HO-Cap: A Capture System and Dataset for 3D Reconstruction and Pose Tracking of Hand-Object Interaction

Jikai Wang[1] Qifan Zhang[1] Yu-Wei Chao[2] Bowen Wen[2]
Xiaohu Guo[1] Yu Xiang[1]

[1] The University of Texas at Dallas, [2] NVIDIA

39th Conference on Neural Information Processing Systems (NeurIPS 2025)

Data Capture Setup

- Multi-view Cameras
- Calibration & Synchronization

Annotation

- 3D Object Reconstruction
- Object Pose Estimation
- Hand Pose Estimation
- Joint Hand-Object Pose Optimization

Dataset

- Dataset Statistics
- Annotation Types

Baseline

- Hand Pose Estimation
- Object Detection
- Object Pose Estimation

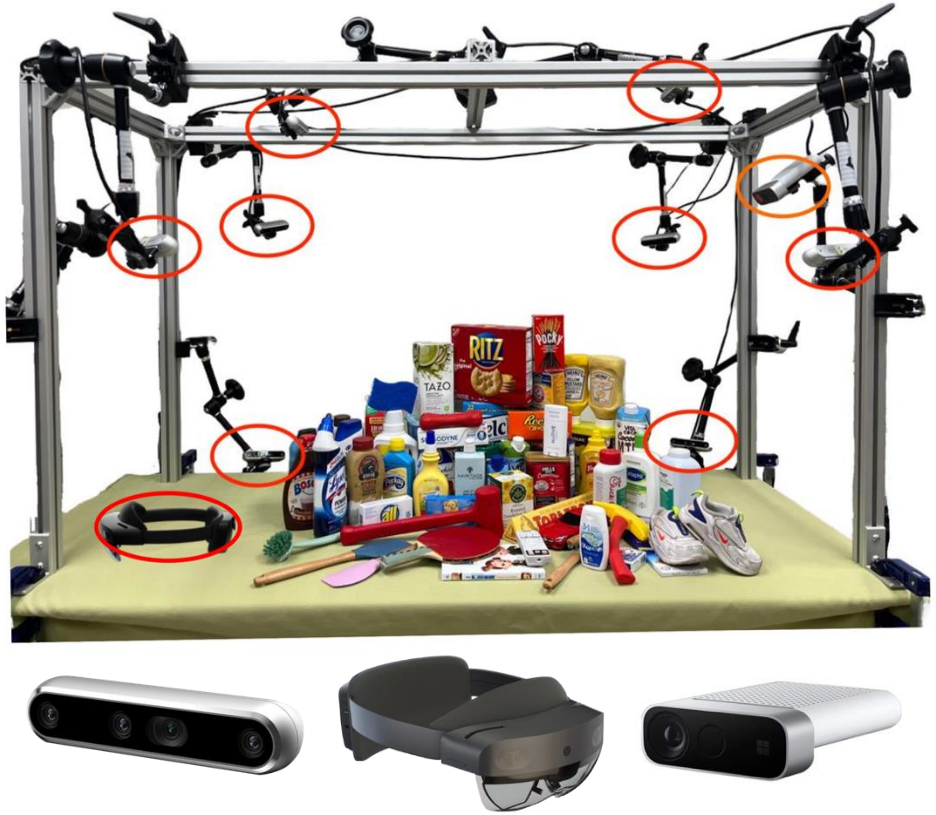
Mulit-View + Egocentric Capture Setup

□ Hardware Configuration

- **8× Intel RealSense D455** - cover the entire workspace from multiple angles.
- **1× Azure Kinect** - provides high-resolution depth for detailed 3D reconstruction.
- **1× HoloLens 2** - records egocentric RGB-D data for first-person analysis.

