Asian Deans' Forum 2026

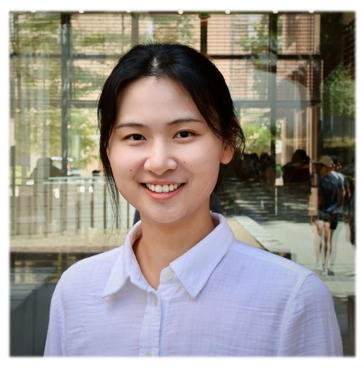
THE RISING STARS

Women in Engineering Workshop

SEP 13 – 15 | HKUST, Hong Kong

Shenghui (Vivian) Chen

CENTER FOR autonomy | TEXAS The University of Texas at Austin



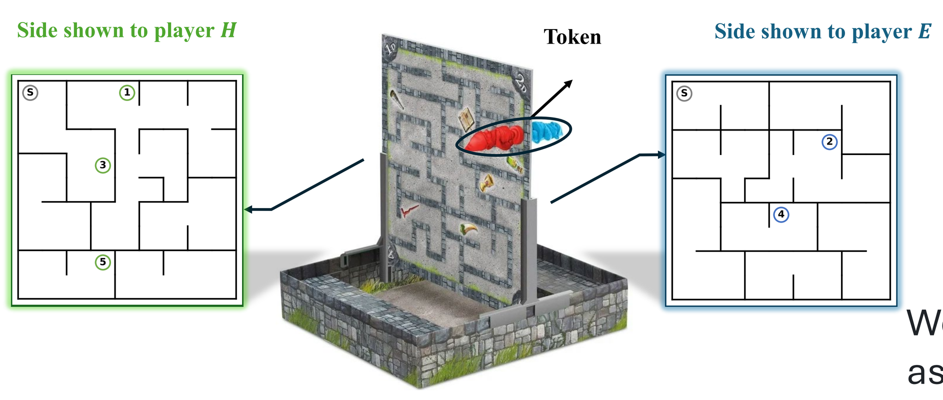
Human-Agent Collaboration | Human-Centered AI | Evaluation

How to enable AI agents to collaborate adaptively with humans under incomplete information?

Coordinate under Incomplete Information

Testbed: Gnomes at Night

Our testbed, Gnomes at Night 🎮, is a maze game where two players alternately control a token to achieve a common goal with incomplete information.



Shared-Control Game

A shared-control game where player E and player H **collectively controls** a token in **alternating turns**:

$\Gamma = (\mathcal{S}, s_{init}, s_{final}, \mathcal{A}^E, \mathcal{A}^H, \mathcal{T}^E, \mathcal{T}^H)$ where

$\mathcal{S}$ is a finite set of states,

s_{init} is an initial state, s_{final} is a final state,

$\mathcal{A}^i$ is a finite set of actions of player $i \in \{E, H\}$,

$\mathcal{T}^i: \mathcal{S} \times \mathcal{A}^i \to \mathcal{S}$ is a **private** deterministic transition function of each player $i \in \{E, H\}$.

We augment Γ with a **common** reward function that assigns equal reward to both : $\mathcal{R}: \mathcal{S} \times \cup_i \mathcal{A}^i \to \mathbb{R}$.

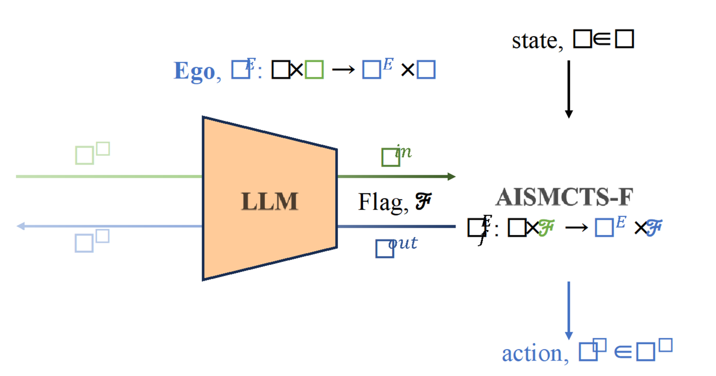
Communication-based Approach

Human-Agent Cooperation in Games under Incomplete Information through Natural Language Communication. IJCAI 2024.

