



Ego4D

Advanced Topics in Embodied Learning and Vision

Ying Wang

2026.02.24

Egocentric videos

Egocentric videos captures the world from a first-person perspective, providing immersive and personalized data.

- **Memory Augmentation:** Log users' daily life and help users recall past experiences on demand.
- **AR/VR:** Enhances immersive experiences and real-time contextual interactions.
- **Human-Robot Interaction:** Enables robots to better understand and collaborate with humans.

...



Challenges?

Limited data

Long videos

Motion blurs

Object occlusions

Partial visibility

Ego4D

A massive-scale egocentric video dataset and benchmark suite.

- 3,025 hours of dailylife activity video
hundreds of scenarios (household, outdoor, workplace, leisure, etc.)
- 855 unique camera wearers
- 74 worldwide locations
- 9 different countries.

<https://ego4d-data.org/>



EGO-EXO4D



A massive-scale multi-view multi-modal dataset

- simultaneous ego and multiple exo videos
- multiple egocentric sensing modalities (audio, IMU, point cloud, eye gaze...)
- 5,035 videos
- 1,286 ego+exo hours
- 740 participants
- 123 sites

<https://ego-exo4d-data.org/>



Let's explore the data!

1. Review and accept the terms of Ego4D license agreement.

<https://ego4d-data.org/docs/start-here/#license-agreement>

2. Use the visualizer tool to explore the data

<https://visualize.ego4d-data.org>

3. For small datasets like EgoSchema, you can download a copy by yourself in Burst. If you need to access the full video dataset, please reach out to TAs. (Torch path: /projects/work/public/ml-datasets/, but the file system is not shared with Burst)
4. Check the official website for documentations <https://ego4d-data.org/docs/>

Ego4D Challenges

At CVPR/ICCV workshops, Meta hosts various Ego4D challenges of Ego4D's five benchmarks.

- Search Ego4D in <https://eval.ai/web/challenges/list> to participate in challenges!
- You can submit (an answer file) to evaluate your model on the private test sets



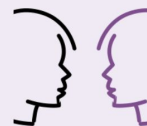
Episodic Memory



Hand-Object Interactions



AV Diarization



Social



Forecasting

Episodic memory

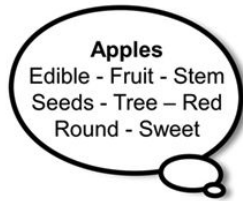
- **Episodic memory:** specific first-person experiences

E.g. What did I eat and who did I sit by on my first flight to France?

- **Semantic memory:** acquired knowledge—memorized facts or information.

E.g. What's the capital of France?

Semantic Memory



object knowledge learned
over many interactions

Episodic Memory



memory for specific events
that you have experienced

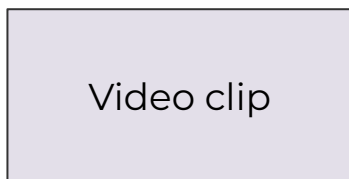


Egocentric video records the who/what/when/where
of an individual's daily life experience
→ ideal for episodic memory!

Applications: personal AI assistant

VQ2D/VQ3D

Visual queries with 2D/3D localization: Given an egocentric video clip and an image crop depicting the query object, return the most recent occurrence of the object in the input video, in terms of contiguous bounding boxes (2D + temporal localization) or the 3D displacement vector from the camera to the object in the environment.



+



+

Query
frame:
14357

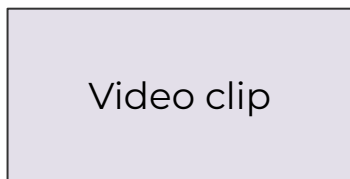


Most recent occurrence
(frame number + bb box)

[\[VQ2D example in visualizer\]](#)

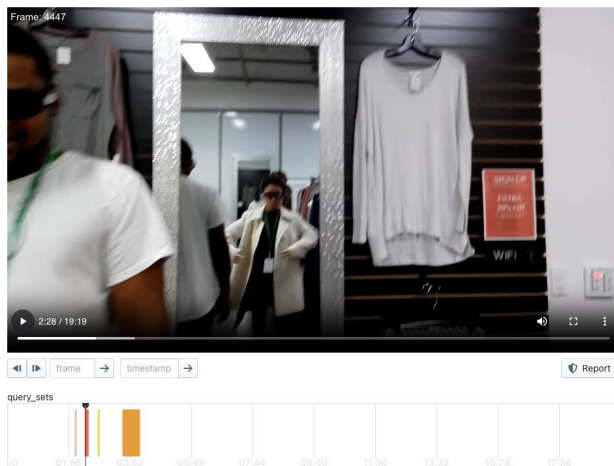
NLQ

Natural Language Query: Given a video clip and a query expressed in natural language, localize the temporal window within all the video history where the answer to the question is evident.



+

Query: Who did I interact with when I looked in the mirror.?



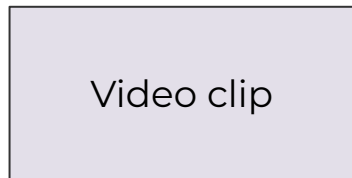
A temporal window that can answer the question:

02:26 - 02:34

[\[NLQ example in visualizer\]](#)

MQ

Moments queries: Given an input video and a query action category, the goal is to retrieve all the instances of this action category in the video. Specifically, it poses the request 'Retrieve all the moments that I do X in the video.', where X comes from a pre-defined taxonomy of action categories.



+

Moment:
chop_/cut_wood_pieces
_using_tool



All temporal windows of the given activity

0:00 - 03:48 chop_

0:15 - 03:45 chop_

02:28 - 02:31 chop_

[\[MQ example in visualizer\]](#)

EgoTracks

EgoTracks: Given an egocentric video and a visual template of an object, localize the bounding box containing the object in each frame of the video along with a confidence score representing the presence of the object.



[\[EgoTrack paper\]](#)

Goal Step

Goal Step: Given an untrimmed egocentric video, identify the temporal action segment corresponding to a natural language description of the step. Specifically, predict the (start_time, end_time) for a given keystone description.



- > 0:00 - 02:16 Stir dough (Prepare the dough mixture)
 - > 02:19 - 02:26 Clean up the kitchen area (Clean the cabinet)
 - > 02:31 - 05:28 Knead the dough until it is smooth (Knead the dough in a bowl)
 - > 05:29 - 05:35 Clean up the kitchen area (Clean the cabinet)
 - > 05:41 - 06:36 Wash hands (Clean hands)
 - > 06:40 - 06:55 Clean up the kitchen area (Clean the cabinet)
 - > 06:57 - 08:42 Roll out the dough on a floured surface (Roll dough into balls)
 - > 08:43 - 08:52 Clean and clear kitchen surfaces (Clean the table)
 - > 08:57 - 09:28 Sprinkle flour onto the cooking surface (Add flour to chopping board)
 - > 09:34 - 14:38 Roll out the dough on a floured surface (Roll out the dough ball)
 - > 14:45 - 15:10 Wash or disinfect chopping board (Clean the chopping board)
- > **substeps** [37]

[\[Goal Step example in visualizer\]](#)

EgoSchema

EgoSchema: Given a 3-minute video clip, one question and 5 possible answer choices, output the index from 0 to 4 indicating which answer choice is the most correct.

? What is the overarching behavior of C and the man in the video?

- 1 C teaches the man game rules but the man seems distracted and is not paying attention
- 2 The man teaches C how to play the card game while organizing the deck for future games
- 3 C and the man are playing a card game while keeping track of it in a notebook
- 4 C shows the man how to properly shuffle cards while the man plays them
- 5 The man shows C a new card game while C takes notes for future reference



[\[EgoSchema website\]](#)

More ego models and data!

