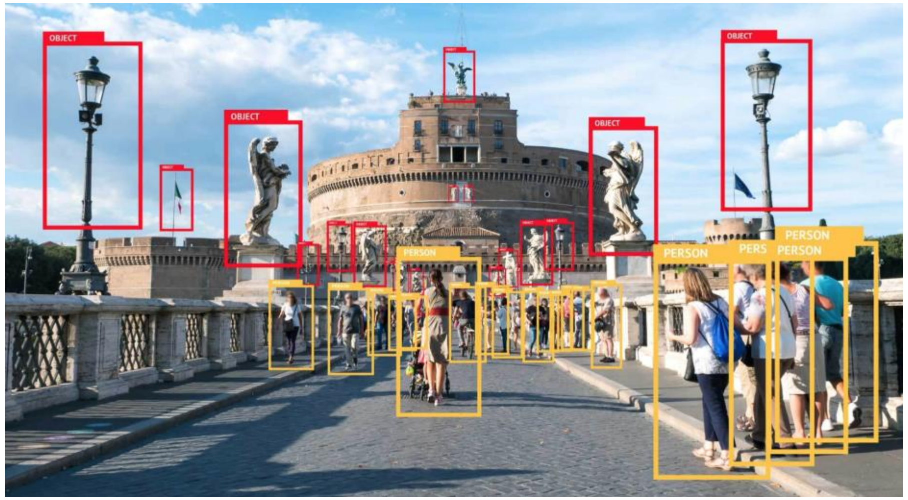
Towards Compositionality in Video Understanding

Roei Herzig

October 26, 2021

Deep Learning

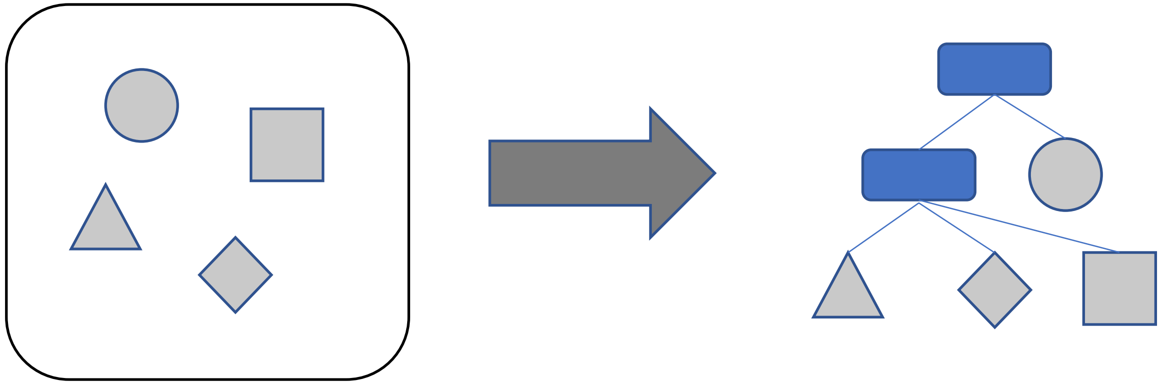
Visual Recognition



Autonomous Driving

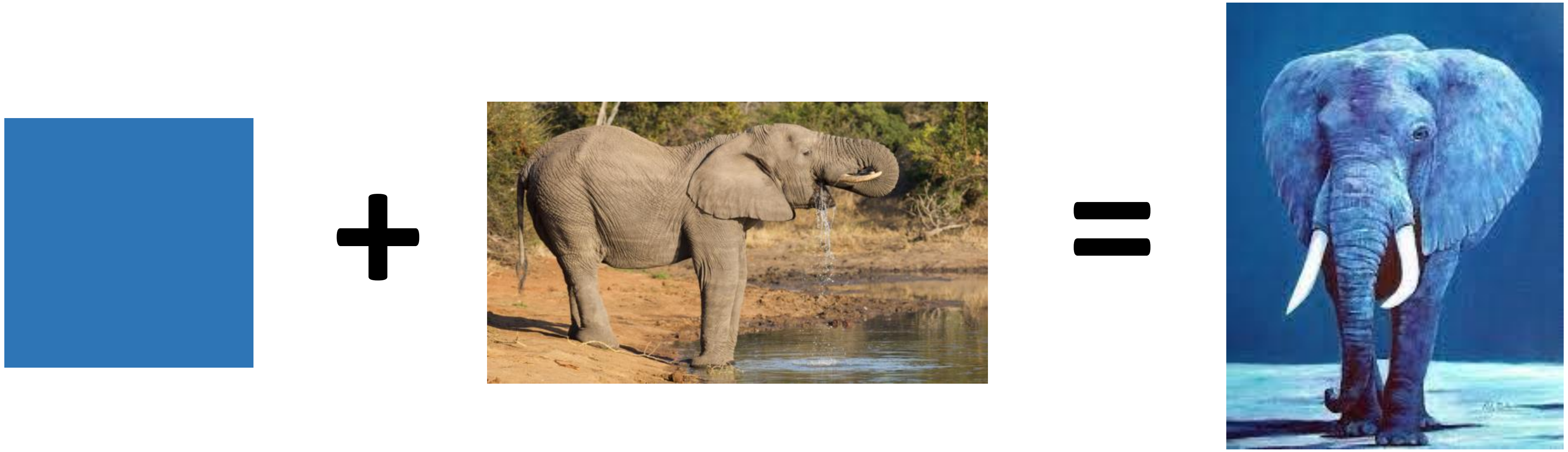
What is missing?

Compositionality



Simple Concepts/Primitives

Complex Concepts



Compositionality

- Many existing vision architectures are not compositional
- Furthermore, we still have open questions:
  - What architectures help models learn compositionality?
  - How do we find the balance between compositional and black-box models?
- We would like to develop **compositional and structured models** that leverage inductive biases into our architectures

Compositionality in Videos

- Actions are performed by objects and create long-range spatio-temporal dependencies
- Composing the actions differently would lead to a different outcome

Compositionality in Videos

Instructions:

Add Eggs

Season with salt and pepper

Whisk the eggs mixture

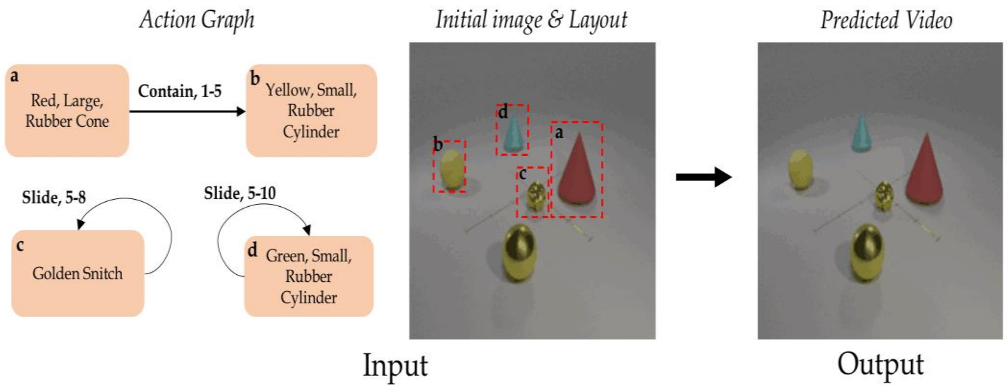
Pour mixture on pan and cook

Add cheese

Towards Compositionality in Video Understanding

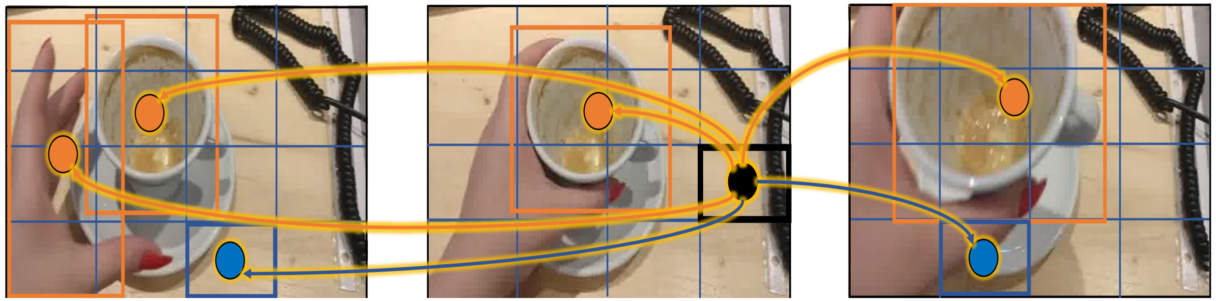
Action Graphs

ICML 2021



Object-Region Video Transformers

Arxiv 2021



Compositional Video Synthesis with Action Graphs

ICML 2021

Amir Bar*, Roei Herzig*,
Xiaolong Wang, Anna Rohrbach, Gal Chechik, Trevor Darrell, Amir Globerson