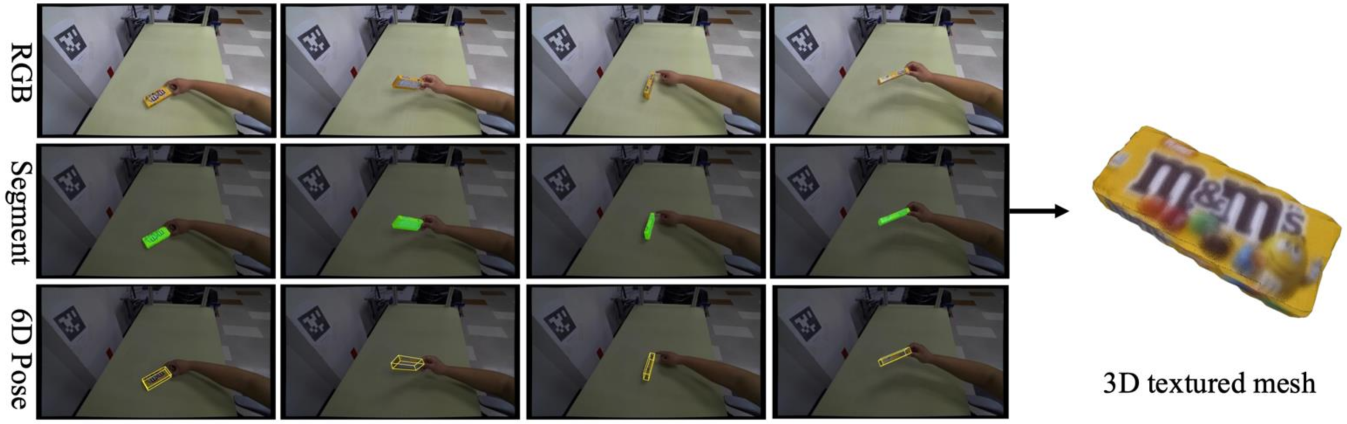
□ Calibration & Fusion

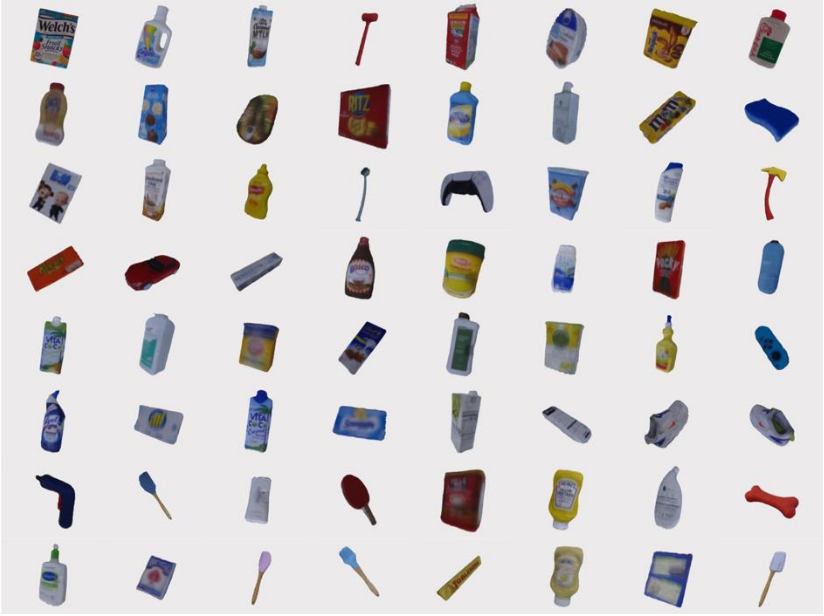
- All RealSense cameras are extrinsically calibrated.
- Head poses are directly obtained from the HoloLens.
- Streams are .synchronized by timestamps and fused into a single world frame



Capture Setup | Annotation | Dataset | Baseline

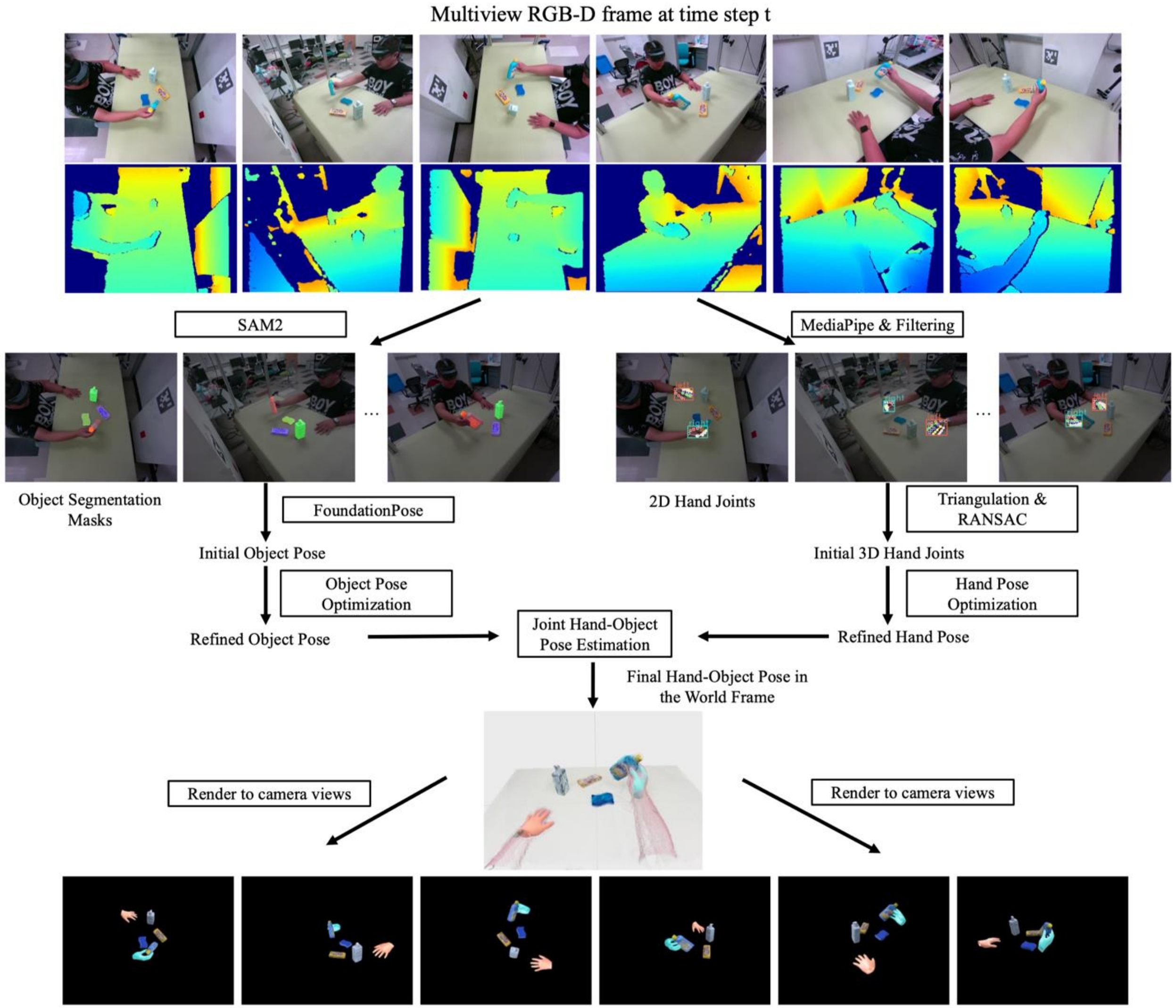
3D Object Reconstruction (BundleSDF)



