VTCD: Understanding Video Transformers via Universal Concept Discovery

Matthew Kowal[1,2,3], Achal Dave[1], Rares Abrus[1], Adrien Gaidon[1]
Konstantinos G. Derpanis[2,3,4] , Pavel Tokmakov[1]
[1]Toyota Research Institute, [2]York University, [3]Vector Institute, [4]Samsung AI Centre Toronto

Highlight CVPR SEATTLE, WA JUNE 17-21, 2024

Project Page
yorkucvil.github.io/VTCD

1. **VTCD** - the first algorithm for unsupervised concept discovery in video transformers
2. We discover common processing patterns among several models
3. We apply VTCD for fine-grained action recognition and zero-shot semi-VOS

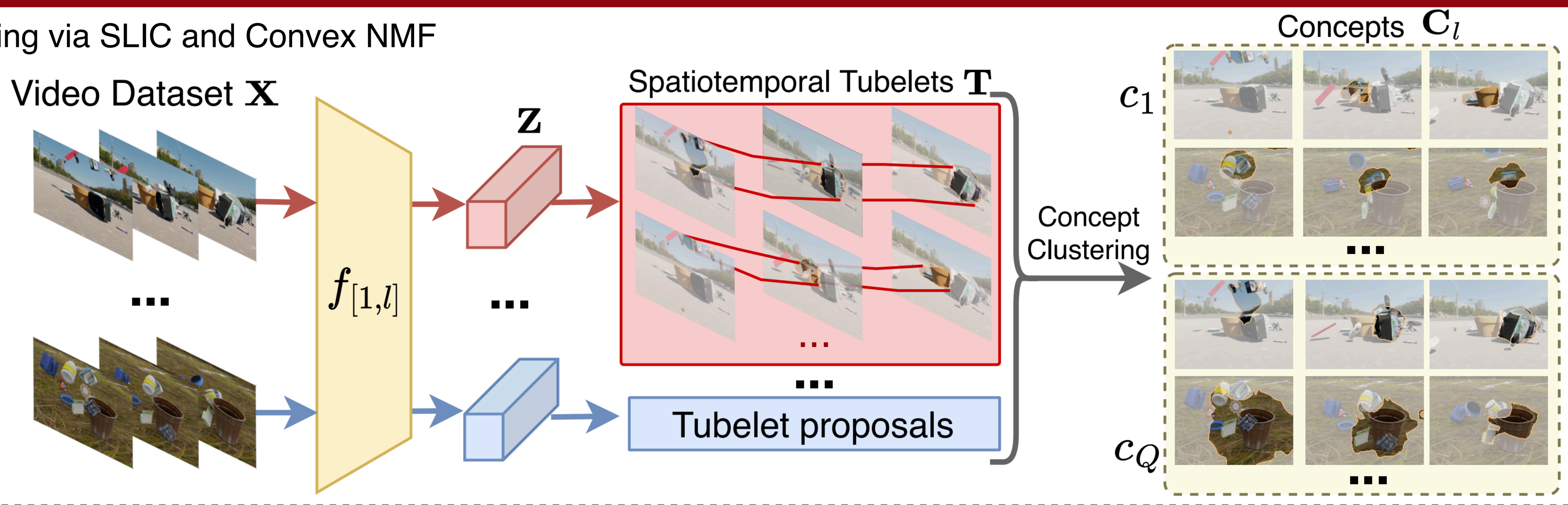
Motivation and Research Questions

This paper introduces a novel concept-discovery approach for video transformers and aims to answer the following questions:

