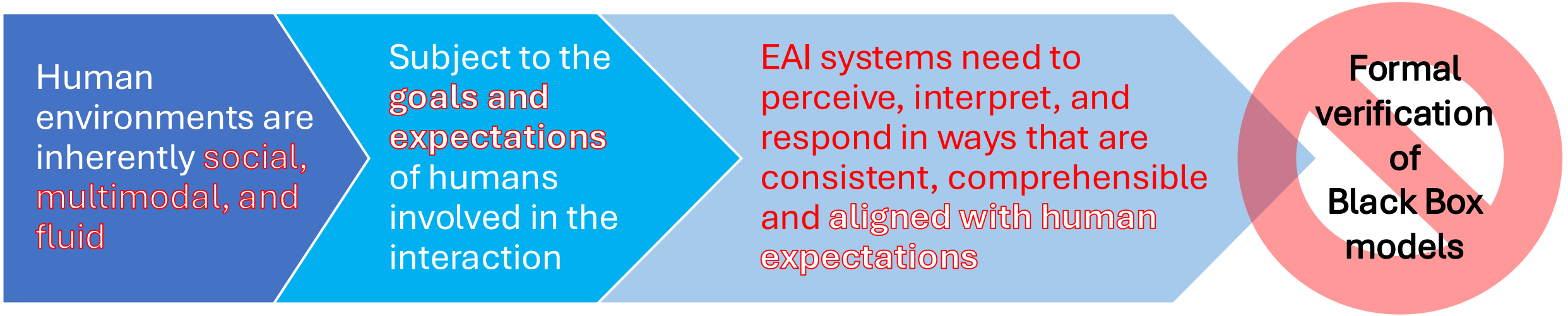
Explicit World Models for Reliable Human-Robot Collaboration

Kenneth Kwok[1], Basura Fernando[1], Qianli Xu[2],
Vigneshwaran Subbaraju [1], Dongkyu Choi[3], Boon Kiat Quek[1]

[1]Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore
[2]Institute for Infocomm Research (I^{2}R), Agency for Science, Technology and Research, Singapore
[3]A-FAB Technology Team, Mechatronics Research, Samsung Electronics, Korea
kenkwok@a-star.edu.sg,fernando_basura@a-star.edu.sg, xu_qianli@a-star.edu.sg,
vigneshwaran_subbaraju@a-star.edu.sg, edc.choi@samsung.com, quekbk@a-star.edu.sg

Reliable HRI with Embodied AI

Reliability in HRI is **contextually determined**:

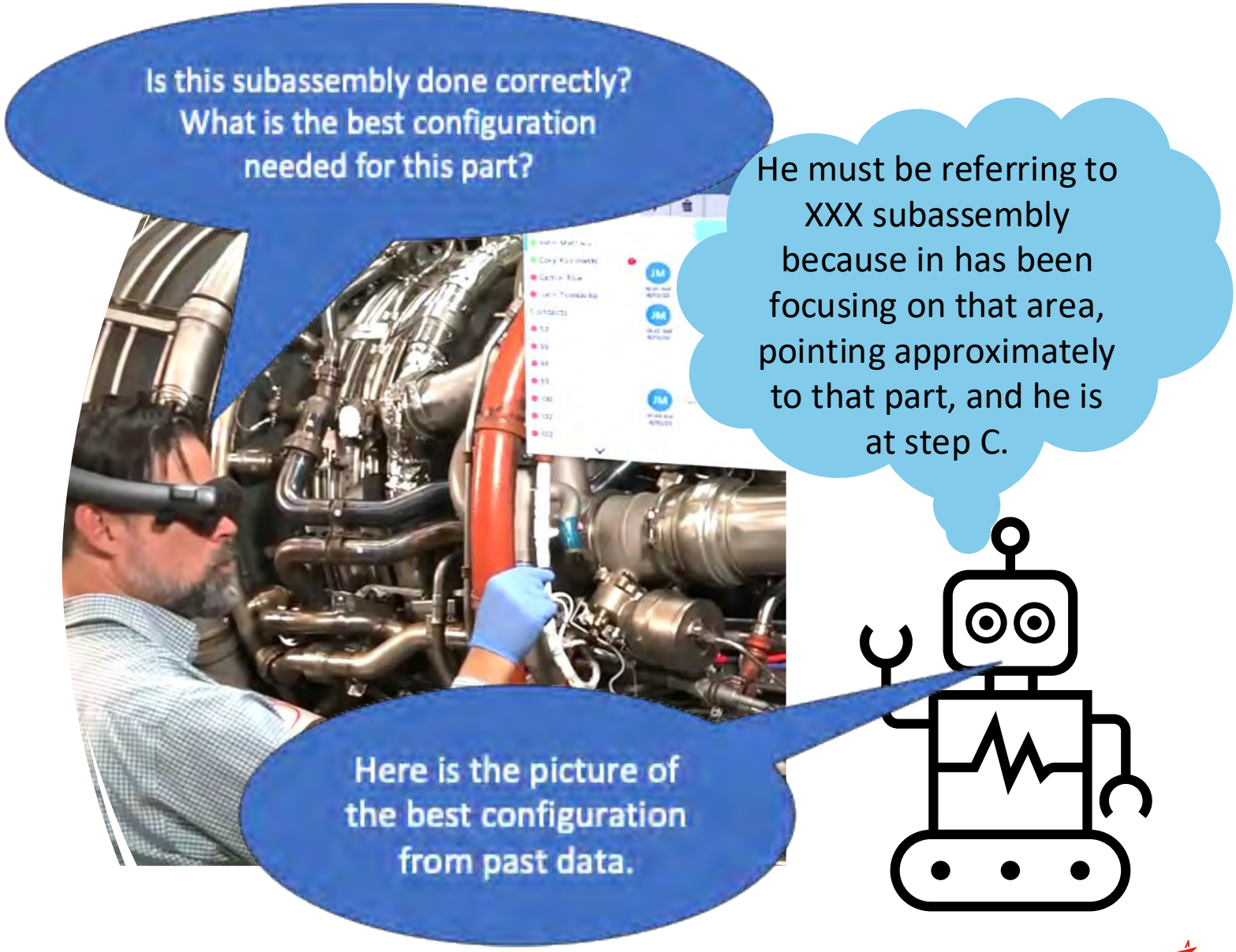


Proposed Approach:

HRI centred on building and updating an accessible **explicit world model**, representing the **common ground** between human and EAI to **align robot behaviours with human expectations**.

Common Ground in Human-Robot Collaboration

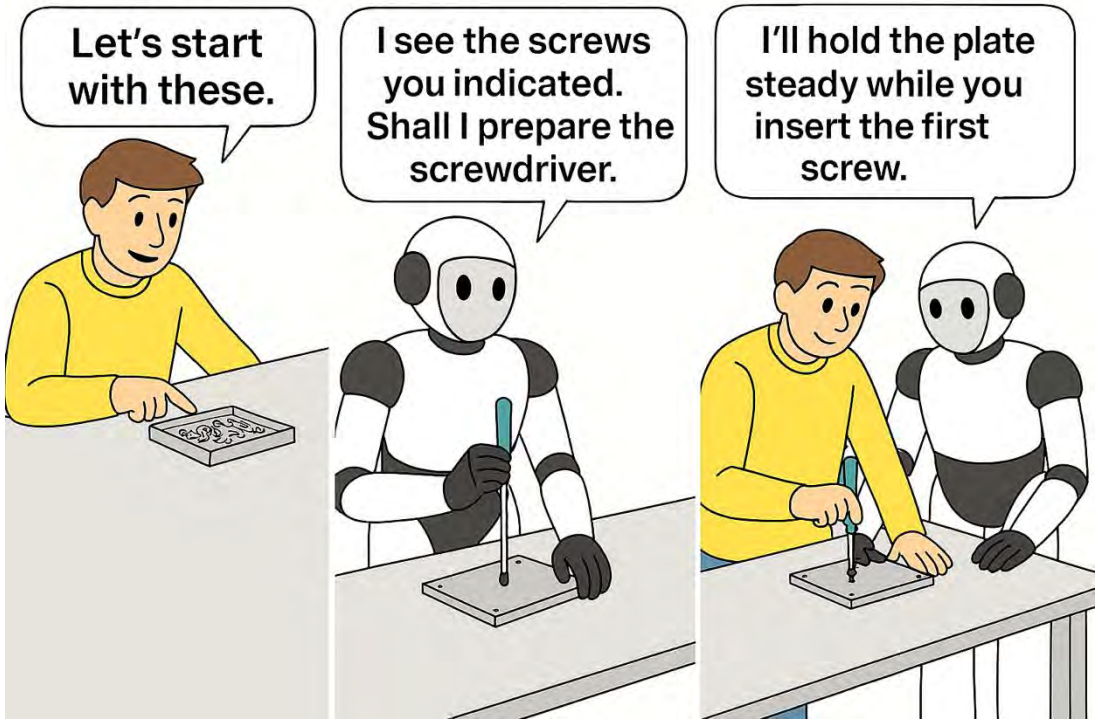
- Common ground = shared understanding of tasks, communications, and environments between agents (Dillenbourg and Traum, 2006)
- Aligns human and robots in terms of perception, cognition, and embodiment
- Forms the basis for cooperative action in Human-Robot Collaboration



