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

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.

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 K tasks at the same time (resembles MTL).

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;
- Mitigate negative transfer when tasks are not strongly related (unlike MTL).

Experiments: We consider 7 diverse egocentric video tasks from Ego4D.
- EgoT2-s selects auxiliary task feature tokens for the primary task (Primary task: L TA, Auxiliary task: AR).
- EgoT2-s reliably adapts the auxiliary tasks to suit the primary task, consistently improving performance across all tasks.

Approach:
- EgoT2-s selectively activates auxiliary task feature tokens for the primary task (Primary task: L TA, Auxiliary task: AR).
- EgoT2-g activates task tokens conditioned on the task prompt.
- 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-g is flexible, accurate and mitigates negative transfer.