Learning Object State Change in Videos: An Open-World Perspective

Zihui (Sherry) Xue[1,2] Kumar Ashutosh[1,2] Kristen Grauman[1,2]

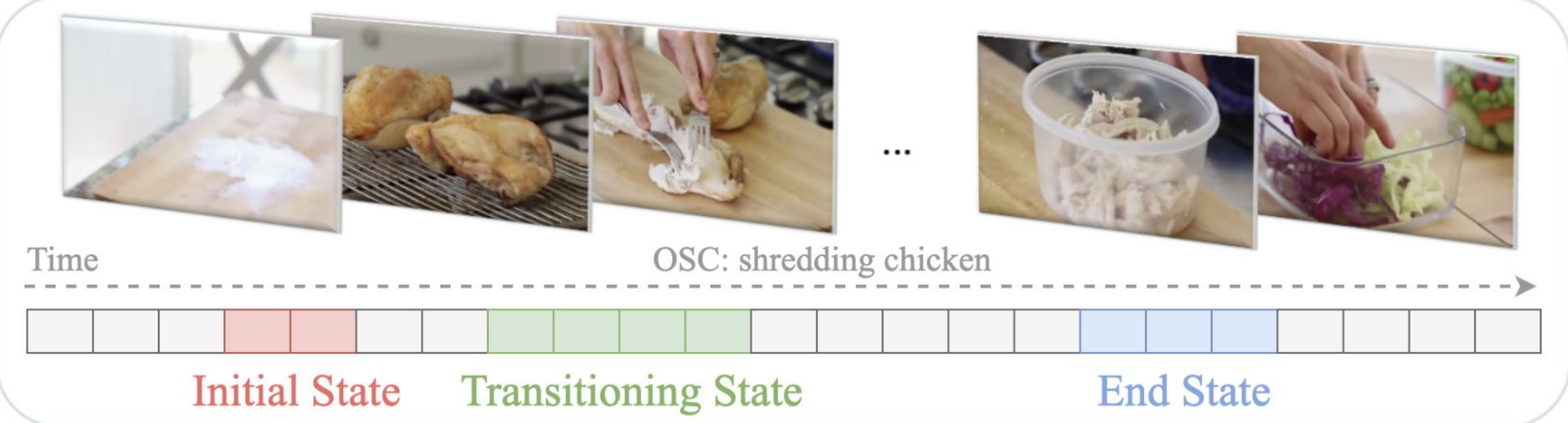
[1]UT Austin [2]FAIR, Meta

CVPR JUNE 17-21, 2024 SEATTLE, WA

