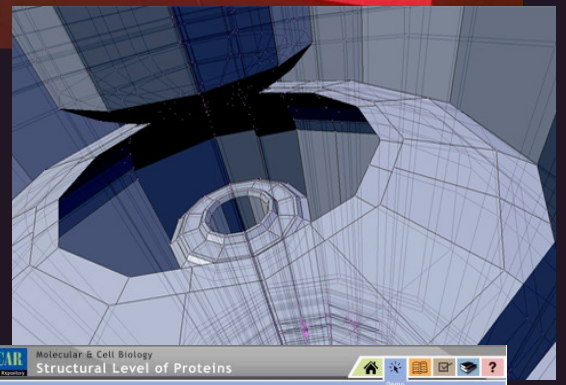
Socially Intelligent Digital Humans

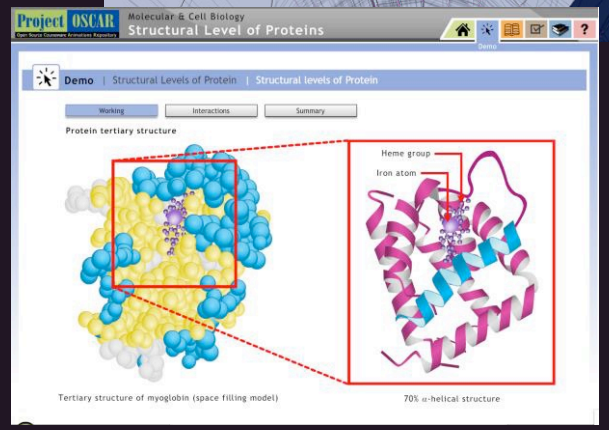
Evaluating and Generating Multiparty Behavior

