Progress-Aware Online Action Segmentation for Egocentric Procedural Task Videos

Yuhan Shen, Ehsan Elhamifar

Northeastern University

MCADS LAB

CVPR JUNE 17-21, 2024 SEATTLE, WA

Overview

What is Online Action Segmentation?

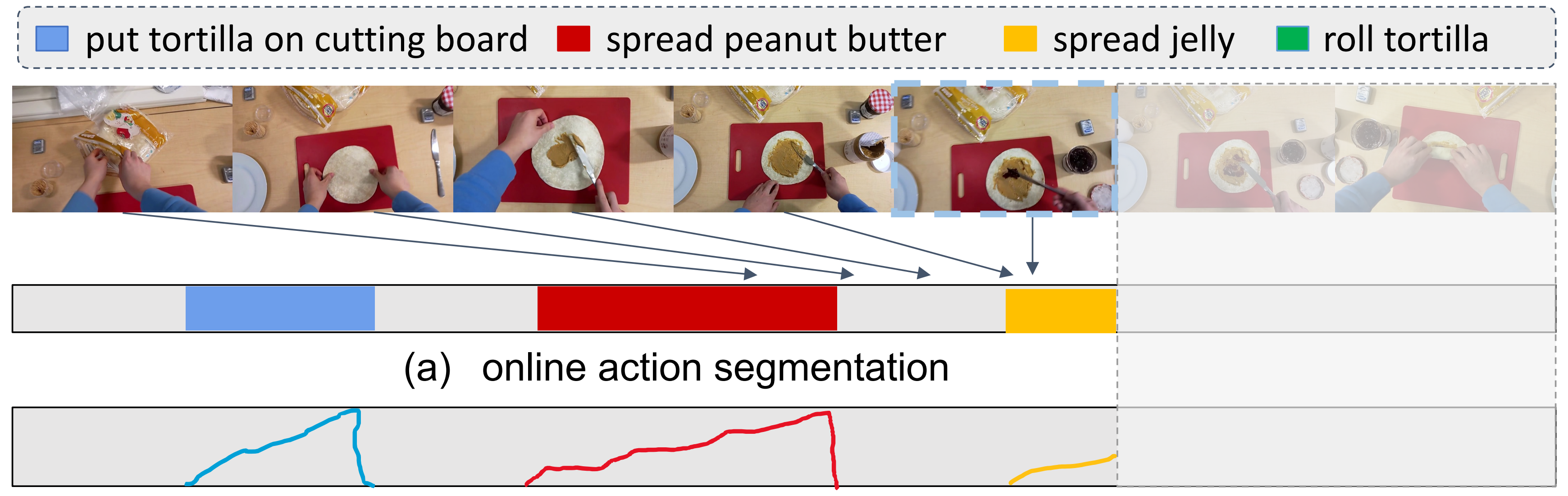
- Recognize and segment actions as frames arrive in real time
- Make predictions without future frame information

Why Online Action Segmentation and Egocentric Videos?

- AR/VR task assistants provide guidance for procedural tasks
- Enable real-time user assistance

How Online Action Segmentation?

- Remove access to future frames during training
- Estimate action progress and leverage task graphs



(a) online action segmentation

(b) action progress prediction

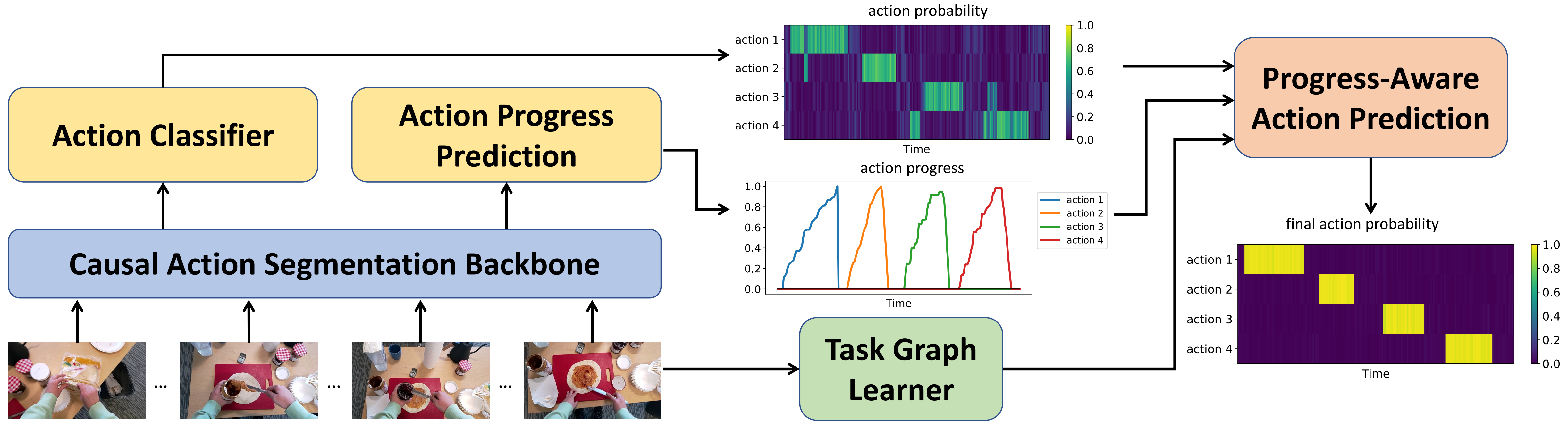
Online recognition step: spread jelly current progress: 20%
"How to complete this step?"
"Here is a reference video for you to follow its progress."
progress: 50%
progress: 75%
progress: 100%
"What should I do next?"
"Roll tortilla."
Task Graph

