Compositional Video Synthesis with Action Graphs

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

ICML 2021

* Equally contributed.

SRVU, ICCV 2021

Our Goal

Synthesize videos of actions

Our Goal

Learn to synthesize videos of actions

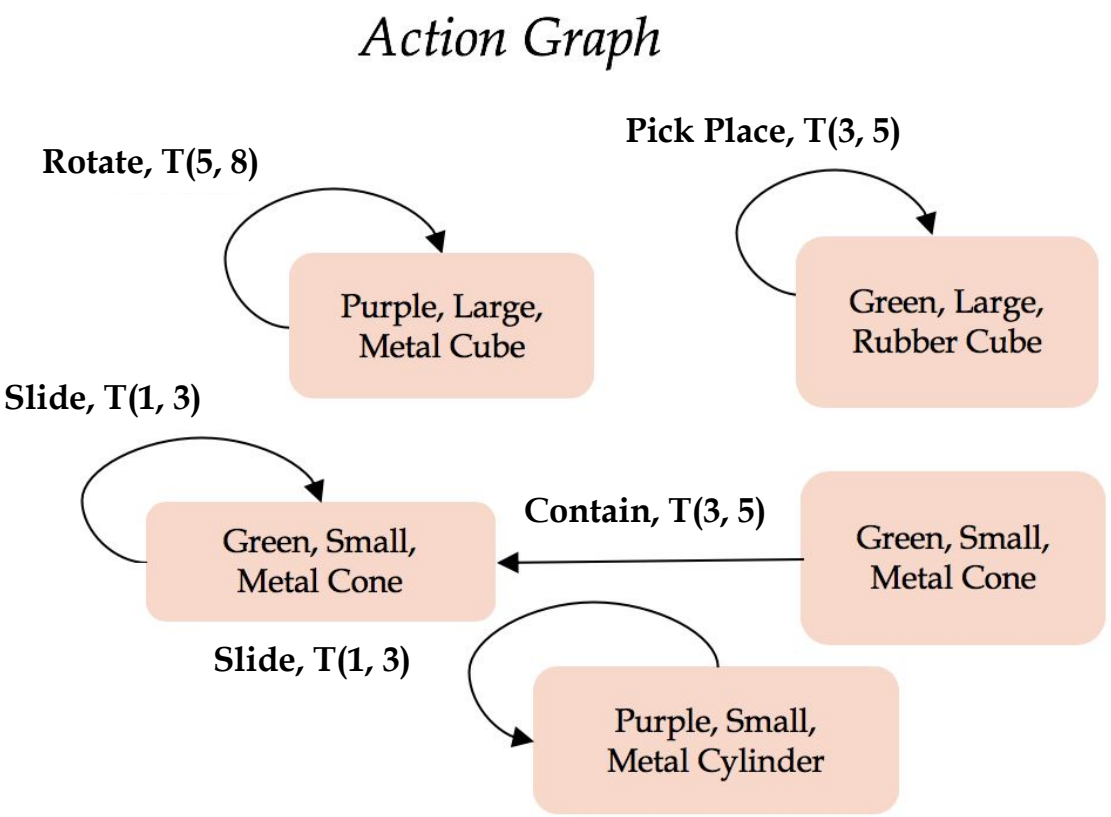
Our model should be able to synthesize:

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

How should we model actions?