See our website for data, code & qualitative videos →

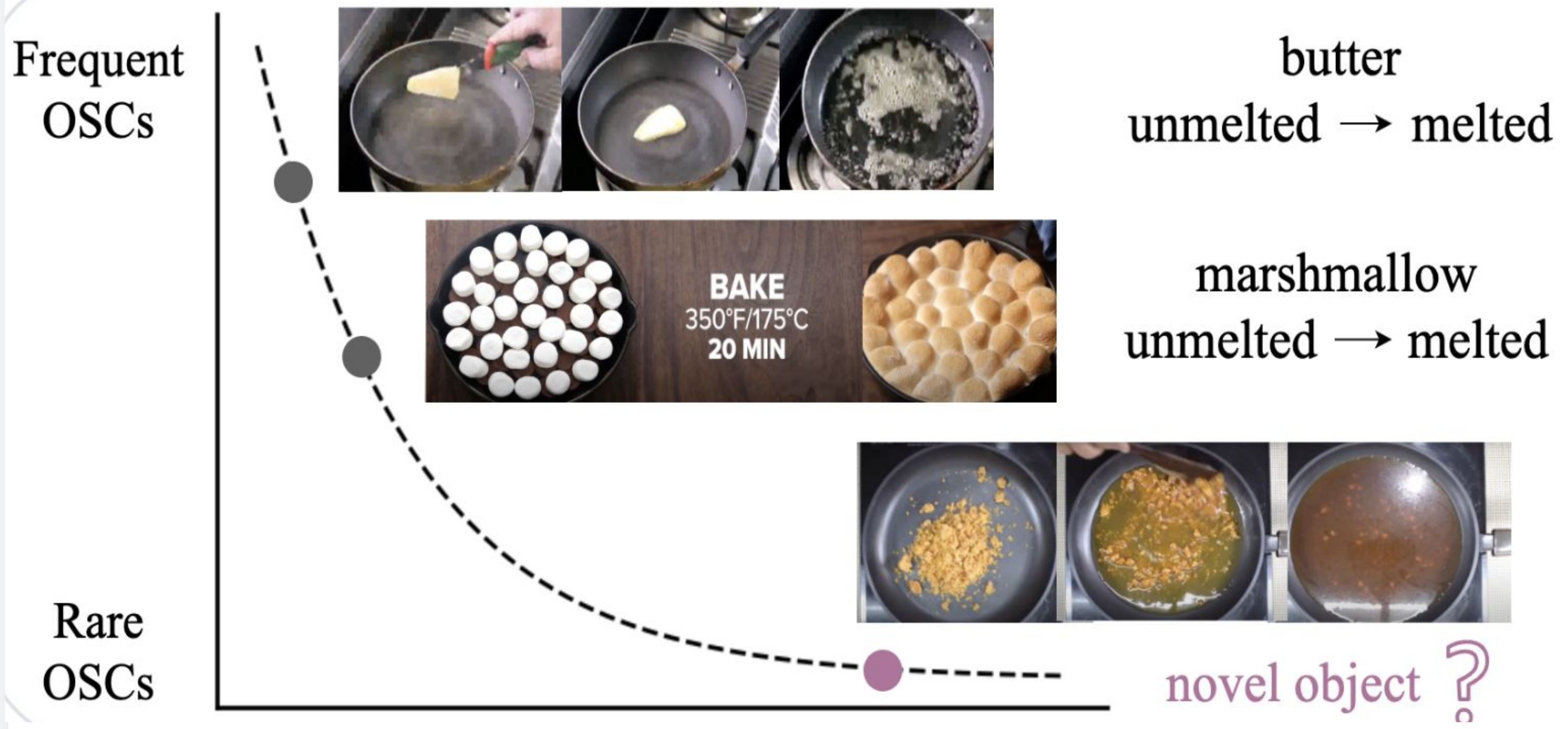
Video Object State Change (OSC)

Objective: temporally localize an object's three states (initial, transitioning and end) from a video