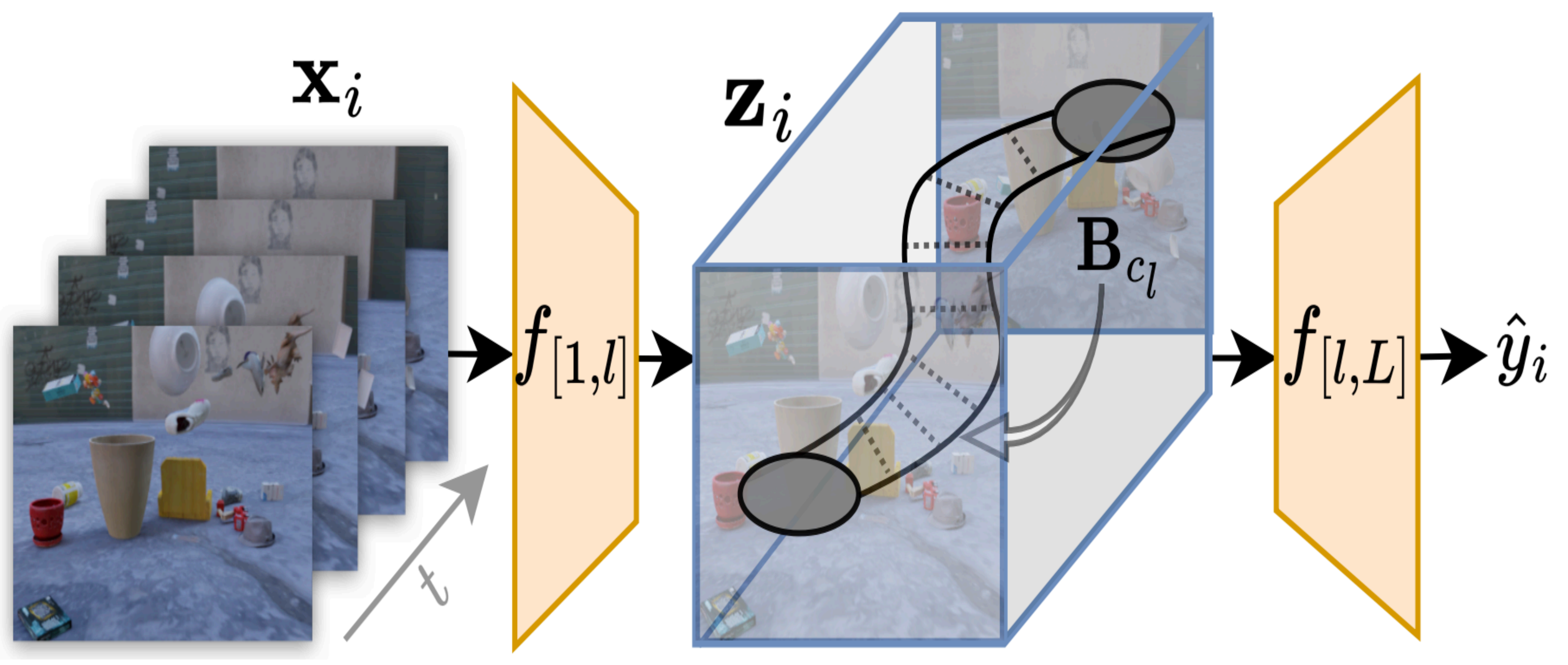
1. What spatiotemporal concepts are used by self-supervised transformers for complex video understanding tasks?
2. Are any concepts universal to models trained with different objectives? (E.g., supervised vs self-supervised)
3. How can these discovered concepts be leveraged for downstream video applications?

Method

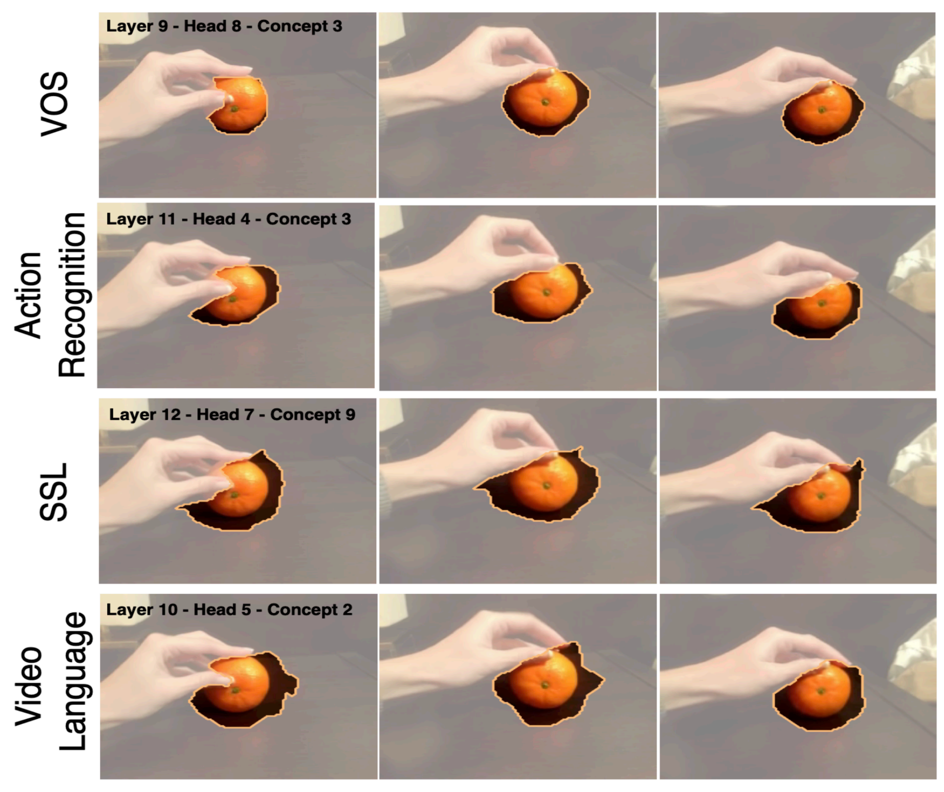
i) Concept clustering via SLIC and Convex NMF



