Video Understanding in Egocentric Vision

Simone Alberto Peirone

PhD Student @ Politecnico di Torino, Italy

March 2025

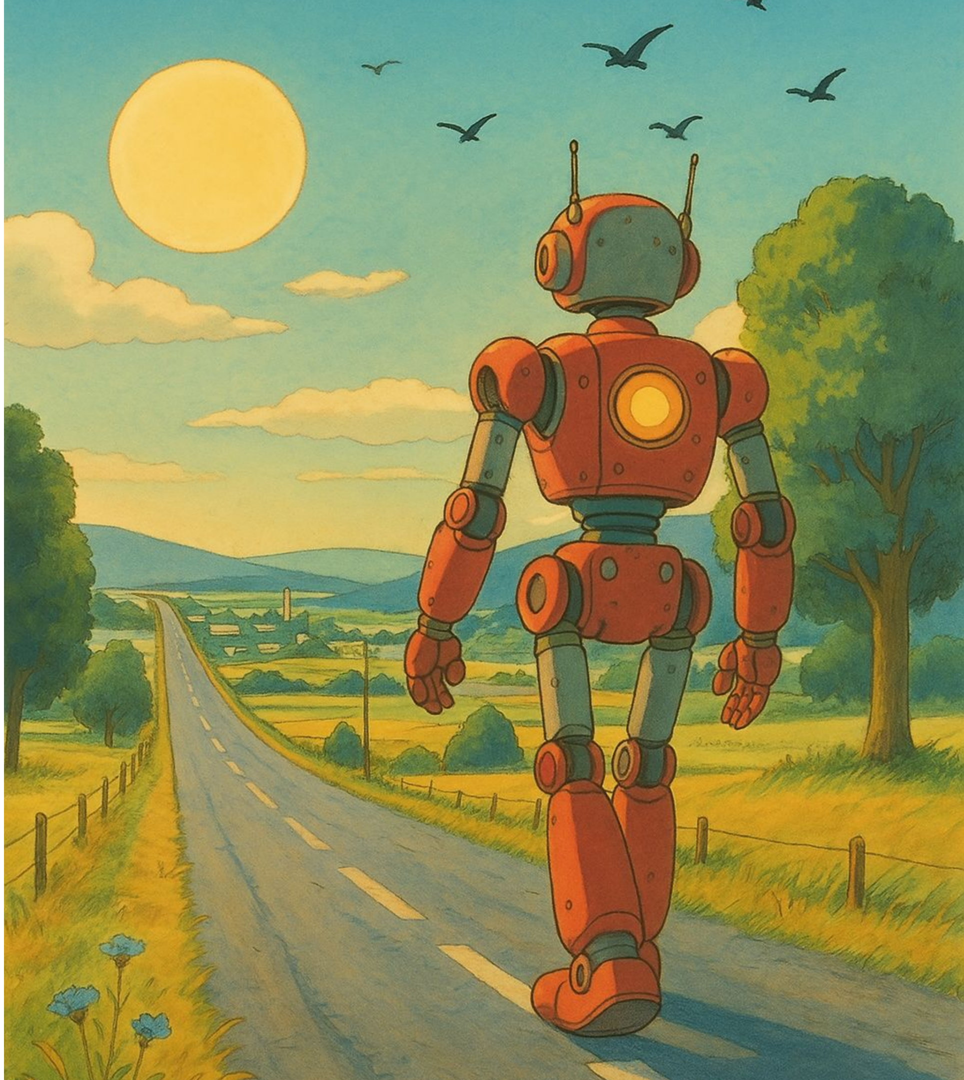
Politecnico di Torino

1859

ellis

European Laboratory for Learning and Intelligent Systems

VANDAL



What is Egocentric Vision?

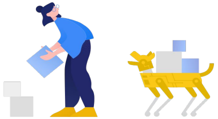
Computer vision from a **human-centric point of view**

Head-mounted cameras

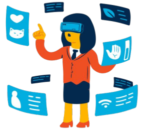
Close to the **action**

…with lots of applications:

Assistive robotics

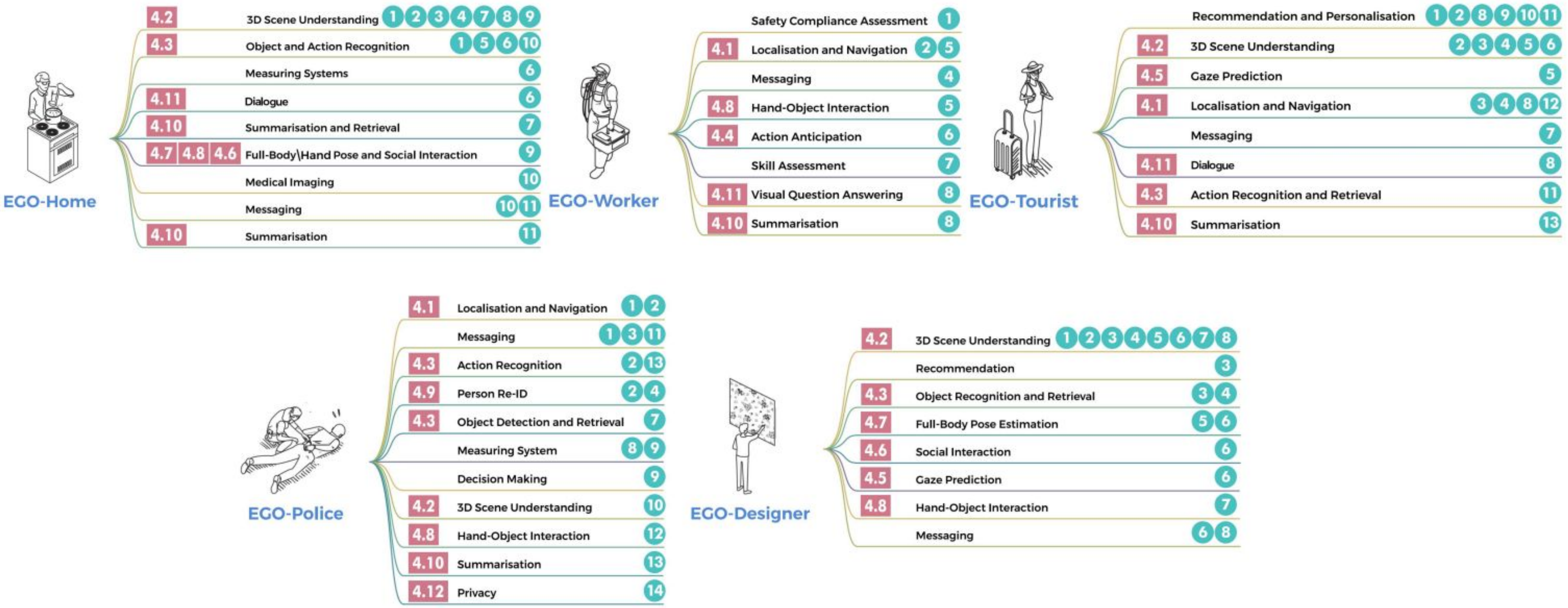


Industrial applications



Augmented reality

Egocentric Vision is pervasive



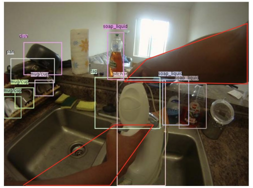
Plizzari, Chiara, et al. "An outlook into the future of egocentric vision." IJCV 2024

The “*Image-Net moment*” in Egocentric Vision

Big focus on cooking as it’s
a very complex human activity

“once we solve cooking, we will have solved video understanding in general”
Dima Damen, EgoVis workshop @ CVPR 2024

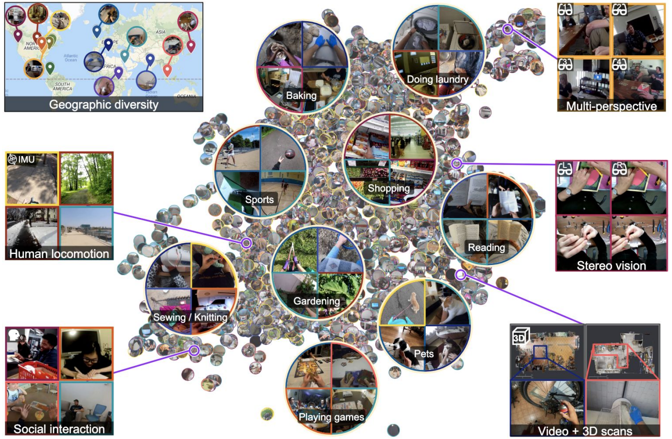
ADL



EGTEA

EPIC-Kitchens

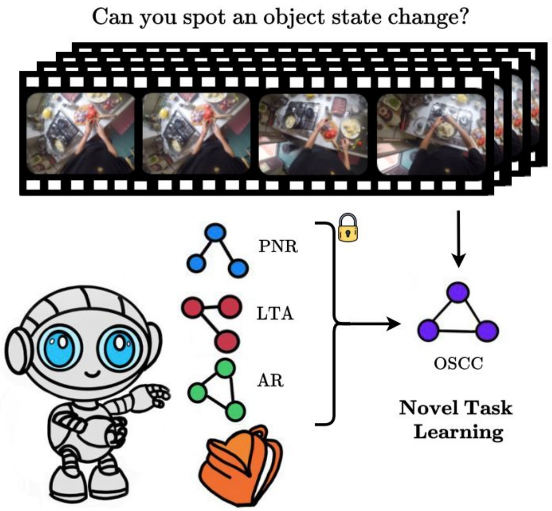
EGO4D

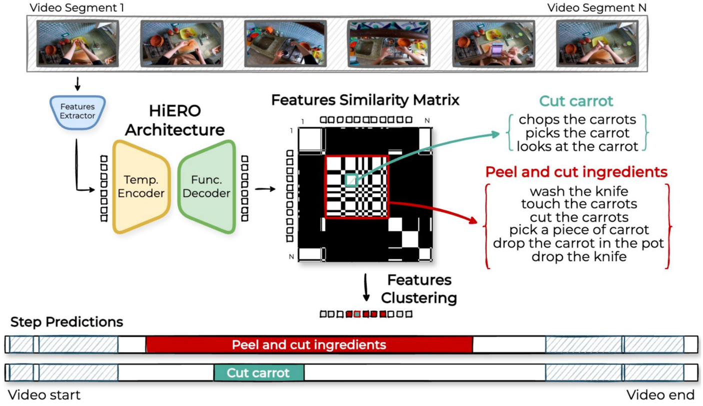


2012 | 2015 | 2018 | 2022

Pirsiavash *et al*. "Detecting activities of daily living in first-person camera views." CVPR2012
Li *et al*. "In the eye of beholder: Joint learning of gaze and actions in first person video." ECCV 2018
Damen *et al*. "Scaling egocentric vision: The epic-kitchens dataset." *ECCV 2018*
Grauman *et al*. "Ego4d: Around the world in 3,000 hours of egocentric video." *CVPR 2022*

The plan for this talk





Learning about human activities from different *perspectives*

PAPERS

1. A backpack full of skills: Egocentric Video Understanding with Diverse Task Perspectives

2. Hier-EgoPack: Hierarchical Egocentric Video Understanding with Diverse Task Perspectives

Learning the *hierarchy* behind human activities

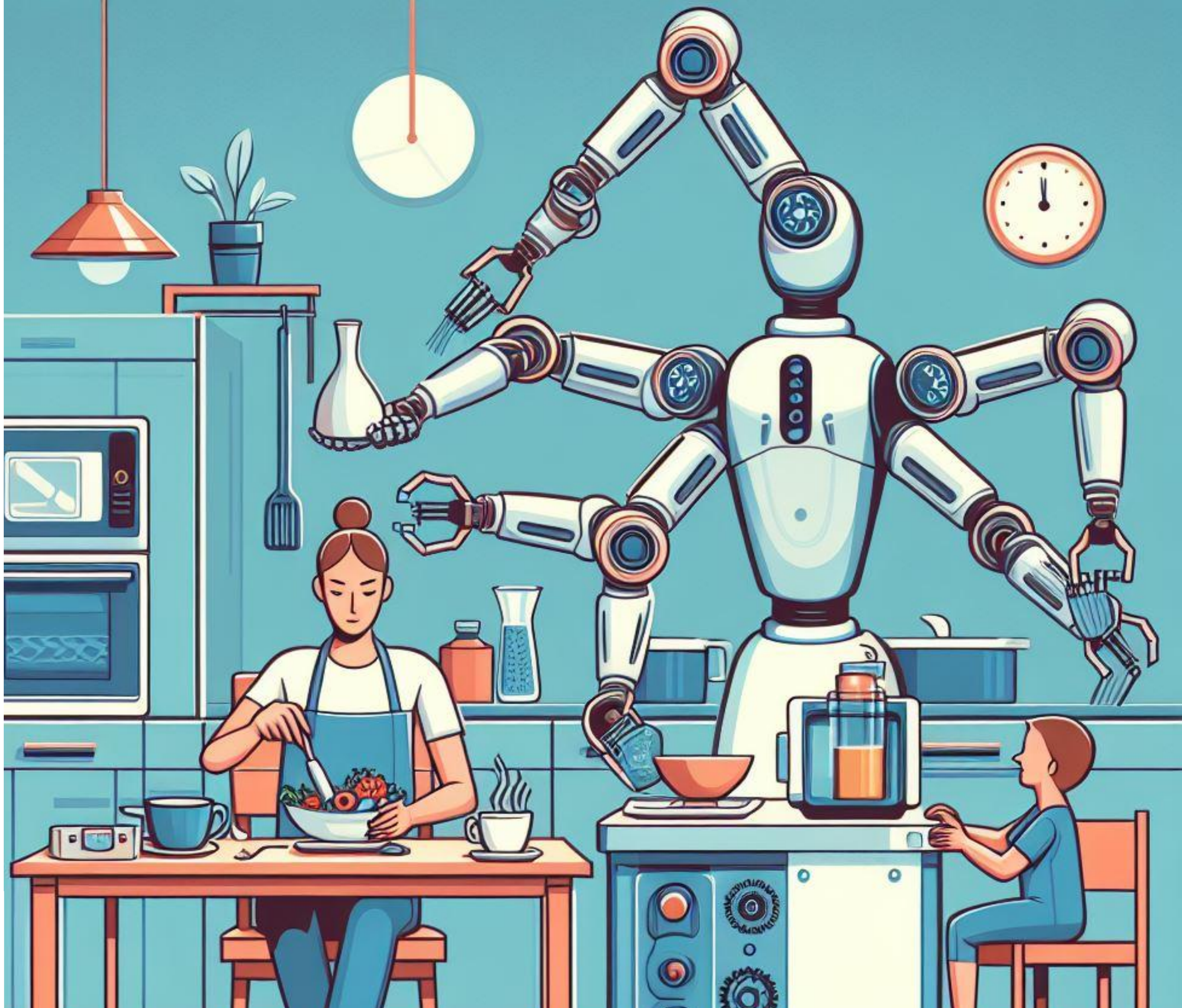
PAPERS

1. HiERO: understanding the hierarchy of human behavior enhances reasoning on egocentric videos

A backpack full of skills: Egocentric Video Understanding with Diverse Task Perspectives

CVPR 2024

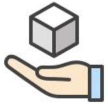
Simone Alberto Peirone, Francesca Pistilli, Antonio Alliegro, Giuseppe Averta



<u>What</u> can we learn from a single video?

