# COMBINING IMAGES AND WORDS IN DEEP NETWORKS THAT IDENTIFY PEOPLE FROM BODY SHAPE


ALICE J. O'TOOLE

*THE UNIVERSITY OF TEXAS AT DALLAS*



ICPRAM 2024

13th International Conference on Pattern Recognition Applications and Methods

Rome, Italy

24 - 26 February, 2024

# ACKNOWLEDGEMENTS


- UT-Dallas
  - Blake A. Myers
  - Matthew Q. Hill
  - Veda Nandan Gandi
  - Thomas Metz
  - Lucas Jaggernauth
  - Madeline Rachow
- Johns Hopkins Univ.
  - Rama Chellappa
  - Carlos D. Castillo (Amazon)
  - Ram Prabhakar
- STR – team lead

- UT-Dallas
  - Matt Hill
  - Carina Hahn
- MPI for Intelligent Machinery
  - Stephan Streuber
  - Michael Black

Research funded by BRIAR Program and Intelligence Advanced Research Projects Activity (IARPA), via [2022-21102100005].

# OVERVIEW

- Problem – person identification based on body shape
  - Biometric Recognition and Identification at Altitude and Range (BRIAR)
    - IARPA https://www.iarpa.gov/research-programs/briar

- Linguistic descriptors to "quantify" body shape
  - psychology, computer graphics

- Body identification networks:
  - linguistic descriptors
  - object-based shape descriptors

- Person recognition = face + body + gait
  - fusion

# PROBLEM

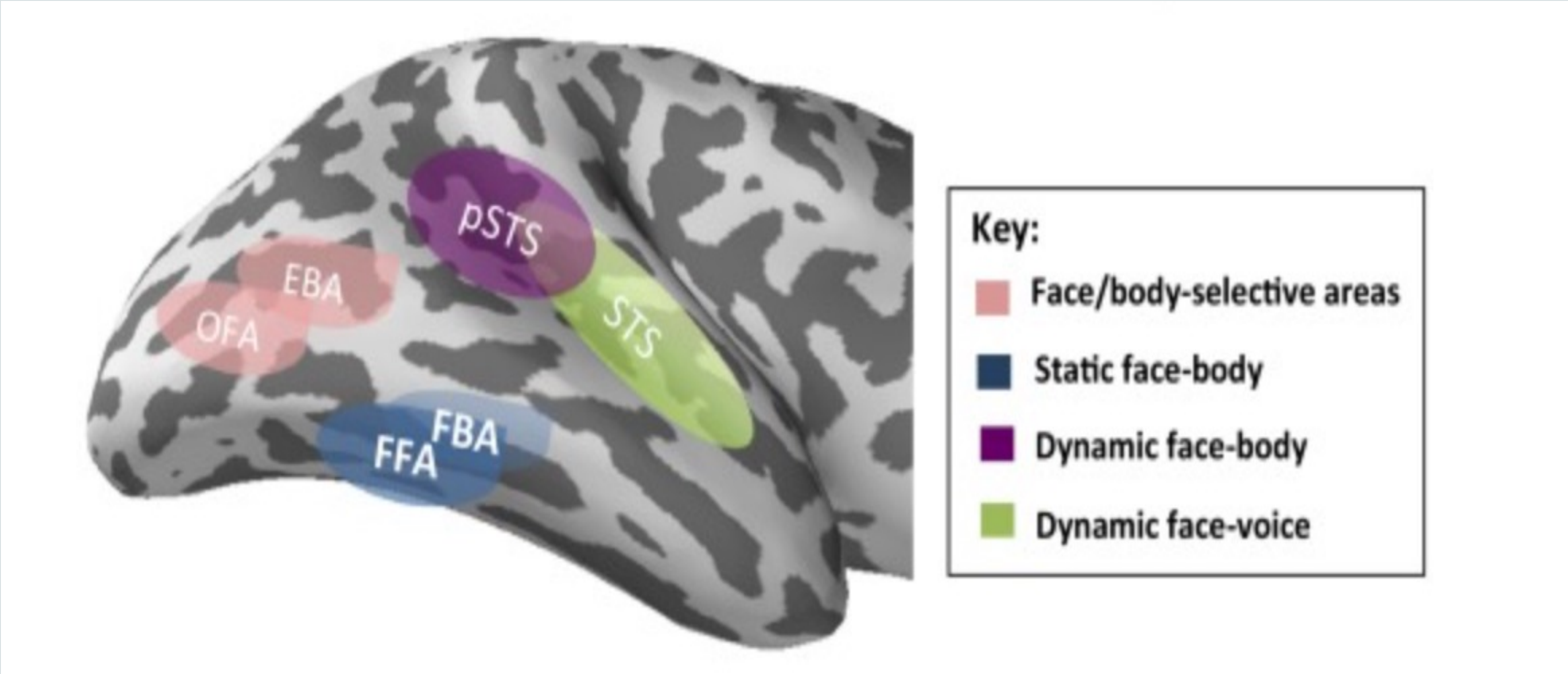
*face*

*body*

*gait*



subjects consented to publication




Yovel & O'Toole (2016)


subject consented to publication

*same person or different people?*

*face*

*body*

*gait*

subjects consented to publication

# BODY AS A BIOMETRIC

- Why use body?
  - visible at large distances
  - subset of cues constant over change in view
    - height, weight, proportions, rough shape
  - “fusability”

  - *You have no other option!*

# BODY IS LEAST COMMON DENOMINATOR

subjects consented to publication

# BODY AS A BIOMETRIC

- Why not use body?
  - not unique
  - lack of a body algorithms
    - face
    - gait
    - body

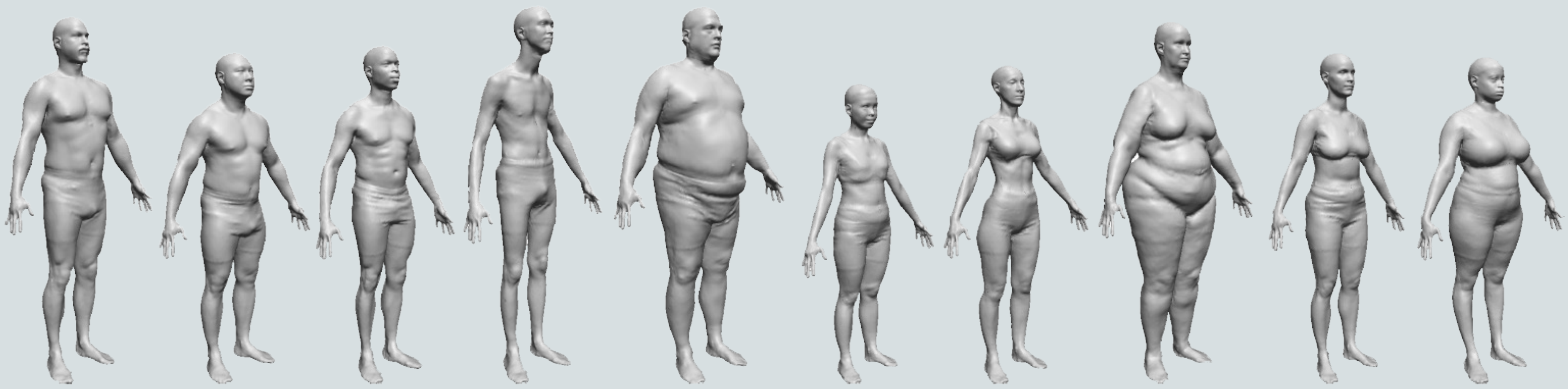
***Bo*** ***n***

# LINGUISTIC DESCRIPTIONS
# &
# 3D BODY SHAPES

Hill, Matthew Q., et al. "Creating body shapes from verbal descriptions by linking similarity spaces." *Psychological science* 27.11 (2016): 1486-1497.

Streuber, Stephan, et al. "Body talk: Crowdshaping realistic 3D avatars with words." *ACM Transactions on Graphics (TOG)* 35.4 (2016): 1-14.

# HUMAN BODY SHAPE



- body = complex 3D shape
  - Laser scan = 12500 vertices and 25000 facets

# LINGUISTIC DESCRIPTIONS OF BODIES

- muscular, athletic
- stout, portly
- shapely, hourglass

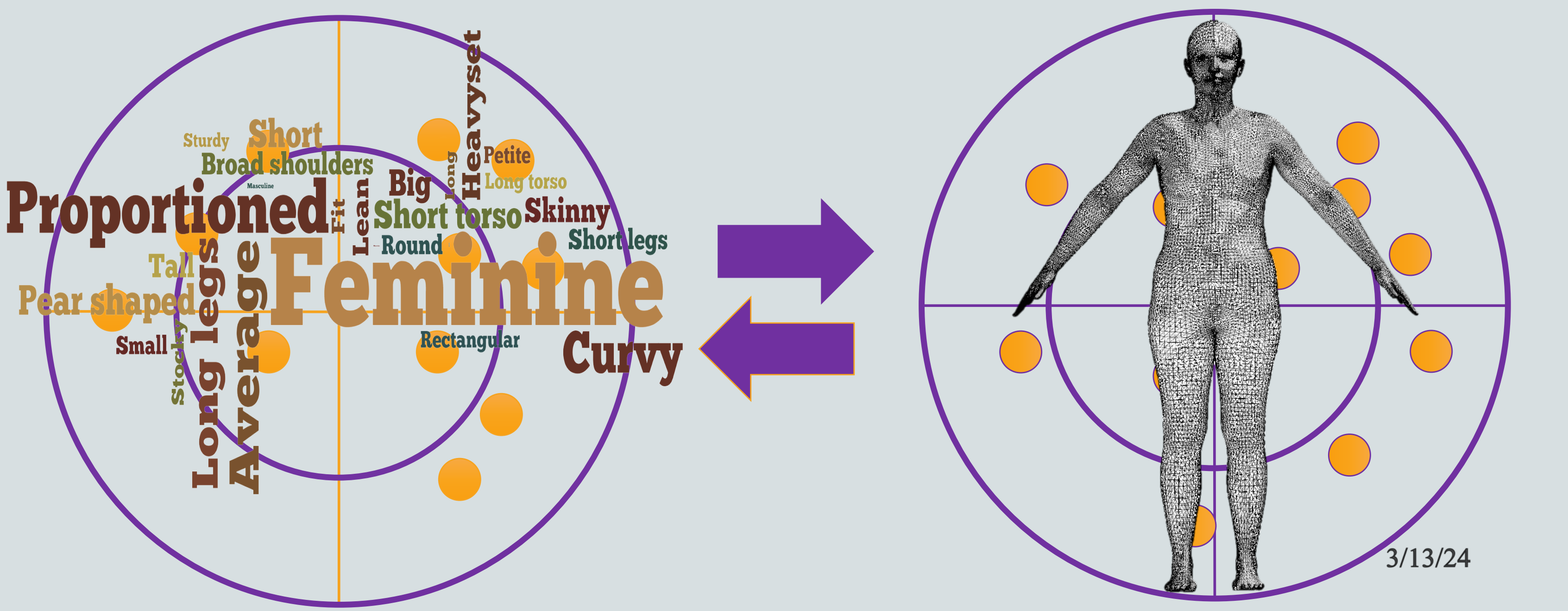
# RATIONALE

- human language and vision
  - evolved a long time ago
    - 50K and 2 million years ago
    - words don't leave fossils or tool fragments
  - language communicates information efficiently
    - not every vertex in the laser scan carries a lot of information

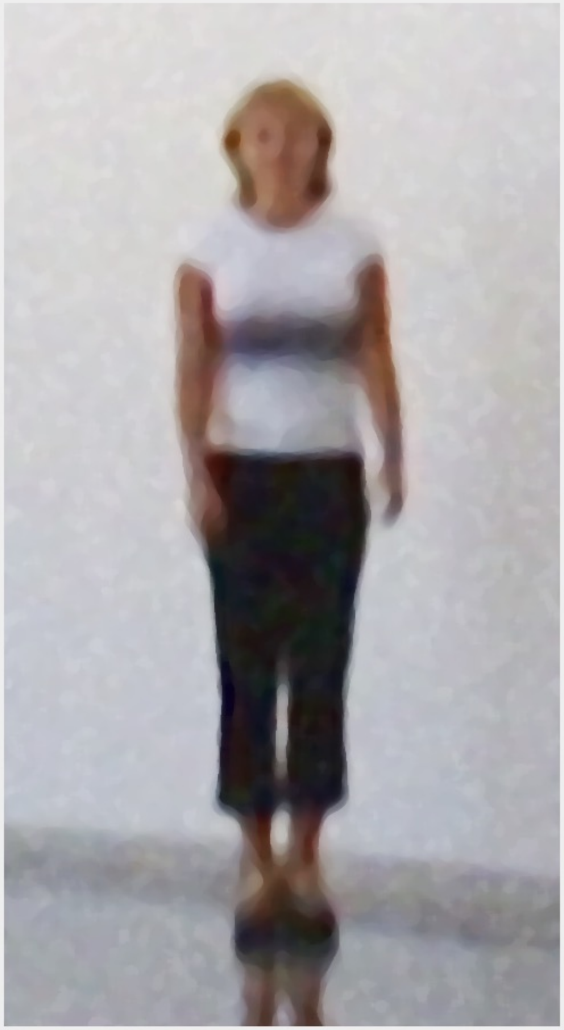
long legs lean curvy

# APPROACH

- human descriptions to create a similarity space
  - point proximity -> similarity between ***body descriptions***
- geometric shape space to *ground-truth* description space
  - point proximity -> similarity between ***body shapes***



