

# Appendix

## A. Video Model Implementation

We adapted the pretrained Stable Video Diffusion model [1], which generates 25-frame videos at a time. In our adaptation, the first 13 frames correspond to the stereo view 1, and the last 12 frames correspond to the stereo view 2, captured from the two cameras. To condition the video model to generate a stereo video, we modified the per-frame image embedding based on the viewing angle of each output frame. Since each frame of the stereo videos should be paired but the video model generates an odd number of frames, the first frame of the video model output is always the same as the input and discarded at test time. The model training hyperparameters are given in Table 1. During inference, we use 30 denoising steps with a constant classifier-free guidance of 1.0. Additional qualitative results for the four tasks are shown in Fig 1, 2, 3 and 4.

| H-Param            | Res     | Lr   | Batch Size | Train Steps | Clip Duration | Fps | MotScr |
|--------------------|---------|------|------------|-------------|---------------|-----|--------|
| Rotation (Full DS) | 768×448 | 1e-5 | 4          | 16384       | 2.0           | 6   | 200    |
| Rotation (2/3 DS)  | 768×448 | 1e-5 | 4          | 16384       | 2.0           | 6   | 200    |
| Rotation (1/3 DS)  | 768×448 | 1e-5 | 3          | 15360       | 2.0           | 6   | 200    |
| Scooping           | 768×448 | 1e-5 | 4          | 16384       | 3.0           | 5   | 200    |
| Sweeping           | 768×448 | 1e-5 | 4          | 16384       | 3.0           | 5   | 200    |
| Push-Shape         | 768×448 | 1e-5 | 4          | 17408       | 2.0           | 6   | 200    |

Table 1: **Hyperparameters for Video Model Training.** Res: image and video resolution, Lr: learning rate, Batch Size: batch size, Training Steps: training steps for the evaluation checkpoint, Clip Duration: single demonstration video length in seconds, Fps: the video sub-sampling frame rate and model fps parameter, MotScr: model motion score parameter.

## B. Experimental Setup

The stereo camera setup consists of two Intel RealSense D435i cameras spaced approximately 660 mm apart at a 45° angle. The distance between the cameras and the table is about 760 mm. The real-world data collection and the robot experiment setups are shown in Fig. 9, 10. The training videos are recorded at a resolution of 1280×720 and are then cropped and resized to the appropriate resolution for model input. The table surface used for data collection and experiments is covered with a black cloth, which introduces variations in friction and increases uncertainty in the Push-Shape experiments. For the rotation and scooping experiments, UFACTORY xArm 7 robots are used, while UR5 robots are used for the sweeping and Push-Shape experiments. For calculating the mIoU in the Push-Shape experiment, the view from the stereo camera 1 is used. In each trial with multiple steps, the resulting image with the highest IoU with the target is used to calculate the rotation error. In sweeping and push-shape experiments, the robot end-effector height is limited to avoid robot collision with the table top.

## C. Data Collection

For all tasks, the first frame of the human demonstration video is an image of the scene. The subsequent frames include the human demonstrator using the tool to perform the manipulation. In the Push-Shape demonstration, an object is pushed to a location in multiple steps. The final position of the object is used as a mask and blended with the entire video for the target position. The objects used in training and testing for different tasks are shown in Fig. 5, 6, 7 and 8.

## D. Object Tracking

In the videos, the tool is tracked using MegaPose [2]. Utilizing a stereo setup, the center of the tracked object from each camera are projected into 3D space as a straight line. The translation component of the object in 3D space is determined by finding the midpoint between the projected lines from the two cameras. The rotation component of the object is obtained by averaging the

object rotations from the two views. This refined object pose from the stereo setup enhances the accuracy of the object’s depth measurement from the cameras. In the scooping task, only the handle of the scooper is tracked to avoid inaccuracies due to occlusion by particles. In the sweeping and Push-Shape tasks, the tool without the handle is tracked, as the handle is occluded by the human hand. To obtain the tool trajectories for training the Diffusion Policy, the same stereo tracking is applied to the demonstration videos.

### E. Diffusion Policy Baseline

We use a CNN-based Diffusion Policy as our baseline, employing two pretrained ResNet-18 [3] image encoders to process the stereo images of the scene. The input images have a resolution of  $384 \times 224$ , similar to the original implementation resolution. We found that higher resolution input images did not improve model performance.

