Egocentric Video Task Translation

Zihui (Sherry) Xue[1,2], Yale Song[2], Kristen Grauman[1,2], Lorenzo Torresani[2]

[1]UT Austin [2]FAIR, Meta AI

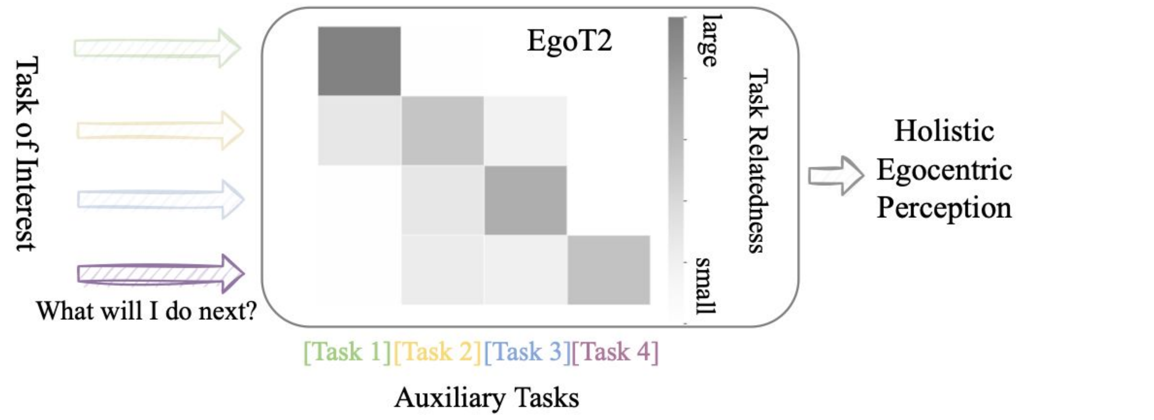
TEXAS The University of Texas at Austin | Meta AI

For more details, our released code and model checkpoints, see our website →

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Main idea: We propose task translation as a new learning paradigm to leverage synergies across different video tasks.

The attention maps produced by EgoT2 offer good interpretability on inherent task relations.



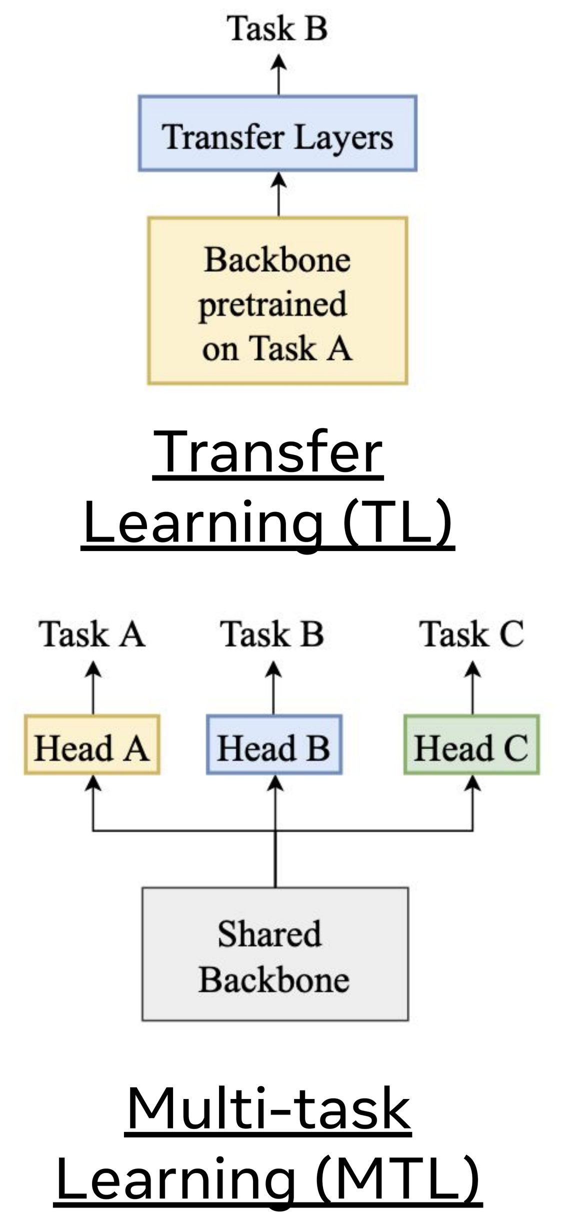
Motivation:

- Traditionally, third-person video understanding tasks are studied in isolation;
- Recent egocentric datasets provide suites of tasks associated with various human-human and human-object interactions;
- Strong synergies exist among these egocentric tasks.

