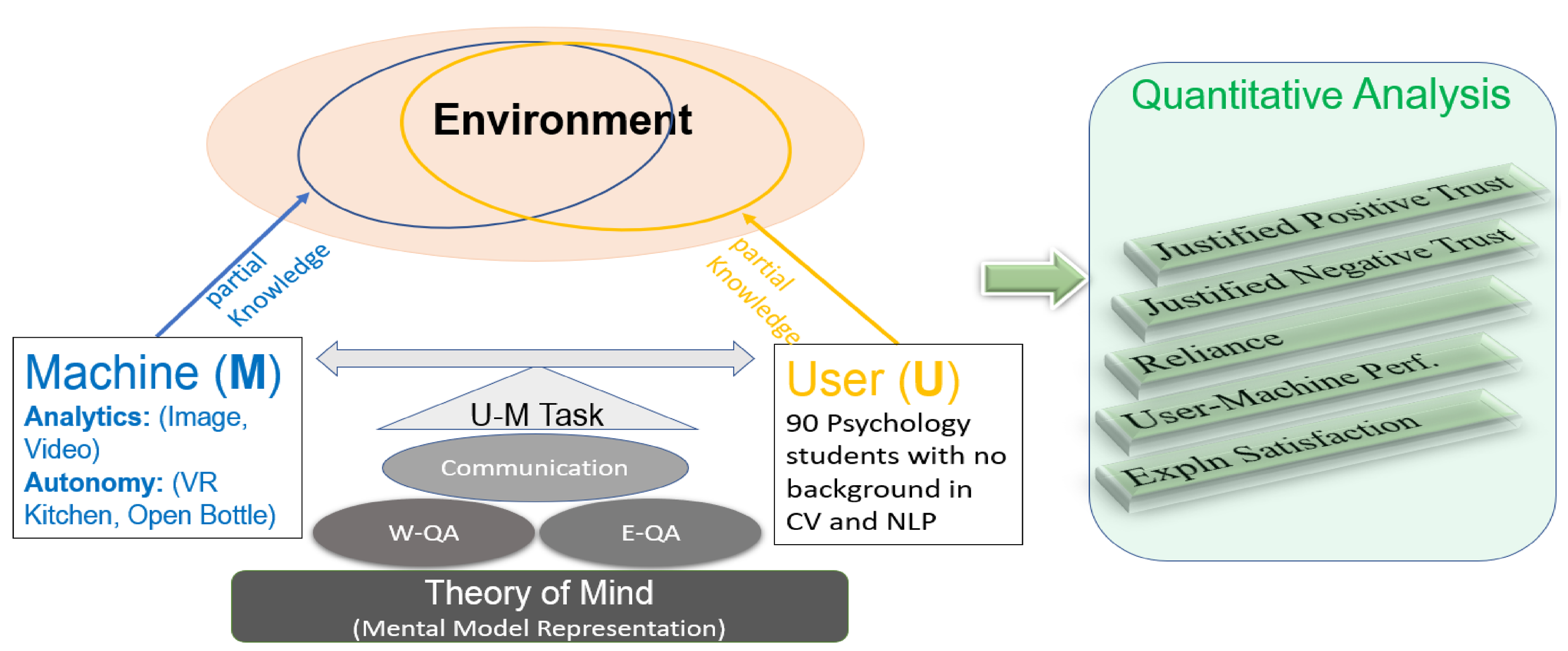
Learning and Communicating Explainable Representations for Analytics and Autonomy

Arjun Akula (UCLA), Sinisa Todorovic (OSU), Joyce Chai (MSU), Song-Chun Zhu (UCLA)