## References

- [1] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.
- [2] Y. Labbé, L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, J. Carpentier, M. Aubry, D. Fox, and J. Sivic. Megapose: 6d pose estimation of novel objects via render & compare. *arXiv preprint arXiv:2212.06870*, 2022.
- [3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

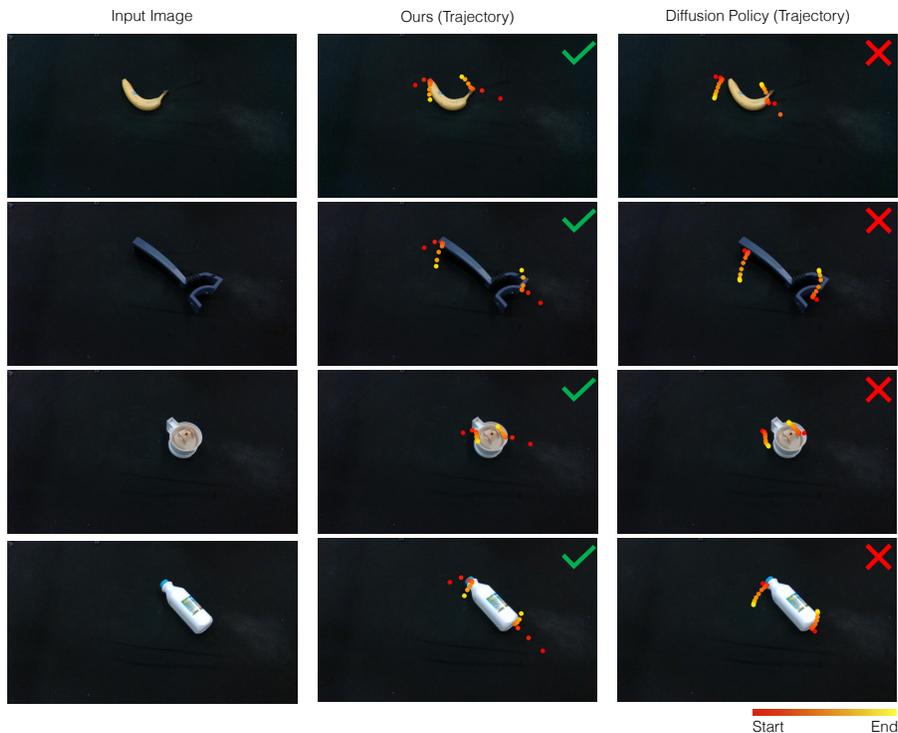


Figure 1: **Additional Rotation Qualitative Results.** The trajectories of the end-effectors are projected onto the input image.

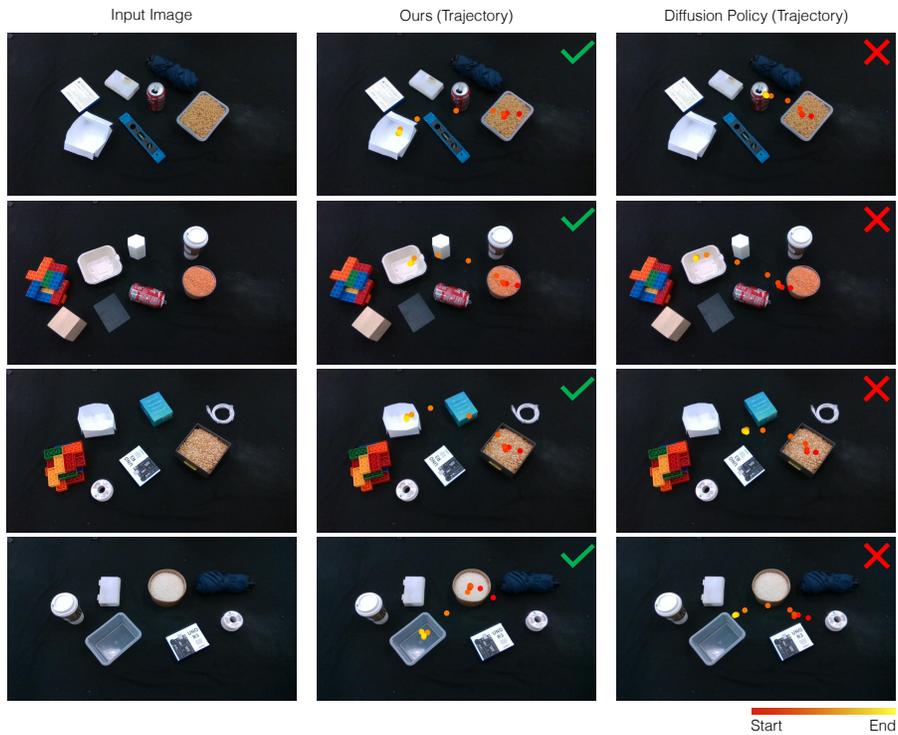


Figure 2: **Additional Scooping Qualitative Results.** The trajectory of the end-effector is projected onto the input image.