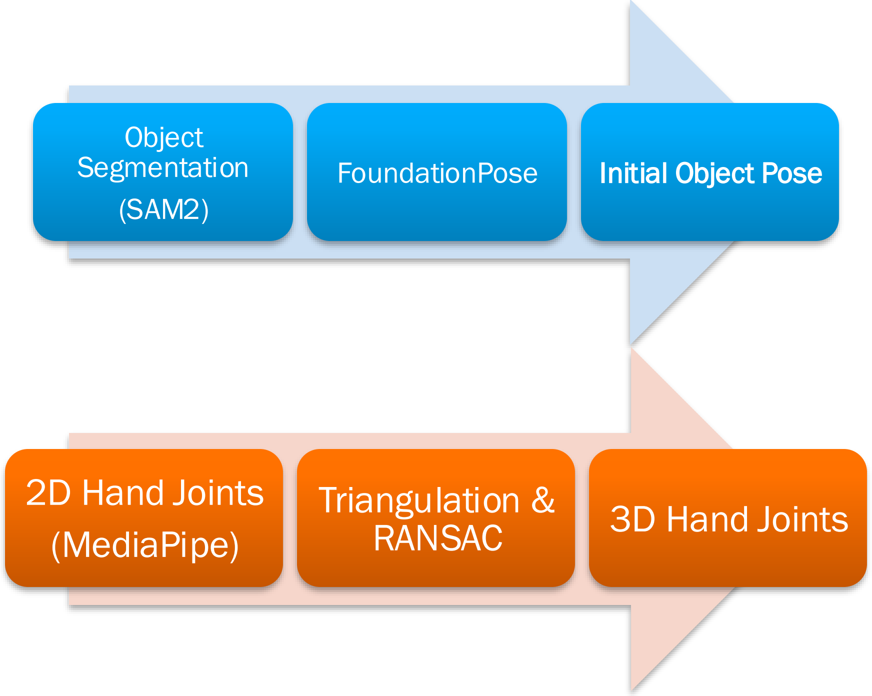


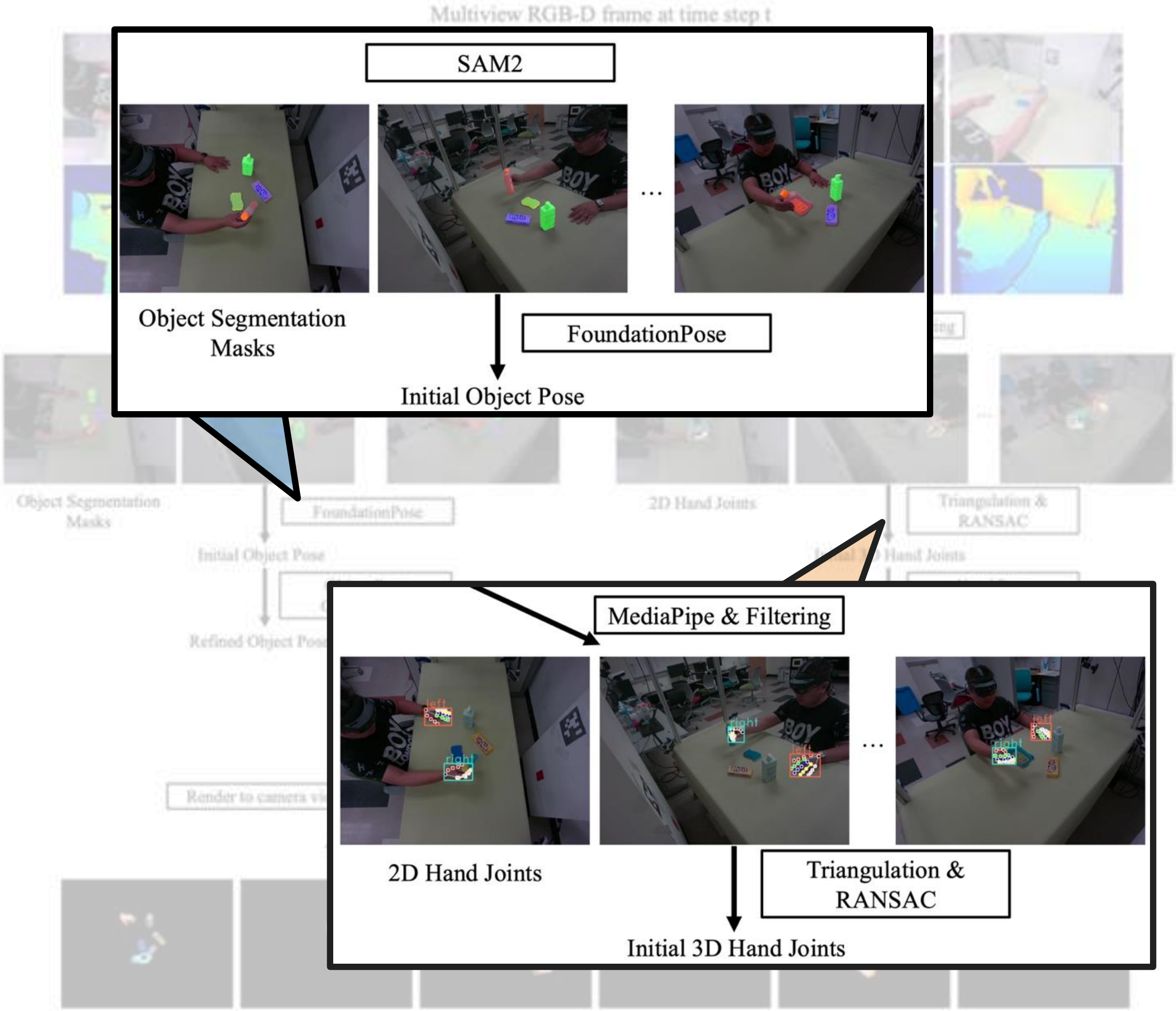
Annotation Pipeline

We propose a semi-automatic annotation pipeline leverages large pre-trained models along with SDF-based optimization, requiring only minimal manual initialization.

