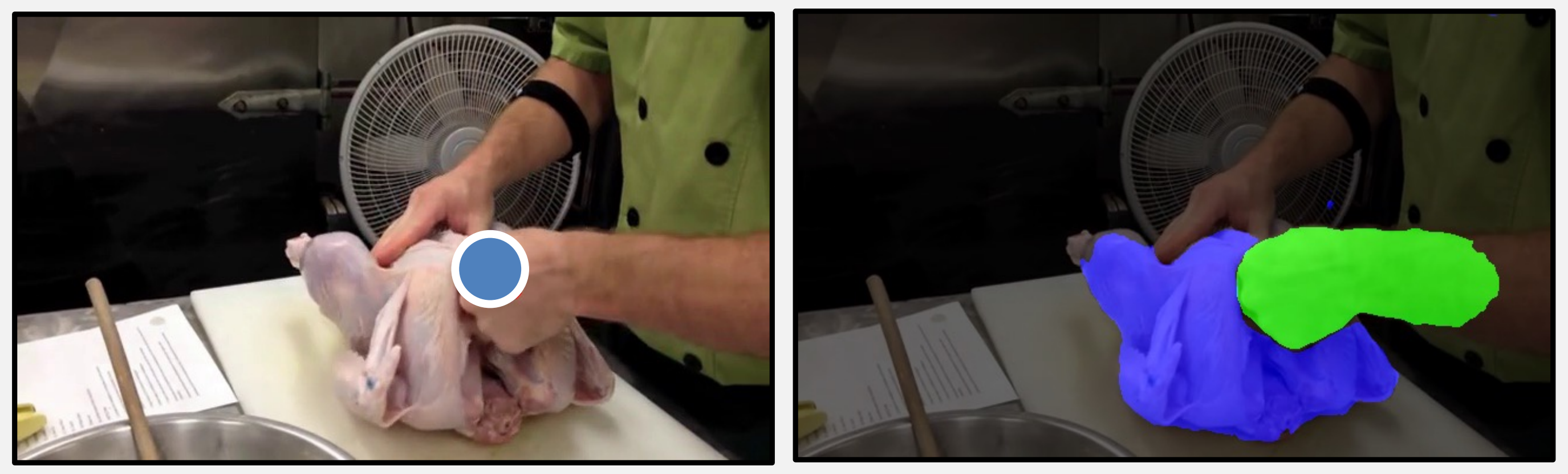
COHESIV: Contrastive Object and Hand Embeddings for Segmentation In Video

Dandan Shan*[1], Richard EL. Higgins*[1], David F. Fouhey[1]
University of Michigan[1]

Overview

Goal: Given a single RGB image and a 2D hand location as input, we aim to segment hands and hand-held objects to better understand contact regions.



For learning, we generate *responsibility maps* in video and use them as pseudo-labels.

