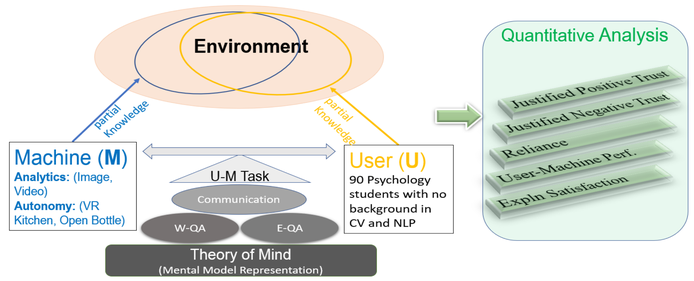
Explainable AI as Collaborative Task Solving

Arjun Akula[1] Changsong Liu[1] Sinisa Todorovic[2] Joyce Chai[3] Song-Chun Zhu[1]

[1] UCLA Center for Vision, Cognition, Learning, and Autonomy
[2] Oregon State University, [3] Michigan State University

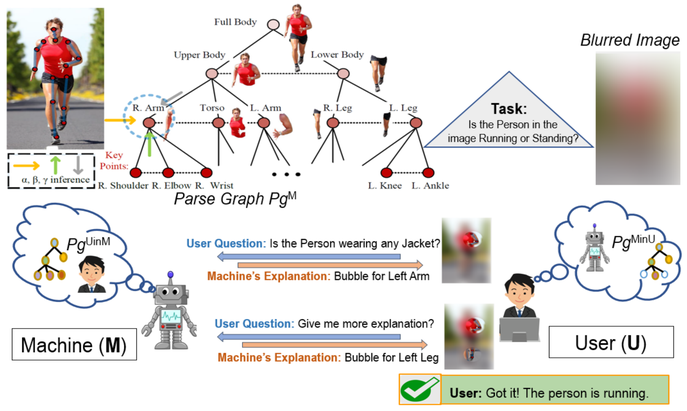
CVPR LONG BEACH CALIFORNIA June 16-20, 2019

X-ToM:Explanation with Theory-of-Mind



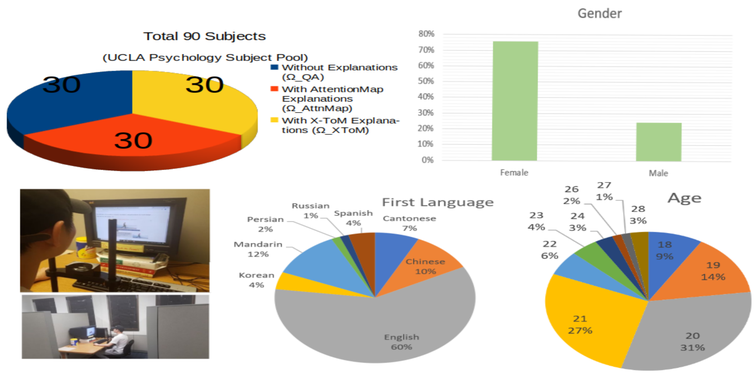
X-ToM for optimizing the dialog with a user towards estimating and increasing human trust.

XAI as Collaborative Task Solving

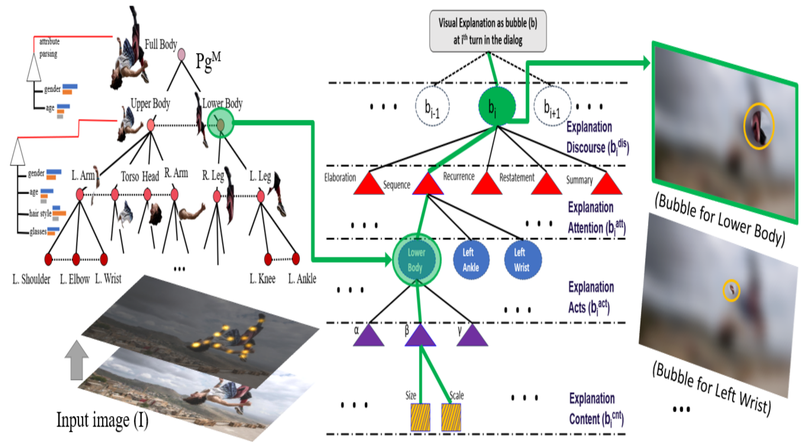


Through a dialog, we estimate Trust and Reliance in terms of pg^M, pg^{UinM} and pg^{MinU}.

Subject Pool



Explanation Generation in X-ToM



Both the Machine and the User solve the image recognition tasks. The Machine interprets the image I as pg^M. The Human receives visual explanations – bubbles – optimized by the X-ToM Explainer.