Dillenbourg, P., & Traum, D. (2006). Sharing Solutions: Persistence and Grounding in Multimodal Collaborative Problem Solving. Journal of the Learning Sciences, 15(1), 121–151. https://doi.org/10.1207/s15327809jls1501_9

Building Common Ground

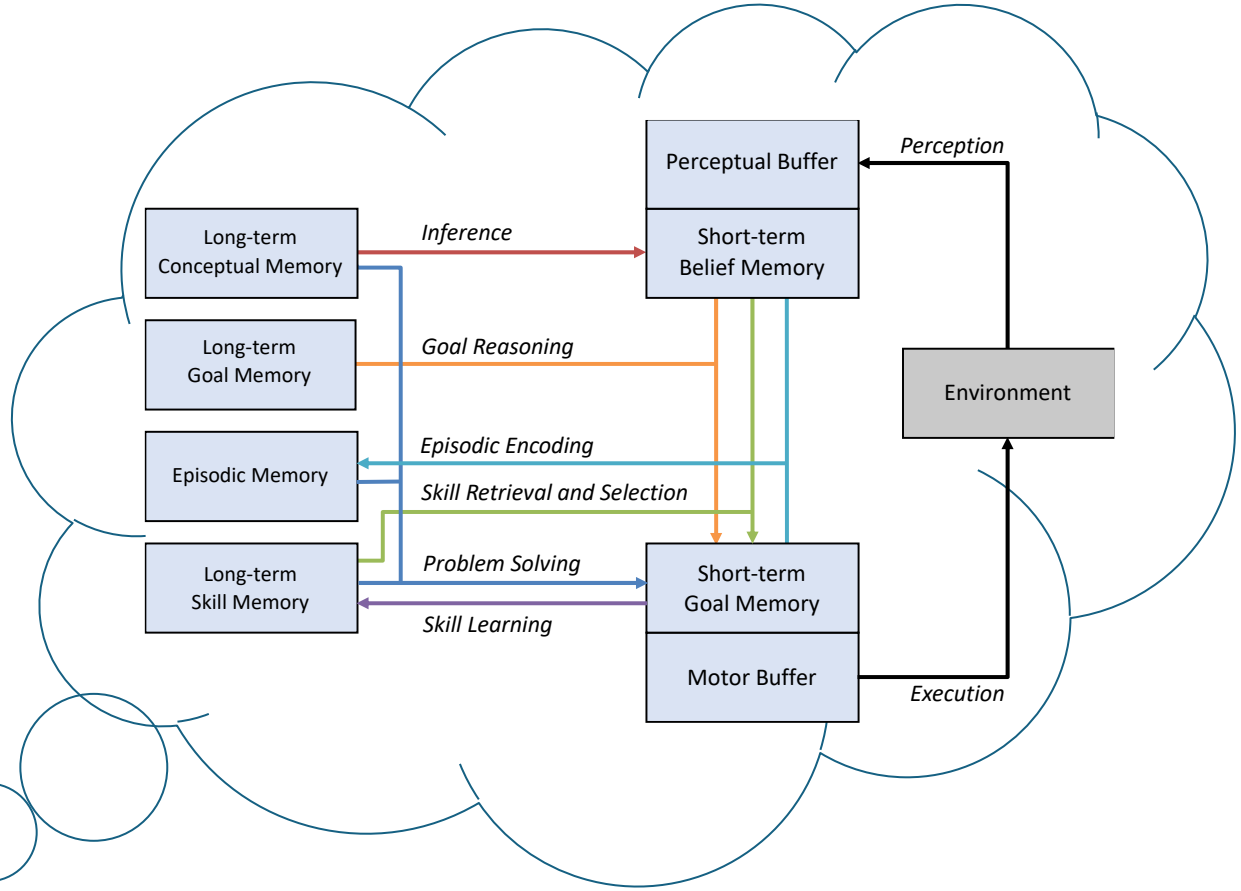
- Perceptual Grounding
  - Objects
  - Actors
  - Activity (Task), Situation (State)
- Joint Attention and Embodied Communication
  - Monitoring speech and behaviours (gaze, gestures etc.)
  - Continuous intention inference
  - Acting expressively to convey intentions

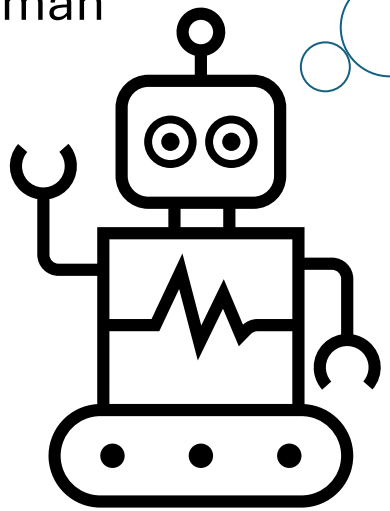


STEP	HUMAN ACTION	ROBOT ACTION	COMMON GROUND
1	Points to screws, says "Let's start with these."	Interprets gesture + speech, confirms understanding	Initial alignment on task goal
2	Nods, positions base plate	Retrieves screwdriver, avoids workspace obstruction	Dynamic coordination based on cues
3	Prepares to insert screw	Holds plate steady, verbal update	Maintains shared world model and predictability

Explicit World Model

- **Cognitive Architectures (CAs)**
  - Use explicit world models to represent the environment, relational concepts, and executable procedures
  - Raw sensory data is transformed into symbols and rules – a high level, interpretable and structured abstraction of the environment – for cognitive processing
  - Heavily dependent on human handcrafting

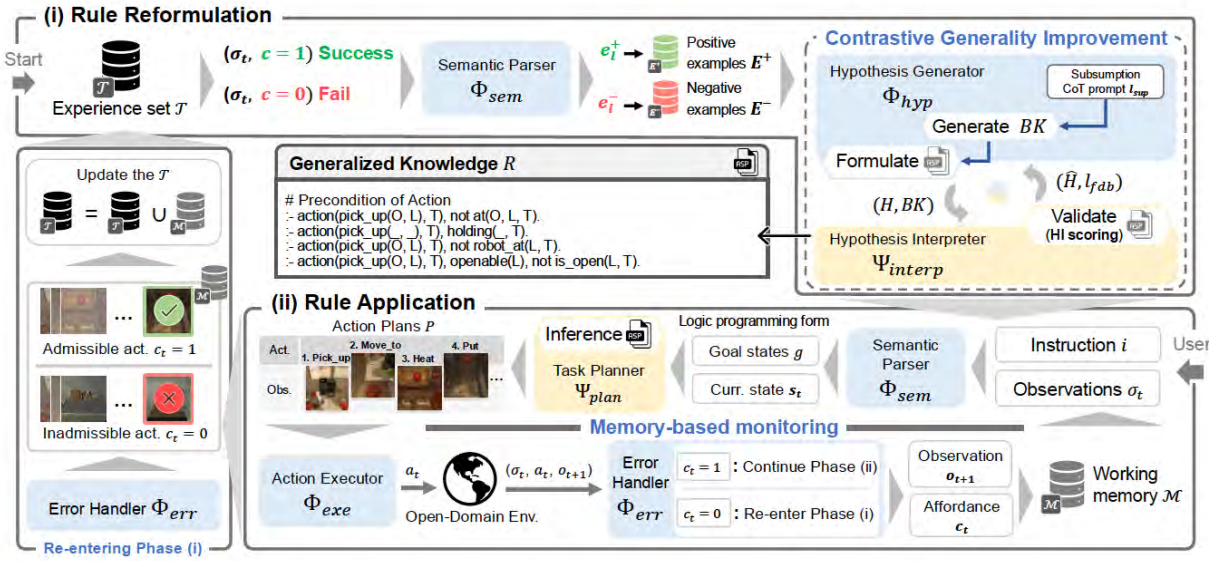


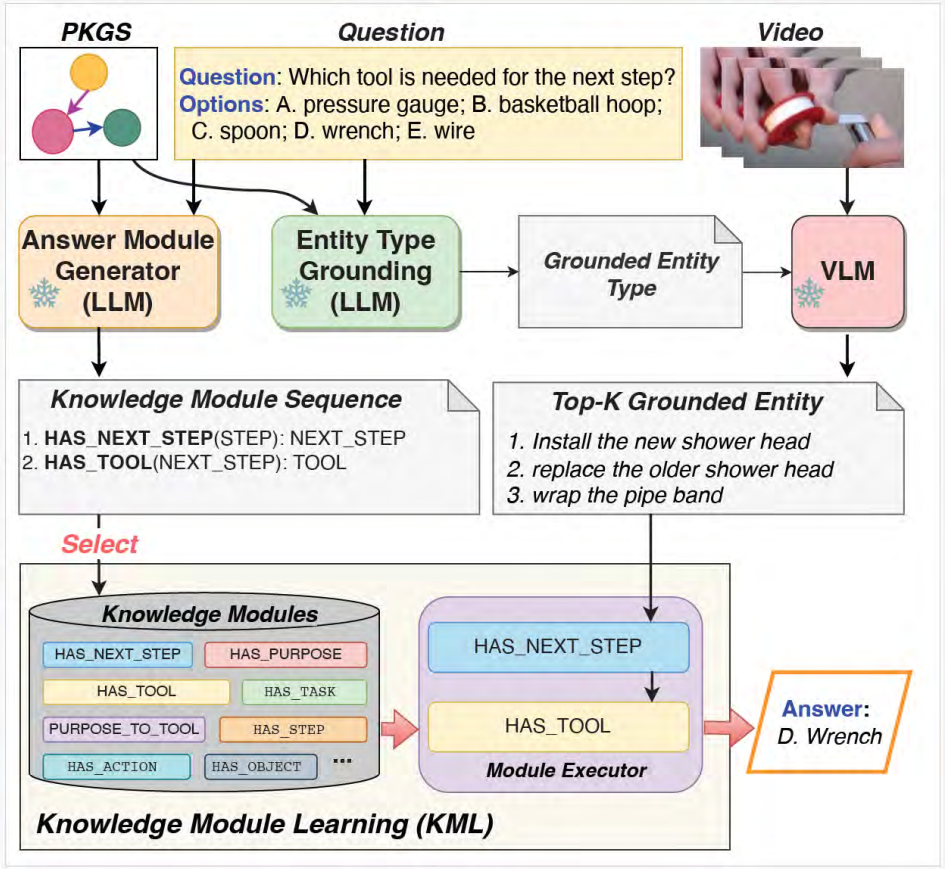


Langley, P. & Choi, D (2006). "A Unified Cognitive Architecture for Physical Agents." In *Proceedings of the 21st National Conference on Artificial Intelligence*, Vol 2, 1468-1474