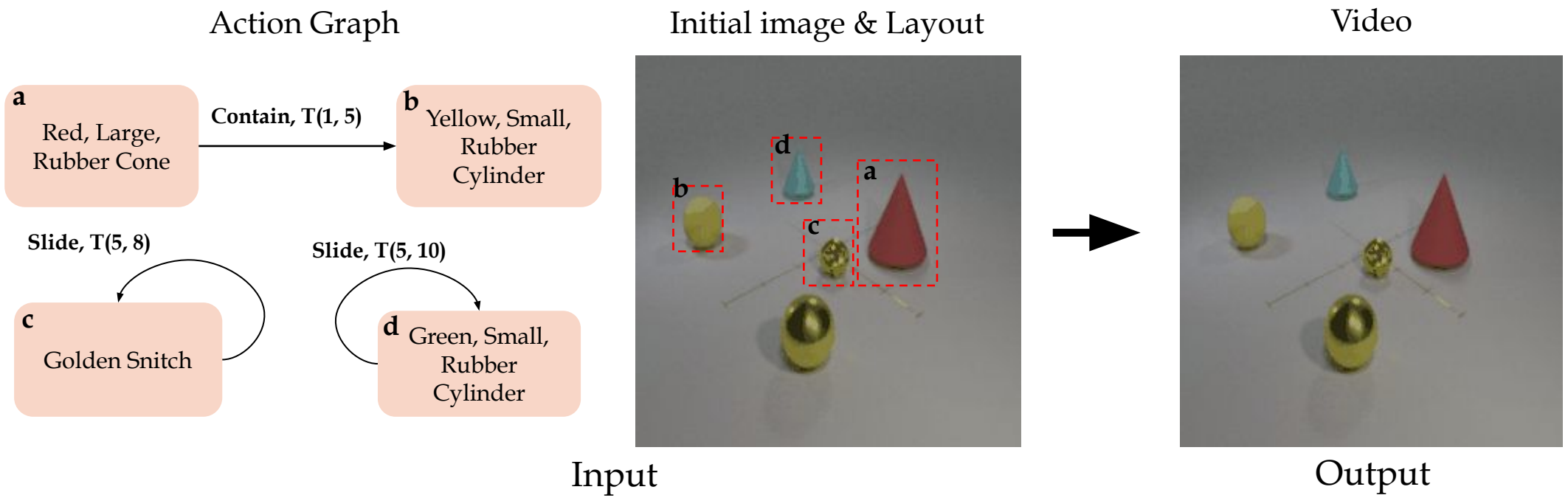
The Action Graph Representation

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



Video

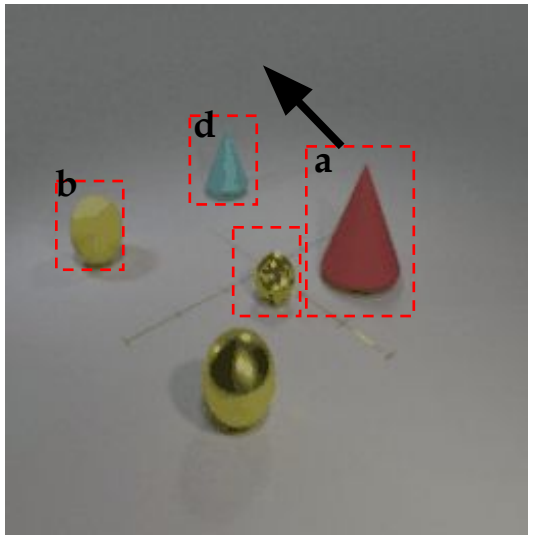
New Task Setting: Action-Graph-to-Video



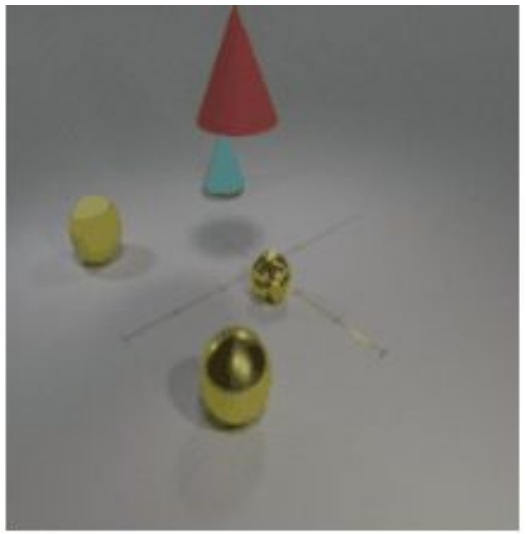
The Action Graph to Video Model

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

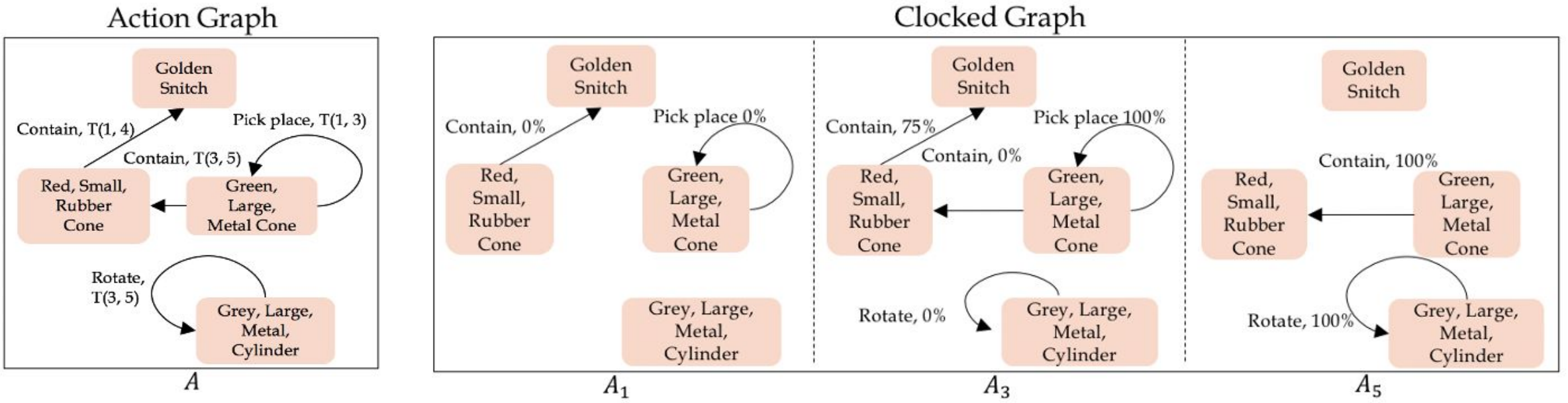


Previous image and layout


Next frame

Scheduling Actions via “Clocked Edges”