Our Goal

Synthesize videos of actions

Our Goal

Learn to synthesize videos of actions

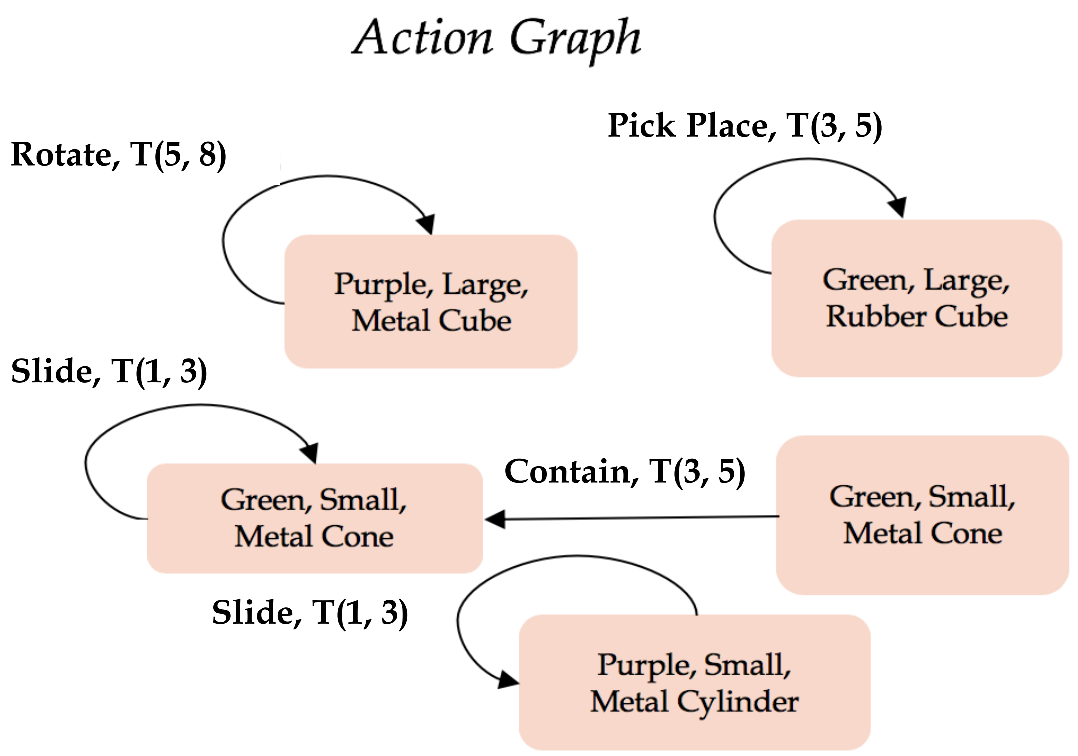
Our model should be able to synthesize:

- Multiple actions and objects
- Potentially simultanious actions
- Coordinated and timed actions

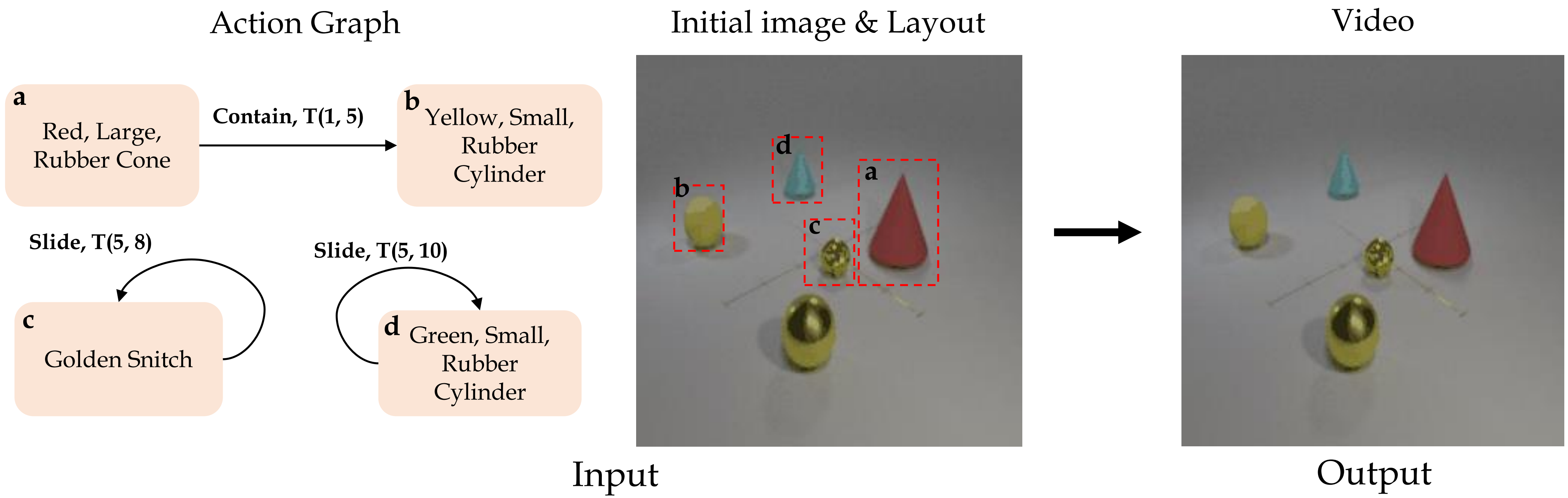
How should we model actions?

The Action Graph Representation

- Nodes are objects
- Edges are timed actions
- Each action is annotated with a a start and end time



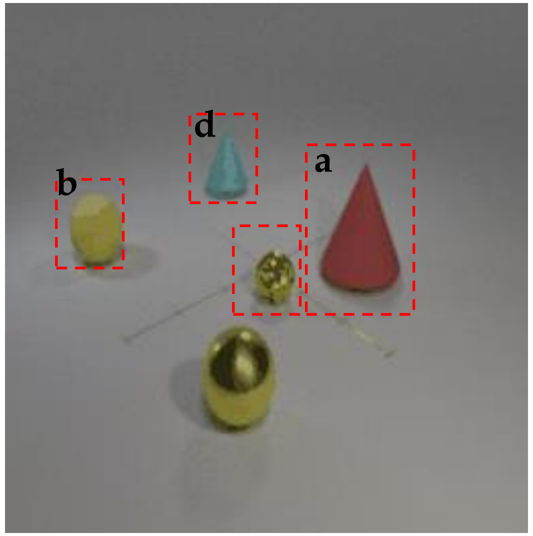
Task Setting: Action-Graph-to-Video



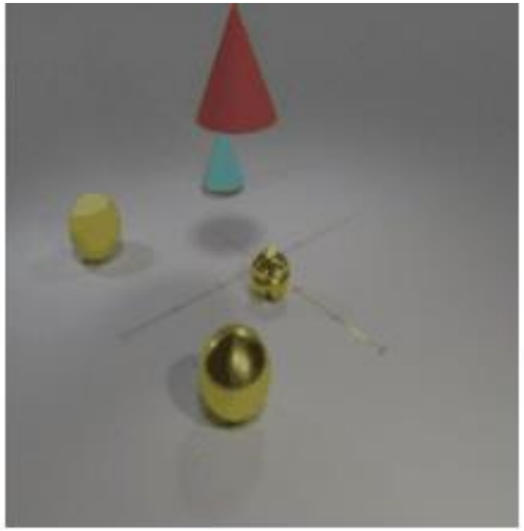
The Action Graph to Video Model

Synthesize next frame in a coarse-to-fine manner

- Action execution schedule, given Action Graph
- Given the schedule, predict how should object moves
- Then, predict how should pixels move



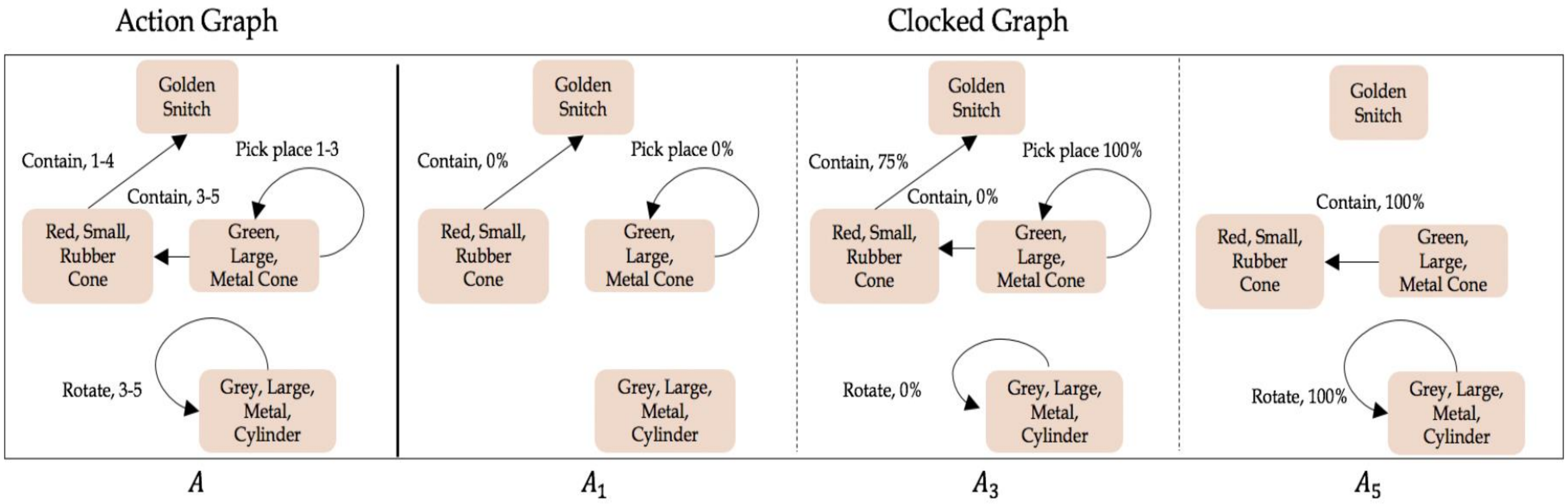
Previous image and layout



Next frame

Scheduling Actions via “Clocked Edges”