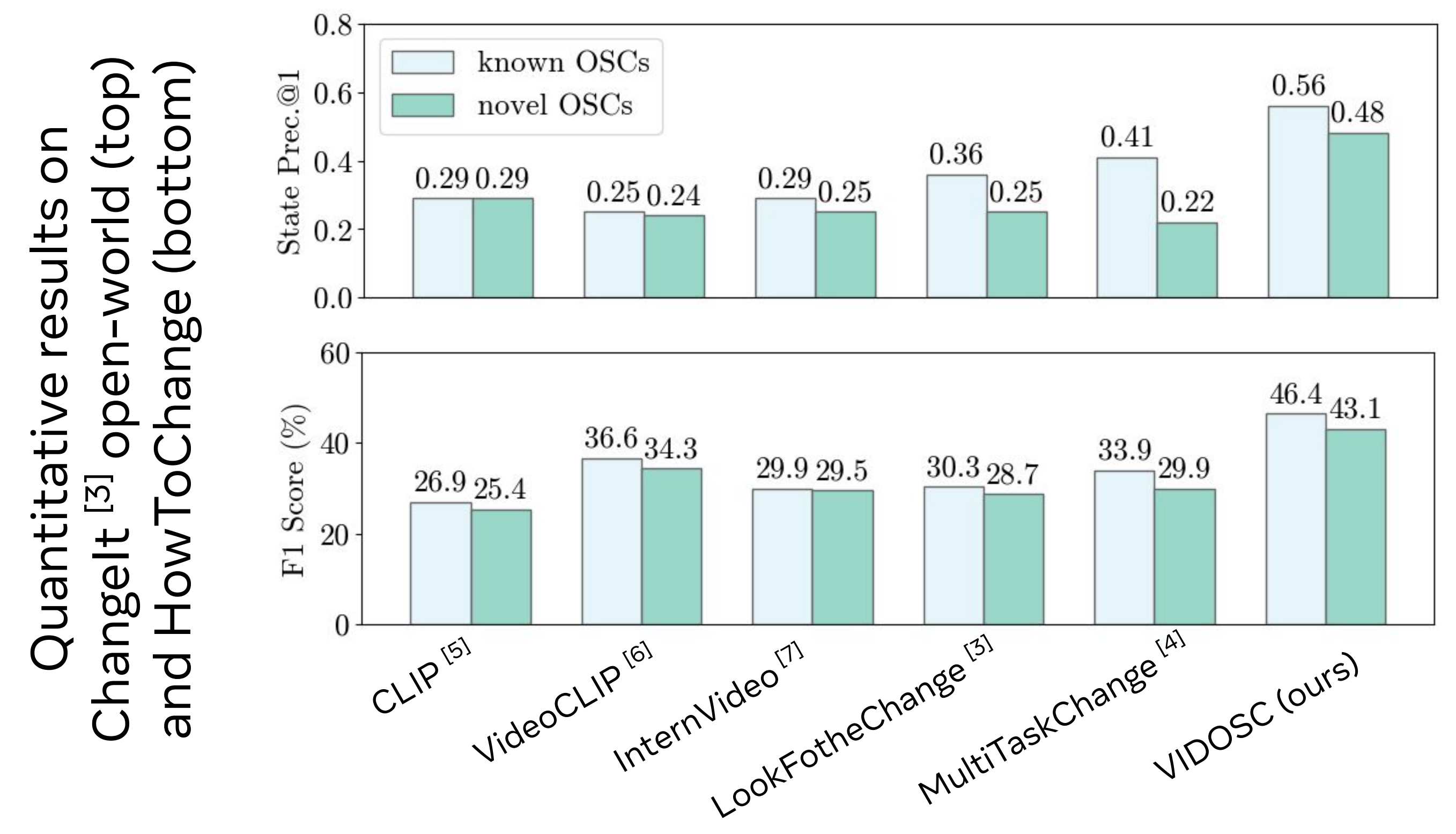
Motivation

- OSCs naturally exhibit a **long-tail**. Certain OSCs, such as melting butter, are frequently shown in videos while others like melting jaggery might be rarely seen.
- Prior works assume a **closed** vocabulary, limited to identifying state changes for objects observed during training.