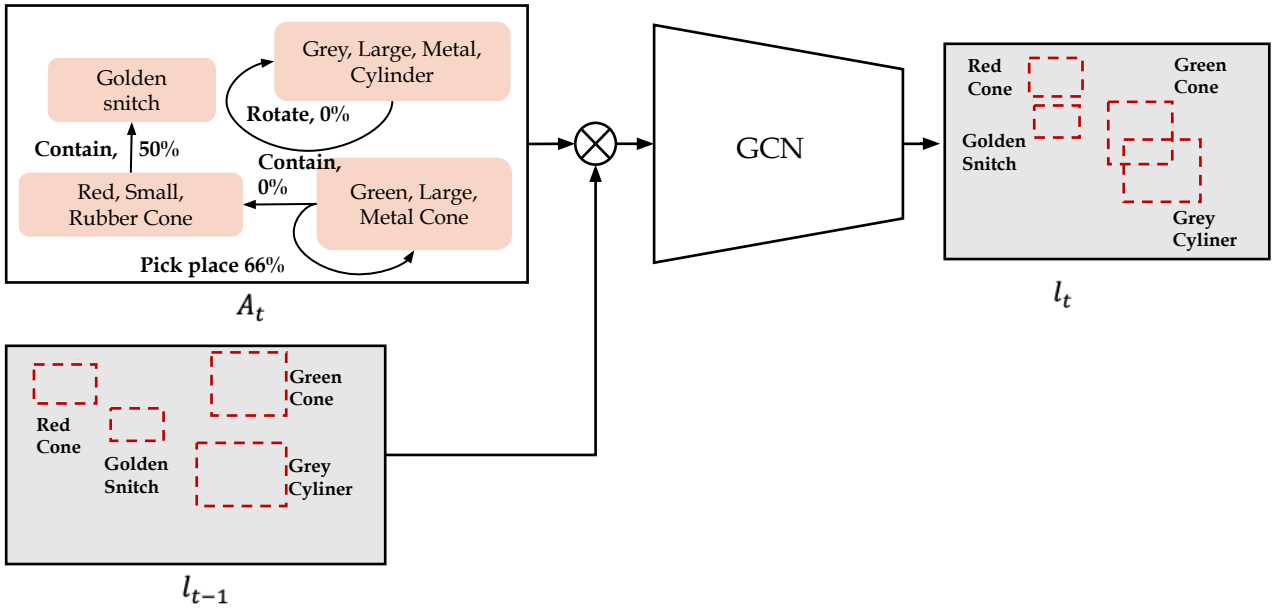
How to synchronize and schedule multiple actions?