# BODY DESCRIPTIONS



| | Does Not Apply | Applies Somewhat | Applies Perfectly |
|---|---|---|---|
| Proportioned | ◉ | ○ | ○ |
| Rectangular | ◉ | ○ | ○ |
| Stocky | ◉ | ○ | ○ |
| Short legs | ◉ | ○ | ○ |
| Muscular | ◉ | ○ | ○ |
| Average | ◉ | ○ | ○ |
| Tall | ◉ | ○ | ○ |
| Sturdy | ◉ | ○ | ○ |
| Big | ◉ | ○ | ○ |
| Long legs | ◉ | ○ | ○ |
| Lean | ◉ | ○ | ○ |
| Short torso | ◉ | ○ | ○ |
| Pear shaped | ◉ | ○ | ○ |
| Petite | ◉ | ○ | ○ |
| Broad shoulders | ◉ | ○ | ○ |
| Heavyset | ◉ | ○ | ○ |
| Long | ◉ | ○ | ○ |
| Long torso | ◉ | ○ | ○ |
| Round (Apple) | ◉ | ○ | ○ |
| Built | ◉ | ○ | ○ |
| Fit | ◉ | ○ | ○ |
| Skinny | ◉ | ○ | ○ |
| Masculine | ◉ | ○ | ○ |
| Small | ◉ | ○ | ○ |
| Short | ◉ | ○ | ○ |
| Feminine | ◉ | ○ | ○ |
| Curvy | ◉ | ○ | ○ |

NEXT PAGE



subjects consented to publication

| Descriptor Terms | | | |
|---|---|---|---|
| proportioned | sturdy | broad shoulders | skinny |
| rectangular | big | heavyset | masculine |
| stocky | long legs | long | small |
| short legs | lean | long torso | short |
| muscular | short torso | round (apple) | feminine |
| average | pear shaped | built | curvy |
| tall | petite | fit | |

# LANGUAGE SPACE DATA

- Body representations:
  - descriptions made from <u>images</u> of people

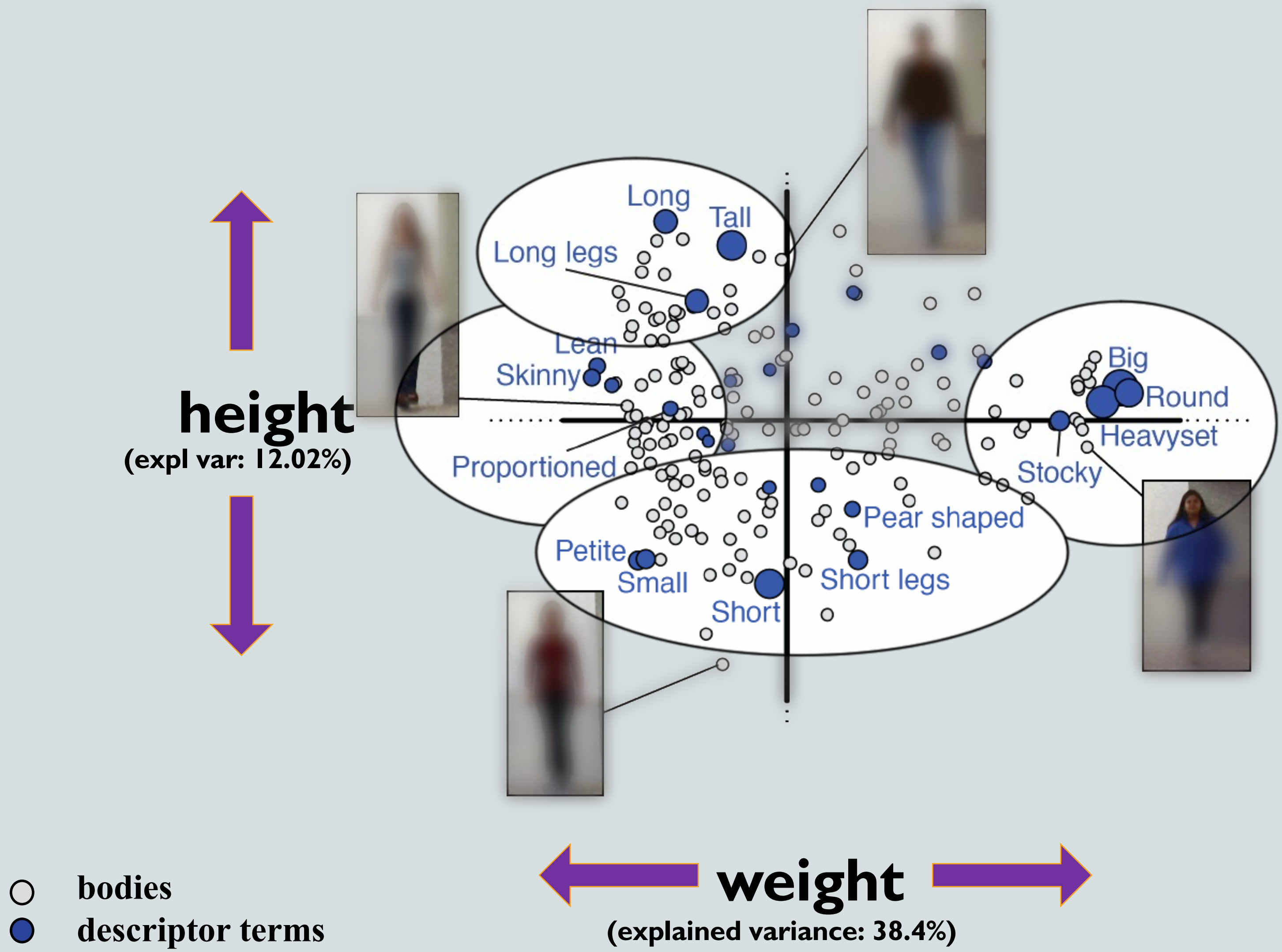
# LANGUAGE SIMILARITY SPACE

- applied **correspondence analysis** to:
  - descriptor vectors for the 164 female bodies
    - 27 elements - terms that "applied perfectly" to the body

- Correspondence Analysis (CA) (Benzicri, 1973)
  - multivariate analysis analogous to PCA, but for categorical data
  - allows observations (bodies) to be plotted in the same space as the variables (descriptor terms)

