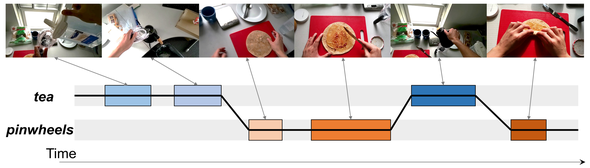
Understanding Multi-Task Activities From Single-Task Videos

Yuhan Shen and Ehsan Elhamifar

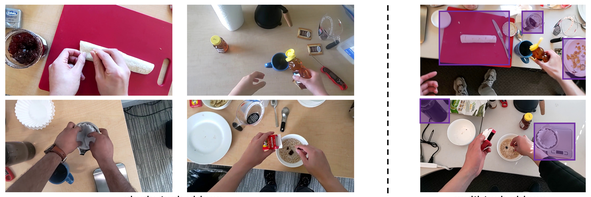
Overview

- Traditional Temporal Action Segmentation (TAS): *single-task per video*
- **Multi-Task Temporal Action Segmentation (MT-TAS):** segment videos where users perform *multiple tasks* by *interleaving their steps*

Challenge 1: frequent task switching and resumption



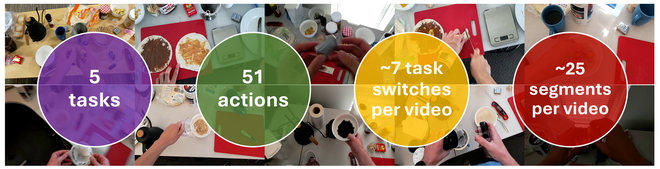
Challenge 2: complex scenes with **objects from multiple tasks**



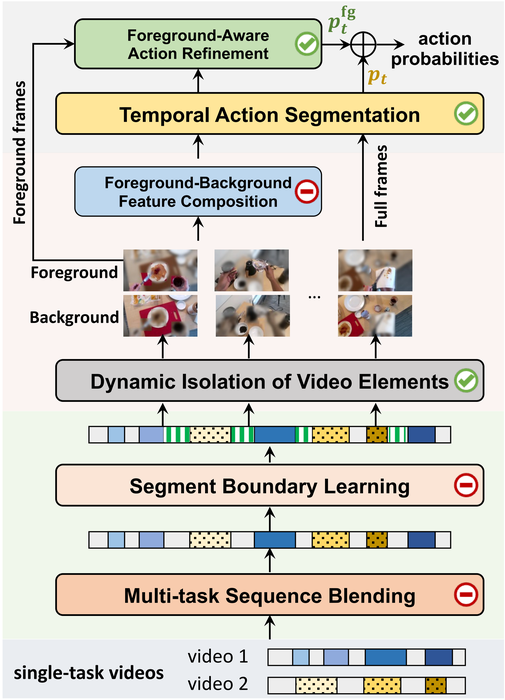
Challenge 3: lack of multi-task data for training and evaluation

MEKA Dataset

- We introduce the ***M**ulti-task **E**gocentric **K**itchen **A**ctivities* (MEKA) dataset as an evaluation benchmark for MT-TAS by extending a single-task TAS dataset EgoPER (Lee et al. CVPR'24)
- MEKA provides multiple modalities: RGB, depth, audio, eye gaze, and hand tracking.



Training Multi-Task TAS Models from Single-Task Videos



Multi-task Sequence Blending (MSB)

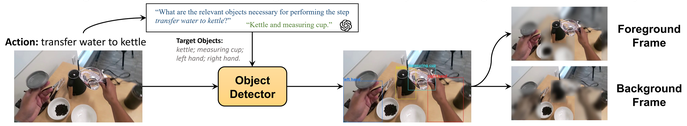
- Create realistic multi-task sequences by asking LLM whether to switch tasks or not

Segment Boundary Learning (SBL)

- Enhance temporal smoothness by predicting boundary feature from its neighbors

Dynamic Isolation of Video Elements (DIVE)