Time specific Action Graphs

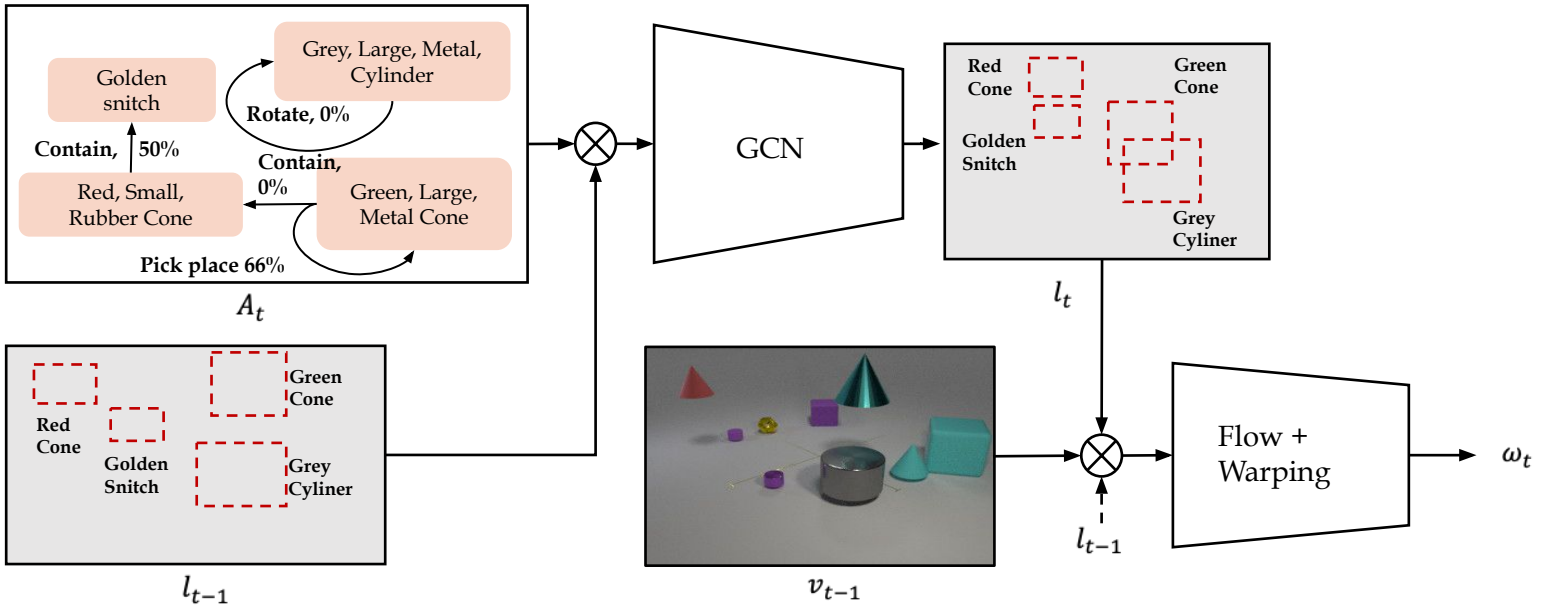
Action Graph to Video

- Predict new scene layout given previous layout and Clocked Action Graph



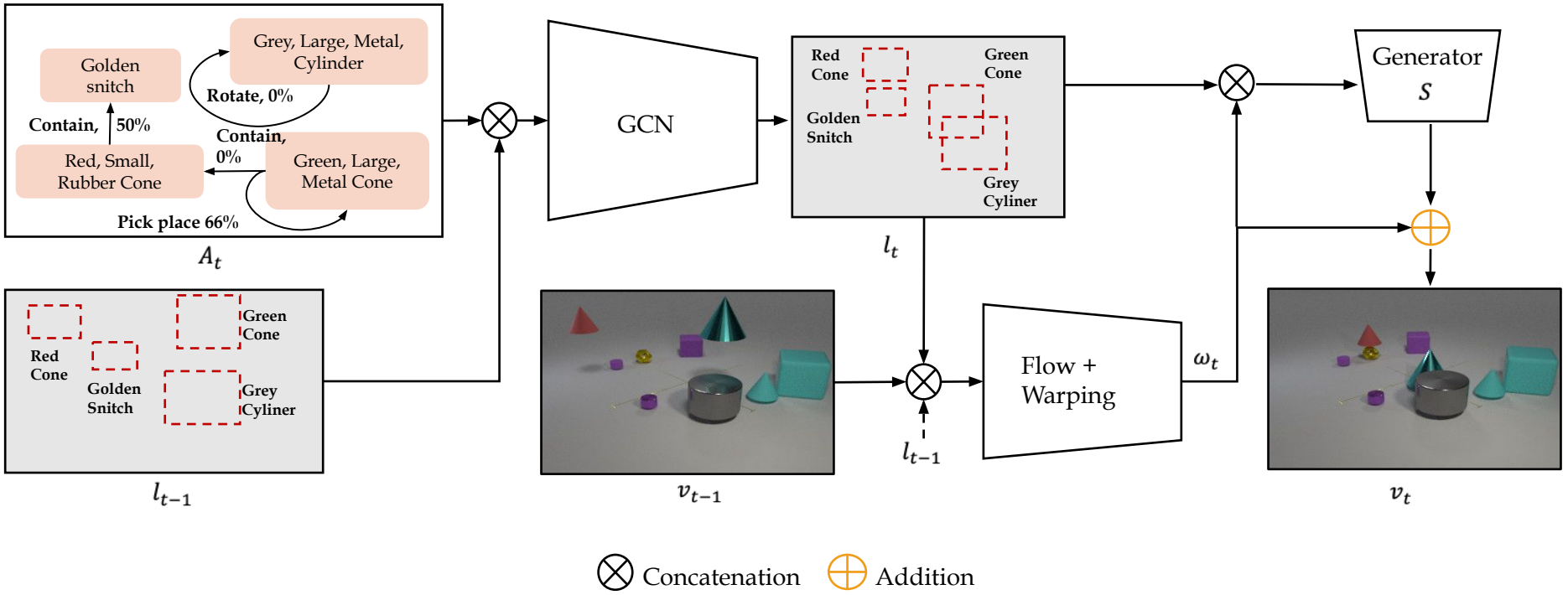
Action Graph to Video

- Predict new scene layout given previous layout and Clocked Action Graph
- Predict the future pixels flow, and warp the previous image