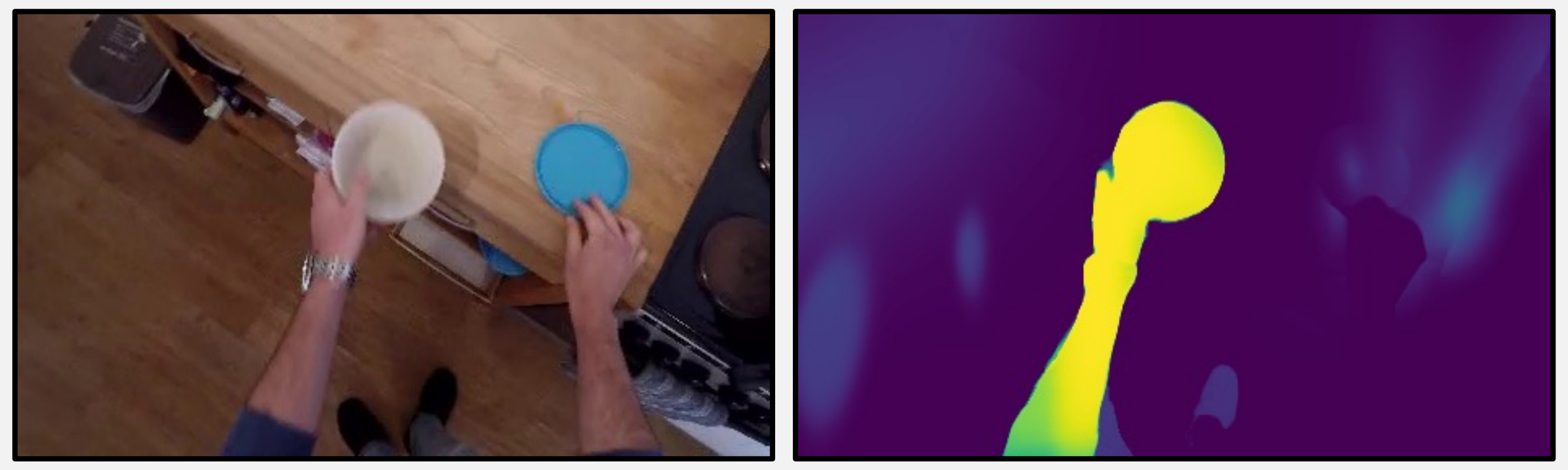
Responsibility Map

We use *responsibility* as the notion of synchronous motion for hand and in-hand object, explaining how well each pixel is explained by each hand's motion model or the background.

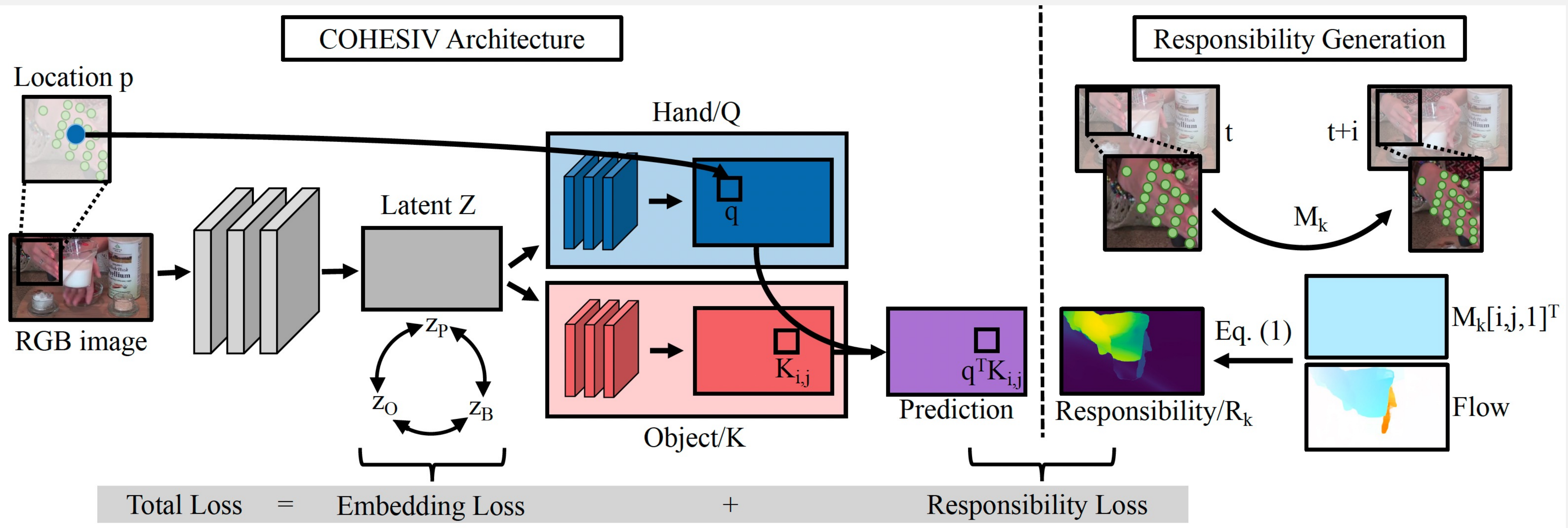
Given a set of N hands, we produce N responsibility maps $R \in \mathbb{R}^{H\times W\times(N+1)}$. For the kth hand,

$$R_{i,j,k} = \frac{exp_t(-d_k(O_{i,j,:}))}{exp_t\left(-d_{BG}(O_{i,j,:})\right) + \sum_{k'=1}^{N} exp_t(-d_k(O_{i,j,:}))}$$

