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

Simone Alberto Peirone, Francesca Pistilli, Giuseppe Averta

name.surname@polito.it (Politecnico di Torino, Italy)

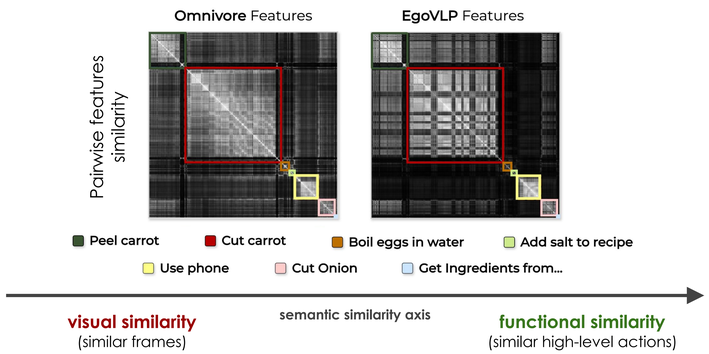
ICCV HONOLULU HAWAII OCT 19-23, 2025

ellis European Laboratory for Learning and Intelligent Systems · Politecnico di Torino 1859 · FAIR Future Artificial Intelligence Research

ArXiv Paper · Code

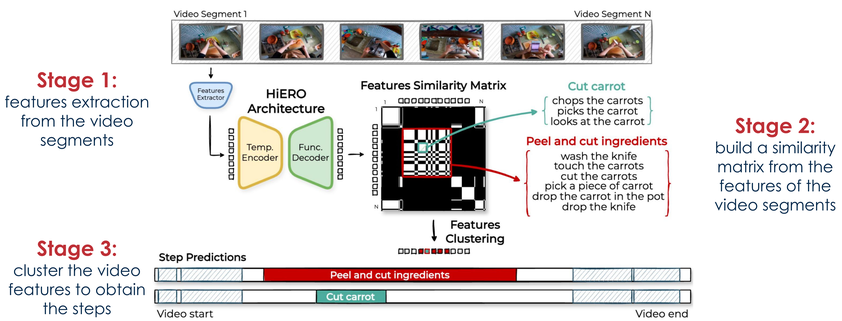
Human activities are goal-oriented and hierarchical

How do video models encode action hierarchies?



Discovering hierarchical activities by *clustering*

We learn a features space where functionally similar actions are mapped to close embeddings



HiERO improves Video-Language Alignment

with little overhead compared to other VLP methods!

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 [57] (CVPR'23)	94.5	63.1	12.05	22.38	7.43	15.44
EgoVLPv2 [39] (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

Action hierarchies help detecting short actions (**EgoMCQ**) and long-range temporal/causal relationships (**EgoNLQ**)

HiERO takes **<20 GPU hours** vs ~1500 GPU hours for EgoVLP

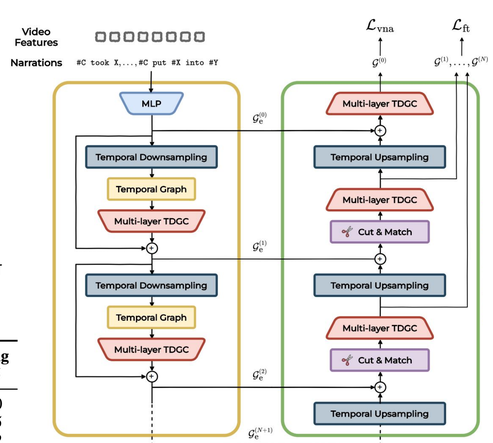
HiERO is designed for hierarchical reasoning on human activities

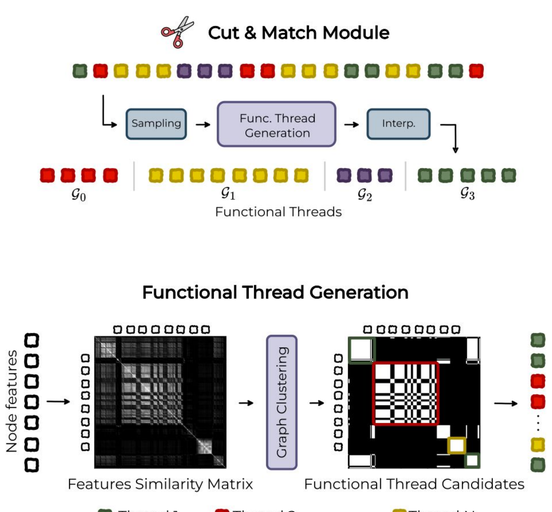
Architecture

- **Temporal Encoder**: temporal reasoning on **local temporal segments**
- **Function-Aware Decoder**: temporal reasoning on **activity-clustered segments**

Objectives

- **Video-narrations** align video clips with their narrated descriptions
- **Functional threads** push segments belonging to the same high-level activity closer





Align Loss	Func. Th. Cluster	Func Th. Loss	EgoMCQ Inter	EgoMCQ Intra	EgoProceL F1	EgoProceL IoU	Step-Grounding R@1	Step-Grounding R@5
✗	✗	✗	90.6	57.2	40.0	21.9	10.73	24.70
✓	✗	✗	**91.8**	59.5	43.8	24.1	11.27	27.35
✓	✓	✗	**91.8**	**59.6**	43.3	24.2	11.44	27.12
✓	✓	✓	91.6	**59.6**	**44.5**	**25.3**	**11.57**	**27.41**

HiERO can detect procedural steps in *zero-shot*

while being competitive (or even better!) compared to supervised baselines

With a simple clustering in the **HiERO features space** we can group actions that are functionally related to each other (**Ego4D Goal-Step** and **EgoProceL**)

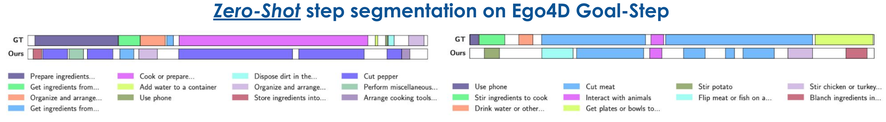
Method	Approach	mIoU@0.3 R@1	mIoU@0.3 R@5	mIoU@0.5 R@1	mIoU@0.5 R@5
Omnivore [48]	Supervised	12.02	19.99	7.71	14.17
Ours (Omnivore)	Supervised	13.02	21.81	8.59	15.98
EgoVLP	Supervised	15.43	25.91	10.95	19.77
Ours (EgoVLP)	Supervised	**15.64**	**26.01**	**11.14**	**20.08**
EgoVLP	Zero-Shot	10.73	24.70	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**

Strong Zero-Shot performance on Step-Grounding

Method	Average F1	Average IoU
Random [8] (NeurIPS'24)	14.8	6.1
CnC [4] (ECCV'22)	22.0	10.7
GPL-2D [5] (WACV'24)	22.0	11.9
GPL [5] (WACV'24)	25.6	13.9
OPEL [8] (NeurIPS'24)	32.0	16.3
Omnivore	39.1	22.0
Ours (Omnivore)	44.0	24.5
EgoVLP	40.0	21.9
Ours (EgoVLP)	**44.5**	**25.3**

+12.5% on Procedure Learning!

Zero-Shot step segmentation on Ego4D Goal-Step



(a) Video 1 (0f07958c-04e3-4be9-9118-f3313c4e183e) (b) Video 2 (4bddae9e-8ffb-4a03-9421-adf6268d91b6)