- Disentangle foreground and background regions by detecting action-relevant objects



Foreground-Background Feature Composition (FBFC)

- Learn to reconstruct original feature from foreground and background features

$$\hat{f}_t = \mathcal{D}(f_t^{\text{fg}}, f_t^{\text{bg}}) \qquad \mathcal{L}_{\text{FBFC}} = \sum_t \left\| \hat{f}_t - f_t \right\|^2$$

- Mix backgrounds from different tasks during training to improve generalization

$$\tilde{f}_t^{\text{bg}} = \beta f_t^{\text{bg}} + (1-\beta) f_{t'}^{\text{bg}} \qquad \tilde{f}_t = \mathcal{D}\left(f_t^{\text{fg}}, \tilde{f}_t^{\text{bg}}\right)$$

Single-Task: coffee
Action: fold paper filter

Single-Task: quesadilla
Action: add banana

Multi-Task: coffee and quesadilla
Action: fold paper filter
(with tortilla and cutting board in background)

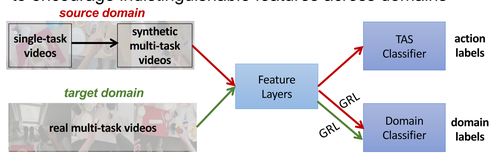
Foreground-Aware Action Refinement (FAAR)

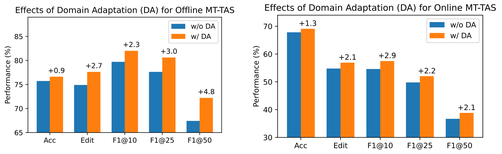
- Apply a light-weight TAS model on foreground frames and combine with initial predictions

MSB	SBL	FBFC	FAAR	Offline TAS model: MSTCN (Farha and Gall. CVPR'19)					Offline TAS model: FACT (Lu and Elhamifar. CVPR'24)					Online TAS model: ProTAS (Shen and Elhamifar. CVPR'24)				
				Acc	Edit	F1@{10,25,50}			Acc	Edit	F1@{10,25,50}			Acc	Edit	F1@{10,25,50}		
baseline				62.3	50.1	52.6	49.3	39.7	53.7	42.9	44.6	41.6	31.7	49.1	24.6	24.1	20.0	12.4
✓				67.8	64.2	68.9	66.8	56.0	72.8	73.1	73.1	70.0	58.4	51.1	30.7	26.5	22.1	12.8
✓	✓			68.8	64.7	70.8	68.8	58.1	73.6	74.5	76.1	73.6	62.3	51.1	33.1	29.3	24.5	14.6
✓	✓	✓		73.7	69.5	73.0	69.9	58.8	73.8	74.4	79.6	76.7	65.5	55.2	37.5	32.8	27.1	18.3
✓	✓	✓	✓	**75.7**	**74.9**	**79.7**	**77.6**	**67.4**	**75.7**	**76.0**	**81.2**	**79.3**	**69.9**	**67.8**	**54.8**	**54.6**	**49.8**	**36.7**

Domain Adaptive MT-TAS

- Leverage **unlabeled multi-task videos** for domain adaptation
- Add a domain classifier with **Gradient Reversal Layer (GRL)** to encourage indistinguishable features across domains

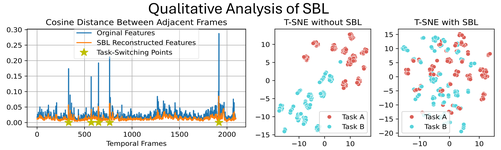
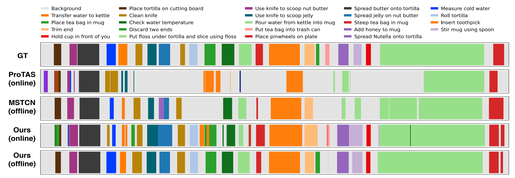




Qualitative Results

Qualitative Analysis of FBFC

Qualitative Analysis of SBL



Visualization of Action Segmentation on a Testing Video