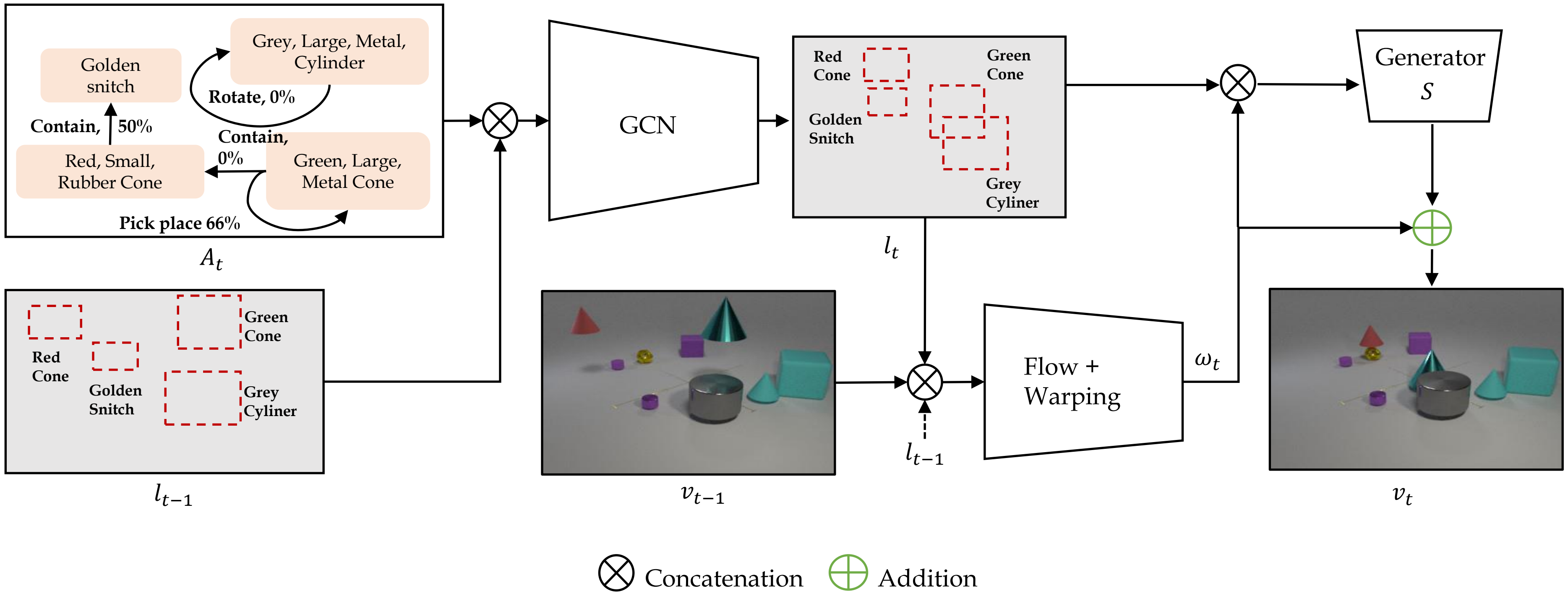
How to synchronize and schedule multiple actions?



Time Specific Action Graphs

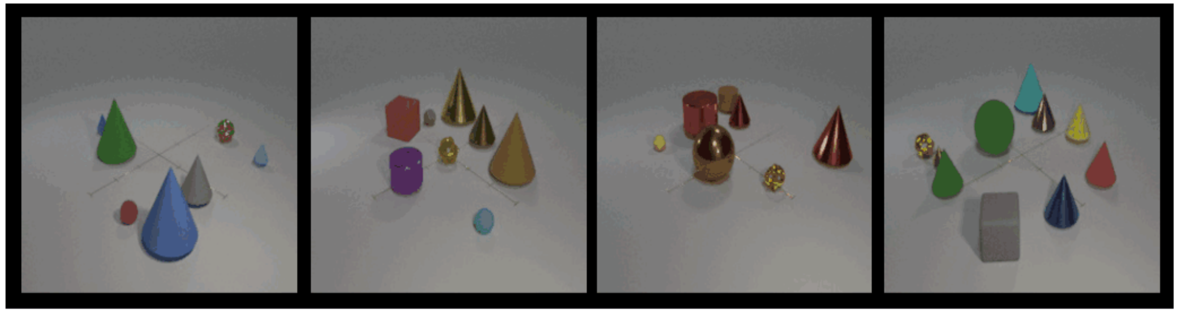
Action Graph to Video

- Predict new scene layout given previous layout and Clocked Action Graph
- Predict the future pixels flow, and warp the previous image
- Refine the warped image via a SPADE Generator

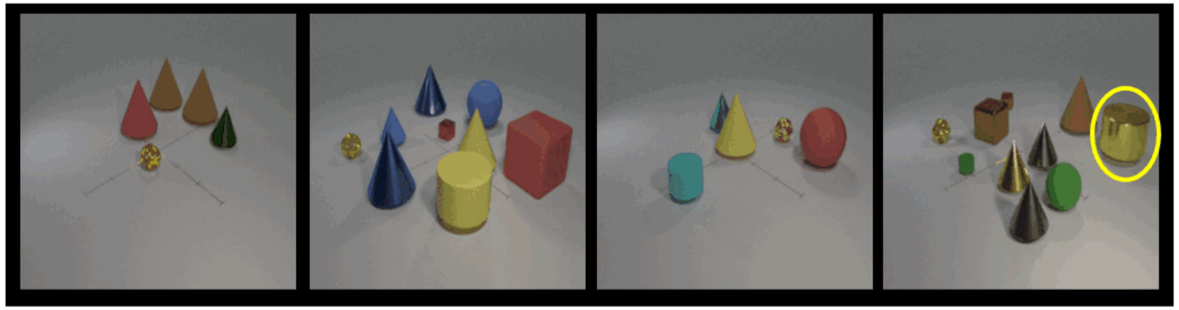


Results

Actions in CATER



Multiple Simultaneous Actions



Slide *Contain* *Pick Place* *Rotate*

Actions in Something Something

Push Left *Move Down* *Uncover* *Push*

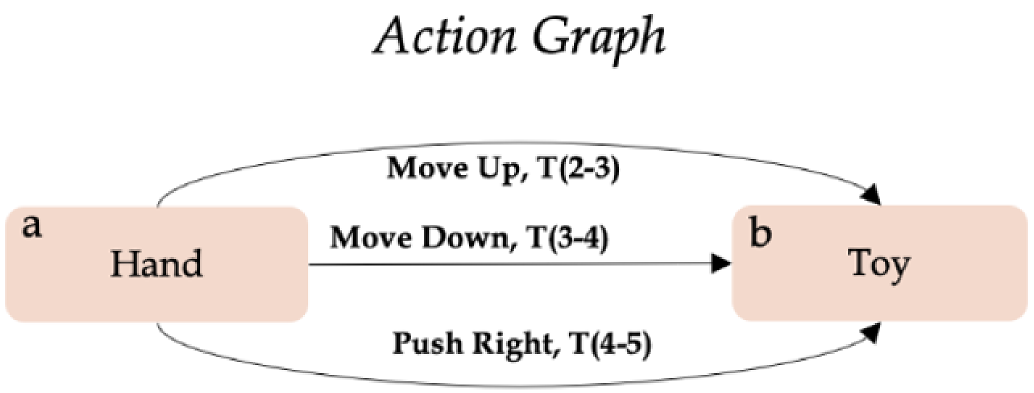
Push Right *Move Up* *Cover* *Take*

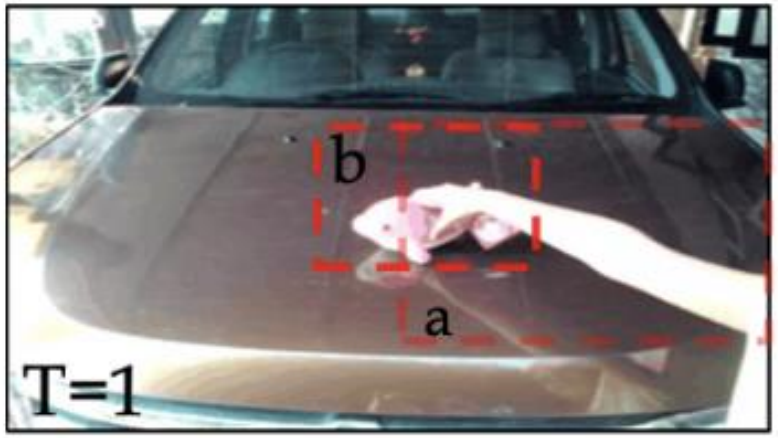
Zero-shot synthesis

So far, we've showed that our model can synthesize the actions present in the training data.

