NYU

Ego4D

Advanced Topics in Embodied Learning and Vision

Ying Wang

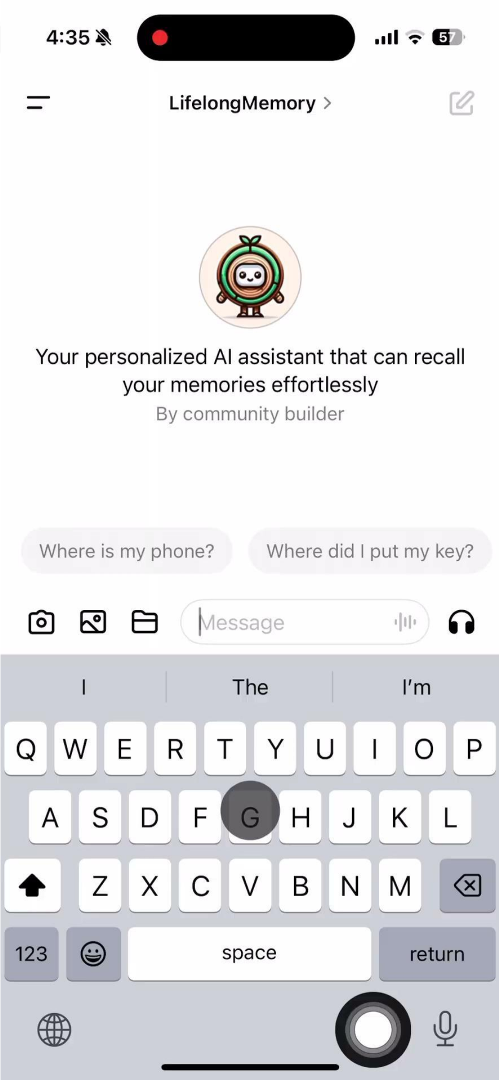
2025.02.13

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

Meta

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. [Greene] HPC team has stored Ego4D data under /vast/work/public/ml-datasets/ego4d

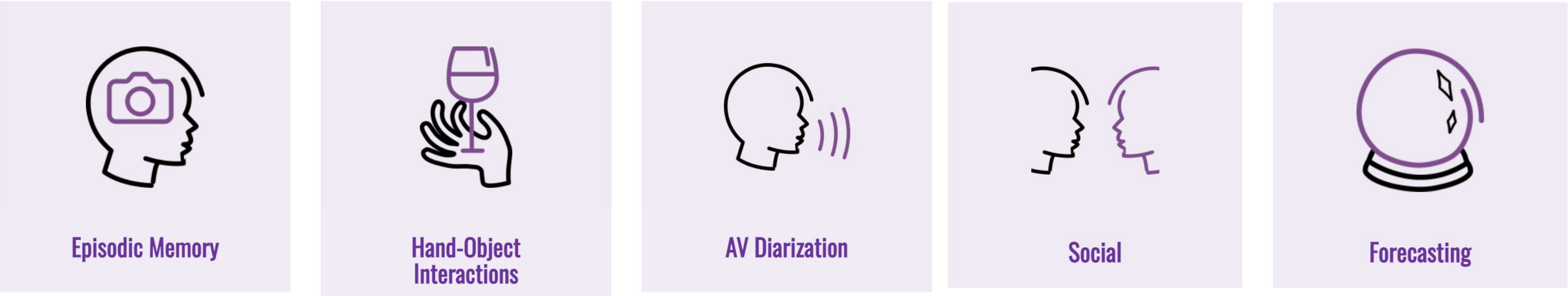
   Unfortunately, the file system in Burst is independent of Greene. For small datasets like EgoSchema, you can download a copy by yourself in Burst. If you need to access the full video dataset, we recommend using Greene.

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



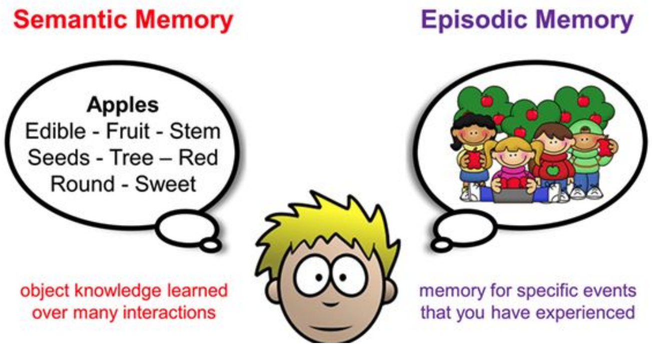
Episodic memory

- **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?