Chirag Raman
Asst. Professor, Tapri Lab

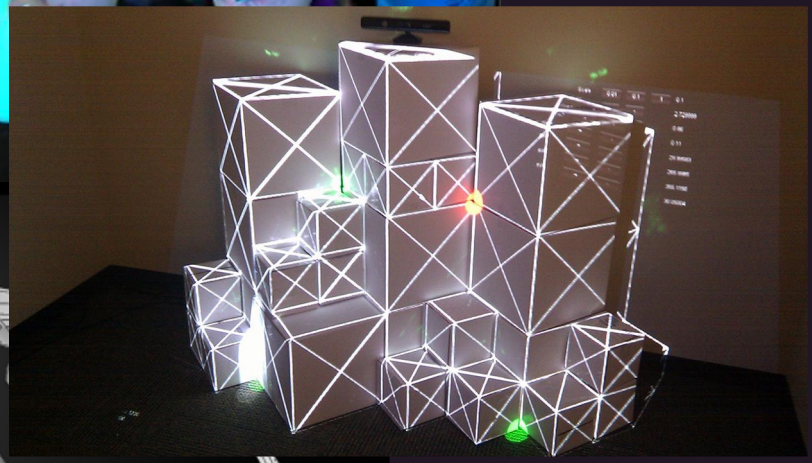
Spring School on Social XR, 2026

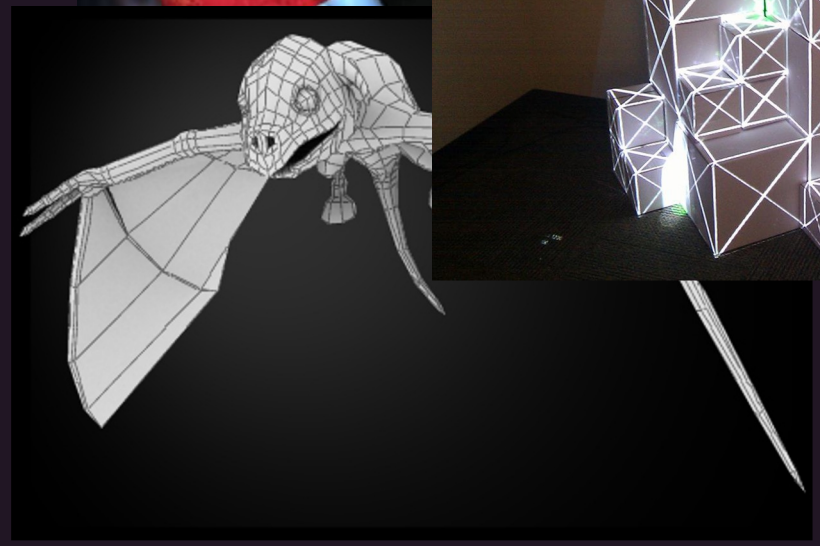
Introduction



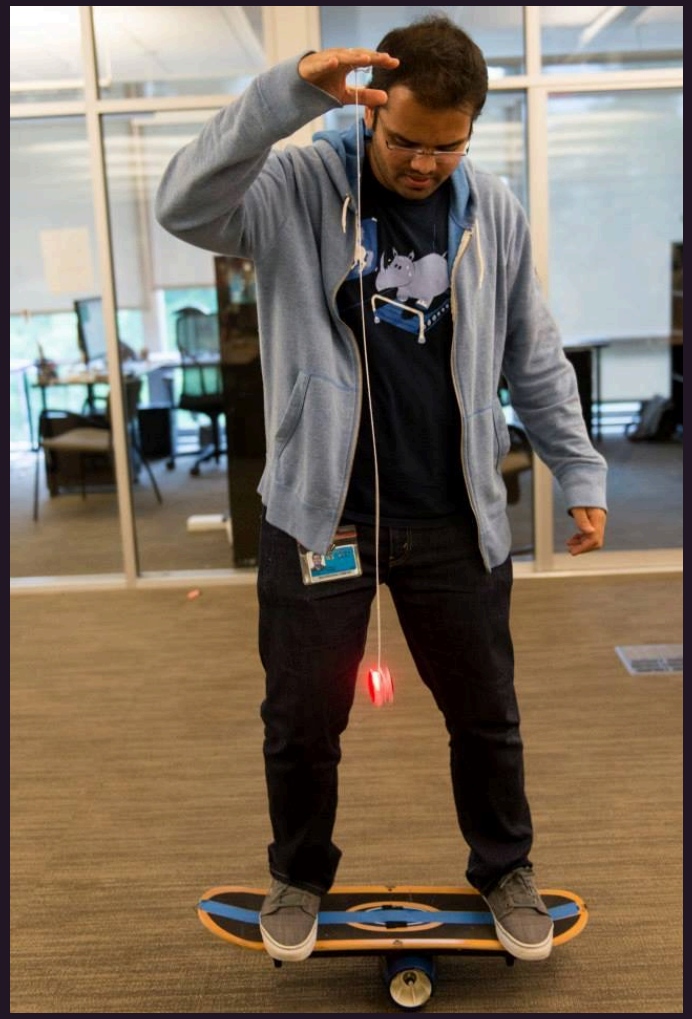


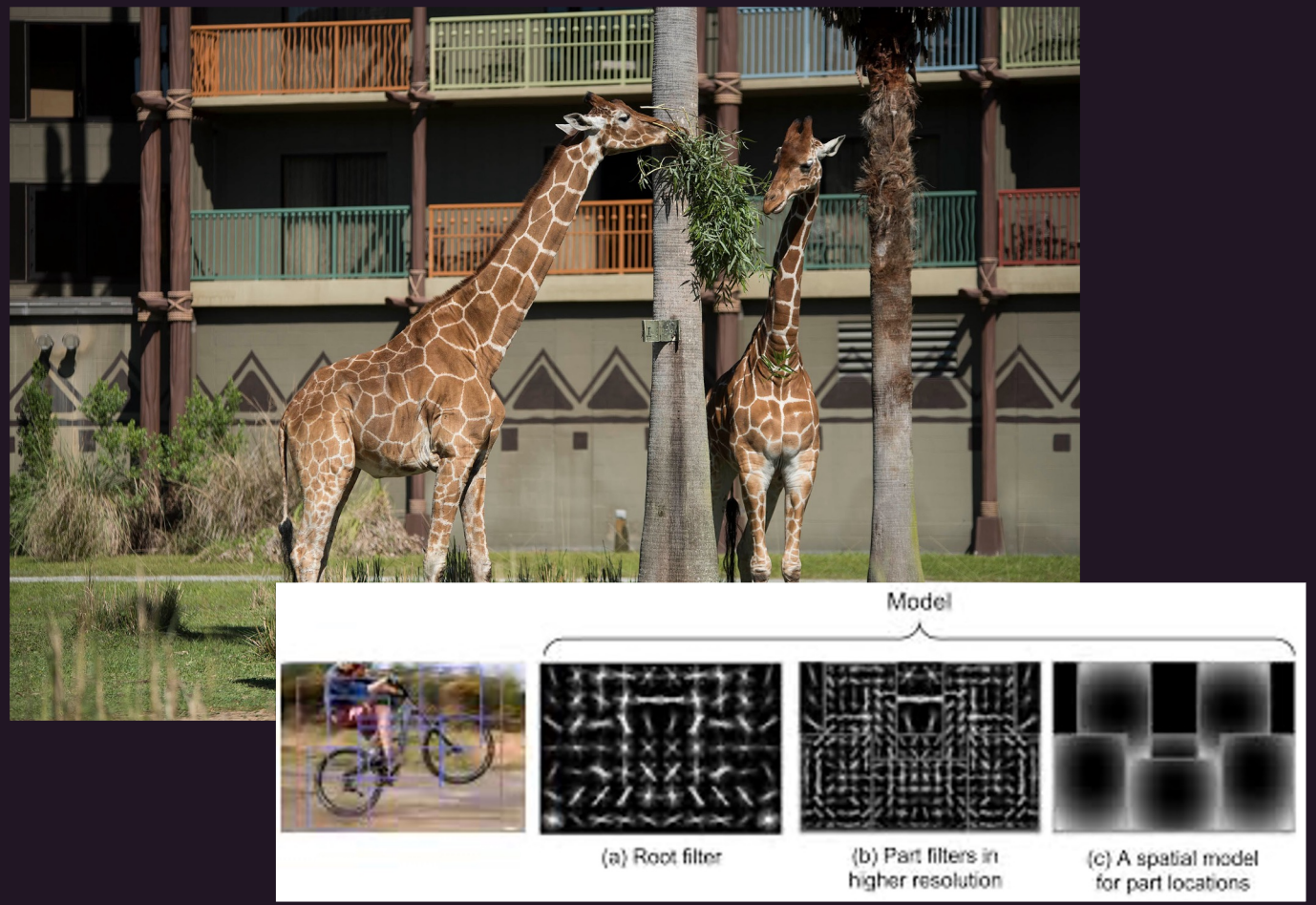
Introduction



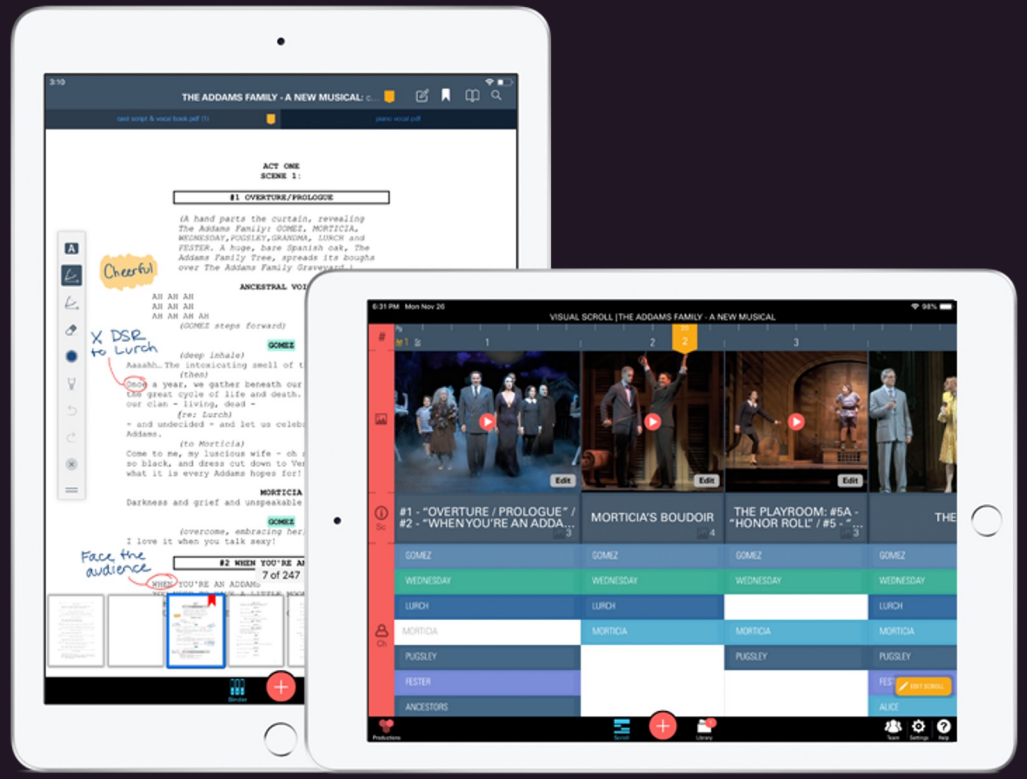


Introduction

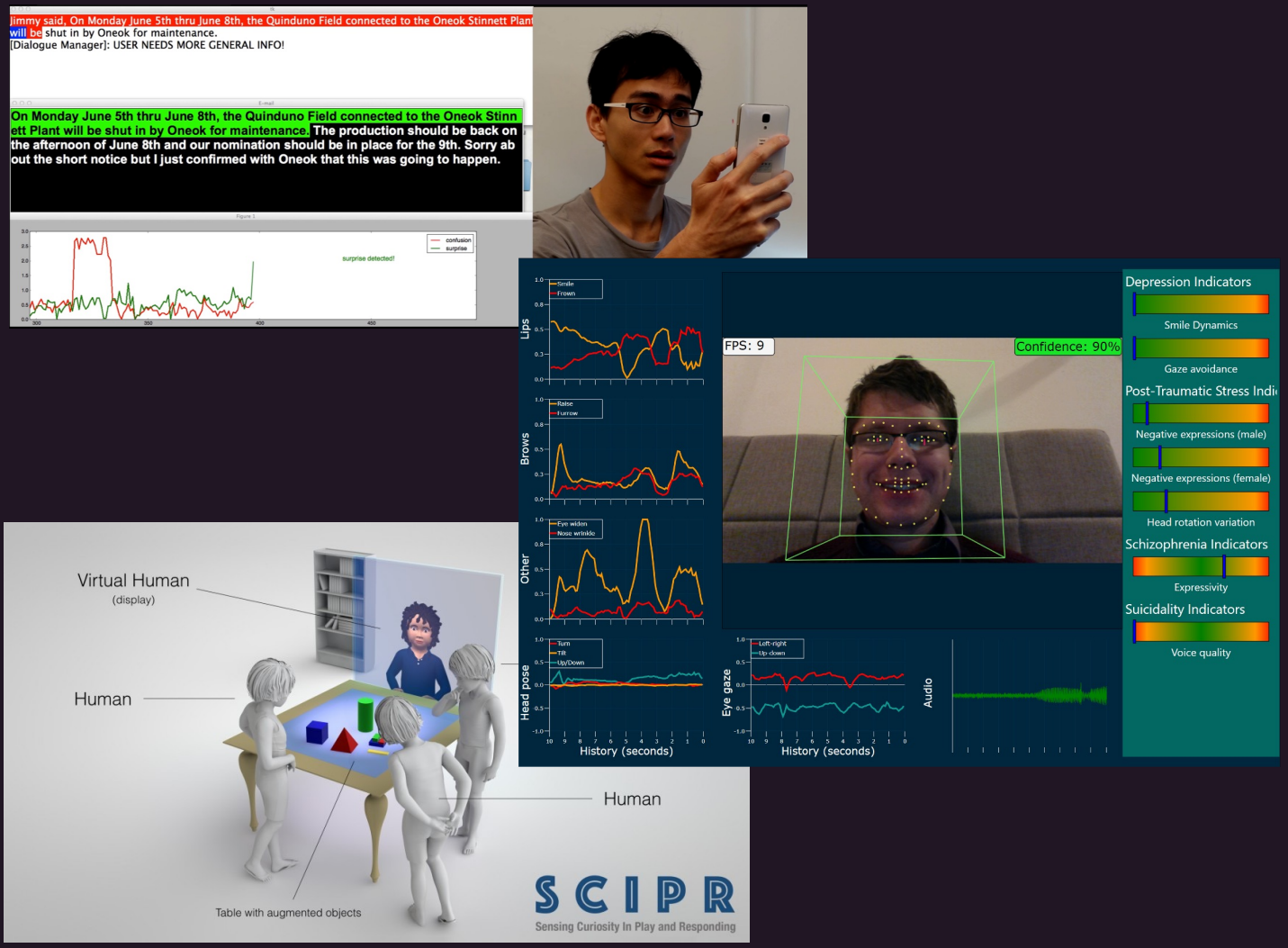




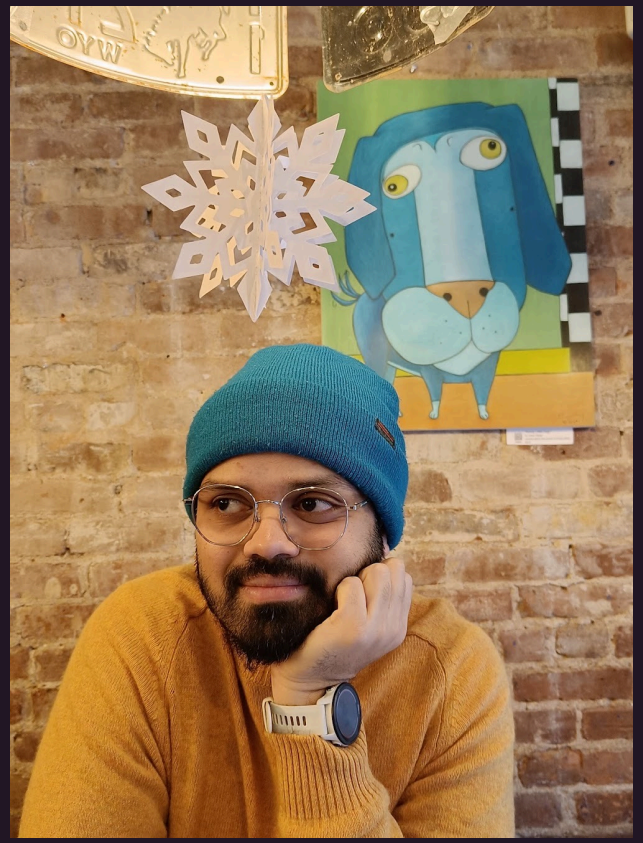
Introduction

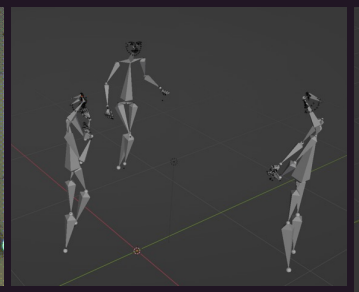


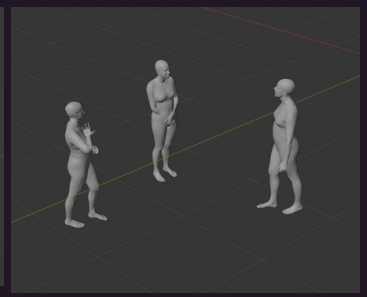
Introduction



Introduction



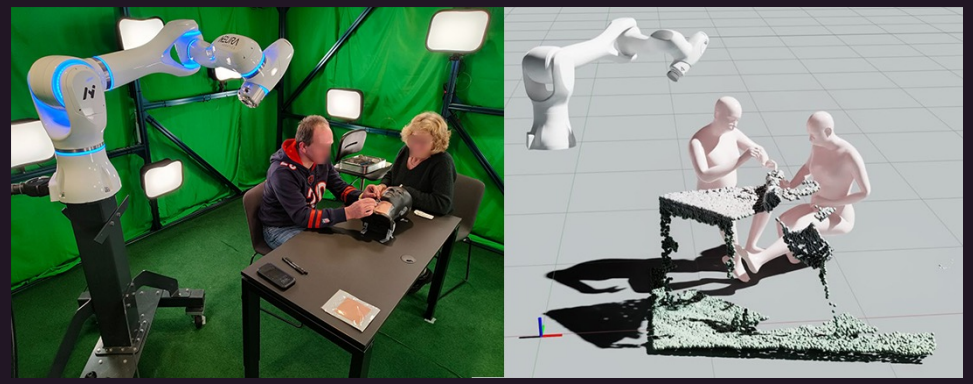




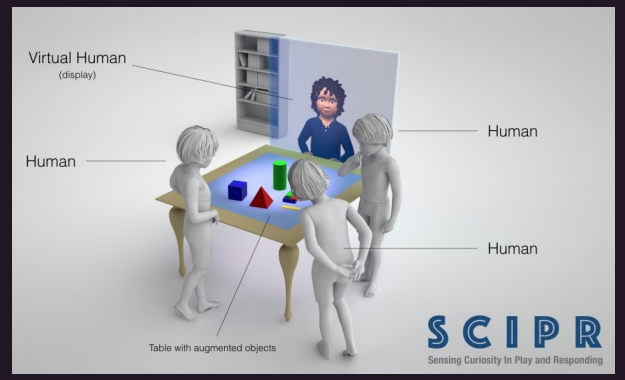
Social intelligence is inseparable from task-based competence

T. Korkiakangas, et al. (2014). “Nurse–surgeon object transfer: Video analysis of communication and situation awareness in the operating theatre”.

