Towards Embodied Agent Intent Explanation in Human-Robot Collaboration: ACT Error Analysis and Solution Conceptualization

Amanuel Ergogo

Zhao Han

UNIVERSITY of SOUTH FLORIDA

RARE LAB

May 19, 2025
HCRL-ICRA 2025 at ICRA 2025

Motivation & Problem

- Robots using learned policies (e.g., ACT) are **opaque (lack transparency)**
- Humans may struggle to predict robot actions to collaborate on the fly

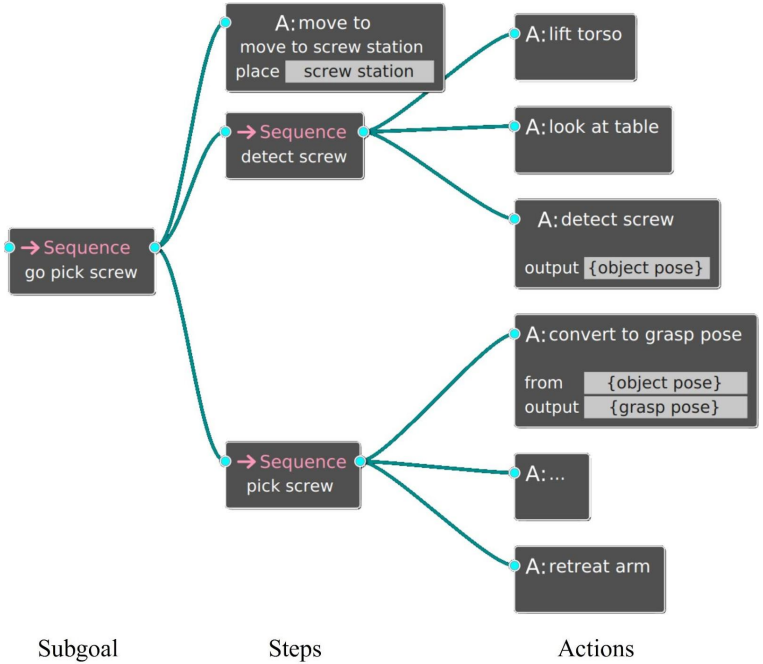
Approaches to Explainability

What methods exist for generating robot intent explanations?

Inherently Interpretable Methods (BTs, Graphs)

Simplify BT: From actions to goals

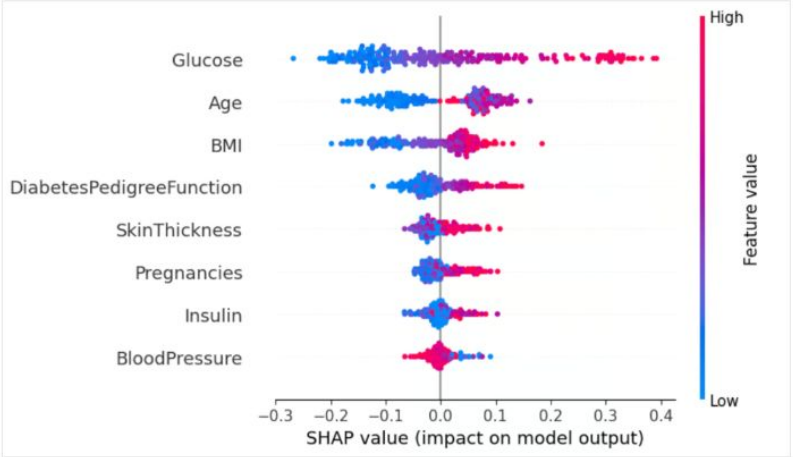
- Easy to justify behaviors
- Easy to generate hierarchical, concise explanations
- **Require hand-crafted logic**
- **Not applicable to learned policies**

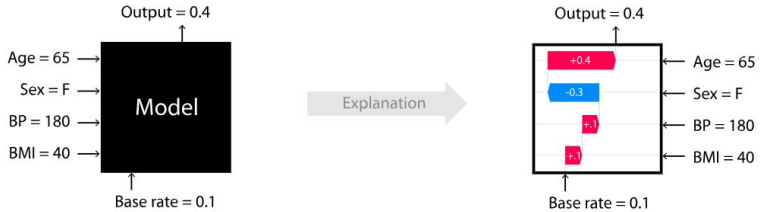


Post-Hoc XAI (e.g., Saliency, LIME, SHAP)