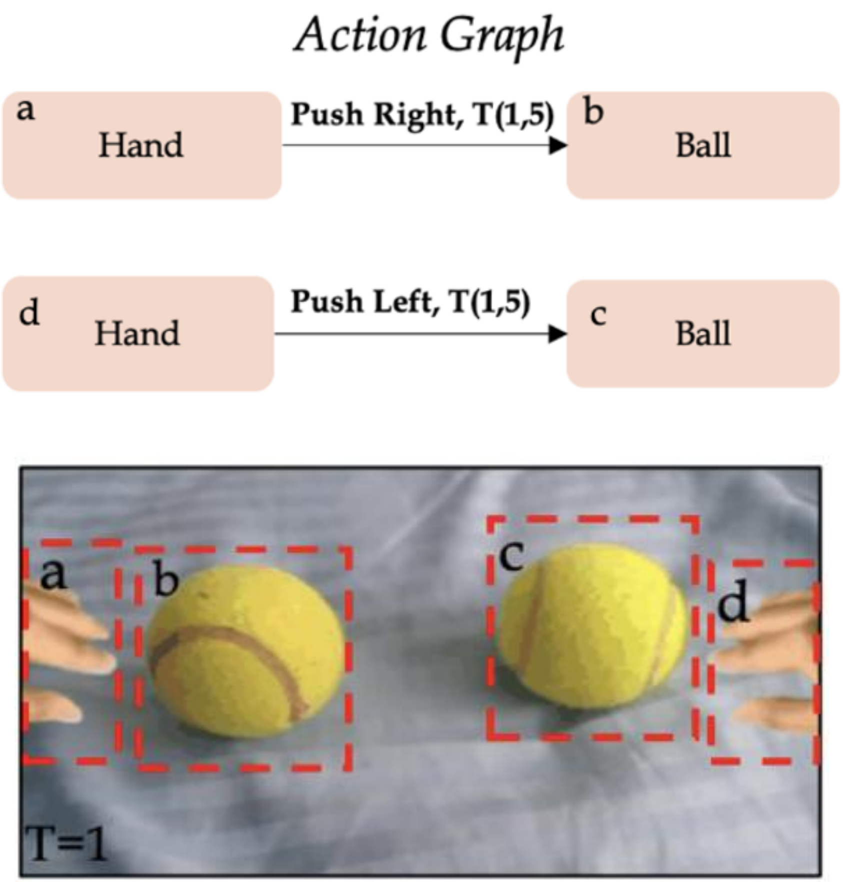
Can we use this approach to synthesize more complex videos?

Synthesizing zero-shot *sequential* actions

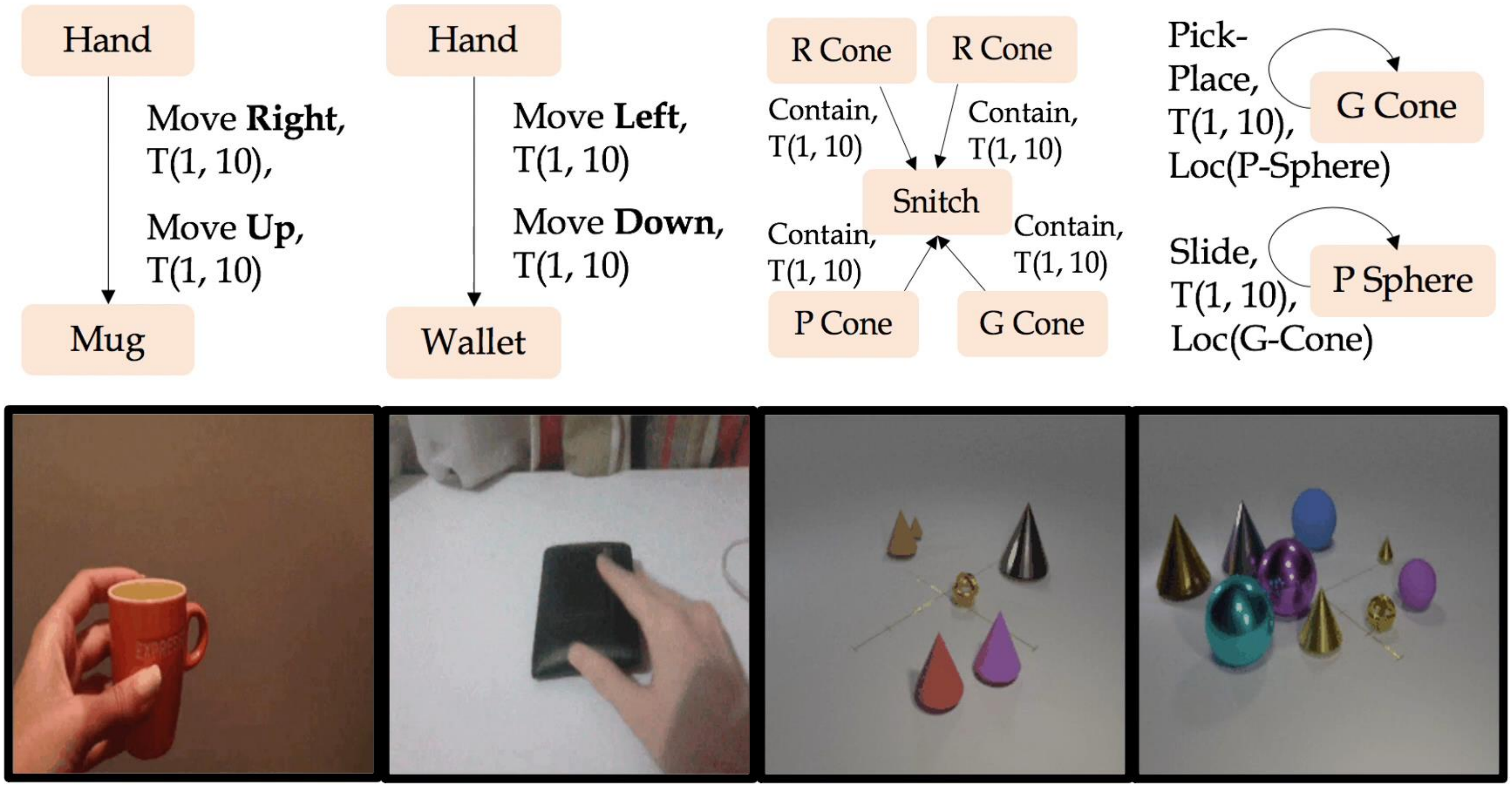




Synthesizing zero-shot *simultaneous* actions



Synthesizing new action composites



Right Up *Left Down* *Huddle* *Swap*

Object-Region Video Transformers

Arxiv, 2021

Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam,
Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, Amir Globerson

Motivation

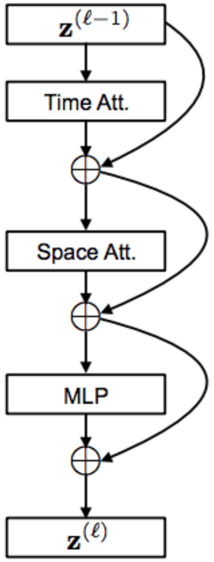
"Picking up a coffee cup"

How can humans recognize actions in videos?

- An action is roughly composed by:
  - What the objects are
  - How do they interact
  - How do they move

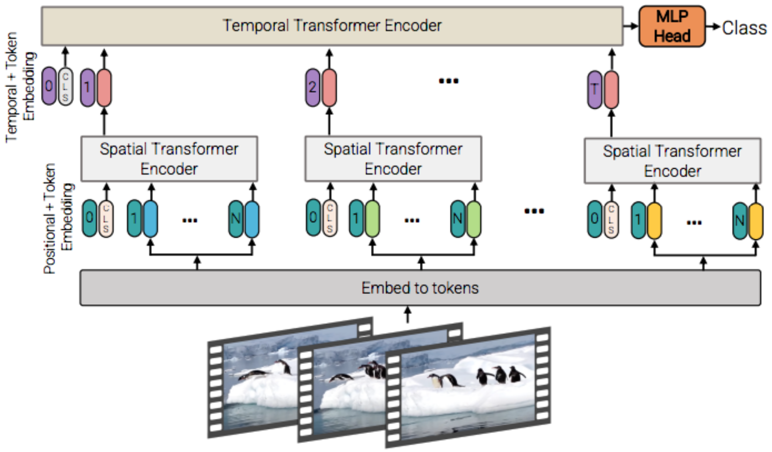
Video Transformer Models

TimeSformer [1]

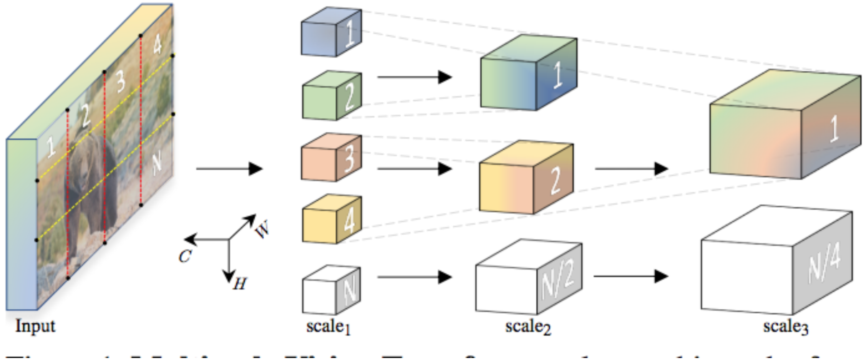


Divided Space-Time Attention (T+S)

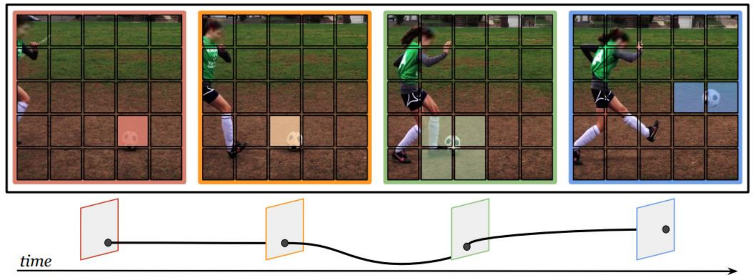
ViViT [2]



MViT [3]