(c) AR assistant for procedural tasks

Contributions

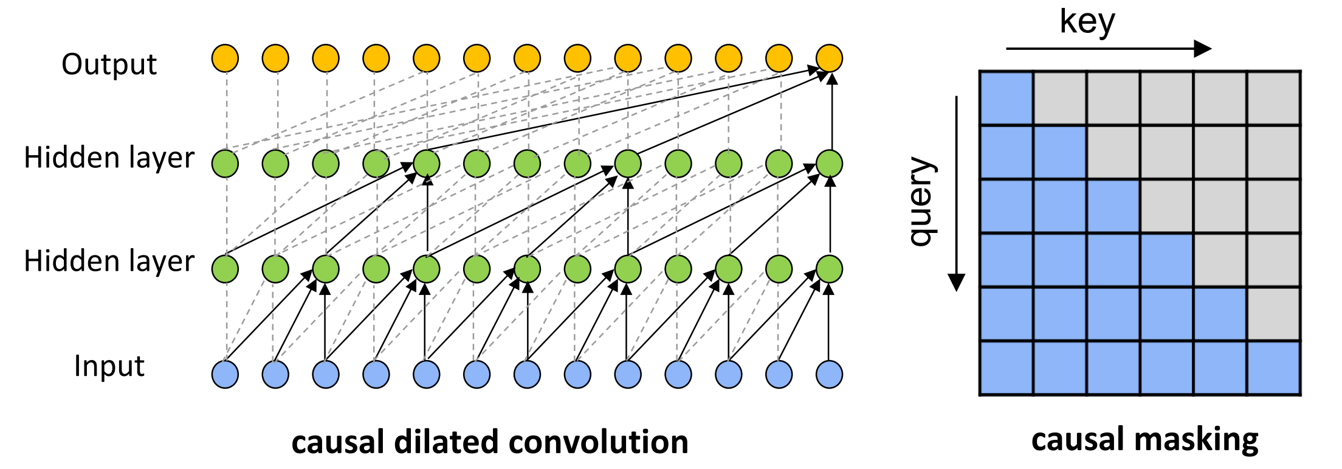
- ○ PRogress-aware Online Temporal Action Segmentation (ProTAS)
  - ○ Address online action segmentation in egocentric procedural task videos
- ○ Leverage task graph learner for online action segmentation
- ○ Achieve significant improvements on three datasets

PRogress-aware Online Temporal Action Segmentation (ProTAS)