Post-hoc XAI (e.g., Saliency, LIME):

- Static or offline explanations
- **Require model access**
- **Not suitable for real-time explanation**





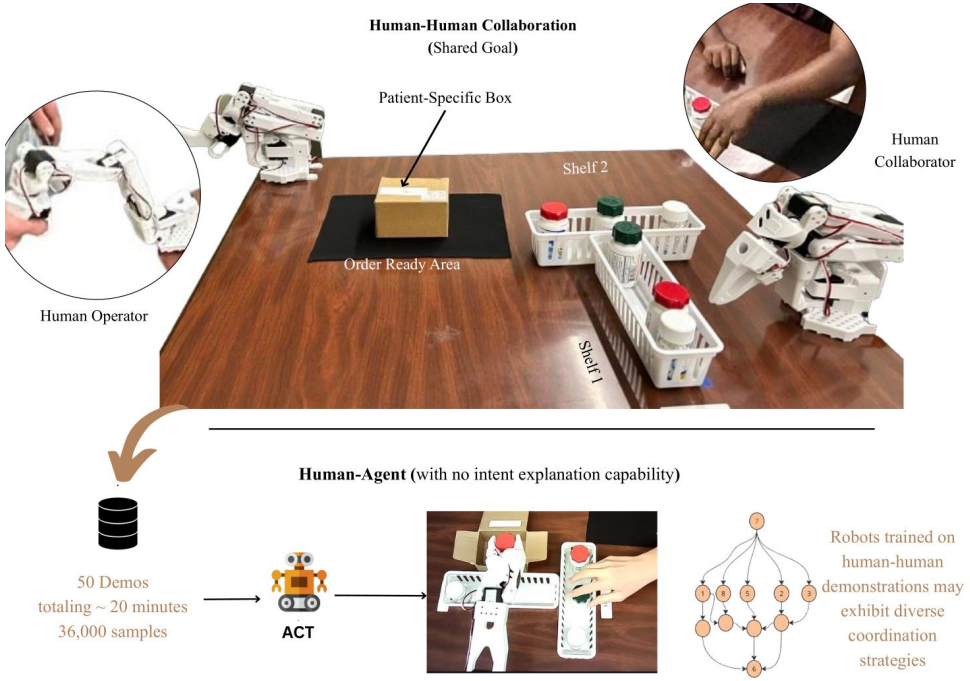
Lundberg, Scott M and Lee, Su-In— A Unified Approach to Interpreting Model Predictions

Empirical Study & CRIE: Conceptual Solution

1. Would a high-performing robot fail at teamwork?
2. How can we enable real-time, model-agnostic robot intent explanation without altering the policy?