Explicit World Model

- **Neuro-symbolic Architectures**
  - Explicit, interpretable, and self-evolving world models
  - Develop internal representations without human handcrafting





Knowledge Module Learning (Nguyen et al. 2025)

NeSyC (Choi et al., in press)

Choi,W.; Park, J.; Ahn, S.; Lee, D.; and Woo, H. In Press. "A Neuro-Symbolic Continual Learner for Complex Embodied Tasks in Open Domains." In 2025 13th International Conference on Learning Representations (ICLR).

Nguyen, T.-S.; Yang, H.; Neoh, T. Y.; Zhang, H.; Ee, Y. K.; and Fernando, B. 2025. "Neuro-Symbolic Knowledge Reasoning for Procedural Video Question Answering with Knowledge Module Learning (KML)." ArXiv, abs/2503.14957.

Spatial-Temporal World Models

1. World models as interface (Li Fei Fei's Marble)
   - Words and flat media → 3D assets humans can edit and share
2. World models as simulator (Deepmind's Genie)
   - Continuous, controllable video worlds that agents can interact with
   - SIMA 2-style agents built on top
3. World models as cognition (LeCun-style architectures)
   - Multimodal perception → latent variables and transition functions
   - Internal predictive state

"Why Fei-Fei Li, Yann LeCun and DeepMind Are All Betting on "World Models" — and How Their Bets Differ" https://entropytown.com/articles/2025-11-13-world-model-lecun-feifei-li/

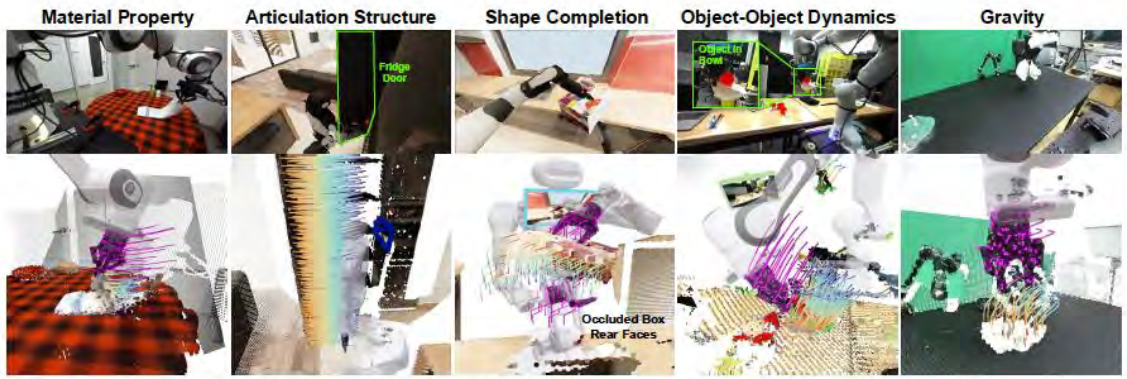


Figure 3. **Rich Supervision of 3D World Modeling for Physical Interactions**, when conditioned on 3D robot point flows and partial observable RGB-D. The 3D world modeling objective enjoys dense pixel-level supervision while encoding a wide range of capabilities central to robotic manipulation. To predict full-scene evolution, the model needs to implicitly segment objects of interest, identify material property and/or articulation structure, perform implicit shape completion for contact reasoning, propagate robot-object interaction for object-object dynamics, and simultaneously considering the effects of gravity, encapsulated all in a single forward pass of the learned model.

Huang, W., Chao, Y. W., Mousavian, A., Liu, M. Y., Fox, D., Mo, K., & Fei-Fei, L. (2026). PointWorld: Scaling 3D World Models for In-The-Wild Robotic Manipulation. arXiv preprint arXiv:2601.03782.

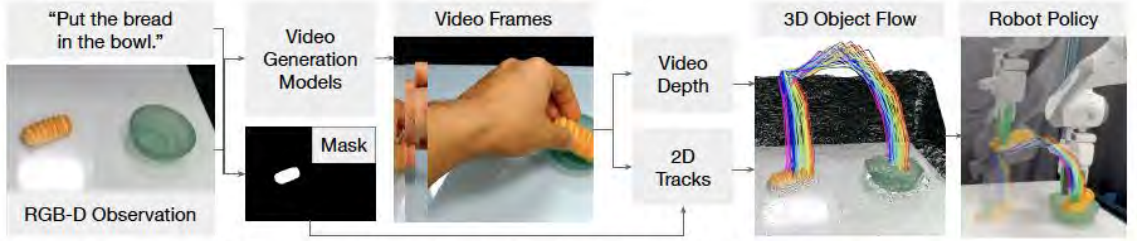
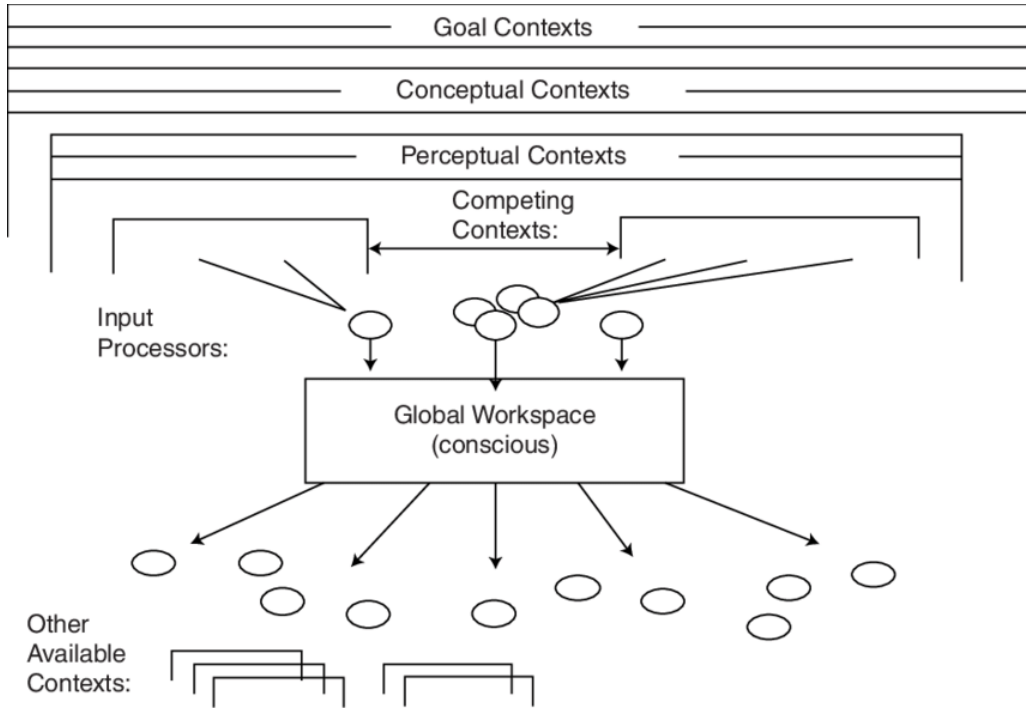


Fig. 2: **An overview of Dream2Flow.** Given a task instruction and an initial RGB-D observation, an image-to-video model synthesizes video frames conditioned on the instruction. We additionally obtain object masks, video depth, and point tracking from vision foundation models, which are used to reconstruct 3D object flow. Finally, a robot policy generates executable actions that track the 3D object flow using trajectory optimization or reinforcement learning.

Dharmarajan, K., Huang, W., Wu, J., Fei-Fei, L., & Zhang, R. (2025). Dream2Flow: Bridging Video Generation and Open-World Manipulation with 3D Object Flow. arXiv preprint arXiv:2512.24766.

World Models as Global Workspaces

- **Global Workspace Theory** (Baars 2005) is a foundational cognitive architecture that describes *global-information sharing* mechanisms in the brain.
- World Models function as global workspaces to provide a unified, interpretable representation of the current state of perception, goals, and context that is broadcast across specialised cognitive modules — essentially acting as a *shared internal model* that enables coordination and decision-making.