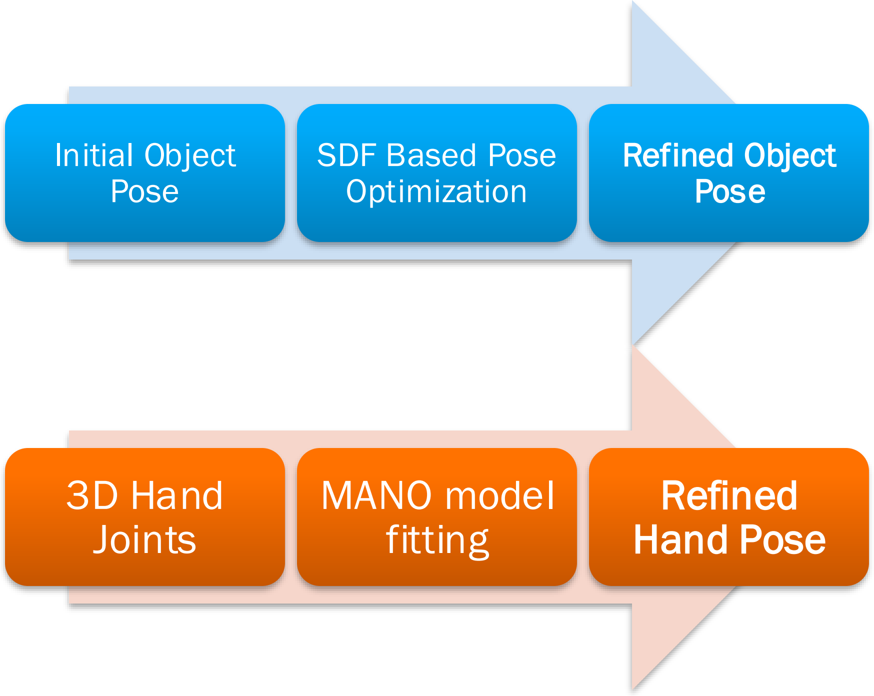


Stage One: Initial Object Poses and 3D Hand Joints Estimation



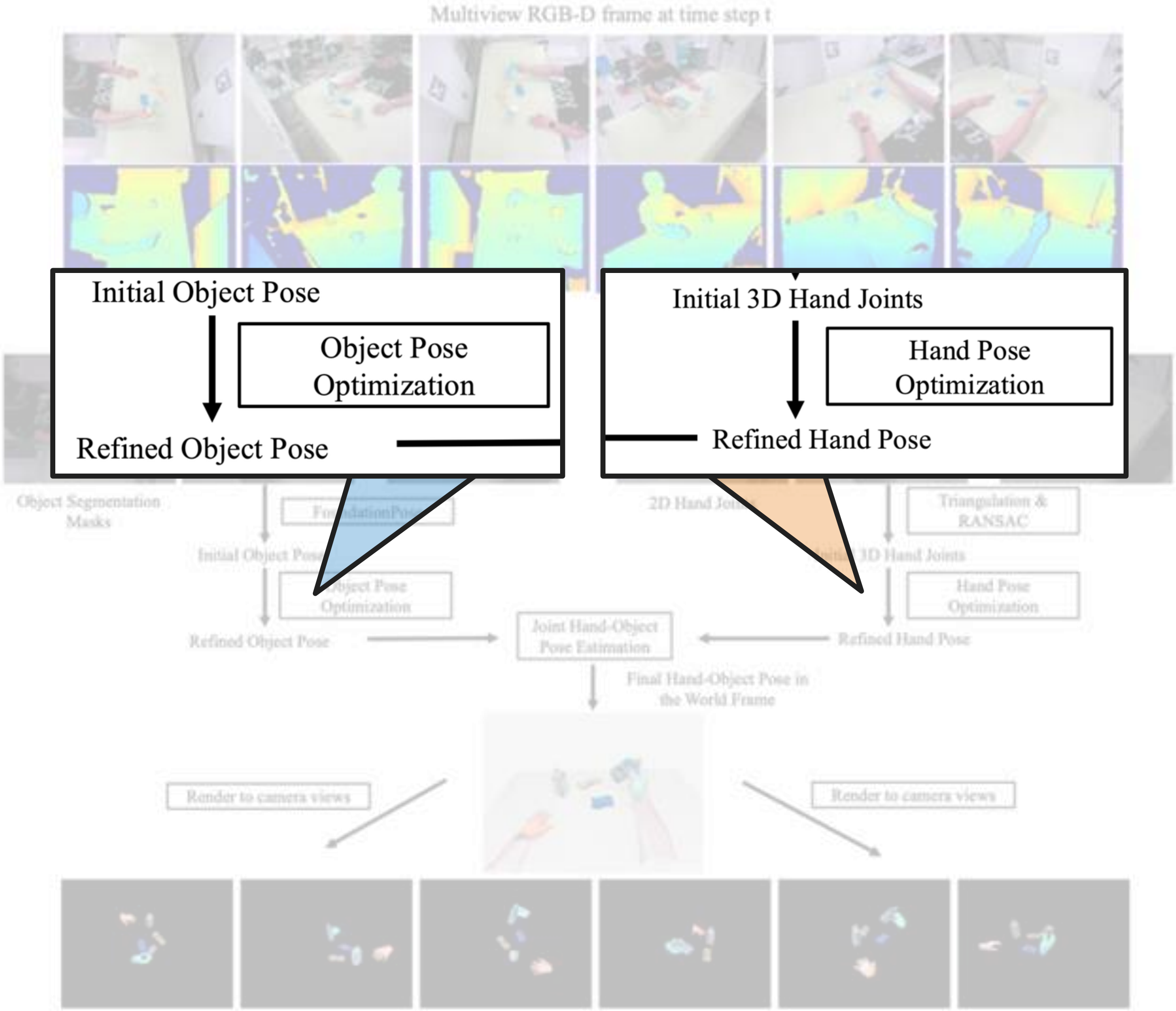