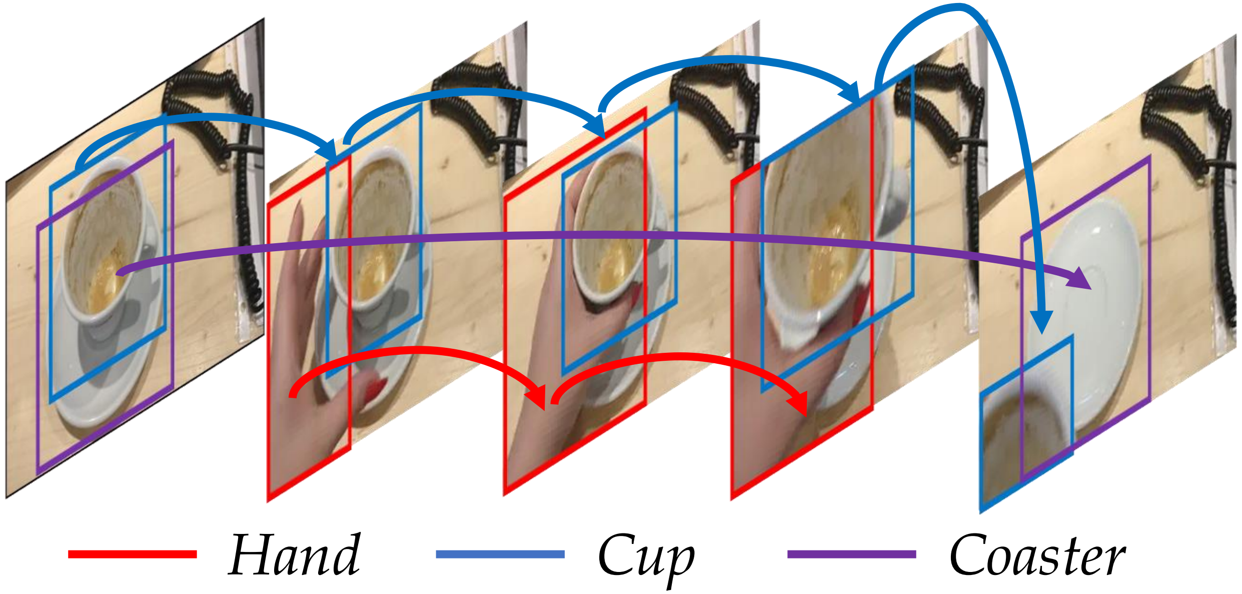
MotionFormer [4]



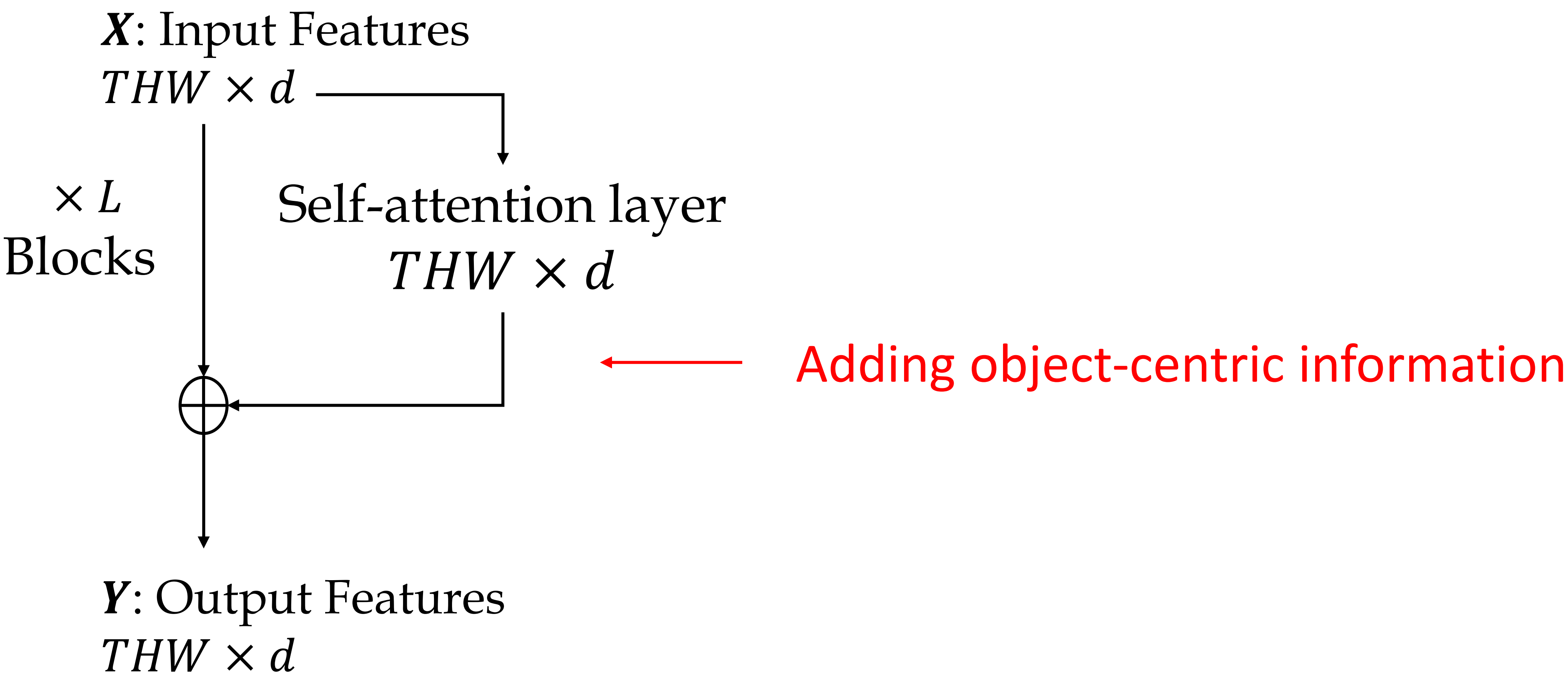
[1] Is Space-Time Attention All You Need for Video Understanding?, ICML21
[2] ViViT: A Video Vision Transformer, ICCV21
[3] Multiscale Vision Transformers, ICCV21
[4] Keeping Your Eye on the Ball: Trajectory Attention in Video Transformers, NeurIPS21

Object-Centric Approach

- Objects are key to understanding actions
- Our question: How can this be captured by Video Transformer Models?