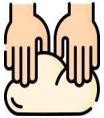
Different video tasks = different, possibly complementary, perspectives

Actions Recognition (AR)



Object State Change Classification (OSCC)

Long Term Action Anticipation (LTA)



Point of No Return (PNR)

Human-Object Interaction (HOI) Tasks from Ego4D

Action Recognition (AR)

"Knead dough"

Given an short clip, predict the action being performed

Long Term Anticipation (LTA)

Given an input video, predict the next K actions the person will perform

Object State Change Classification (OSCC) / Point of No Return (PNR)

State-change: Wood smoothed

State-change: Plant removed from ground

OSCC: predict <u>if</u> there is an object state change in the video.

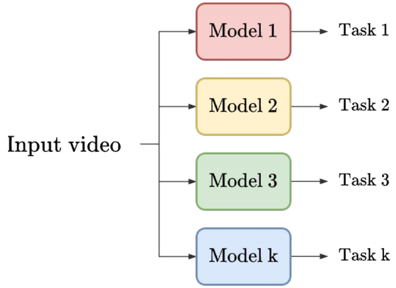
PNR: predict the <u>timestamp</u> of the state change.

Grauman, Kristen, et al. "Ego4d: Around the world in 3,000 hours of egocentric video." *CVPR 2022*

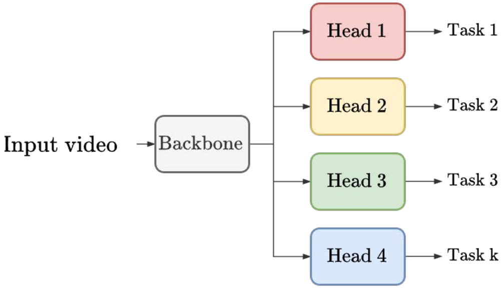
How can we learn from these perspectives?

Main approaches from the literature:

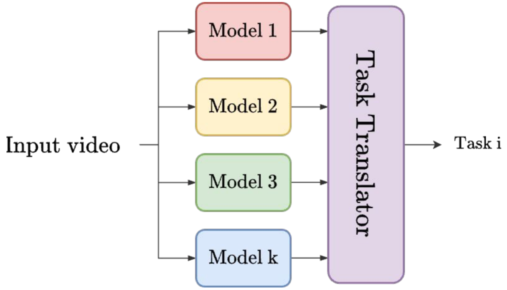
Single Task models



Multi-Task Learning



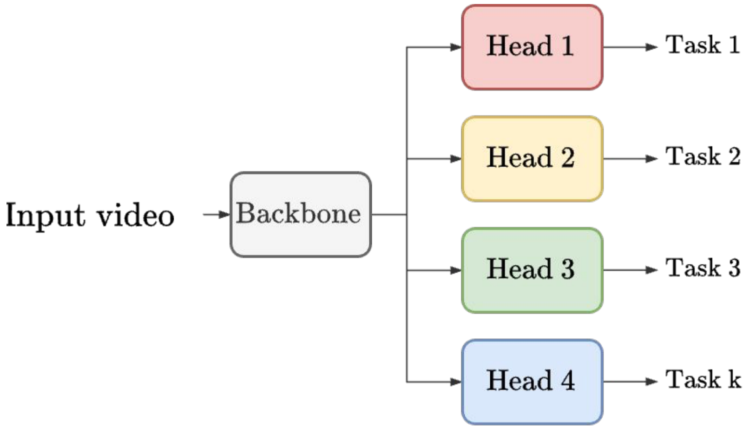
Task Translation



<u>How</u> can we learn from videos? - Multi-Task Learning

Jointly learn multiple tasks using a shared backbone and a set of task-specific heads

- + **Same model is shared across different tasks**
- - **Does not *<u>explicitly</u>* model task sinergies**
- - **May suffer of negative transfer between tasks**

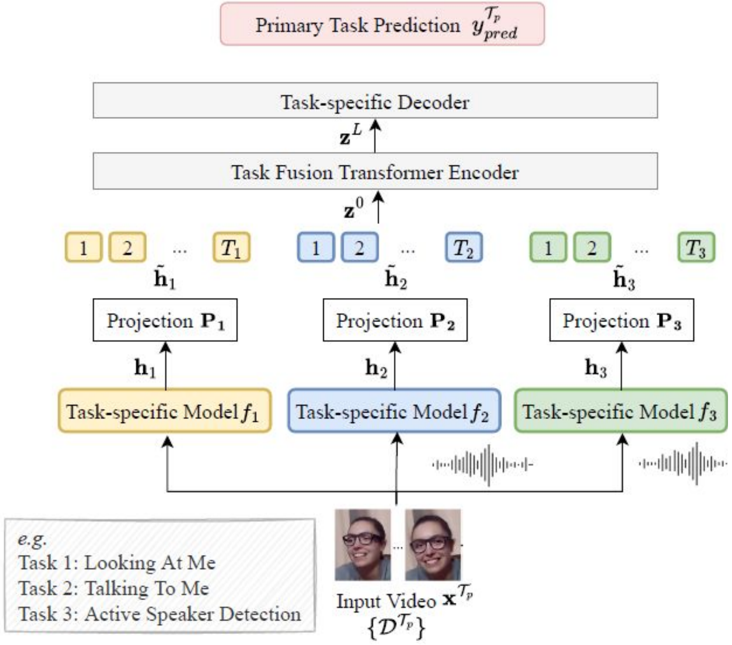


How can we learn from videos? - Cross-Task Translation

EgoT2 proposes an innovative approach to leverage cross-task sinergies by learning to "translate" features across different tasks

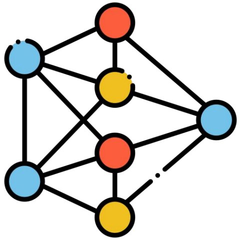
+ **Combine perspectives from different tasks**

- **Need to know all the tasks before-hand**
- **One model for each task**



Xue, Zihui, et al. "Egocentric Video Task Translation" (CVPR 2023)

A new paradigm for Egocentric Video Understanding



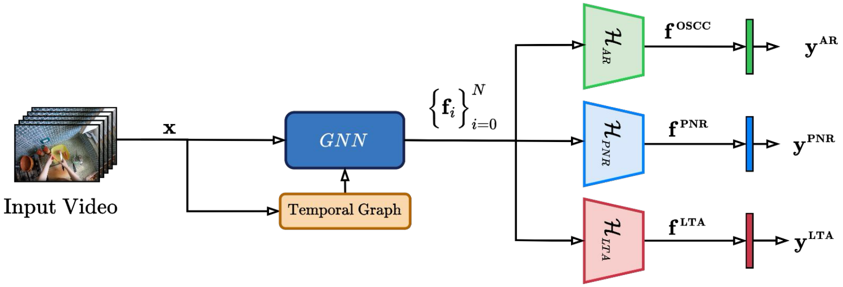
Shared model
for all the tasks

Knowledge reuse
across tasks

Outperform single and
multi-task baselines

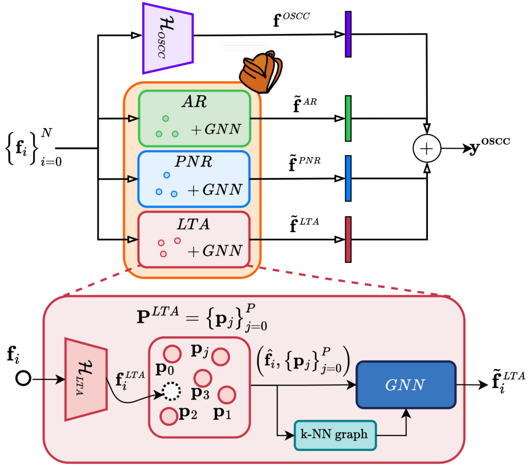
The EgoPack approach

Step 1: MTL Pre-training step



Multi-task pre-training
on a set of known task

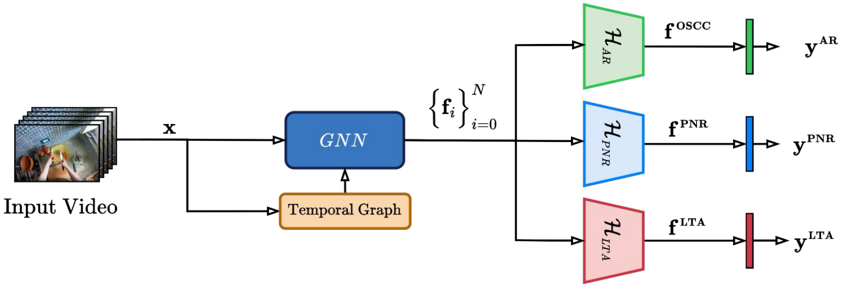
Step 2: Novel Task Learning



Fine-tuning on a novel task
with EgoPack's cross-task interaction

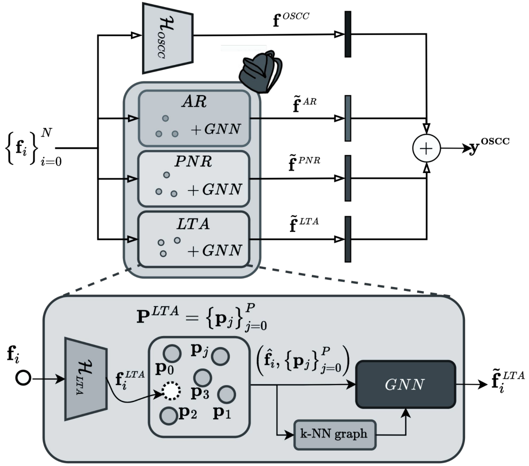
The EgoPack approach

Step 1: MTL Pre-training step



Multi-task pre-training
on a set of known task

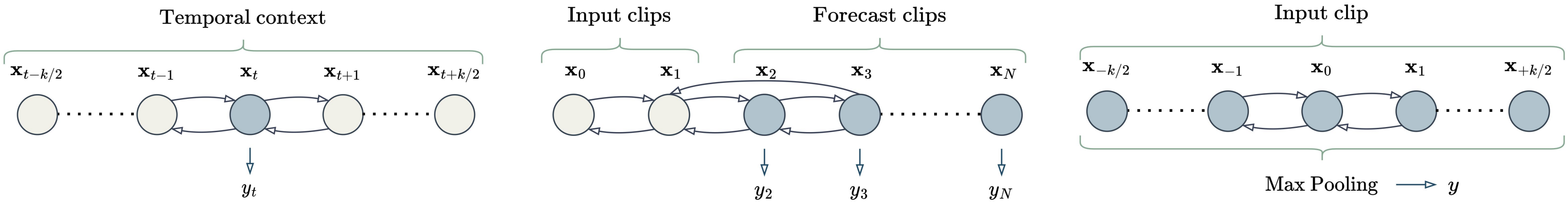
Step 2: Novel Task Learning



Fine-tuning on a novel task
with EgoPack's cross-task interaction

Step 1: A graph-based temporal model

We can model many egocentric vision tasks with a shared graph-based structure…

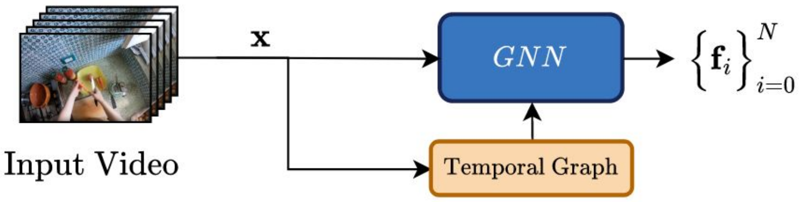


Node Classification (AR, PNR) **Future Node Classification** (LTA) **Graph Classification** (OSCC)

Each node is a temporal segment and

egocentric video tasks become different graph operations

Step 1: Temporal Multi-Task Pre-Training



Input nodes are time segments of the video

$$\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \quad \text{with} \quad \mathbf{x}_i \in \mathbb{R}^D$$