# LANGUAGE SPACE

## (1ST & 2ND AXES)

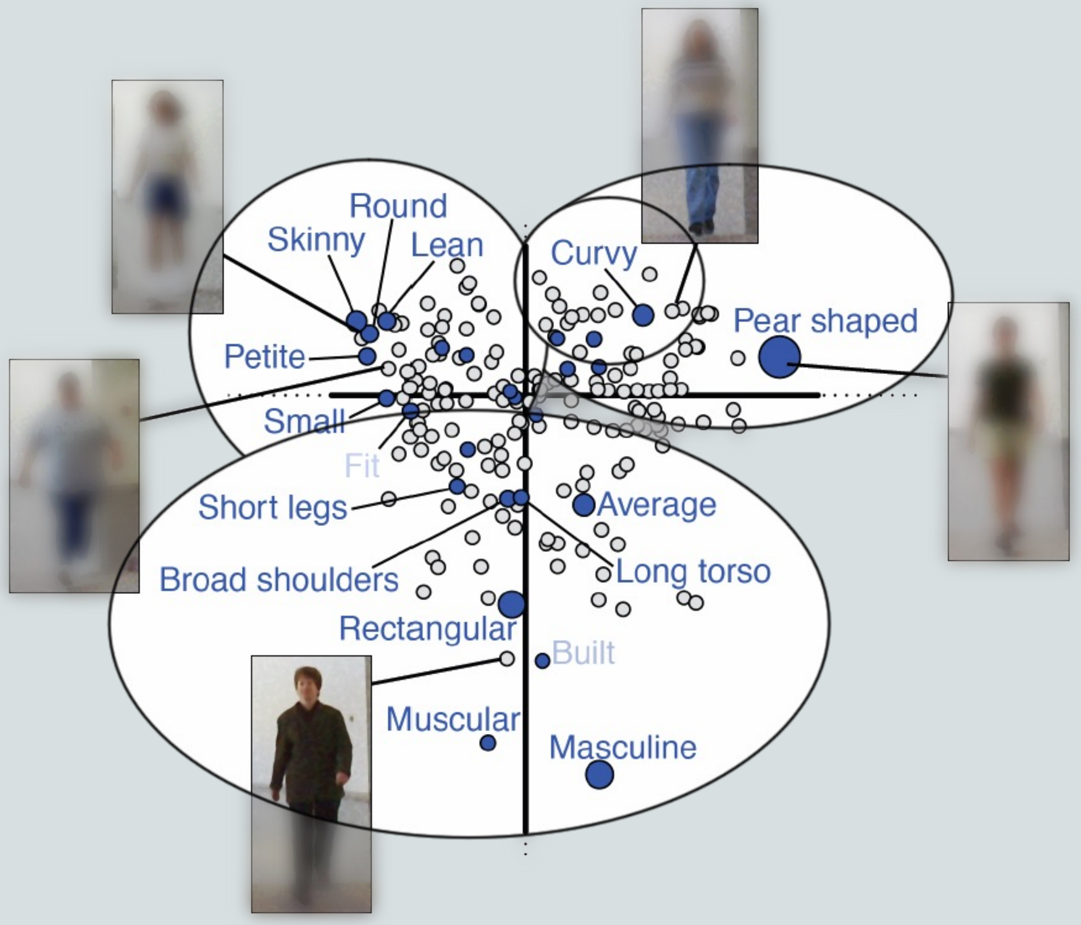


# LANGUAGE SPACE

## (3ST & 4TH AXES)

masculine vs. curvy

**(expl var: 6.73%)**

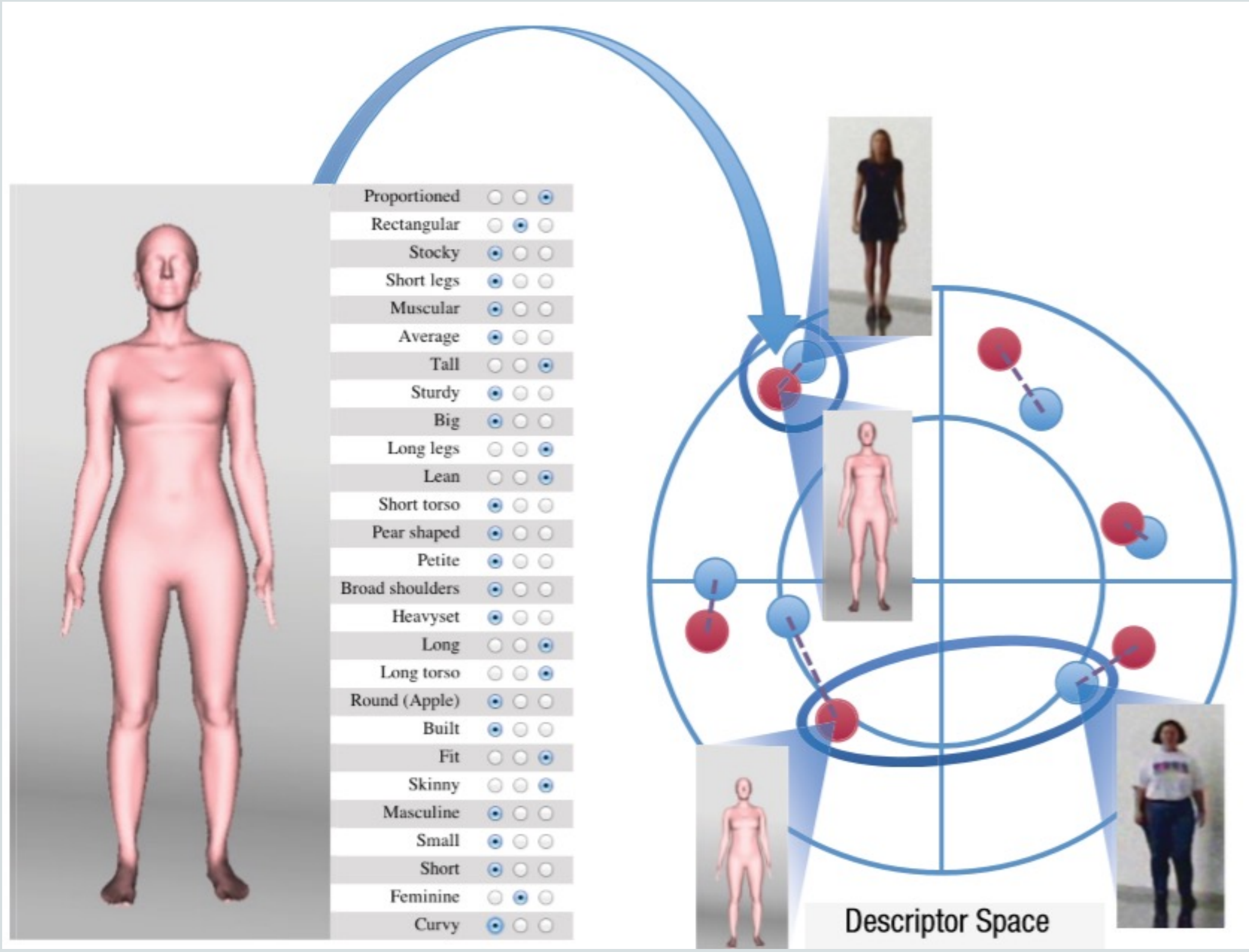



pear shaped vs. other

**(explained variance: 10.78%)**

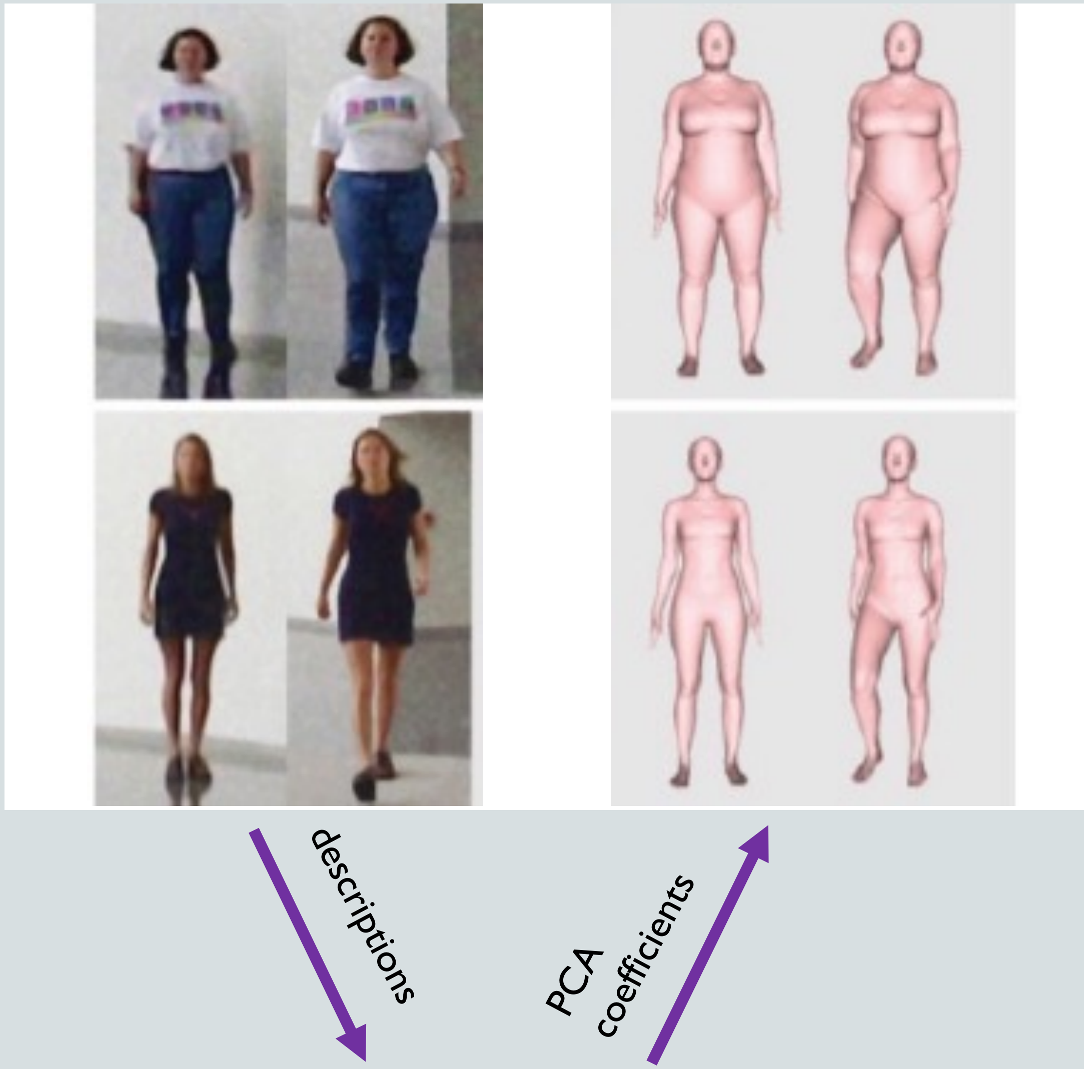
○ bodies
● descriptor terms

# GROUND TRUTH LINGUISTIC DESCRIPTION SPACE WITH 3D BODY MODEL SPACE



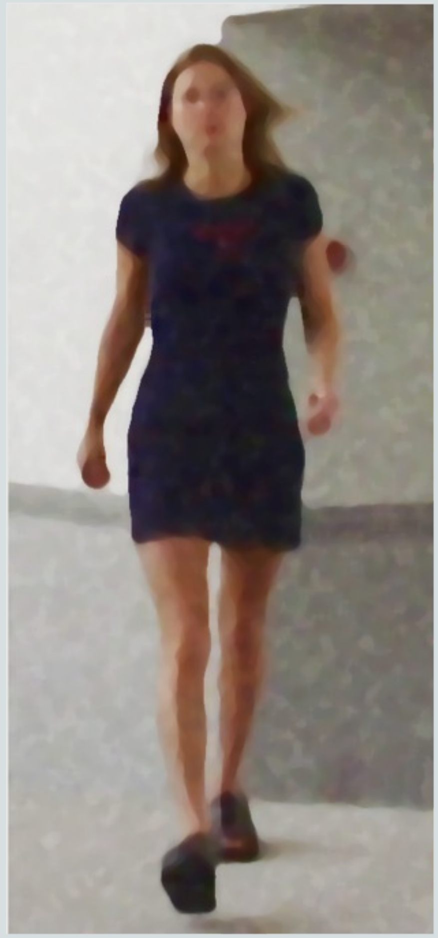

Body model PCA of laser scans of bodies (Loper et al., 2016)

# 3D BODY SYNTHESIS FROM DESCRIPTIONS



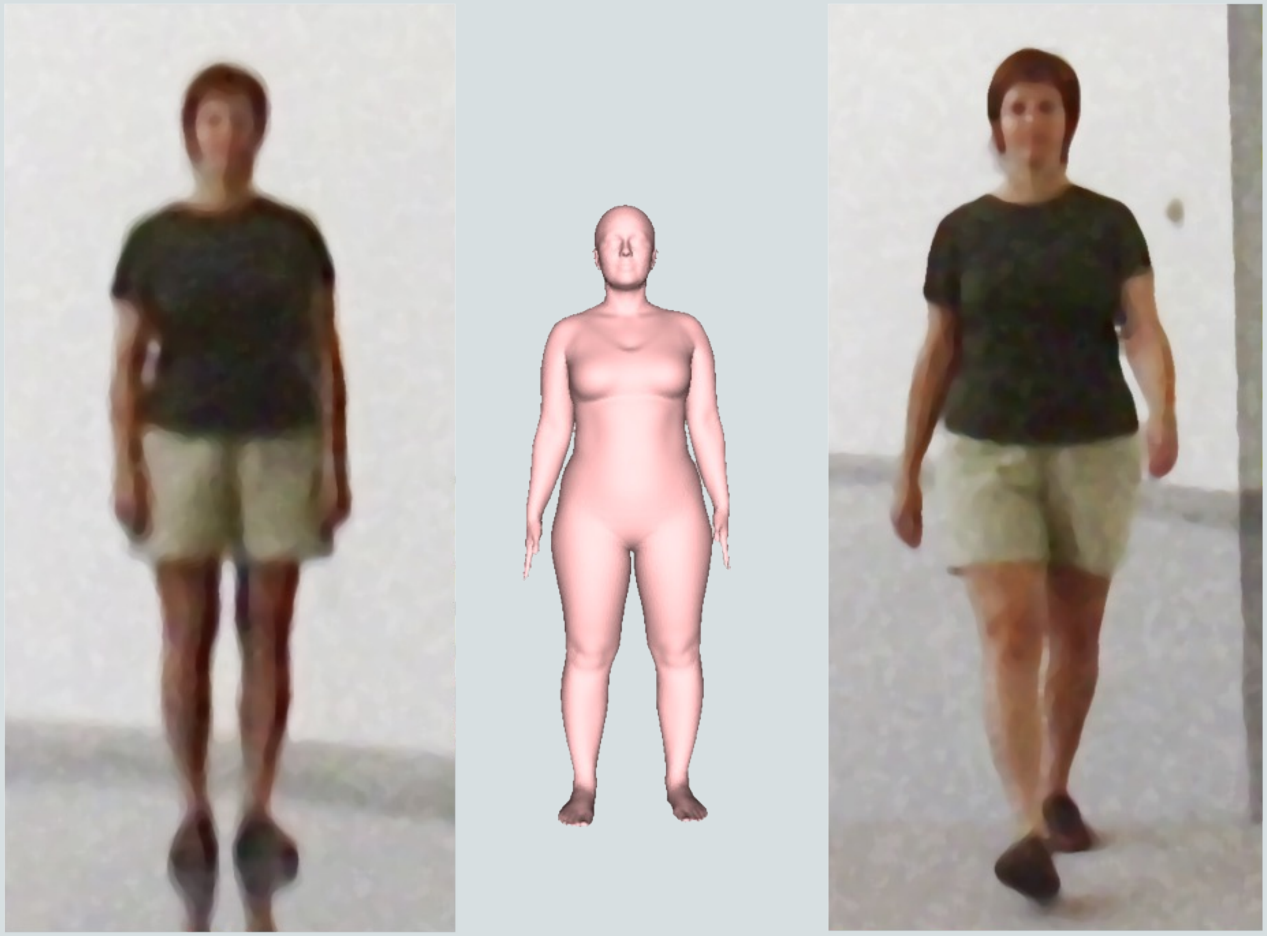

PCA of 3000+ laser scans
SMPL model (Loper et al., 2016)

subjects consented to publication



3/13/24

subject consented to publication

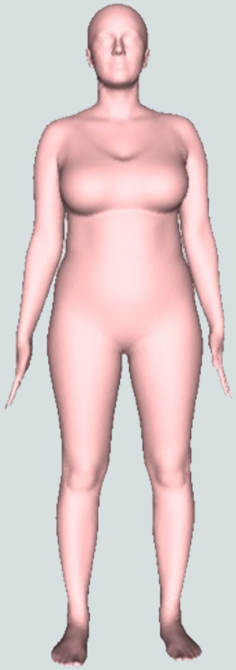




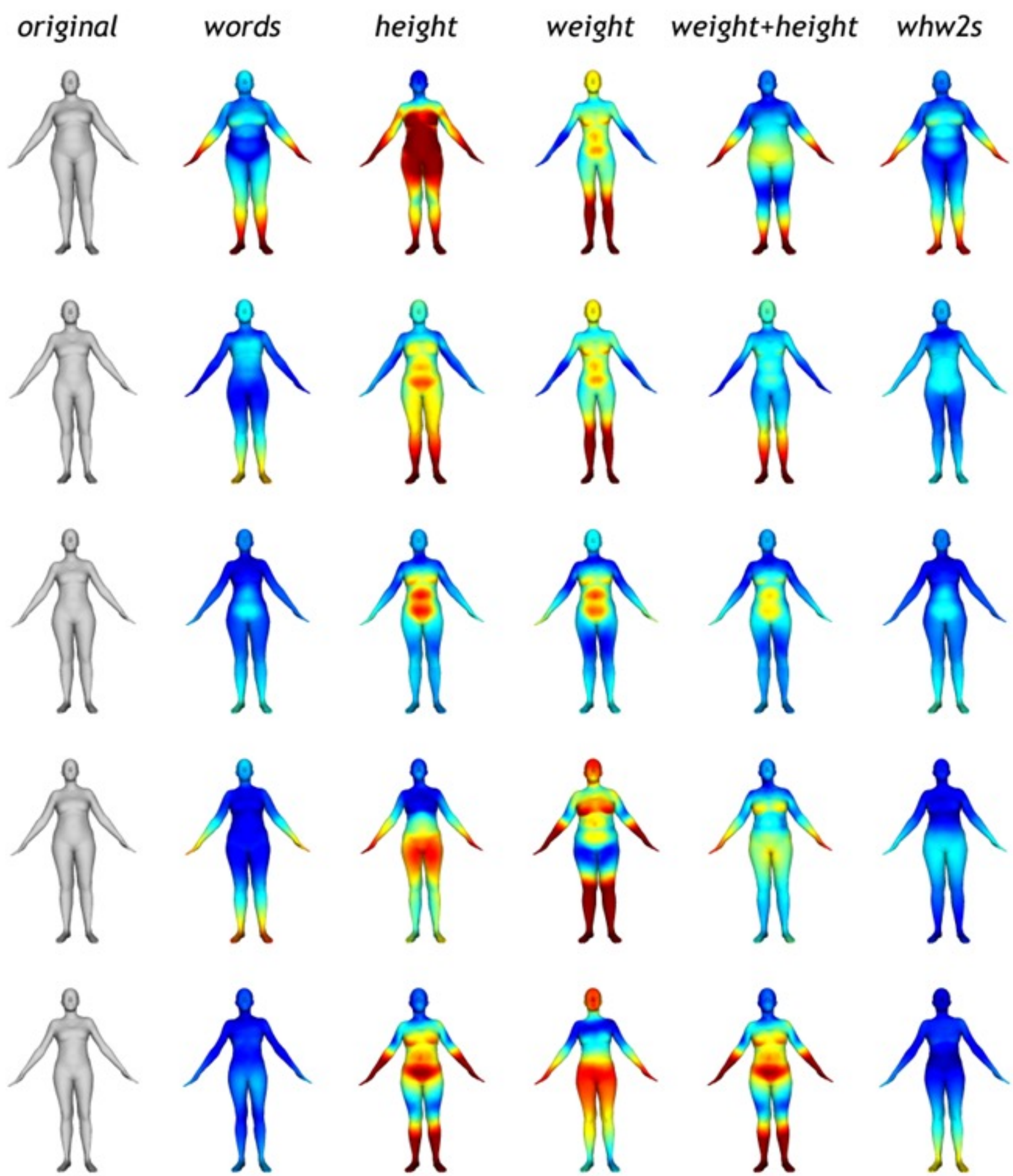
subject consented to publication

3/13/24

subject consented to publication

3/13/24




Streuber et al. (2016)

# CONCLUSIONS

- linguistic descriptions
  - can be used to synthesize 3D bodies

- efficient way to perform a laser scan without a laser scanner ☺

# IDENTIFICATION FROM BODY SHAPE

Myers BA, Jaggernauth L, Metz TM, Hill MQ, Gandi VN, Castillo CD, O'Toole AJ. Recognizing People by Body Shape Using Deep Networks of Images and Words. arXiv:2305.19160. 2023 May 30. *Proc. IEEE International Joint Conference on Biometric, Sept. 2023*