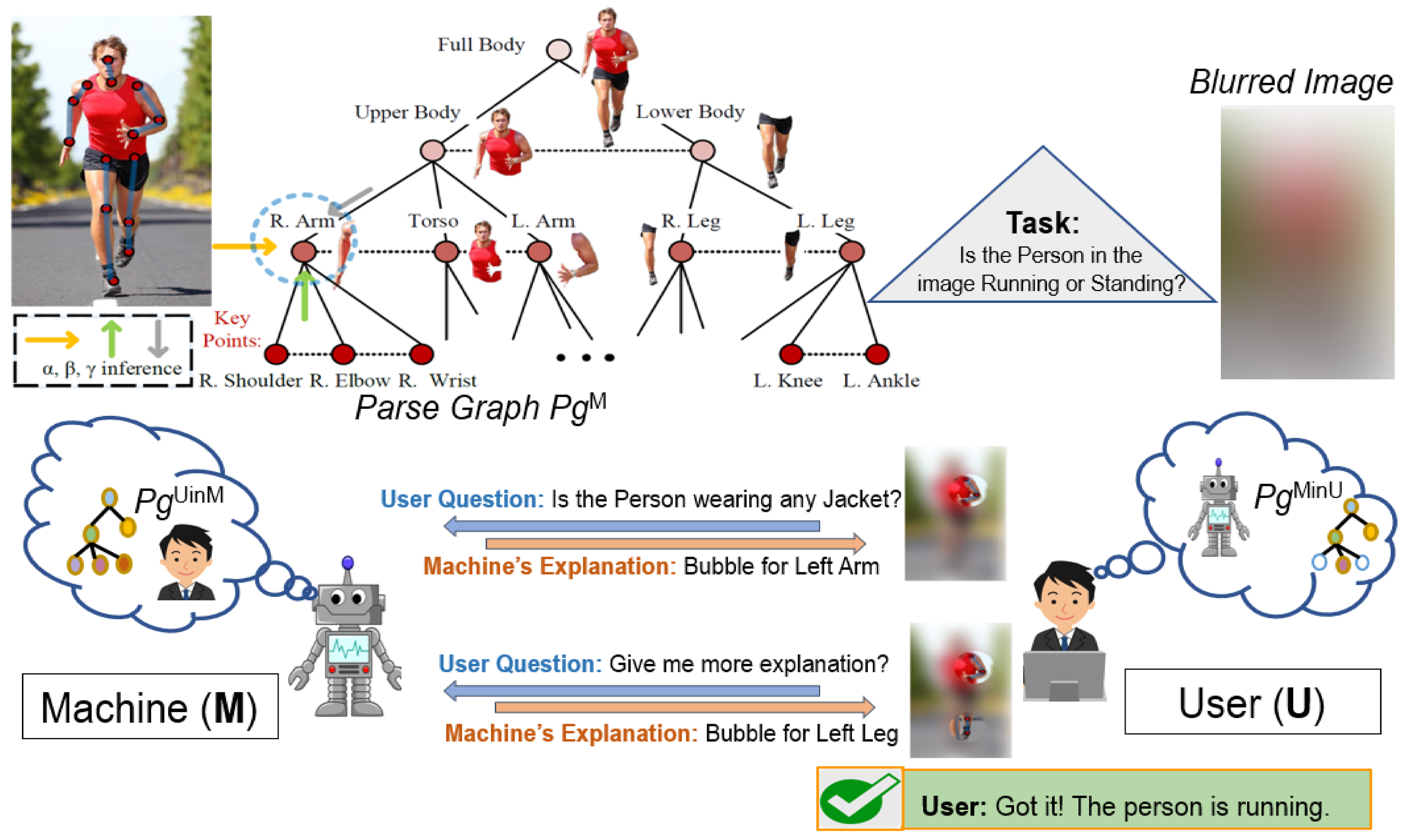
PI meeting, Feb 12-14
UC Berkeley

X-ToM: Explanation with Theory-of-Mind