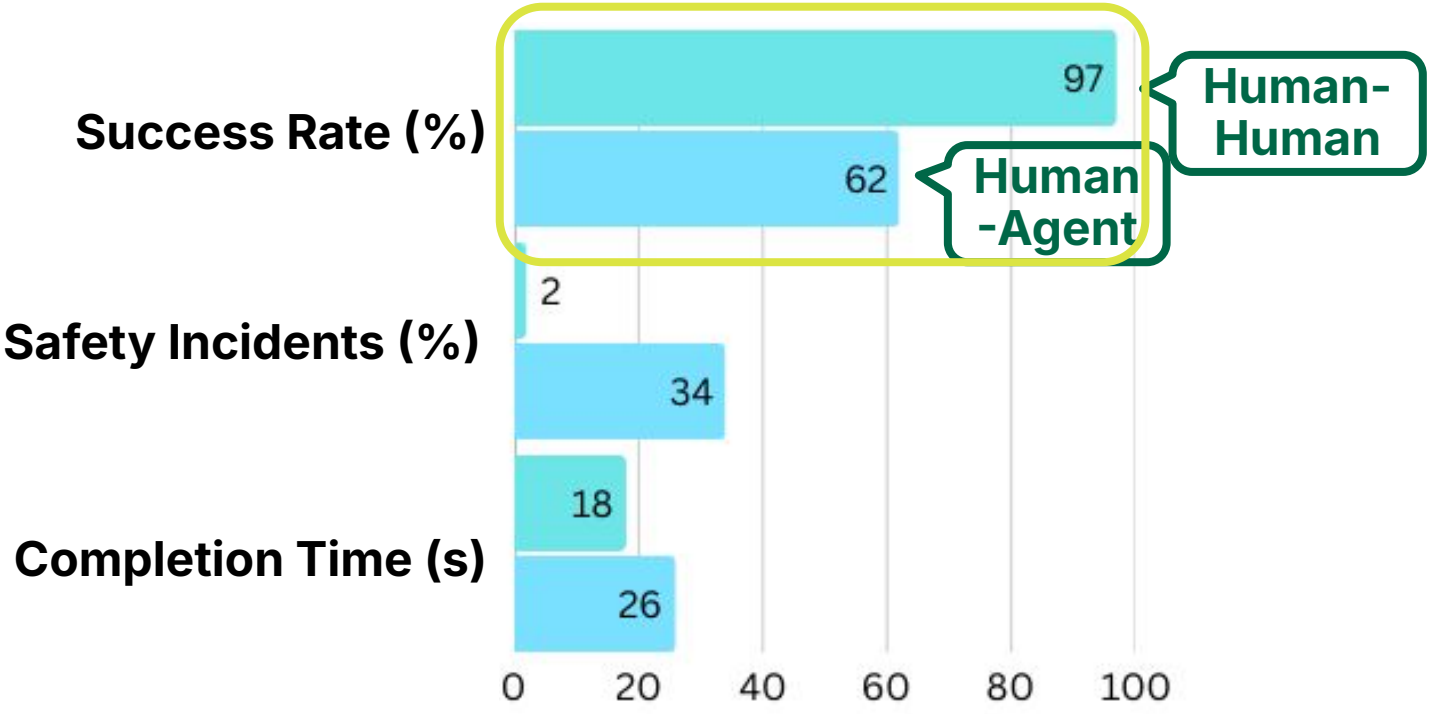
Empirical Analysis: ACT in Medication Dispensing

- Medication-dispensing task: fulfill a shared order
- Conditions:
  - **Human-Human** (baseline)
  - **Human-Agent** (ACT-controlled robot, no explanation)
- Evaluate how well ACT supports coordination without intent explanation