Edges connect temporally close nodes depending on the task

Temporal Reasoning using message passing

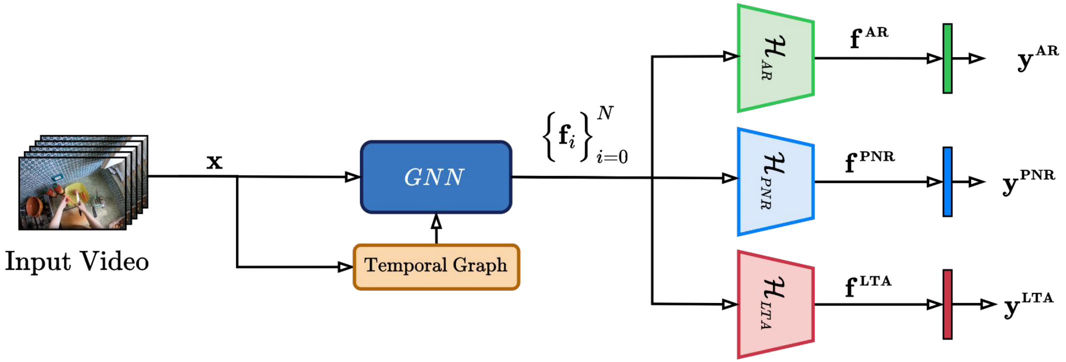
$$\mathbf{g}_i^{(l+1)} = \underset{\mathbf{f}_j \in \mathcal{N}_i}{mean}\left(\phi\left(\mathbf{W}_p^{(l)}\mathbf{f}_j^{(l)} + \mathbf{b}_p^{(l)}\right)\right)$$

$$\mathbf{f}_i^{(l+1)} = \mathbf{W}_r^{(l)}\mathbf{f}_i^{(l)} + \mathbf{W}^{(l)} \cdot \mathbf{g}_i^{(l+1)} + \mathbf{b}^{(l)}$$

This design unifies all tasks under a **shared temporal modelling**

Hamilton et al. "Inductive representation learning on large graphs." NeurIPS 2017

Step 1: Temporal Multi-Task Pre-Training

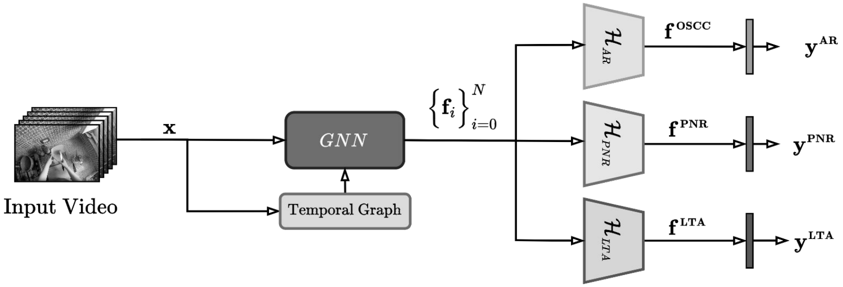


The output of the Temporal Model is specialized into task-specific features using a set of **task-specific heads**

The output are the task logits $\mathbf{y}_i^k \in \mathbb{R}^{D_o^k}$

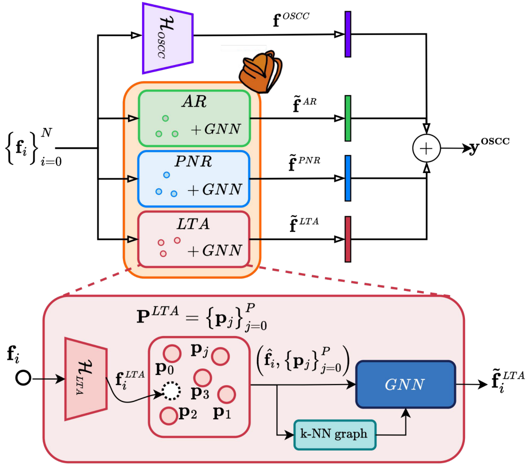
The EgoPack approach

Step 1: MTL Pre-training step



Multi-task pre-training
on a set of known task

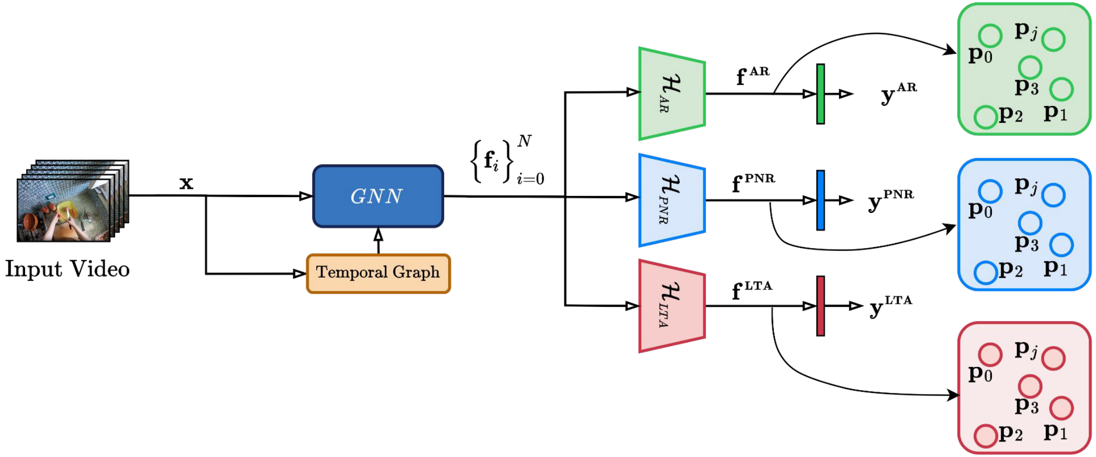
Step 2: Novel Task Learning



Fine-tuning on a novel task
with EgoPack's cross-task interaction

Step 2: Novel Task Learning with EgoPack

Given as input the same video, the model's heads express **different and complementary perspectives** on the content of the video



Step 2.1: Prototypes collection

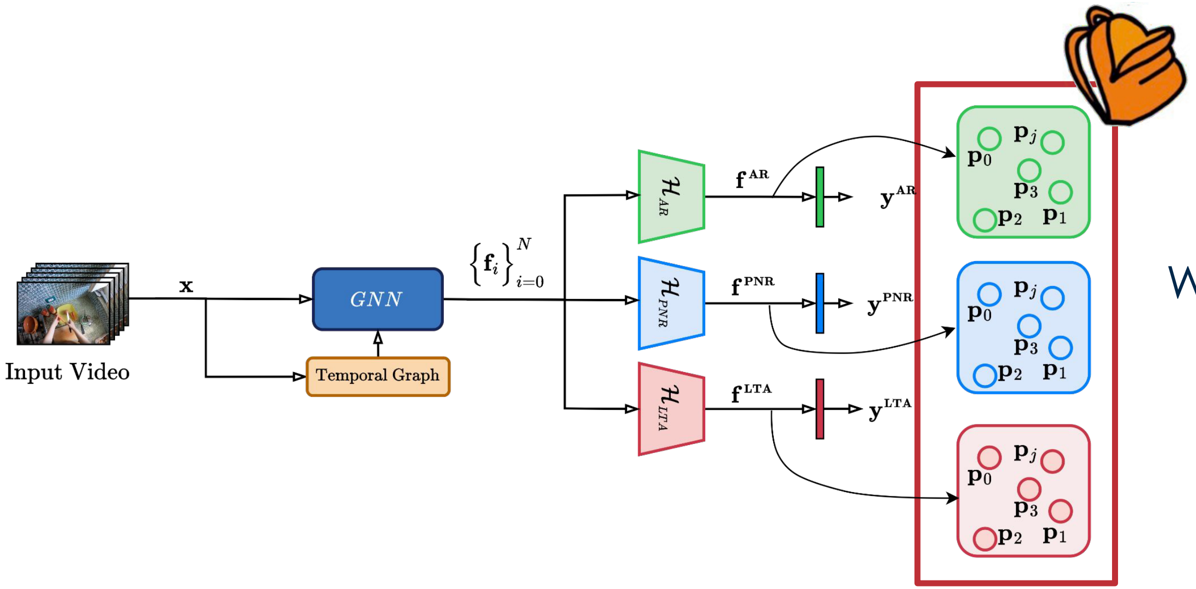
We collect action-wise **task-specific prototypes** by feeding the model with AR videos

$$\mathbf{P}^k = \{\mathbf{p}_0^k, \mathbf{p}_2^k, \ldots, \mathbf{p}_P^k\} \in \mathbb{R}^{P \times D_k}$$

for each task $\mathcal{T}_k$

Step 2: Novel Task Learning with EgoPack

Given with the same video, the model's heads express **different and complementary perspectives** on the content of the video

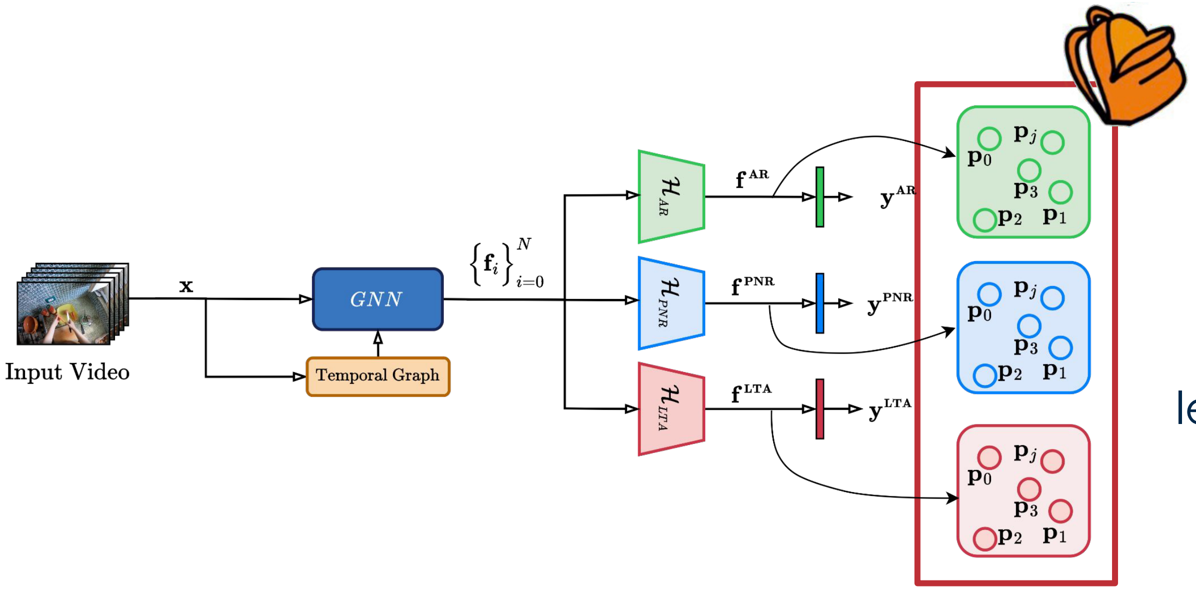


Step 2.1: Prototypes collection

We call these prototypes a ***"backpack of skills"***

Step 2: Novel Task Learning with EgoPack

Given with the same video, the model's heads express **different and complementary perspectives** on the content of the video

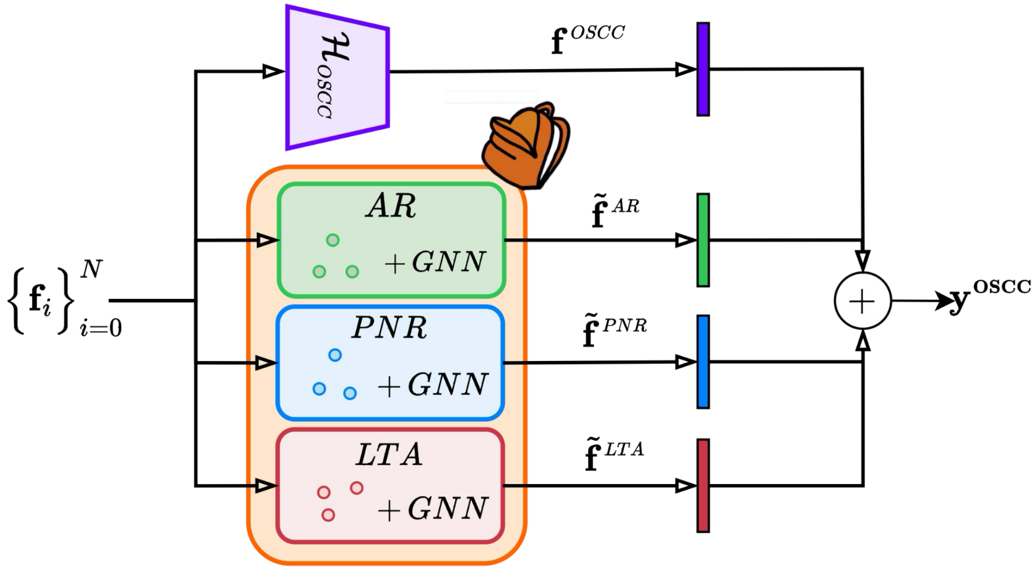


Step 2.1: Prototypes collection

They represent the **previous experience** learnt in the pre-training phase

Step 2: Novel Task Learning with EgoPack

To learn a **novel task**, e.g., **Object State Change Classification**,
we add the corresponding head and exploit the synergies with the previous tasks.