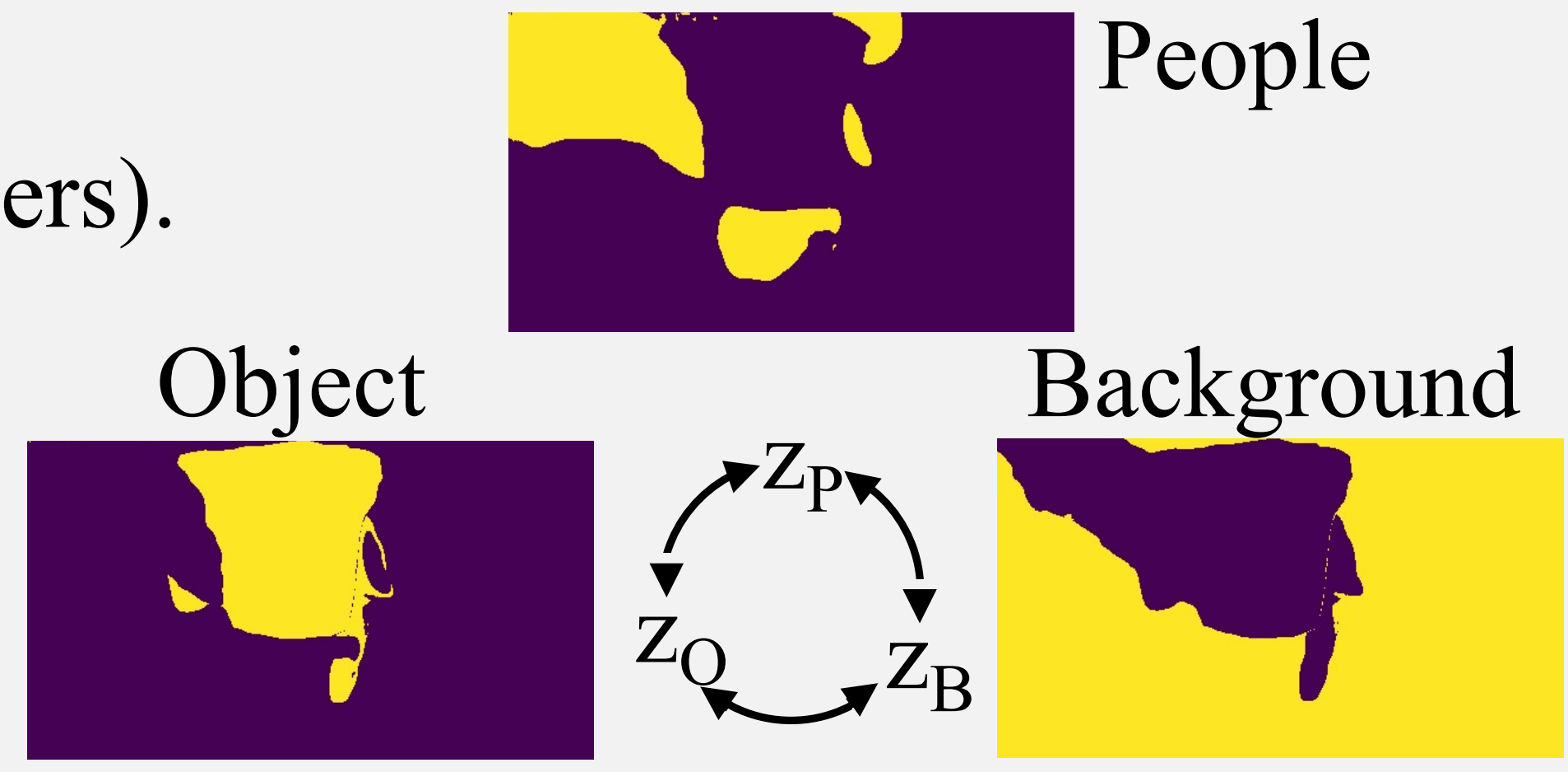
- $O \in \mathbb{R}^{H\times W\times 2}$: Optical flow.
- d_{BG}, d_k : Distances between an optical flow vector and a model.
- Hand vertices between 2 frames are used to fit a Homography.