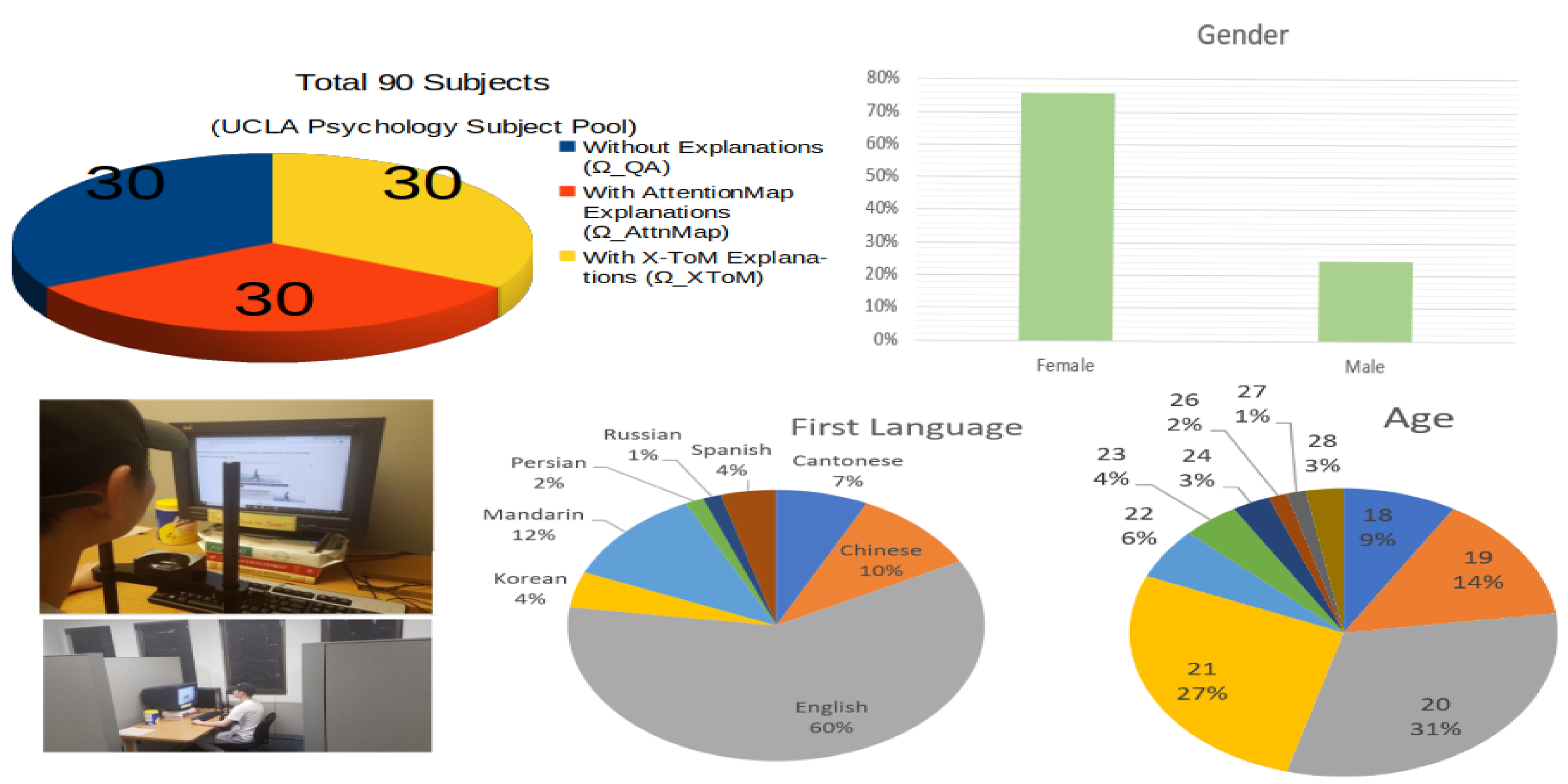
X-ToM for optimizing the dialog with a user towards estimating and increasing human trust.

