Graph-Structured Visual Imitation
Maximilian Sieb*, Zhou Xian*, Audrey Huang, Oliver Kroemer, Katerina Fragkiadaki
Learn policy from visual input so that the robot’s actions create the same effect on the environment as the human.
Focus of this work:

Find interpretable state representation that allows for sample-efficient visual imitation learning from single demonstration.
Visual Entity Graphs for Visual Imitation

1. Detect **object** and **hand entities** in image

2. Detect object-level **pixel entities**

3. Establish **geometric relations** between the entities

4. Place **attention** on “important” edges
Establish correspondence between demonstration and imitation.
From Visual Entity Graphs to Policy Learning

$$
C(G^t_D, G^t_I) = \sum_{i,j,i<j} w(E^t_{(i,j)}) \cdot att(E^t_{(i,j)}) \cdot \|x^t_D - x^t_I - (x^t_i - x^t_j)\|.
$$

Visual Entity Graph Encoding $G$

Demonstration

Imitation

Encode

Cost $(G^t_D, G^t_I)$

Policy Learning

$G^t_D$

$G^t_I$
| Demonstration | Imitation | Imitation different object instance |
|---------------|-----------|------------------------------------|
| **Pushing**   |           |                                    |
| **Pushing – Direction Change** |           |                                    |
| **Stacking**  |           |                                    |
| **Pouring**   |           |                                    |