# WORDS FOR BODY IDENTIFICATION

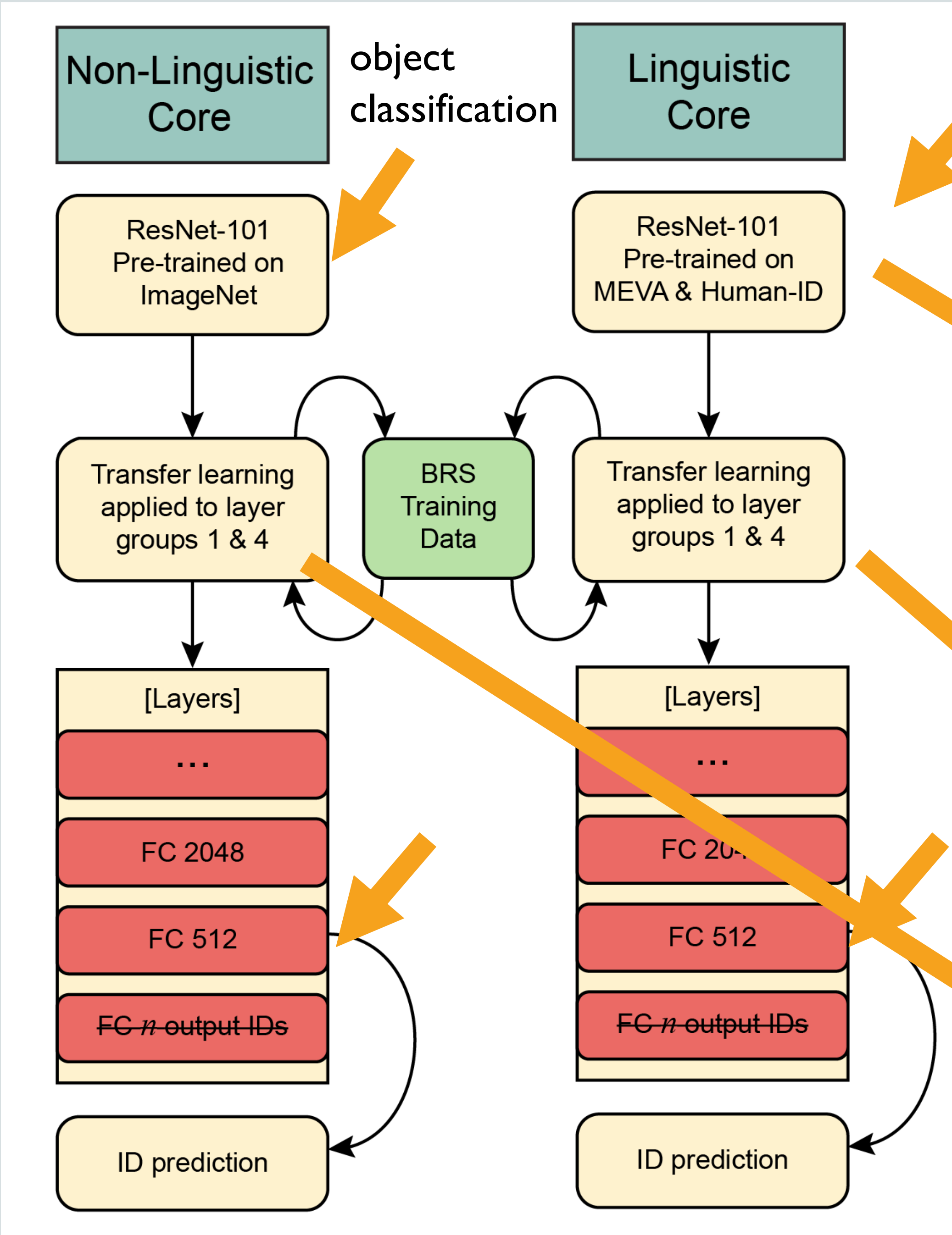
- Rationale
  - **descriptors sufficient to synthesize 3D body**
  - descriptor-based representation for identification?

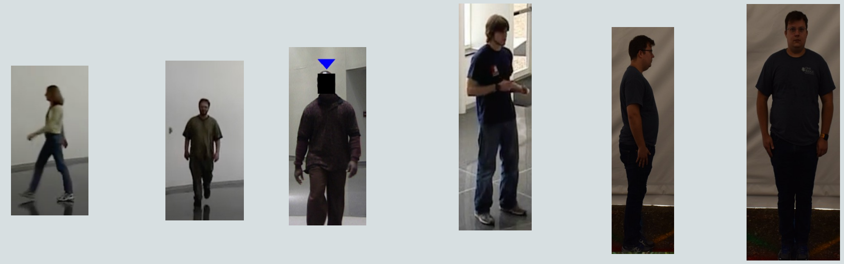
- Advantages
  - robust across large distances
  - generalize across yaw and pitch (curvy, tall, stout, long legs,)
    - accessible across a range of view
  - (relatively) clothing independent

  - Explainable AI??

# CURRENT PROBLEM

- learn mapping from images to descriptors
  - pretraining – to categorize body shape
- image to identity
  - transfer learning – image to identity
  - fine tuning within a category



curvy,
tall,
stocky,
short legs
muscular,…..etc.

**identity trained**
close range,
UAV
100m, 200m, etc….1000m

**identity trained**
close range,
UAV
100m, 200m, etc….1000m

# MODELS

- linguistic body model (LCRIM)
  - linguistic core model
    - body image to linguistic description
  - identity-tuning
    - body image to identity

- non-linguistic body model (NLCRIM)
  - pre-trained object classification core model
    - ImageNet trained
  - identity-tuning
    - body image to identity

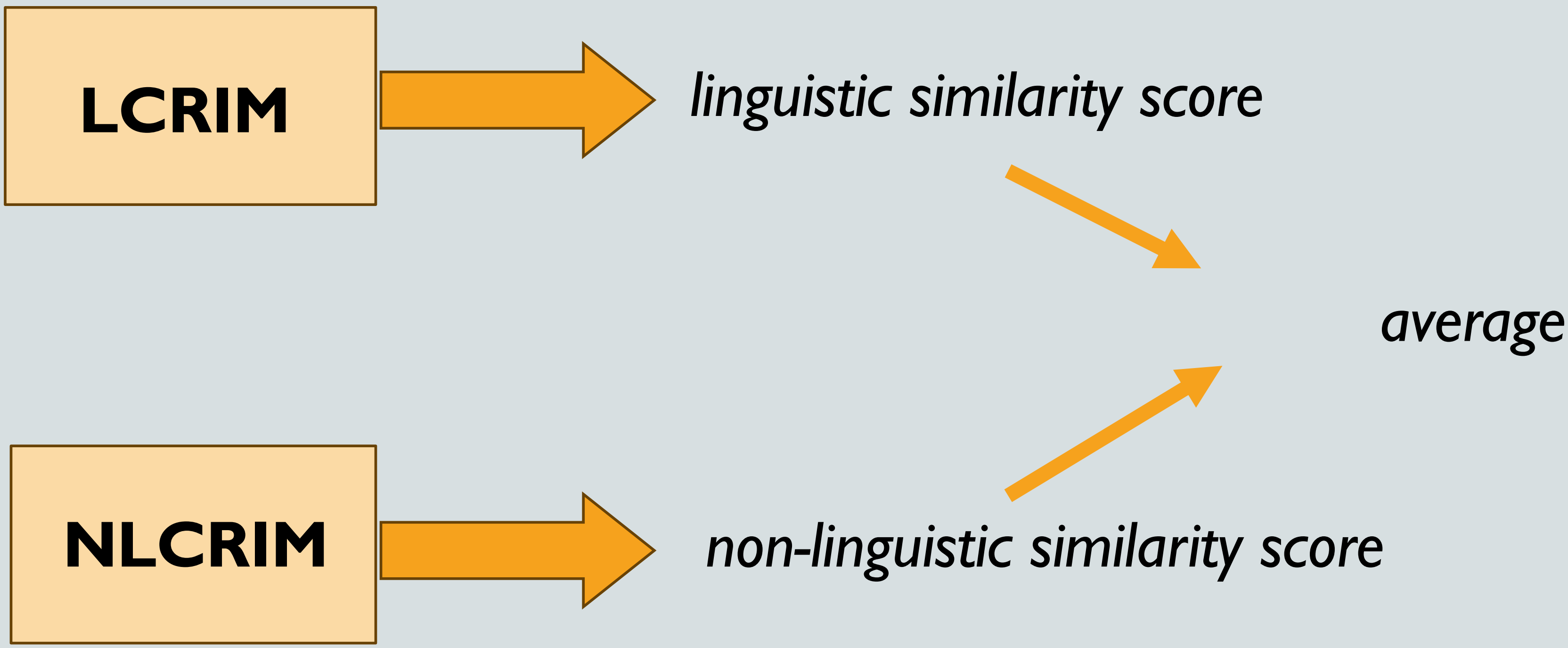
- Fusion = LCRIM + NLCRIM


- training
  - 577 IDs
    - 242,386 images
- test
  - 485 gallery IDs
    - 43,722 images
  - 260 probe IDs
    - 2,192,305 image frames from 9,795 videos

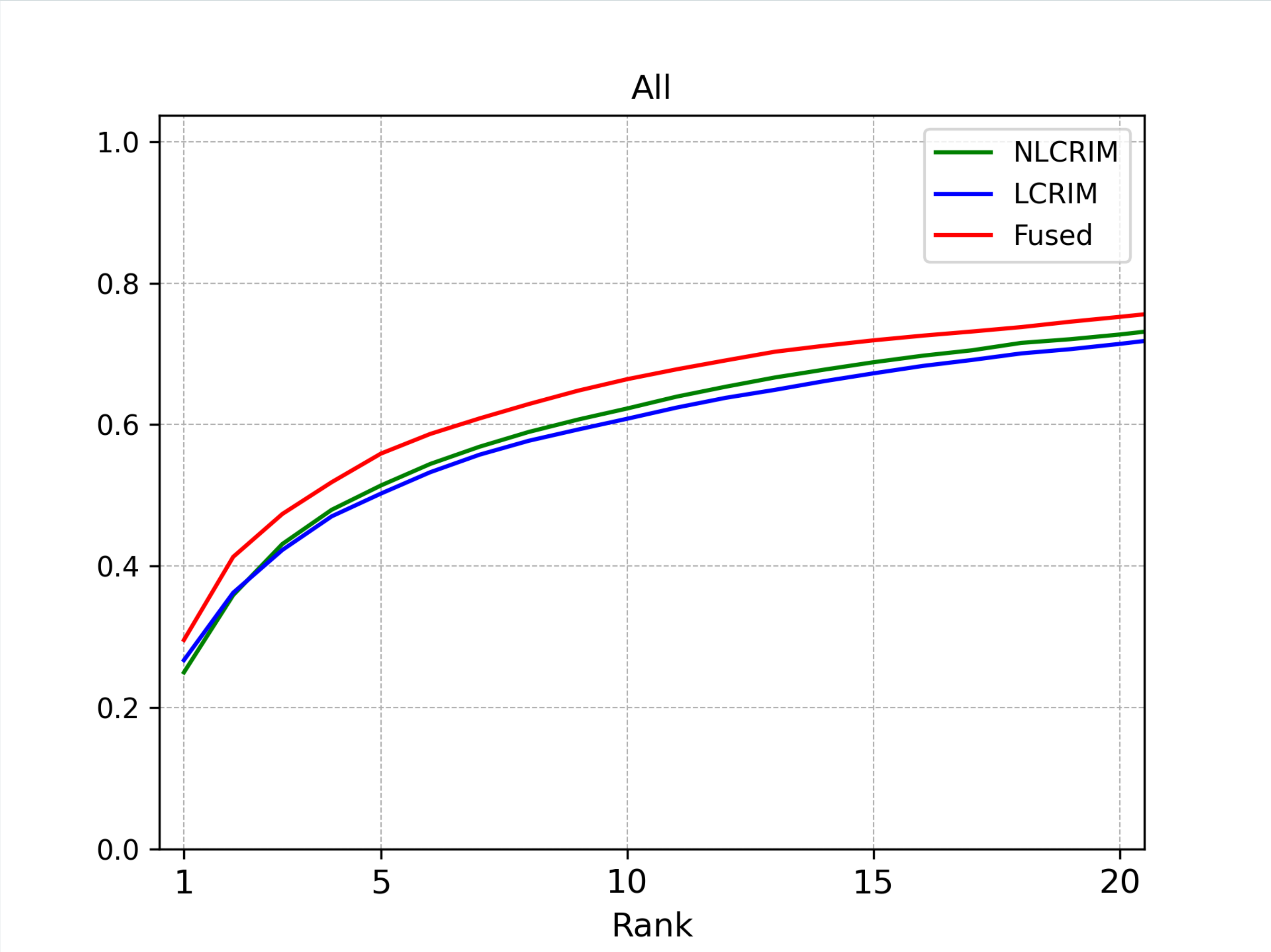


(BRS-BTS dataset, Cornett, et. al., 2022)