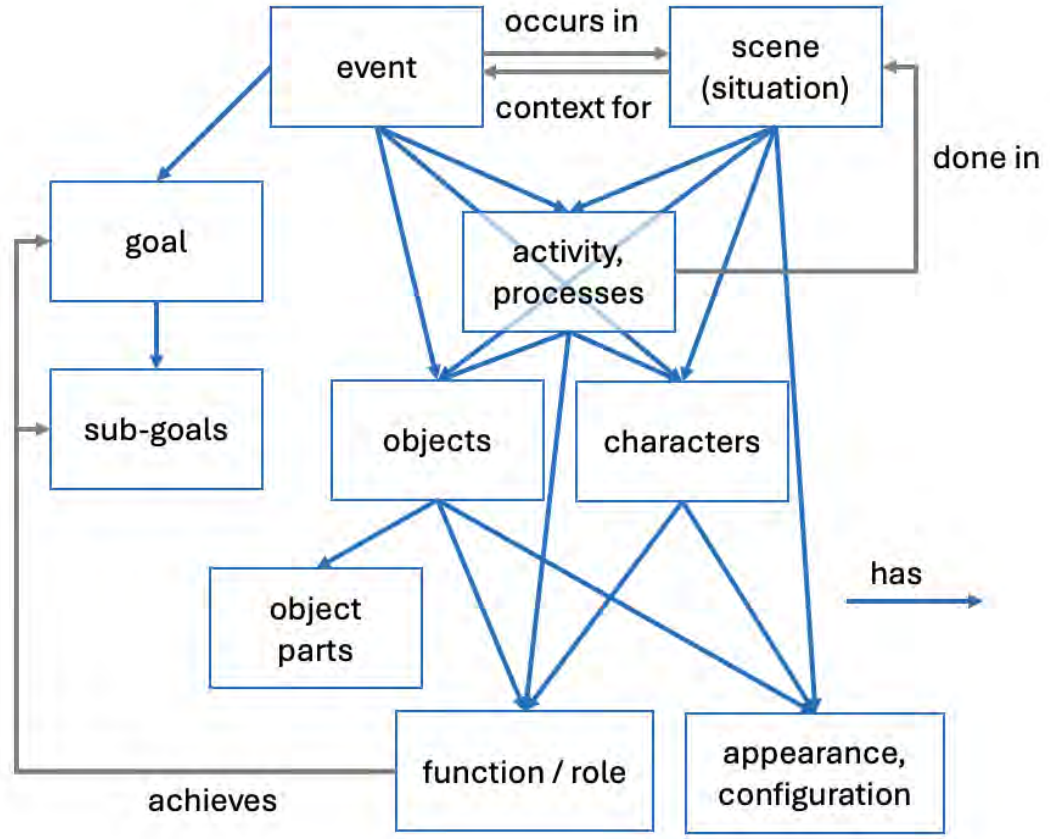
Baars Global Workspace Theory as depicted in Sun, et.al. (2007)

Baars, B. J. (2005). Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience. In S. Laureys (Ed.), Progress in brain research (Vol. 150, pp. 45–53). Elsevier. https://doi.org/10.1016/S0079-6123(05)50004-9

Sun, Ron & Franklin, Stan. (2007). Computational Models of Consciousness: A Taxonomy and Some Examples. Cambridge Handbook of Consciousness. 10.1017/CBO9780511816789.008.

World Models as Global Workspaces

- Episodic (event-centric, relational) vs Spatial-Temporal (flow, physical)
- Event-centric Representation
  - Event, Scene (time, space, causes, characters and goals)
  - Hierarchical time scales
    - Extended (coarse grained)
    - Brief (fine grained)
- Constructing World Models
  - Event Segmentation
  - Goal Inference
  - Event Model

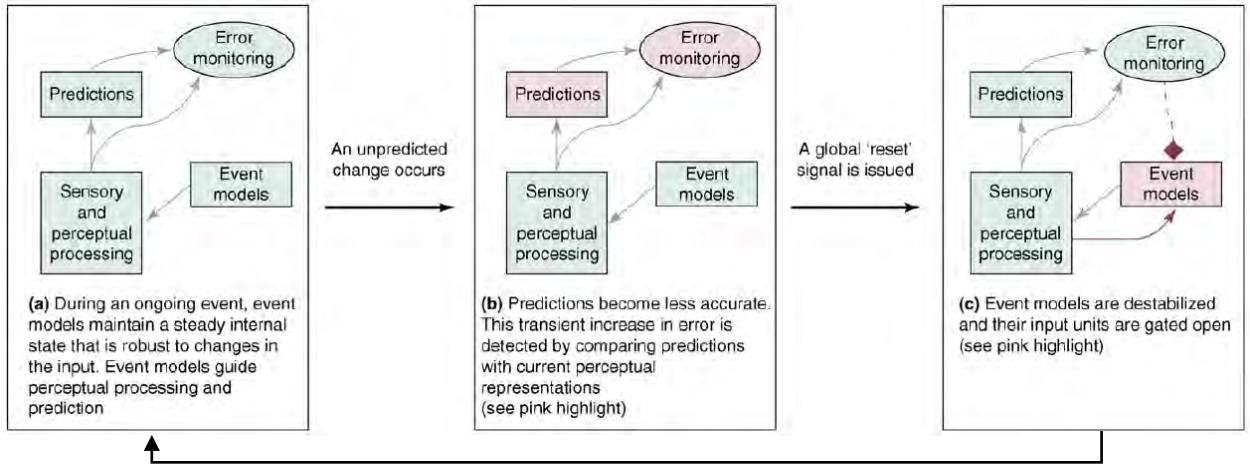


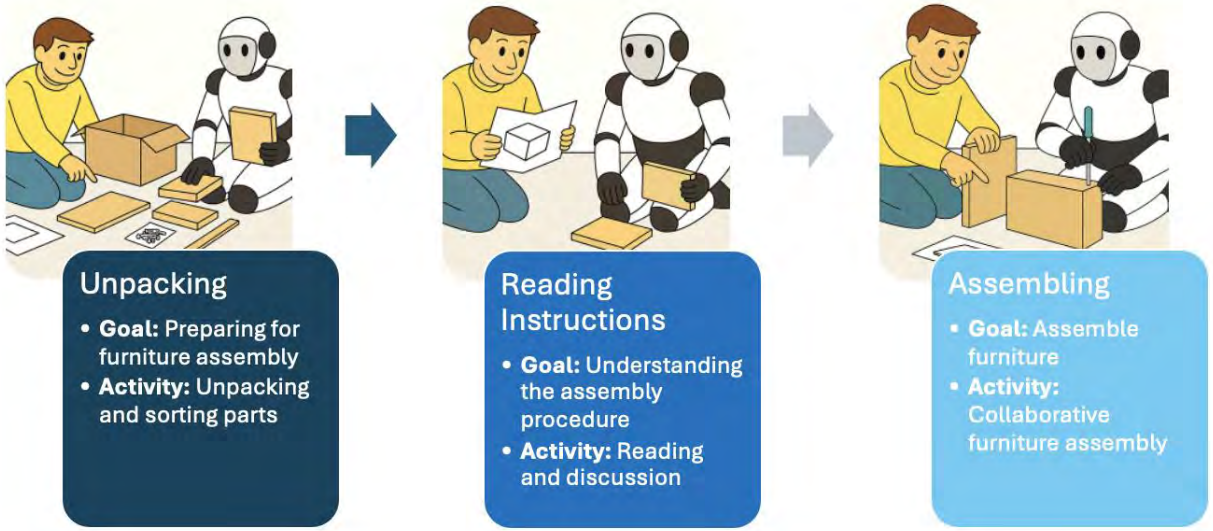
Tversky, B., Zacks, J. M., & Hard, B. M. (2008). The structure of experience. In T. F. Shipley & J. M. Zacks (Eds.), *Understanding events: From perception to action* (pp. 436–464). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195188370.003.0019

Event Perception and Segmentation

Event Segmentation Theory (EST) proposes that perceptual systems spontaneously segment an activity into events as a side effect of trying to anticipate upcoming information.

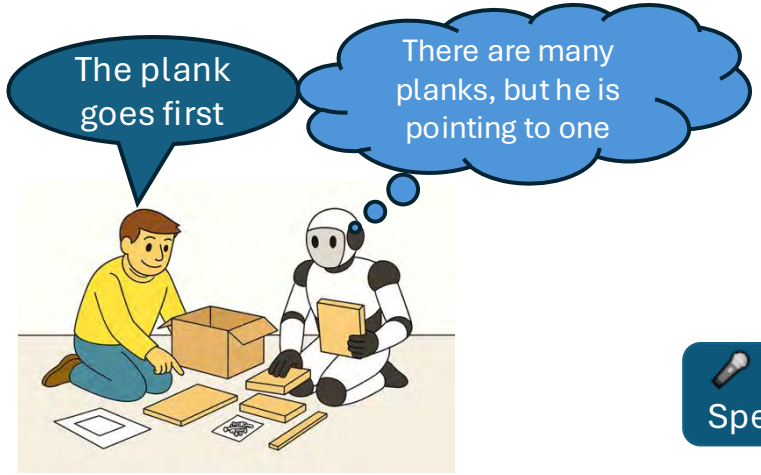
- Event boundaries are associated with changes in time, space, causes, characters and goals
- Actions are hierarchically organized by goals and subgoals
  - **Coarse-grained events** focus on objects, using more precise nouns and less precise verbs.
  - **Fine-grained events** focus on actions on those objects, using more precise verbs but specifying the objects less precisely.