ii) Concept Randomized Importance Sampling (CRIS)



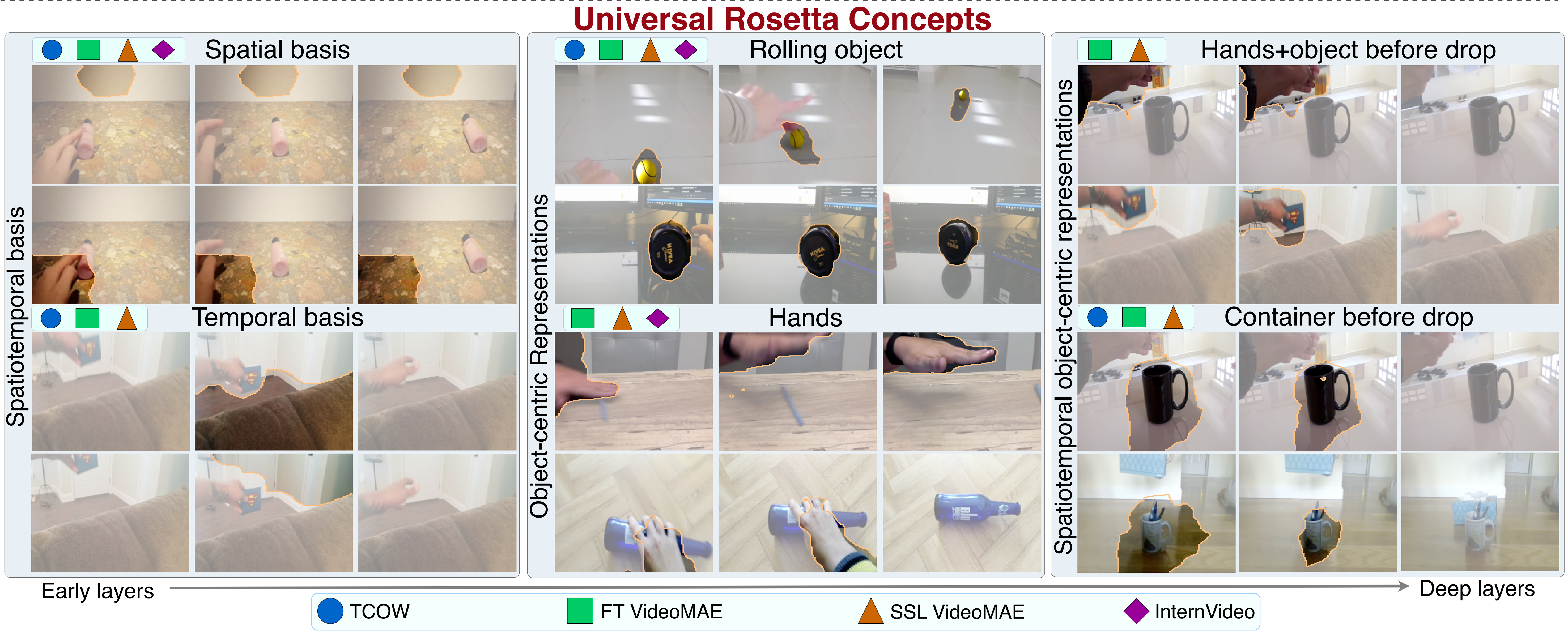
iii) Find universal *Rosetta* concepts across models



