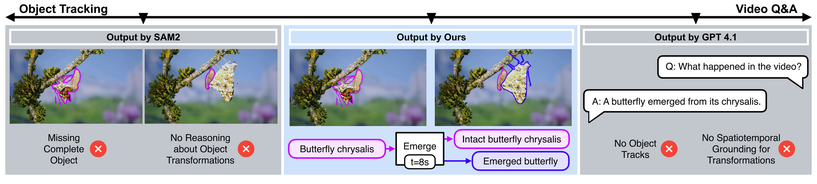
Tracking and Understanding Object Transformations

Yihong Sun, Xinyu Yang, Jennifer J. Sun, Bharath Hariharan
Cornell University

Project Page

Introduction



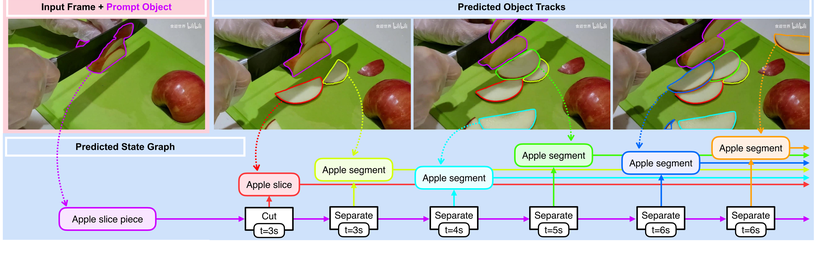
Motivation:

- Existing trackers **fail** when objects transform (E.g., chrysalis → butterfly).
- We seek a system that can output **complete object tracks** with **spatiotemporal grounding** of transformation **descriptions**.

Insight:

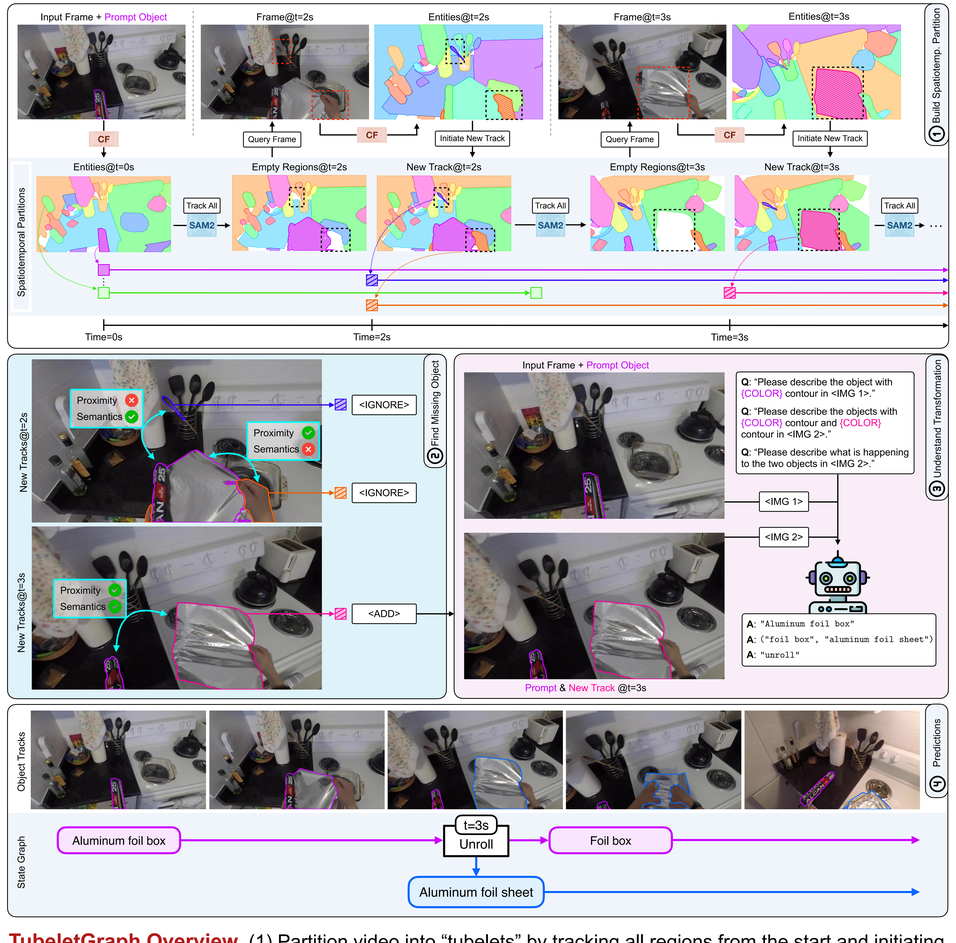
- Object tracking errors are **asymmetric**: false negatives >> false positives
- The missing objects (FN) are often **caused by transformations**.
- Recovering them reveals **when and where** transformations occurred!