Task 1 (Looking-at-me Cls.) Is the person looking at me?	Task 1 (Object State Change Cls.) Does the object go through a state change in the video?
Task 2 (Talking-to-me Cls.) Is the person talking to me?	Task 2 (Action Anticipation) What activities will happen in the future?

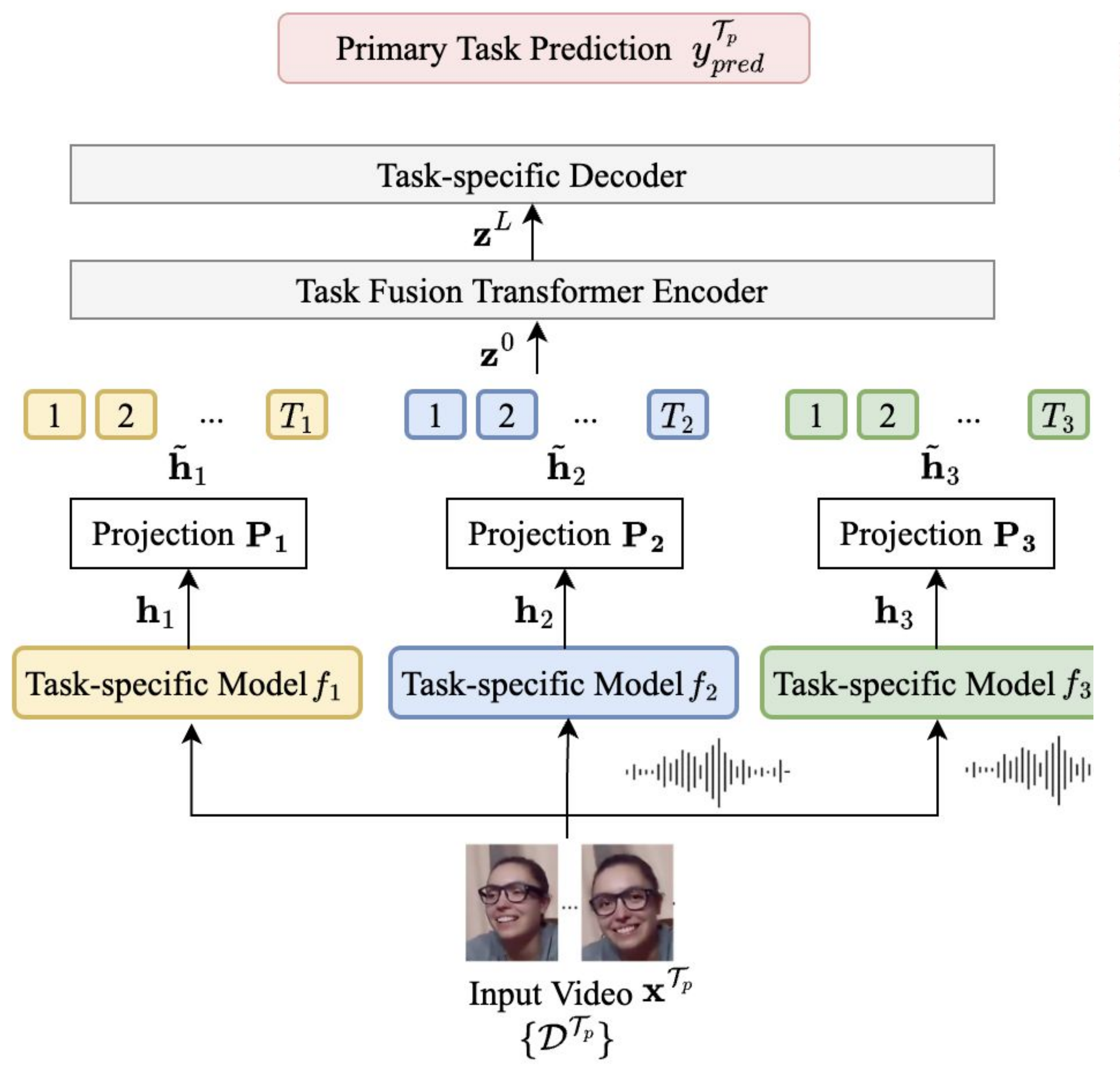
Approach: Given K video tasks, we propose two designs with distinct advantages: EgoT2-s & EgoT2-g.