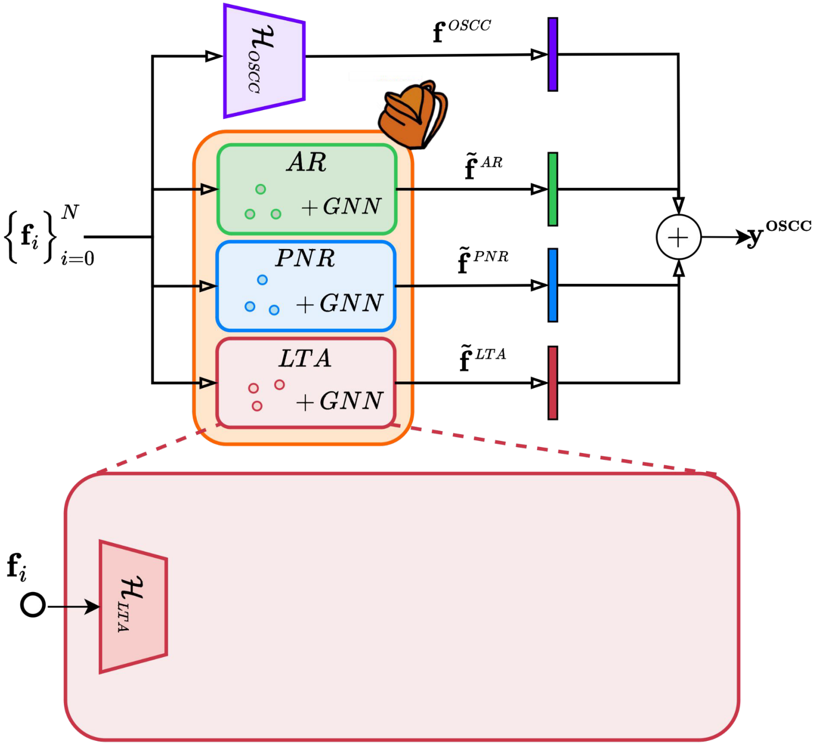


New head trained for the novel task (OSCC)

Knowledge reuse from previous tasks using the prototypes from AR, PNR and LTA

Step 2: Novel Task Learning with EgoPack

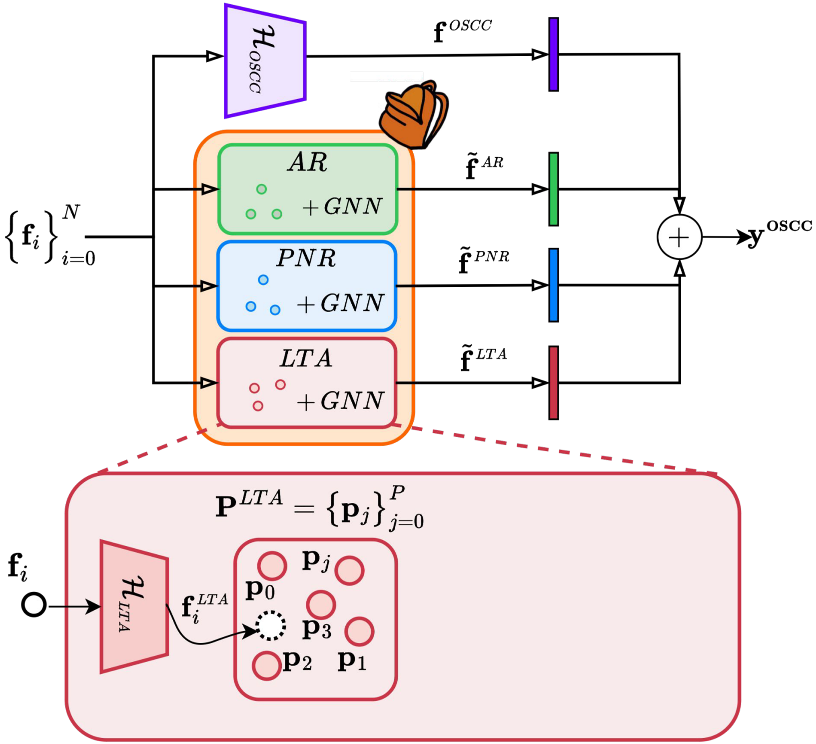
When learning a novel task, **we feed the temporal features through the task-specific heads** of the K pre-training tasks to obtain f_i^k.



Step 2: Novel Task Learning with EgoPack

When learning a novel task, **we feed the temporal features through the task-specific heads** of the pre-training tasks to obtain f_i^k.

These features act as queries to look for the **closest matching prototypes** using k-NN in the features space.



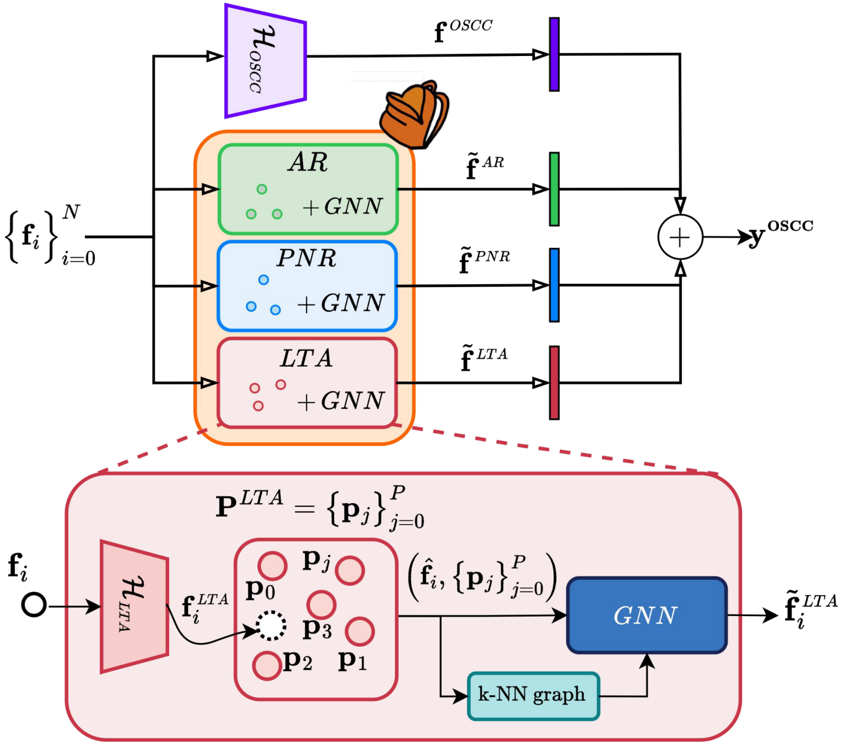
Step 2: Novel Task Learning with EgoPack

When learning a novel task, **we feed the temporal features through the task-specific heads** of the pre-training tasks to obtain f_i^k.

These features act as queries to look for the closest matching prototypes using k-NN in the features space.

We refine the task features using **Message Passing with task prototypes**.

$$\mathbf{f}_i^{(l+1),k} = \mathbf{W}_r^{(l)} \mathbf{f}_i^{(l),k} + \mathbf{W}^{(l)} \cdot \max_{\mathbf{p}_j^k \in \mathcal{N}_i^{(l),k}} \mathbf{p}_j^k$$



Experimental Results - Ego4D HOI Tasks

We validate EgoPack on AR, OSCC, PNR and LTA from Ego4D.

	Trained on frozen features	AR		OSCC	LTA		PNR
		Verbs Top-1 (%)	Nouns Top-1 (%)	Acc. (%)	Verbs ED (↓)	Nouns ED (↓)	Loc. Err. (s) (↓)
Ego4D Baselines	✗	22.18	21.55	68.22	0.746	0.789	0.62
EgoT2s	✗	23.04	23.28	**72.69**	0.731	0.769	**0.61**
MLP	✓	24.08	30.45	70.47	0.763	**0.742**	1.76
Temporal Graph	✓	24.25	30.43	71.26	0.754	0.752	**0.61**
Multi-Task Learning	✓	22.05	29.44	71.10	0.740	0.746	0.62
Task Translation†	✓	23.68	28.28	71.48	0.740	0.756	**0.61**
EgoPack	✓	**25.10**	**31.10**	71.83	**0.728**	0.752	**0.61**

Metrics: accuracy for AR and OSCC, Edit Distance for LTA and Temporal Localization Error (in seconds) for PNR.

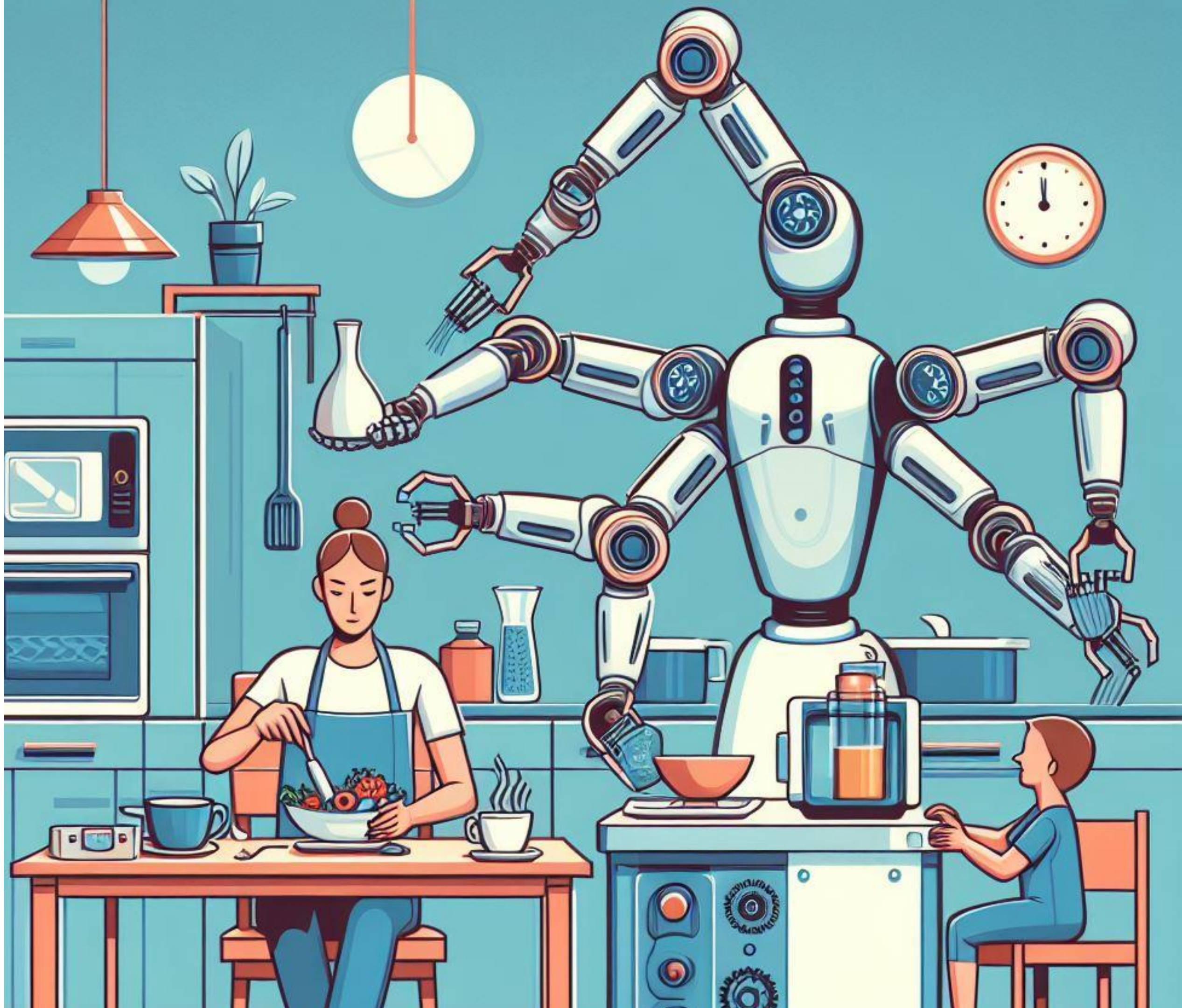
Politecnico di Torino

1859

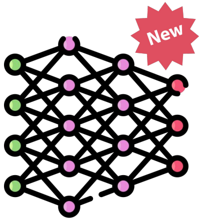
Hier-EgoPack: Hierarchical Egocentric Video Understanding with Diverse Task Perspectives

Journal Extension (*under review*)

Simone Alberto Peirone, Francesca Pistilli, Antonio Alliegro, Tatiana Tommasi, Giuseppe Averta

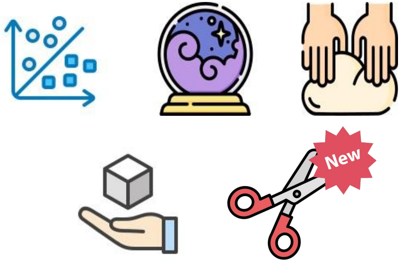


Hier-EgoPack: Hierarchical Egocentric Video Understanding with Diverse Task Perspectives



A newly crafted GNN for multi-scale temporal reasoning

that supports strong hierarchical temporal reasoning



Extension to variable temporal range tasks

which requires to incorporate long-term temporal reasoning

The Moment Queries (MQ) Task

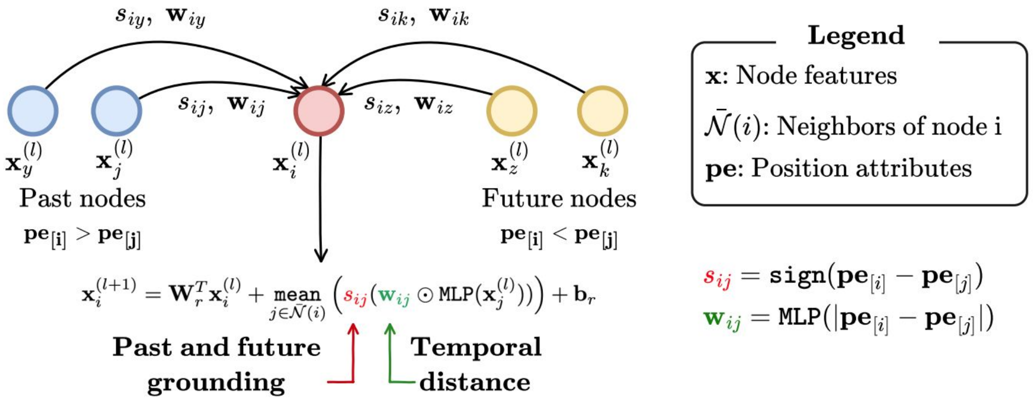
Find the segment of a video that matches a given activity query.