Results

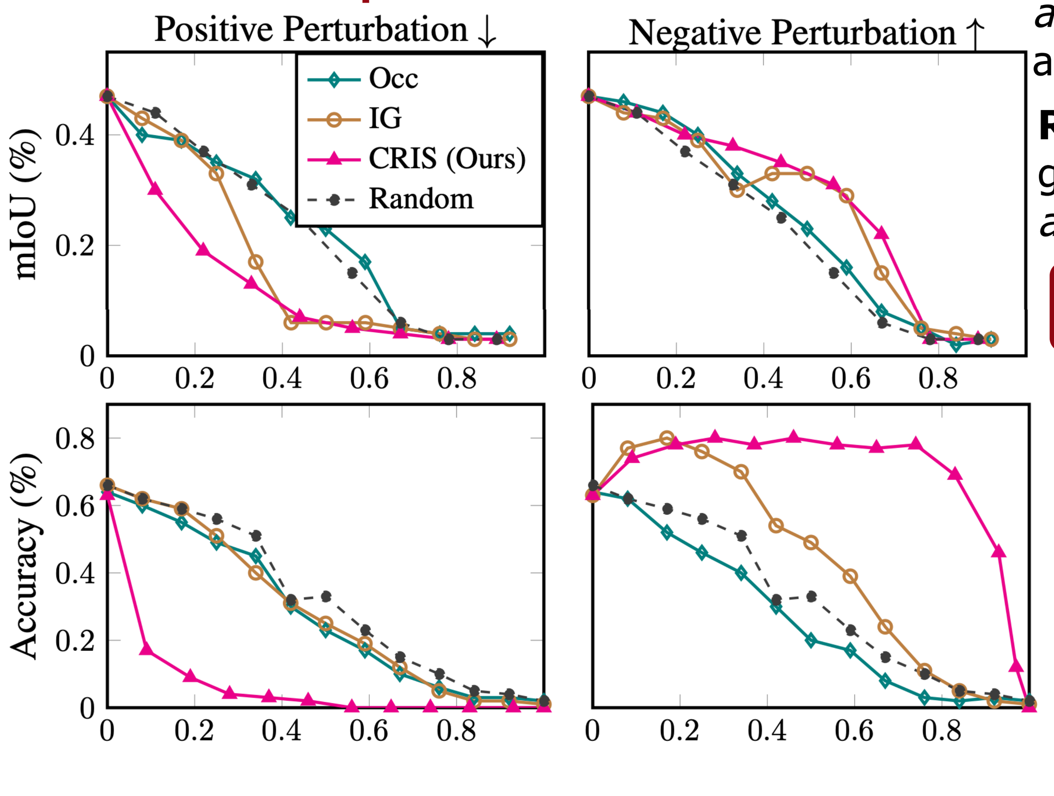
Most Important Concepts

Universal Rosetta Concepts



VTCD Validation

Concept Attribution Curves



Left: We validate VTCD with *concept attribution curves* - outperforming occlusion and gradient-based methods

Right: Comparing discovered concepts to groundtruth masks show our *tubelets + CRIS are superior* to baselines

Tubelet Proposal Validation

Model	TCOW Positive ↓	TCOW Negative ↑	VideoMAE Positive ↓	VideoMAE Negative ↑
Baseline + Occ	0.174	0.274	0.240	0.300
Baseline + CRIS	0.166	0.284	0.157	0.607
VTCD (Ours)	**0.102**	**0.288**	**0.094**	**0.625**

Applications

Zero-Shot Semi-VOS (DAVIS16)

Features	VTCD	VTCD + SAM
VideoMAE-SSL	45.0	68.1
VideoMAE	43.1	66.6
InternVideo	45.8	68.0

We use VTCD concepts to perform zero-shot semi-VOS on DAVIS16 with models *not trained for segmentation*

Model Pruning (SSv2)

Model	Accuracy ↑	GFLOPs ↓
Baseline	37.1	180.5
VTCD 33% Pruned	**41.4**	121.5
VTCD 50% Pruned	37.8	**91.1**

Pruning the least important heads for a subset of SSv2 classes improves *efficiency* and *performance*