Conventional approaches



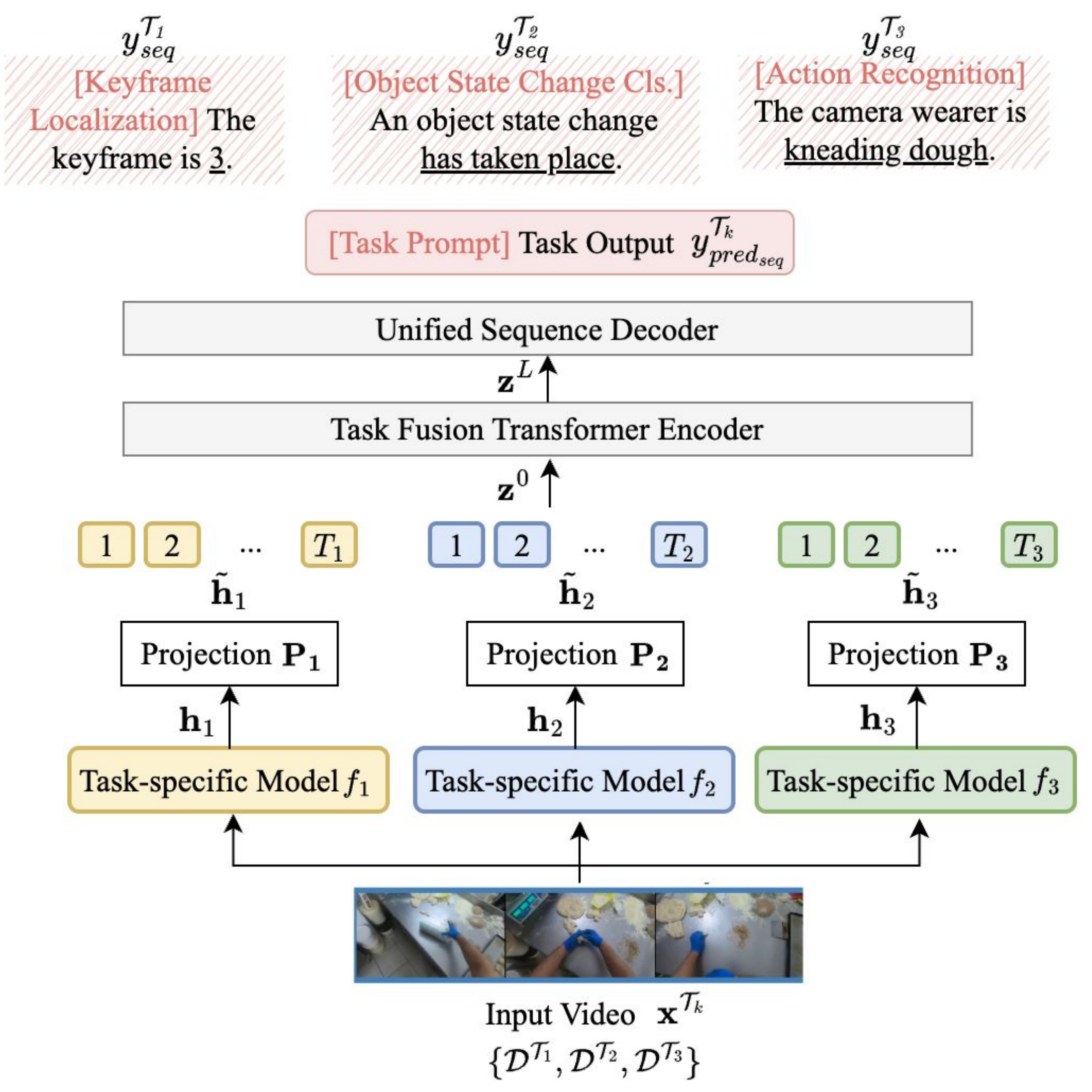
EgoT2-s (task-specific translation)

Objective: improve 1 primary task with K-1 auxiliary tasks (resembles TL)



EgoT2-g (task-general translation)

Objective: improve all K tasks at the same time (resembles MTL)