Activities may be seconds or minutes long → ***long-range temporal reasoning***

A new GNN for multi-scale temporal reasoning

Temporal Distance-Gated Convolution (TDGC)

A temporal-aware GNN that leverages the temporal distance between nodes to rescale their contributions during message passing.

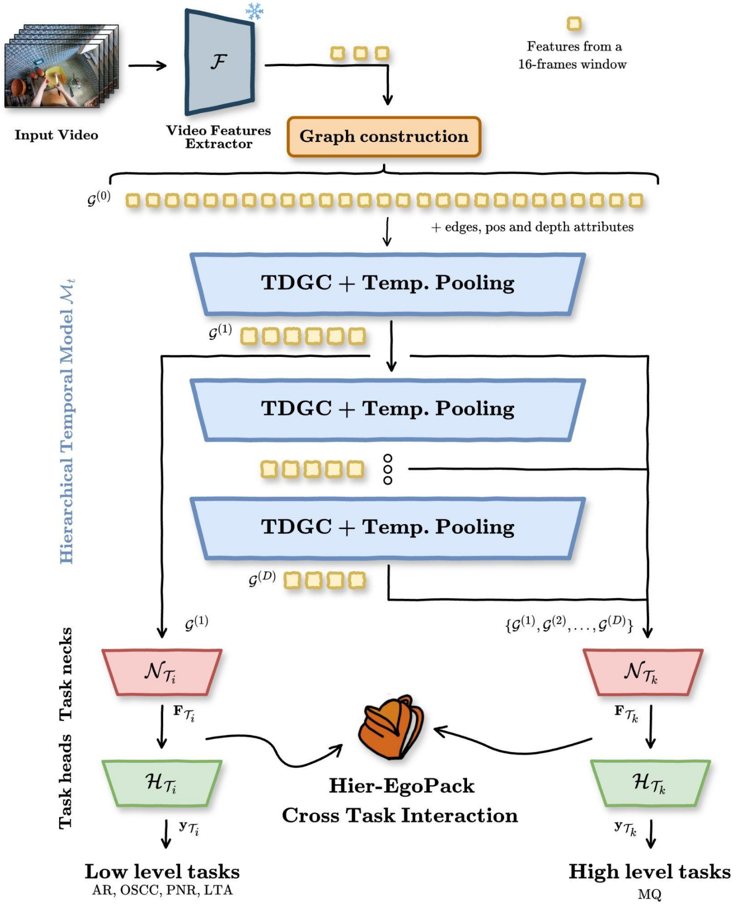


$$s_{ij} = \mathtt{sign}(\mathbf{pe}_{[i]} - \mathbf{pe}_{[j]})$$
$$\mathbf{w}_{ij} = \mathtt{MLP}(|\mathbf{pe}_{[i]} - \mathbf{pe}_{[j]}|)$$

Extension to temporal hierarchical tasks

We extend the temporal backbone to support **hierarchical and long term reasoning**

The hierarchical GNN progressively aggregates temporal information, moving from *fine-grained* temporal segments to more *coarse* representations



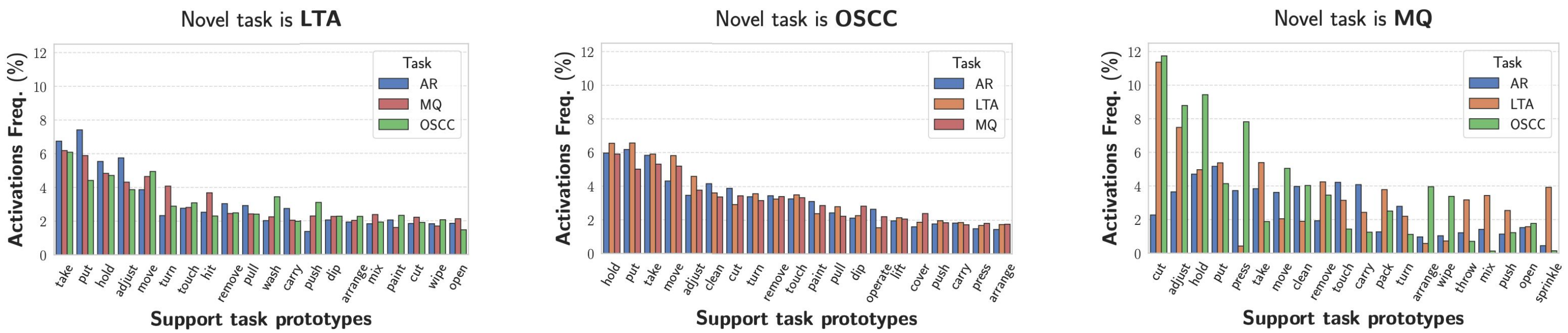
Experimental results on Ego4D tasks

	AR		OSCC	LTA		PNR	MQ
	Verbs Top-1 (%)	Nouns Top-1 (%)	Acc. (%)	Verbs ED (↓)	Nouns ED (↓)	Loc. Err. (↓)	mAP
Ego4D Baselines [8]	22.18	21.55	68.22	0.746	0.789	0.62	6.03
EgoT2s [5]	23.04	23.28	72.69	0.731	0.769	**0.61**	N/A
EgoPack [6]	25.10	31.10	71.83	0.728	0.752	**0.61**	N/A
Single Task	26.93	33.50	75.22	0.728	0.752	0.62	20.2
MTL	26.31	33.90	74.79	0.730	0.754	0.62	18.5
MTL + FT	26.71	33.51	75.00	0.728	0.749	**0.61**	19.9
MTL + HT	26.07	33.20	74.27	0.729	0.748	0.62	N/A
Task-Translation†	26.10	33.83	**76.42**	0.729	0.750	0.63	20.5
Hier-EgoPack	**27.30**	**34.65**	75.60	**0.725**	**0.741**	**0.61**	**21.0**

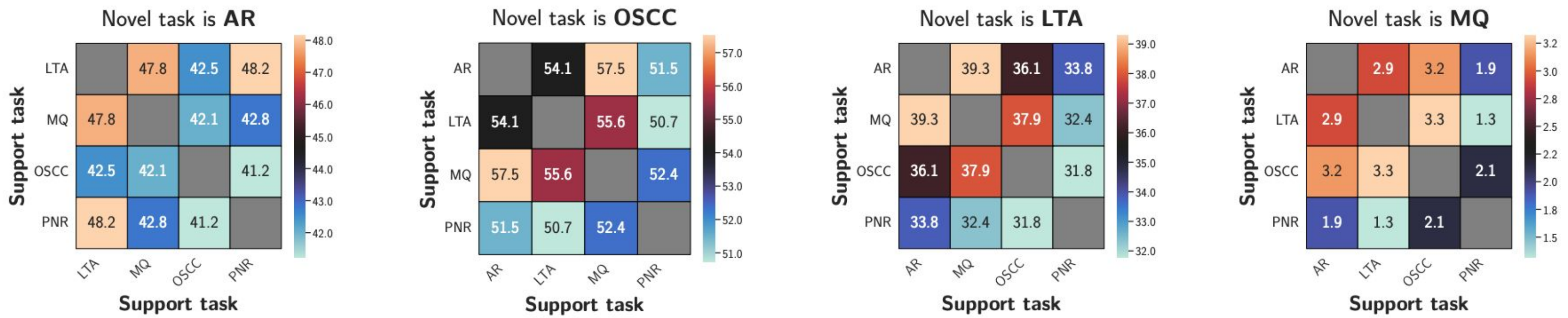
Metrics: accuracy for AR and OSCC, Edit Distance for LTA, Temporal Localization Error (in seconds) for PNR and mAP for Moment Queries (MQ).

Qualitative Visualizations

Activation frequency for the task-specific prototypes from different support tasks



Activations consensus for different novel tasks



Hier-EgoPack: Hierarchical Egocentric Video Understanding with Diverse Task Perspectives

Simone Alberto Peirone, Francesca Pistilli, Antonio Alliegro, Tatiana Tommasi, Giuseppe Averta

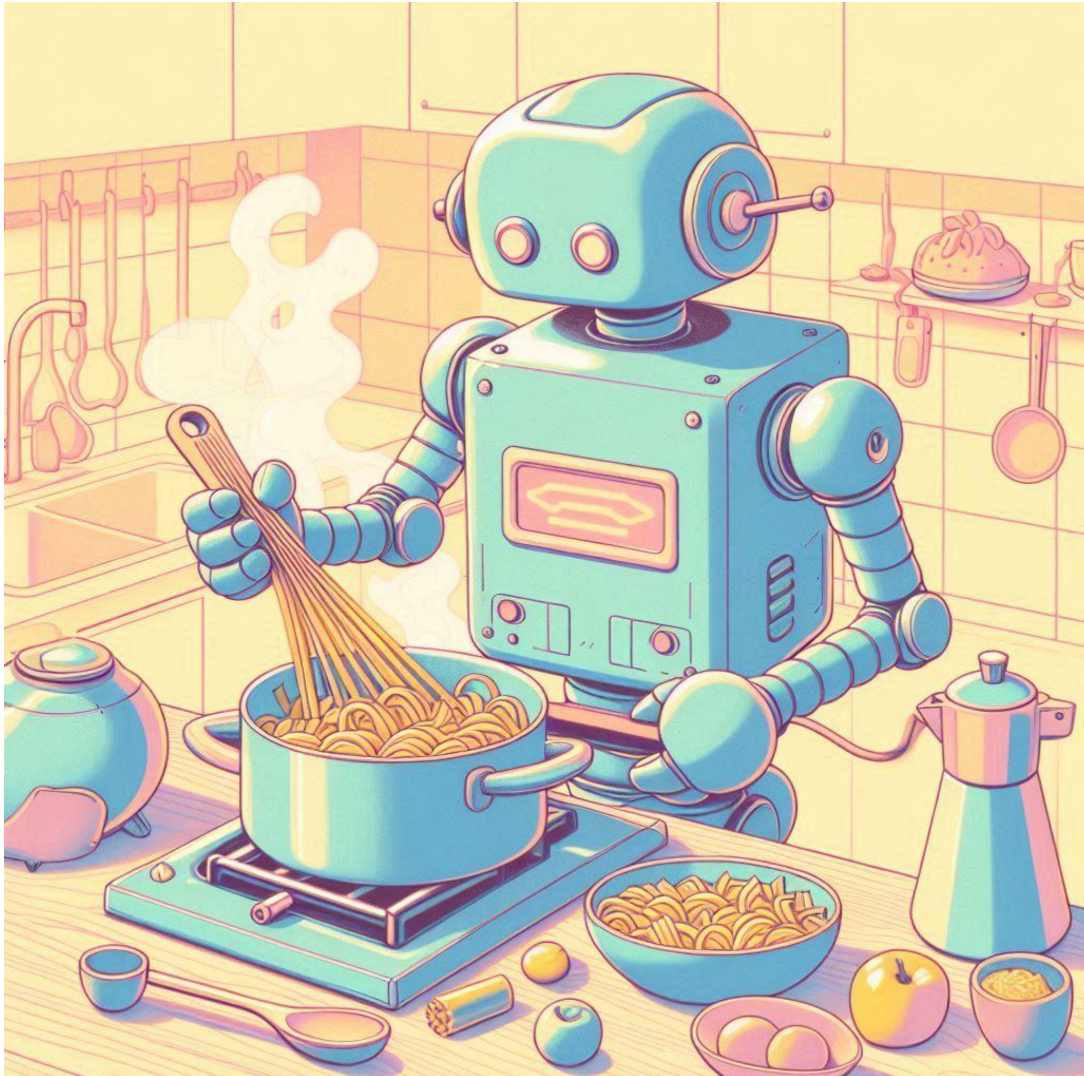
Learn more at **sapeirone.github.io/hier-egopack/**

Politecnico di Torino

HiERO:
understanding the hierarchy of human behavior enhances reasoning on egocentric videos