The Future – Human-AI Coupled Systems



ESI Lab, TU Delft



SCIPR, CMU

Perfect World Games, NVIDIA ACE

m-Team, Michigan Med

Teacher Training, TU Delft

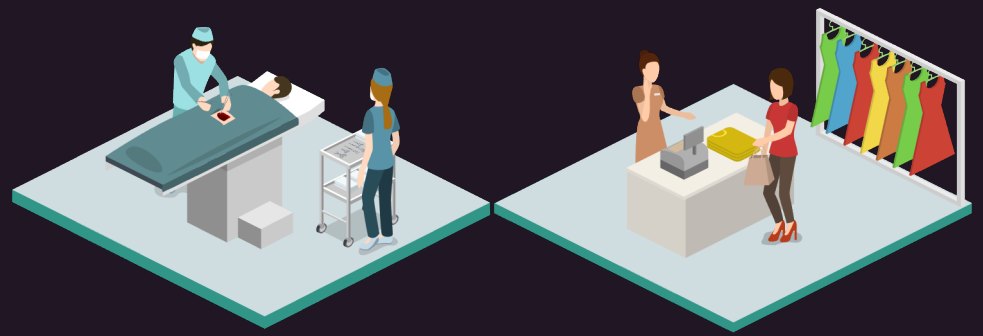
Unreal Metahumans

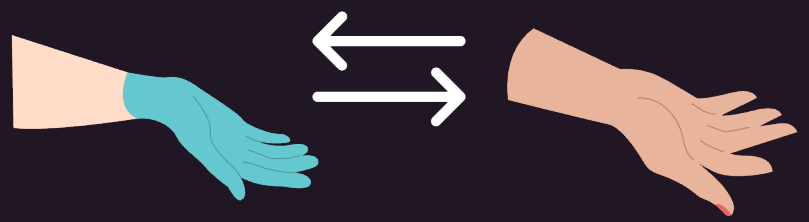
Healthcare

Education

Entertainment

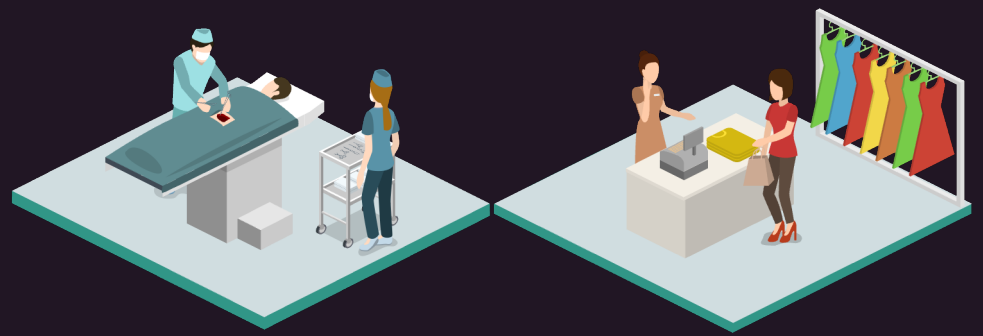
Humans generalize & evolve. AI reacts after retraining.

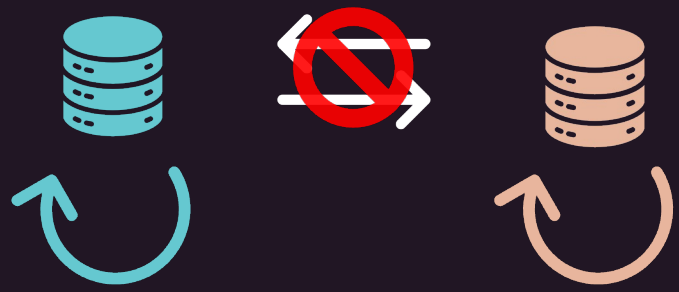




learn from outcome

Humans infer hidden physical and social causes





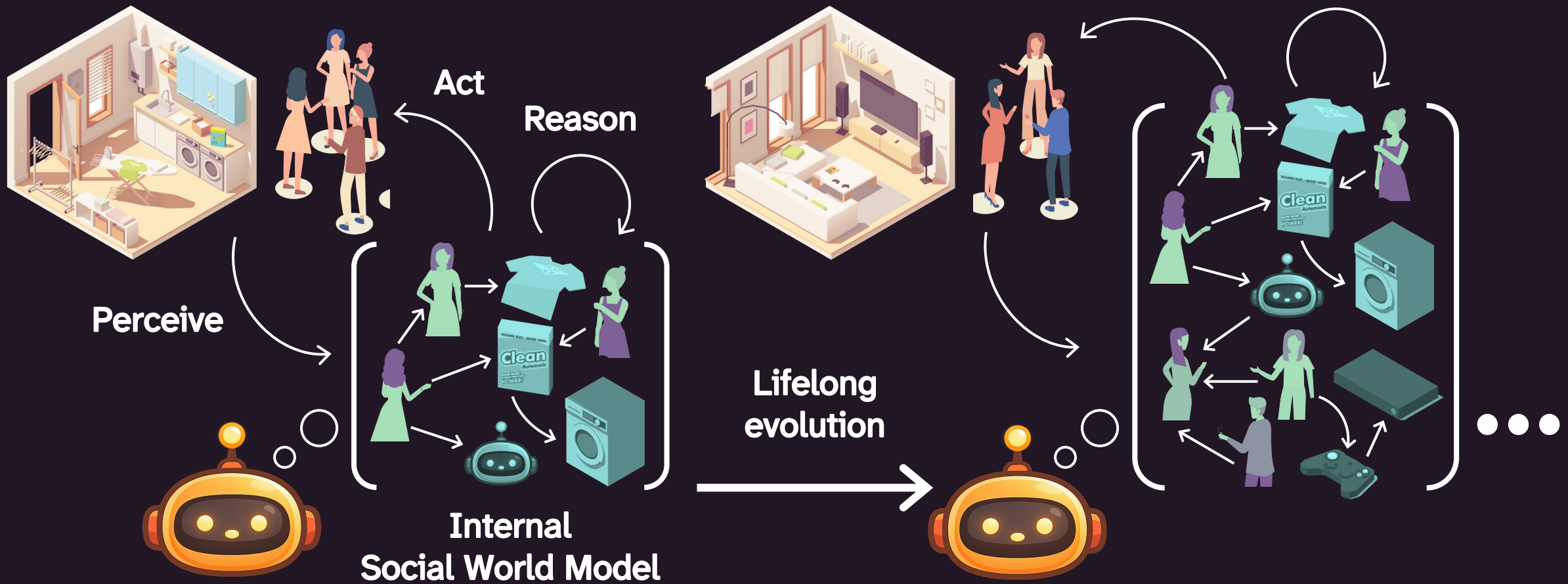
train train

Need retraining per context Fail to infer mental states

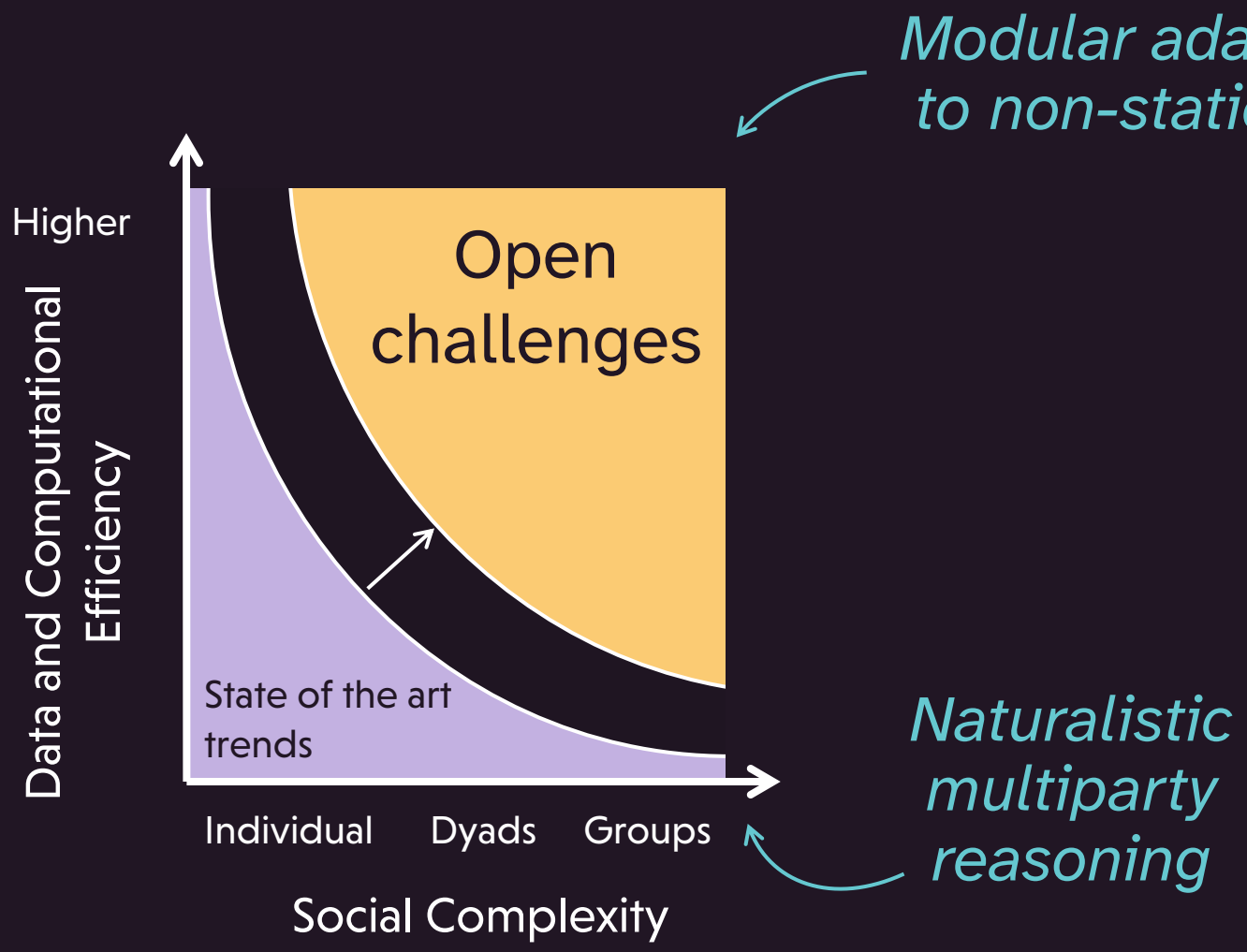
Gerstenberg, T., & Tenenbaum, J. B. (2017). “Intuitive theories”. *The Oxford Handbook of Causal Reasoning.*