Stage Two: Refined Hand and Object Poses



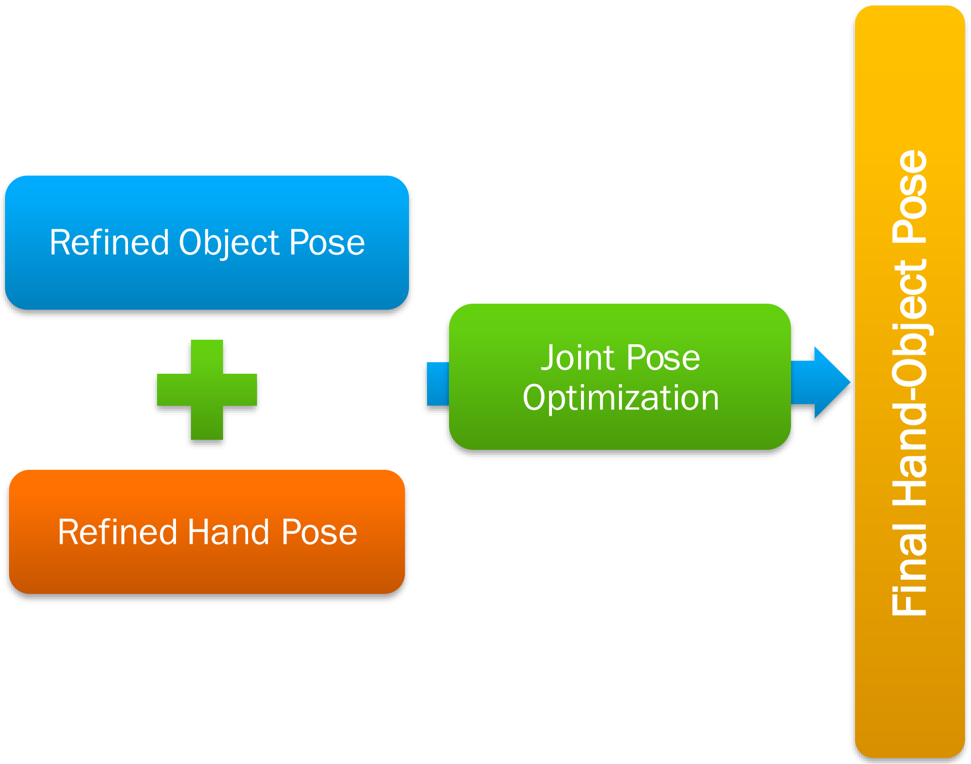
Initial Object Pose → SDF Based Pose Optimization → **Refined Object Pose**