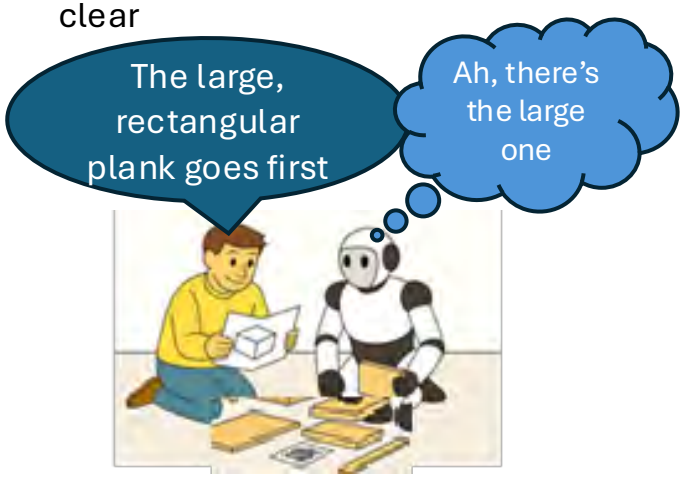


Kurby CA, Zacks JM. Segmentation in the perception and memory of events. Trends Cogn Sci. 2008 Feb;12(2):72-9. doi: 10.1016/j.tics.2007.11.004. PMID: 18178125; PMCID: PMC2263140.

Human State Sensing / Goal Inference

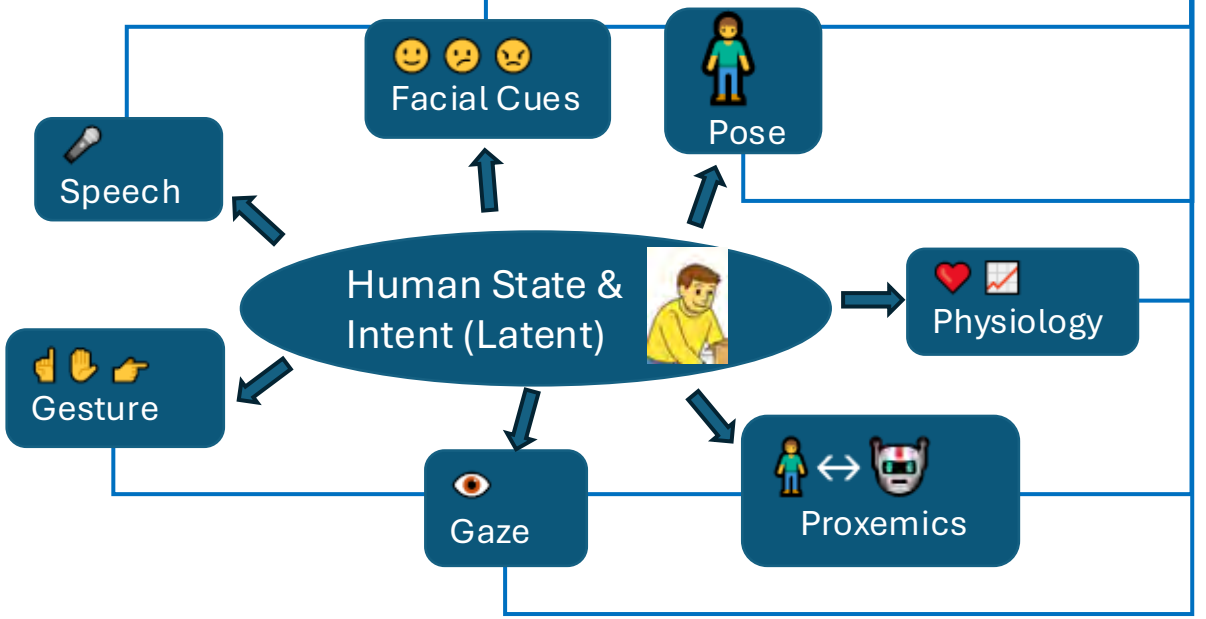


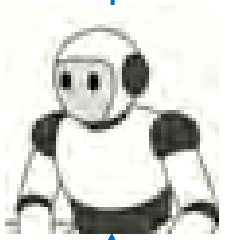
Ambiguous verbal instructions, but pointing/gaze makes it clear



Hands and eyes are busy reading but verbal instruction is specific

Human States / Intent expressed multimodally
Speech • Gesture • Gaze • Touch
Pose • Proxemics• Facial cues





Robot infers human state:
- Fatigue
- Cognitive load
- Anxiety
- Confusion
- Comfort
- Approval

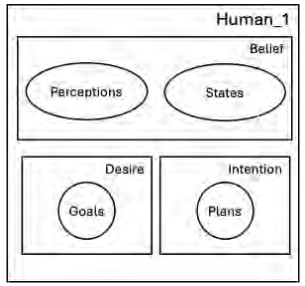
Robot infers intent / goal:
- Object
- Action
- Task
- Situation

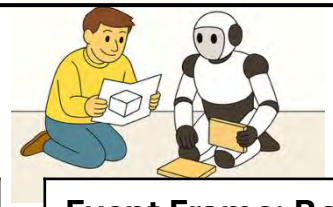
Monitoring human state (facial expressions, nodding etc.) can also help in predicting errors and misalignment.
➢ Enables flexible (non-task specific) error management framework

Stiber, M., Taylor, R. H., & Huang, C. M. (2023, March). On using social signals to enable flexible error-aware HRI. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 222-230).
Bremers, A., Pabst, A., Parreira, M. T., & Ju, W. (2023). Using Social Cues to Recognize Task Failures for HRI: Overview, State-of-the-Art, and Future Directions. *arXiv preprint arXiv:2301.11972*.

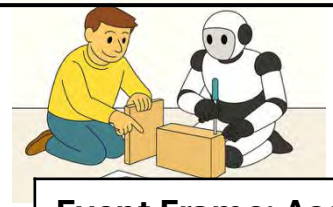
Explicit World Model Representation

Event Frame: Assembling Furniture	
Time	Morning
Location	Living Room
Goal	Assemble Furniture
Characters	Human_1, Robot_1

Event Frame: Unpacking	
Time	{Time 1}
Space	{Configuration 1: characters, objects}
Sub-Goal	Identify and layout parts for assembly
States	{Pre: Parts in box; Post: Parts organised on the floor}  }



Event Frame: Reading Instructions	
Time	{Time 2}
Space	{Configuration 2}
Sub-Goal	Understand assembly procedure
States	…



Event Frame: Assembling	
Time	{Time 3}
Space	{Configuration 3}
Sub-Goal	Assemble parts
States	…

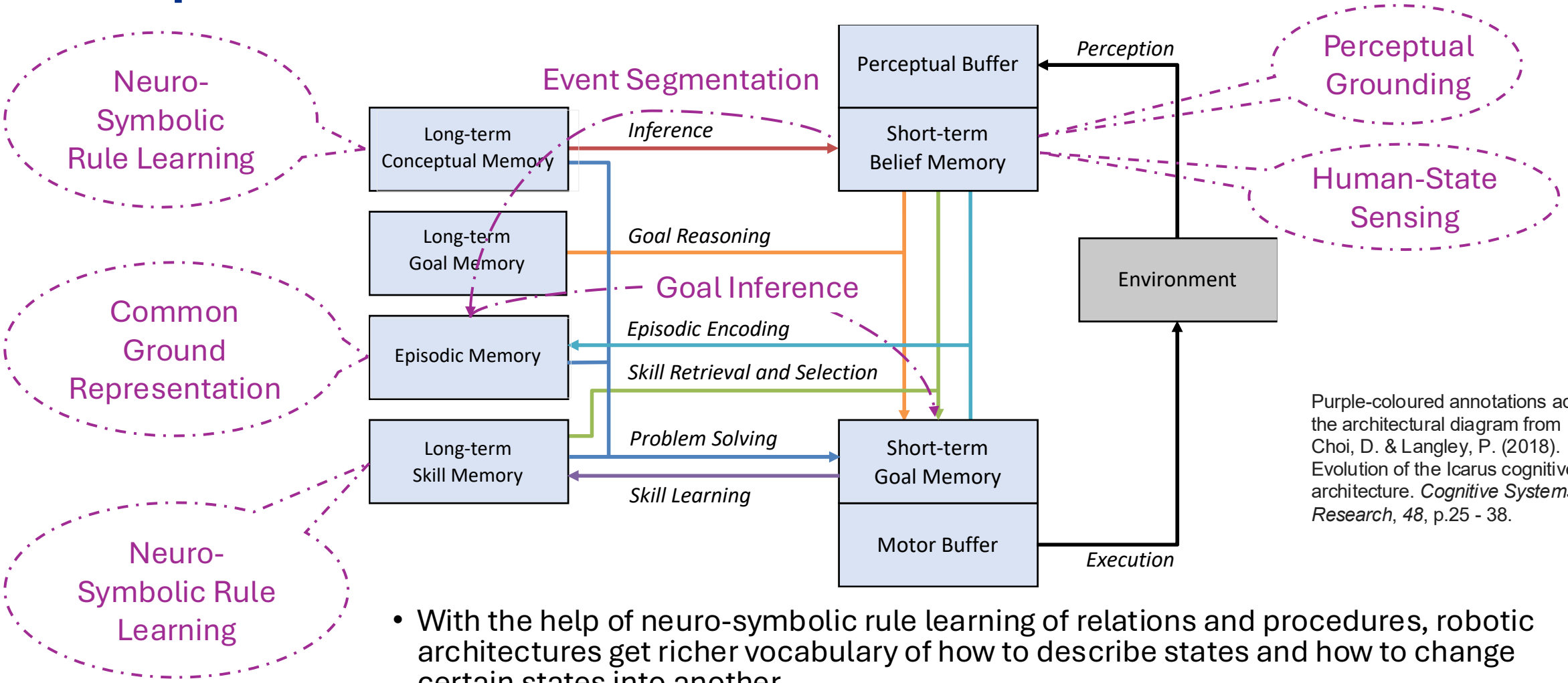
Event Frame: Step 1	
Time	{Time 3.1}
Space	{Configuration 3.1}
Sub-Goal	Assemble part 1
States	…

Event Frame: Step 2	
Time	{Time 3.2}
Space	{Configuration 3.2}
Sub-Goal	Assemble part 2
States	…

…

Event Frame: Step N	
Time	{Time 3.N}
Space	{Configuration 3.N}
Sub-Goal	Assemble parts together
States	…

Implications for Robotic Architectures



Purple-coloured annotations added to the architectural diagram from Choi, D. & Langley, P. (2018). Evolution of the Icarus cognitive architecture. *Cognitive Systems Research*, *48*, p.25 - 38.