Duéñez-Guzmán, E. A., et al. (2023). “A social path to human-like artificial intelligence”. *Nature Machine Intelligence.*

AI must simulate social worlds,
not just predict patterns from static data

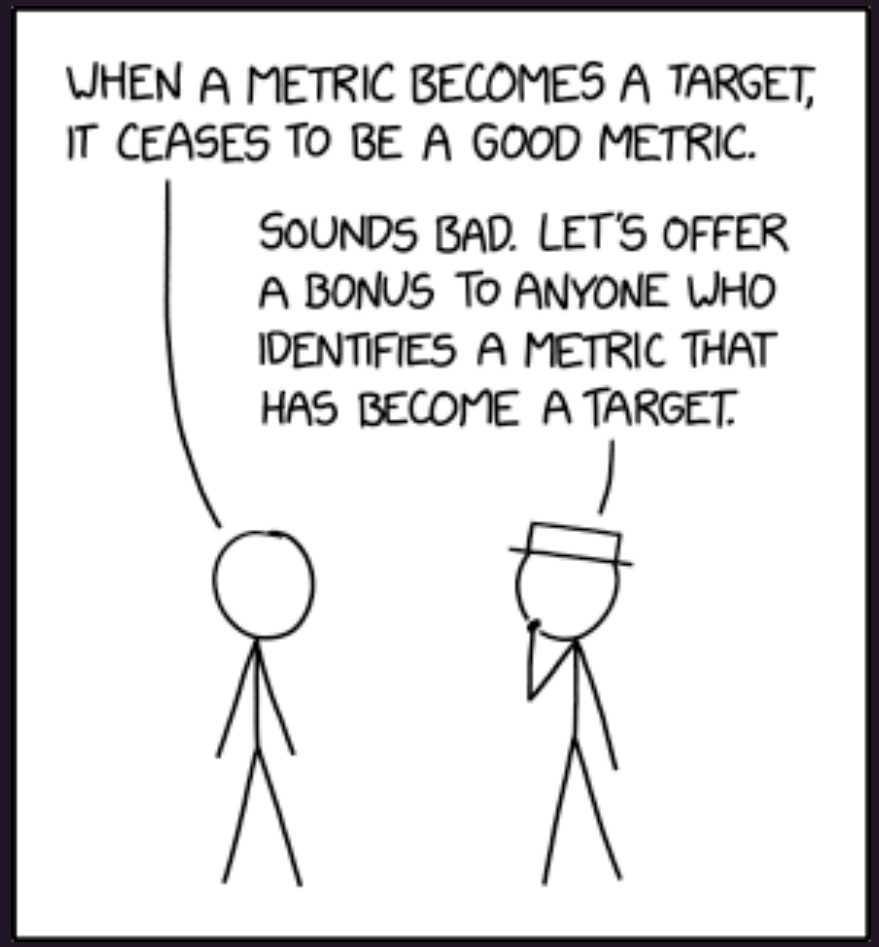

Act
Reason
Perceive
Clean
Lifelong
evolution
Internal
Social World Model

Toward Socially Intelligent, Efficient AI



Evaluation of Multiparty Social Behavior

Core challenges



[later] I'm pleased to report we're now identifying and replacing hundreds of outdated metrics per hour

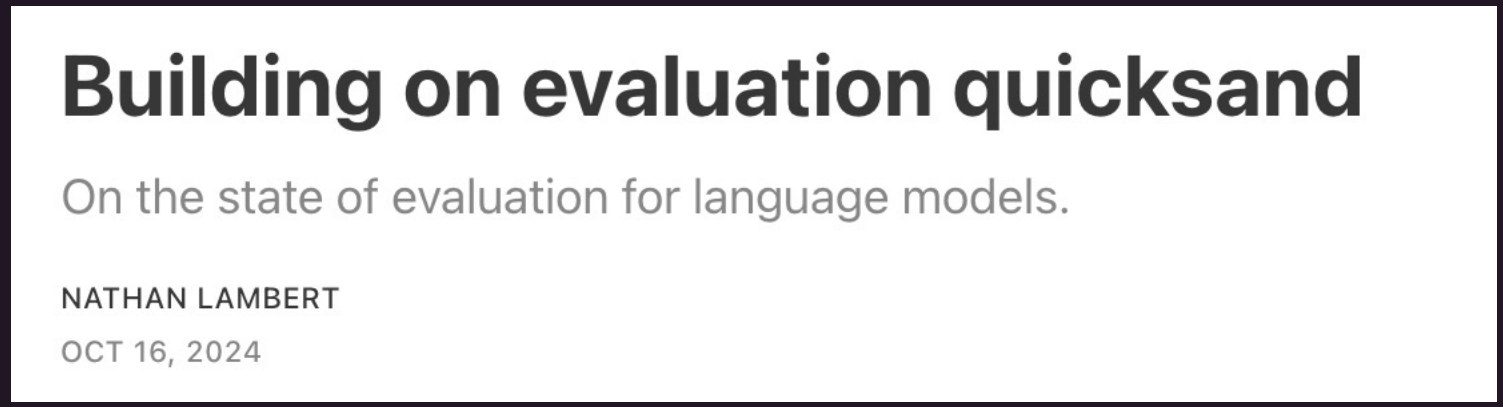
Building on evaluation quicksand

On the state of evaluation for language models.

NATHAN LAMBERT

OCT 16, 2024

Core challenges



Behavior is stochastic: multiple valid responses for a stimulus

- No internal consistency between questionnaire items[1]
- Objective metrics do not agree with subjective evaluations![2]

[1] Wolfert, P., et al. (2022) “A Review of Evaluation Practices of Gesture Generation in Embodied Conversational Agents.”. *IEEE Transactions on Human-Machine Systems*.
[2] Kucherenko, T. et al. (2022) “The GENEA Challenge 2022: A large evaluation of data-driven co-speech gesture generation.” *ICMI*.

Research Questions

1. Which objective measures are appropriate for evaluation and training of models for multimodal behavior generation?

2. How do these objective measures relate to subjective perception of synthesized behavior?

Coordination to Perception

- Do you see the subtle change in Dua's expressions?
- The coordination of her head bobs change towards the end
- Both inter/intra-person coordination affects perception
  e.g., freezing → perceived anxiety

Treffner, P., Peter, M. & Kleidon, M. Gestures and phases: The dynamics of speech-hand communication (2008).

Roelofs, K., Hagenaars, M. A. & Stins, J. Facing freeze: social threat induces bodily freeze in humans: Social threat induces bodily freeze in humans (2010).

Bernieri, F. J. & Rosenthal, R. Interpersonal coordination: Behavior matching and interactional synchrony. Fundamentals of nonverbal behavior. 511, 401–432 (1991).

Objective Metrics of Behavior Coordination

Cross Recurrance Quantification Analysis (CRQA)

Transient synchrony

Measures stability and predictability

Beat Consistency

Empirical Mode Decomposition (EMD) across different rhythmic frequencies

Handles non-stationary signals without assuming fixed basis

Soft-Dynamic Time Warping

Compares the shape of motion or pitch contours rather than rigid clock time

Subjective Measures of Perception

Perceived Conversation Quality	Artificial Social Agent Questionnaire
Inter-personal Relationships	Human-likeness
Nature of Interaction	Natural Appearance
Equal Opportunity	Social Presence

Raman, Chirag, Navin Raj Prabhu, and Hayley Hung. “Perceived Conversation Quality in Spontaneous Interactions.” IEEE Transactions on Affective Computing, 2023, 1–13. https://doi.org/10.1109/taffc.2023.3233950.

Fitrianie, Siska, Merijn Bruijnes, Fengxiang Li, Amal Abdulrahman, and Willem-Paul Brinkman. “The Artificial-Social-Agent Questionnaire: Establishing the Long and Short Questionnaire Versions.” In Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents. New York, NY, USA: ACM, 2022. https://doi.org/10.1145/3514197.3549612.

How do we understand coordination?

Objective Metrics

- CRQA
- Beat Consistency
- Soft-DTW

Subjective Measures

- Perceived Conversation Quality
- Artificial Social Agent Questionnaire

How do we understand coordination?

Objective Metrics

Subjective Measures

CRQA

Beat Consistency

Soft-DTW

Perceived Conversation Quality

Artificial Social Agent Questionnaire

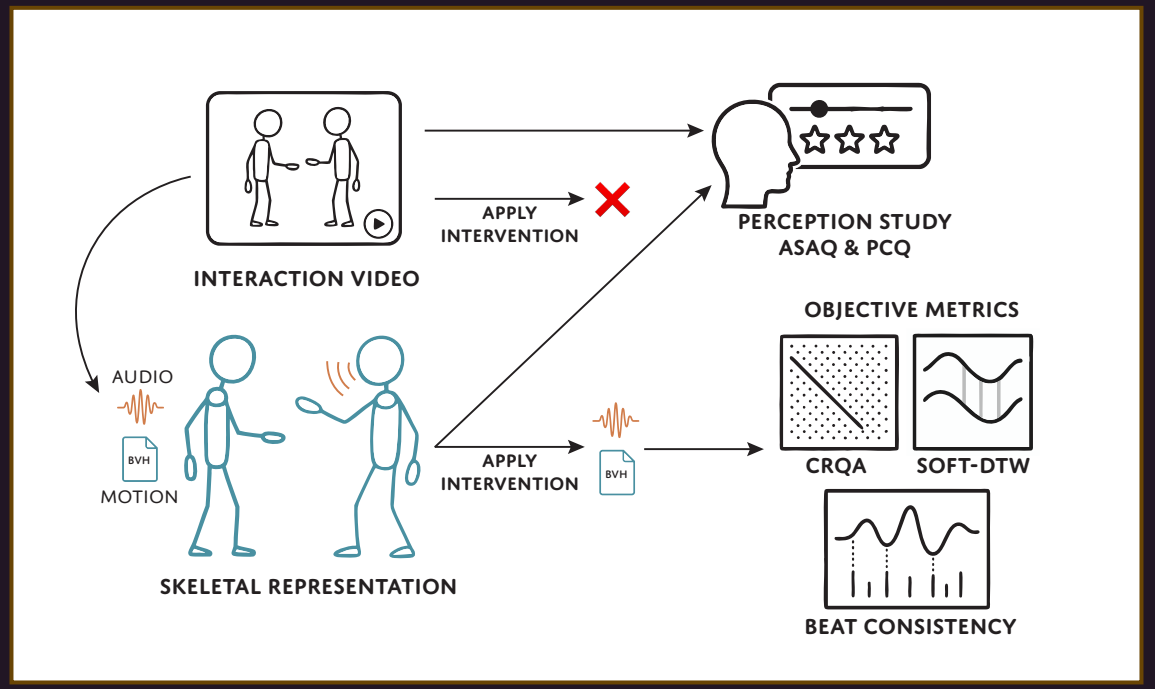
But First, Metric Sensitivity

We investigate how the metrics affect each other

Interventional Study!