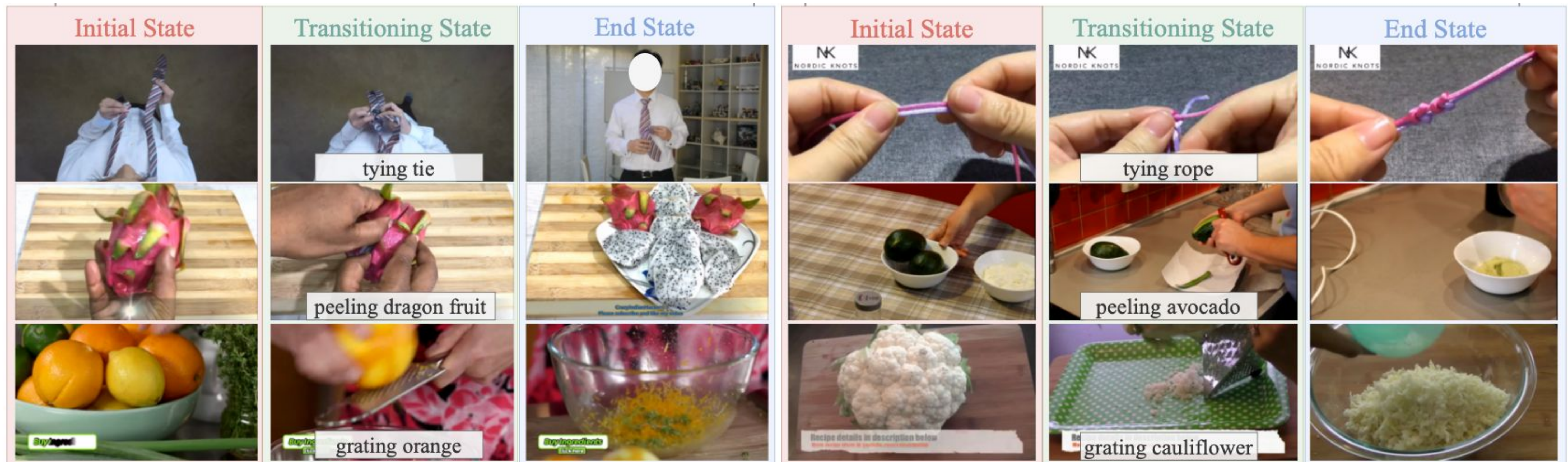


Open-world formulation

Quantitative Results



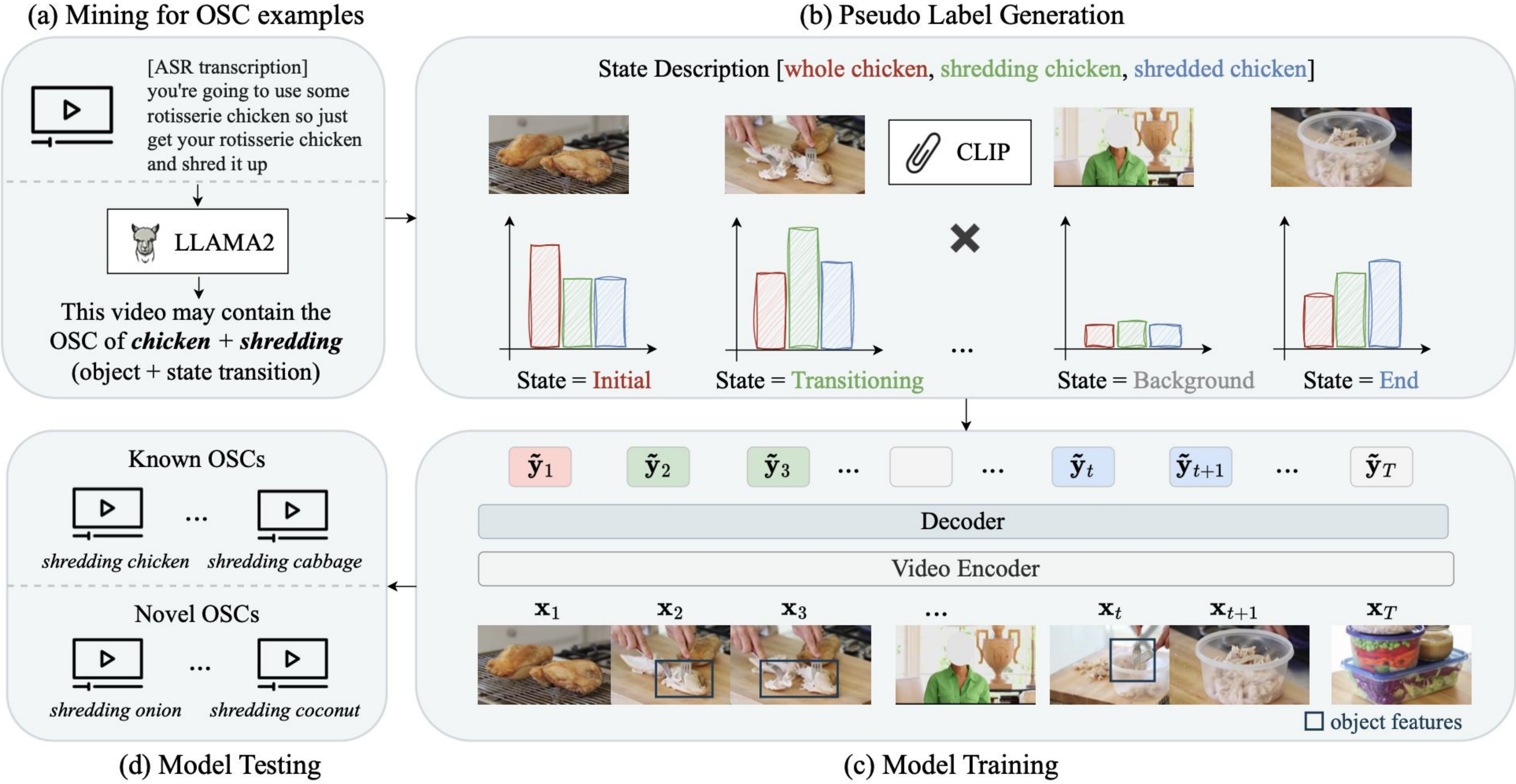
Qualitative Results



Top-1 frame predictions (known OSCs)

Top-1 frame predictions (novel OSCs)