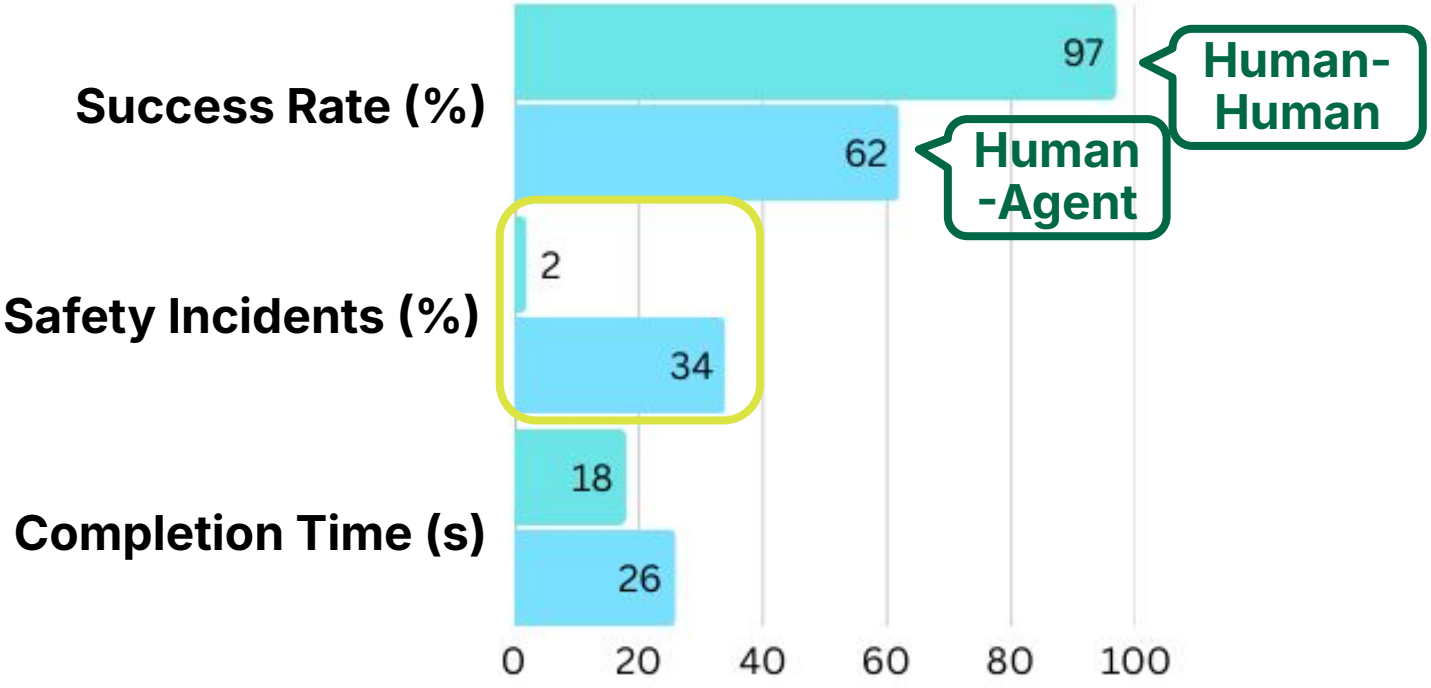
Results — Human-Human vs. Human-Agent

Teamwork Performance Comparison (15 matched trials):



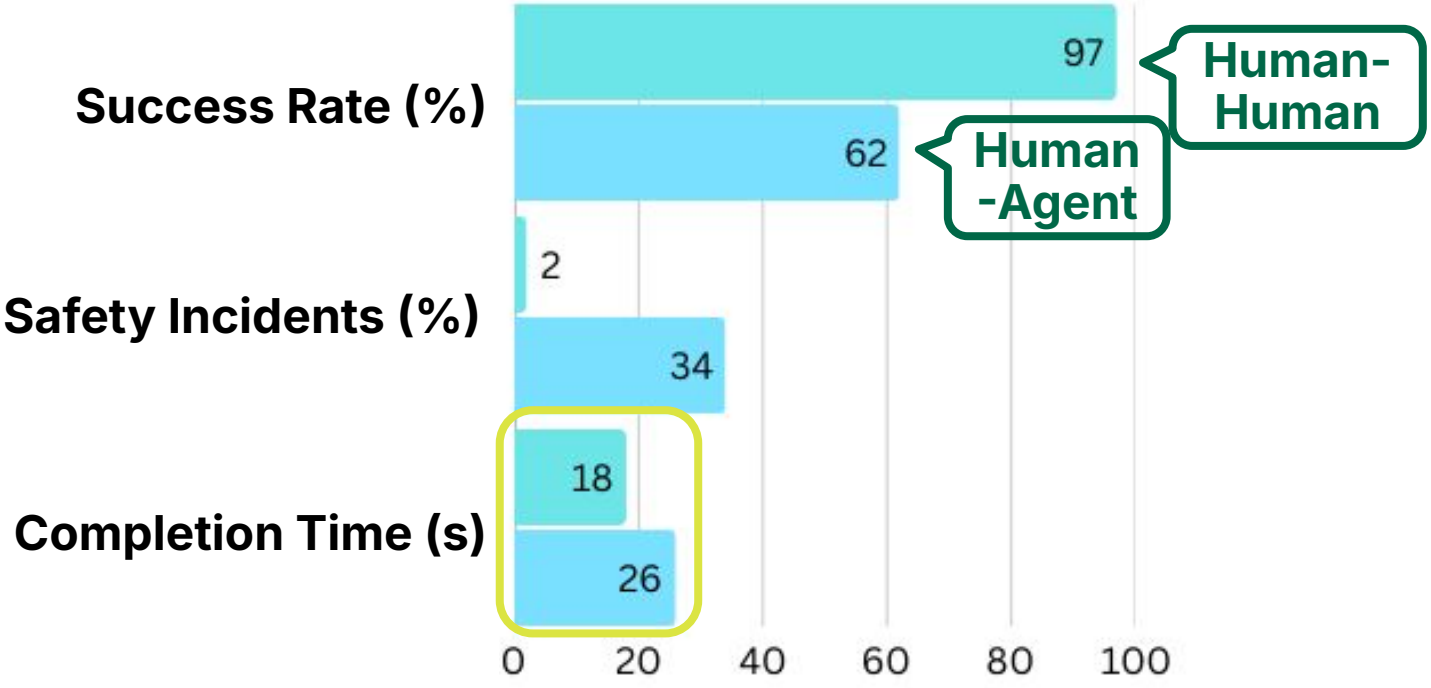
Results — Human-Human vs. Human-Agent

Teamwork Performance Comparison (15 matched trials):