- Audio-Gesture Desync
- Pitch Variance Limits
- Gesture Dampening

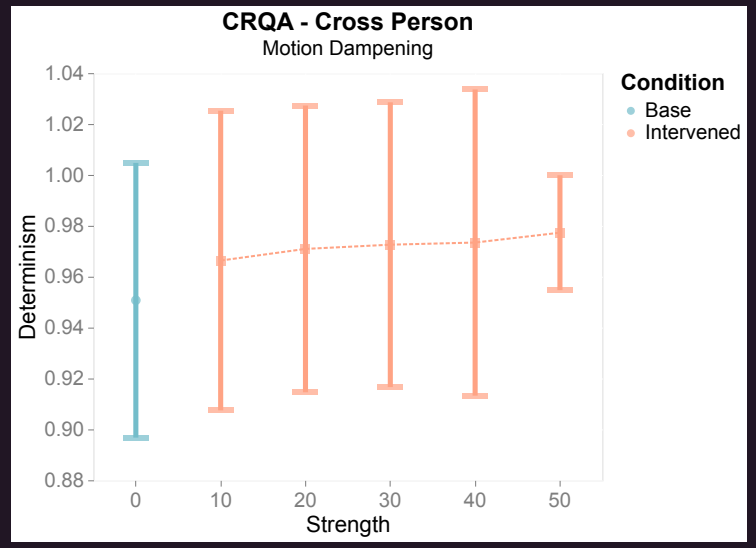
Applied at various strength levels

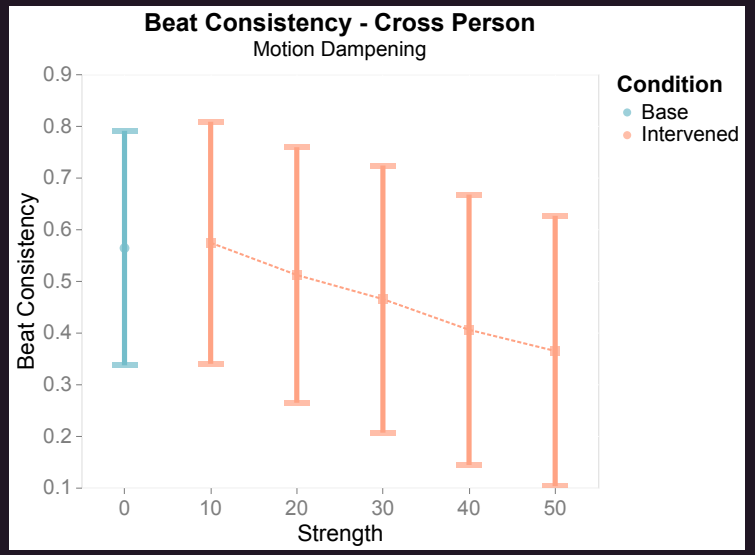


Metric Sensitivity: What did we find?

Kinematic Dampening (Smoothing hand movements):

- CRQA: determinism (%DET) increased
- Beat Consistency: decreased



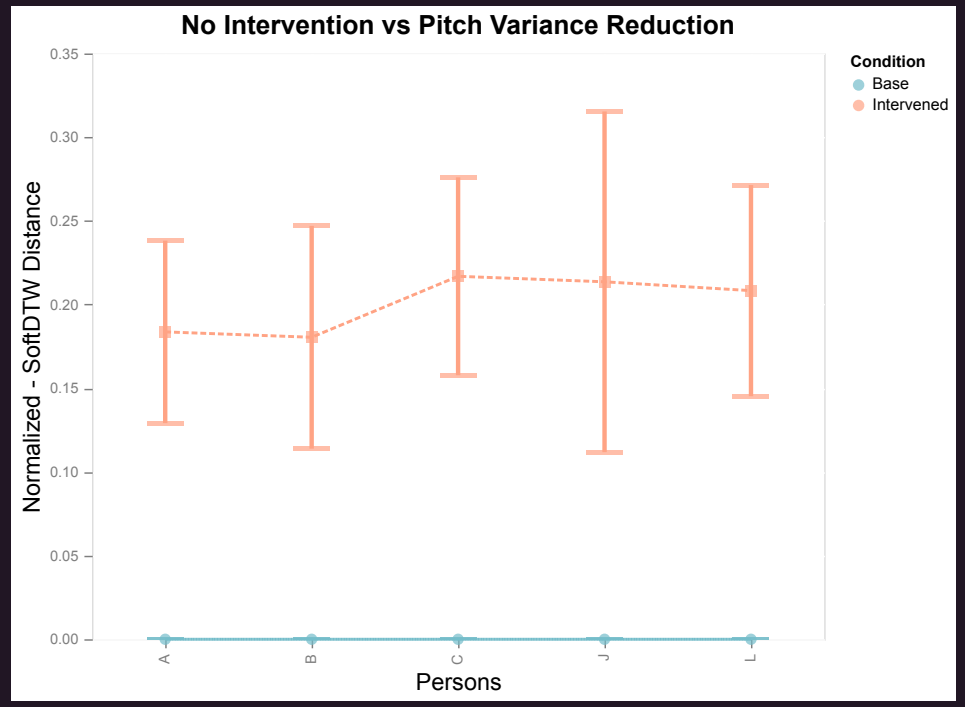


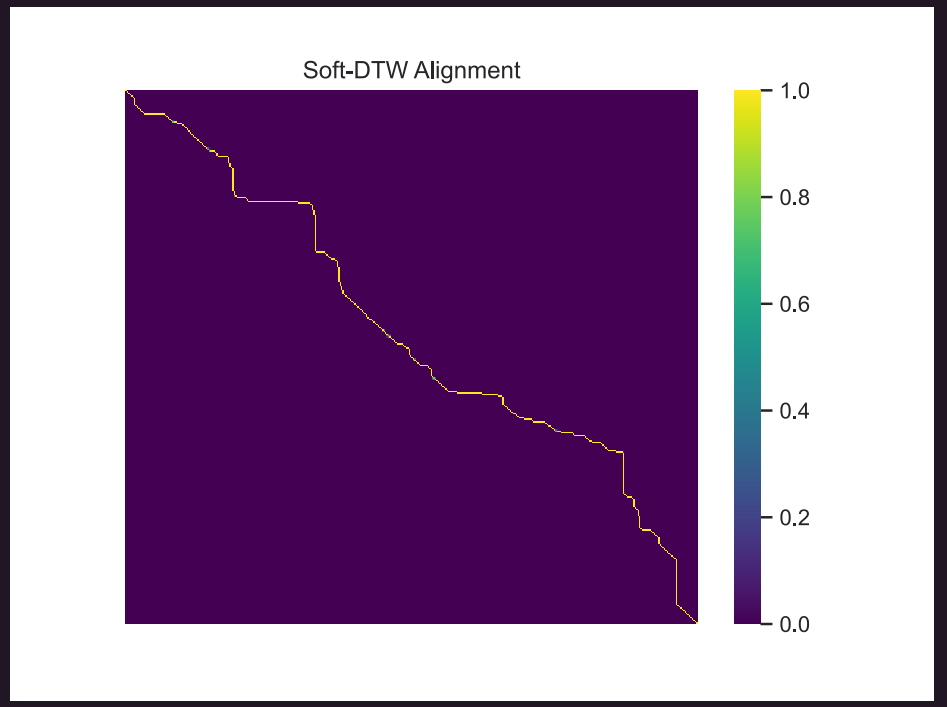
Shirekar, Ojas, Wim Pouw, Chenxu Hao, Vrushank Phadnis, Thabo Beeler, and Chirag Raman. “Multimodal Quantitative Measures for Multiparty Behavior Evaluation.” In Proceedings of the 27th International Conference on Multimodal Interaction, 249–64. New York, NY, USA: ACM, 2025. https://doi.org/10.1145/3716553.3750752.

Metric Sensitivity: What did we find?

Pitch-Variance Reduction (flattening prosody):

- Soft-DTW (alignment): decreased



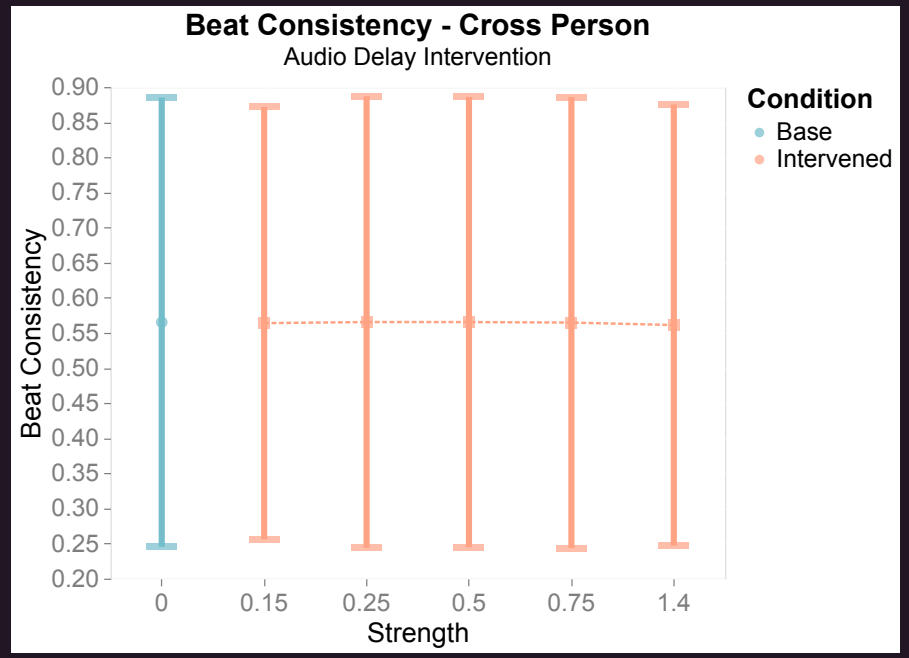


Shirekar, Ojas, Wim Pouw, Chenxu Hao, Vrushank Phadnis, Thabo Beeler, and Chirag Raman. “Multimodal Quantitative Measures for Multiparty Behavior Evaluation.” In Proceedings of the 27th International Conference on Multimodal Interaction, 249–64. New York, NY, USA: ACM, 2025. https://doi.org/10.1145/3716553.3750752.