Trust Estimation

It is possible for humans to feel positive trust with respect to certain tasks, while feeling negative trust (i.e. mistrust) on some other tasks.

Justified Positive Trust (JPT):

$$\text{JPT} = \frac{1}{N}\sum_i \sum_{z=\alpha,\beta,\gamma} \Delta\text{JPT}(i,z),$$

$$\Delta\text{JPT}(i,z) = \frac{\|pg^{MinU}_{i,z,+} \cap pg^{M}_{i,+}\|}{\|pg^{M}_{i,+}\|}$$

Justified Negative Trust (JNT):

$$\text{JNT} = \frac{1}{N}\sum_i \sum_{z=\alpha,\beta,\gamma} \Delta\text{JNT}(i,z),$$

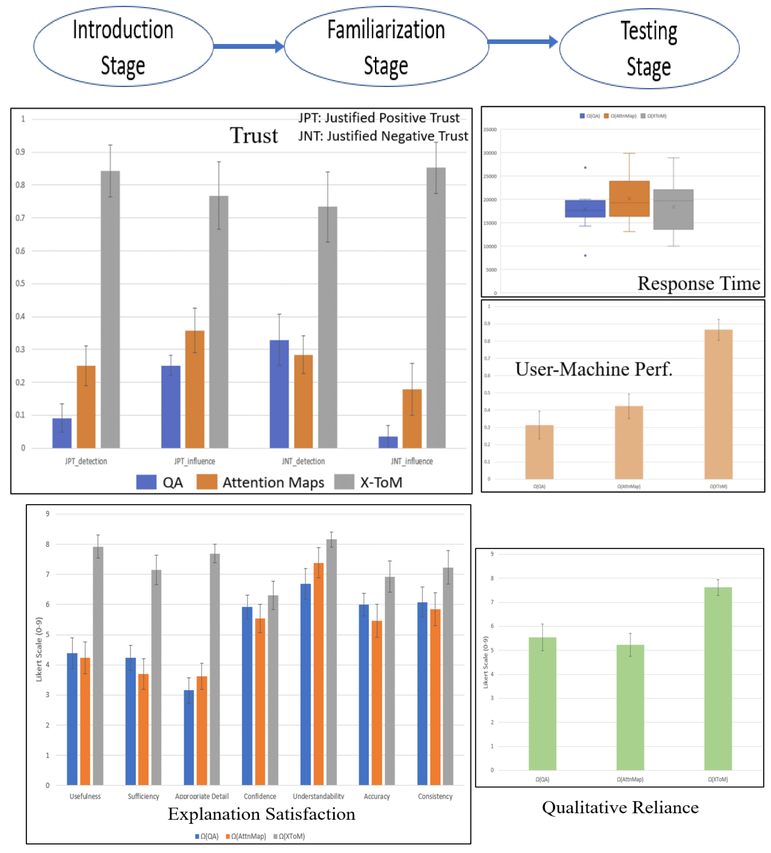
$$\Delta\text{JNT}(i,z) = \frac{\|pg^{MinU}_{i,z,-} \cap pg^{M}_{i,-}\|}{\|pg^{M}_{i,-}\|}$$

Reliance (Rc): Reliance captures the extent to which a human can accurately predict the performer's inference results without over- or under-estimation.

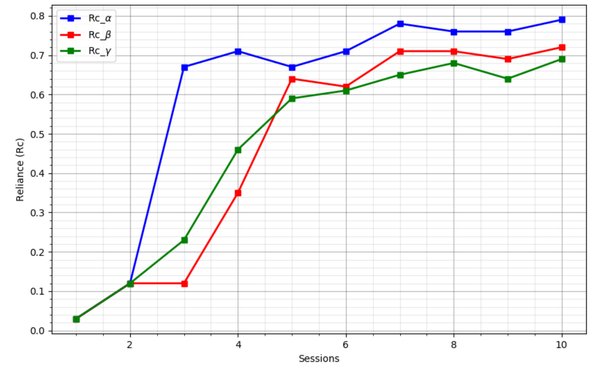
$$\text{Rc} = \frac{1}{N}\sum_i \sum_{z=\alpha,\beta,\gamma} \Delta\text{Rc}(i,z),$$

$$\Delta\text{Rc}(i,z) = \frac{\|pg^{MinU}_{i,z} \cap pg^{M}_{i,z}\|}{\|pg^{M}_{i}\|}$$

Results of our Human Study



X-ToM significantly outperformed ($p < 0.01$) baselines (QA, Attention Maps) in terms of Appropriate Trust, Reliance, User-Machine Performance and Satisfaction.



Gain in Reliance over sessions w.r.t α, β and γ processes. As we can see, with more sessions, we can further improve human reliance.