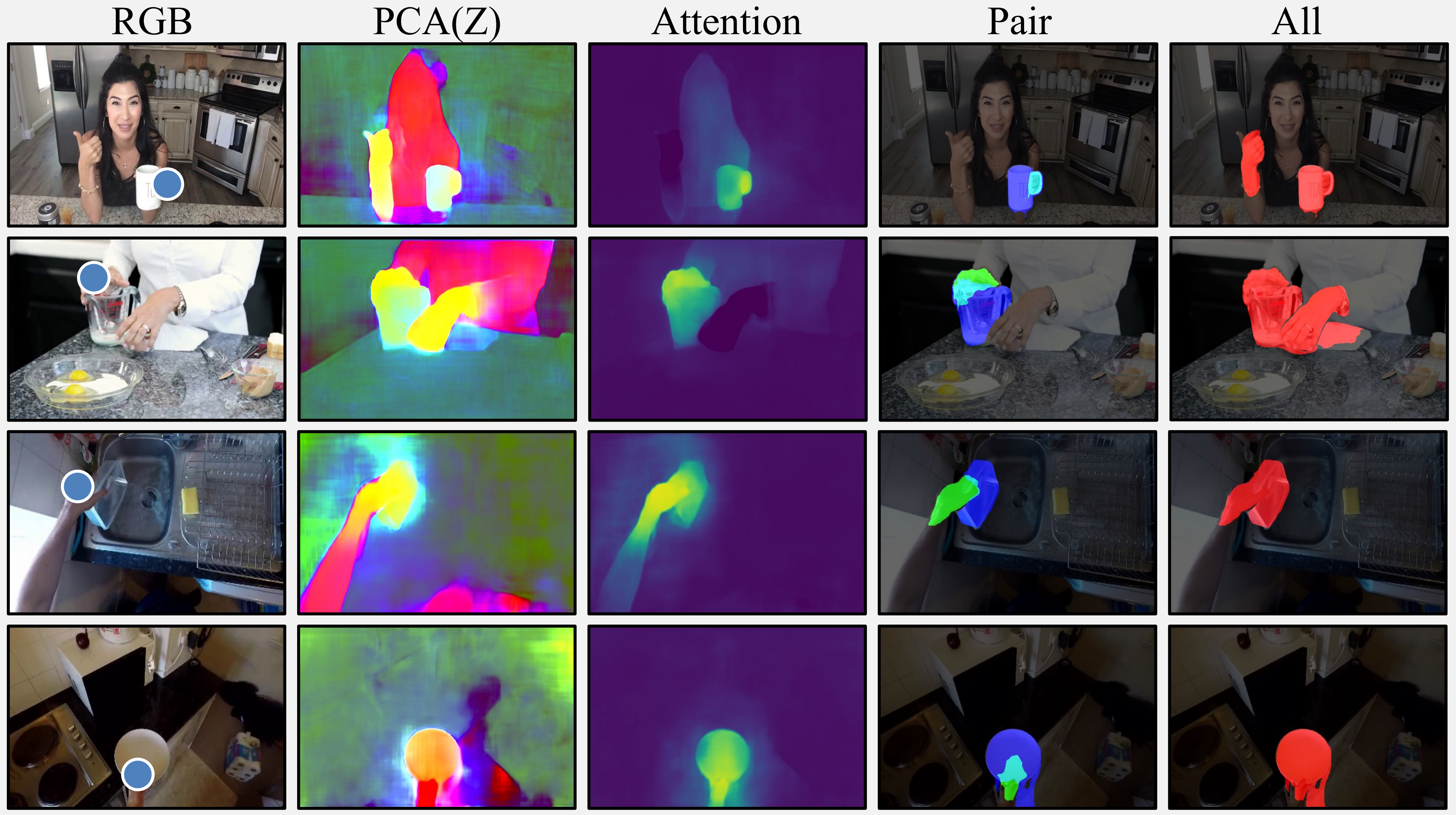
Method



- Feature Embedding (Z, a standard U-Net-style network):
  - Backbone: SE-Net (se-resnext50-4d) with ImageNet pretrained weights.
- Hand Branch (Q) and Object Branch (K):
  - Both are light-weight branches (2 3x3 conv layers).
- Objective Function:
  - Contrastive loss: 3-way contrastive.
  - Responsibility loss: supervise responsibility.