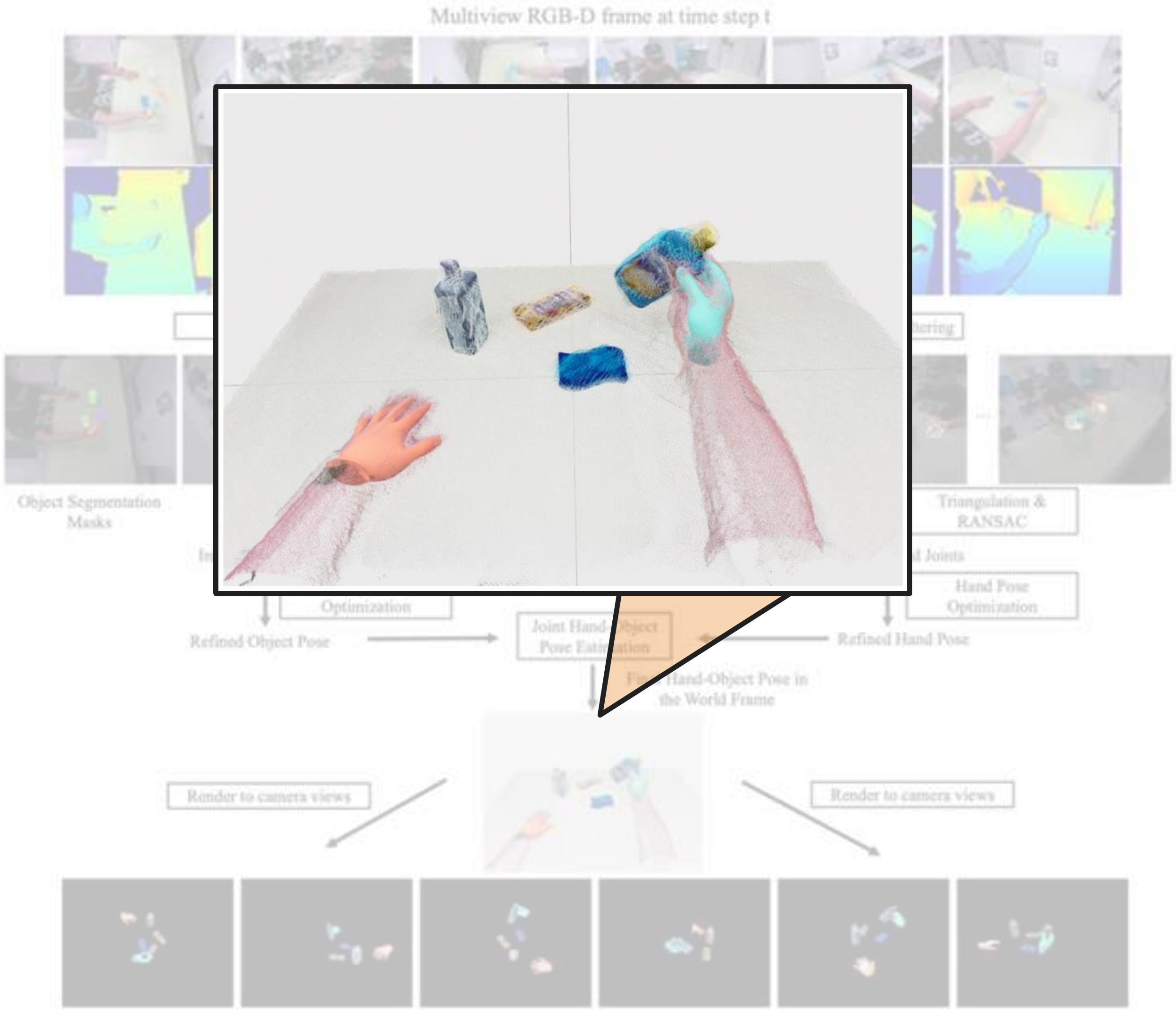
3D Hand Joints → MANO model fitting → **Refined Hand Pose**



