1 Controlling the World by Sleight of Hand: Supplementary Material

Please see `coshand.cs.columbia.edu` for webpage and code.

1.1 Dataset Details

SomethingSomethingv2 The dataset contains over 180k videos of humans performing pre-defined, basic actions with everyday objects. The actions fall under 174 prescribed categories. We exclude videos that belong to the 'pretending' category and videos that do not include hands, as these are not relevant to our task. We utilize hand bounding box annotations provided in the Something-Else dataset [3], however we note that bounding boxes can quite easily and accurately be obtained from off-the-shelf methods such as GroundingDINO [1] making this approach scalable to larger-scale datasets.

In-the-wild Dataset To test the robustness of `CoSHAND` outside of the SomethingSomethingv2 dataset, we collected 45 videos of ourselves interacting with objects in our lab and home settings against various backgrounds. These interactions are listed below (note that some of these actions have multiple videos associated with them, hence only 25 actions are present but with possibly multiple instances per action): opening books, knocking a bottle over, pushing a bottle, lifting a box, opening a container, pulling a chair, pushing a chair, opening a drawer, moving a keyboard across the table, putting pack of gum inside drawer, squishing a soft ball, poking play dough, rotating a bulb, pushing a pencil which then pushes a box, stacking eraser and spray, stacking eraser and marker, separating two markers, moving an eraser across the whiteboard, moving a marker to the side, putting an eraser inside a mug, putting a box inside a mug, moving a pencil behind a mug, separating a picture frame and a bottle , moving a bottle to the side, squishing a pillow.

1.2 Training Details:

We finetuned our model on an 8×A100-80GB machine for 7 days and use AdamW [2] with a learning rate of 10^{-4} for training. We reduce the image size to 256 × 256 (with a 32 × 32 latent dimension), in order to be able to increase the batch size to 192 because larger batch sizes allows for increased training stability and convergence rates. (Note that with more compute and a larger latent space, we could achieve higher quality results as more spatial information would be preserved). We experiment with variations of the classifier free guidance scale at test-time (see **??**). UCG, TCG, and, `CoSHAND` are initialized with the Stable Diffusion provided checkpoint.

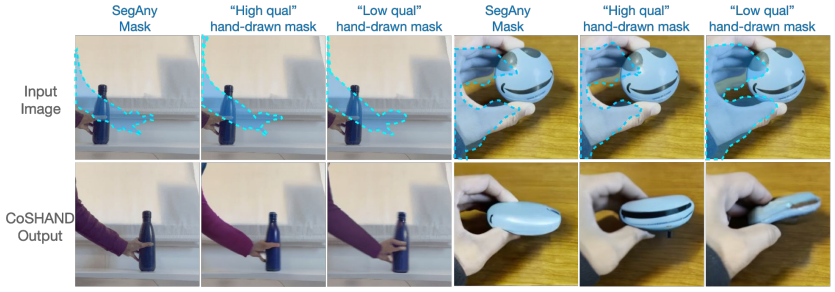

Fig. 1: Robustness to handmasks: `CoSHAND` is robust to query handmasks obtained from different sources. Here we show three sources including: off-the-shelf segmentation models ('SegAny'), manually-drawn 'high quality' query hand-masks, and manually-drawn 'low quality' query hand-masks (where details of hand/fingers are not captured in mask).

References

1. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection (2023)
2. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2019)
3. Materzynska, J., Xiao, T., Herzig, R., Xu, H., Wang, X., Darrell, T.: Something-else: Compositional action recognition with spatial-temporal interaction networks. In: CVPR (2020)