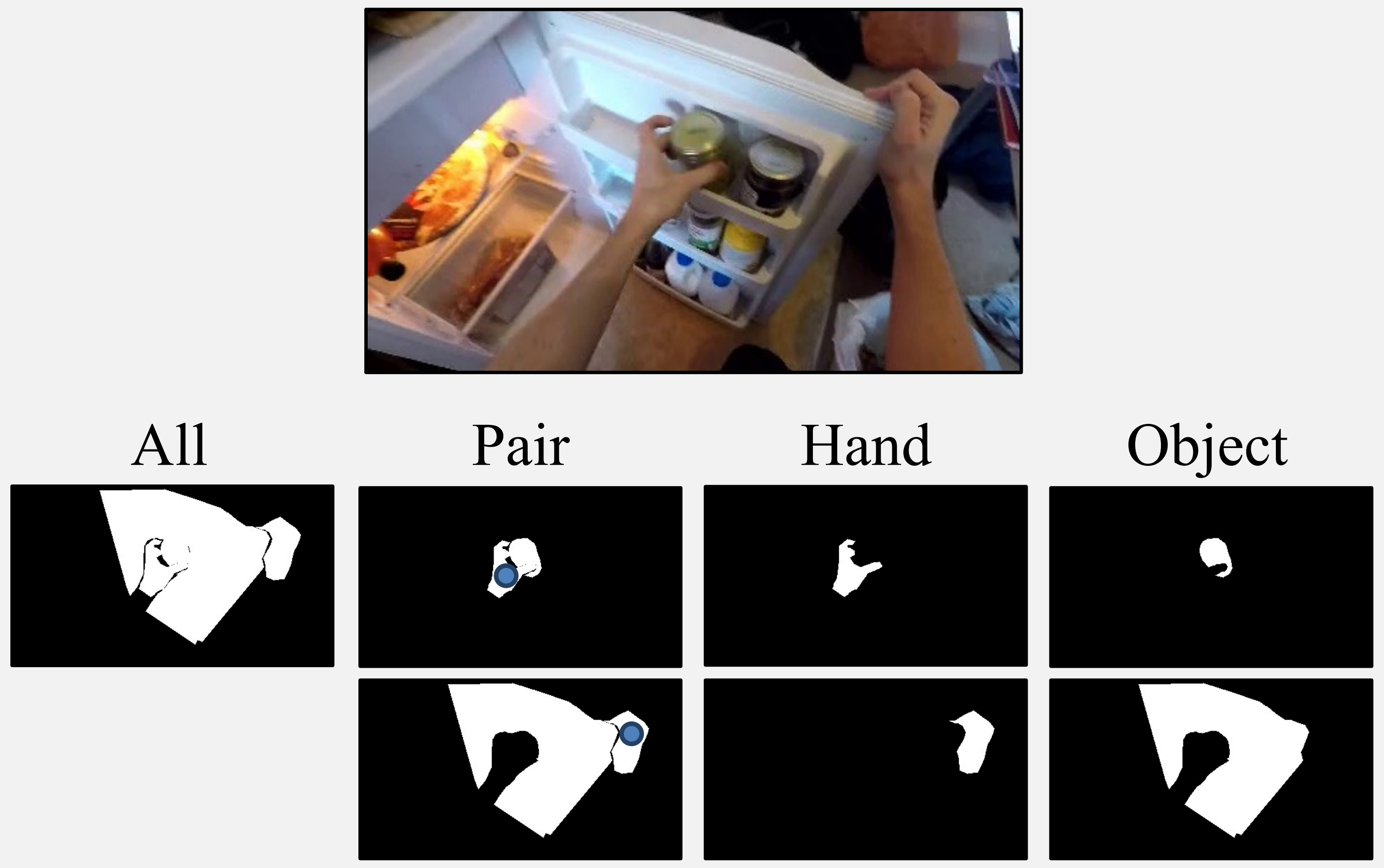
Qualitative Results



Experiments

Tasks: We evaluate the predictions on 4 tasks: hand, object, pair, and all segmentation.