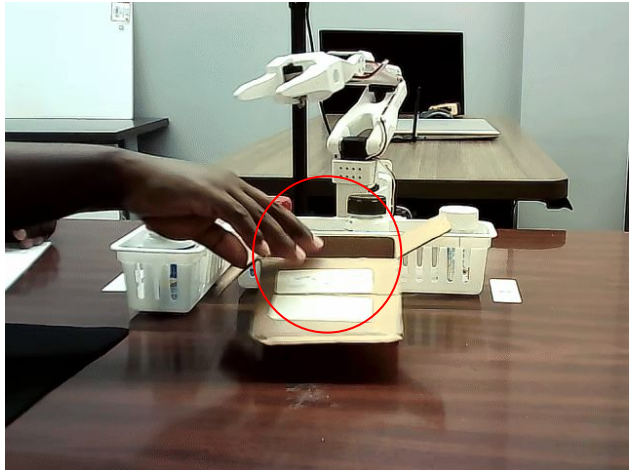


Results — Human-Human vs. Human-Agent

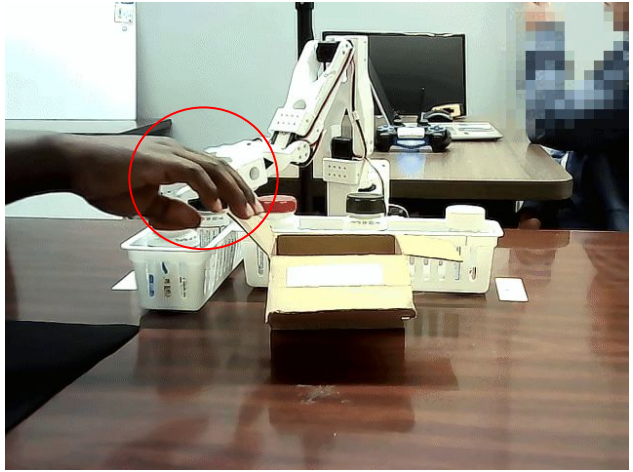
Teamwork Performance Comparison (15 matched trials):