- **EPIC-KITCHENS-100**: 20M frames of egocentric footage of first-person view kitchen activity, captured in an unscripted manner. <https://epic-kitchens.github.io/2025>
- **Aria Digital Twin**: 200 sequences of real-world activities conducted by Aria wearers in two real indoor scenes with 398 object instances. <https://www.projectaria.com/datasets/adt/>
- **EgoBody**: Large-scale dataset capturing ground-truth 3D human motions during social interactions in 3D scenes. <https://sanweiliti.github.io/egobody/egobody.html>
- **HoloAssist**: Egocentric human interaction dataset, where two people collaboratively complete physical manipulation tasks. <https://holoassist.github.io/>
- **EgoLife**: multimodal daily activities of six participants over a week <https://github.com/EvolvingLMMs-Lab/EgoLife>
-

Accept (Oral)

Accept (Spotlight)

Accept (Poster)

Accept (conditional oral)

Accept (conditional spotlight)

Accept (conditional poster)

Reject

Withdrawn Submissions

Desk Rejected Submissions

ego

**Scalable Benchmarking and Robust Learning for Noise-Free Ego-Motion and 3D Reconstruction from Noisy Video**

Xiaohao Xu, Tianyi Zhang, Shibo Zhao, Xiang Li, Sibow Wang, Yongqi Chen, Ye Li, Bhiksha Raj, Matthew Johnson-Roberson, Sebastian Scherer, Xiaonan Huang

Published: 22 Jan 2025, Last Modified: 11 Feb 2025 ICLR 2025 Poster Readers: Everyone

Show details

MMEgo: Towards Building Egocentric Multimodal LLMs

Hanrong Ye, Haotian Zhang, Erik Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Haoxuan You, Dan Xu, Zhe Gan, Jiasen Lu, Yinfei Yang, Bowen Zhang

Published: 22 Jan 2025, Last Modified: 11 Feb 2025 ICLR 2025 Poster Readers: Everyone

Show details

Do Egocentric Video-Language Models Truly Understand Hand-Object Interactions?

Boshen Xu, Ziheng Wang, Yang Du, Zhinan Song, Sipeng Zheng, Qin Jin

Published: 22 Jan 2025, Last Modified: 11 Feb 2025 ICLR 2025 Poster Readers: Everyone

Show details

X-Gen: Ego-centric Video Prediction by Watching Exo-centric Videos

Jilan Xu, Yifei Huang, Baoqi Pei, Junlin Hou, Qingqiu Li, Guo Chen, Yuejie Zhang, Rui Feng, Weidi Xie

Published: 22 Jan 2025, Last Modified: 11 Feb 2025 ICLR 2025 Poster Readers: Everyone

Show details

Leveraging Driver Field-of-View for Multimodal Ego-Trajectory Prediction

M. Eren Akbiyik, Nedko Savov, Danda Pani Paudel, Nikola Popovic, Christian Vater, Otmar Hilliges, Luc Van Gool, Xi Wang

Published: 22 Jan 2025, Last Modified: 11 Feb 2025 ICLR 2025 Poster Readers: Everyone

Show details

ego

**Spatial Reasoning with Vision-Language Models in Ego-Centric Multi-View Scenes**

Mohsen Gholami, Ahmad Rezaei, Zhou Weimin, Sitong Mao, Shunbo Zhou, Yong Zhang, Mohammad Akbari

Published: 26 Jan 2026, Last Modified: 11 Feb 2026 ICLR 2026 Poster Readers: Everyone

Show details

EgoHandICL: Egocentric 3D Hand Reconstruction with In-Context Learning

Binzhu Xie, Shi Qiu, Sicheng Zhang, Yinqiao Wang, Hao Xu, Muzammal Naseer, Chi-Wing Fu, Pheng-Ann Heng

Published: 26 Jan 2026, Last Modified: 11 Feb 2026 ICLR 2026 Poster Readers: Everyone

Show details

Temporal Slowness in Central Vision Drives Semantic Object Learning

Timothy Schaumlöffel, Arthur Aubret, Gemma Raig, Jochen Triesch

Published: 26 Jan 2026, Last Modified: 11 Feb 2026 ICLR 2026 Poster Readers: Everyone

Show details

EgoWorld: Translating Exocentric View to Egocentric View using Rich Exocentric Observations

Junho Park, Andrew Sangwoo Ye, Taemin Kwon

Published: 26 Jan 2026, Last Modified: 23 Feb 2026 ICLR 2026 Poster Readers: Everyone

Show details

EgoDex: Learning Dexterous Manipulation from Large-Scale Egocentric Video

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, Jian Zhang

Published: 26 Jan 2026, Last Modified: 11 Feb 2026 ICLR 2026 Poster Readers: Everyone

Show details

FlowAD: Ego-Scene Interactive Modeling for Autonomous Driving

Mingzhe Guo, Yixiang Yang, Chuanrong Han, Rufeng Zhang, Shirui Li, Ji Wan, Zhipeng Zhang

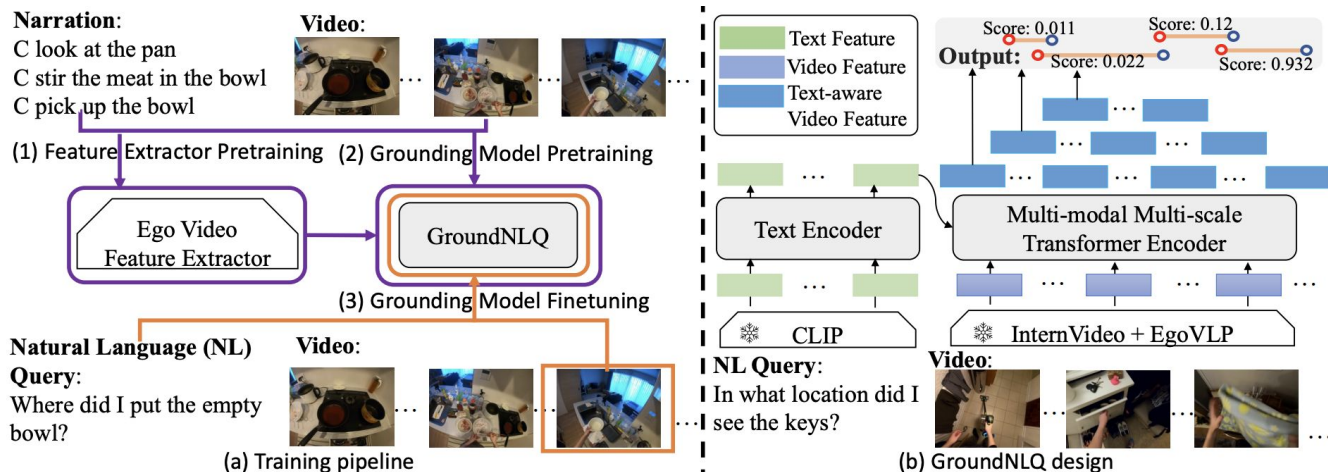
Published: 26 Jan 2026, Last Modified: 11 Feb 2026 ICLR 2026 Poster Readers: Everyone