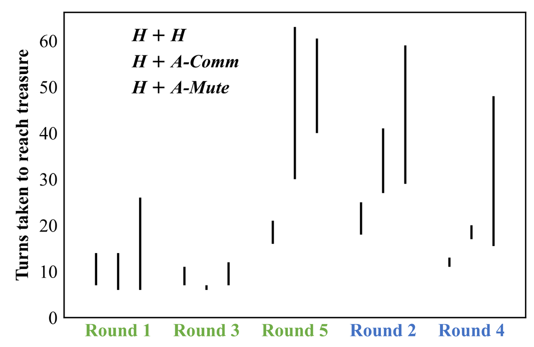
Language module: translates natural language messages into and from a finite set of *flags*, a compact representation defined to capture player intents. (We implement using few shot-prompting with gpt-3.5-turbo.)

Planning module: leverages these flags to compute a policy using an algorithm called ***AISMCTS-F*** that we introduce.

Flags: $\mathcal{F} = \mathcal{A}^H \cup \mathcal{A}^E \cup$ {Accept, Reject, Inquiry, None}

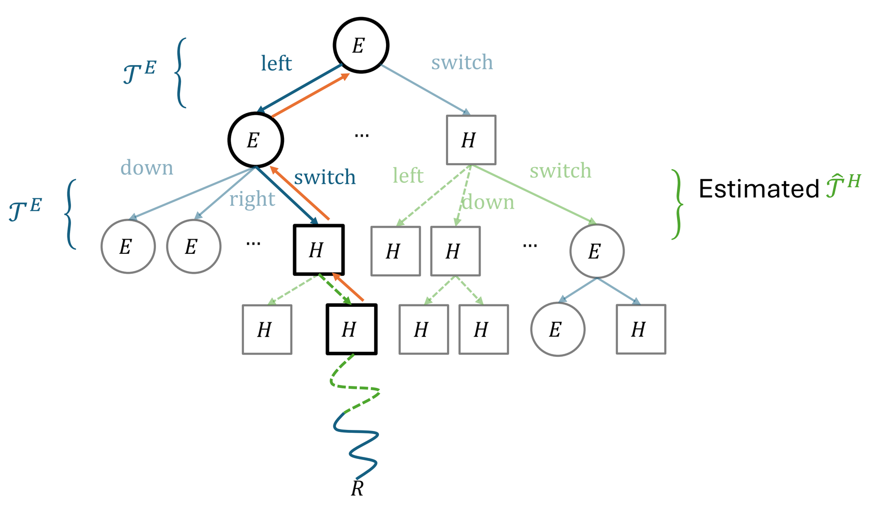


User study. With 150 Prolific participants (mean age = 34.71; 53% female), communication-enabled agents outperformed mute agents in 4 of 5 rounds, requiring significantly fewer turns (ANOVA: F(2,627) = 10.53, p < .01; Tukey HSD: p < .05)

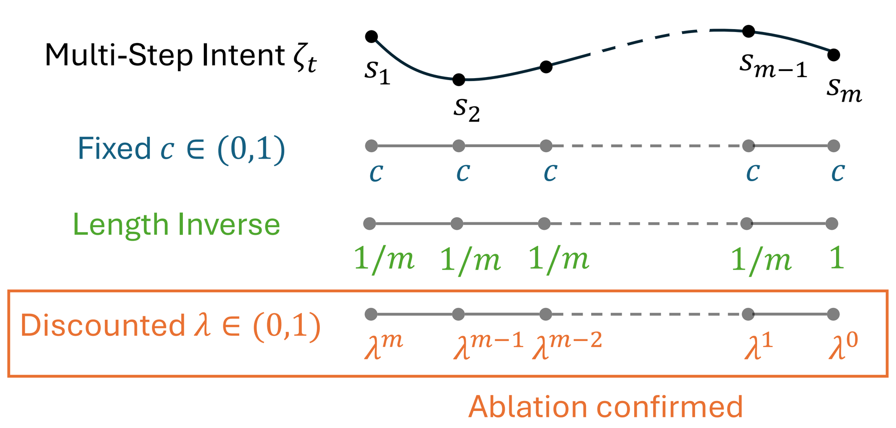


IntentMCTS

Human-Agent Coordination in Games under Incomplete Information via Multi-Step Intent. AAMAS 2025.