Metric Sensitivity: What did we find?

Speech-Gesture Delays:

- Beat Consistency: decreased, but effect is less significant



Shirekar, Ojas, Wim Pouw, Chenxu Hao, Vrushank Phadnis, Thabo Beeler, and Chirag Raman. “Multimodal Quantitative Measures for Multiparty Behavior Evaluation.” In Proceedings of the 27th International Conference on Multimodal Interaction, 249–64. New York, NY, USA: ACM, 2025. https://doi.org/10.1145/3716553.3750752.

What's next?

Objective Metrics

CRQA

Beat Consistency

Soft-DTW

Subjective Measures

Perceived Conversation Quality

Artificial Social Agent Questionnaire

Large-scale citizen-science perception study

Dr. Wim Pouw
Cognitive Science
Tilburg University

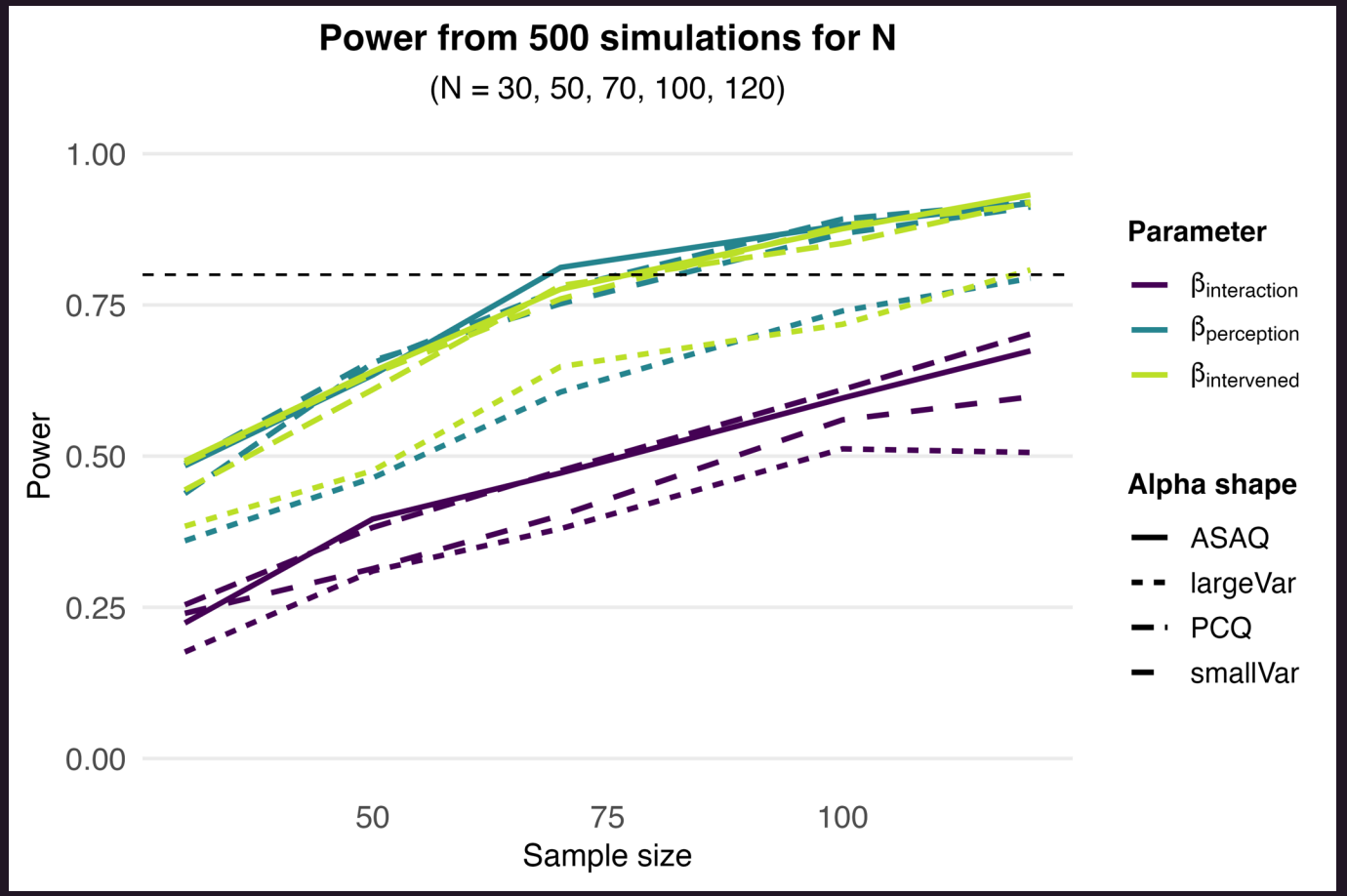
Sample Stimulus

Mughal, Muhammad Hamza, Rishabh Dabral, Ikhsanul Habibie, Lucia Donatelli, Marc Habermann, and Christian Theobalt. “ConvoFusion: Multi-Modal Conversational Diffusion for Co-Speech Gesture Synthesis.” arXiv [Cs.CV], March 26, 2024. arXiv. http://arxiv.org/abs/2403.17936.

Sample Stimulus

Power Analysis

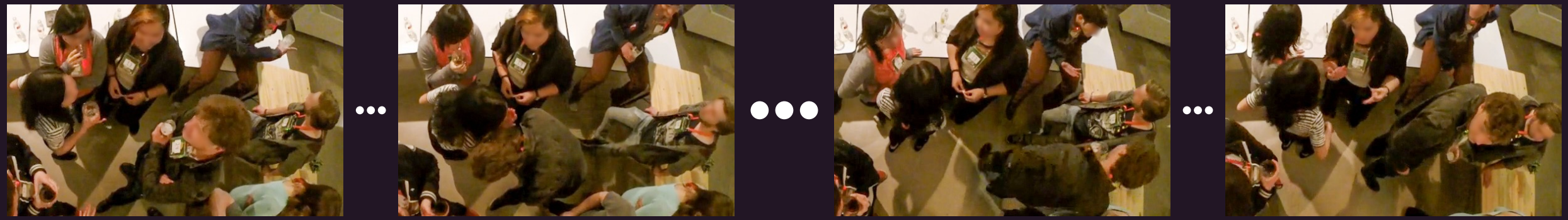
We need roughly 128 people!



Generating Multiparty Social Behavior

Modeling in-the-wild interactions

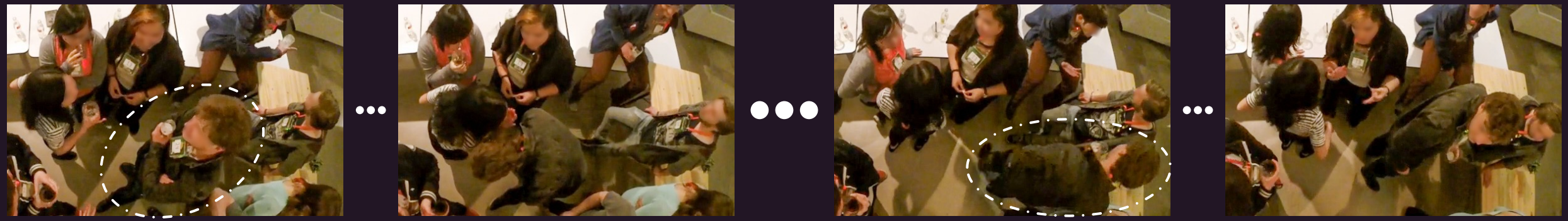
MatchNMingle Dataset



observed **future**

Modeling in-the-wild interactions

MatchNMingle Dataset

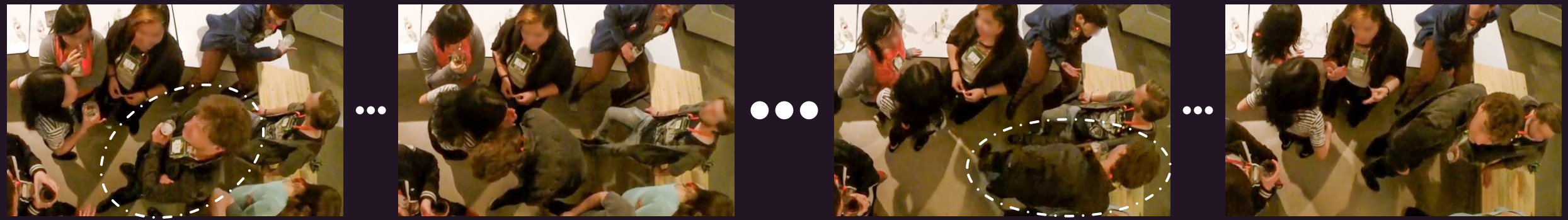


observed

future

Modeling in-the-wild interactions

MatchNMingle Dataset



observed

future

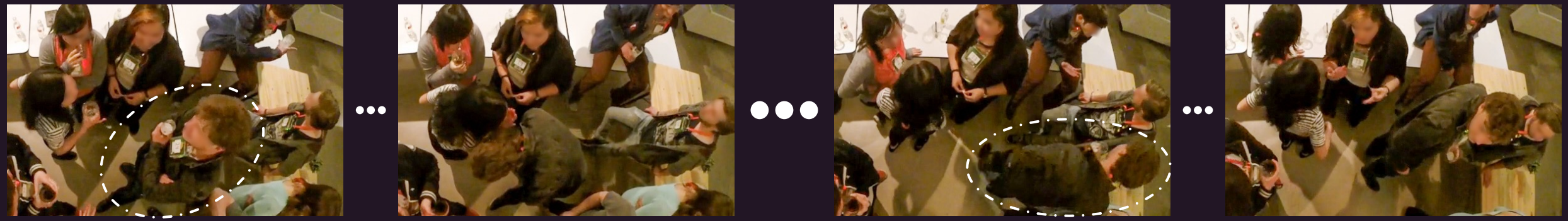
group leaving : y = 1

200 instances over 90 mins
[Van Doorn 2018]