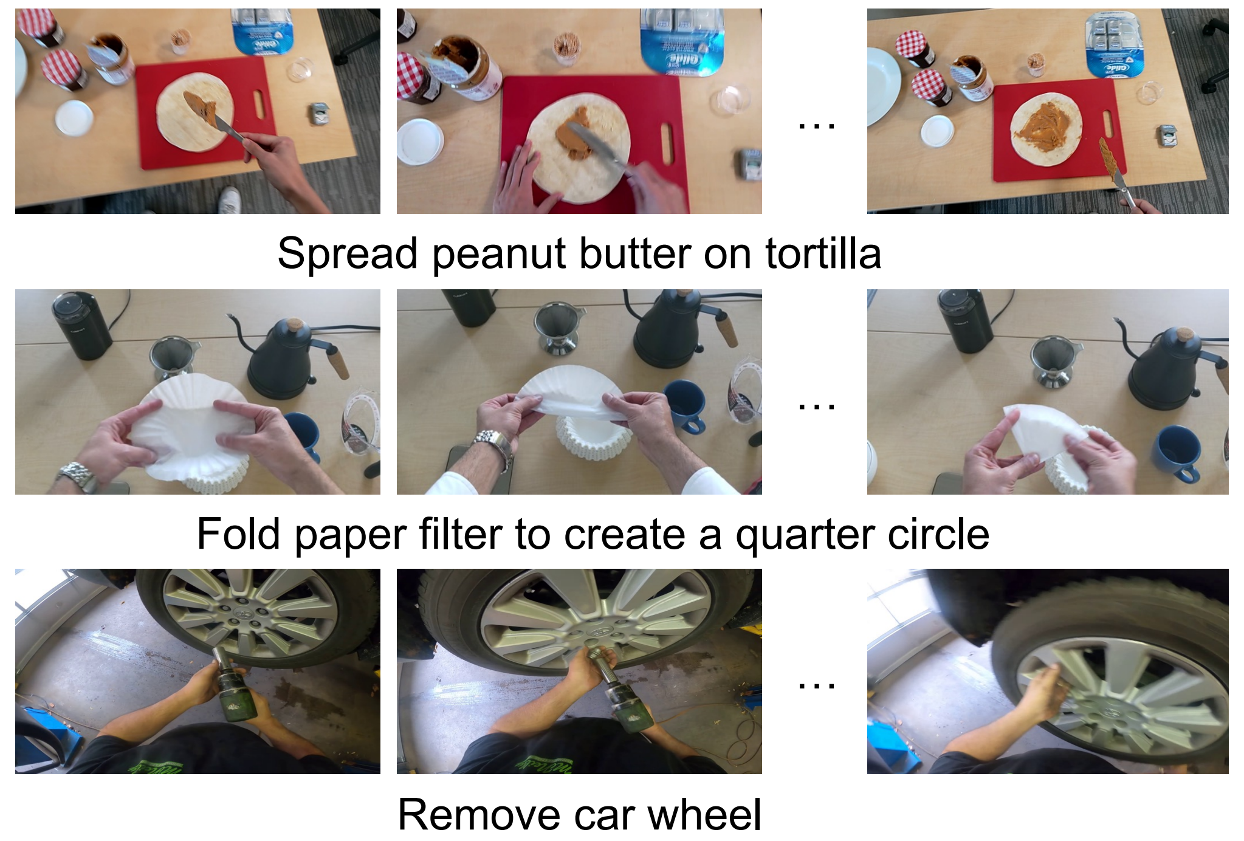
Causal Action Segmentation (CAS)

- Modify existing architectures (TCN-based and Transformer-based) to make them causal
- CE and smoothing loss: $\mathcal{L}_{\text{cls}} + \lambda_{\text{smo}}\mathcal{L}_{\text{smo}}$



causal dilated convolution

causal masking



Spread peanut butter on tortilla

Fold paper filter to create a quarter circle

Remove car wheel

Action Progress Prediction (APP)

- Dynamically estimate progress of ongoing actions via a GRU layer to refine CAS predictions
- Target linear progress: $p^*_{t,k} = \dfrac{t - t_s}{t_e - t_s} \in [0,1]$
- Progress prediction loss:

$$\mathcal{L}_{\text{prog}} = \frac{1}{TK}\sum_{t,k}\left(p_{t,k} - p^*_{t,k}\right)^2$$