Incorporate multi-step intent via reward augmentation

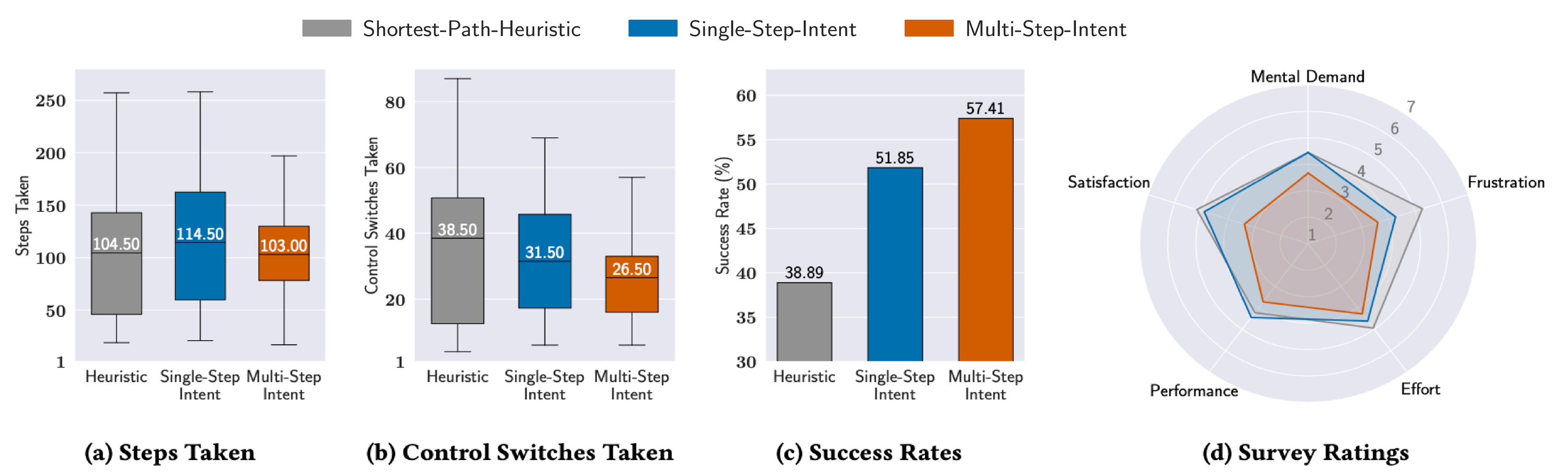


Rollout with current best knowledge (known $\mathcal{T}^E$ & estimated $\hat{\mathcal{T}}^H$ transitions)

From leaf to root: $N(v) \leftarrow N(v) + 1$, $Q(v) \leftarrow Q(v) + q_{sample}$, $q_{sample} = r + \gamma\left[\delta' q'_{sample} + (1-\delta')\frac{Q(v)}{N(v)}\right]$

Step reward; Transition feasible: future return; Transition infeasible: Current value estimate

User study. With 18 graduate students (mean age = 26.28), Multi-Step-Intent achieved fewer steps and control switches, higher success rates, and lower cognitive-load ratings than both baselines.