XAI as Collaborative Task Solving

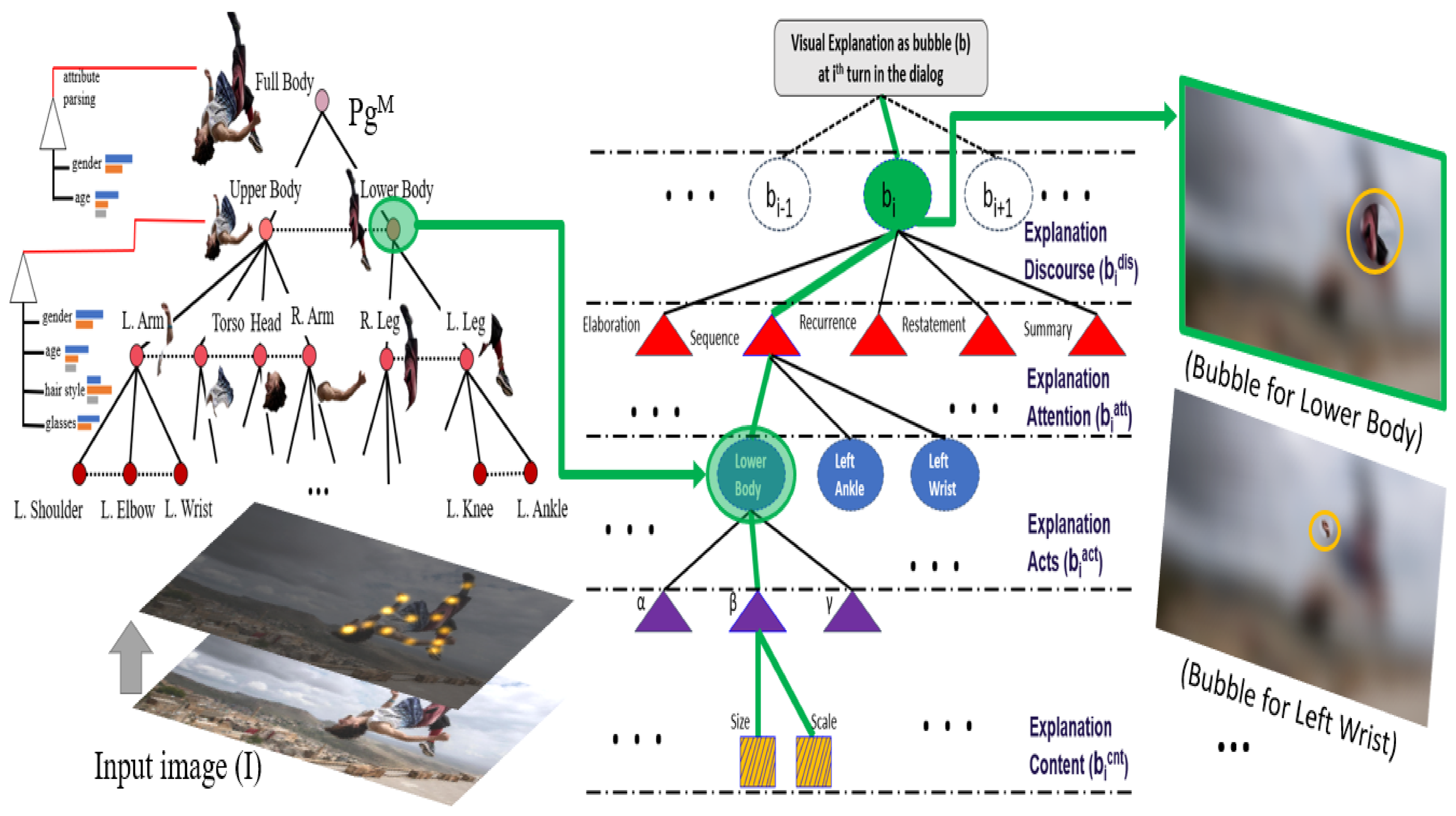


Through a dialog, we estimate Trust and Reliance in terms of pg^M, pg^{UinM} and pg^{MinU}.

