SERUM: State Extraction and Refinement for User Modeling

A method for achieving meaningful user understanding from raw footage

Andy J. Phu (phu00003@umn.edu), Karin de Langis, James Mooney, Khanh Chi Le, Dongyeop Kang

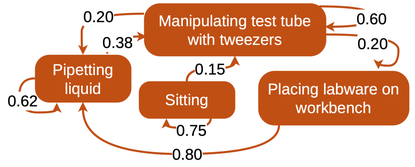
MINNESOTA NLP

UNIVERSITY OF MINNESOTA

https://minnesotanlp.github.io/SERUM-web

Current:

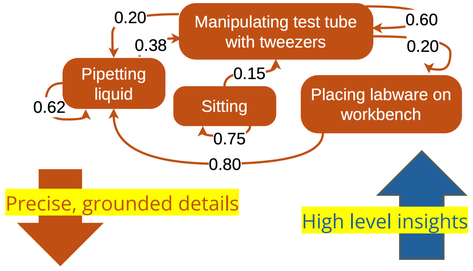
1. For every frame, in one pass
2. ask LLM to annotate what is going on



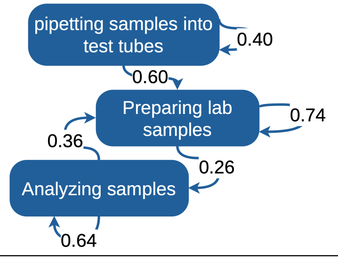
SERUM:

1. For every frame, over **multiple** passes
2. **And considering neighboring frames**
3. ask LLM what is going on and *why*

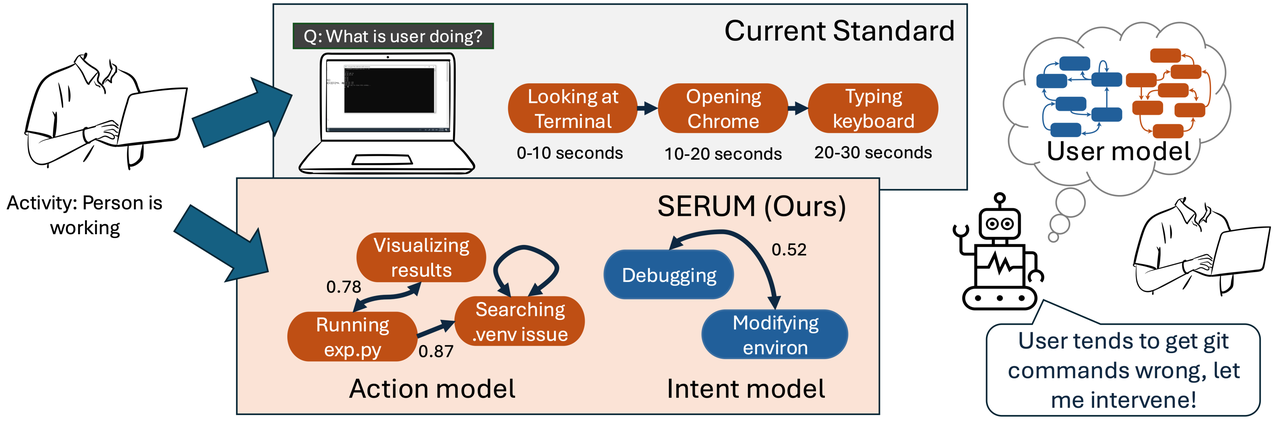
Pass 2i: Hey LLM, what is going on?



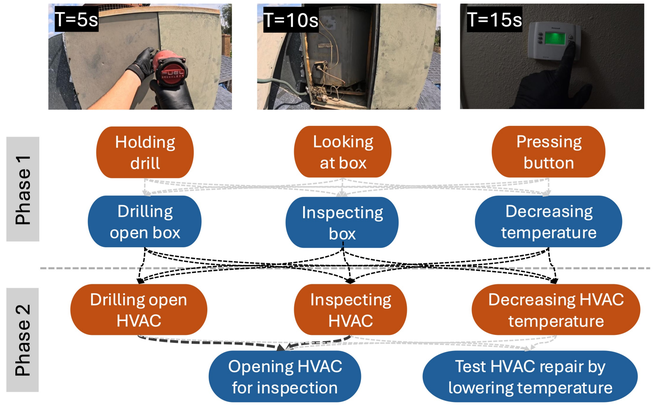
Pass 2i+1: Hey LLM, Why is user doing this?



Overview: SERUM watches what users do, and over multiple refinement passes asks "why"?

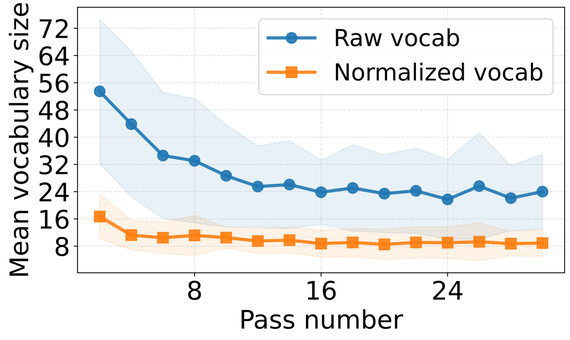


Neighbor Correction: Neighboring frame labels correct poorly supported/hallucinated labels

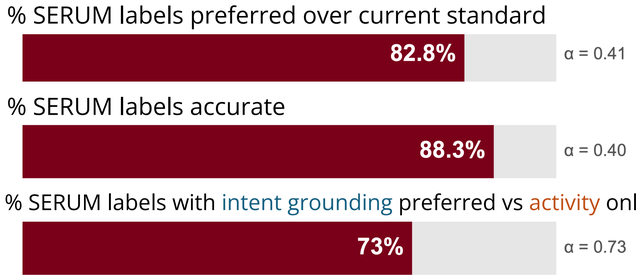


Schematic equilibrium: Neighbor correction and iterative refinement stabilize every video's vocabulary to a schematic equilibrium.

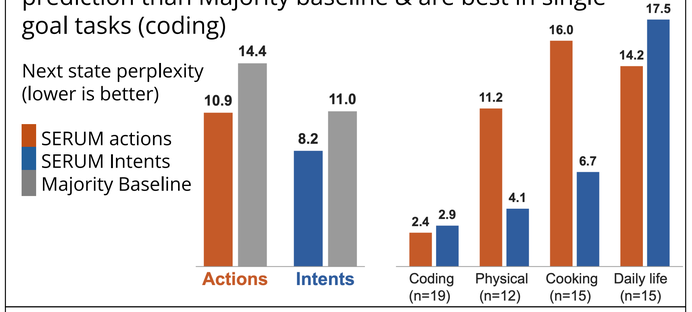
Normalized: semantically identical labels merged (−46%) for more meaningful data analysis.



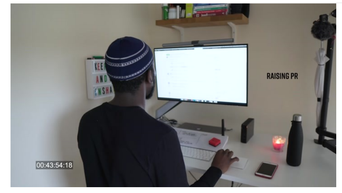
Humans prefer SERUM's labels



SERUM user models perform better: at next-state prediction than Majority baseline & are best in single goal tasks (coding)



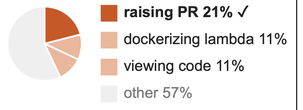
SERUM user models in action:



Now: raising PR

Actual next:
raising PR ✓

SERUM predicts



Majority baseline predicts