Contributions



- We introduce **Track Any State**: tracking objects through transformations while detecting and describing state changes, accompanied by **VOST-TAS**, a new benchmark dataset.
- We propose **TubeletGraph**: a zero-shot framework that recovers missing objects post-transformation by using a **spatiotemporal partition** of the video and constructs a **state graph** to detect and describe the underlying transformations.
- We demonstrate both **state-of-the-art** tracking performance under transformations as well as **effective** detection and description of the transformation itself.

Methods



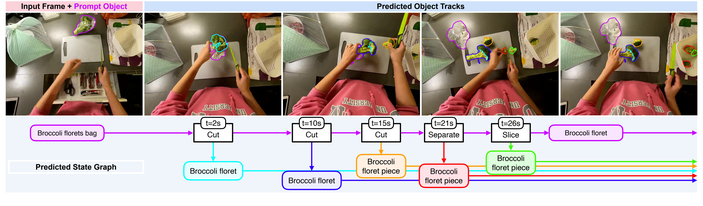
TubeletGraph Overview. (1) Partition video into "tubelets" by tracking all regions from the start and initiating new tracks when untracked pixels emerge; (2) Filter candidates using semantic and spatial proximity priors; (3) Query VLM to describe transformations; (4) Predict state graph with action verbs and object descriptions.

Results

Method	Tracking						State Graph					
	VOST		VSCOS		M^3-VOS		Sem. Acc.		Temp. Loc.		Overall	
	$\mathcal{J}$	$\mathcal{J}_{tr}$	$\mathcal{J}$	$\mathcal{J}_{tr}$	$\mathcal{J}$	$\mathcal{J}_{tr}$	$\mathcal{S}_V$	$\mathcal{S}_O$	$\mathcal{T}_P$	$\mathcal{T}_R$	$\mathcal{H}_{ST}$	$\mathcal{H}$
SAM2	46.1	29.4	72.5	67.1	71.3	59.8	-	-	-	-	-	-
SAM2Long	46.4	29.1	73.0	68.6	70.2	58.7	-	-	-	-	-	-
SAM2.1	48.4	32.4	72.0	66.9	71.3	59.3	-	-	-	-	-	-
DAM4SAM	48.8	33.6	71.3	66.0	72.2	61.3	-	-	-	-	-	-
SAMURAI	49.8	34.0	71.8	66.9	72.6	61.6	-	-	-	-	-	-
Ours	**51.0**	**36.9**	**75.9**	**72.2**	**74.2**	**64.4**	81.8	72.3	**43.1**	**20.4**	**12.0**	**6.5**

Quantitative Results

- **State-of-the-art tracking**: Outperforms all baselines on VOST & VSCOS
- **State graph quality**: While semantic accuracy and temporal localization are promising, spatiotemporal (H_{ST}) and overall (H) recall of transformation remains challenging due to passive detection limitations.



Qualitative Results

- TubeletGraph successfully recovers missing object components post-transformation while simultaneously generating accurate state graphs.

Conclusions

- First system to jointly track objects through transformations while detecting and describing state changes with spatiotemporal grounding.
- Future directions include improving transformation recall and reducing computational cost of exhaustive tubelet tracking.

Acknowledgements

- This research is based upon work supported in part by the National Science Foundation (IIS-2144117, IIS-2107161 and IIS-2505098).
- Yihong Sun is supported in part by an NSF graduate research fellowship.