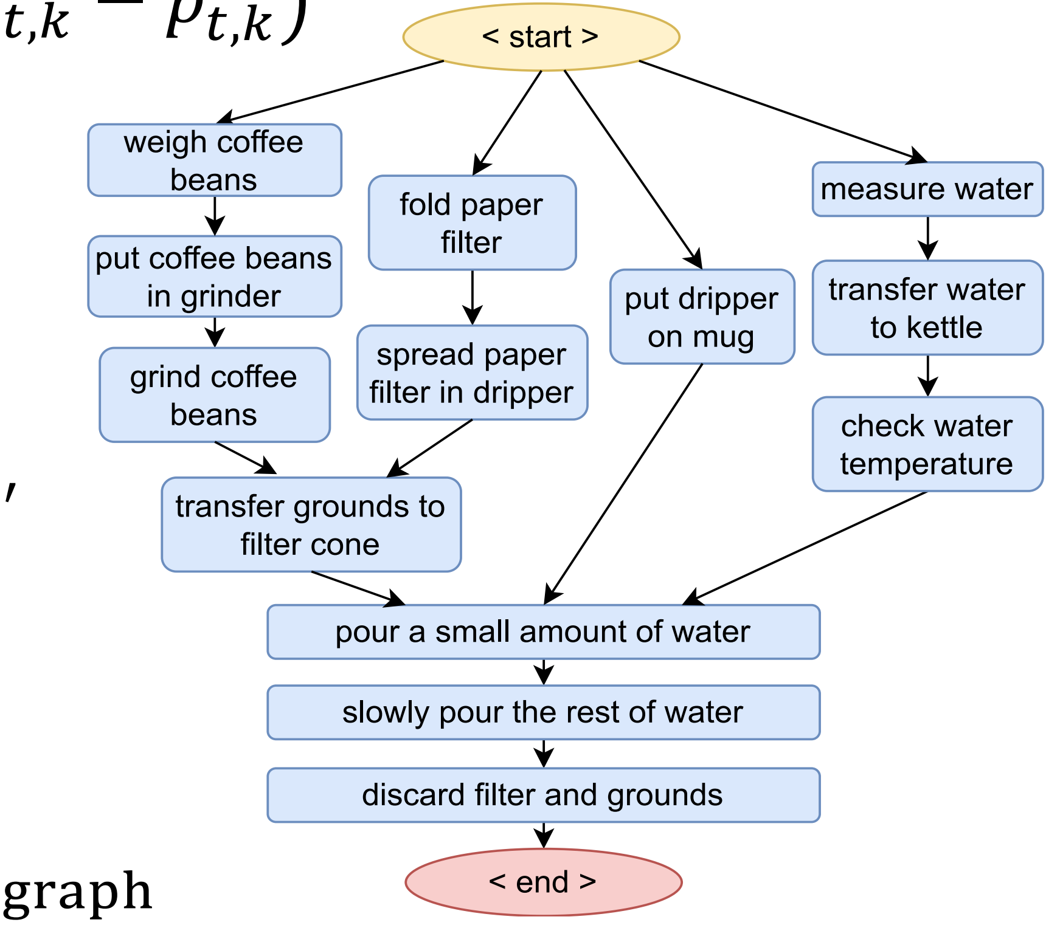
Task Graph (TG)

- Calculate penalty for an action using the completion state of its predecessors and successors:

$$\alpha^p_{t,k} = \sum_{k' \in Predecessor(k)}\left(1 - c_{t,k'}\right),\ \alpha^s_{t,k} = \sum_{k' \in Sucessor(k)} c_{t,k'}$$

- Encourage predictions aligned with task graph:

$$\mathcal{L}_{\text{graph}} = \frac{1}{TK}\sum_{t,k}\left(\alpha^p_{t,k} + \alpha^s_{t,k}\right) \cdot y_{t,k}$$


< start >
weigh coffee beans
put coffee beans in grinder
grind coffee beans
transfer grounds to filter cone
fold paper filter
spread paper filter in dripper
put dripper on mug
measure water
transfer water to kettle
check water temperature
pour a small amount of water
slowly pour the rest of water
discard filter and grounds
< end >

Training Loss: $\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{smo}}\mathcal{L}_{\text{smo}} + \lambda_{\text{prog}}\mathcal{L}_{\text{prog}} + \lambda_{\text{graph}}\mathcal{L}_{\text{graph}}$

Experimental Results

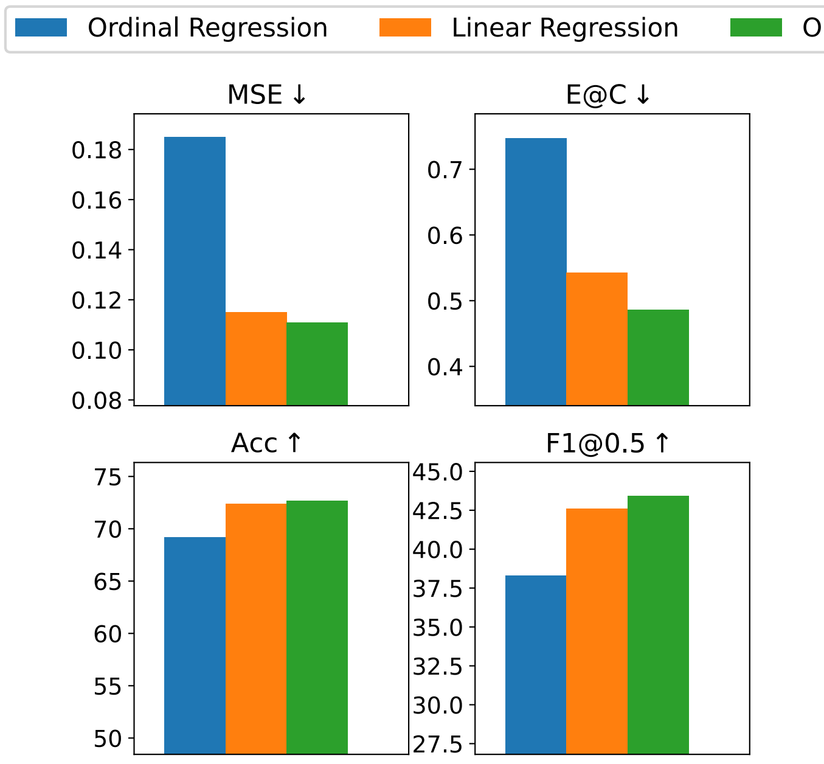
Results on three datasets, MSTCN and ASFormer as backbones

Method	Inference	GTEA			EgoProceL			EgoPER		
		Acc	Edit	F1@0.5	Acc	Edit	F1@0.5	Acc	Edit	F1@0.5
Use MSTCN as backbone										
Base	Offline	76.3	79.0	69.8	69.2	56.9	45.9	83.0	85.9	77.3
Base	Online	47.0	58.8	38.7	18.3	19.9	8.8	20.2	31.0	11.9
CAS	Online	74.0	64.4	56.0	64.5	42.5	33.0	71.8	48.9	39.4
CAS+APP	Online	**76.0**	67.0	57.9	66.3	47.1	35.2	**72.7**	55.0	43.4
CAS+APP+TG	Online	74.3	**69.2**	**59.7**	**67.8**	**48.8**	**35.7**	70.2	**60.7**	**46.3**
Use ASFormer as backbone										
Base	Offline	83.4	84.6	78.9	69.5	59.8	48.8	81.8	88.8	79.9
Base	Online	36.2	48.2	28.3	13.2	17.6	5.4	19.8	24.3	8.8
CAS	Online	77.2	73.3	65.0	64.8	48.1	35.4	70.3	60.6	44.7
CAS+APP	Online	**77.3**	74.0	65.4	66.7	50.7	36.1	70.6	61.2	46.9
CAS+APP+TG	Online	77.0	**74.1**	**66.1**	**68.5**	**52.1**	**36.8**	**71.7**	**62.4**	**48.6**

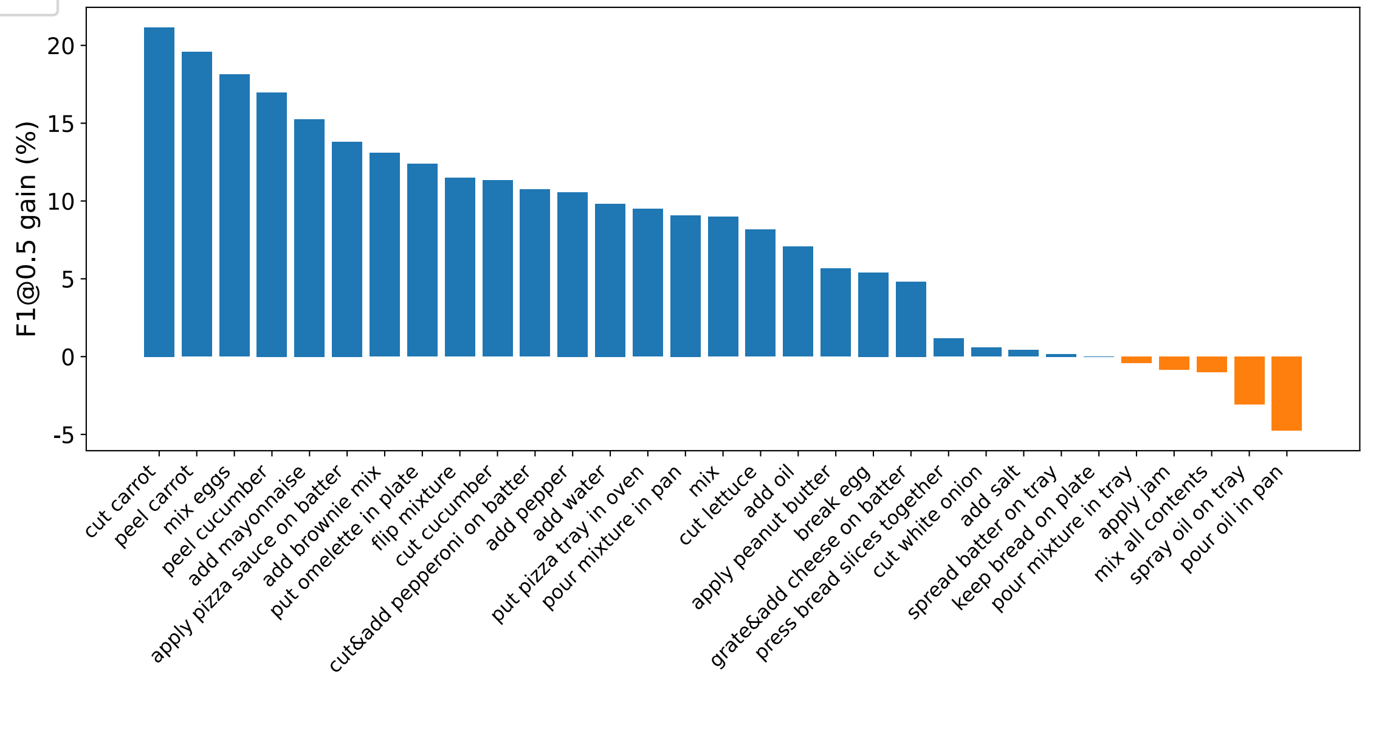
Comparison of different ways of constructing task graph

Task Graph	GTEA			EgoProceL			EgoPER		
	Acc	Edit	F1@0.5	Acc	Edit	F1@0.5	Acc	Edit	F1@0.5
transcript	74.3	69.2	59.7	67.8	48.8	**35.7**	70.2	60.7	46.3
manual	—	—	—	—	—	—	70.4	61.0	46.5
learnable	**74.5**	**69.3**	**60.2**	**68.0**	**48.9**	34.8	**70.6**	**61.4**	**47.1**

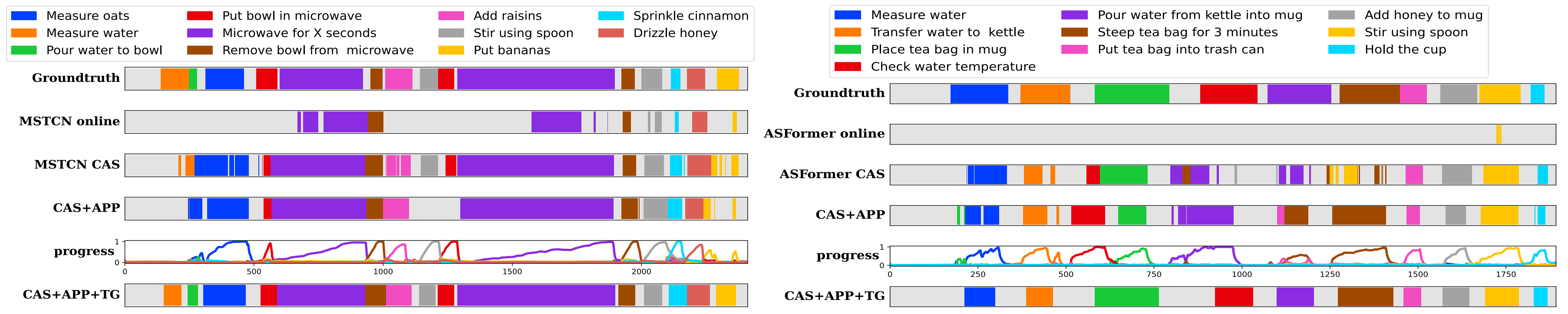
Model designs for APP



Action-wise performance gain



Qualitative results


Measure oats
Measure water
Pour water to bowl
Put bowl in microwave
Microwave for X seconds
Remove bowl from microwave
Add raisins
Stir using spoon
Put bananas
Sprinkle cinnamon
Drizzle honey
Groundtruth
MSTCN online
MSTCN CAS
CAS+APP
progress
CAS+APP+TG
Measure water
Transfer water to kettle
Place tea bag in mug
Check water temperature
Pour water from kettle into mug
Steep tea bag for 3 minutes
Put tea bag into trash can
Add honey to mug
Stir using spoon
Hold the cup
Groundtruth
ASFormer online
ASFormer CAS
CAS+APP
progress
CAS+APP+TG