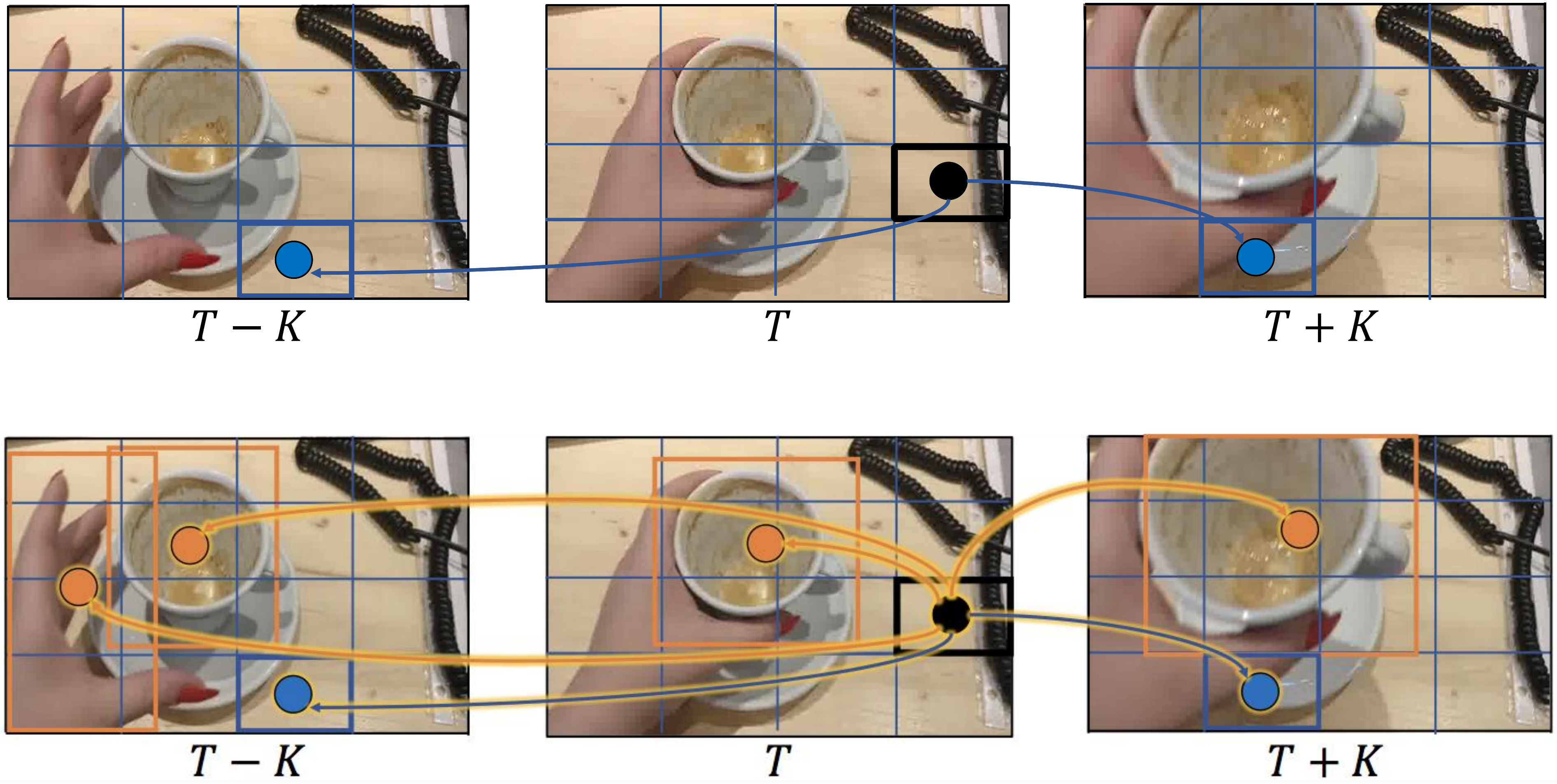
Object-Centric Approach



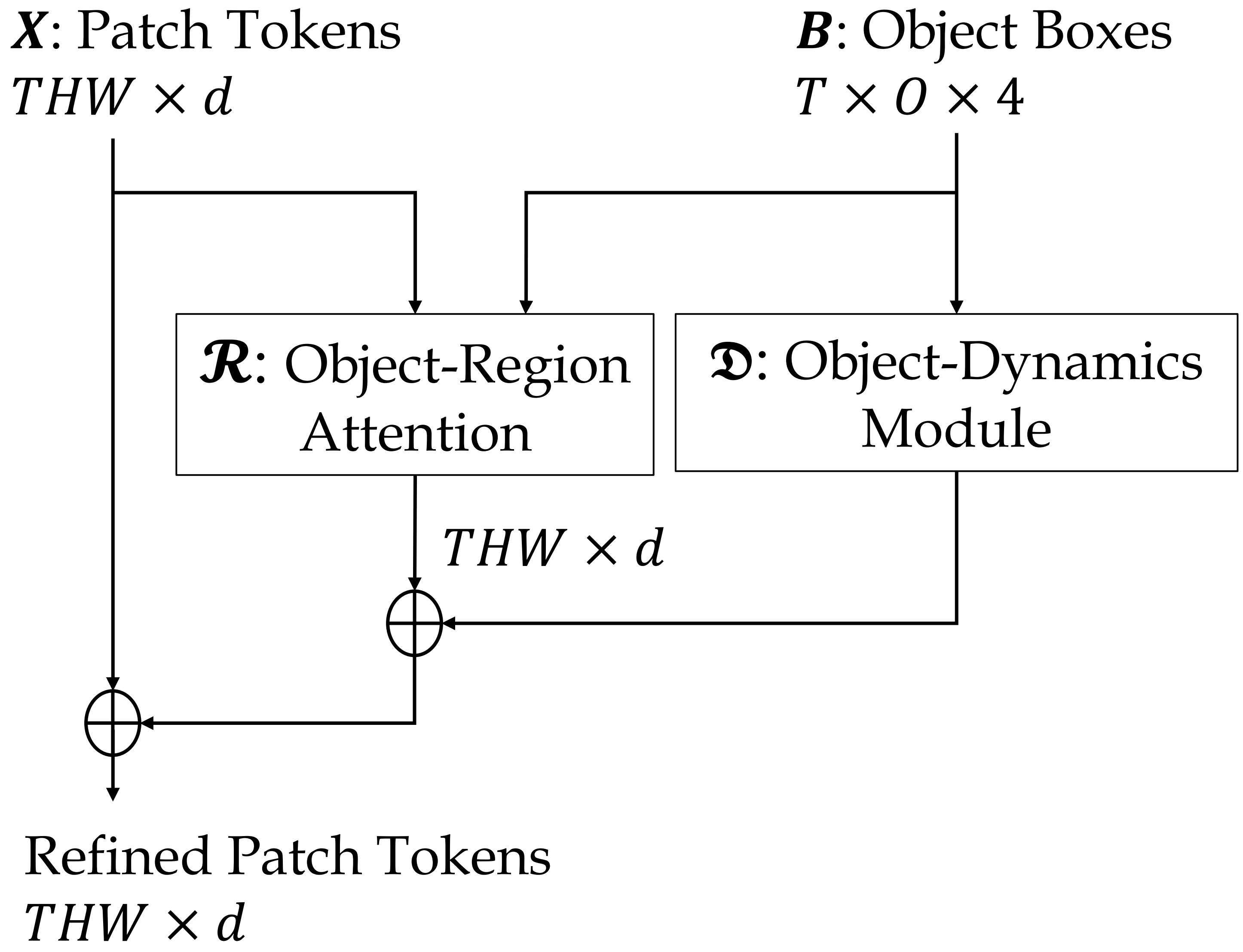
Objects as Transformer Tokens

Standard Transformer

Our Model

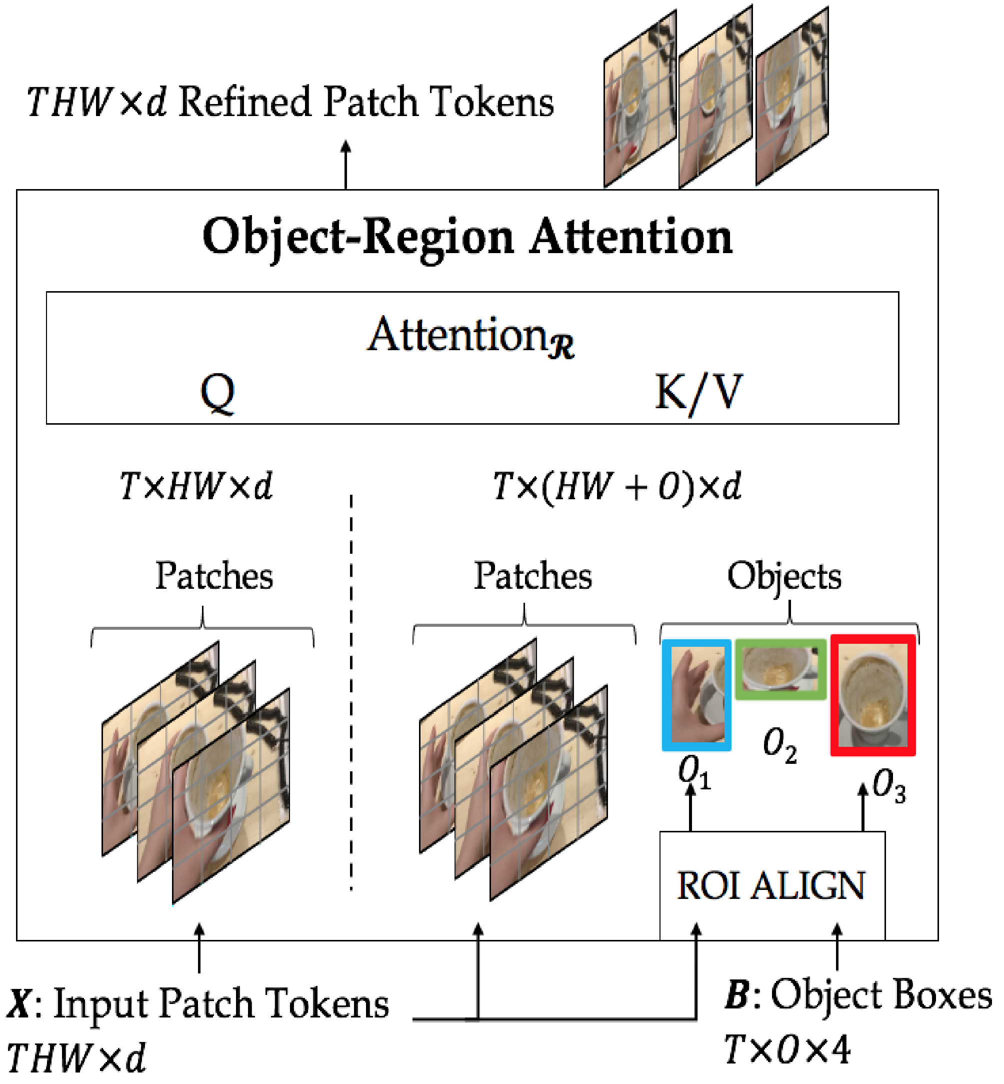


ORViT Block

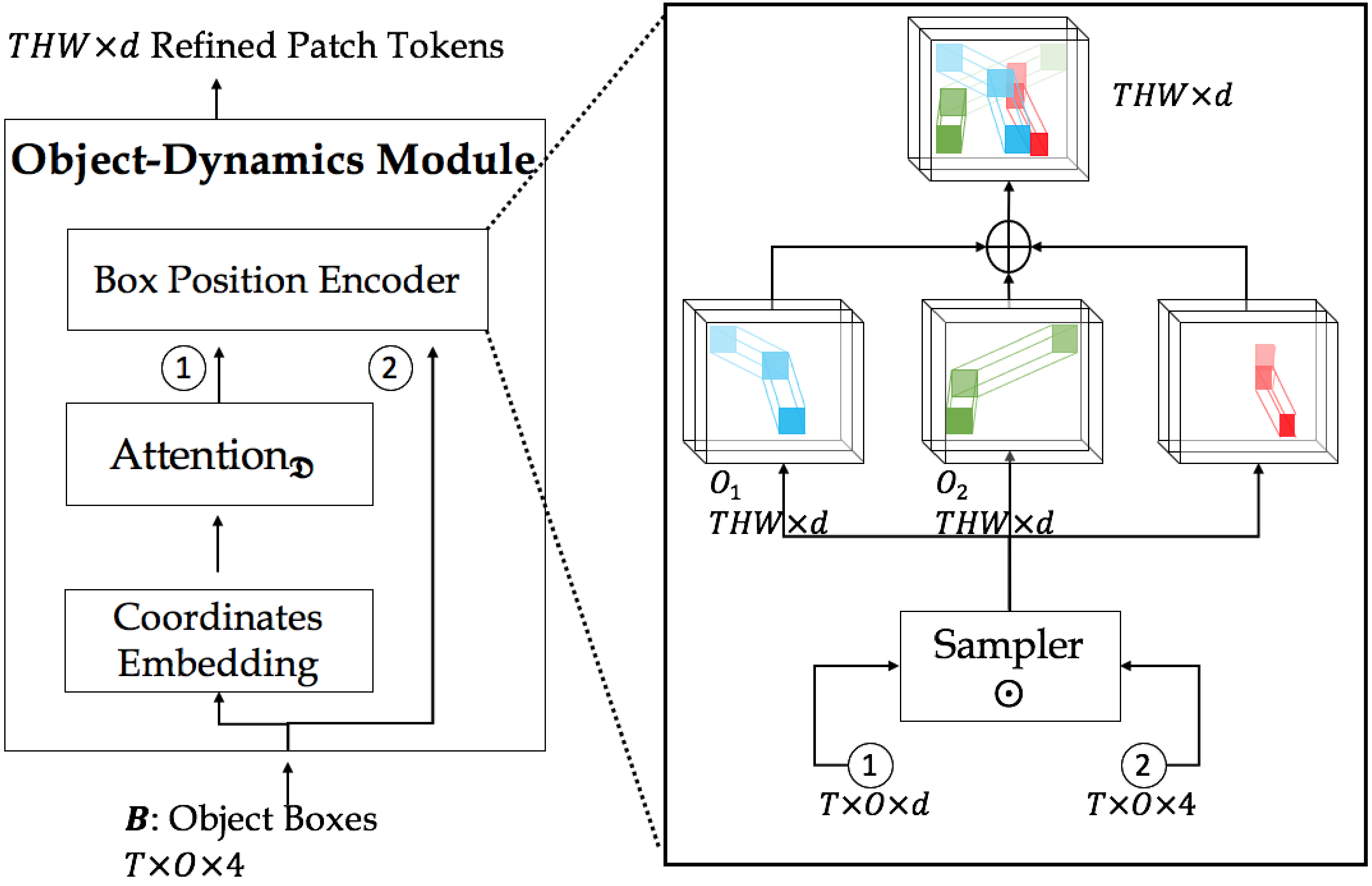

$\boldsymbol{X}$: Patch Tokens
$THW \times d$
$\boldsymbol{B}$: Object Boxes
$T \times O \times 4$
$\mathcal{R}$: Object-Region Attention
$\mathfrak{D}$: Object-Dynamics Module
$THW \times d$
Refined Patch Tokens
$THW \times d$

Object-Region Attention

- Assume access to object boxes in time
- Use these as additional “spatial regions” in the transformer self-attention
- Boxes are also used to extract trajectory information in a separate stream, and re-integrated with the self-attention output



Object Dynamics



Results

Compositional and Few-Shot Action Recognition on SomethingElse

Model	Modality		Compositional		Base		Few-Shot	
	RGB	Boxes	Top-1	Top-5	Top-1	Top-5	5-Shot	10-Shot
I3D (Carreira & Zisserman, 2017)	✓	✗	42.8	71.3	73.6	92.2	21.8	26.7
SF (Feichtenhofer et al., 2019)	✓	✗	45.2	73.4	76.1	93.4	22.4	29.2
TimeSformer (Bertasius et al., 2021)	✓	✗	44.2	76.8	79.5	95.6	24.6	33.8
Mformer (Patrick et al., 2021)	✓	✗	60.2	85.8	82.8	96.2	28.9	33.8
STRG (\w SF) (Wang & Gupta, 2018)	✓	✓	52.3	78.3	75.4	92.7	24.8	29.9
STIN (\w SF) (Materzynska et al., 2020)	✓	✓	54.6	79.4	77.4	95.0	23.0	33.4
Mformer+STRG+STIN	✓	✓	62.3	86.0	83.7	96.8	29.8	36.5
ORViT Mformer (ours)	✓	✓	**69.7**	**91.0**	**87.1**	**97.6**	**33.3**	**40.2**

+15 improvement compared to other graph-based methods
+9.2 improvement compared to the Mformer model