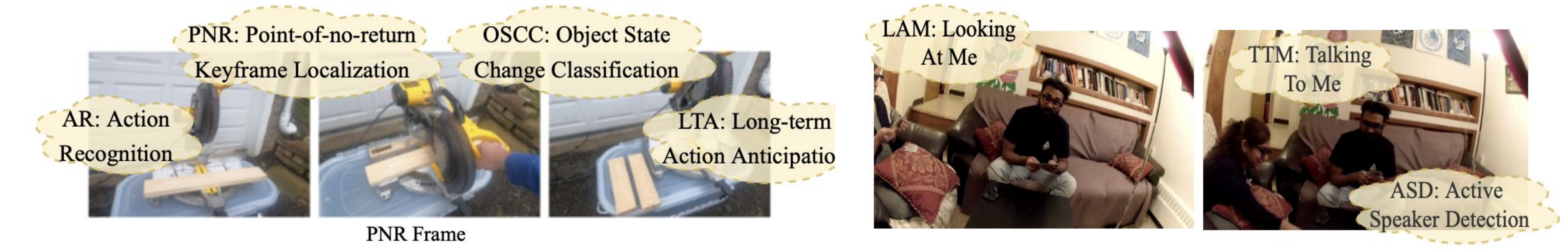
EgoT2:

- "Flipped" design: multiple task-specific (TS) backbones & a single task translator;
- Two-stage training: 1) optimize K TS backbones→2) optimize the task-specific/task-general translator.

Key advantages:

- Backbones and inputs (modality / temporal resolutions) can be selected optimally for each task;
- Do not require a common training set for all tasks;
- Leverage multiple auxiliary tasks simultaneously (unlike TL);
- Mitigate negative transfer when tasks are not strongly related (unlike MTL).

Experiments: We consider 7 diverse egocentric video tasks from Ego4D.



EgoT2-s: leads to top performance as the task translator is individually optimized for each primary task.

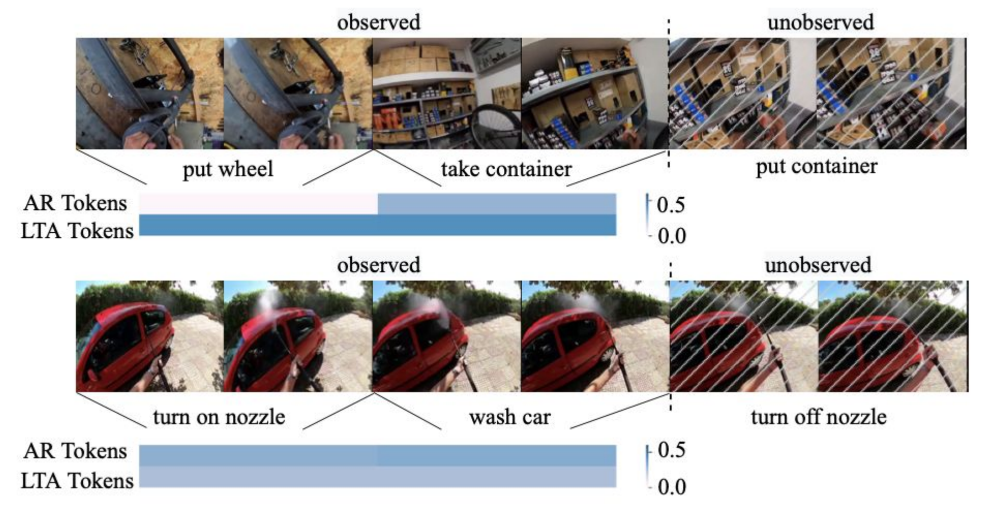
EgoT2-g: a unified framework for all task translation simultaneously, providing added flexibility.

EgoT2-s reliably adapts the auxiliary tasks to suit the primary task, consistently improving performance across all tasks.

	PNR ↓	OSCC ↑	AR ↑	LTA ↓	TTM ↑	ASD ↑
TS model	0.615	68.22	21.87	0.768	58.91	79.05
TL (best)	0.611	70.98	14.81	0.776	63.59	71.06
Late Fusion	**0.610**	72.10	20.17	0.766	64.29	77.54
EgoT2-s	**0.610**	**72.69**	**23.16**	**0.750**	**66.54**	**79.38**

EgoT2-s achieves 1st place in TTM and 3rd place in PNR at the 2022 Ego4D ECCV challenge.

EgoT2-s selectively activates auxiliary task feature tokens for the primary task (Primary task: LTA, Auxiliary task: AR).



EgoT2-g is flexible, accurate and mitigates negative transfer.

	PNR ↓	OSCC ↑	AR ↑	LTA ↑	LAM ↑	TTM ↑	ASD ↑
TS model	0.615	68.2	21.87	21.31	**77.79**	58.91	79.05
MTL	0.617	66.0	N/A	N/A	60.53	61.91	N/A
EgoT2-g	**0.611**	**71.7**	**22.33**	**22.76**	77.63	**64.49**	**79.06**

EgoT2-g activates task tokens conditioned on the task prompt.