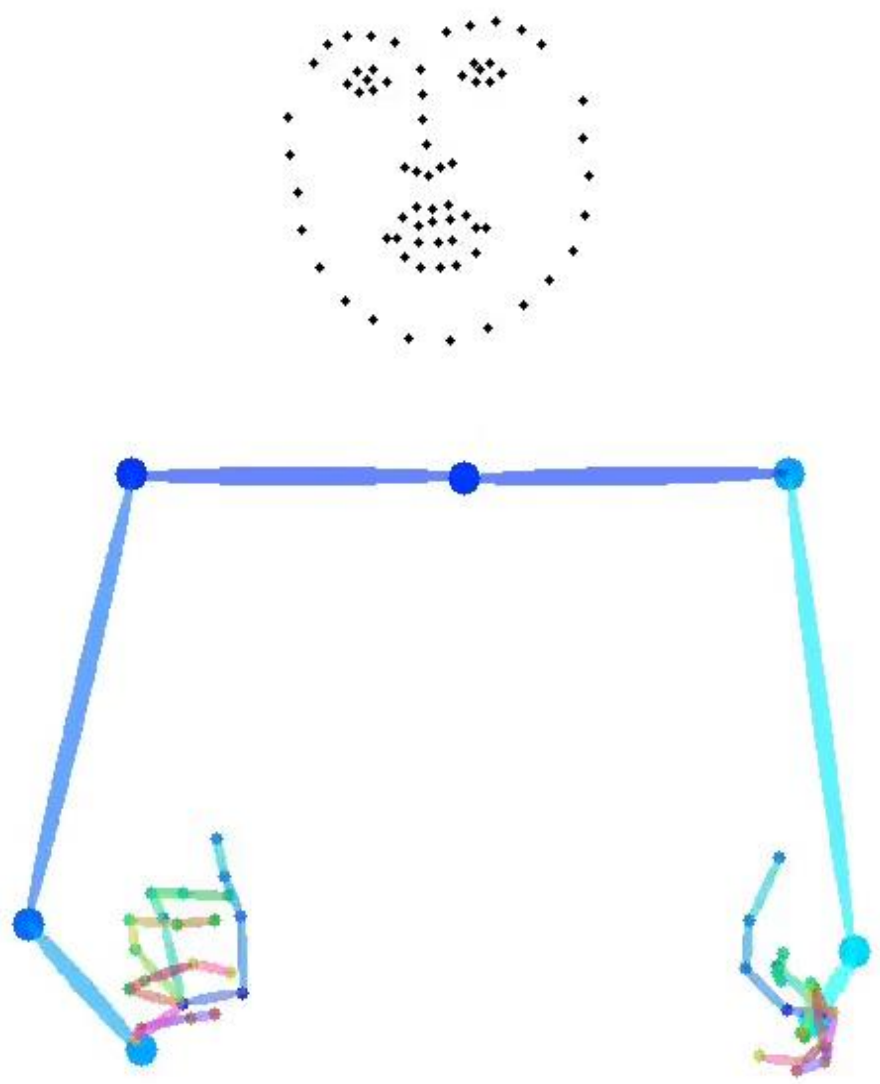




Vladlen-to-Kubinec

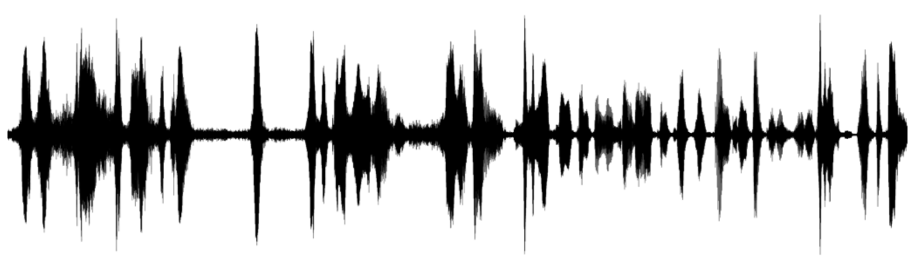
Poster #102
Tuesday, 15:20 –18:00
Exhibit Hall

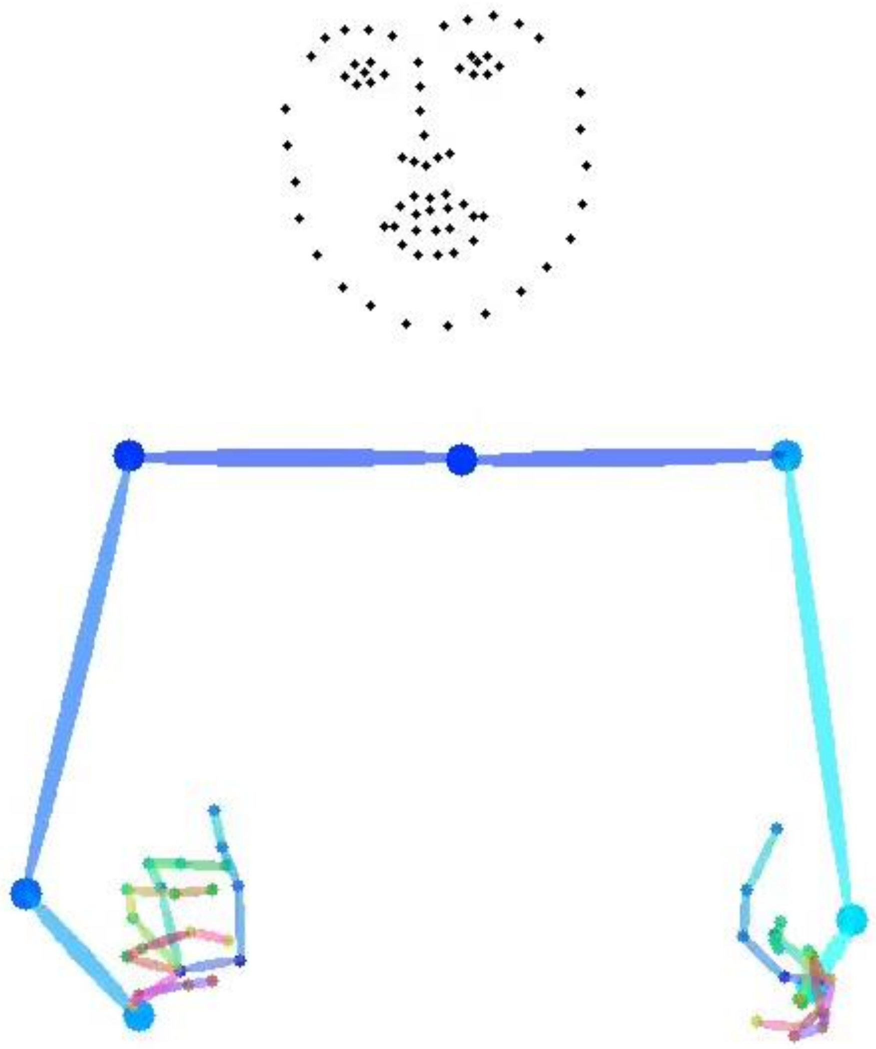




http://people.eecs.berkeley.edu/~shiry/speech2gesture/

Thank you!





http://people.eecs.berkeley.edu/~shiry/speech2gesture/

Open questions in social perception

- Computers today have pitifully low “social intelligence”
- We need to understand the internal state of humans as they interact with each other and the external world
- This includes emotional state, body language, current goals. We currently are able to capture “surface manifestations”
- Hierarchical understanding of human behavior is the holy grail. Computer Vision can contribute to it and be aided by advances in it.