X := { cues over observed seq. }

Modeling in-the-wild interactions

MatchNMingle Dataset



observed

future

$X :=$ { cues over observed seq. }

group leaving : y = 1

200 instances over 90 mins
[Van Doorn 2018]

$Y :=$ { cues over future seq }

Key Observation 1

The *social signal*[1] —the high-level attitudes and social meaning transferred in interactions—is embedded in the low-level cues[2]

1. Ambady et al. 2000
2. Vinciarelli et al. 2009

Challenges and Design Considerations

Multiple socially valid futures possible for an observed sequence

$$\boldsymbol{Y} = f(\boldsymbol{X}) \rightarrow \text{model } p(\boldsymbol{Y}|\boldsymbol{X})$$

Participants' behaviors are interdependent - [Goffman 1963, Kendon 1990]

Capture uncertainty at *global* (group) level –
jointly forecast one future for all participants at a time

Key Observation 2

Social dynamics & behavior coordination are unique for every unique grouping of individuals

Moore 2013, Kendon 1990

Key Idea

View unique groups as meta-learning *tasks* to generalize to unseen groups

Model $p(\boldsymbol{Y}|\boldsymbol{X}, C)$,

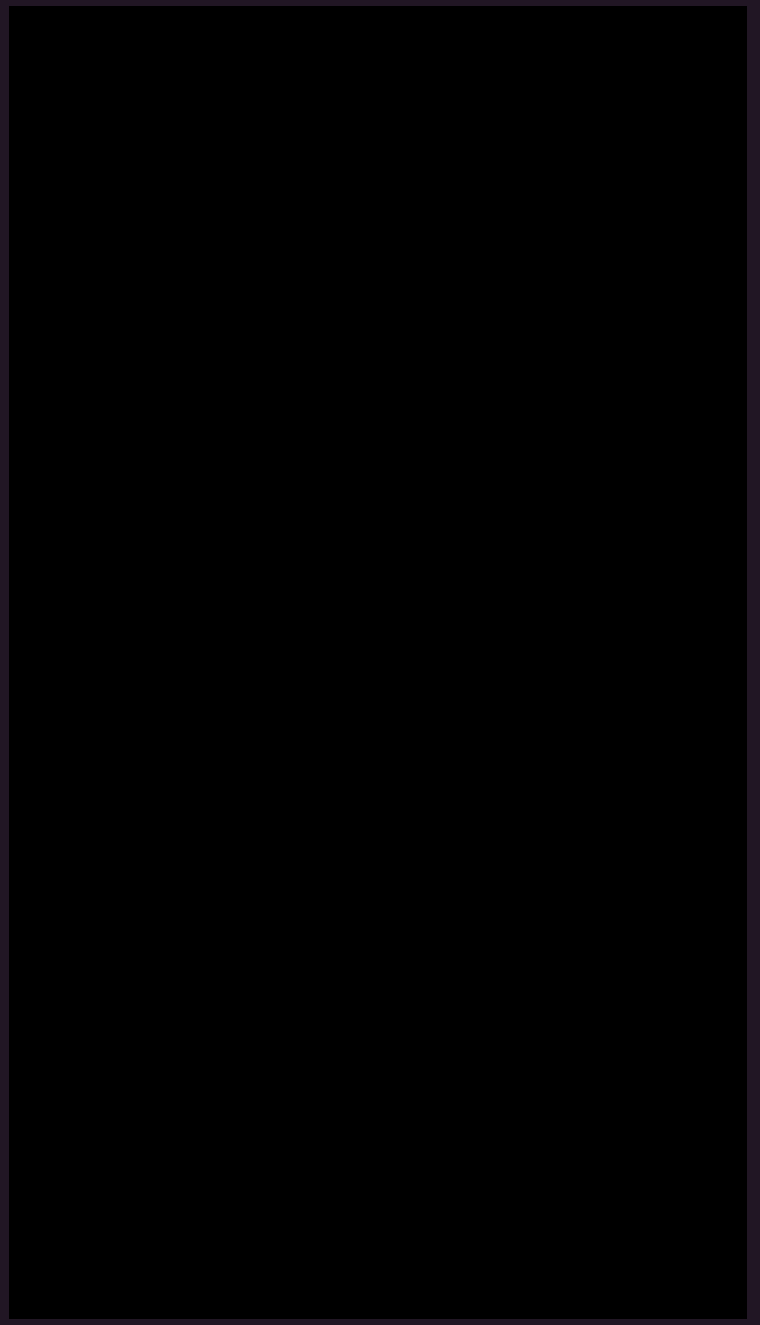
where C denotes a set of context interaction sequences for a group

Social processes: Self-supervised meta-learning over conversational groups for forecasting nonverbal social cues – Raman et al.

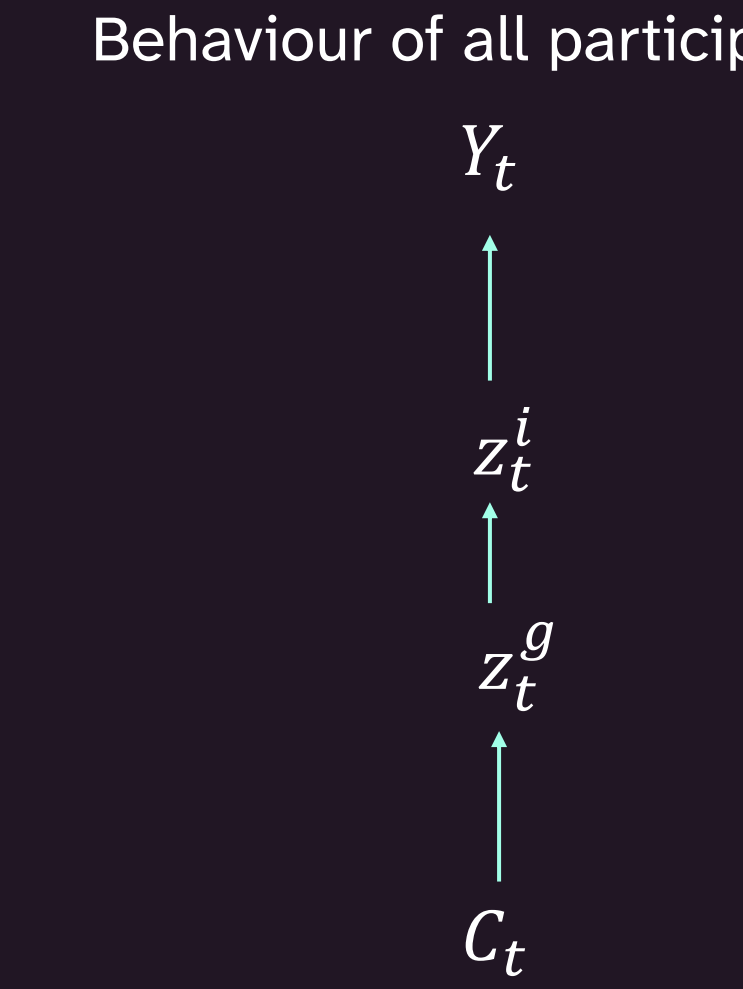
Behavior Interdependence

What do we see here?

- First case: group affects individual
- Second case: individual affects the group