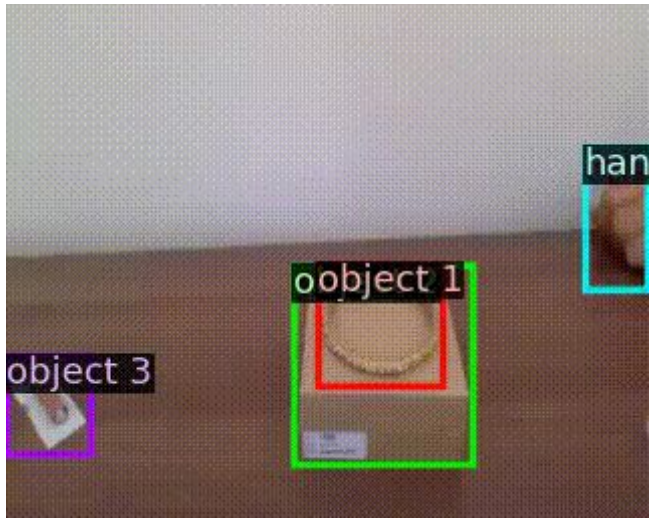
Action Graph to Video

- 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



Datasets of atomic actions

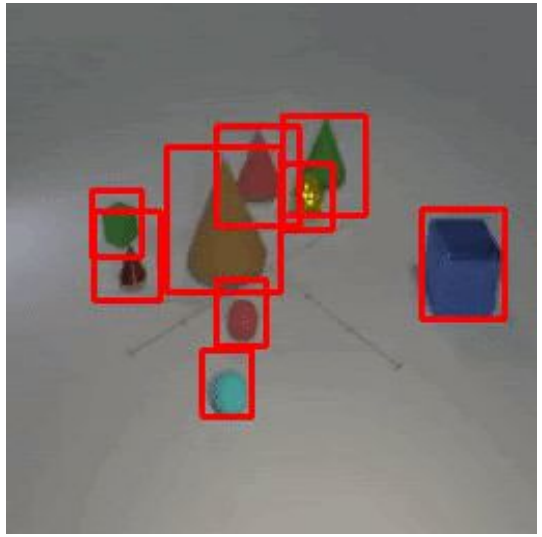
Something Something V2



Move down
Move up
Push right
Push left
Take
Cover
Uncover
Put

Videos of *single* actions

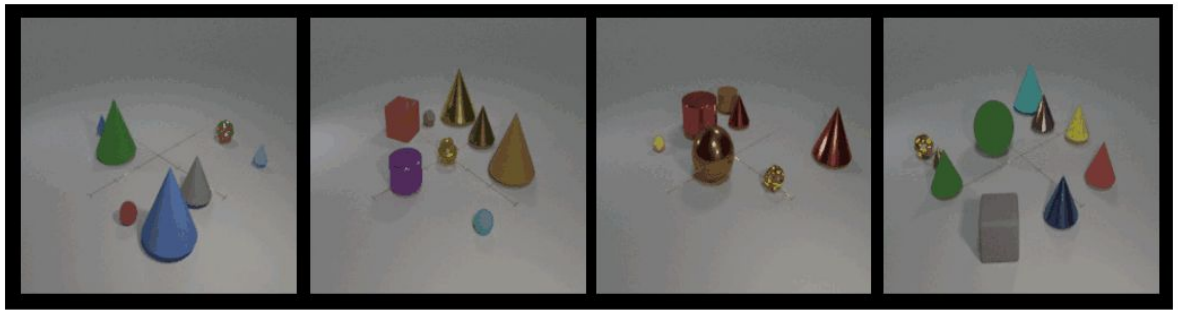
CATER



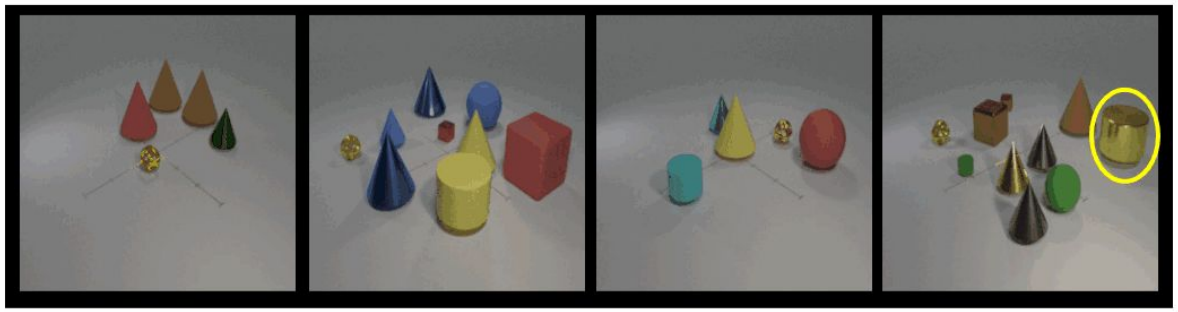
Slide
Contain
Rotate
Pick Place

Videos of *multiple* actions

Actions in CATER

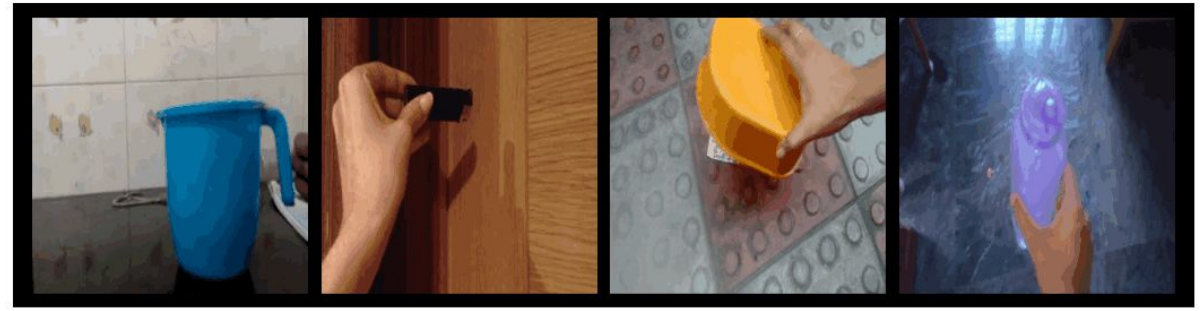


Multiple Simultaneous Actions



Slide ***Contain*** *Pick Place* ***Rotate***

Actions in Something Something



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

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

Human evaluation of the synthesized videos