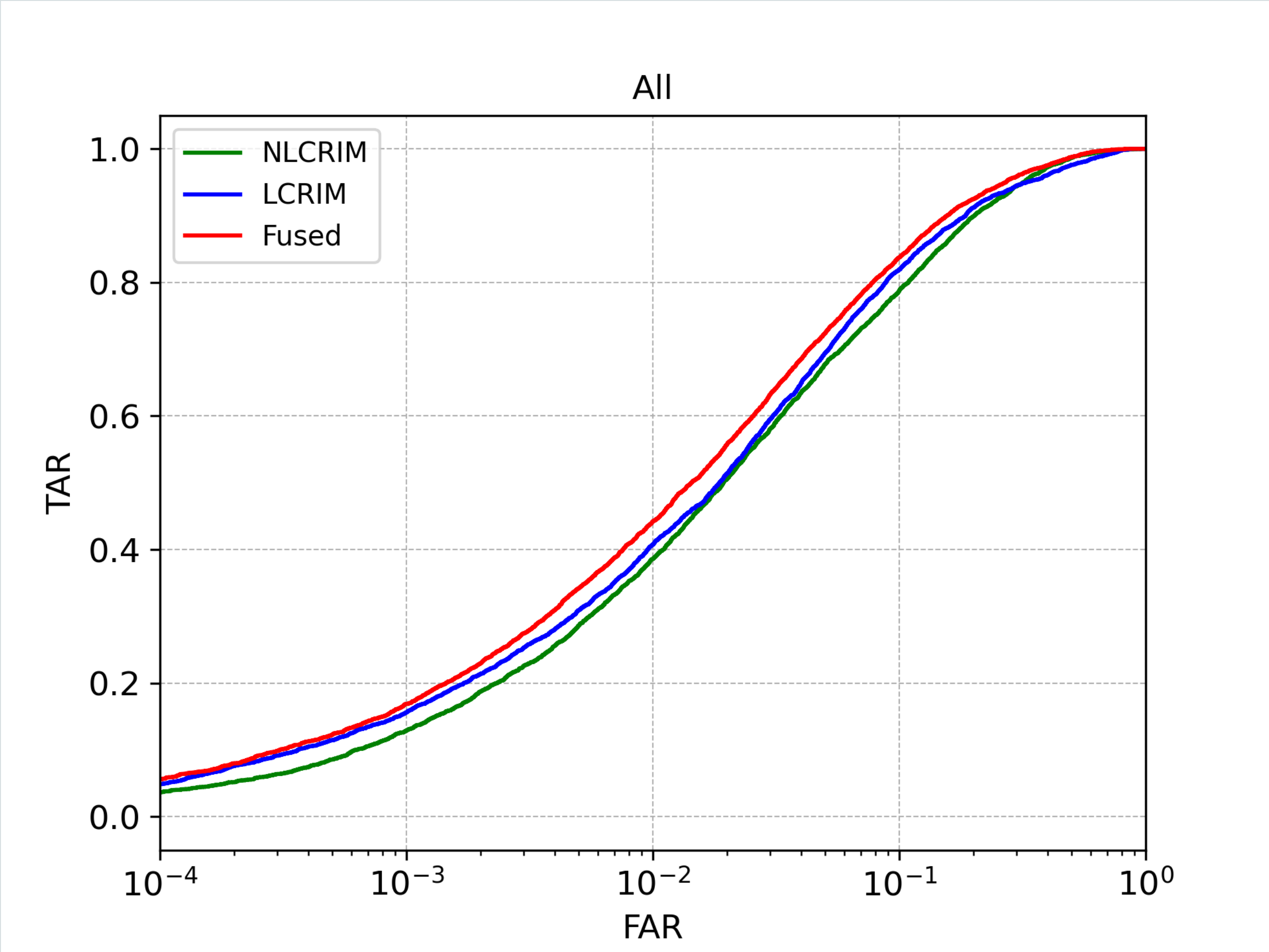
# FUSION


LCRIM
linguistic similarity score
average
NLCRIM
non-linguistic similarity score

# Cumulative Match Characteristic

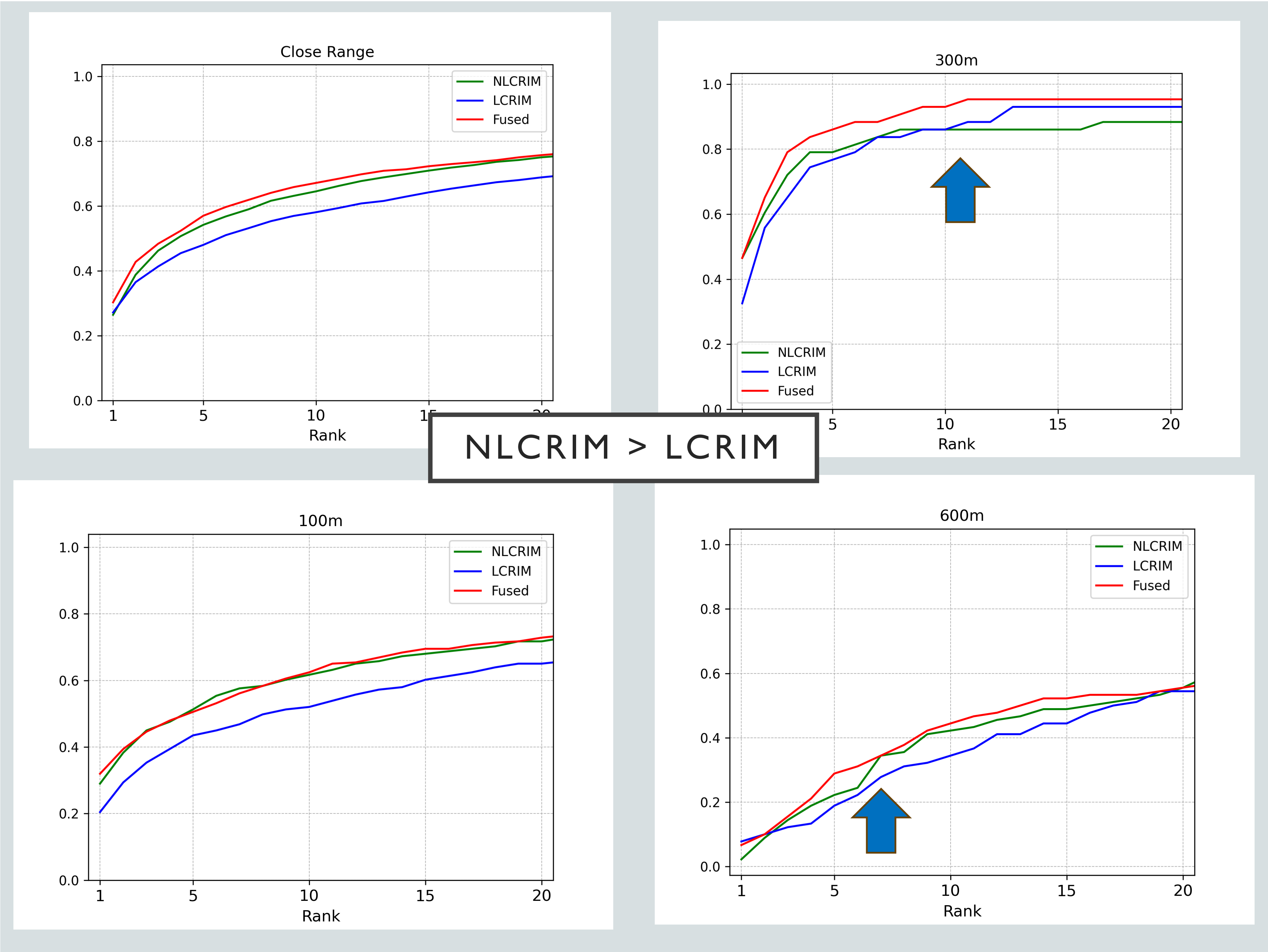


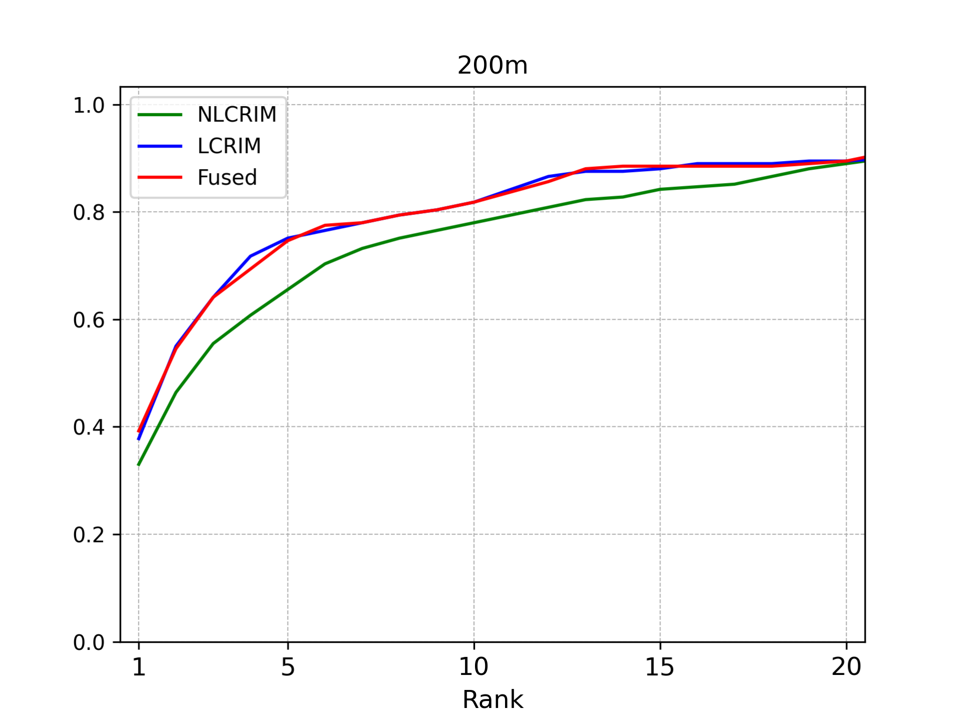
# Receiver Operating Characteristic Curve

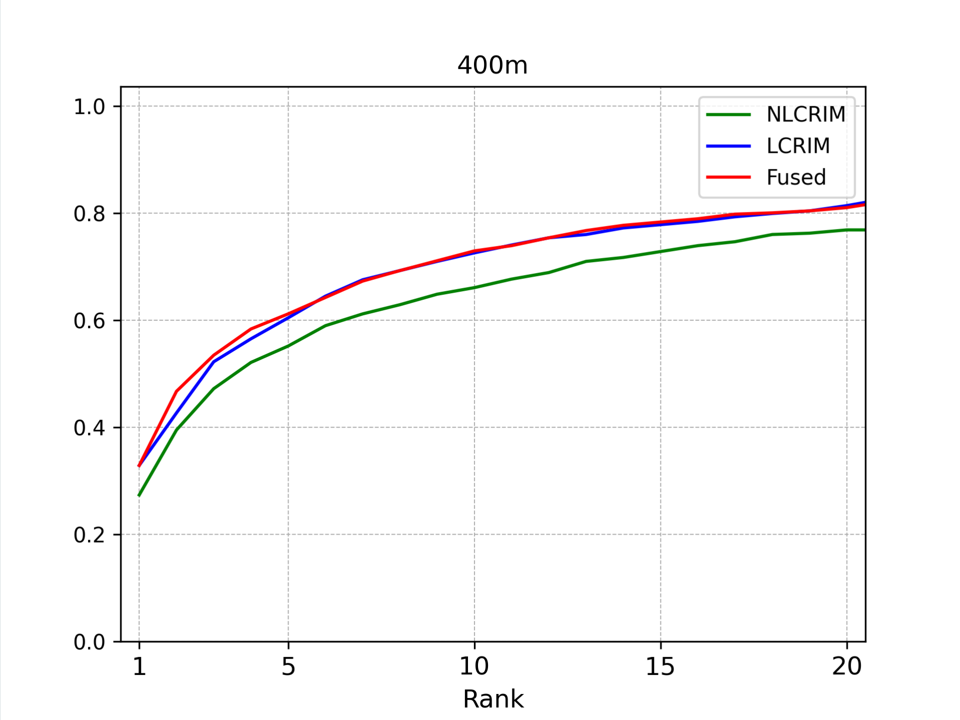


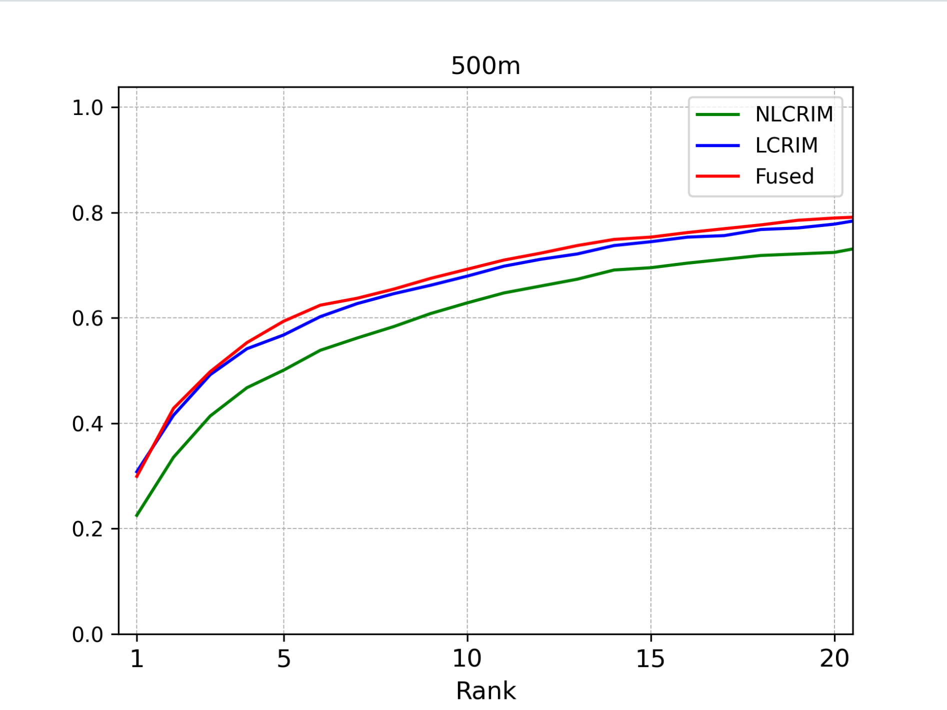
# DISTANCE CONDITIONS

- Linguistic > as views and pitch get more extreme?