- With the help of neuro-symbolic rule learning of relations and procedures, robotic architectures get richer vocabulary of how to describe states and how to change certain states into another.
- These rules constitute an explicit world model that the robotic agents can use to infer abstract states, form a common ground with another agent, select relevant goals, and execute for the chosen goals.

Call To Action – Capability Gaps

- Robust Perceptual Grounding
  - Visual scene parsing
  - Situation recognition
  - Human state sensing/recognition
- Accessible Common Ground Representation
  - Expressive (capture rich social, multimodal, and fluid nature of HRI)
  - Efficient (lightweight enough for real-time updating)
- Implementation in Embodied Cognitive Architecture
  - Event perception and segmentation
  - Human goal inference (Theory of Mind)
  - Neuro-symbolic rule learning from continuous experience
  - Common ground influence on robot's decision making

Thank you!

AAAI 2026 Summer Symposium Series

Sponsored by the Association for the Advancement of Artificial Intelligence

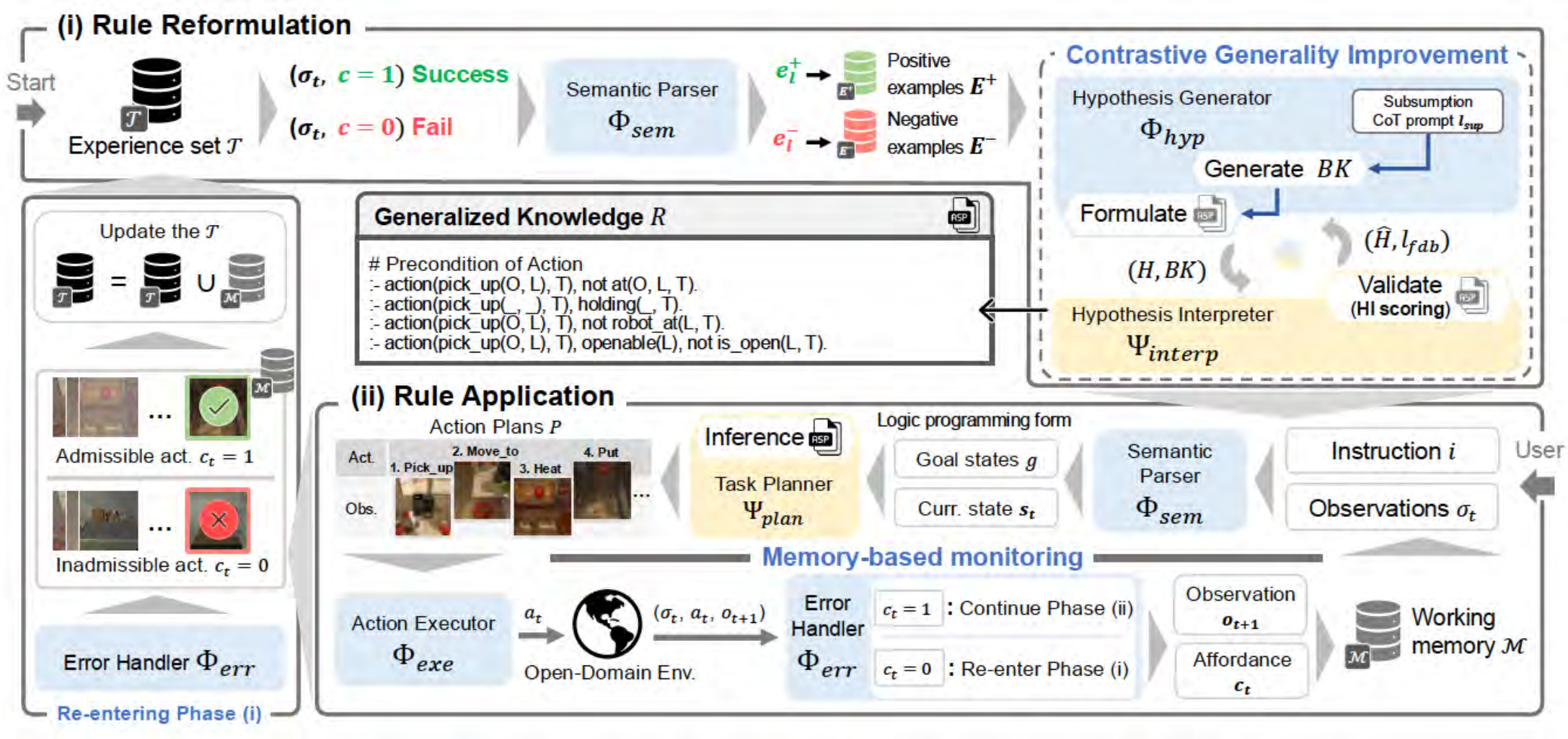
June 22-June 24, 2026
Dongguk University | Seoul, South Korea

Architectures for Embodied Agents: A Synergy of Classic and Foundation Model Paradigms

Organised by:

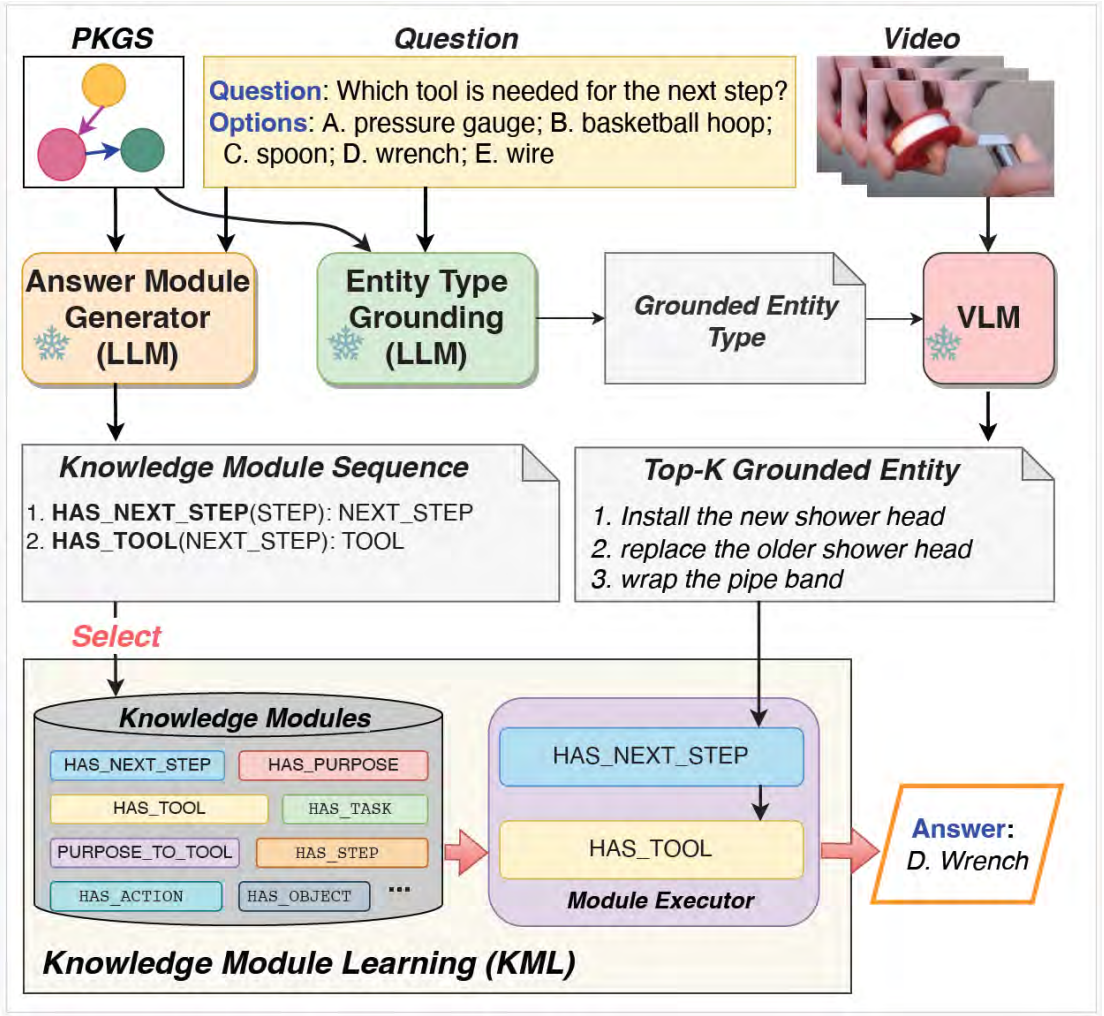
Dongkyu Choi, Samsung Electronics
Pat Langley, Georgia Tech Research Institute
Jaeheung Park, Seoul National University
Kenneth Kwok, Agency for Science, Technology and Research
Sanjay Oruganti, Rensselaer Polytechnic Institute