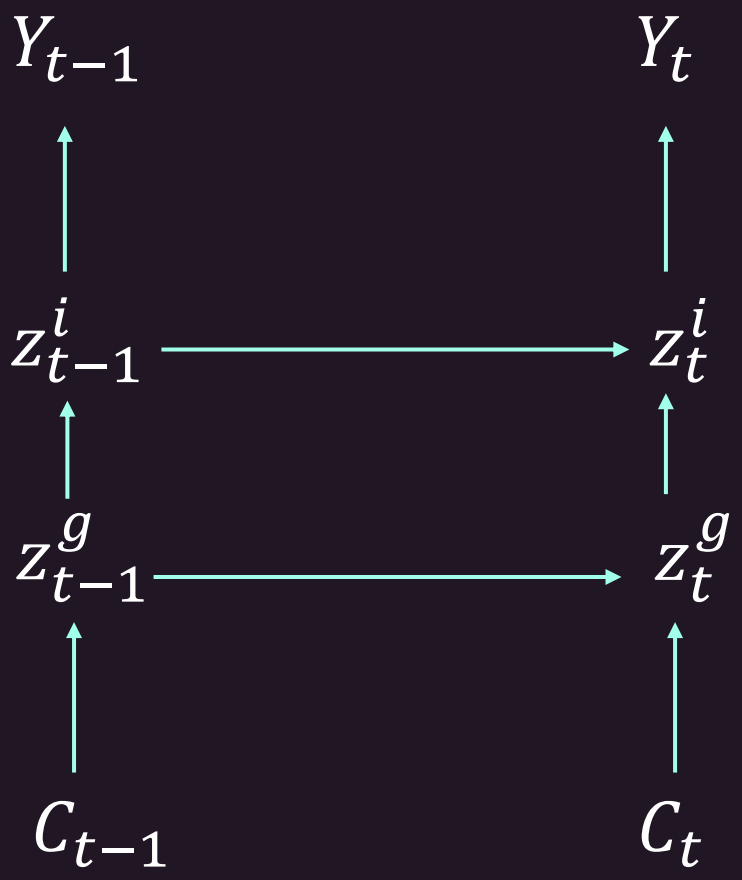


Hierarchical Sequential Neural Processes



Hierarchical Sequential Neural Processes

Behaviour of all participants

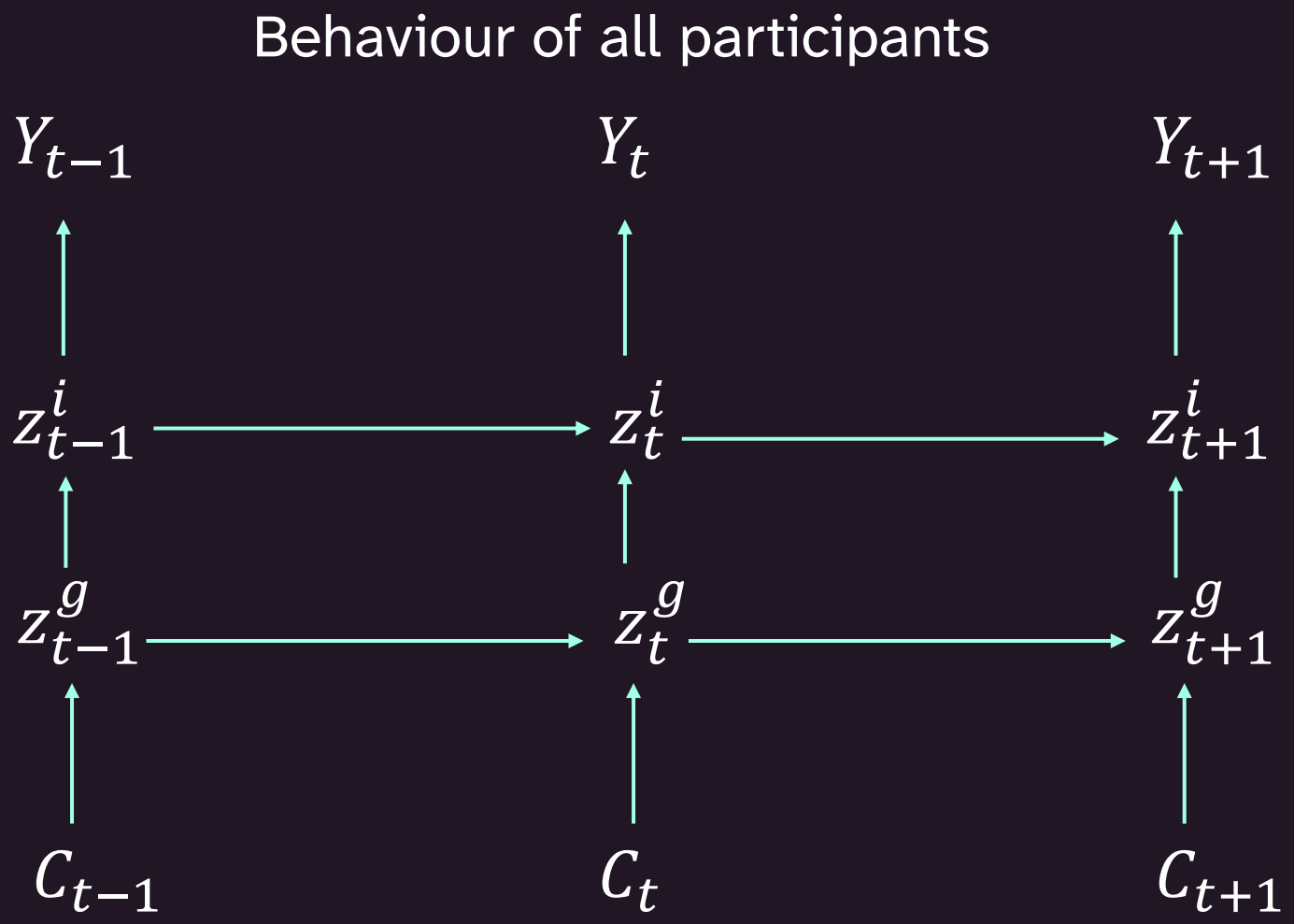


Learn z_t evolution, capturing the underlying data dynamics

Social Context (Partially Observed Features); C_t may be null

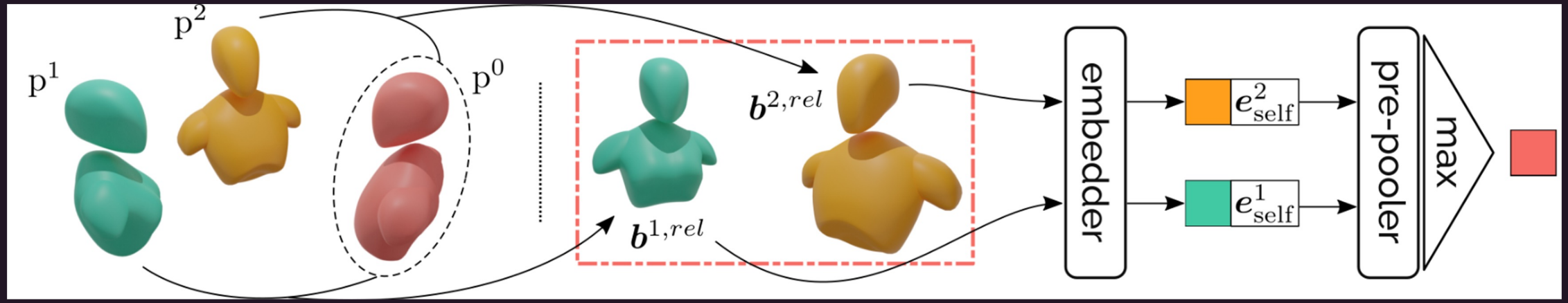
Hierarchical Sequential Neural Processes

Learn z_t evolution, capturing the underlying data dynamics

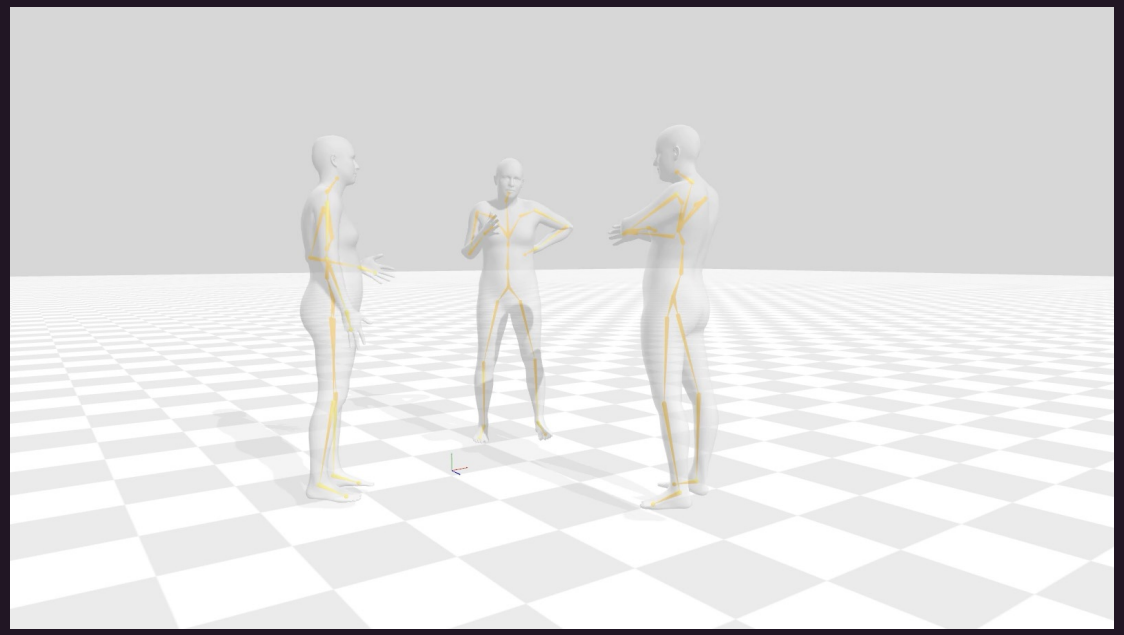


Incorporating partner behavior

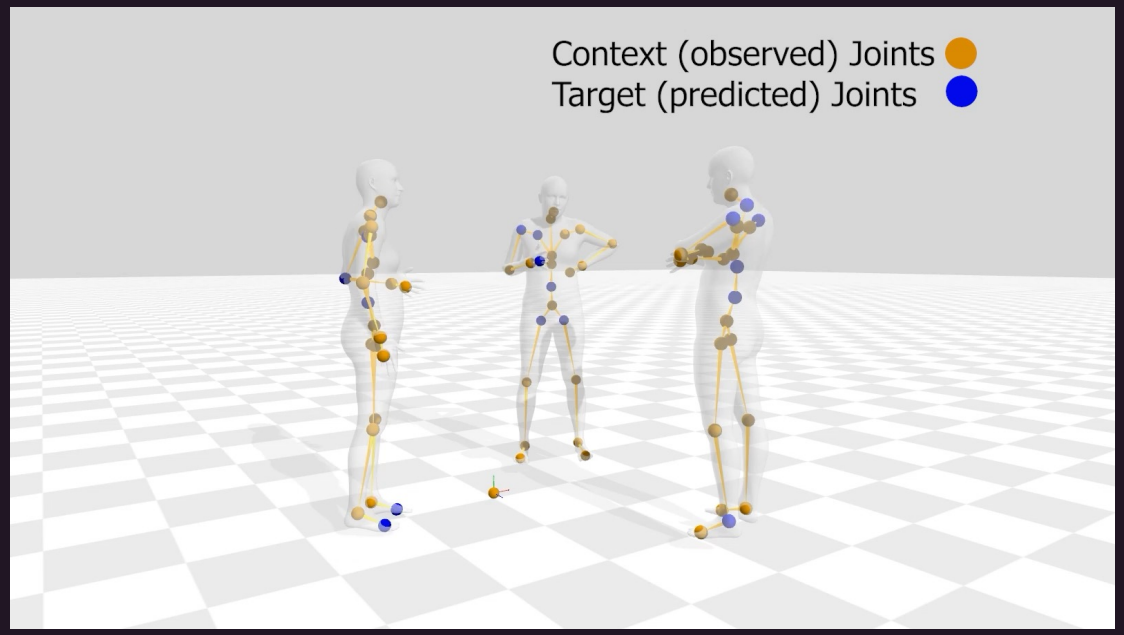
Encode the interaction from the perspective of each participant



WIP Predictions



Ground Truth



Generation

Recap

- Socially intelligent agents need to simulate social worlds and support lifelong evolution from experience
- Evaluation of simulated multiparty behavior remains challenging
  - Metric Sensitivity [CRQA, Beat Consistency, Soft-DTW]
  - Mapping low-level metrics to higher-order perception
- A hierarchical, sequential latent-variable model for capturing individual and group dynamics in generating behavior

Thanks! Questions?

Ojas Shirekar
Computer Science
TU Delft

Dr. Chenxu Hao
Cognitive Psychology
TU Delft

Dr. Vrushank Phadnis
UX & HCI
Google Research

Dr. Wim Pouw
Cognitive Science
Tilburg University