Show details

<https://openreview.net/group?id=ICLR.cc%2F2026%2FConference#tab-accept-poster>

Video features

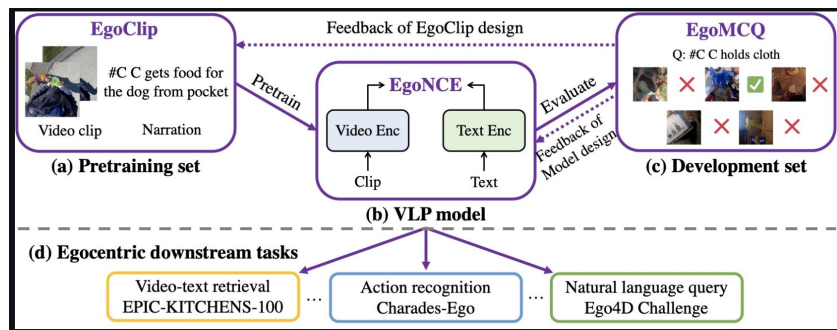
1. Extract video features (slowfast, omnivore...) and text features using pretrained encoders
2. Fuse features and feed them into a base video model



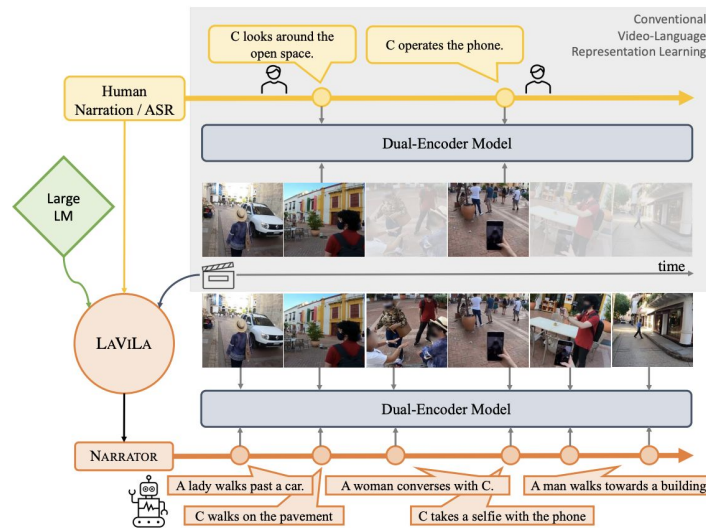
Example: GroundNLQ (winning solution for NLQ2023) <https://arxiv.org/pdf/2306.15255>

Ego foundation models

EgoVLP: “CLIP” trained on video-narration data constructed from Ego4D.



LaViLa: Use LLMs to densely narrate long videos, then use those narrations to train a dual-encoder.

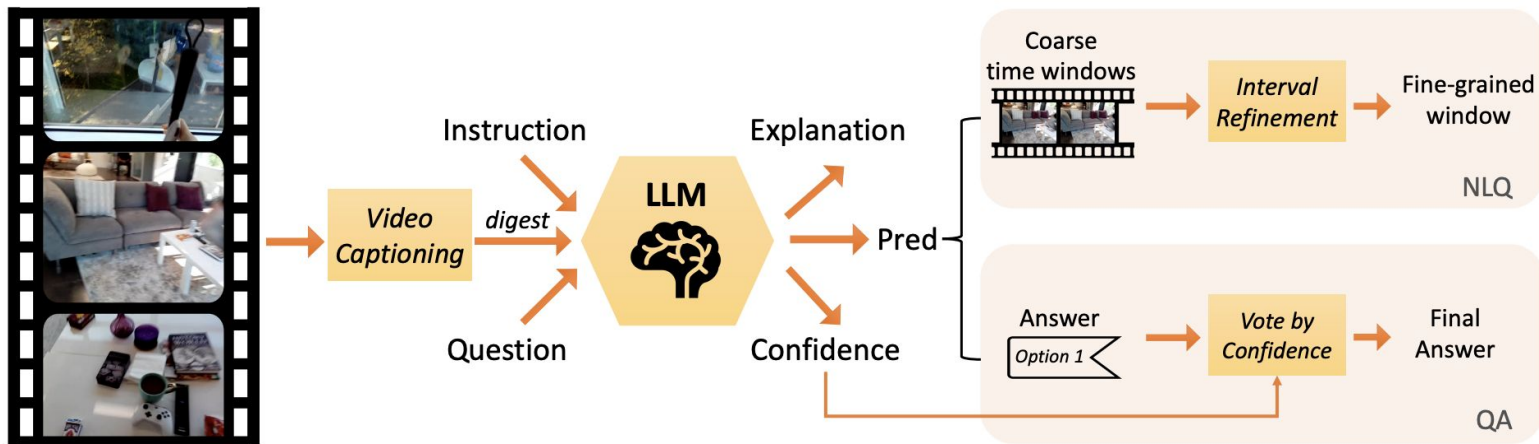


- <https://qhlin.me/EgoVLP/>
- <https://shramanpramanick.github.io/EgoVLPv2/>

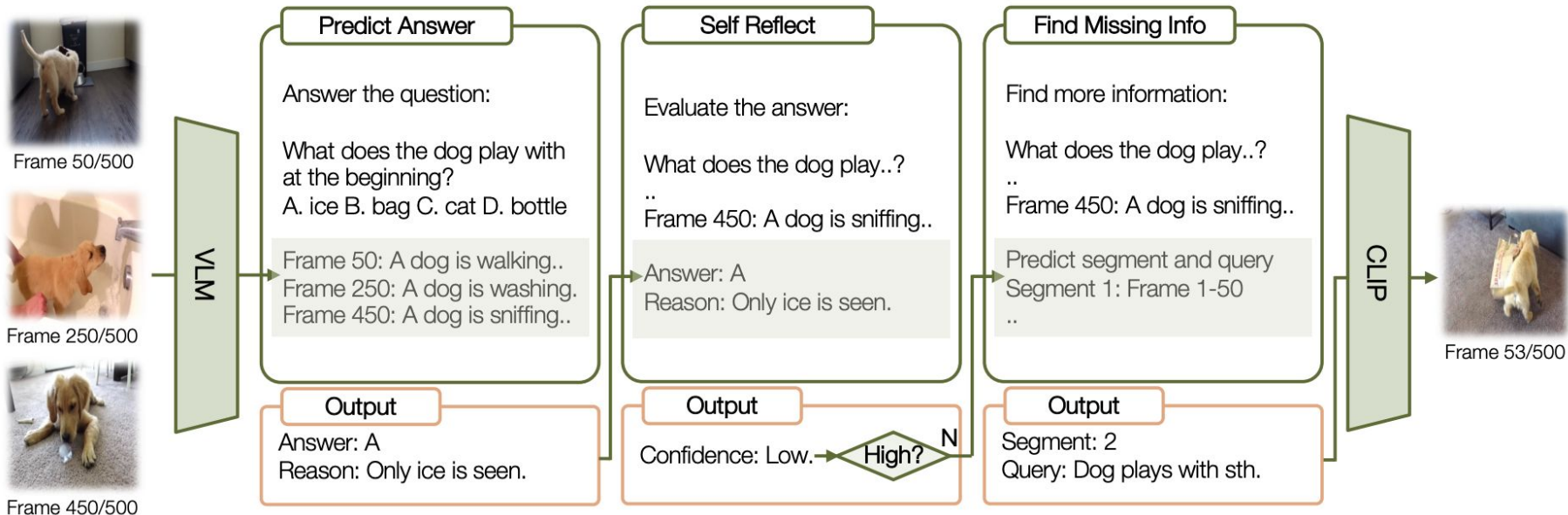
<https://facebookresearch.github.io/LaViLa/>