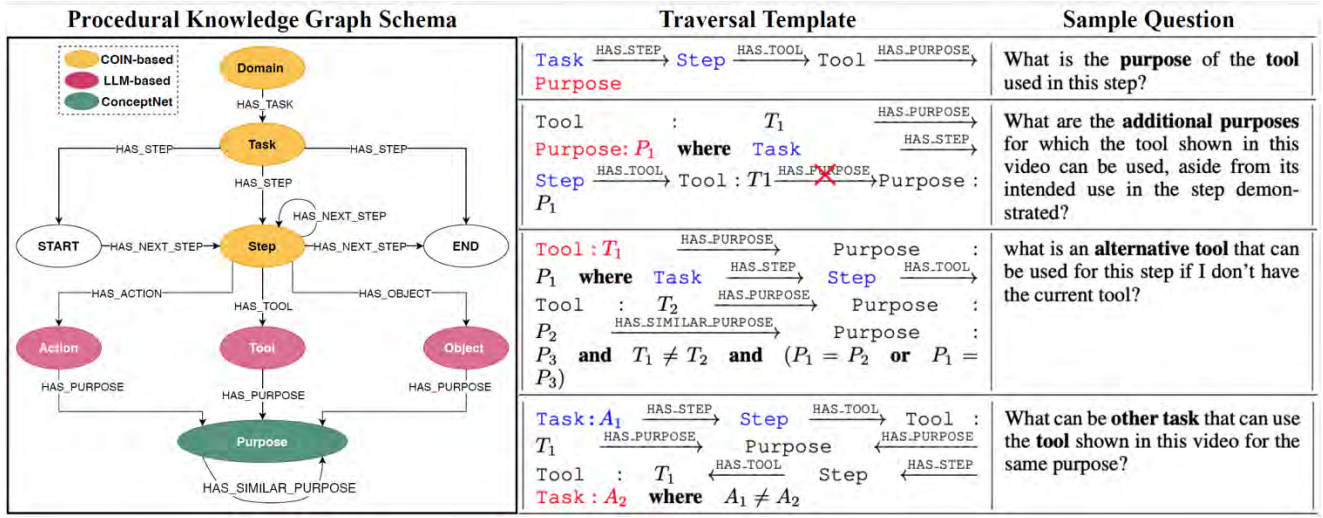
NESYC: A NEURO-SYMBOLIC CONTINUAL LEARNER FOR COMPLEX EMBODIED TASKS IN OPEN DOMAINS

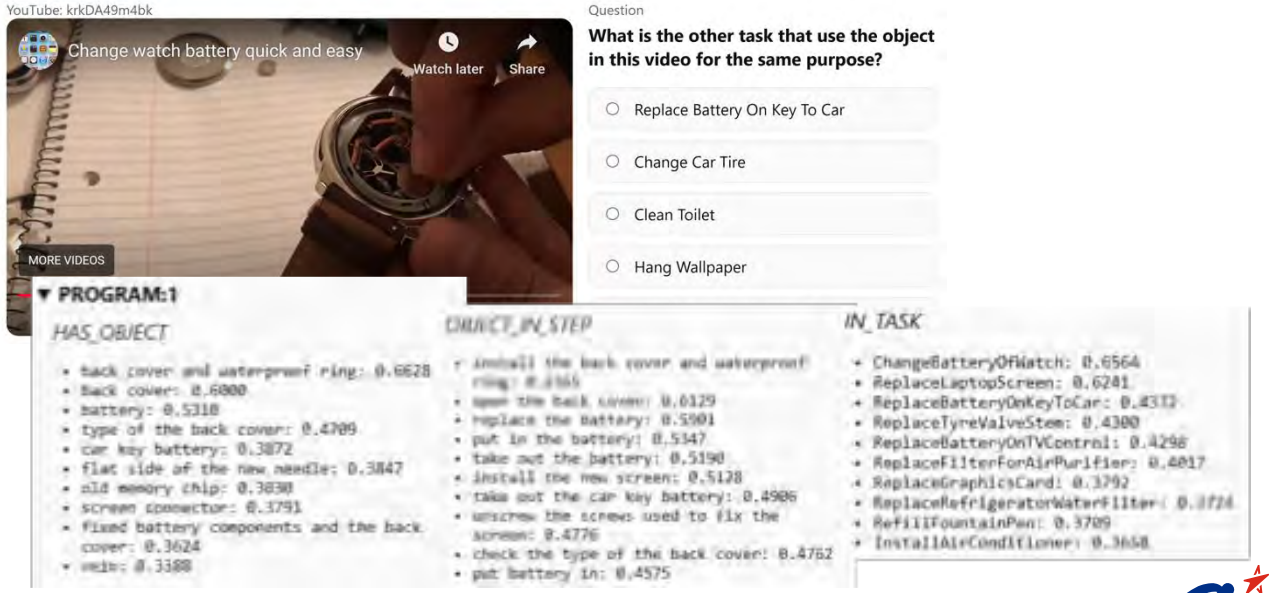


Choi,W.; Park, J.; Ahn, S.; Lee, D.; and Woo, H. In Press. “A Neuro-Symbolic Continual Learner for Complex Embodied Tasks in Open Domains.” In 2025 13th International Conference on Learning Representations (ICLR).

Knowledge Module Learning



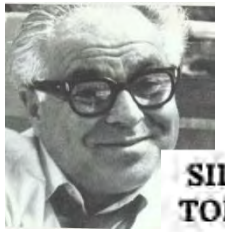




Nguyen, T.-S.; Yang, H.; Neoh, T. Y.; Zhang, H.; Ee, Y. K.; and Fernando, B. 2025. "Neuro-Symbolic Knowledge Reasoning for Procedural Video Question Answering with Knowledge Module Learning (KML)." ArXiv, abs/2503.14957.

Event Model - Representation

- Scripts and Frames



SILVAN TOMKIN

ROGER SCHANK

ROBERT ABELSON

Script: RESTAURANT
Track: Coffee Shop
Props: Tables, Menu, F-Food, Check, Money
Roles: S-Customer, W-Waiter, C-Cook, M-Cashier, O-Owner

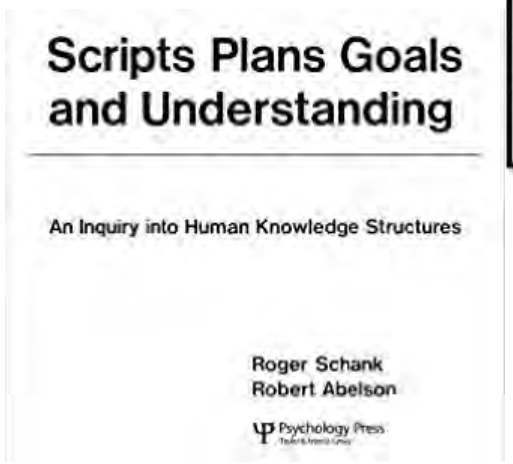
Entry conditions: S is hungry. S has money.
Results: S has less money. O has more money. S is not hungry. S is pleased (optional).

Scene 1: Entering
S PTRANS S into restaurant
S ATTEND eyes to tables
S MBUILD where to sit
S PTRANS S to table
S MOVE S to sitting position

Scene 2: Ordering
(menu on table) (W brings menu) (S asks for menu)
S PTRANS menu to S
S MTRANS signal to W
W PTRANS W to table
S MTRANS 'need menu' to W
W PTRANS W to menu
W PTRANS W to table
W ATRANS menu to S
S MTRANS food list to CP(S)
S MBUILD choice of F
S MTRANS signal to W
W PTRANS W to table
S MTRANS 'I want F' to W
W PTRANS W to C
W MTRANS (ATRANS F) to C
C MTRANS 'no F' to W
W PTRANS W to S
W MTRANS 'no F' to S
(go back to *) or
(go to Scene 4 at no pay path)
C DO (prepare F script) to Scene 3

Scene 3: Eating
C ATRANS F to W
W ATRANS F to S
S INGEST F
(Optionally return to Scene 2 to order more; otherwise go to Scene 4)

Scene 4: Exiting
S MTRANS to W
(W ATRANS check to S)
W MOVE (write check)
W PTRANS W to S
W ATRANS check to S
S ATRANS tip to W
S PTRANS S to M
S ATRANS money to M
(no pay path) S PTRANS S to out of restaurant



CHAIR frame	
A-kind-of:	furniture
number-of-legs:	an integer (default=4)
style-of-back:	straight, cushioned, ...
number-of-arms:	0,1,2

John's-chair frame	
a-kind-of:	furniture
number-of-legs:	4
style-of-back:	cushioned
number-of-arms:	0