Figure 3: **Additional Sweeping Qualitative Results.** The trajectory of the end-effector is projected onto the input image.

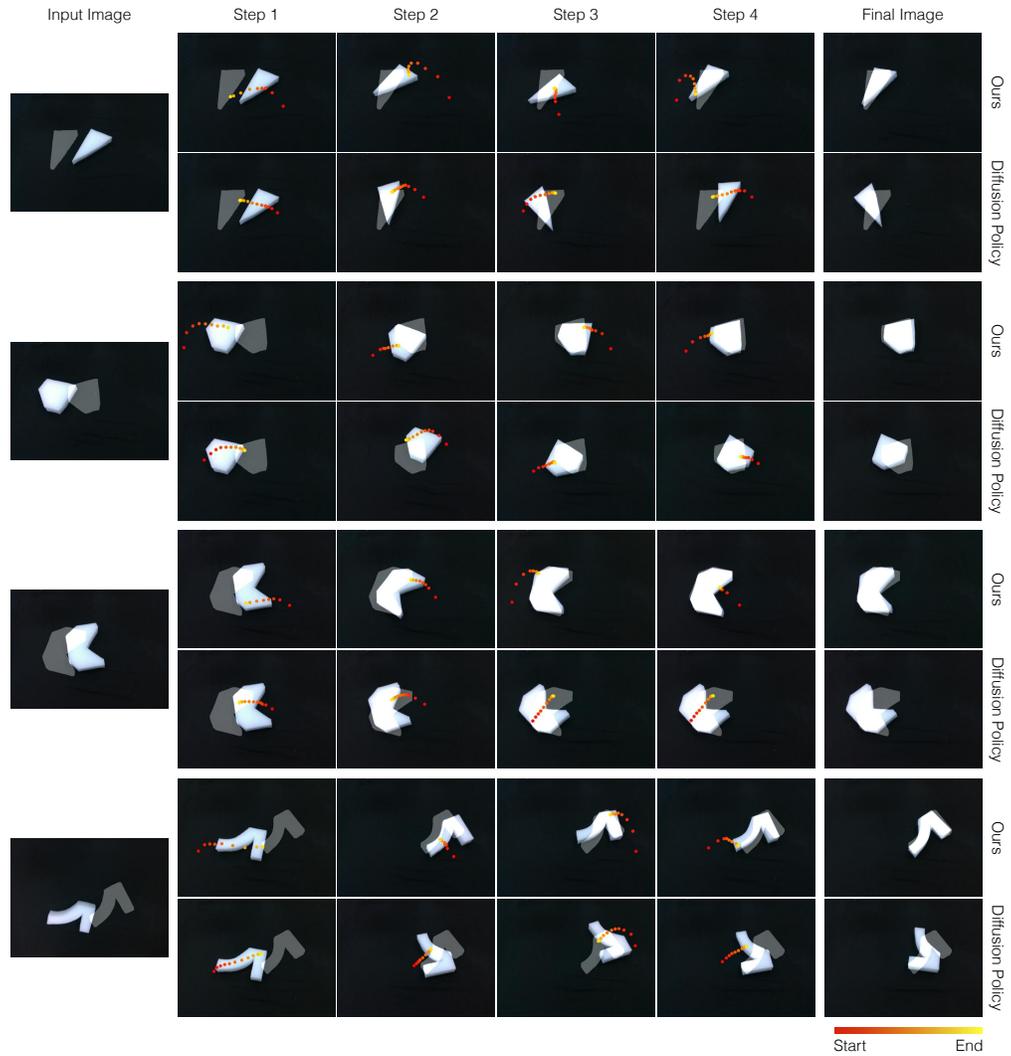


Figure 4: **Additional Push-Shape Qualitative Results.** The trajectory of the end-effector is projected onto the input image.



Figure 5: **Rotation Objects.** The training set includes 14 real-world objects and 17 custom colored shapes made out of foam. The testing set includes 10 challenging real-world objects.