Results – Standard Action Recognition

(a) **Something–Something V2**

Model	Boxes	Pretrain	Top-1	Top-5	GFLOPs×views (10^9)	Param (10^6)
SlowFast, R101	✗	K400	63.1	87.6	106×3×1	53.3
TimeSformer-L	✗	IN	62.5	-	1703×3×1	121.4
ViViT-L	✗	IN+K400	65.4	89.8	3992×4×3	-
MViT-B, 64	✗	K600	68.7	91.5	236×3×1	53.2
Mformer	✗	IN+K400	66.5	90.1	369.5 × 3 × 1	109
Mformer-L	✗	IN+K400	68.1	91.2	1185.1 × 3 × 1	109
Mformer + STRG + STIN	GT	IN+K400	69.2	90.9	375 × 3 × 1	119
ORViT Mformer (Ours)	Detected	IN+K400	**67.9** (+1.4)	90.5 (+0.4)	405 × 3 × 1	148
ORViT Mformer (Ours)	GT	IN+K400	**73.8** (+7.3)	**93.6** (+3.5)	405 × 3 × 1	148
ORViT Mformer-L (Ours)	Detected	IN+K400	**69.5** (+1.4)	**91.5** (+0.3)	1259 × 3 × 1	148.2
ORViT Mformer-L (Ours)	GT	IN+K400	**74.9** (+6.7)	**94.2** (+3.0)	1259 × 3 × 1	148.2

(b) **Diving48**

Model	Pretrain	Frames	Top-1
SlowFast, R101	K400	16	77.6
TimeSformer	IN	16	74.9
TimeSformer-L	IN	96	81.0
TQN	K400	ALL	81.8
TimeSformer	IN	32	80.0
TimeSformer + STRG + STIN	IN	32	83.5
ORViT TimeSformer (Ours)	IN	32	**88.0** (+8.0)

(c) **Epic-Kitchens100**

Method	Pretrain	A	V	N
SlowFast, R50	K400	38.5	65.6	50.0
ViViT-L	IN+K400	44.0	66.4	56.8
Mformer	IN+K400	43.1	66.7	56.5
Mformer-L	IN+K400	44.1	67.1	57.6
Mformer-HR	IN+K400	44.5	67.0	58.5
MF-HR + STRG + STIN	IN+K400	44.1	66.9	57.8
ORViT Mformer-HR (Ours)	IN+K400	**45.7** (+1.2)	**68.4** (+1.4)	**58.7** (+.2)

Results

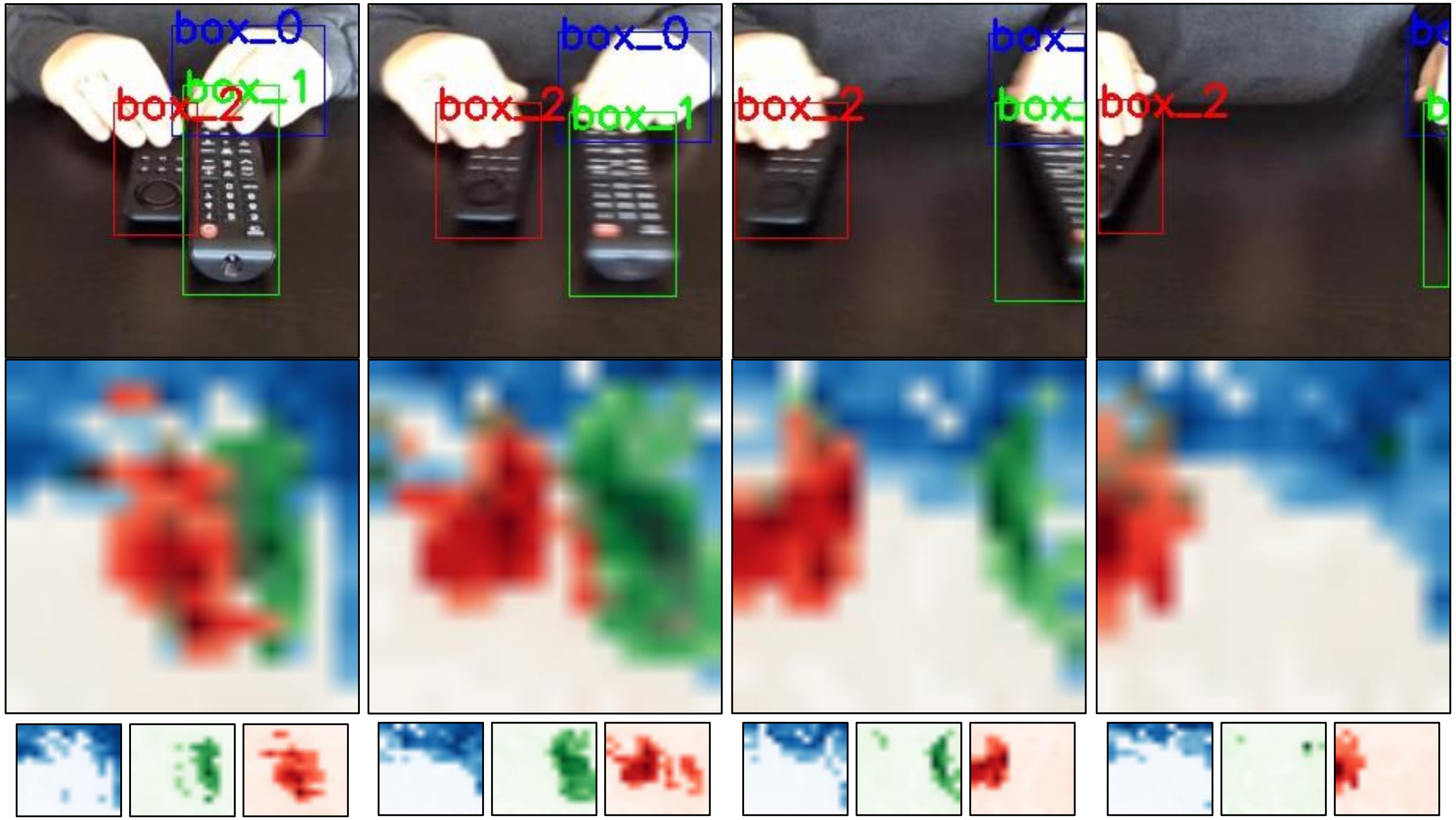
Spatio-Temporal Action detection on AVA

Model	Pretrain	mAP	Param
SlowFast, 4×16, R50	K400	21.9	33.7
SlowFast, 8×8, R101	K400	23.8	53.0
MViT-B, 16×4	K400	25.5	36.4
ORViT MViT-B (Ours)	K400	**26.6**	49.8