On ArXiv soon

Simone Alberto Peirone, Francesca Pistilli, Giuseppe Averta



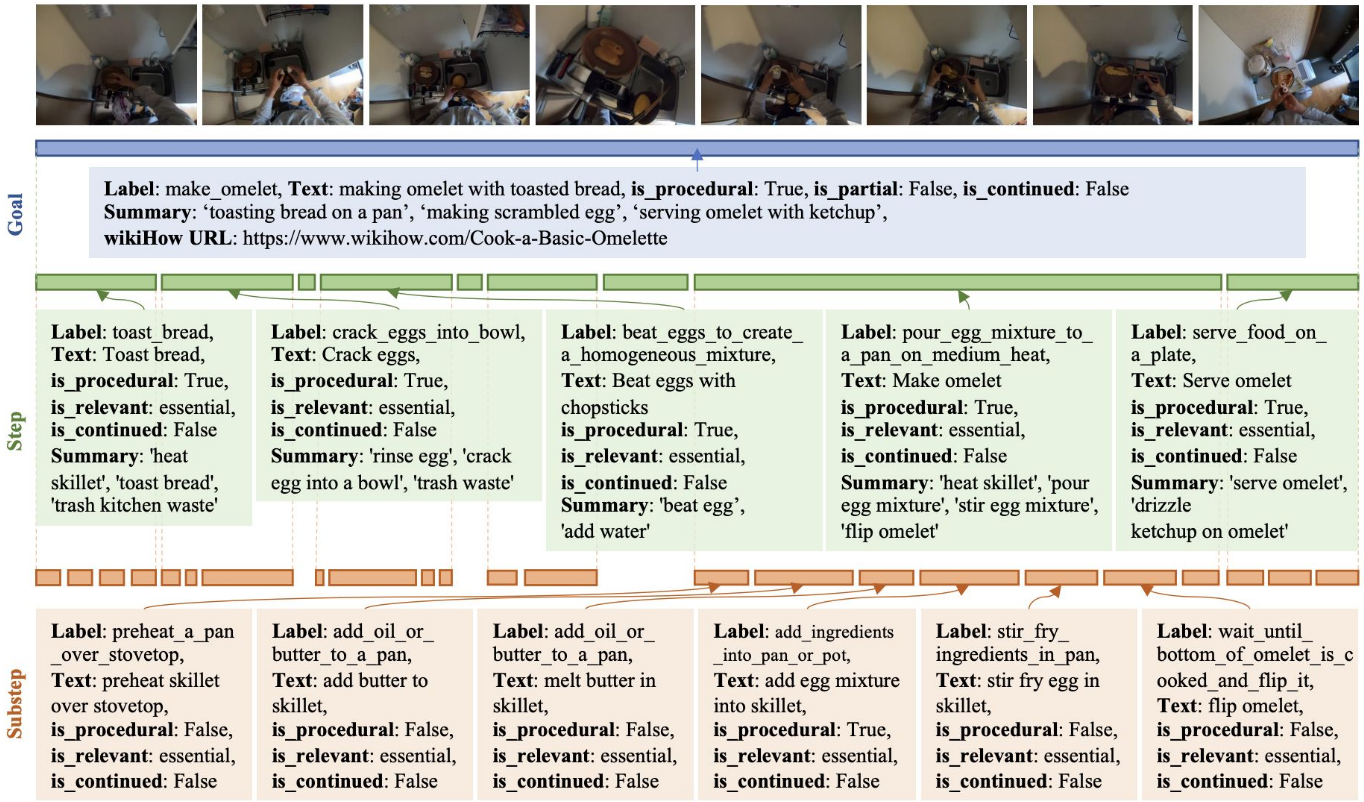
Understanding the hierarchy of human behavior

Human activities are complex and variable…



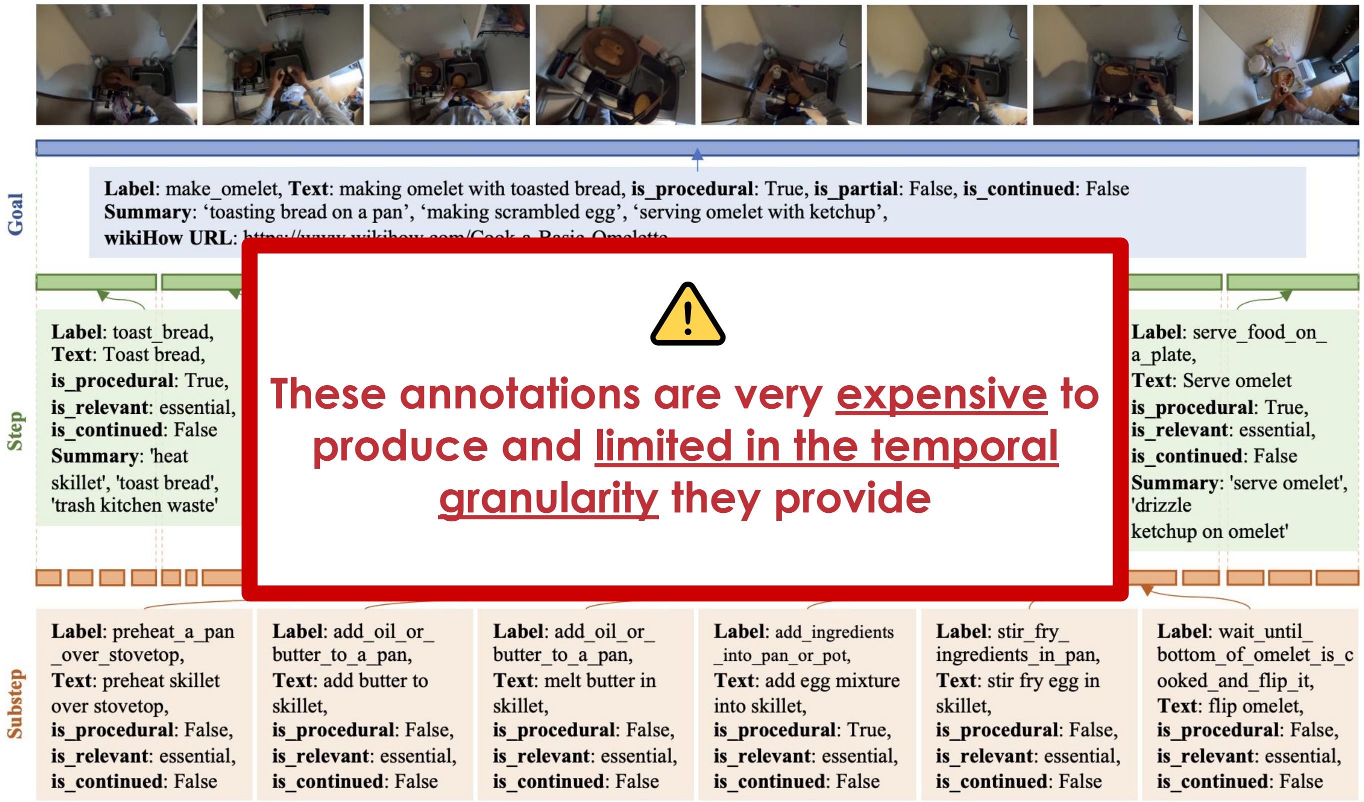
Different examples of human activities *in-the-wild*,
showing a large variety of interactions and actions

Human activities are hierarchical and goal-oriented



Song, Yale, et al. "Ego4d goal-step: Toward hierarchical understanding of procedural activities." *NeurIPS 2023*

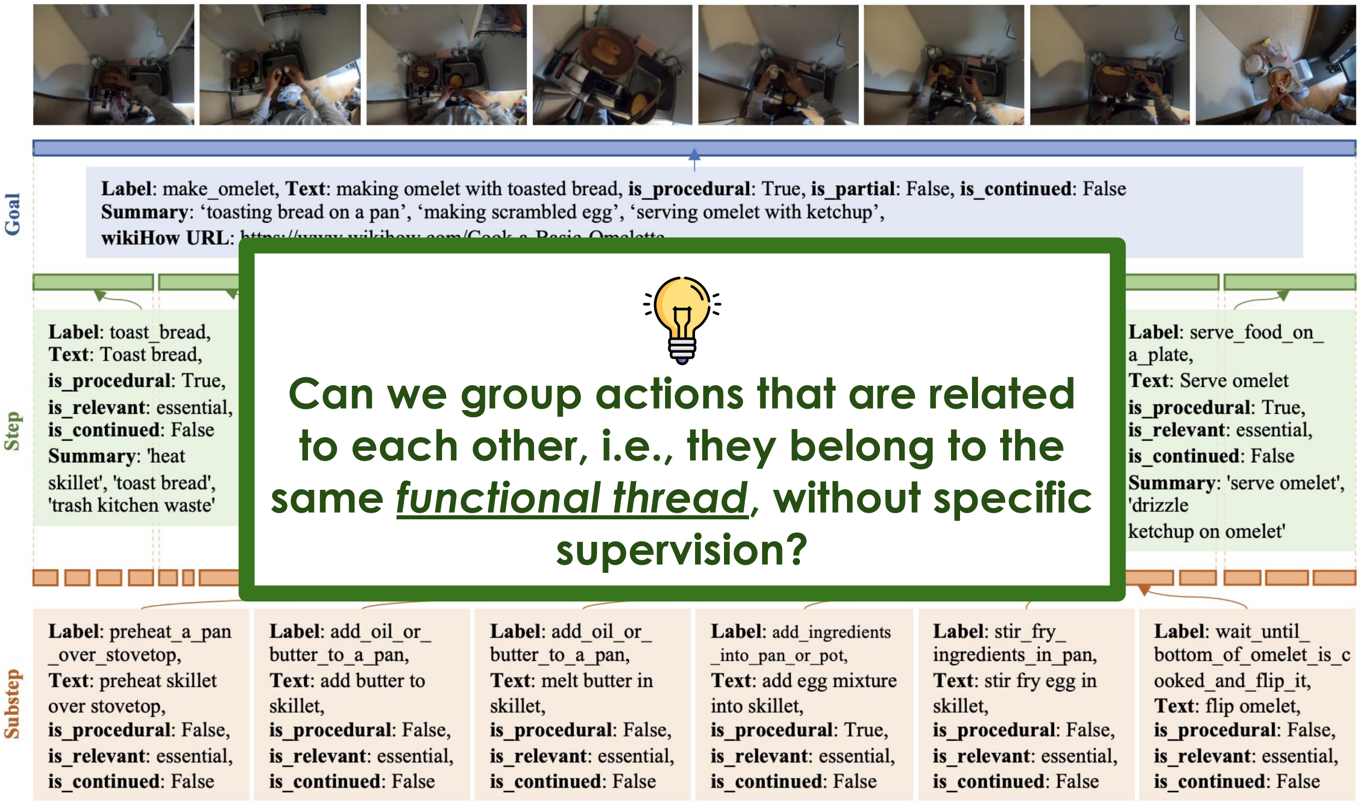
Human activities are hierarchical and goal-oriented



⚠ These annotations are very <u>expensive</u> to produce and <u>limited in the temporal granularity</u> they provide

Song, Yale, et al. "Ego4d goal-step: Toward hierarchical understanding of procedural activities." *NeurIPS 2023*

Human activities are hierarchical and goal-oriented

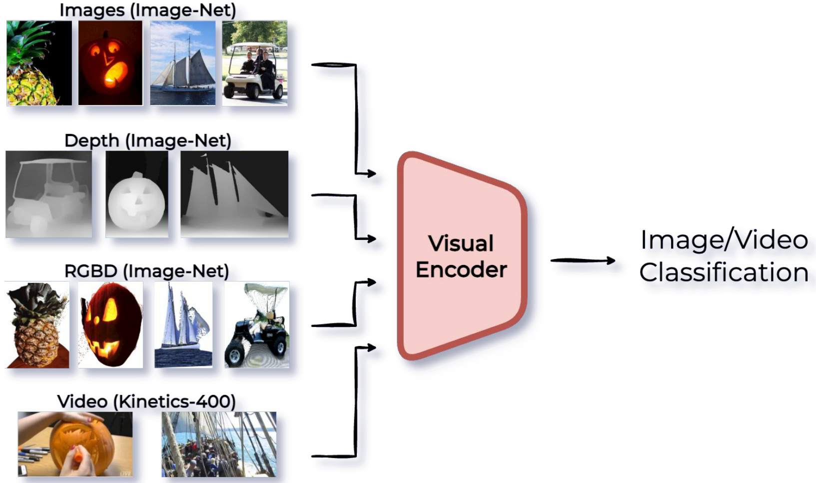


Can we group actions that are related to each other, i.e., they belong to the same *functional thread*, without specific supervision?