Capture Setup | Annotation | Dataset | Baseline

Final Hand and Object Poses

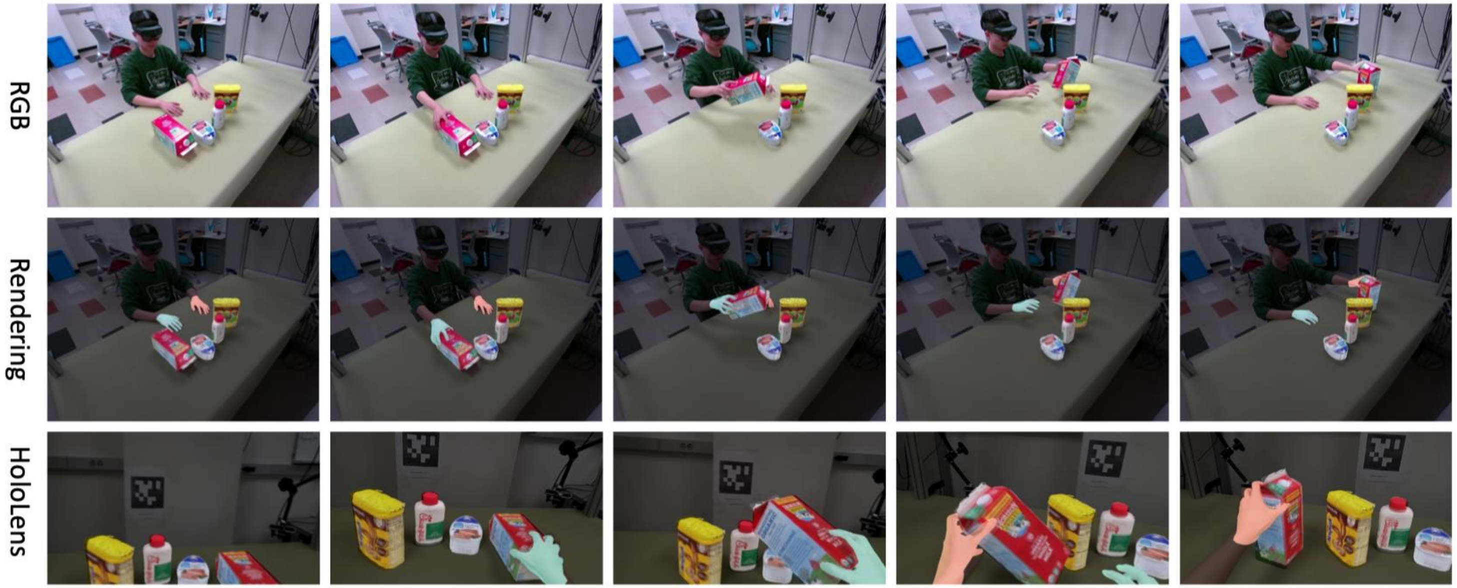




Dataset Statistics

❑ 9 Camera, 9 Subjects, 64 Unique Objects, 64 Video Sequences, 3 Tasks

❑ ~ 656K Markerless RGB-D Frames



Capture Setup

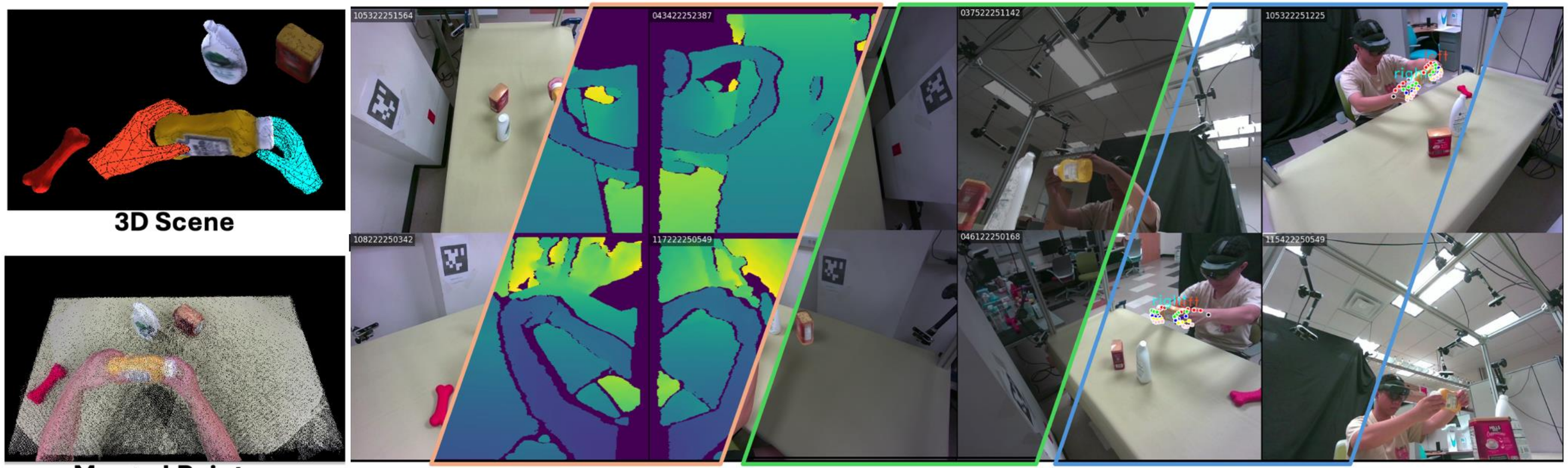
Annotation

Dataset

Baseline

Annotation Types

- ❑ MANO-based 3D Hand Pose, 6D Object Poses, 2D Hand Joint Keypoints, 6D Head Poses
- ❑ Hand and Object Segmentation Masks



3D Scene

Merged Points

Depth

Object Pose

Hand Pose

Benchmarks & Baselines

We provide a benchmark with baseline results for:

- □ Hand Pose Estimation
- □ Object Detection
- □ Object Pose Estimation

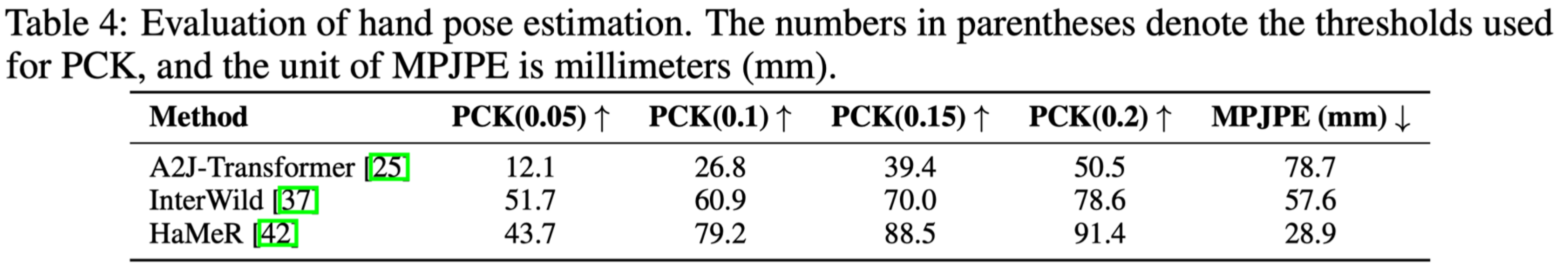


Table 4: Evaluation of hand pose estimation. The numbers in parentheses denote the thresholds used for PCK, and the unit of MPJPE is millimeters (mm).