+1.1 improvement compared to the MViT-B model

Visualizations

"Moving something and something away from each other"



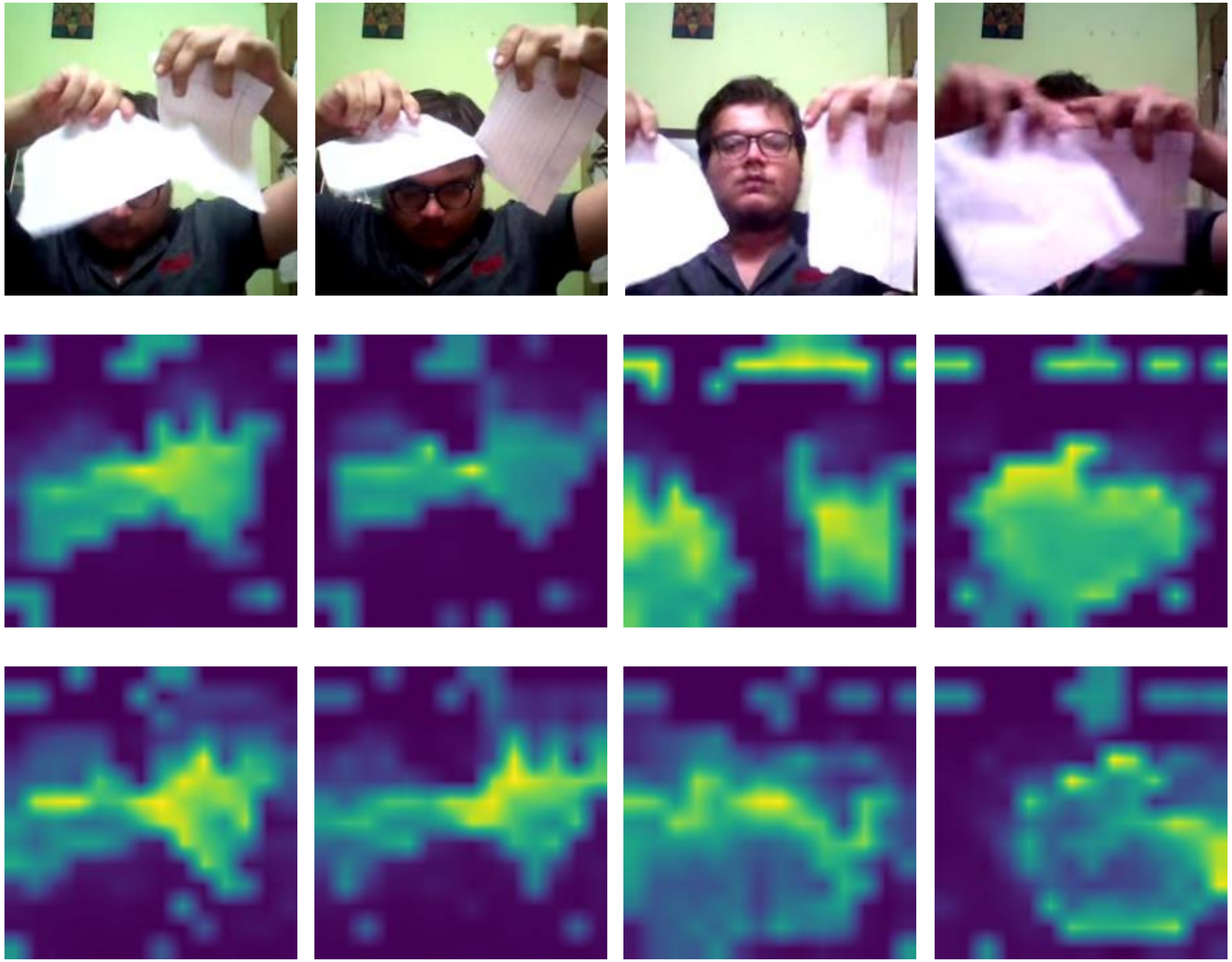
Box 1 Box 2 Box 3

Visualizations

"Tearing something into two pieces"

ORViT-Mformer

Mformer

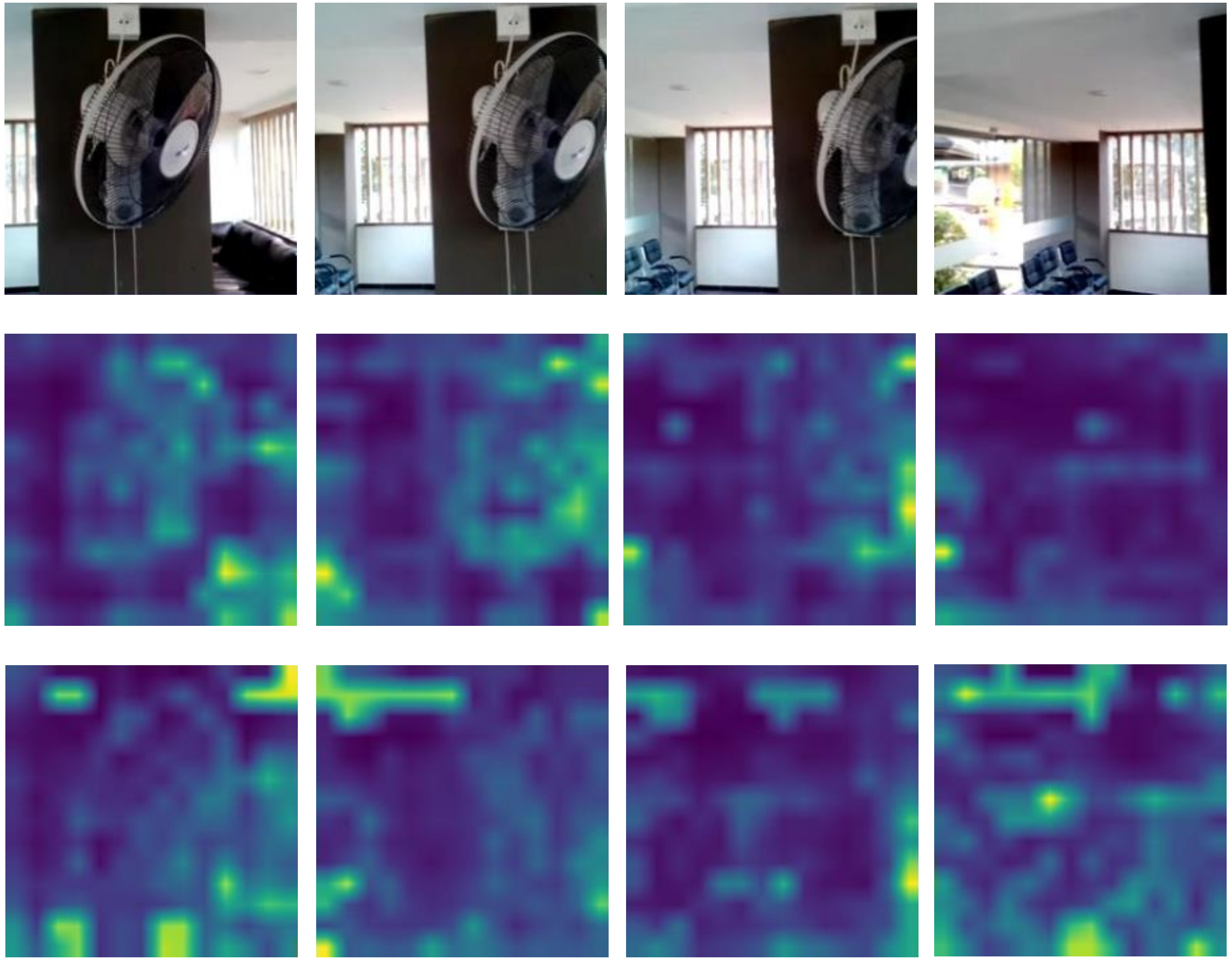


Visualizations

“Turning the camera left while filming something”

ORViT-Mformer

Mformer



Thank you!

Roei Herzig
TAU

Amir Globerson
TAU

Trevor Darrell
Berkeley

Webpage: https://roeiherz.github.io/