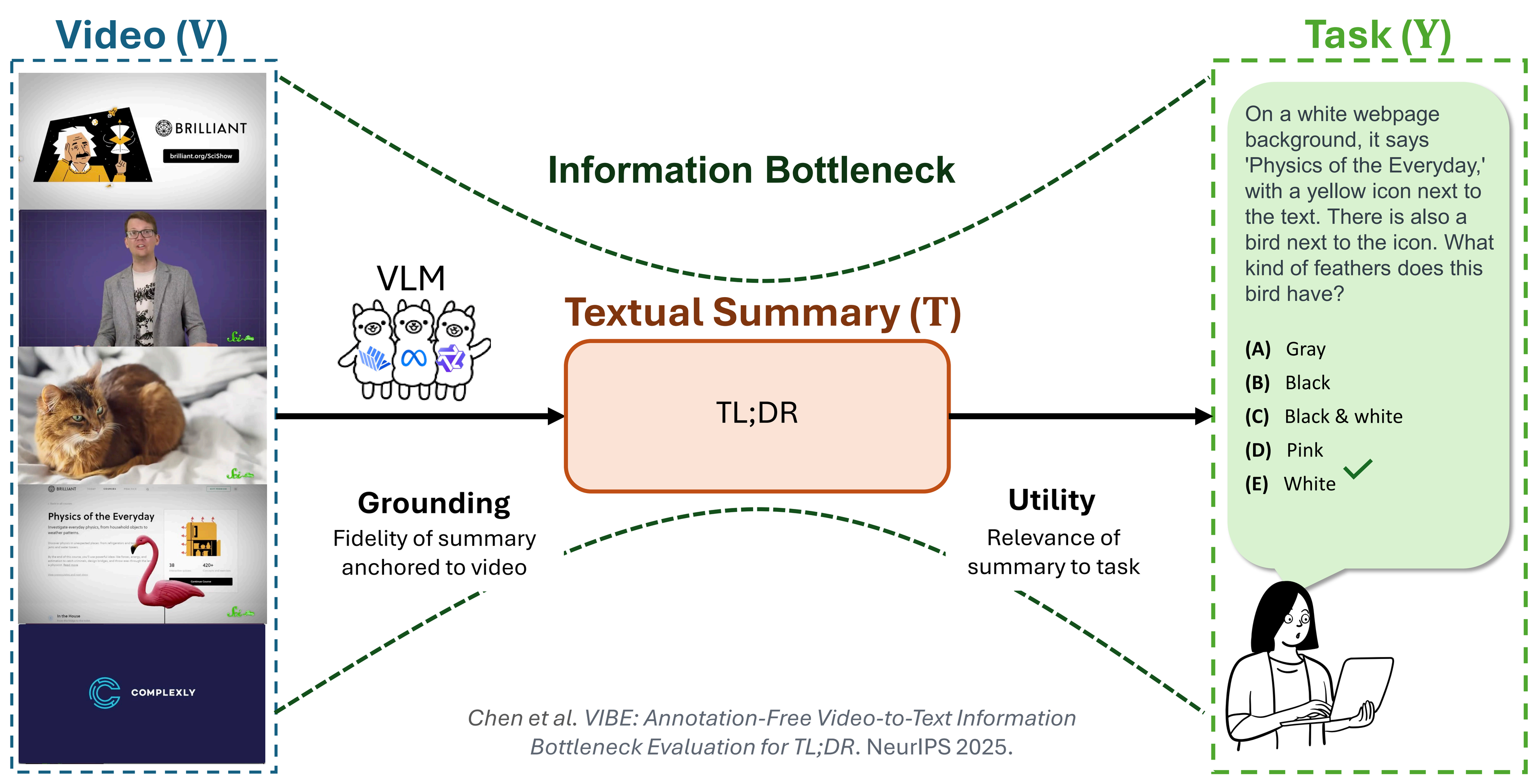


(a) Steps Taken (b) Control Switches Taken (c) Success Rates (d) Survey Ratings

Information-theoretic Multimodal Evaluation

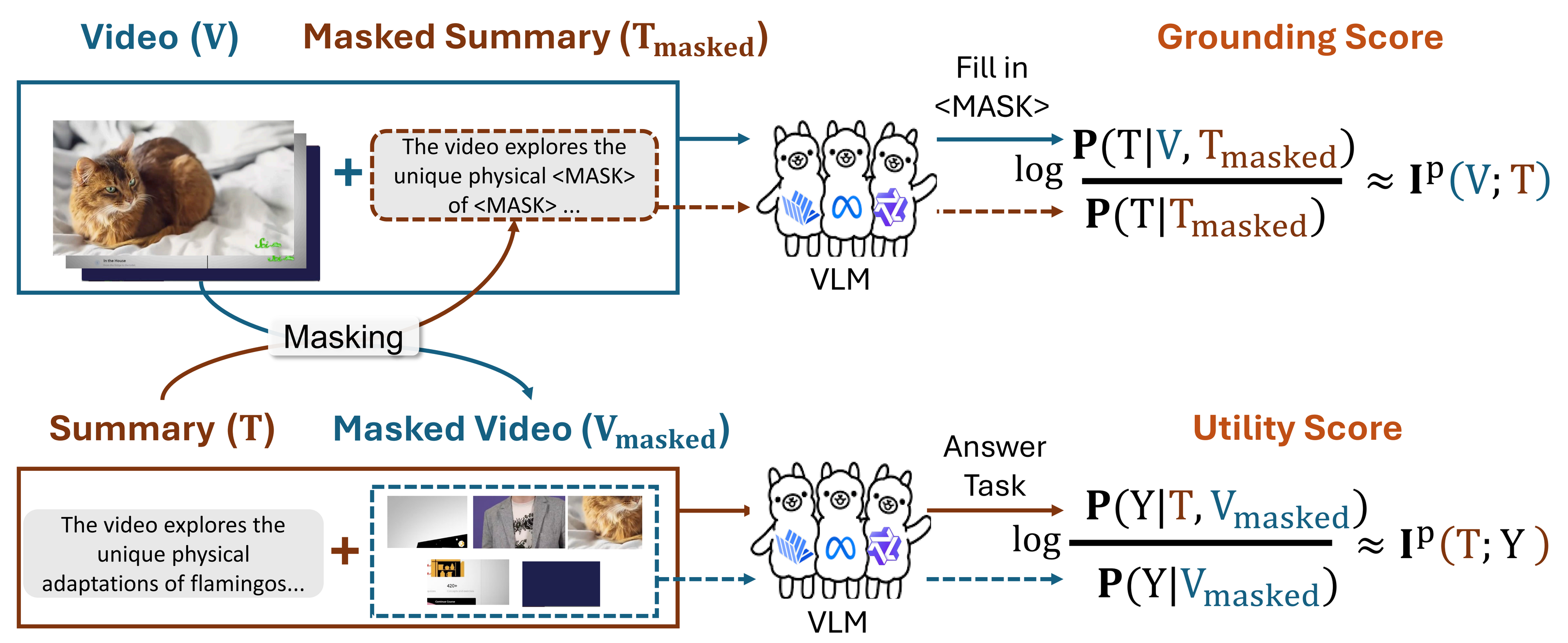
Problem: Manual video review slows decision-making in domains like traffic monitoring, where human oversight remains essential. Yet VLM-generated summaries are verbose and poorly aligned with downstream tasks, reducing human efficiency.

Challenge: Reference-based evaluations (*e.g.*, ROUGE and BLEU) require **costly human annotations** and **focus on text similarities**, limiting scalability and grounding in visual information.



Chen et al. VIBE: Annotation-Free Video-to-Text Information Bottleneck Evaluation for TL;DR. NeurIPS 2025.

VIBE applies **Information Bottleneck** principles to evaluate textual summaries on grounding (video-summary alignment) and utility (summary-task relevance). It computes scores using multimodal masking with pointwise mutual information.