AG2Vid vs. Baseline	Semantic Accuracy		Visual Quality	
	CATER	SmthV2	CATER	SmthV2
CVP (Ye et al., 2019)	**85.7**	**90.6**	**76.2**	**93.8**
HG (Nawhal et al., 2020a)	–	**84.6**	–	**88.5**
V2V (Wang et al., 2018a)	**68.8**	**84.4**	**68.8**	**96.9**
RNN	**56.0**	**80.6**	**52.0**	**77.8**
AG2Vid-GTL	48.6	46.2	42.9	50.0

Each cell is the %times AG2Vid model is better

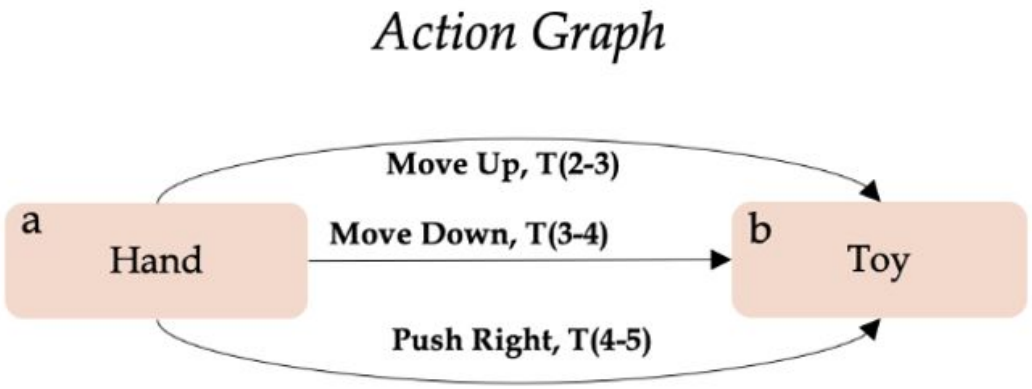
Our model outperforms baselines by synthesizing videos that are **more semantically correct** and have **better visual quality**.

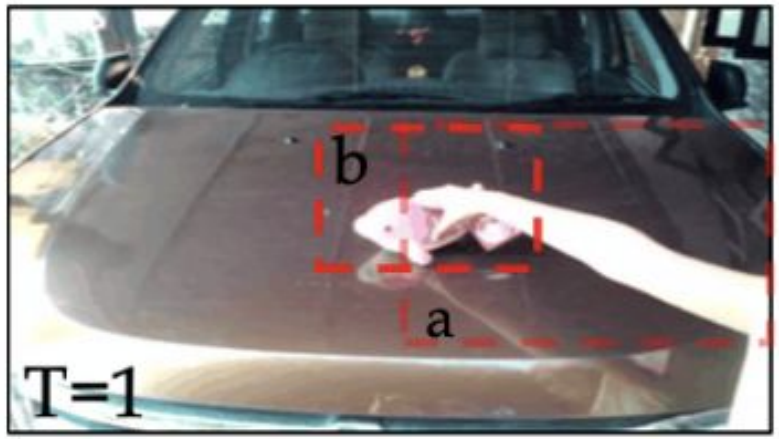
Zero-shot synthesis

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

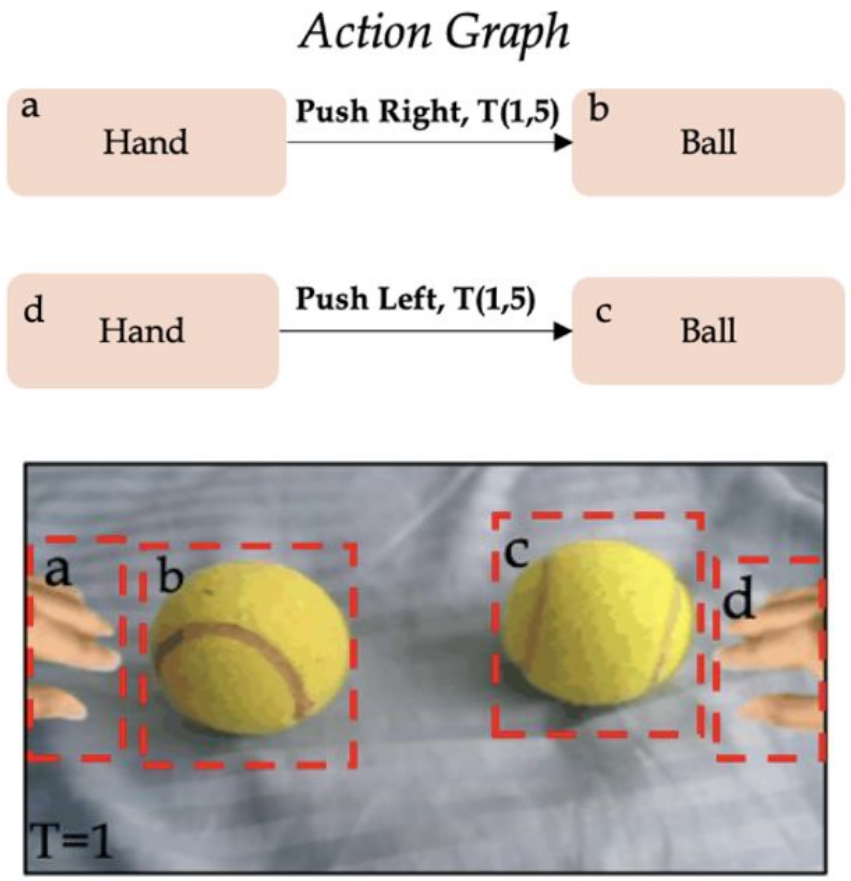
Can we use this approach to synthesize more complex videos?

Synthesizing zero-shot *sequential* actions

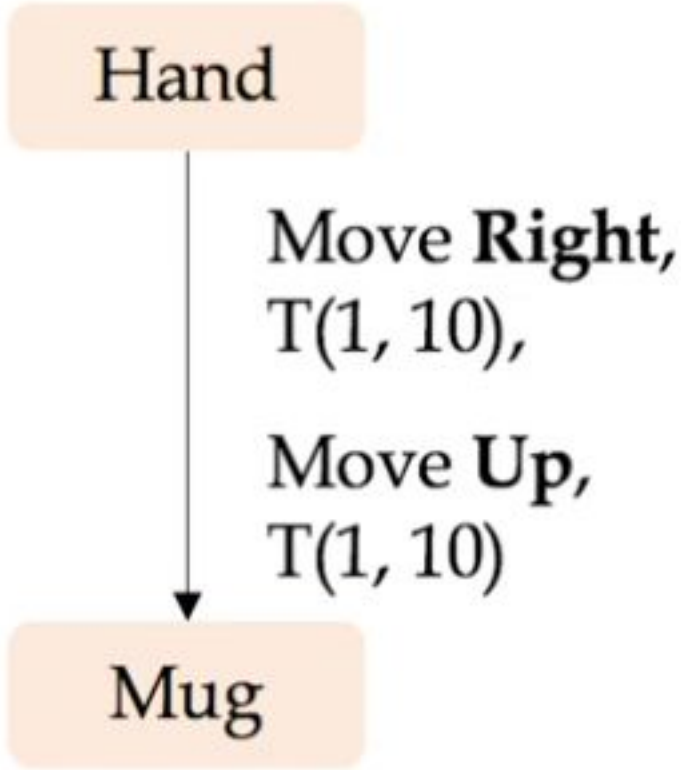




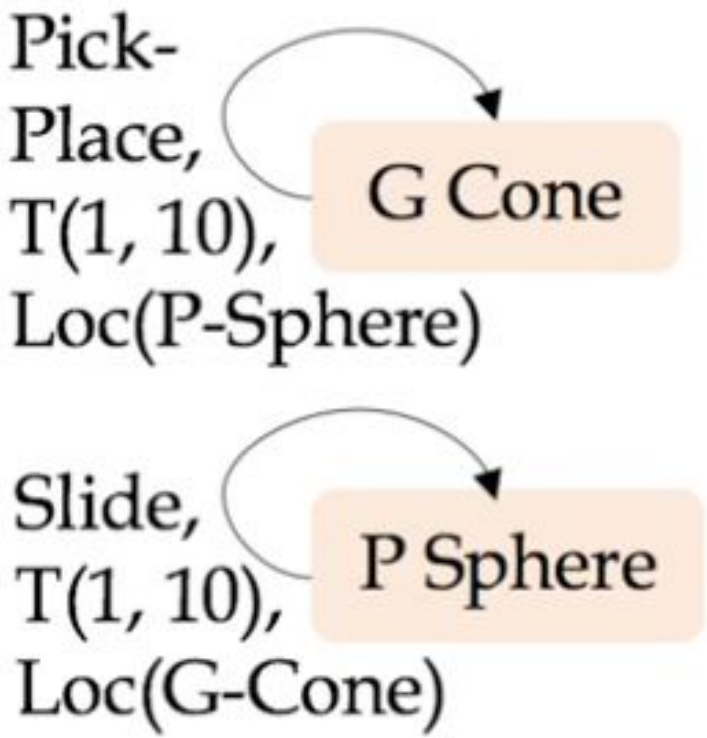
Synthesizing zero-shot *simultaneous* actions