Captions + LLMs

1. Convert videos into a textual log using a captioning model.
2. Use LLM to process the captions and answer queries.



Agents





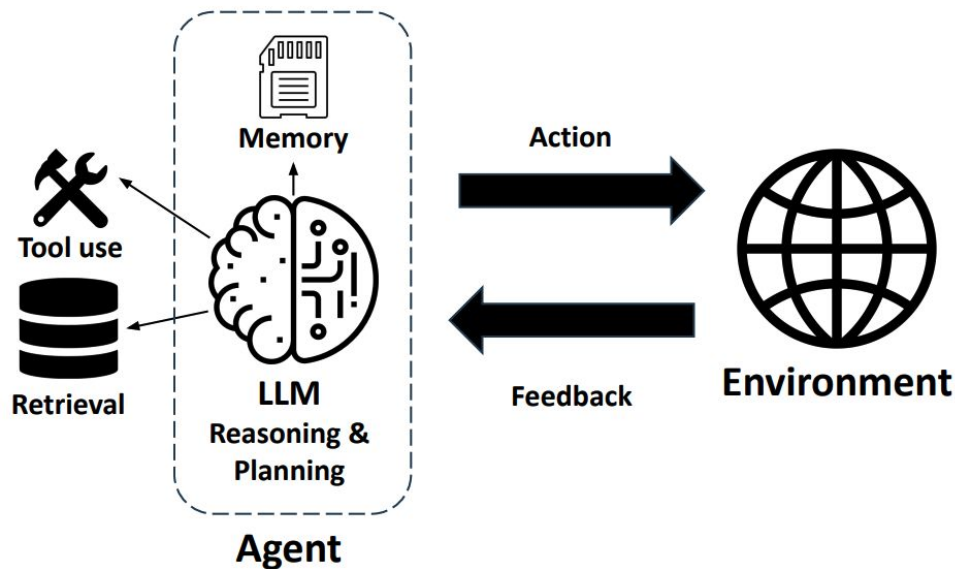
LLM Agent

Advanced Topics in Embodied Learning and Vision

Ying Wang

LLM Agents

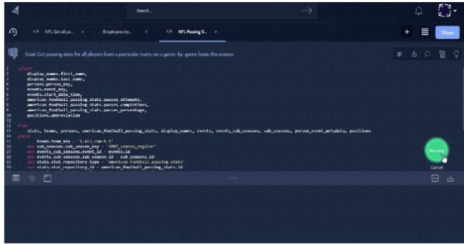
“An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators.” — Russell & Norvig, AI: A Modern Approach (2020)



Why LLM Agents?

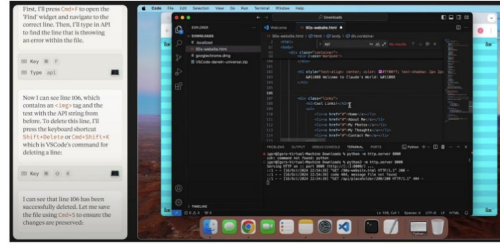
- Solving real-world tasks typically involves a trial-and-error process
- Leveraging external tools and retrieving from external knowledge expand LLM's capabilities
- Agent workflow facilitates complex tasks
 - Task decomposition
 - Allocation of subtasks to specialized modules
 - Division of labor for project collaboration
 - Multi-agent generation inspires better responses

Applications



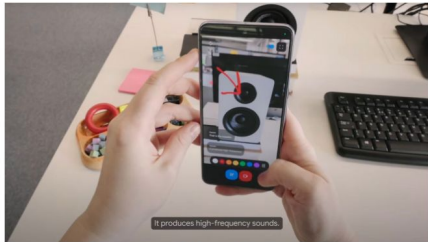
Code generation

Cursor, GitHub Copilot, Devin, Google Jules...



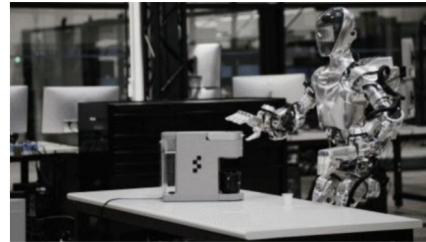
Computer use

Anthropic Claude, Google Jarvis, OpenAI Operator



Personal assistant

Google Astra, OpenAI GPT-4o,...



Robotics

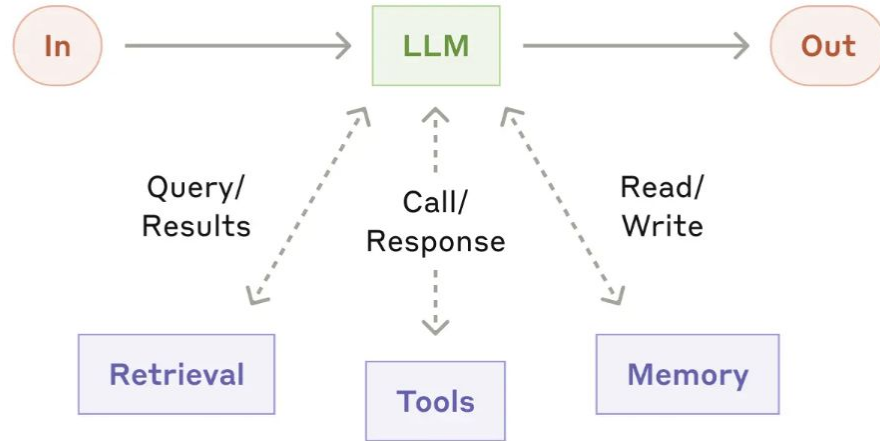
Figure AI, Tesla Optimus, NVIDIA GR00T...

- Education
- Law
- Finance
- Healthcare
- Cybersecurity
- ...

Agenda

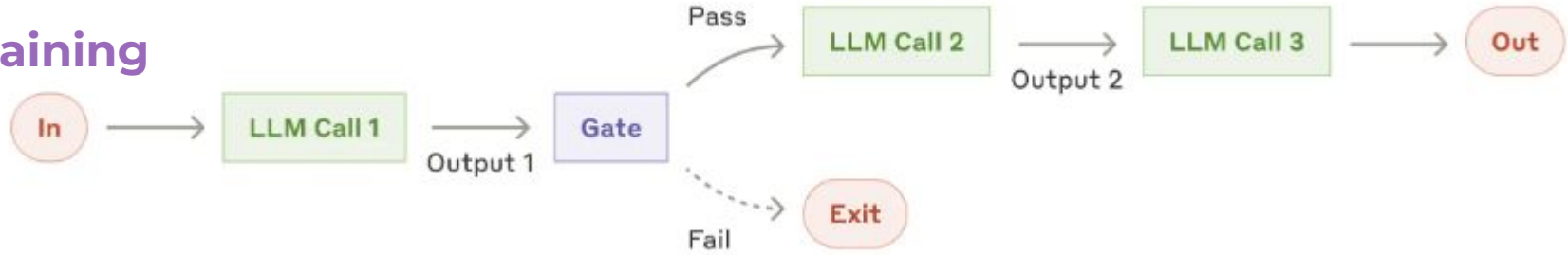
- Introduction
- Building block, workflows, agent
- Cognitive architectures for language agents
- LLM Agent Environments
 - Embodied Agent Interface
 - AgentBoard