Four Evaluation Tasks



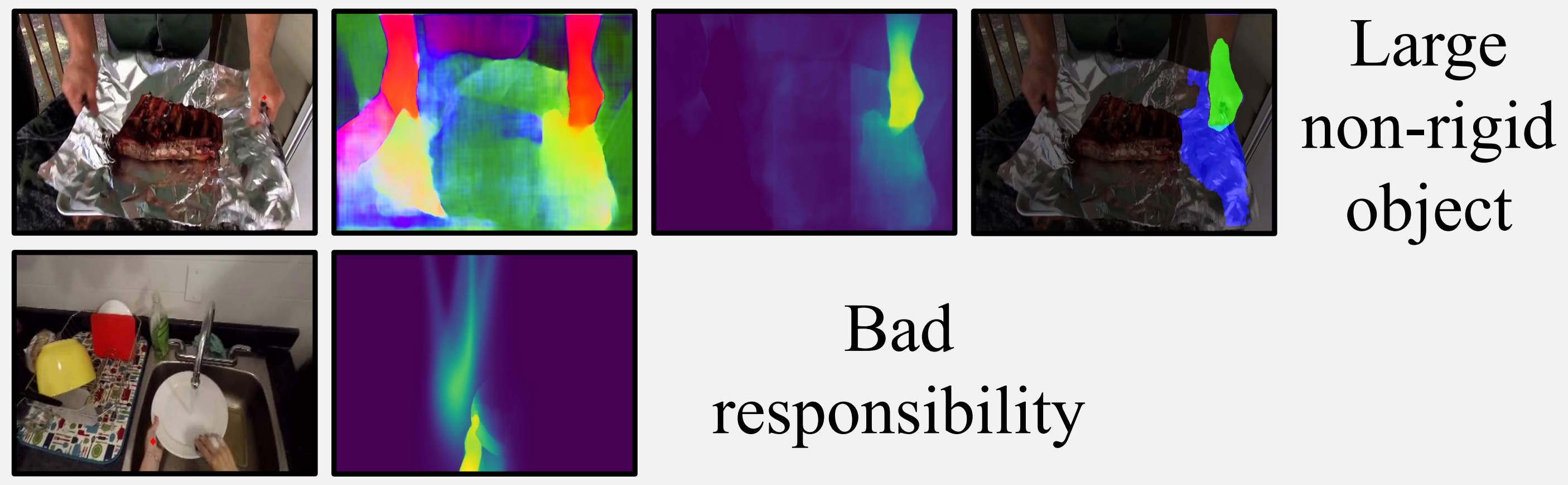
Baselines

	100DOH				EPICK			
	All	Pair	Hand	Obj	All	Pair	Hand	Obj
COHESIV	51.9	46.5	53.6	29.3	43.2	42.1	60.7	19.5
Saliency	25.2	20.1	8.6	17.0	21.6	15.9	6.0	11.7
Flow	29.3	21.5	12.9	12.1	15.4	11.9	6.2	6.6
Thresholded Responsibility	44.5	37.0	-	-	42.9	30.0	-	-
Supervised Bounding Box	56.9	47.0	56.5	34.9	54.3	44.8	53.8	34.4

Ablations

	100DOH				EPICK			
	All	Pair	Hand	Obj	All	Pair	Hand	Obj
COHESIV	51.9	46.5	53.6	29.3	43.2	42.1	60.7	19.5
Attention-Only	42.8	40.0	-	-	38.1	37.8	-	-
Embeddings-Only	25.7	18.3	13.2	22.9	30.0	20.8	24.6	14.4
COHESIV w/ ResNet Backbone	45.8	41.2	48.1	25.2	39.8	39.1	55.2	17.9
COHESIV w/ Predicted Query	47.7	42.8	47.8	28.1	40.0	38.6	55.1	19.4

Failure Cases



* indicates equal contribution.