Song, Yale, et al. "Ego4d goal-step: Toward hierarchical understanding of procedural activities." *NeurIPS 2023*

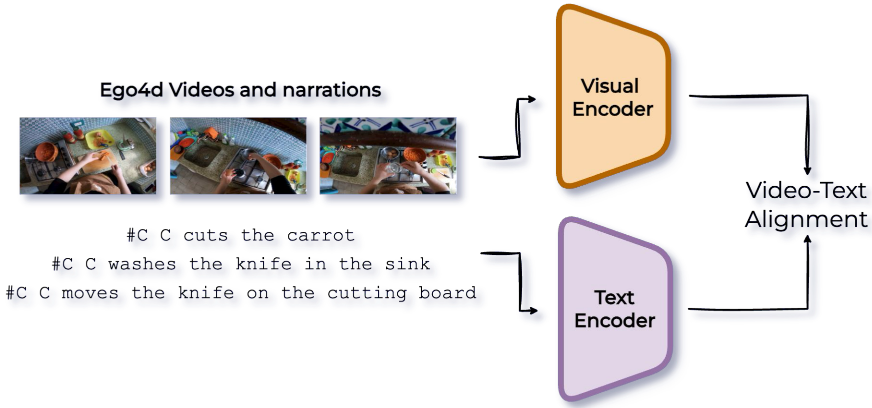
Discovering functional threads in videos from features similarity

Let's take two **video feature extractors** and look at the similarity matrix for the segments of a Ego4D video



Omnivore

Girdhar, Rohit, et al. "Omnivore: A single model for many visual modalities." *CVPR 2022*

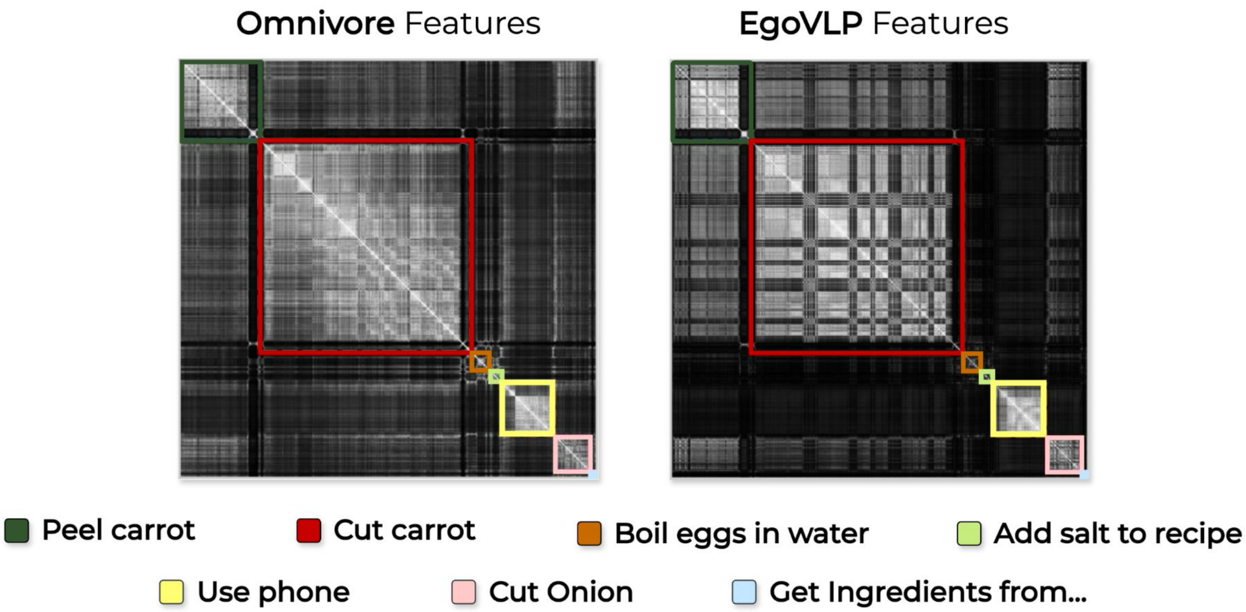


EgoVLP

Lin, Kevin Qinghong, et al. "Egocentric video-language pretraining." *NeurIPS 2022*

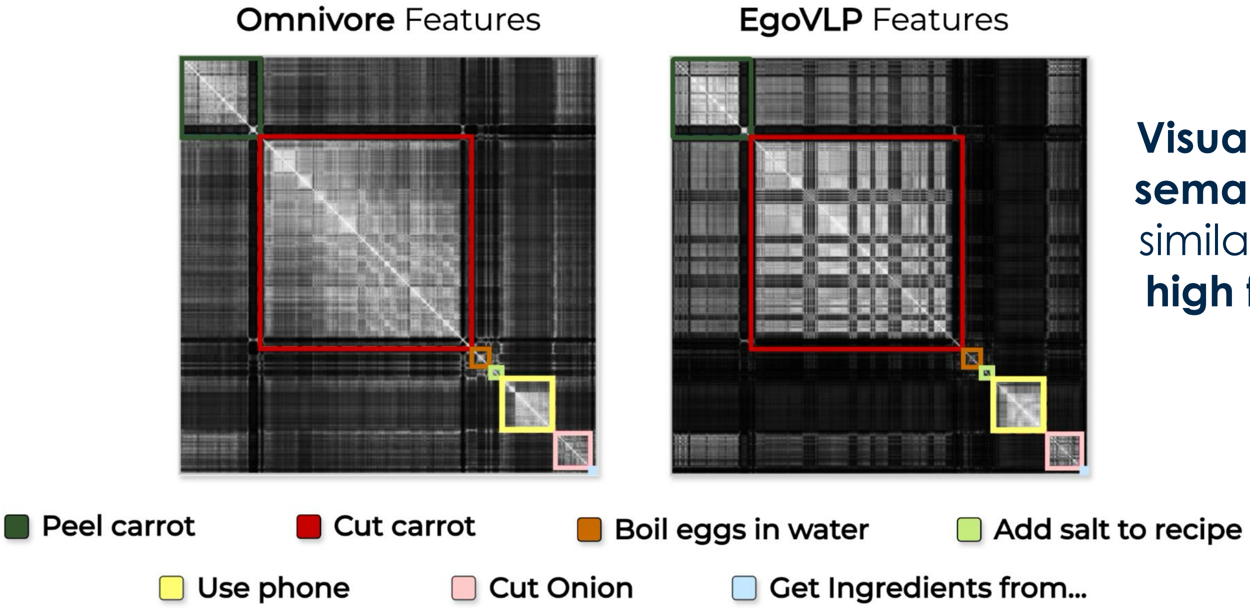
Discovering functional threads in videos from features similarity

Let’s take two **video feature extractors** and look at the similarity matrix for the segments of a Ego4D video



Discovering functional threads in videos from features similarity

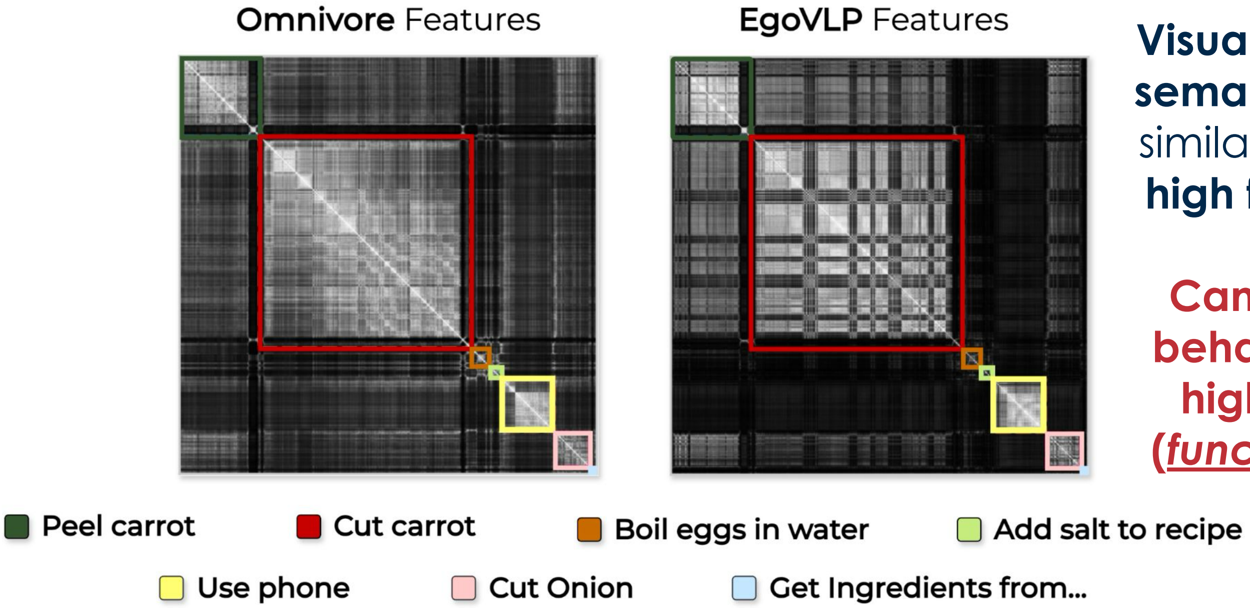
Let's take two **video feature extractors** and look at the similarity matrix for the segments of a Ego4D video



Visually (Omnivore) or **semantically** (EgoVLP) similar segments have **high feature similarity**

Discovering functional threads in videos from features similarity

Let’s take two **video feature extractors** and look at the similarity matrix for the segments of a Ego4D video

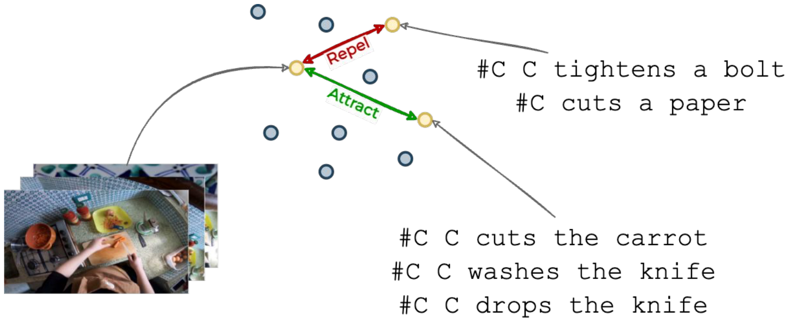


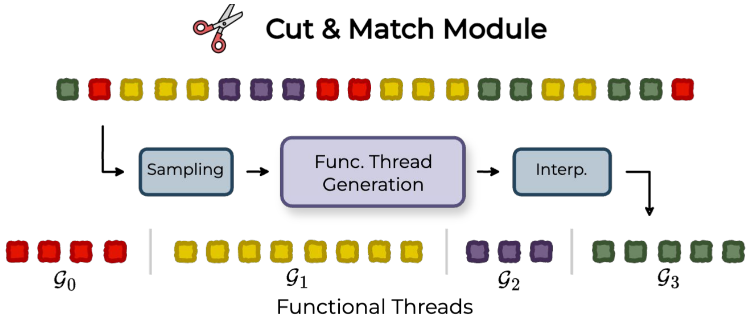
Visually (Omnivore) or **semantically** (EgoVLP) similar segments have **high feature similarity**

Can we exploit this behavior to discover high-level actions (*functional threads*)?

The HiERO Architecture

HiERO learns to **map close** in feature space **actions** corresponding to the **same functional thread** via two objectives:





1. Clip - Narrations alignment

align segments of a video with their corresponding narrations

2. Functional threads clustering

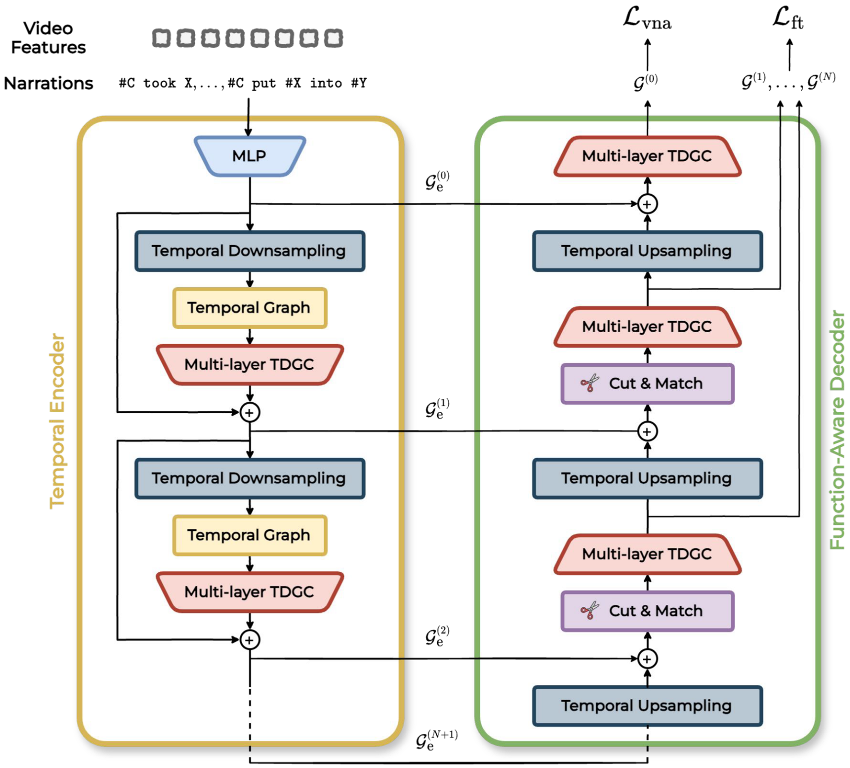
groups segments of the video that encode ***functionally similar*** actions

The HiERO Architecture

HiERO is built on two components:

1. A **Temporal Encoder** gradually aggregates temporal information in the video

2. A **Function-Aware Decoder** discovers strongly connected regions in the input videos that correspond to functional threads using the **Cut & Match** module



The *Cut & Match* Module

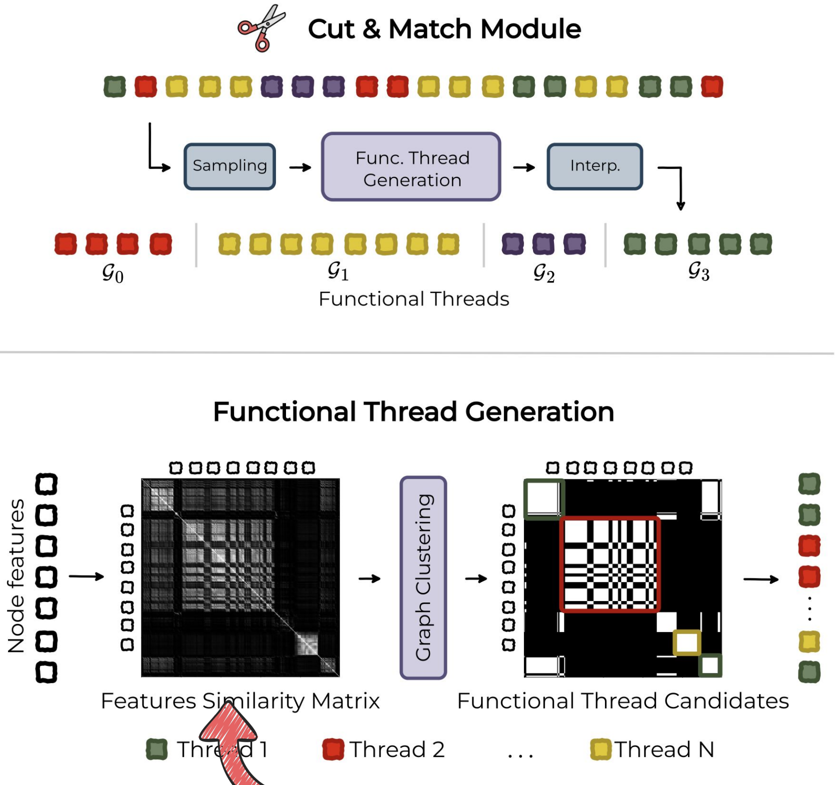
Building the graph based on **temporal connectivity is limited** as it looks at local temporal portions of the video