Synthesizing new *action composites*

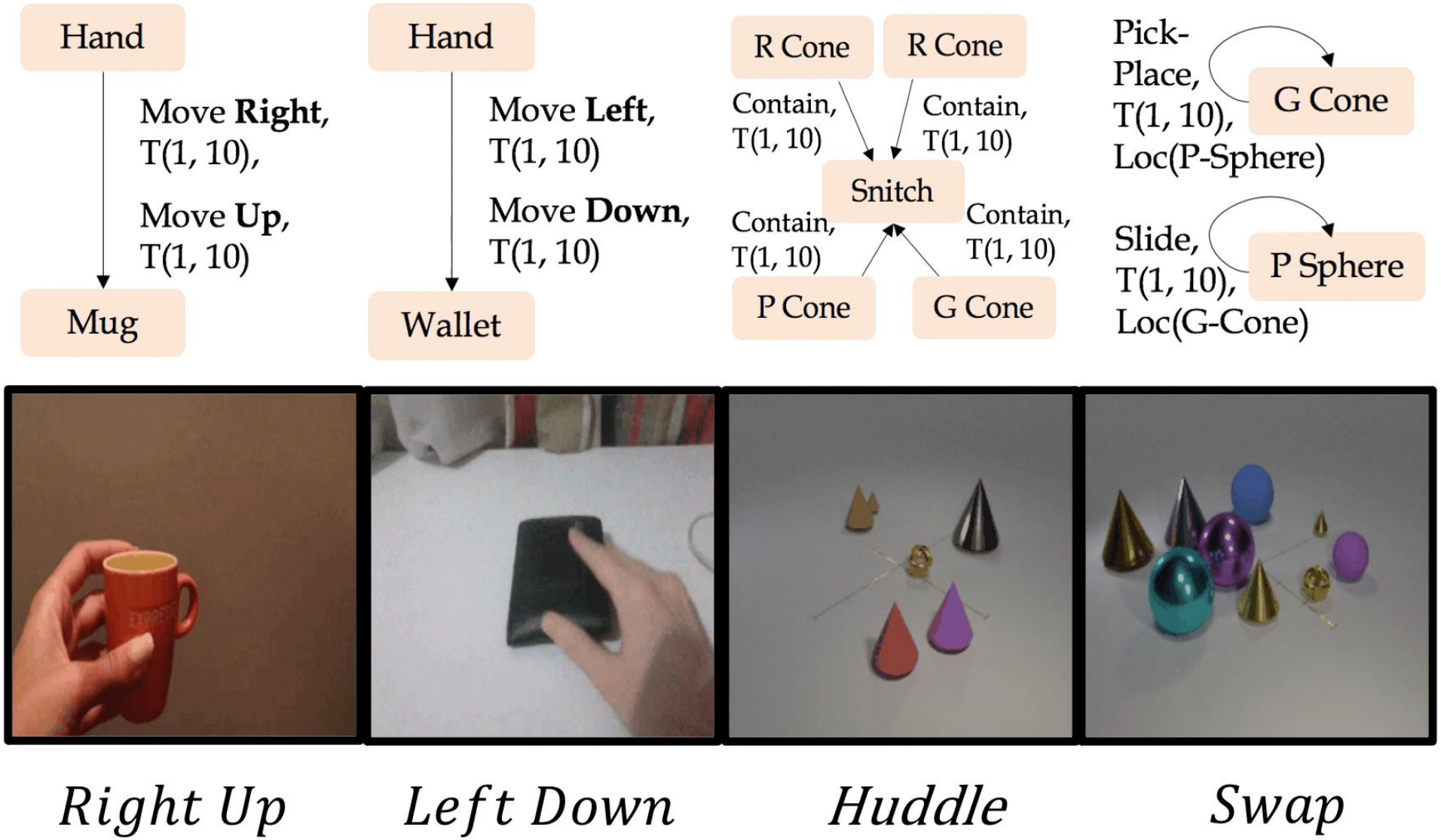


New action: **Move Right-Up**



New action: **Swap Place**

Synthesizing new action composites

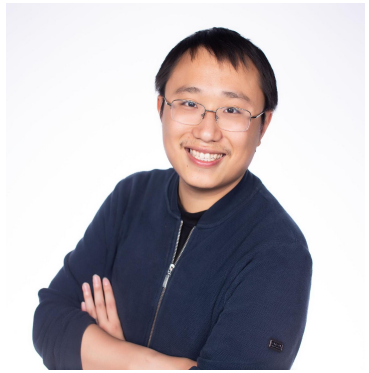


TEL AVIV UNIVERSITY אוניברסיטת תל אביב

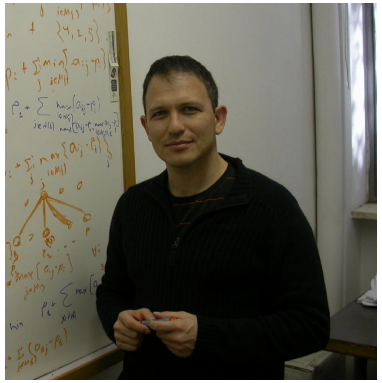
NVIDIA

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH









See project page for code and models:
roeiherz.github.io/AG2Video

Thank you!