Subject Pool

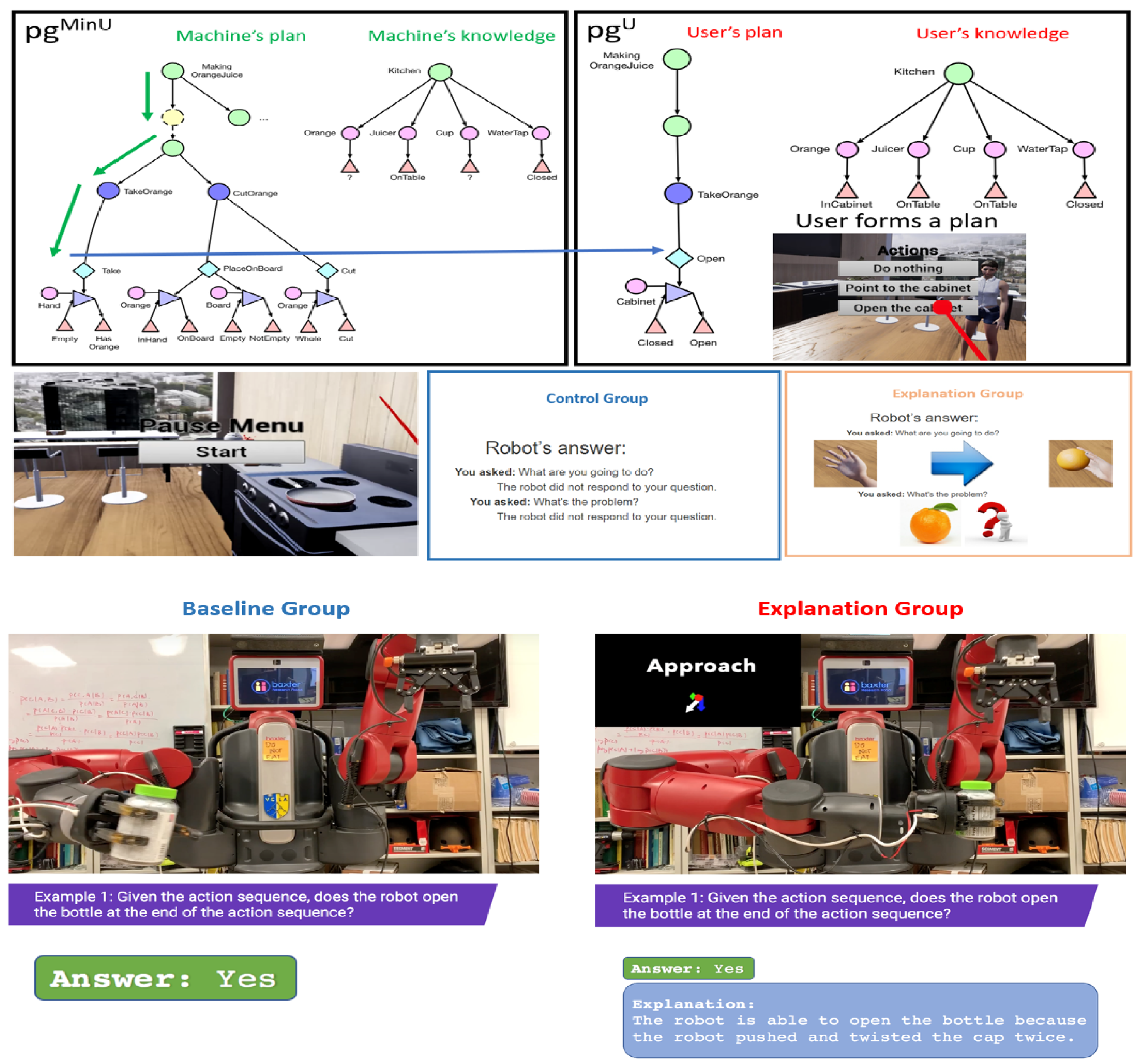


X-ToM for Image Analytics



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.