Close Range
300m
100m
600m
NLCRIM
LCRIM
Fused
Rank
1.0
0.8
0.6
0.4
0.2
0.0
1
5
10
15
20
NLCRIM > LCRIM

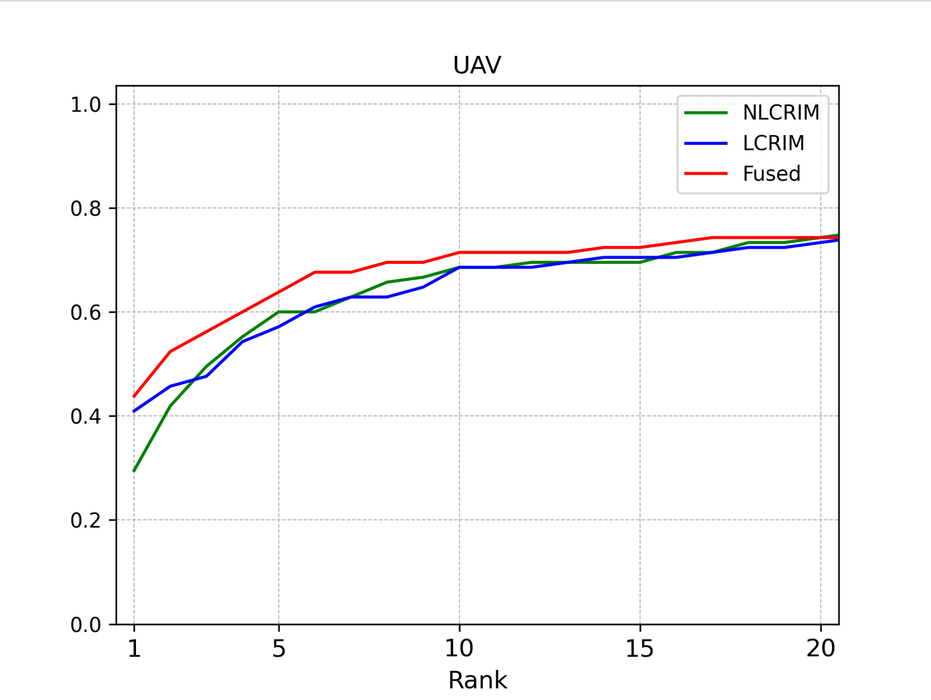


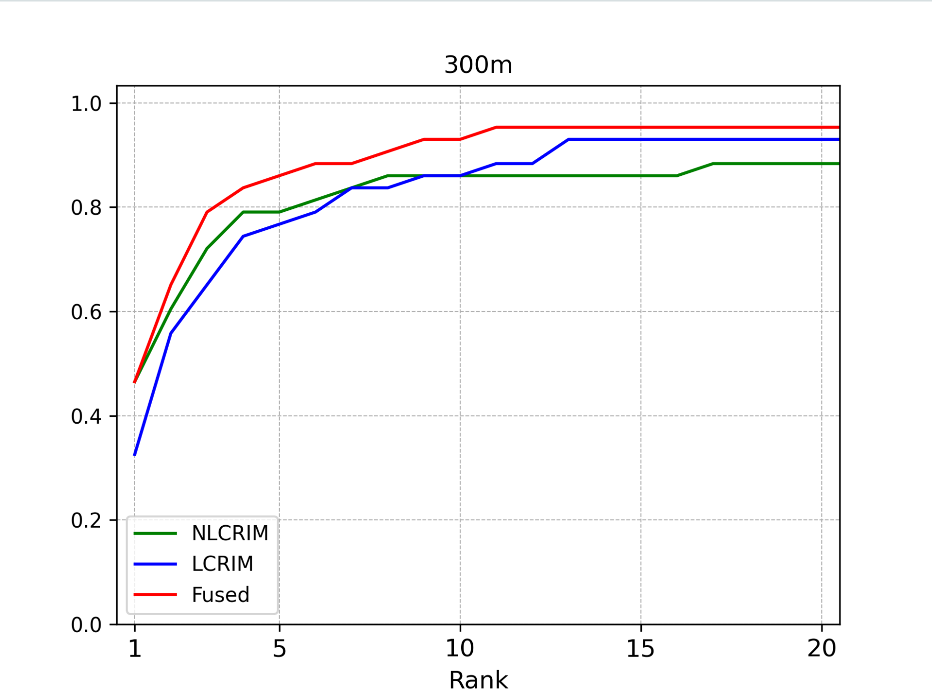




LCRIM > NLCRIM

# FUSION >> (NLCRIM OR LCRIM)



# LINGUISTIC? NON-LINGUISTIC? FUSED?

- condition-dependent
  - fusion *almost* always best

  - linguistic/non-linguistic
    - less predictable

# CONCLUSIONS

- Linguistic descriptors
  - complement body shape representations
  - better at further distances (tentatively)
  - tap similar types of information

# PERSON = FACE + BODY + GAIT

FUSION, VARIANCE, QUALITY

# LIMITS OF THE BODY


*same person or different people?*

*same person or different people?*

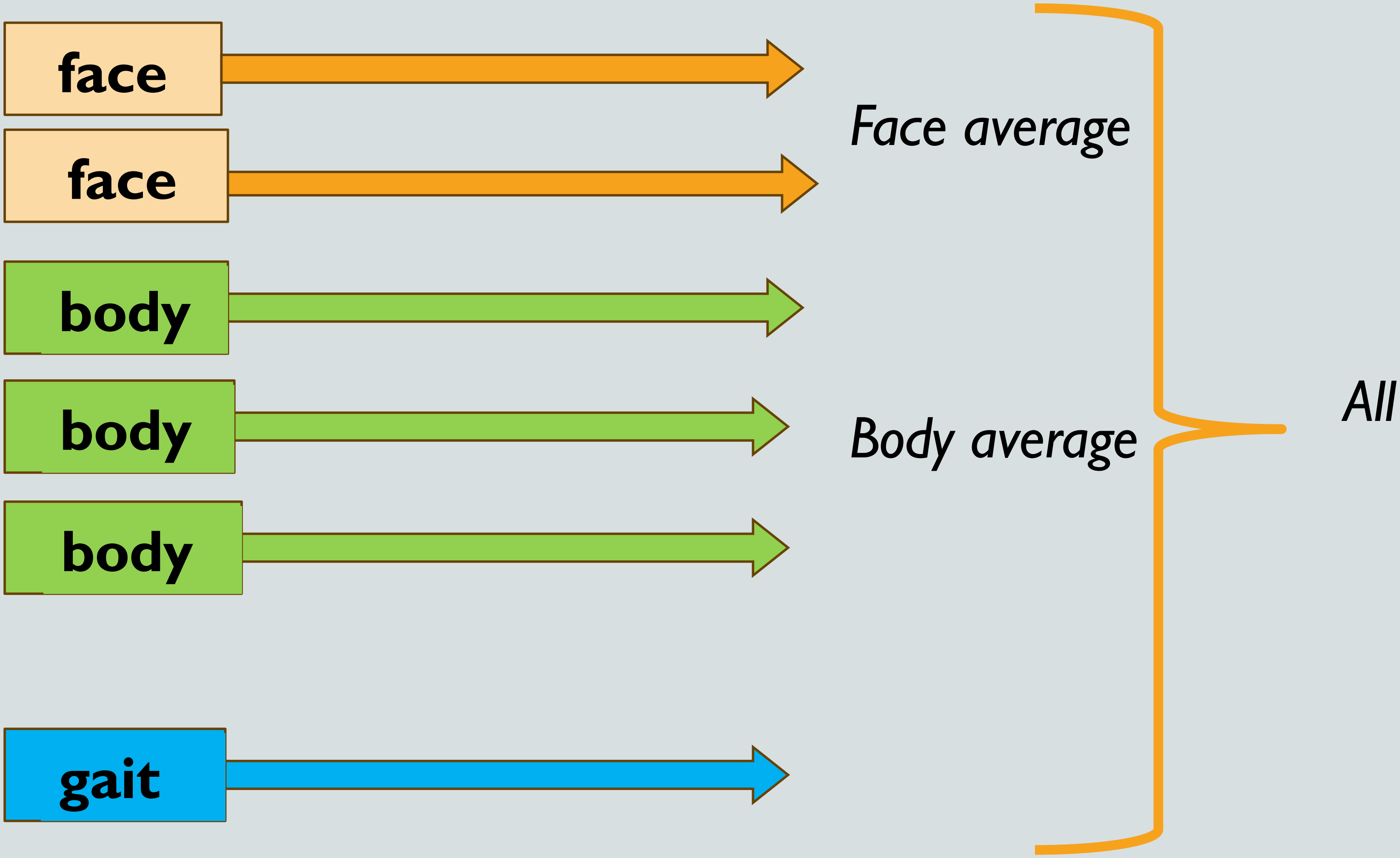
*same person or different people?*

subject consented to publication

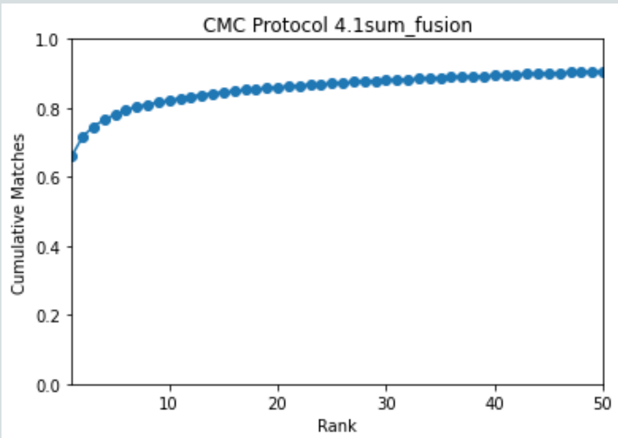
# FACE, BODY, & GAIT: MODEL (DIS)AGREEMENT

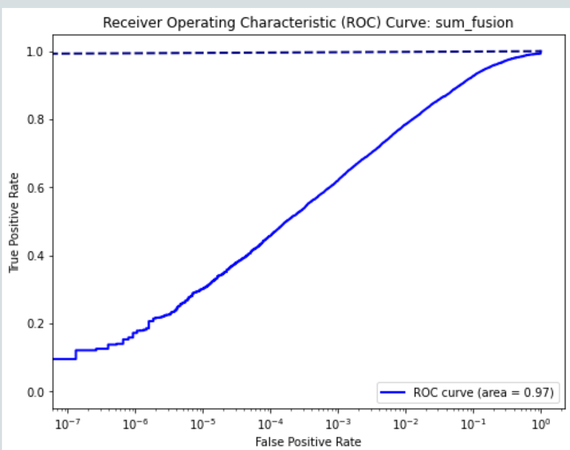
| | | Body 1 | Body 2 | Body 3 | Face 1 | Face 2 | Gait |
|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 |
| Body 1 | 0 | 1 | 0.44643 | 0.446626 | 0.186319 | 0.185375 | 0.15463 |
| Body 2 | 1 | 0.44643 | 1 | 0.785518 | 0.25484 | 0.255659 | 0.346526 |
| Body 3 | 2 | 0.446626 | 0.785518 | 1 | 0.250673 | 0.251236 | 0.352391 |
| Face 1 | 3 | 0.186319 | 0.25484 | 0.250673 | 1 | 0.660499 | 0.127177 |
| Face 2 | 4 | 0.185375 | 0.255659 | 0.251236 | 0.660499 | 1 | 0.132766 |
| Gait | 5 | 0.15463 | 0.346526 | 0.352391 | 0.127177 | 0.132766 | 1 |