Generic Restaurant Frame	
a-kind-of:	Business Establishment
Types:	Range: (Cafeteria, Seat-Yourself, Wait-to-be-seated, Fastfood) Default: IF plastic-orange-counter THEN fastfood IF stack-of-trays THEN cafeteria IF wait-for-waitress-sign OR reservation-made THEN wait-to-be-seated OTHERWISE seat_yourself
Location:	Range: an ADDRESS if-needed: (Look at the menu)
Name:	if needed: (Look at the menu)
Food-style:	Range: (Burgers, Chinese, American, Seafood, French) Default: Chinese if-added: (Update Alternative of Restaurant)
Time-of-Operation:	Range: a time-of-day Default: open evenings except Mondays
Payment form:	Range: (Cash, CreditC Script)
Event-Sequence:	Default: Eat-at-Restau

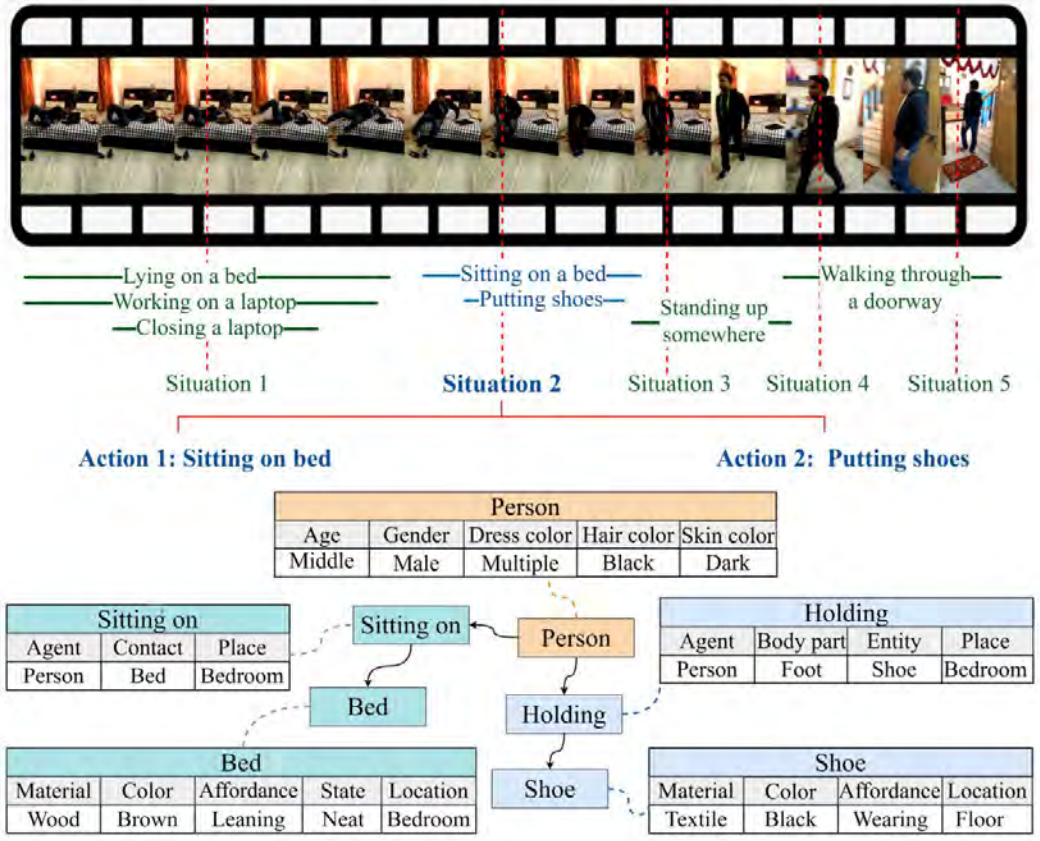
A Framework for Representing Knowledge

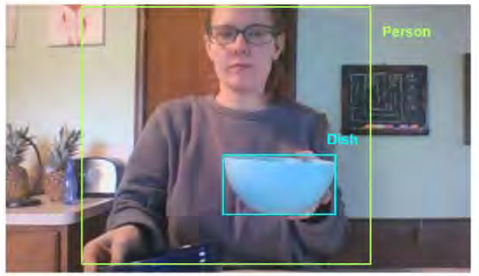
Marvin Minsky

MIT-AI Laboratory Memo 306, June, 1974.

Event Model - Representation

- Situational Scene Graphs

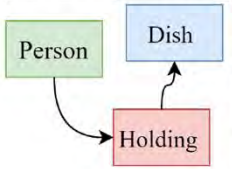




Action: Holding a dish

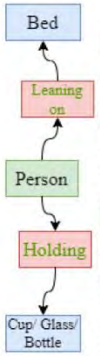
Dish	
Content	Unsure
Material	Ceramic
Color	Blue
Affordance	Holding
Location	Hand

Person	
Age	Middle
Gender	Female
Dress color	Ash
Hair color	Brown
Skin color	Fair



Holding	
Agent	Person
Body part	Hand
Entity	Dish
Place	Kitchen

Actions: Lying on a bed, Holding a cup/glass/bottle.



Cup/Glass/Bottle	
Content	Liquid
Material	Metal
Color	Silver
Container	Cup
Affordance	Drinking
Location	Hand

Person	
Age	Middle
Gender	Male
Dress color	Multiple
Hair color	Black
Skin color	Brwon

Leaning on	
Agent	Person
Contact	Bed
Place	Living room

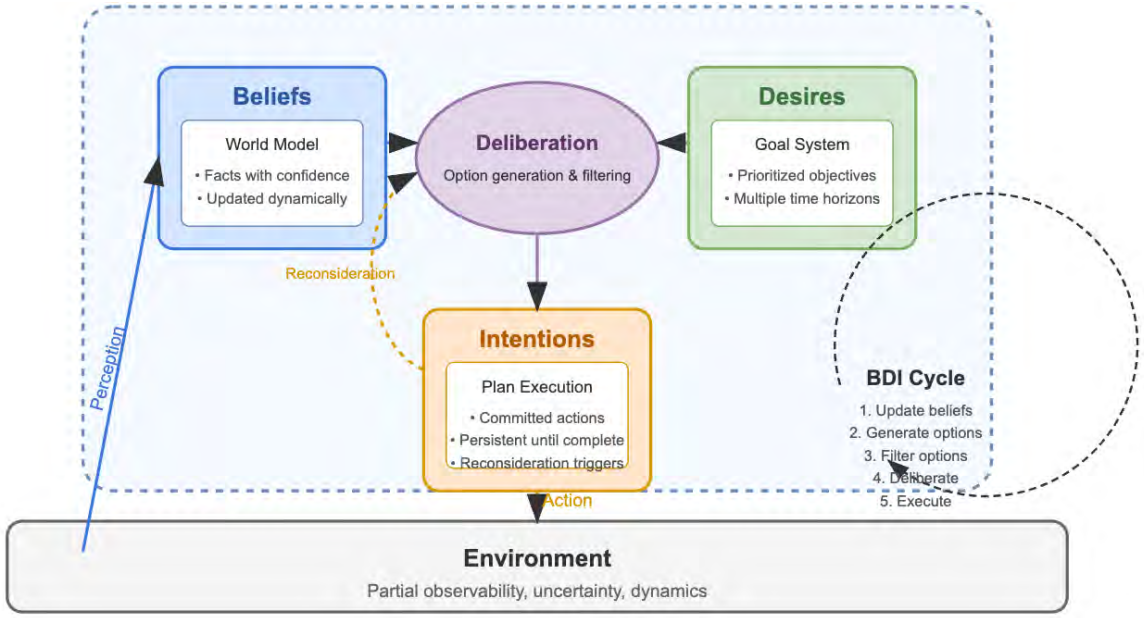
Holding	
Agent	Person
Body part	Hand
Entity	Cup
Place	Living room

Bed	
Material	Cloth
Color	Green
Affordance	Sitting
Location	Living room

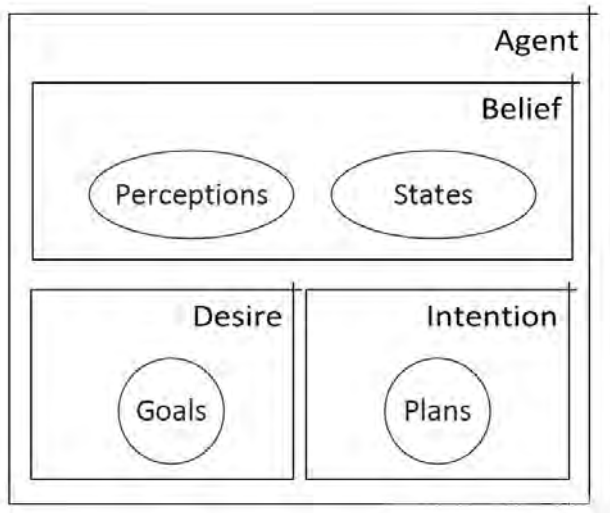
Event Model - Representation

Belief-Desire-Intention (BDI) Agent Model

- **Belief:** Agent's understanding of the world, including itself and other agents
- **Desire:** Agent's goals, preferences, and values
- **Intentions:** Agent's plans, strategies, and actions



Michael Bratman



Agent State expressed as Beliefs-Desires-Intentions (BDI)

Mccann, H., & Bratman, M.E. (1991). Intention, Plans, and Practical Reason.