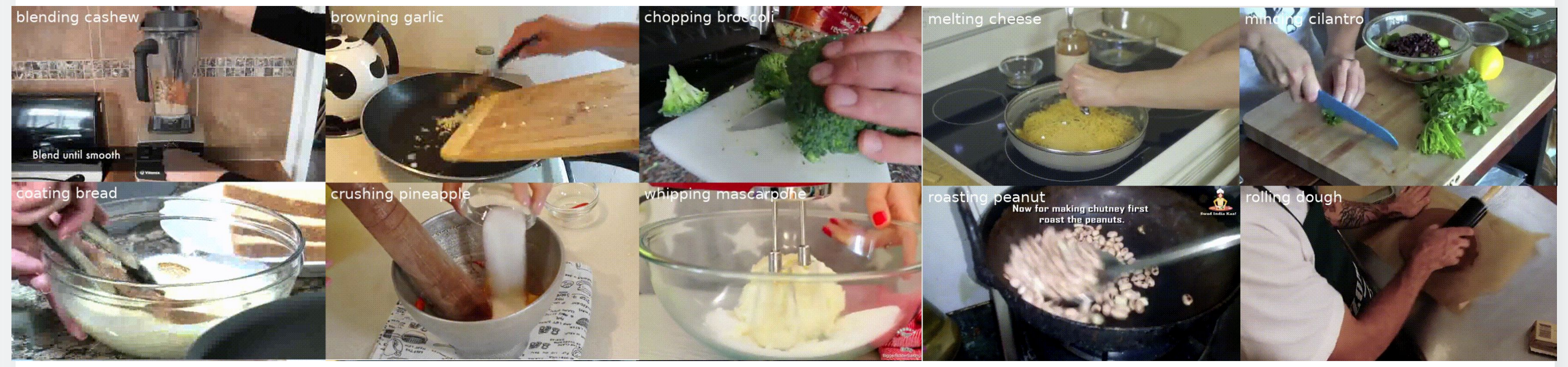
Framework Overview


(a) Mining for OSC examples
[ASR transcription] you're going to use some rotisserie chicken so just get your rotisserie chicken and shred it up
LLAMA2
This video may contain the OSC of chicken + shredding (object + state transition)
(b) Pseudo Label Generation
State Description [whole chicken, shredding chicken, shredded chicken]
CLIP
State = Initial
State = Transitioning
...
State = Background
State = End
Known OSCs
shredding chicken
...
shredding cabbage
Novel OSCs
shredding onion
...
shredding coconut
(d) Model Testing
ỹ1 ỹ2 ỹ3 ... ... ỹt ỹt+1 ... ỹT
Decoder
Video Encoder
x1 x2 x3 ... xt xt+1 xT
□ object features
(c) Model Training

(a) Leverage ASR transcriptions and LLM for automatic OSC mining from instructional videos;
(b) Employ OSC textual descriptions with VLM for pseudo label generation;
(c) Video OSC model for object-agnostic state predictions (shared state vocabulary, temporal modeling, object-centric features);
(d) Test with open-world formulation, evaluating performance on both known and novel OSCs.

HowToChange Dataset

The first open-world benchmark for video OSC localization

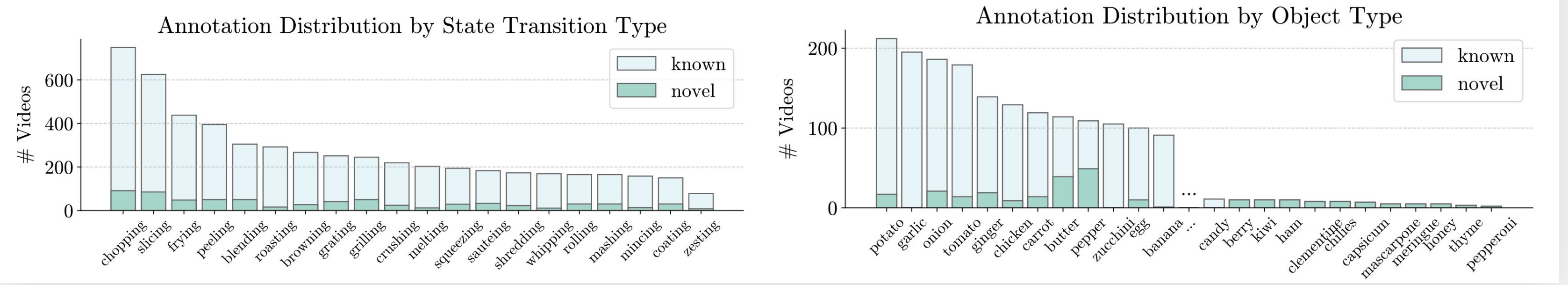


	#Obj	#ST	#OSC	#Obj per ST	#Video	GT Label?
Alayrac et al. [1]	5	6	7	1.2	630	✅
TaskFluent [2]	25	14	32	2.3	809	✅
ChangeIt (Train) [3]	42	27	44	1.6	34,428	
ChangeIt (Eval) [3]	42	27	44	1.6	667	✅
HowToChange (Train)	122	20	318	15.9	36,075	
HowToChange (Eval)	134	20	409	20.5	5,424	✅