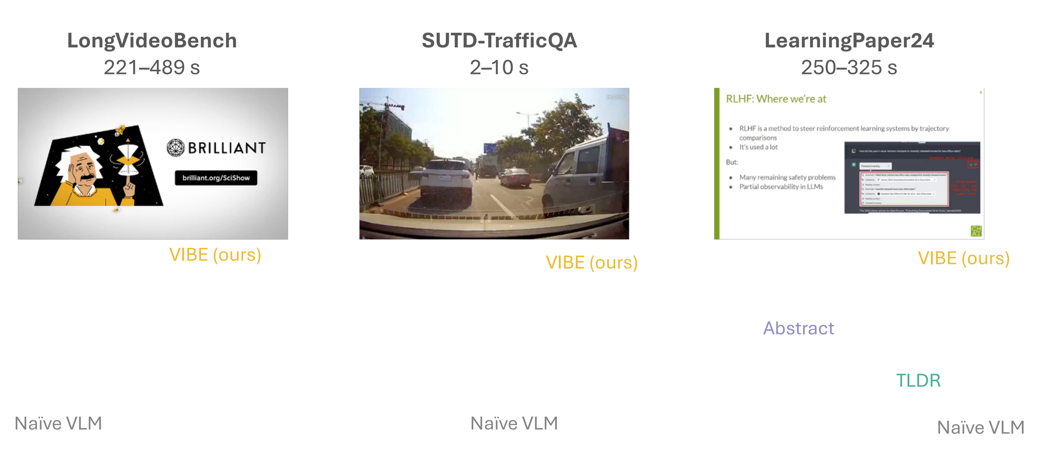
$$\log \frac{P(T|V, T_{masked})}{P(T|T_{masked})} \approx I^P(V; T)$$

$$\log \frac{P(Y|T, V_{masked})}{P(Y|V_{masked})} \approx I^P(T; Y)$$

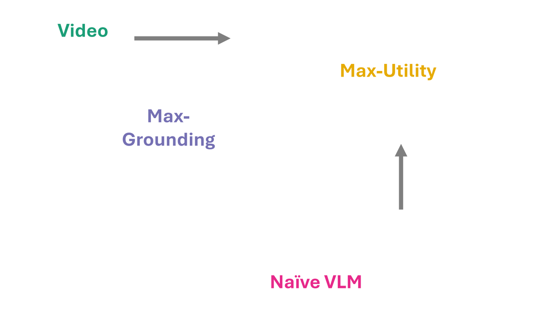
Rejection Sampling: Selects summaries that maximize a weighted sum between grounding and utility scores as optimal captions for humans.