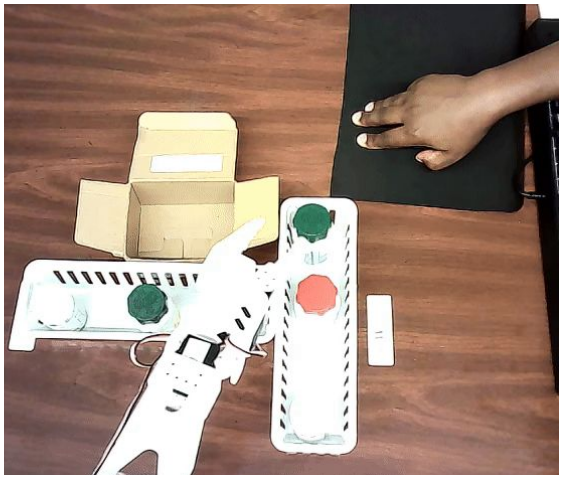
Common Failure Modes



Redundant Retrievals



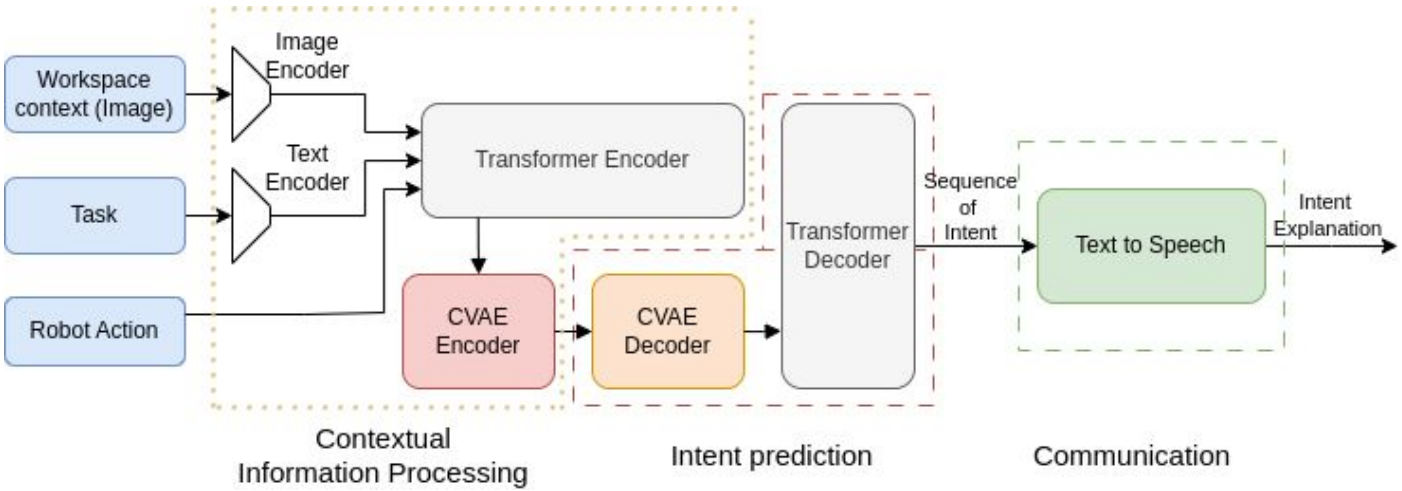
Safety Conflicts



Delays/Hesitation

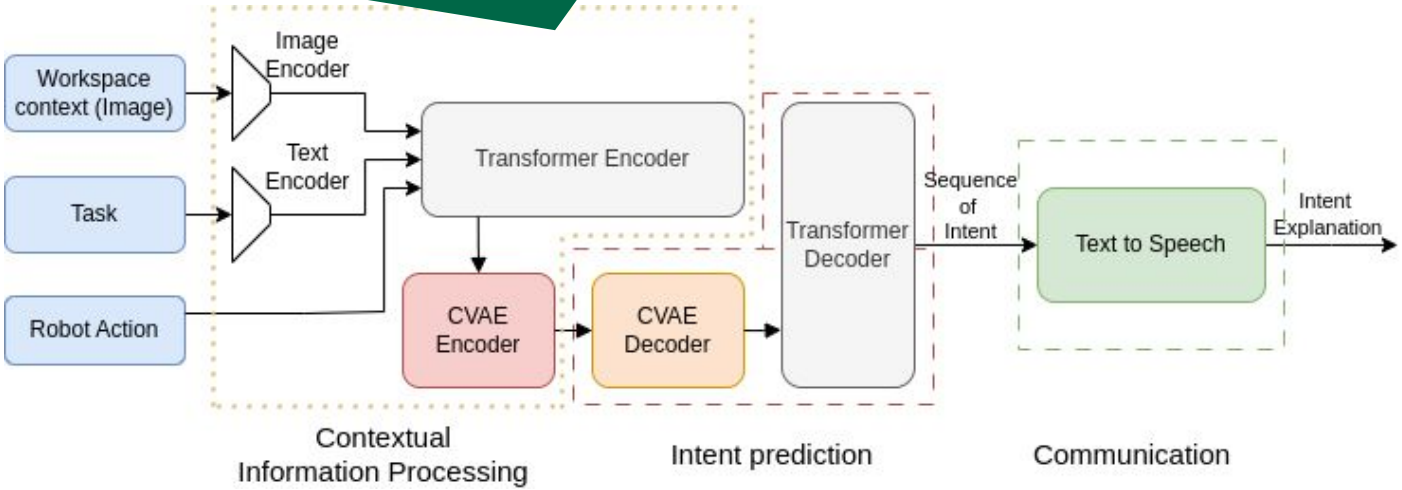
Contextual Robot Intent Explanation (CRIE) System Architecture