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.

Video clip

+

+

Query frame: 14357

Most recent occurrence (frame number + bb box)

[VQ2D example in visualizer]

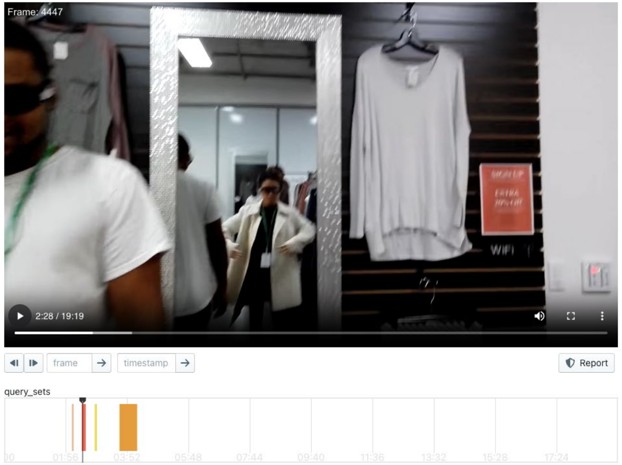
NLQ

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.

Video clip

+

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

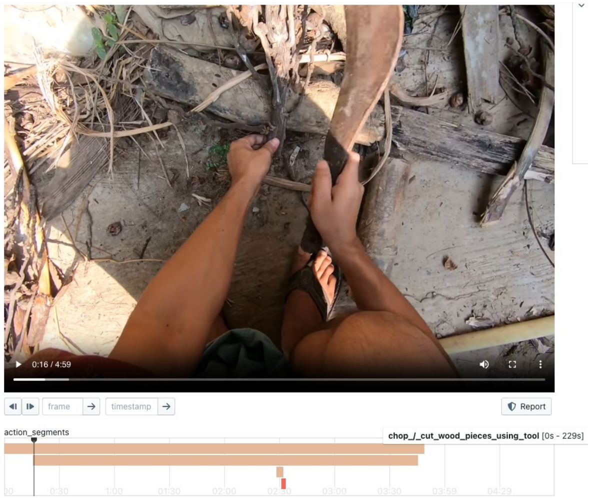
Moments queries: Given an input video and a query action category, the goal is to retrieve all the instances of this action category in the video. Specifically, it poses the request 'Retrieve all the moments that I do X in the video.', where X comes from a pre-defined taxonomy of action categories.

Video clip

+

Moment:
chop_/_cut_wood_pieces
_using_tool

[MQ example in visualizer]



All temporal windows of the given activity

0:00 - 03:48 chop_

0:15 - 03:45 chop_

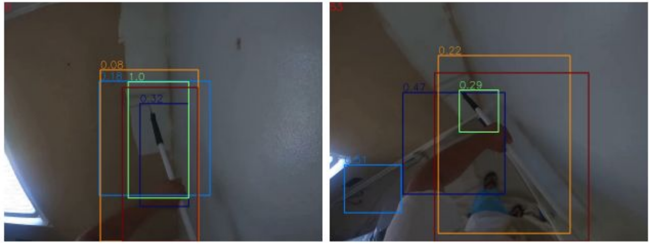
02:28 - 02:31 chop

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.

Video clip

+

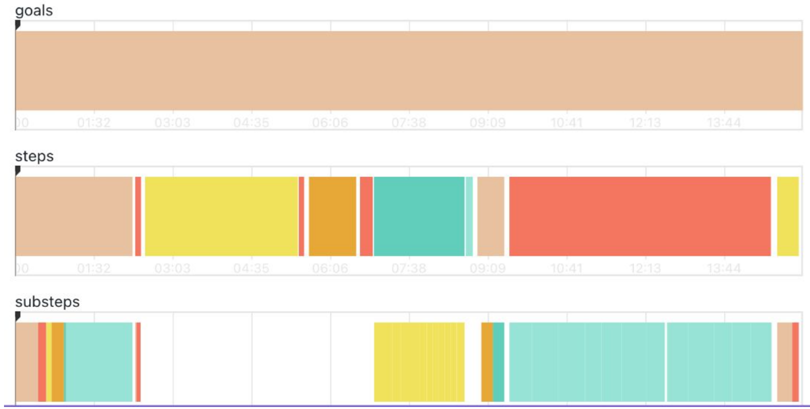


Bounding box + confidence

[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 keystep 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]

Ego4D Challenges (2024)

Episodic memory:

- Visual queries with 2D/3D localization (VQ2D/VQ3D)
- Natural language queries (NLQ)
- Moments queries (MQ)
- EgoTracks
- Goal Step
- Ego Schema

Hands and Objects:

- Temporal localization

Audio-Visual Diarization:

- Localization and Tracking
- Speech transcription

Social Understanding:

- Looking at me
- Talking to me

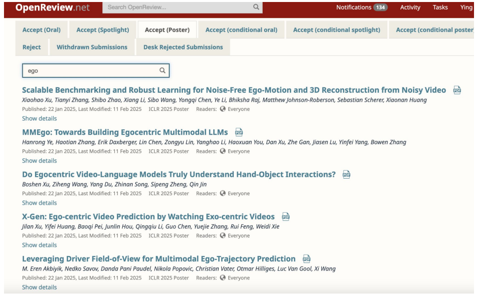
Forecasting:

- Short-term hand object prediction
- Long-term activity prediction

You can also use the Ego4D data to study a novel problem!

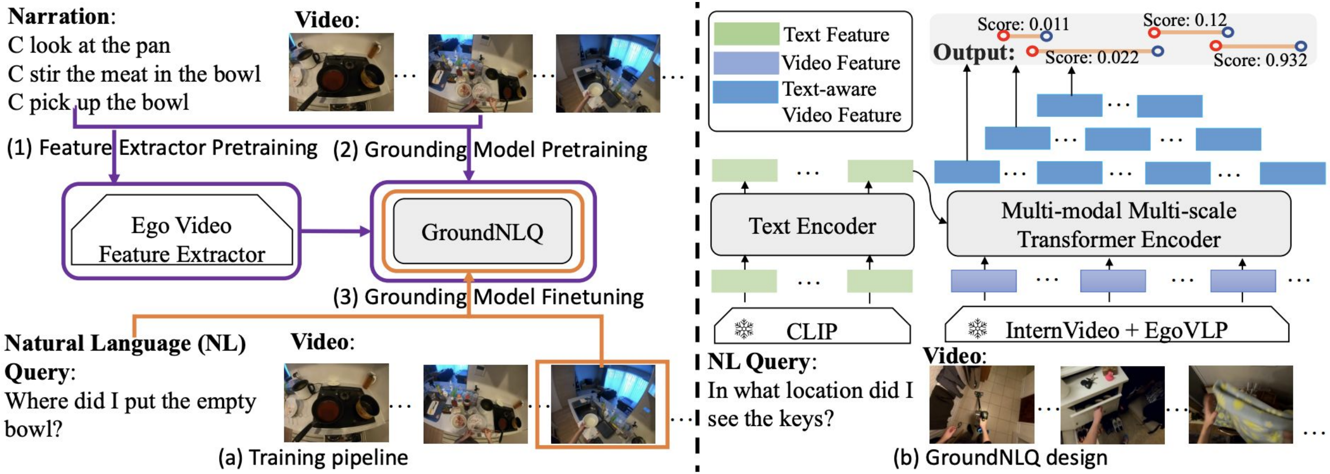
More ego models and data!

- **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/
- …



Video features

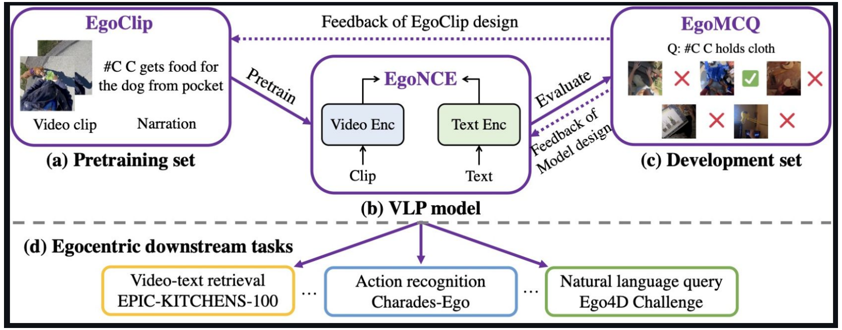
1. Extract video features (slowfast, omnivore…) and text features using pretrained encoders
2. Fuse features and feed them into a base video model



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

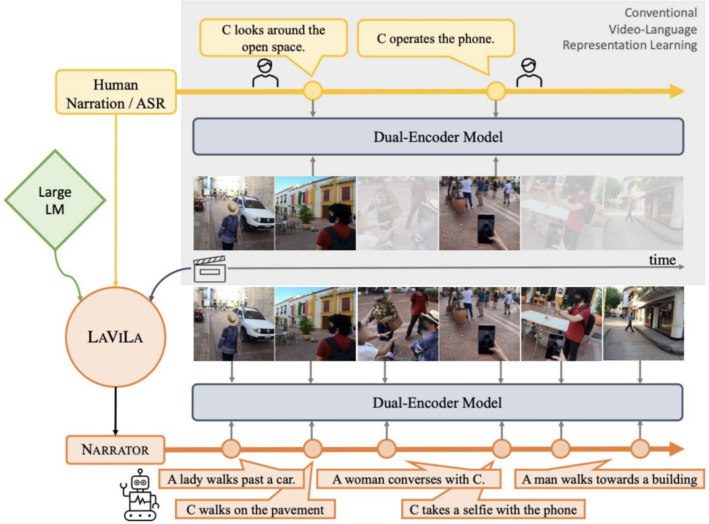
Ego foundation models

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



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

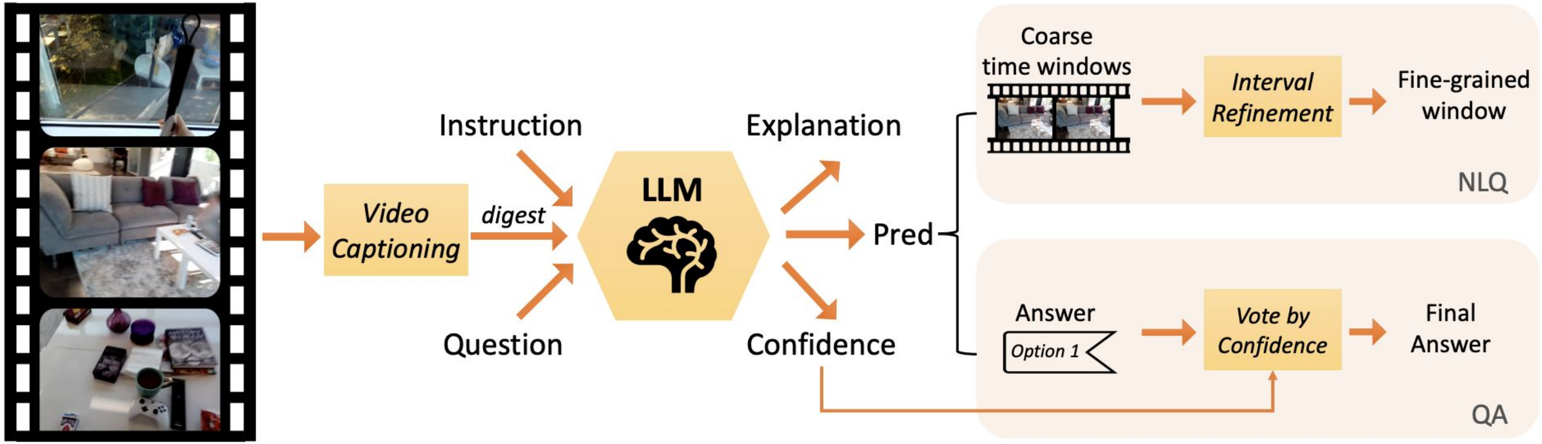
LaViLa: Use LLMs to densely narrate long videos, then use those narrations to train a dual-encoder.



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.



https://agenticlearning.ai/lifelong-memory/

Demo!

Ego4D quick start