Pareto frontier across datasets

reveals tradeoff between grounding and utility



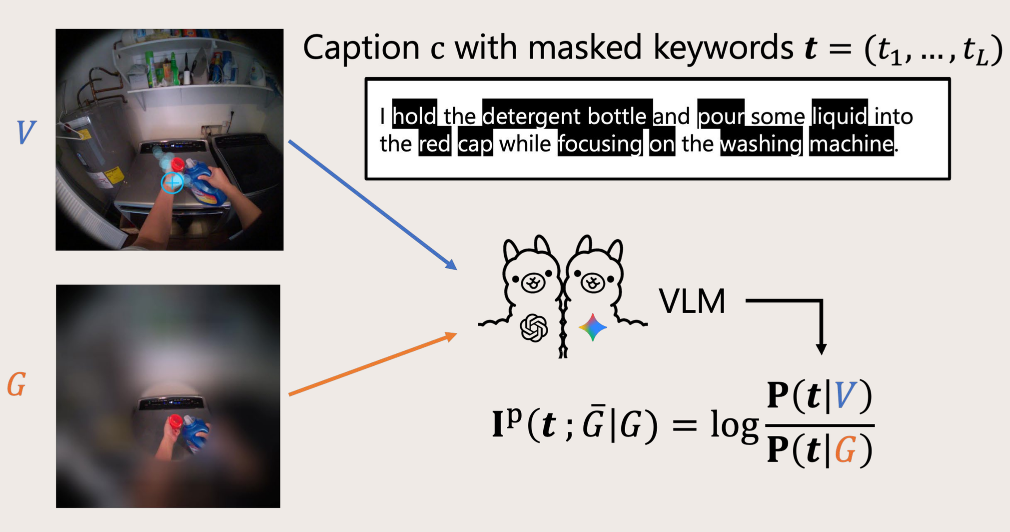
User study



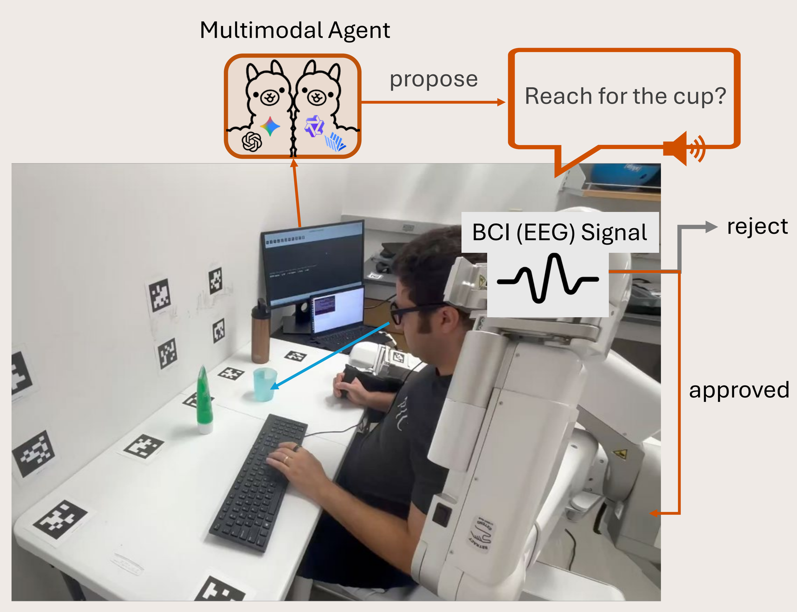
Max-Utility and Max-Grounding summaries are **more accurate** than naive VLM outputs, with **less response time** than raw video.

Ongoing and Future Directions

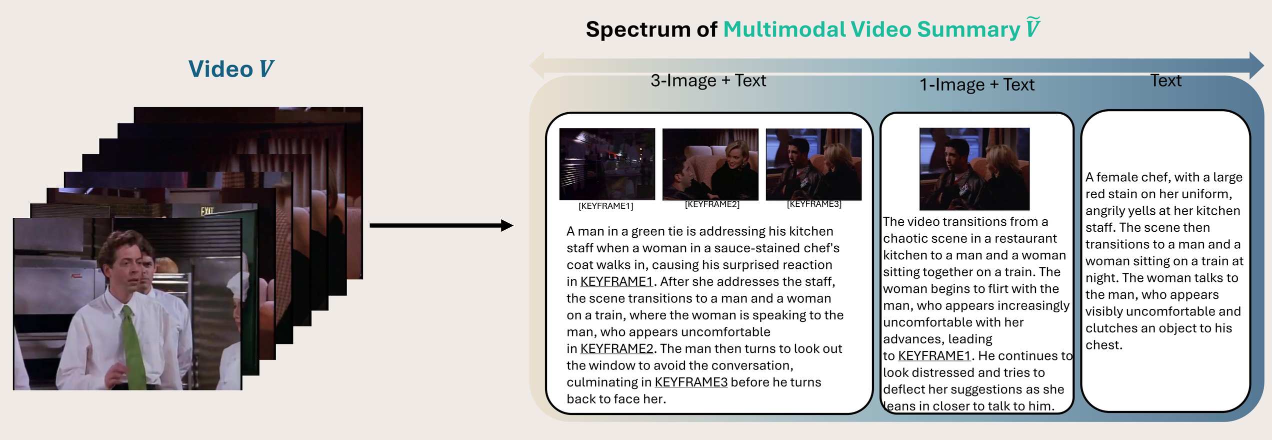
To bring human *behavioral and physiological signals*, such as gaze and BCI, into more *embodied* interaction loop.



VEGAS: Human-Aligned Video Caption Evaluation via Gaze. TMLR (under review).



BCI+Gaze for Assistive Exoskeleton Control



Chen et al. ViSIL: Unified Evaluation of Information Loss in Multimodal Video Captioning. NeurIPS 2026 (under review).

Organizers

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY | UNSW SYDNEY | SEOUL NATIONAL UNIVERSITY | NUS National University of Singapore | 臺灣大學 NTU | UTokyo | 清华大学 Tsinghua University