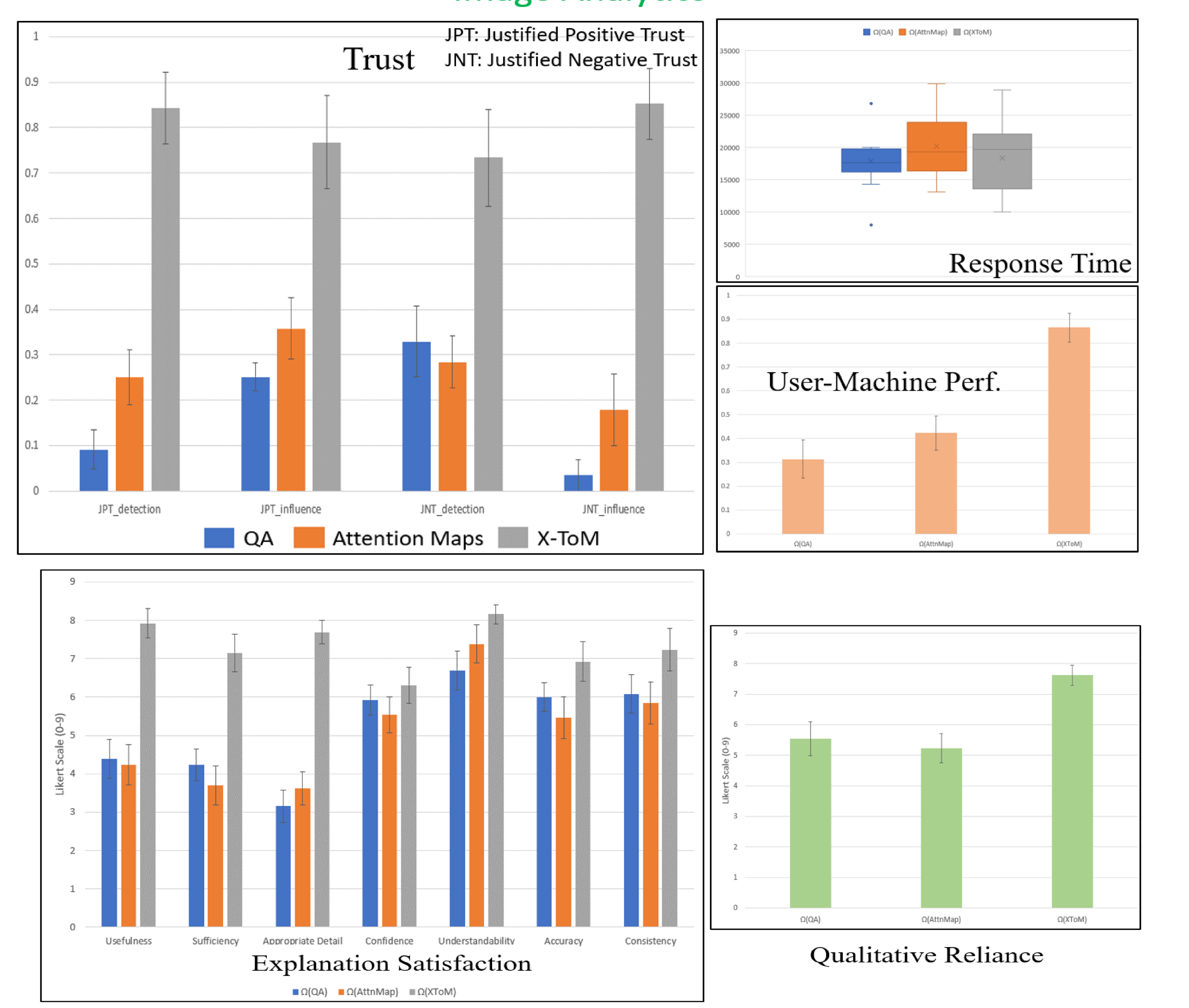
X-ToM for Autonomy

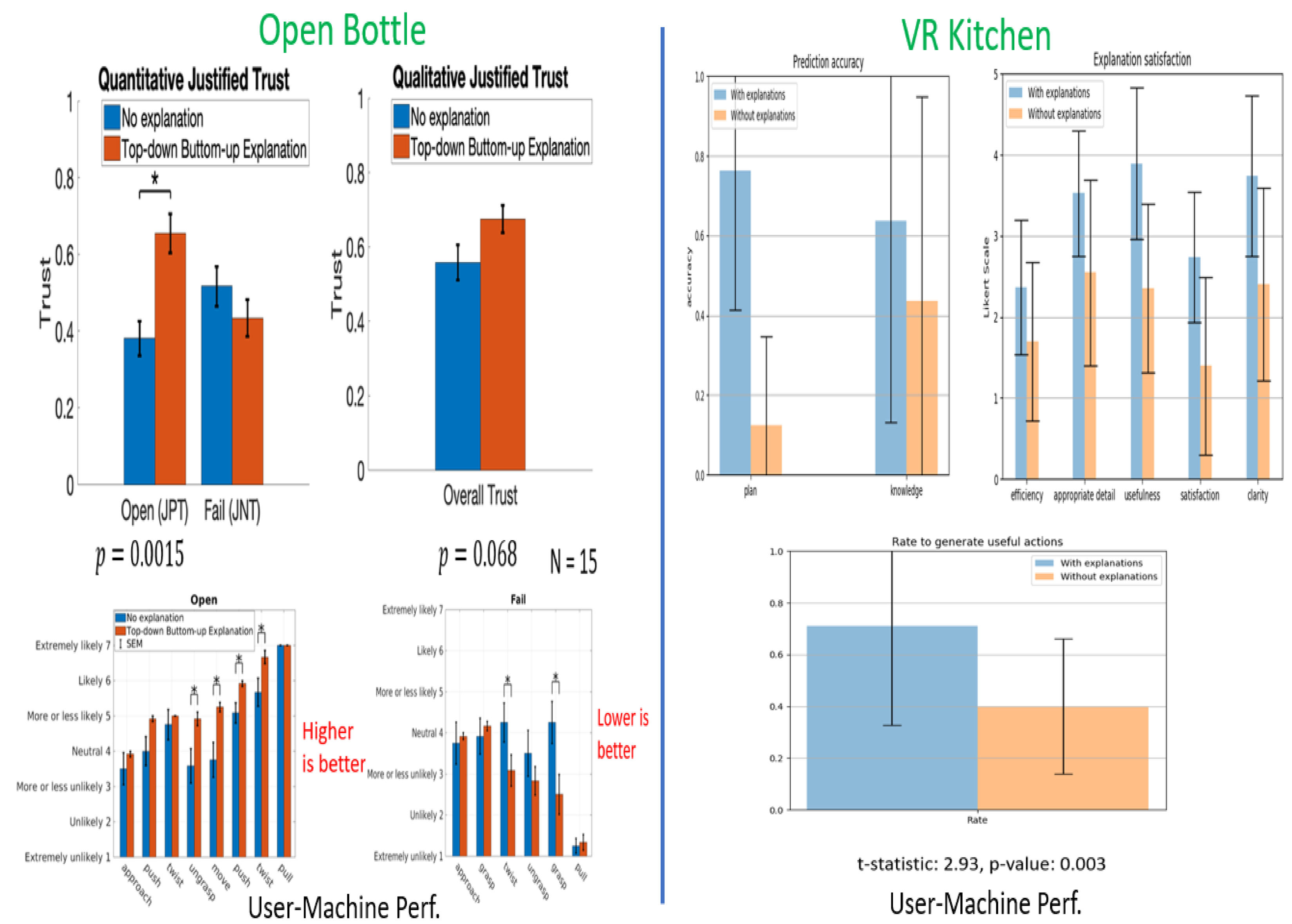


Both the Machine and the User solve common household tasks in VR Kitchen (e.g. `making orange juice`) and robot manipulation tasks (e.g. `opening a bottle`). The Machine provides explanations to the user based on its knowledge and the inferred human's mind.

Results of our Human Study

Image Analytics





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