Method	PCK(0.05) ↑	PCK(0.1) ↑	PCK(0.15) ↑	PCK(0.2) ↑	MPJPE (mm) ↓
A2J-Transformer [25]	12.1	26.8	39.4	50.5	78.7
InterWild [37]	51.7	60.9	70.0	78.6	57.6
HaMeR [42]	43.7	79.2	88.5	91.4	28.9

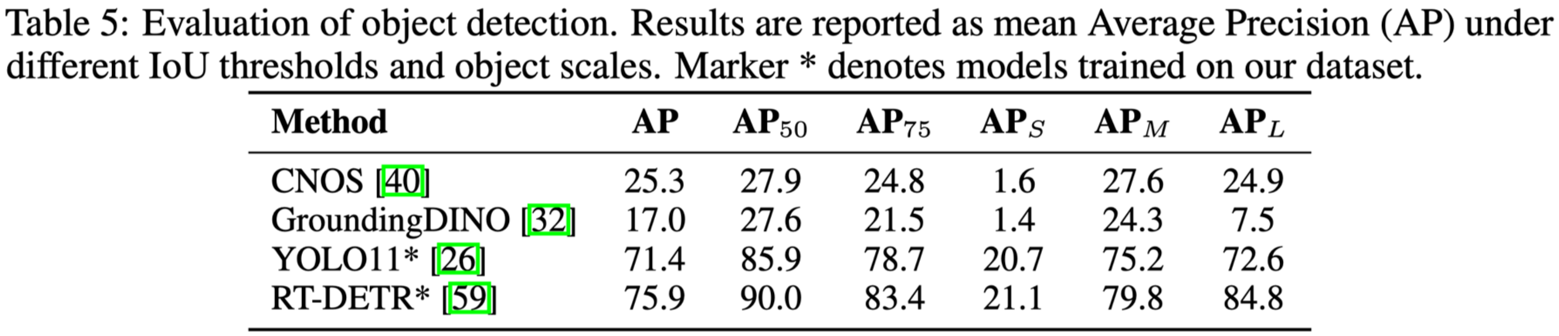


Table 5: Evaluation of object detection. Results are reported as mean Average Precision (AP) under different IoU thresholds and object scales. Marker * denotes models trained on our dataset.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
CNOS [40]	25.3	27.9	24.8	1.6	27.6	24.9
GroundingDINO [32]	17.0	27.6	21.5	1.4	24.3	7.5
YOLO11* [26]	71.4	85.9	78.7	20.7	75.2	72.6
RT-DETR* [59]	75.9	90.0	83.4	21.1	79.8	84.8

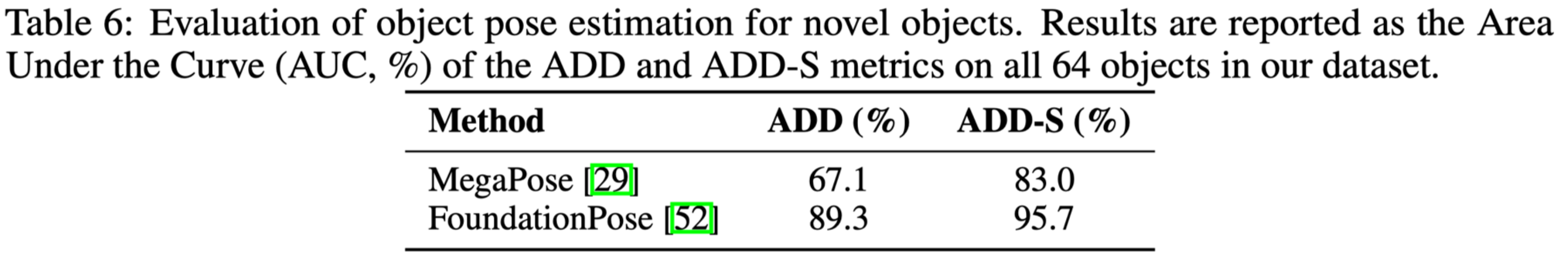
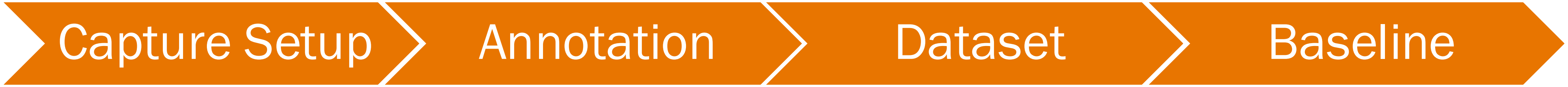


Table 6: Evaluation of object pose estimation for novel objects. Results are reported as the Area Under the Curve (AUC, %) of the ADD and ADD-S metrics on all 64 objects in our dataset.

Method	ADD (%)	ADD-S (%)
MegaPose [29]	67.1	83.0
FoundationPose [52]	89.3	95.7



Conclusion

- □ Multi-view, Markerless 3D hand-object Capture.
- □ Scalable Semi-automatic Annotation Method.
- □ Physically consistent hand-object poses.
- □ Enables Research in 3D Perception, Pose Estimation, and Robot Learning.

Project Website: ***https://irvlutd.github.io/HOCap***