⇩

Redefine the graph connectivity based on segments that share functionally related actions

⇩

Use **Spectral Clustering** to identify strongly connected regions in the graph

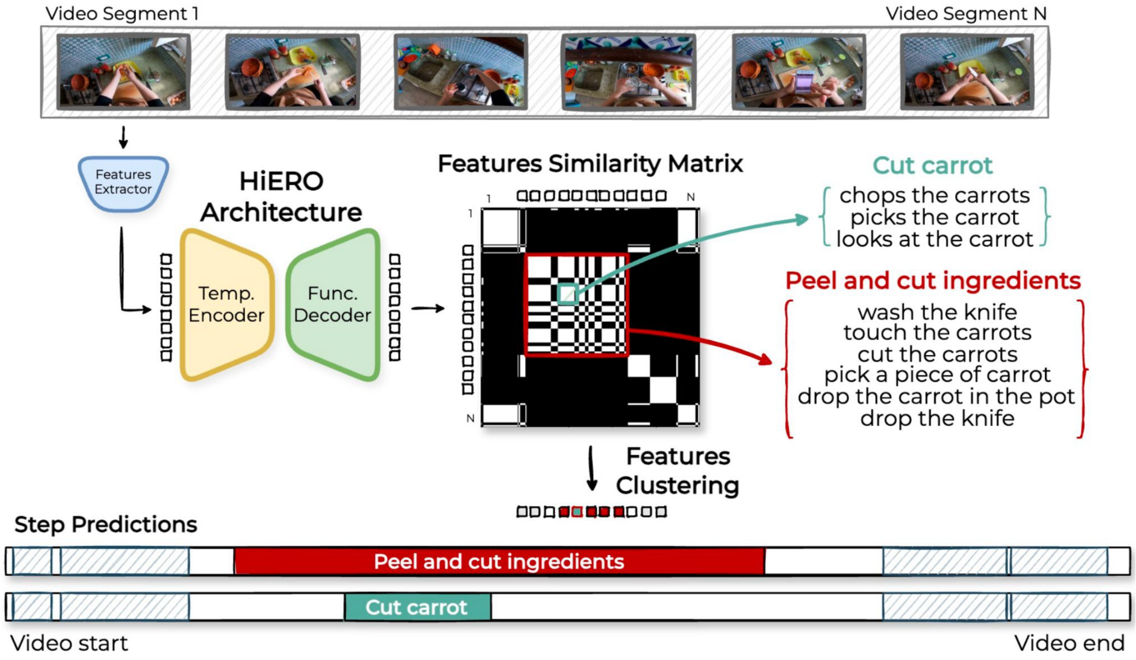


Fully connected graph with edges defined by features similarity

(Zero-Shot) procedure learning tasks with HiERO

In the HiERO's features space, we can
detect functional threads with a simple features clustering operation!

Stage 1:
features extraction from the video segments



Stage 2:
build a similarity matrix from the features of the video segments

Stage 3:
cluster the video features to obtain the steps

Experimental Validation

Three main validation tracks in supervised and zero-shot settings

1. **Video-Text Alignment** on EgoMCQ and EgoNLQ
   a. **EgoMCQ**: text-to-video retrieval
   b. **EgoNLQ**: temporal grounding of natural language queries
2. **Procedure Learning** on EgoProceL
3. **Step localization and grounding** on Ego4D Goal-Step

Zero-Shot!

Experimental Validation (1): Video-Text Alignment

EgoMCQ: given a textual caption and set of five short candidate clips, find the correct match

EgoNLQ: find the temporal segment that answer a textual query.

Observation: HiERO is effective in discriminating short actions (EgoMCQ) and in capturing long-range casual and temporal dependencies (EgoNLQ).

Method	**EgoMCQ**		**EgoNLQ**			
	Accuracy (%)		mIOU@0.3		mIOU@0.5	
	Inter	Intra	R@1	R@5	R@1	R@5
Omnivore [15]† (CVPR'22)	–	–	6.56	12.55	3.59	7.90
SlowFast [13] (ICCV'19)	–	–	5.45	10.74	3.12	6.63
EgoVLP [29] (NIPS'22)	90.6	57.2	10.84	18.84	6.81	13.45
HierVL [2] (CVPR'23)	90.5	52.4	–	–	–	–
LAVILA [56] (CVPR'23)	94.5	63.1	12.05	22.38	7.43	15.44
EgoVLPv2 [38] (ICCV'23)	91.0	60.9	12.95	**23.80**	7.91	**16.11**
Ours (Omnivore)	90.1	53.4	10.27	18.20	6.01	12.52
Ours (EgoVLP)	91.6	59.6	11.41	19.67	7.05	13.91
Ours (LAVILA)	**94.6**	**64.4**	**13.35**	21.12	**8.08**	15.31

Results on the EgoMCQ and EgoNLQ tasks on Ego4D. Performance are reported in terms of accuracy (EgoMCQ) and Recall at different IoU thresholds (EgoNLQ)

Lin, Kevin Qinghong, et al. "Egocentric video-language pretraining." *NeurIPS 2022*

Experimental Validation (2): Procedure Learning

Procedure Learning: given a procedural video, identify all the key-steps (video segments) of the procedure, *without additional training*.

Method	**Average**		**CMU-MMAC** [10]		**EGTEA** [28]		**MECCANO** [40]		**EPIC-Tents** [21]		**PC Ass.** [4]		**PC Disass.** [4]	
	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU
Random [8] (NeurIPS'24)	14.8	6.1	15.7	5.9	15.3	4.6	13.4	5.3	14.1	6.5	15.1	7.2	15.3	7.1
CnC [4] (ECCV'22)	22.0	10.7	22.7	11.1	21.7	9.5	18.1	7.8	17.2	8.3	25.1	12.8	27.0	14.8
GPL-2D [5] (WACV'24)	22.0	11.9	21.8	11.7	23.6	14.3	18.0	8.4	17.4	8.5	24.0	12.6	27.4	15.9
GPL [5] (WACV'24)	25.6	13.9	31.7	17.9	27.1	16.0	20.7	10.0	19.8	9.1	27.5	15.2	26.7	15.2
OPEL [8] (NeurIPS'24)	32.0	16.3	36.5	18.8	29.5	13.2	39.2	20.2	20.7	10.6	33.7	17.9	32.2	16.9
Omnivore	39.1	22.0	44.7	26.8	37.1	19.2	36.0	19.0	40.8	21.9	35.7	21.5	40.3	23.5
Ours (Omnivore)	44.0	24.5	47.2	27.7	**39.7**	**19.9**	**41.6**	**22.1**	**45.3**	**24.3**	43.7	25.1	46.3	27.9
EgoVLP	40.0	21.9	49.2	31.0	36.6	18.3	33.1	16.1	37.4	19.2	38.2	20.8	45.4	25.6
Ours (EgoVLP)	**44.5**	**25.3**	**53.5**	**34.0**	**39.7**	19.6	39.8	20.3	39.0	20.3	**44.9**	**25.6**	**49.9**	**32.1**

Results on the Procedure Learning task on EgoProceL using F1 score and IoU on the discovered key-steps

Bansal, Siddhant, Chetan Arora, and C. V. Jawahar. "My view is the best view: Procedure learning from egocentric videos." *ECCV 2022*
Bansal, Siddhant, Chetan Arora, and C. V. Jawahar. "United we stand, divided we fall: Unitygraph for unsupervised procedure learning from videos." WACV 2024
Chowdhury, Sayeed Shafayet, Soumyadeep Chandra, and Kaushik Roy. "OPEL: Optimal Transport Guided ProcedurE Learning." NeurIPS 2024

Experimental Validation (3): Step localization and grounding

Step Grounding

given a textual description of a step, find the corresponding temporal boundaries in the video

Method	Approach	mIoU@0.3		mIoU@0.5	
		R@1	R@5	R@1	R@5
Omnivore [47]	Supervised	12.02	19.99	7.71	14.17
Ours (Omnivore)	Supervised	13.02	21.81	8.59	15.98
EgoVLP	Supervised	15.43	25.91	10.95	19.77
Ours (EgoVLP)	Supervised	**15.64**	**26.01**	**11.14**	**20.08**
EgoVLP	Zero-Shot	10.73	24.70	7.38	16.53
Ours (Omnivore)	Zero-Shot	9.29	22.89	6.24	15.05
Ours (EgoVLP)	Zero-Shot	**11.57**	**27.41**	**7.87**	**18.70**

Results on the Step Grounding task on Goal-Step, in terms of Recall at different IoU thresholds

Step Localization

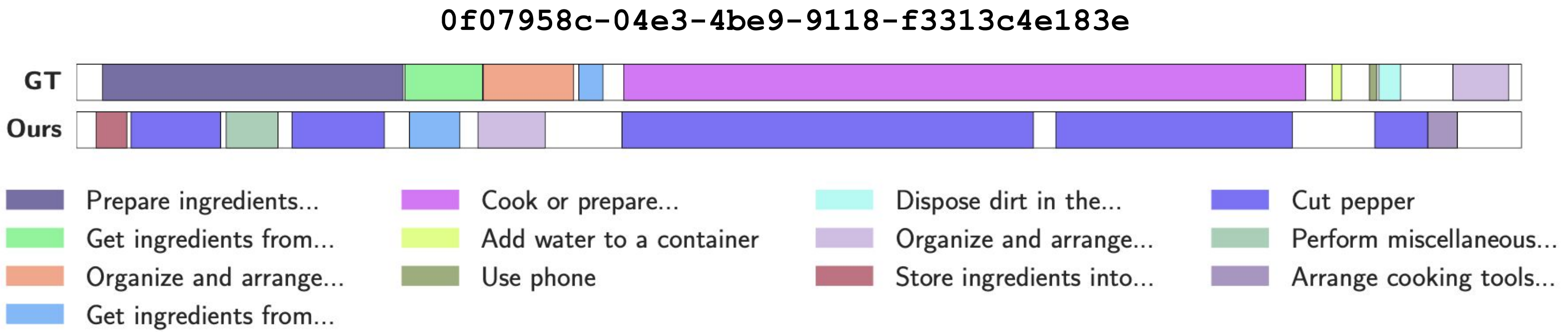
given a procedural video, find all the steps in the video (temporal boundaries and step label)

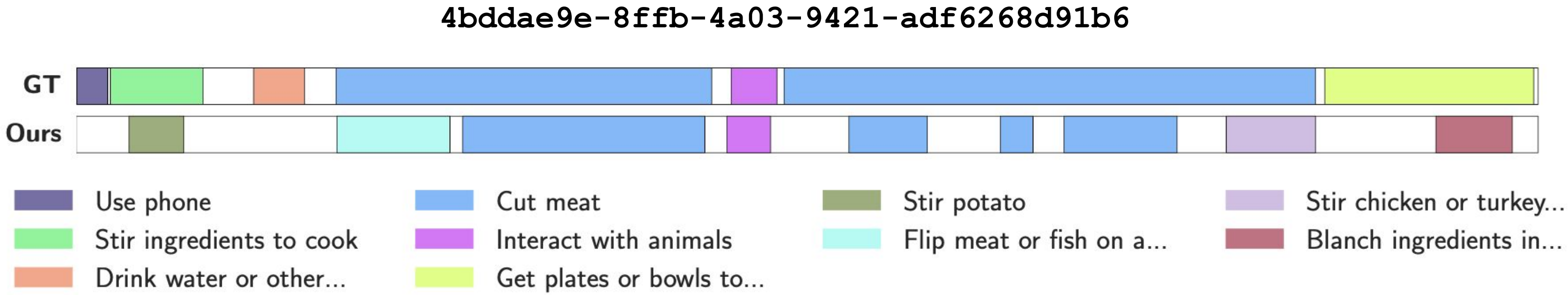
Method	Approach	mAP @ IoU					
		0.1	0.2	0.3	0.4	0.5	Avg
Omnivore [47]	Supervised	–	–	–	–	–	10.3
EgoOnly [47]	Supervised	–	–	–	–	–	**13.6**
EgoVLP	Supervised	13.2	12.2	11.1	10.0	8.6	11.0
Ours (EgoVLP)	Supervised	14.1	13.1	12.1	10.9	9.5	11.9
EgoVLP	Zero-Shot	11.8	9.7	8.3	6.7	5.1	8.3
Ours (EgoVLP)	Zero-Shot	12.0	10.0	8.8	7.3	5.6	8.7

Results on the Step Localization task on Goal-Step, in terms of mAP at different IoU thresholds

Song, Yale, et al. "Ego4d goal-step: Toward hierarchical understanding of procedural activities." *NeurIPS 2023*

Some success and failure cases in zero-shot Step Localization





Observation: several failure cases are linked to *mismatches in the temporal granularity* of the ground truth and the predictions

HiERO: understanding the hierarchy of human behavior enhances reasoning on egocentric videos

Simone Alberto Peirone, Francesca Pistilli, Giuseppe Averta

ArXiv preprint will be out soon…

Thank you!

Simone Alberto Peirone

simone.peirone@{polito.it,epfl.ch}