# FUSION

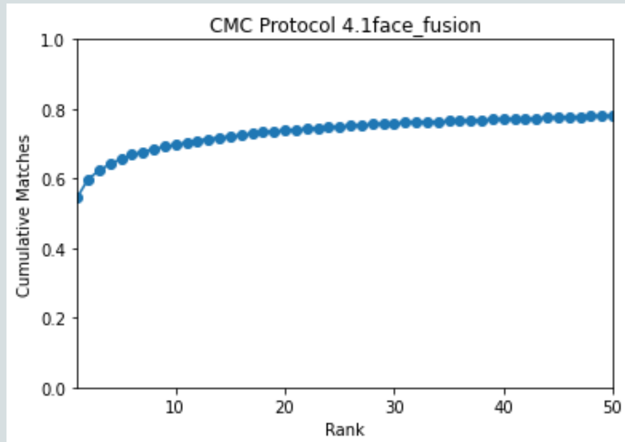


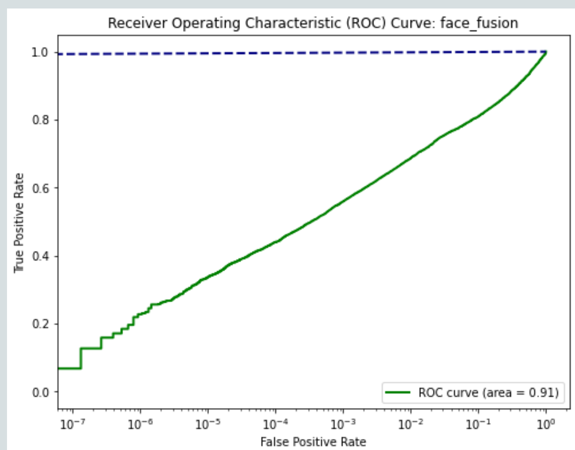
# Fusion



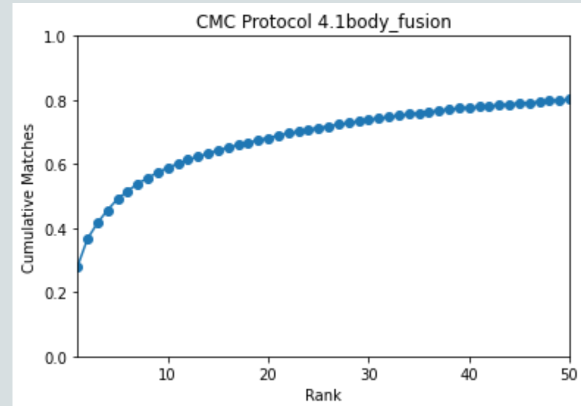



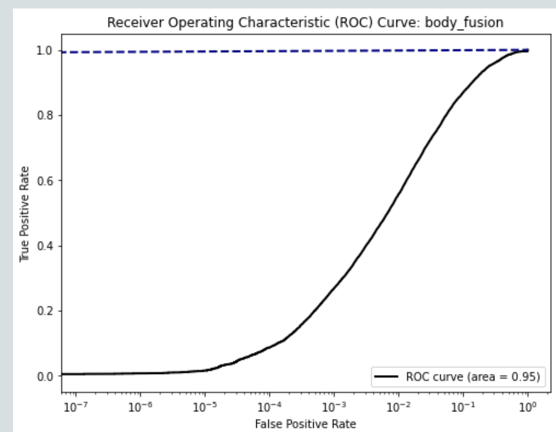

# Face Fusion

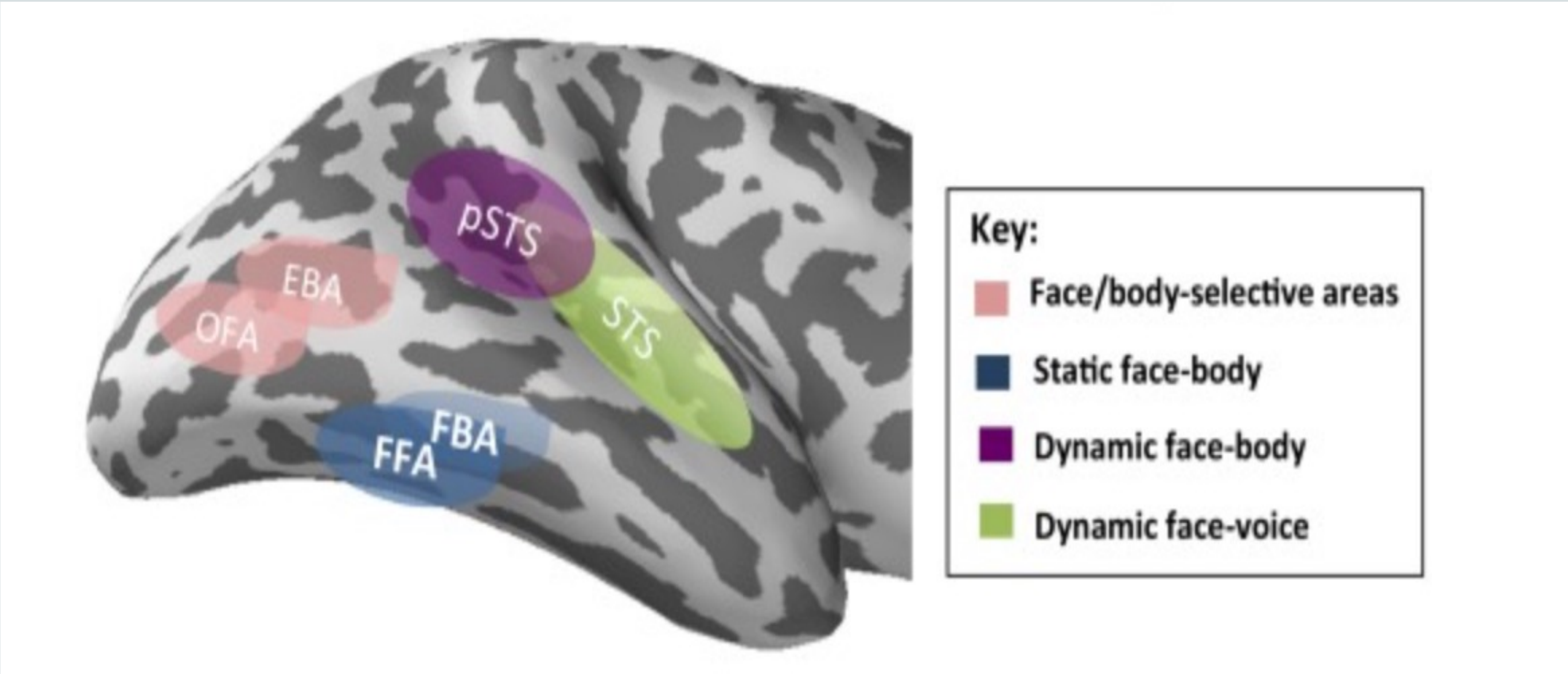



# Body Fusion






Yovel & O'Toole (2016)

# APPROACH

- fusion on a case-by-case basis
  - requires **quality** of face vs. body vs. gait with limited meta-data
- What happens when they do not agree?
  - face with body?
  - face with face? body with body?
  - gait with face or body?
- *Can disagreement be informative of quality???*

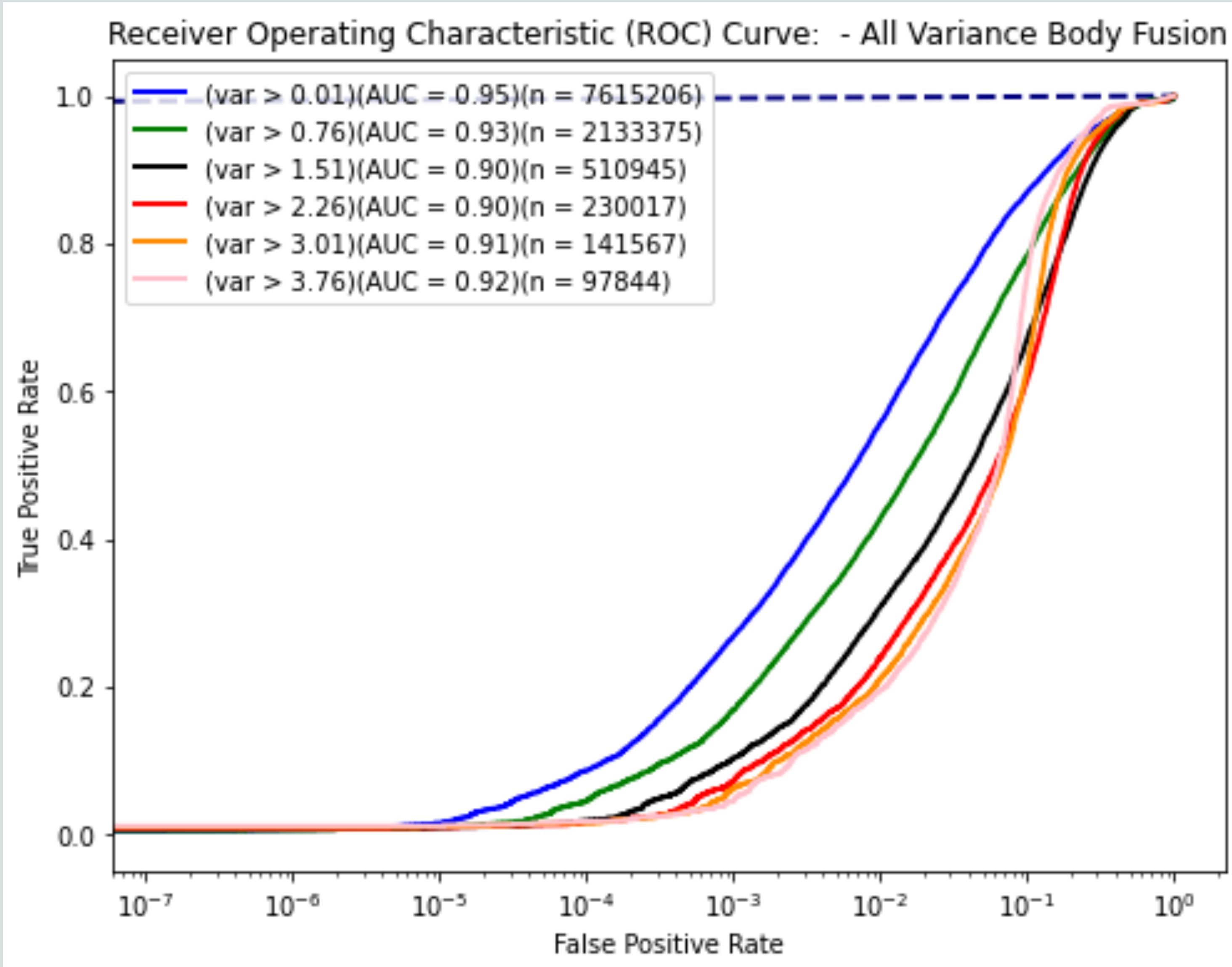
# VARIANCE OF ESTIMATES

- Proposal
  - Can variance of model estimates guide fusion?
- Predict
  - high variance indicates "low quality" and low accuracy
- Prerequisite (sanity test)
  - *Does variance of model estimates relate to accuracy?*

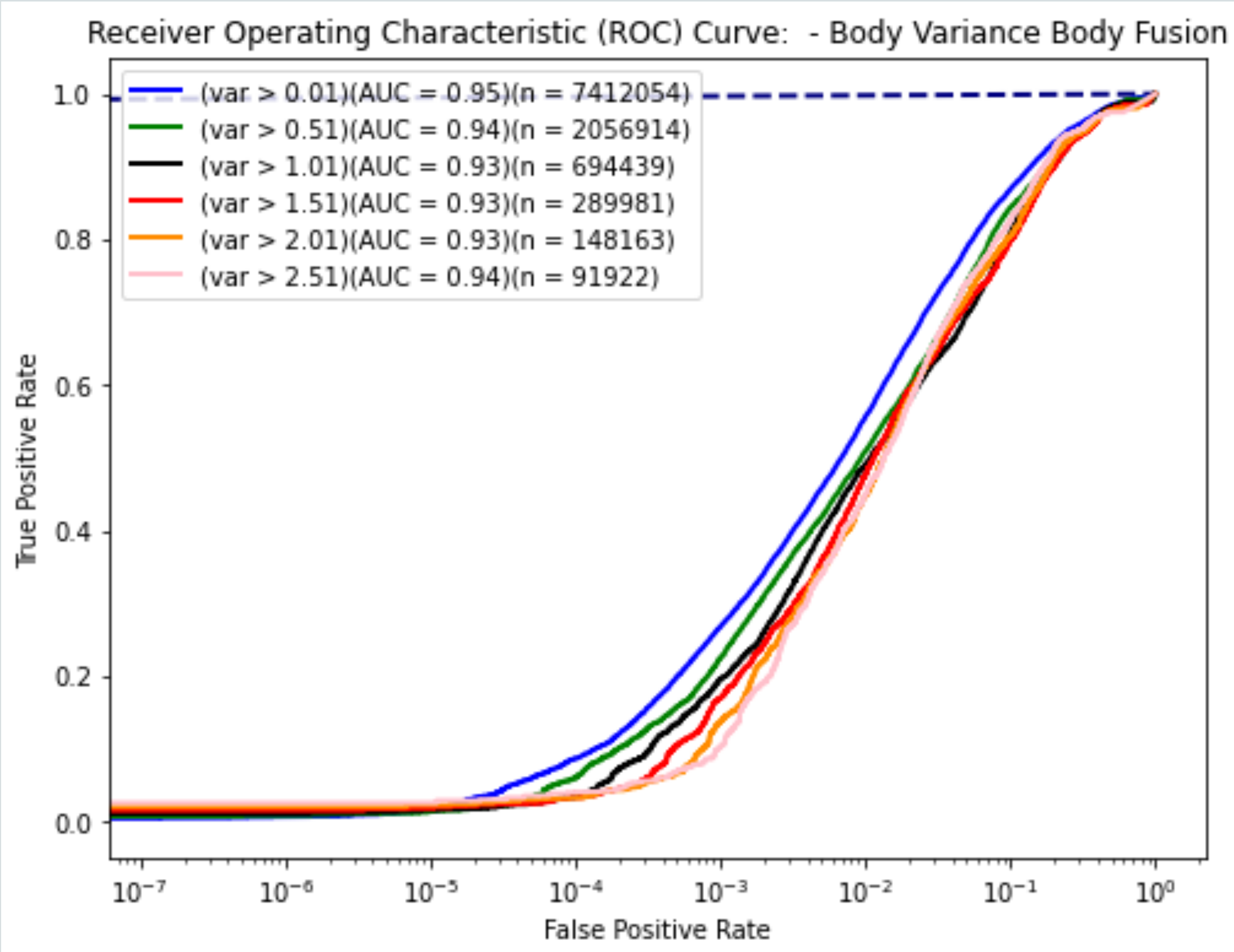
# DOES MODEL VARIANCE PREDICT ACCURACY?

- Variance *on each item*
  - all-model variance
  - face-model variance
  - body-model variance

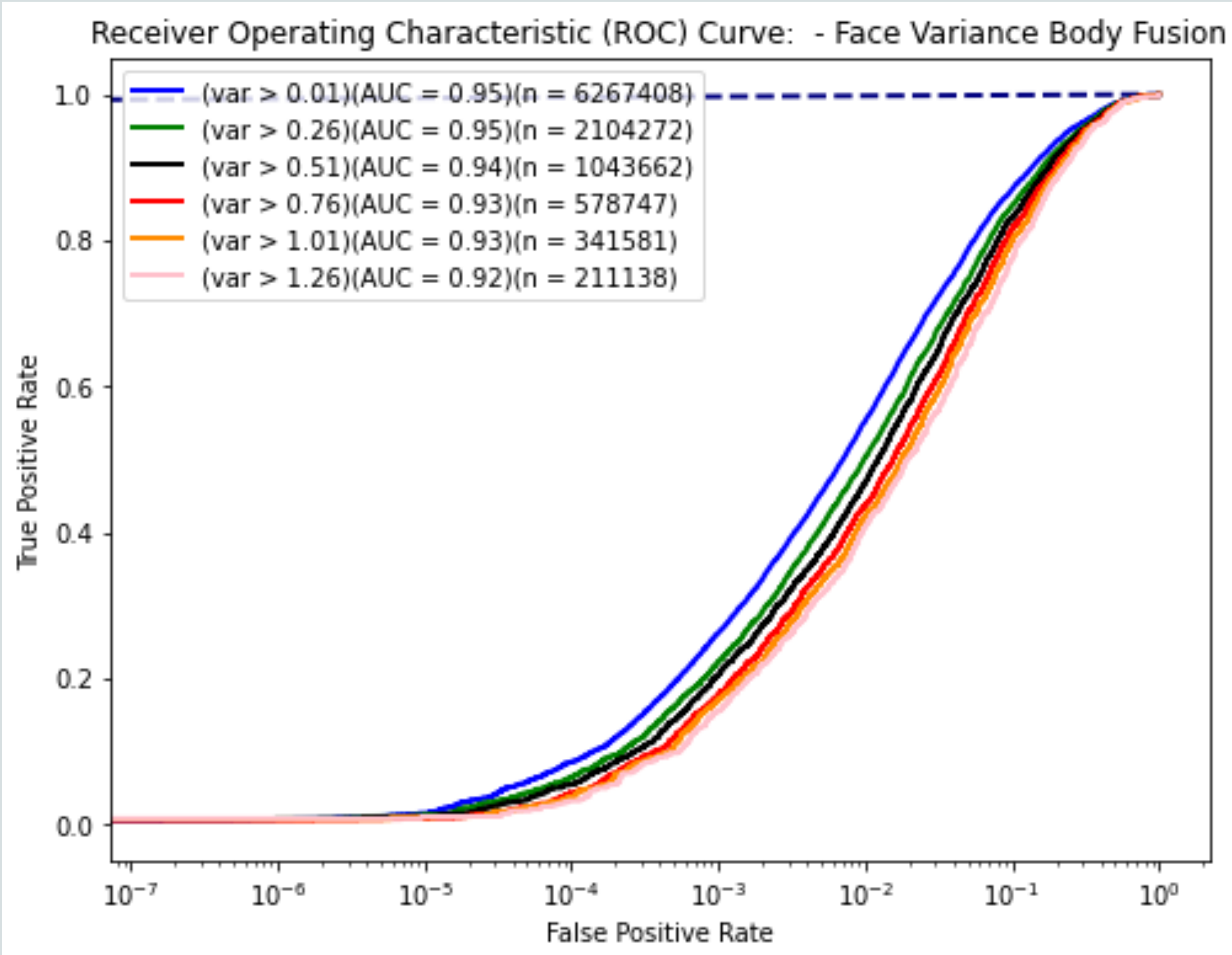
- Performance:
  - face fusion similarity scores (2 face algorithms)
  - body fusion similarity scores (3 body algorithms)