Building Block



Workflows

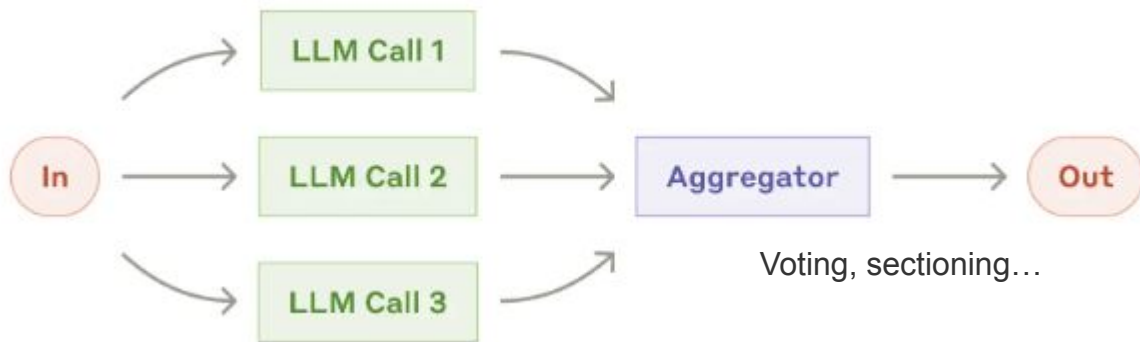
chaining



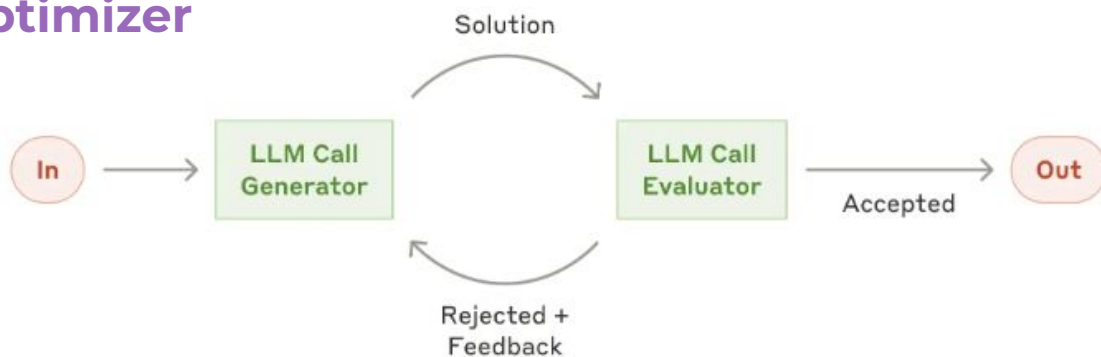
routing



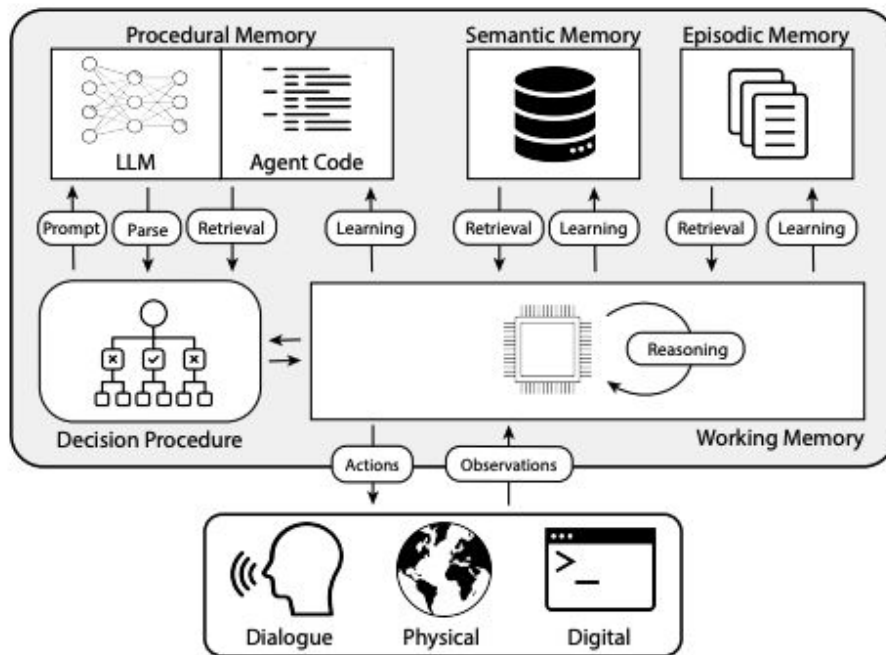
Parallelization



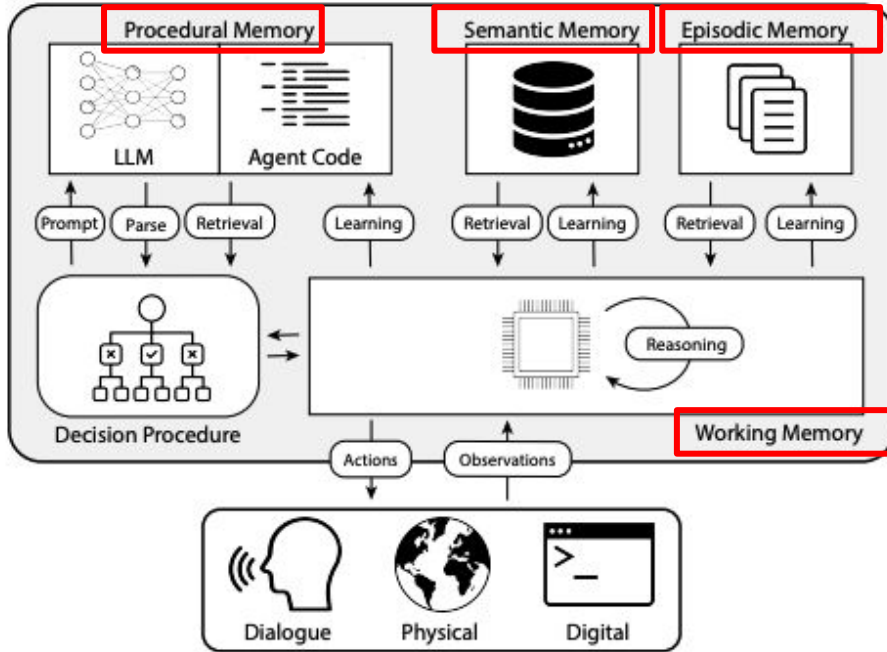
Evaluator-optimizer



Cognitive architectures for language agents



CoALA: Memory



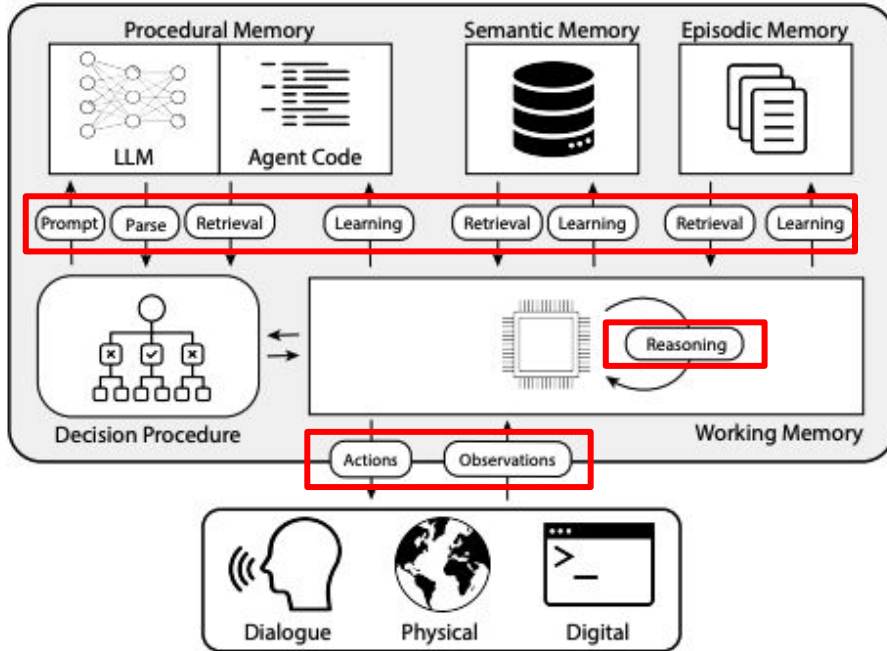
Episodic memory stores experience from earlier decision cycles.

Semantic memory stores an agent's knowledge about the world and itself.

Procedural memory: (i) implicit knowledge stored in the LLM weights; (ii) explicit knowledge written in the agent's code.

Working memory maintains active and readily available information as symbolic variables for the current decision cycle

CoALA: Action

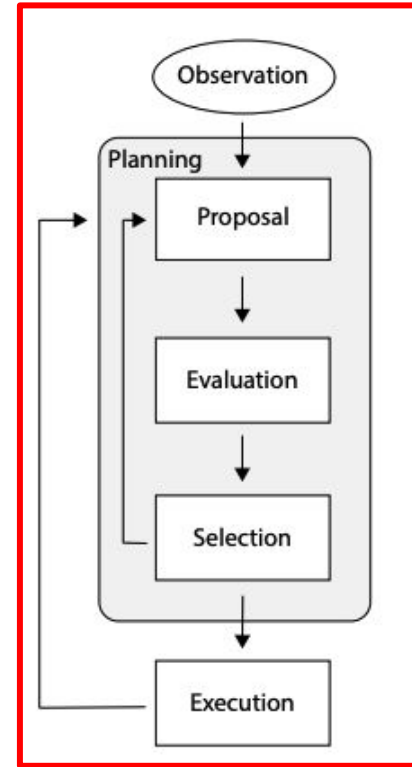
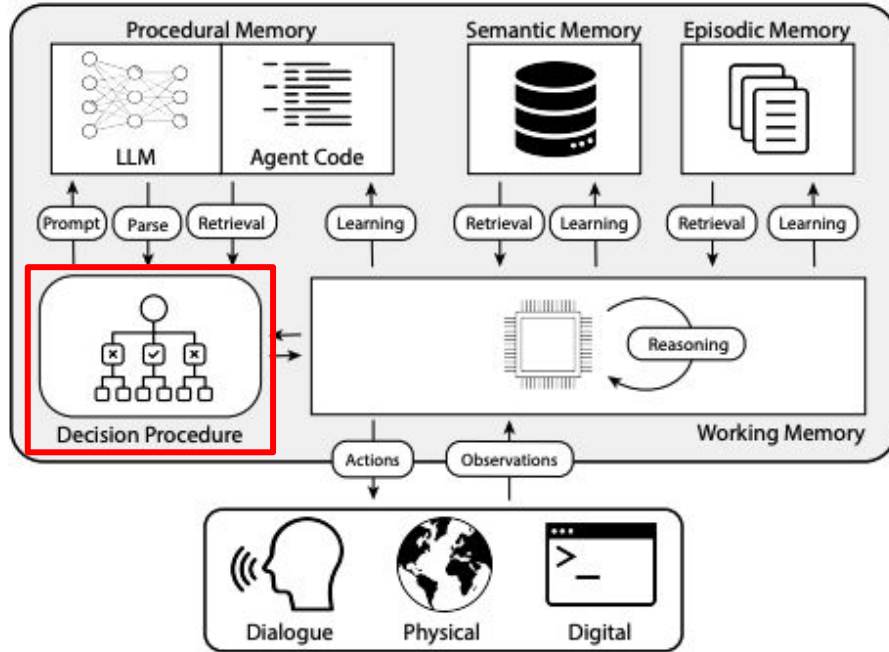


External actions interact with external environments. E.g., control a robot, communicate with a human, navigate a website

Internal actions interact with internal memories.

- retrieval (read from long-term memory)
- learning (write to long-term memory)
- reasoning (update the short-term working memory with LLM)

CoALA: Decision making



Agenda

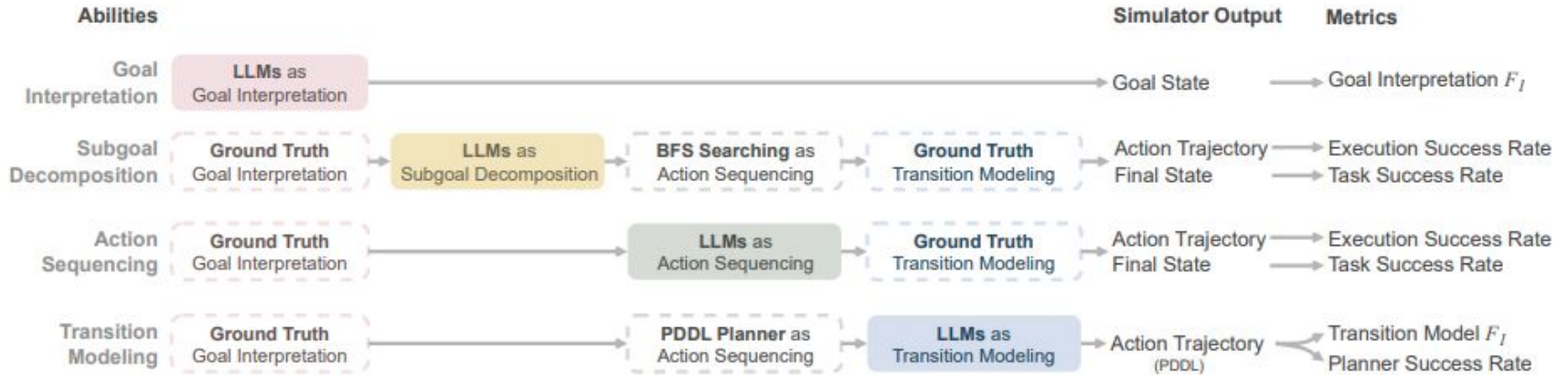
- Introduction
- Building block, workflows, agent
- Cognitive architectures for language agents
- LLM Agent Environments
 - Embodied Agent Interface: Benchmarking LLMs for Embodied Decision Making (NeurIPS 2024) <https://embodied-agent-interface.github.io/>
 - AgentBoard: An Analytical Evaluation Board of Multi-turn LLM Agents (NeurIPS 2024) <https://hkust-nlp.github.io/agentboard/>

Embodied Agent Interface: Benchmarking LLMs for Embodied Decision Making

Unlike existing evaluations rely solely on a final success rate, EAI breaks down the evaluation into

- **four modules** for decision making: goal interpretation, subgoal decomposition, action sequencing, and transition modeling
- a collection of **fine-grained metrics**, such as hallucination errors, affordance errors, various types of planning errors, etc.

Embodied Agent Interface



For each ability module, to provide a comprehensive evaluation for it, we isolate this single module to be handled by the LLMs while using existing data or tools for the other modules.

M1. Goal Interpretation

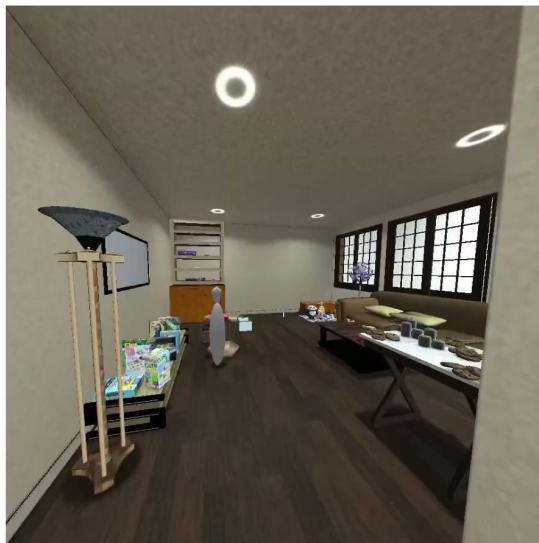
Ground the natural language instruction to the environment representations of objects, states, relations, and actions.

Goal Interpretation

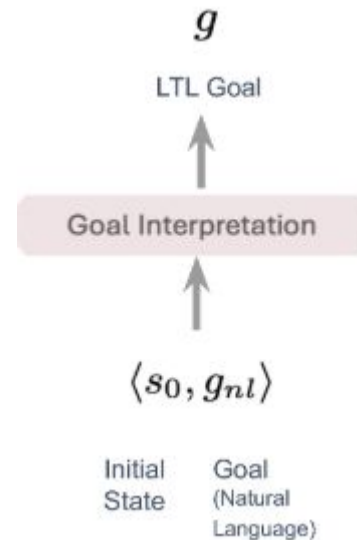
Subgoal Decomposition

Action Sequencing

Transition Modeling



This video is for demonstration only. There're no actual controller-level actions. For action execution examples, visit our repository: <https://github.com/embodyed-agent-interface/embodyed-agent-interface>.



M2. Subgoal Decomposition

Subgoal Decomposition generates a sequence of states, where each state can be a set of objects and their states.

Goal Interpretation **Subgoal Decomposition** Action Sequencing Transition Modeling


Input: Bottling Fruit

Environment

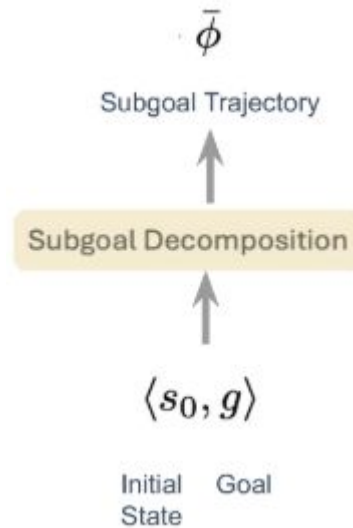
Goal

Subgoal Trajectory

LLM Output



This video is for demonstration only. There're no actual controller-level actions. For action execution examples, visit our repository: <https://github.com/embodyed-agent-interface/embodyed-agent-interface>.



M3. Action Sequencing

Action Sequences are essential to achieve the state transitions identified in Subgoal Decomposition.

Goal Interpretation

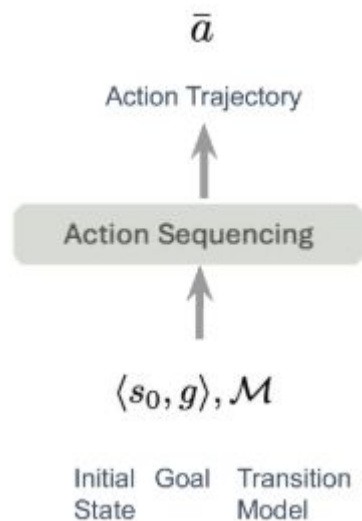
Subgoal Decomposition

Action Sequencing

Transition Modeling



This video is for demonstration only. There're no actual controller-level actions. For action execution examples, visit our repository: <https://github.com/embodied-agent-interface/embodied-agent-interface>.



M4. Transition Modeling

Transition Modeling serves as the low-level controller to guide the simulator in performing state transitions from preconditions to post-effects.

Goal Interpretation

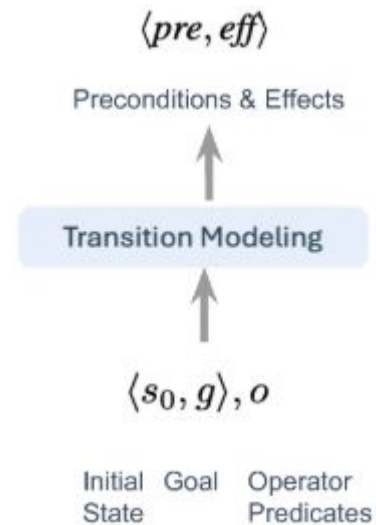
Subgoal Decomposition

Action Sequencing

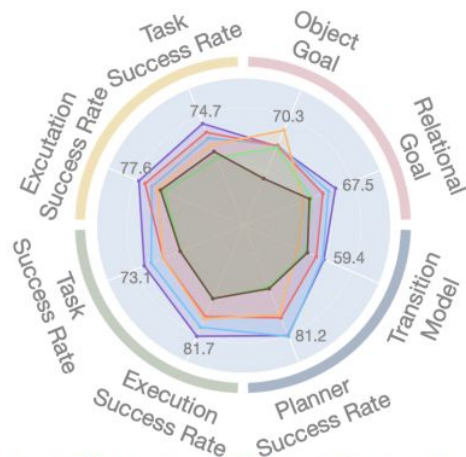
Transition Modeling



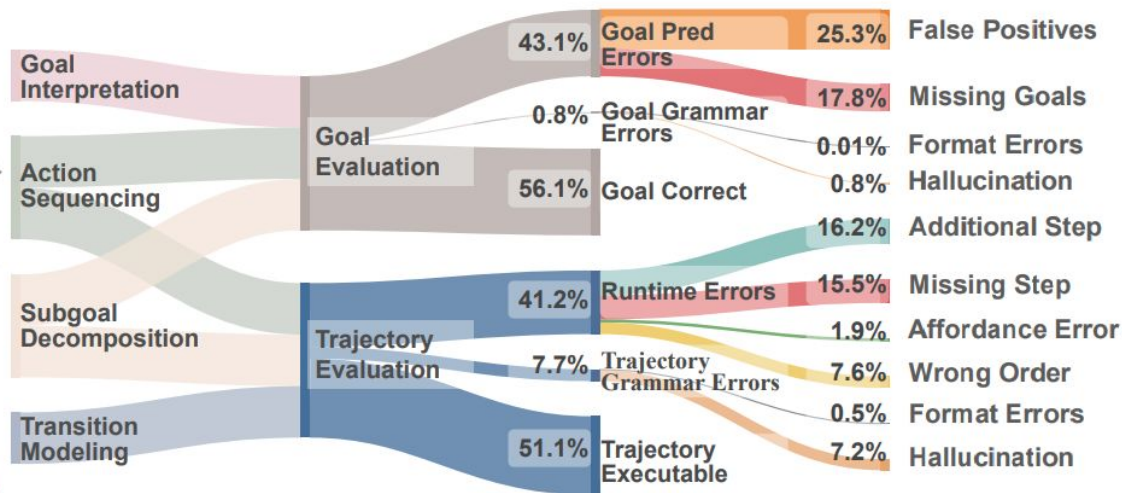
This video is for demonstration only. There're no actual controller-level actions. For action execution examples, visit our repository: <https://github.com/emodied-agent-interface/emodied-agent-interface>.



Evaluation



Gemini1.5-Pro GPT4o Llama3-70B
 Claude3.5-Sonnet o1 Mixtral-8x22B



Grammar Error

Parsing

PLACE_ONFLOOR(floor.0)

✗ Unknown action PLACE_ONFLOOR

Action-Arg Len

GRASP(rag.0, bowl.1)

✗ GRASP only has one param

Hallucination

RINSE(hand.65)

✗ hand.65 is not in the scene

Goal Satisfaction Error

Missing State

Goal

on(television.410) and
facing(agent.65, television.410)

LLM Output

...
FIND(television.410)
SWITCH_ON(television.410)

Error Info: State Unsatisfied

✗ Missing Final State
facing(agent.65, television)

Missing Relation

Goal

next_to(plywood.78, plywood.79) and
next_to(plywood.79, plywood.80)

LLM Output

...LEFT_PLACE_NEXTTO(plywood.79)
LEFT_GRASP(plywood.79)
LEFT_PLACE_NEXTTO(plywood.80)

Error Info: Relation Unsatisfied

✗ Missing Final Relation
next_to(plywood.78,plywood.79)

Missing Goal Action

Goal

TOUCH(cat)

LLM Output

...
FIND(cat.1000)
TURN_TO(cat.1000)

Error Info: Action Unsatisfied

✗ Missing Goal Action
TOUCH(cat.1000)

Trajectory – Runtime Error

Wrong Order

WALK(table.355)
SIT(chair.356)
FIND(novel.1000)
GRAB(novel.1000)



VirtualHome

✗ Precondition
not sitting(agent.65) = False
✓ Historical State
not sitting(agent.65) = True

Missing Step

...
CLOSE(fridge.0)
SLICE(strawberry.0)
SLICE(peach.0)



BEHAVIOR

✗ Precondition
holding(knife.0) = False
✗ Historical State
holding(knife.0) = False

Affordance Error

LEFT_RELEASE
OPEN(shelf.16)
LEFT_RELEASE
LEFT_GRASP(pool.50)



BEHAVIOR

✗ Precondition
shelf.16 not openable
✗ Precondition
pool.50 not grabbable

Additional Step

OPEN(top_cabinet.27)
RIGHT_GRASP(soap.79)
...
OPEN(top_cabinet.27)

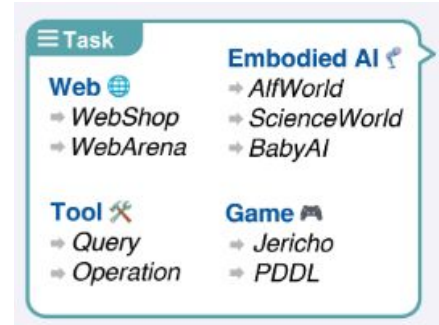


BEHAVIOR

✗ Current State
open(top_cabinet.27) = True
! Expected State
open(top_cabinet.27) = False

AgentBoard: An Analytical Evaluation Board of Multi-turn LLM Agents

- **Task diversity** is necessary to cover various agent tasks such as embodied, web, and tool agents.
- **Multi-round** interaction is critical to mimic realistic scenarios.
- Evaluating agents in **partially-observable** environments, where they must actively explore to understand their surroundings, is essential for practical assessments.



AlfWorld

Interactive text environment that require agents to explore surroundings and perform commonsense tasks like “put two soap bars in garbagecan”.



Goal: heat some apple and put it in fridge.

> check valid actions

Choose an action from these valid actions: go to cabinet 1, go to cabinet 2, go to cabinet 3, ...

> go to fridge 1

The fridge 1 is closed.

> open fridge 1

You open the fridge 1. The fridge 1 is open. In it, you see a apple 1, a bowl 3, a cup 2, a cup 1, a egg 3, a lettuce 1, a potato 2, a potato 1, and a tomato 1. (reward: 0.25)

...

ScienceWorld

Interactive text environment testing scientific commonsense, e.g. “measure the melting point of the orange juice”.

Goal: Your task is to find the animal with the longest life span.

...

> go to outside

You move to the outside. (reward: 0.3333333333333333)

> look around

This outside location is called the outside. Here you see: the agent a substance called air an axe a baby brown bear ...

> focus on baby brown bear

You focus on the baby brown bear. (reward: 0.6666666666666666)

...



BabyAI

Interactive 20x20 grid environment where agents navigate and interact with objects within a limited sight range.

Goal: Open the red door, and open the blue door

> check valid actions

You can take the following actions: turn left, turn right, move forward, toggle and go through blue closed door 1, go to blue closed door 1, check available actions

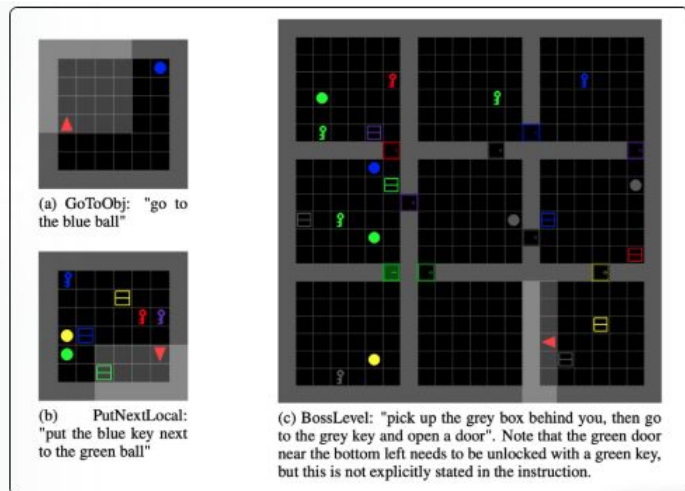
> go to blue closed door 1

In front of you in this room, you can see several objects: There is a blue closed door 1 right in front of you 1 steps away. The room has walls around you. You are not carrying anything. (reward: 0.25)

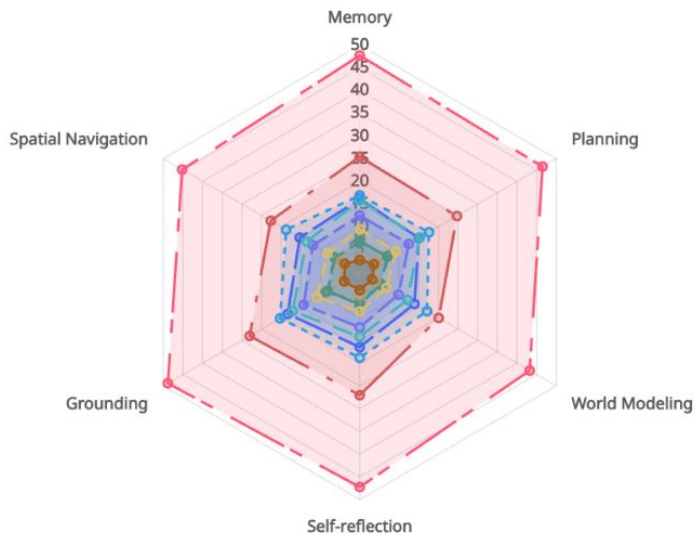
> toggle and go through blue door 1

The action is not recognized. Please check valid actions.

...



Evaluation



	AlfWorld	ScienceWorld	BabyAI
Memory			
1. Could finish tasks within 2k tokens	1	2	1
2. Could finish task within 4k tokens			
3. Otherwise			
Planning			
1. ≤ 3 subgoals on average	1	2	2
2. ≤ 5 subgoals on average			
3. Otherwise			
World Modeling			
1. Requires no additional knowledge other than instruction			
2. Requires knowledge of the environment from exploration	3	3	2
3. Requires commonsense knowledge in addition to knowledge from environment			
Self-Reflection			
1. Detailed feedback and error message with instruction for the next step.			
2. Not very detailed feedback and error message	3	2	2
3. No error message, e.g. "no change in state"			
Grounding			
1. No specific action format is required, could recognize similar actions	2	3	2
2. Action format is required			
3. Action format hard to follow			
Spatial Navigation			
0. No spatial navigation	1	1	1
1. 2D navigation			