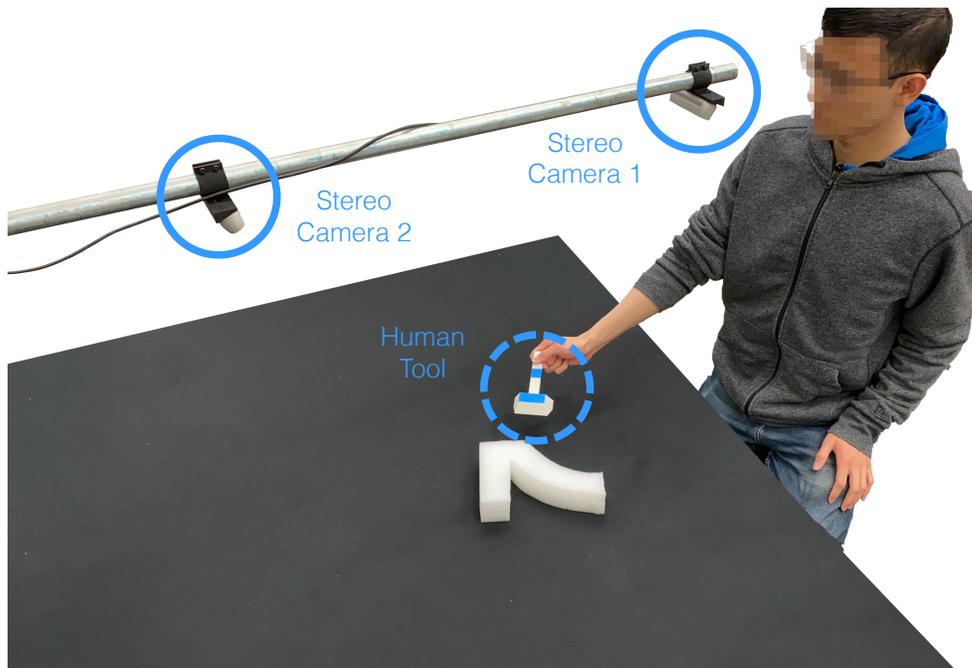


Figure 9: **Real-world Data Collection Setup.** The data collection setup has the same camera arrangement as the robot experiment setup.

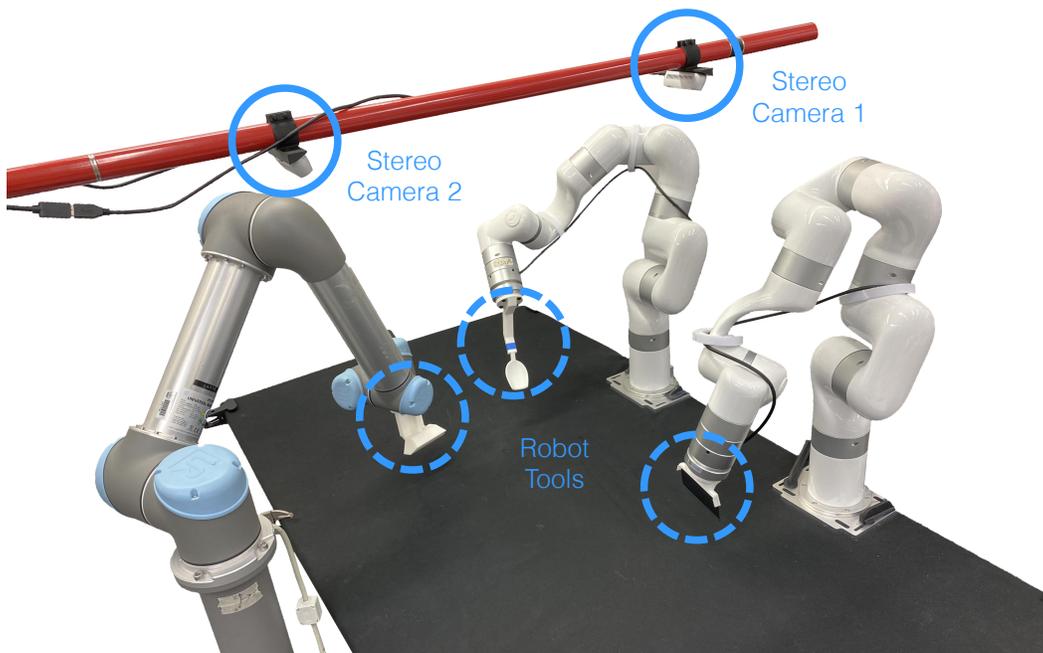


Figure 10: **Real-world Robot Experiment Setup.** The robot experiment setup includes the 3 robots to perform all the experiments.