Encodes actions, context, goals & progression

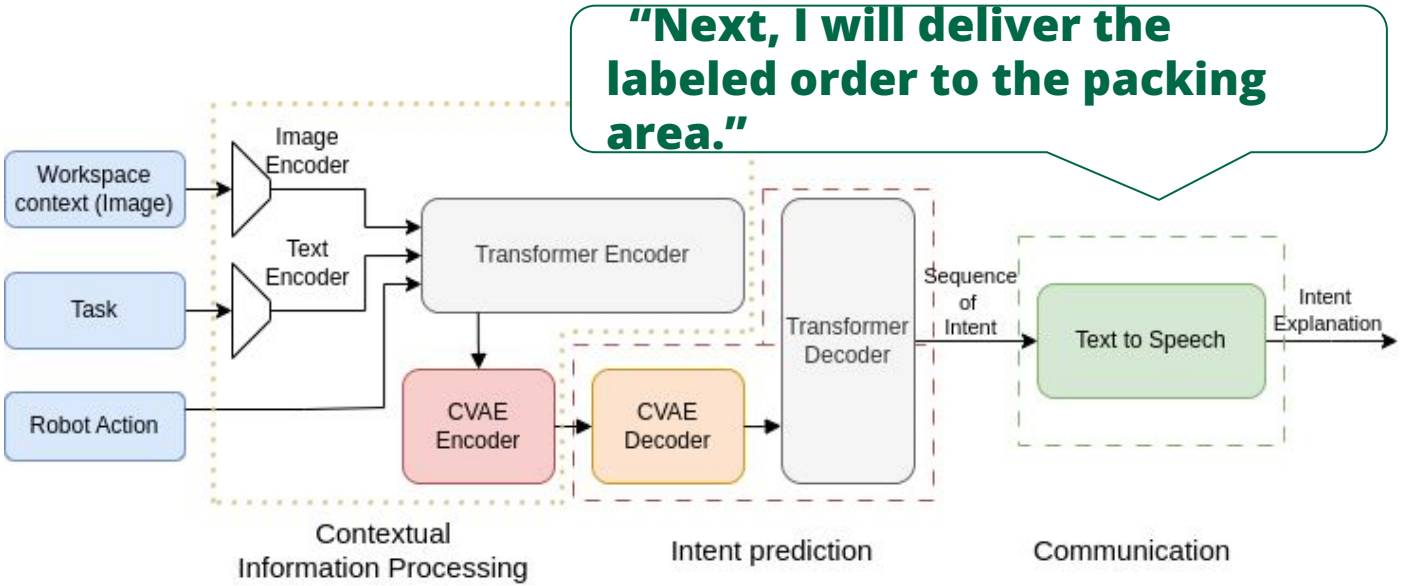


Contextual Robot Intent Explanation (CRIE) System Archit

Uses Transformer and CVAE to process contextual inputs into a latent representation of subtask intent and decode it into a symbolic subtask labels



Contextual Robot Intent Explanation (CRIE) System Architecture



Towards Embodied Agent Intent Explanation in Human-Robot Collaboration: ACT Error Analysis and Solution Conceptualization

Amanuel Ergogo

Zhao Han

UNIVERSITY of SOUTH FLORIDA

RARE LAB

therarelab.com

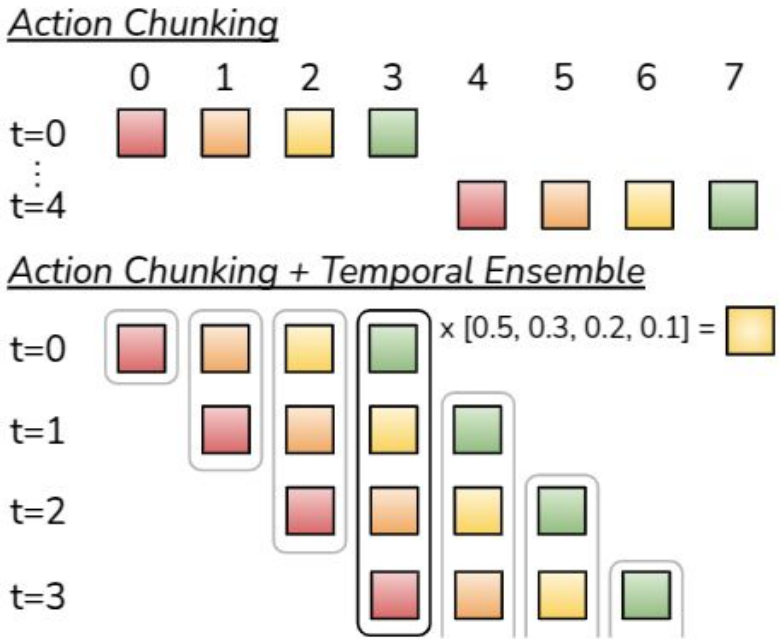
Key Takeaways

1. State-of-art robot policies limit coordination and safety during collaboration
2. Transparent robot intent is essential for teamwork
3. CRIE will enable real-time & policy-agnostic intent explanations for fluent collaborations

Built-in Structure (ACT)

- Chunking actions → supports short-horizon intent prediction

- **Policy-specific**
- **No symbolic subtask labels or natural language**



Zhao, Kumar, Levine, Finn Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware