Dec 2025

TEXAS
The University of Texas at Austin

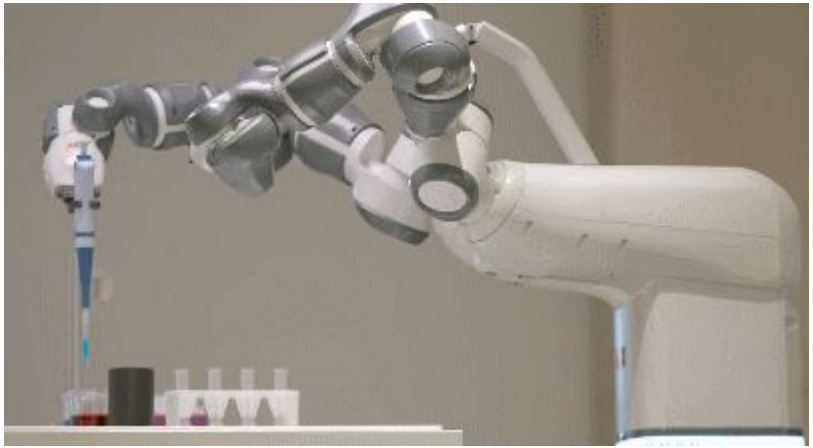
Online State Transition Classification for Procedural Activities

Akhil, Kate, Kelly, Omatharv

The University of Texas at Austin

Procedures!

Pipetting and sorting test tubes



Assembly line monitoring

Actions are decided by object state changes, not just by motion!

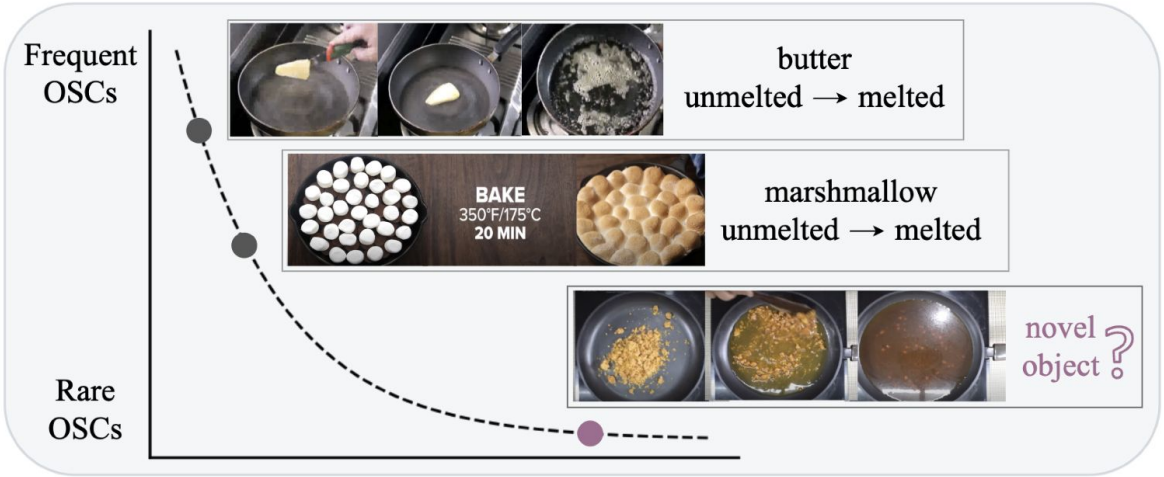
Cutting Tomato =

Uncut → Sliced

Pouring water =

Empty → Full

VidOSC: Learning 3 Stages of an OSC



Object agnostic state prediction

Requires seeing future frames to predict current state

Xue, Zihui, et. al. "Learning object state changes in videos: An open-world perspective." Proceedings of the CVPR Conference. 2024.

Detecting Object state change ***IN REAL TIME*** becomes important!

Why Online?

Why Online?

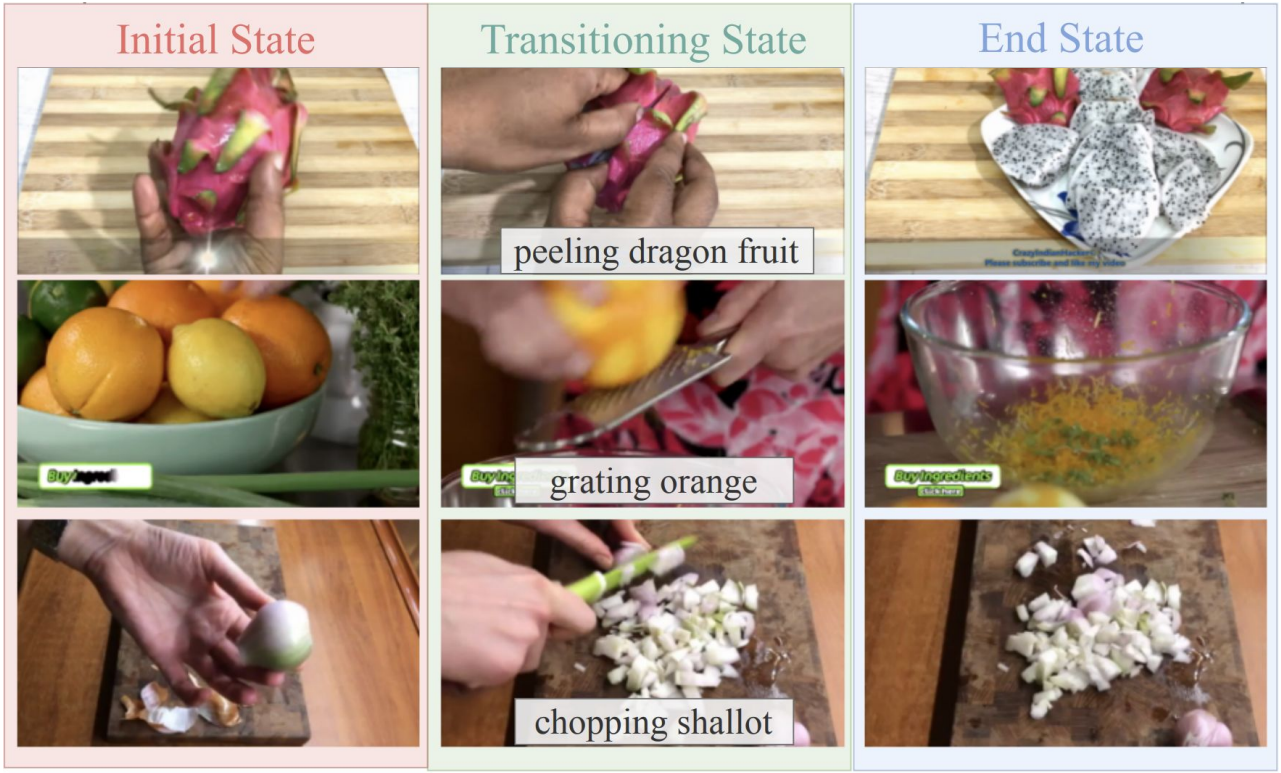
Wait until the end for feedback.

"The butter is melted. Let's proceed!"

Get feedback immediately!

GOAL: provide live feedback to move on from the current step

HowToChange Dataset



HowTo100M Food & Entertaining Category

Actions and passive transformations

5000+ eval videos
400+ unique OSCs

HowToChange Dataset

[0, 13.02] [13.82, 25.49] [29.62, 35.36]

Frame-Wise HowToChange Dataset

[0, 13.02] [13.82, 25.49] [29.62, 35.36]

1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 0 2 2 2 2 2 2 2 2 2 0

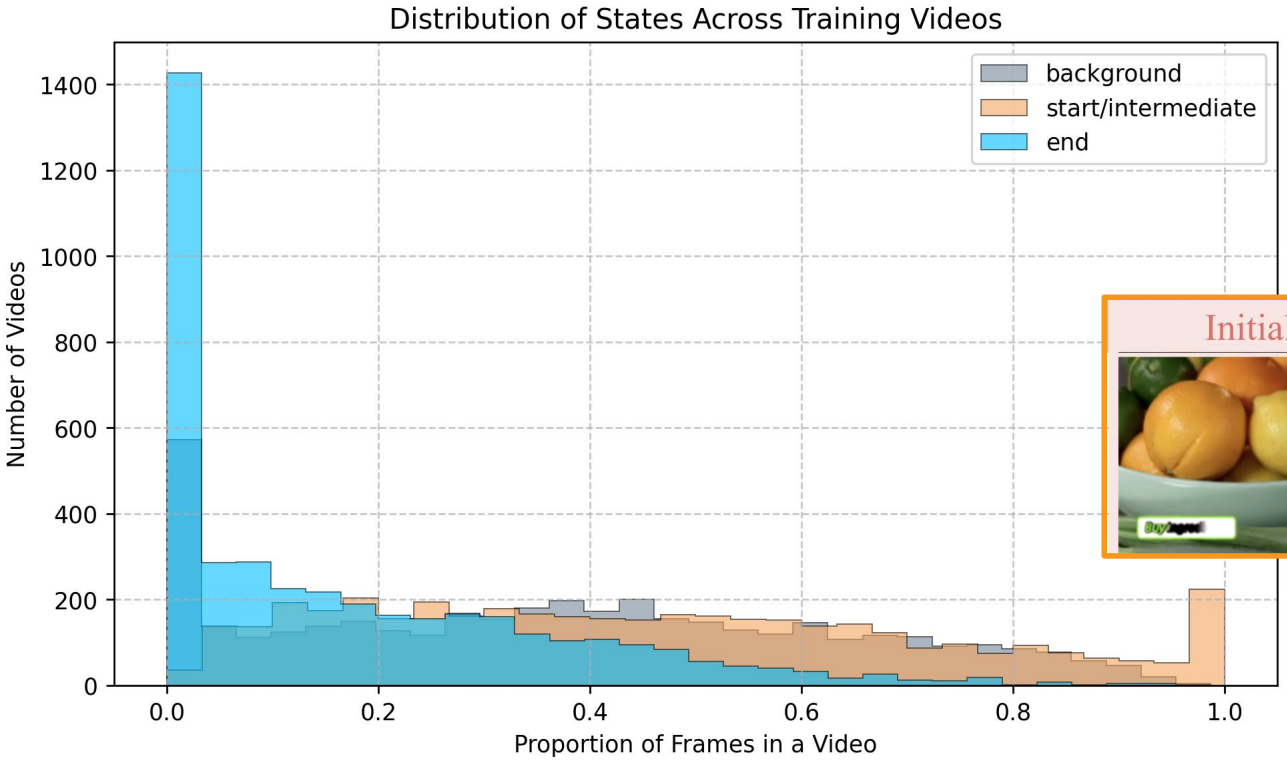
Background State

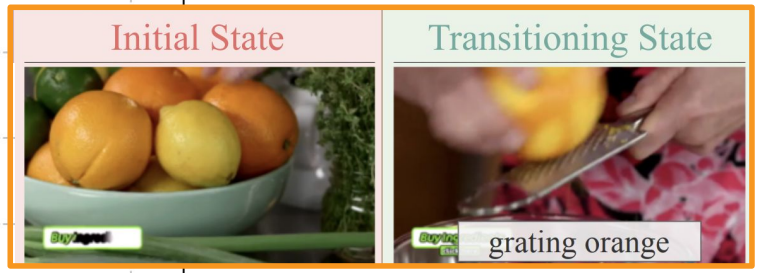
Frame-Wise HowToChange Dataset

1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 0 0 2 2 2 2 2 2 2 2 2 0

Background: **14483 segments**
Initial/transition: **11514 segments**
End: **5757 segments**

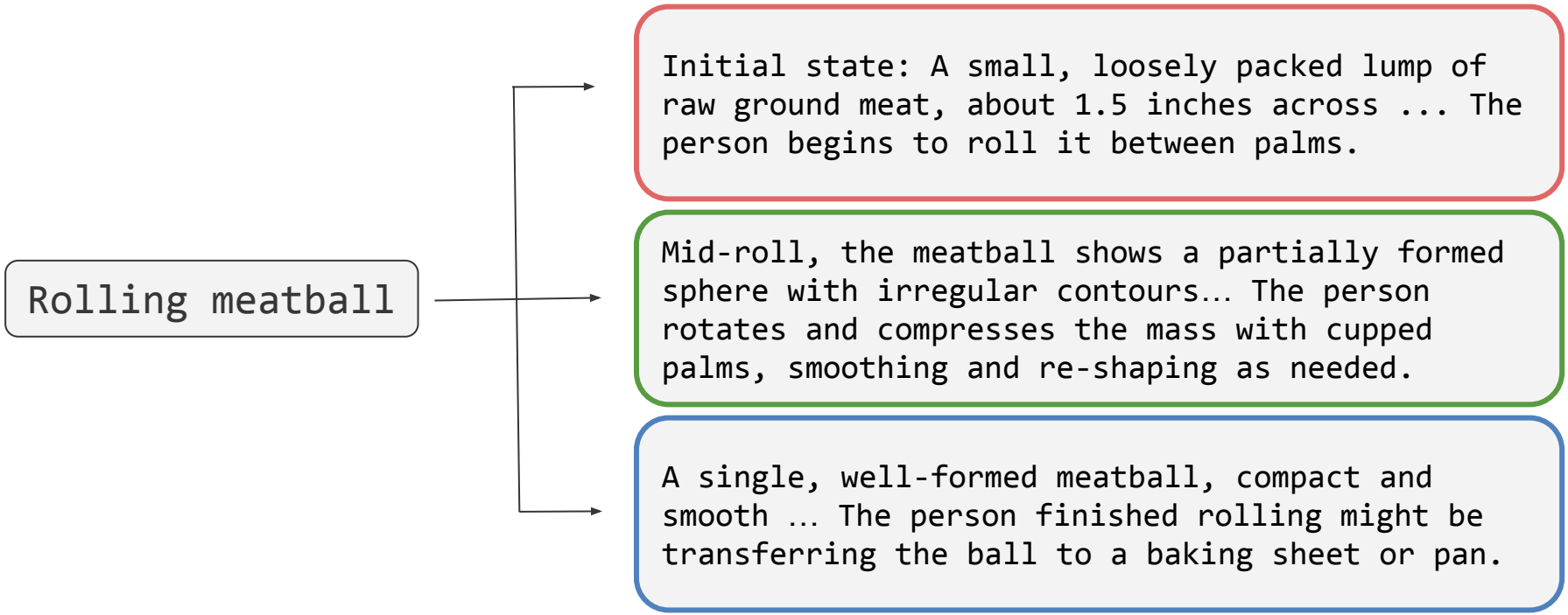
Frame-Wise HowToChange Dataset



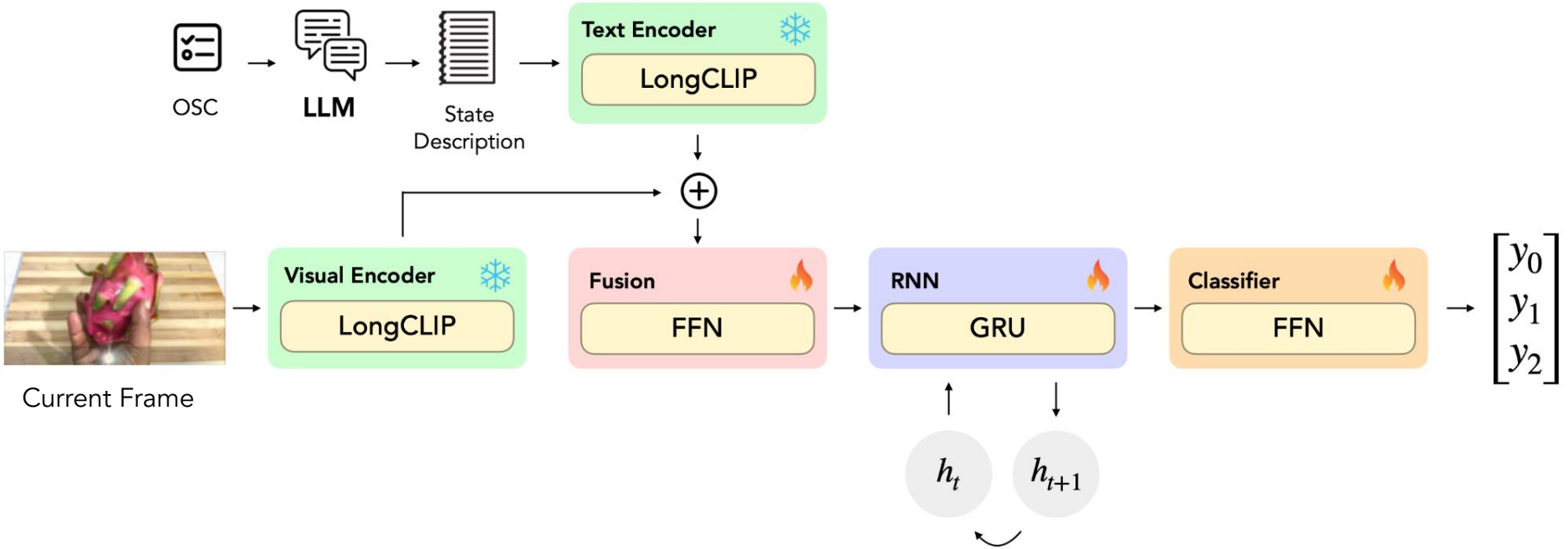


Dataset: Description Bank

Maps OSC to LLM state descriptions



Architecture



Current Frame

Training

Frame-wise prediction on train set of ~4000 videos

Sampled each video at 1 FPS

Loss = Weighted cross-entropy loss + $\lambda_E \cdot$ Entropy Loss + $\lambda_M \cdot$ Monotonicity Loss averaged over all frames

Parameter	**Value**
Learning Rate	0.001
Monotonicity λ	0.1
Entropy λ	0.1
Window size	5
Epochs	1
Batching	False

Results

Evaluated on test set of ~700 videos

	Acc (%) ↑	F1 (%) ↑	End State F1 (%) ↑	End State IoU (%) ↑
VidOSC (Multitask)	38.07	17.21	1.37	1.64
VidOSC (Single task)	53.15	34.23	31.51	36.91
Ours	49.50	31.91	31.25	36.92
Ours (+ LLM desc)	51.54	41.80	34.73	41.36

Results

Evaluated on test set of ~700 videos

	Acc (%) ↑	F1 (%) ↑	End State F1 (%) ↑	End State IoU (%) ↑
VidOSC (Multitask)	38.07	17.21	1.37	1.64
VidOSC (Single task)	**53.15**	34.23	31.51	36.91
Ours	49.50	31.91	31.25	36.92
Ours (+ LLM desc)	51.54	**41.80**	**34.73**	**41.36**

Success Example: Slicing Lemon

Failure Example: Roasting Coconut

Future Work

- Unedited, egocentric video
- Long stream videos
- Unseen or novel state changes
- Non-cooking tasks
- Self or weak supervision

…

Conclusion

- **Motivation**: provide live feedback to move on from the current step based on state change detection
- **Method**: supervised learning + recurrent network + LLM-generated state description as knowledge bank
- **Results**: outperformed offline OSC segmentation baselines on end state detection