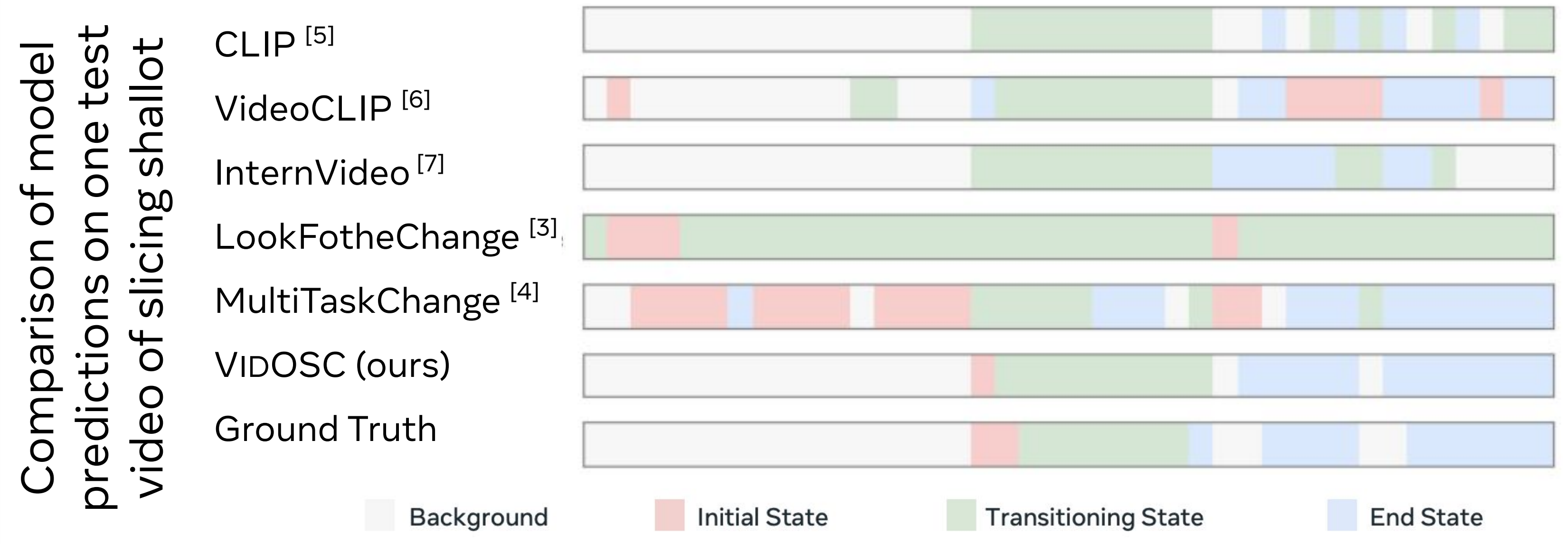
Obj = object, ST = state transition

Features:
- Scale jump in the number of OSCs (#OSC) and videos (#Video)
- Wide variety of objects per state transition (#Obj per ST)

Ground truth annotation distribution



Qualitative Results



[1] Alayrac et al., Joint discovery of object states and manipulation actions, ICCV 17.
[2] Liu et al., Jointly recognizing object fluents and tasks in egocentric videos, ICCV 17.
[3] Soucek et al., Look for the Change: Learning Object States and State-Modifying Actions from Untrimmed Web Videos, CVPR 22.
[4] Soucek et al., Multi-Task Learning of Object State Changes from Uncurated Videos, arXiv 22.
[5] Radford et al., Learning transferable visual models from natural language supervision, ICML 21.
[6] Xu et al., VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding, EMNLP 21.
[7] Wang et al., InternVideo: General Video Foundation Models via Generative and Discriminative Learning, arXiv 22