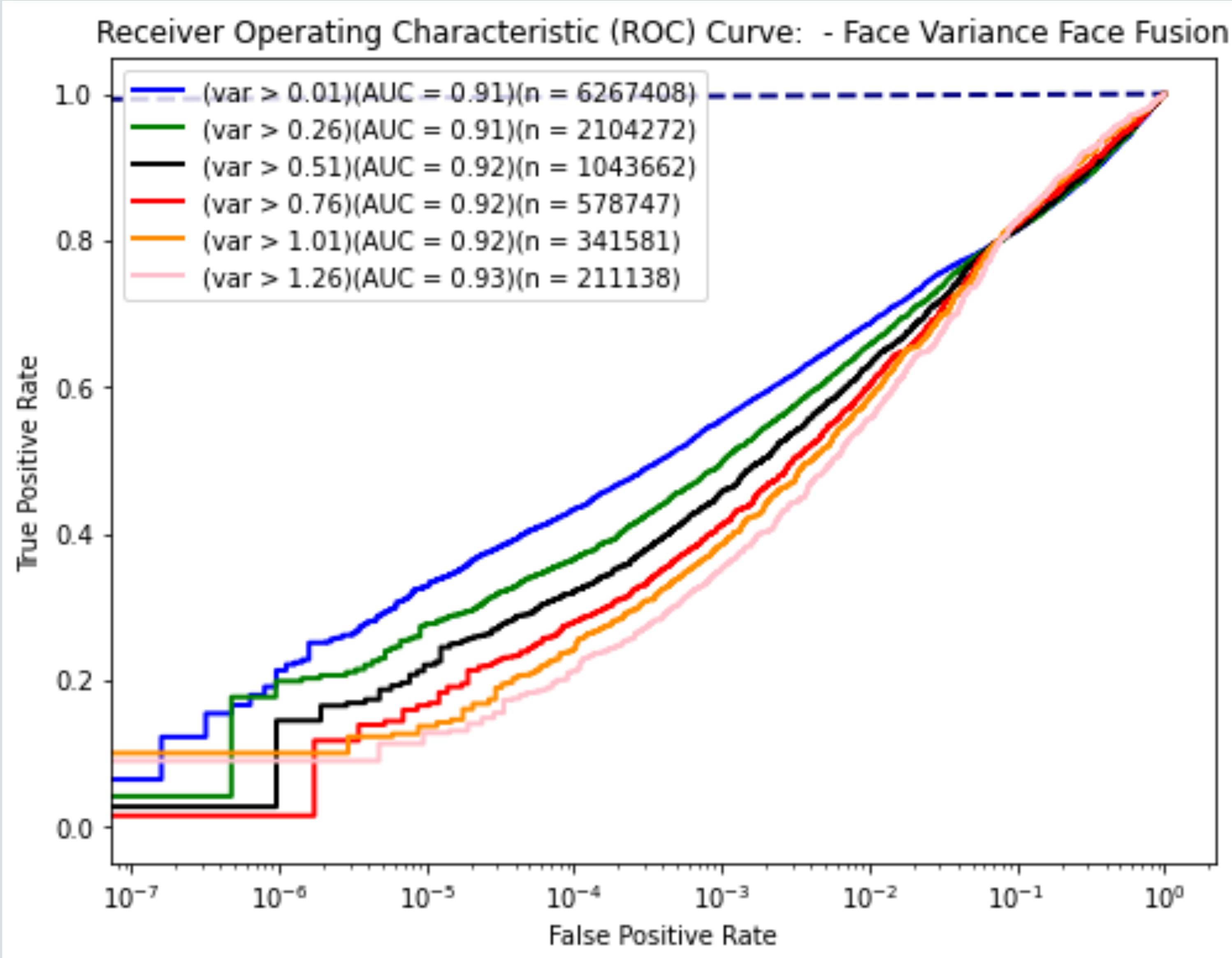

Low variability model scores - better performance with body information




Low variability body models better performance with body

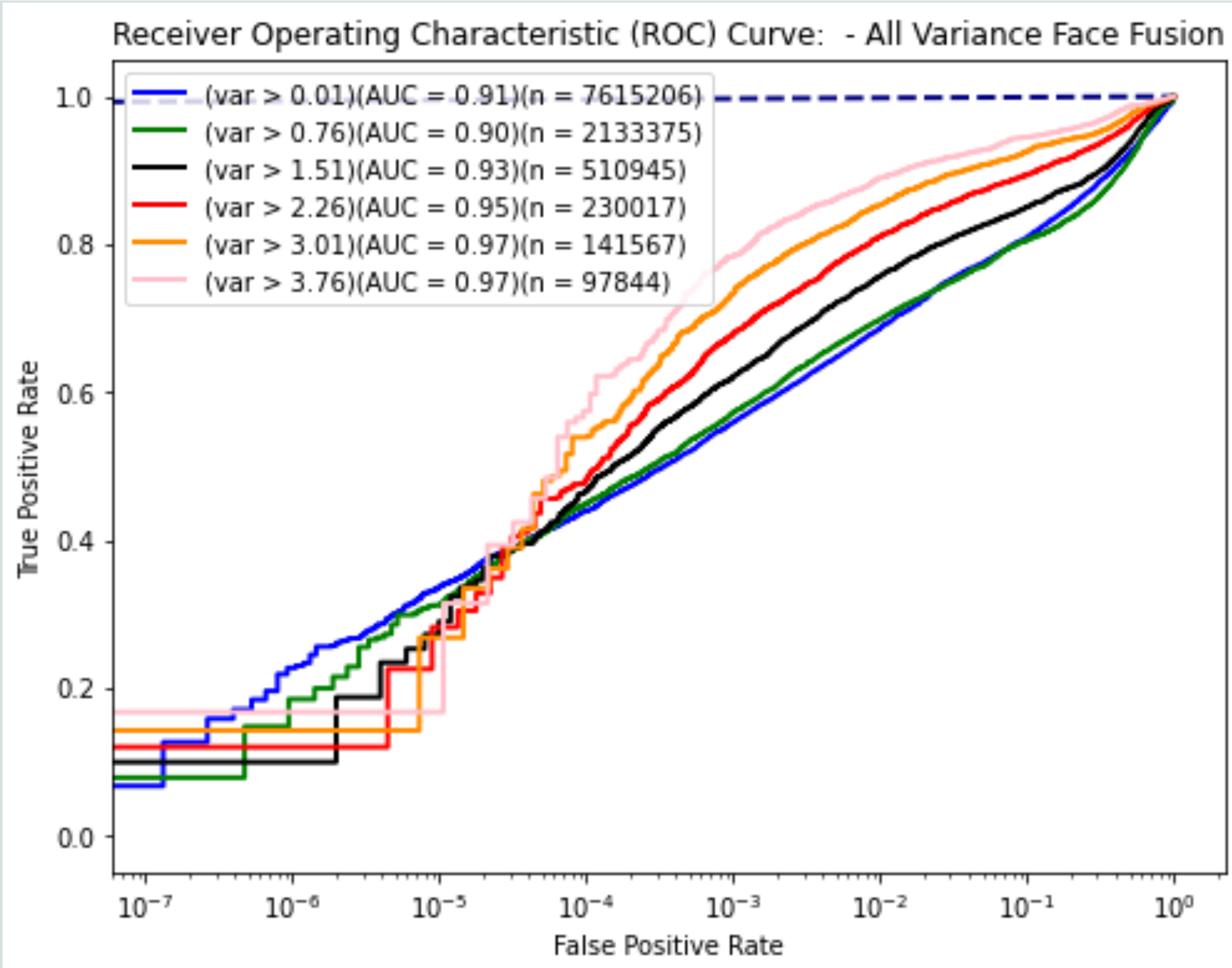



Low variability face models better performance with body



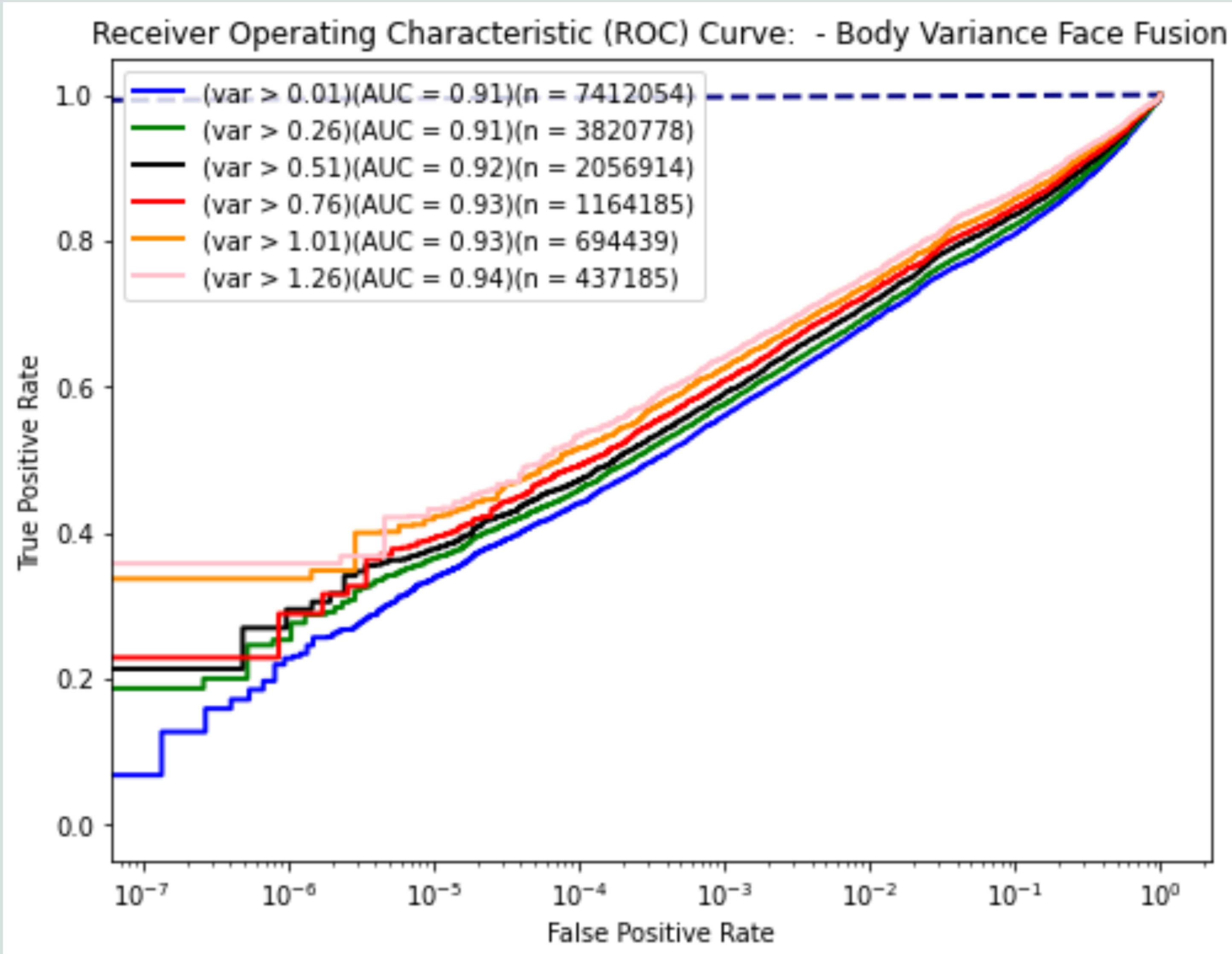

Low variability face models better performance with the face

# 2 INVERSIONS




High variability model scores ***better*** performance with the face

(except at very low FP)




High variability body estimates better performance with face!

# TAKE HOME MESSAGE

- Biometrics has ignored the body on the (correct) premise that it is not "unique"
  - not unique $\neq$ not helpful
- Linguistic descriptions of bodies
  - graphics, shape classification, identification
- Body algorithms boost identification over
  - face
  - gait
- Quality estimates from model discord within/across modalities

# ACKNOWLEDGEMENTS

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via [2022-21102100005]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S.

Government. The US. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

# THANK YOU!