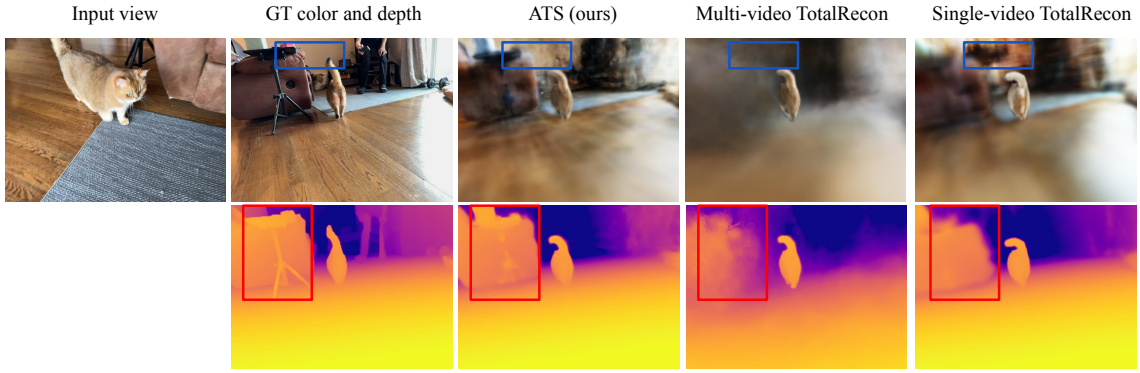

Figure A: **Qualitative comparison on 4D reconstruction (Tab. C).** We compare with TotalRecon on 4D reconstruction quality. We show novel views rendered with a held-out camera that looks from the opposite side. ATS is able to leverage multiple videos captured at different times to reconstruct the wall (blue box) and the tripod stand (red box) even they are not visible in the input views. Multi-video TotalRecon produces blurry RGB and depth due to bad camera registration. The original TotalRecon takes a single video as input and therefore fails to reconstruct the regions (the tripod and the wall) that are not visible in the input video.

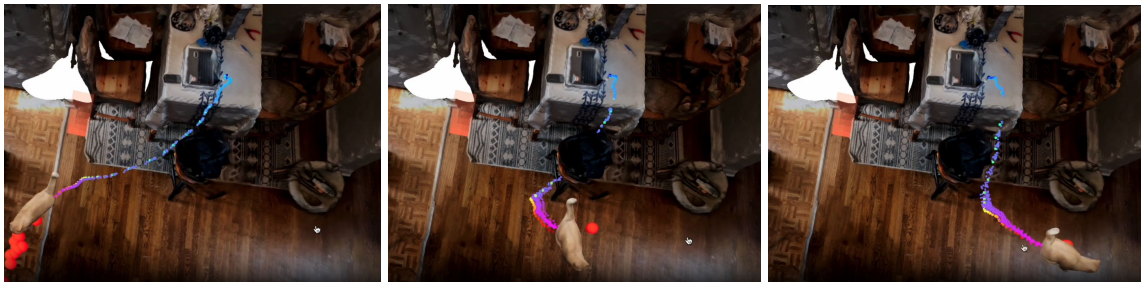

Figure B: **Generalization ability of the behavior model.** Thanks to the ego-centric encoding design (Eq. 12), a specific behavior can be learned and generalized to novel situations even it was seen once. Although there's only one data point where the cat jumps off the dining table, our method can generate diverse motion of cat jumping off the table while landing at different locations (to the left, middle, and right of the table) as shown in the visual.

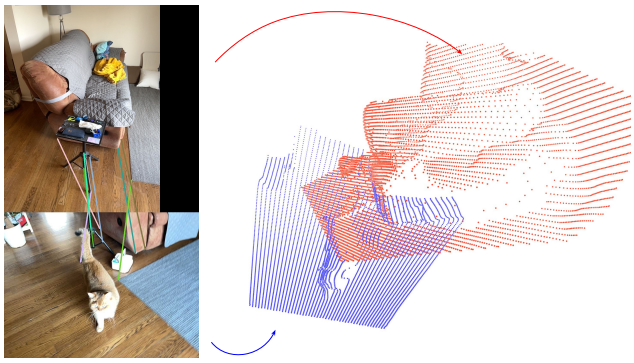

Figure C: **GT correspondence and 3D alignment.** Left: Annotated 2D correspondence between the canonical scene (top) and the input image (bottom). Right: we visualize the GT camera registration by transforming the input frame 3D points (blue, back-projected from depth) to the canonical frame (red). The points align visually.

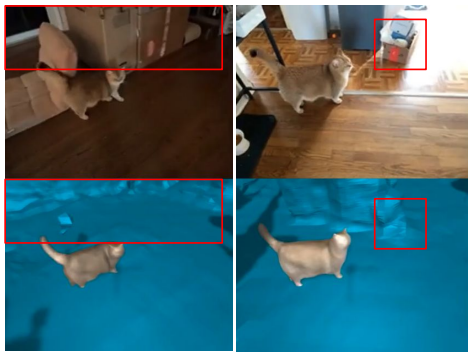

Figure D: **Robustness to layout changes.** We find our camera localization to be robust to layout changes, e.g., the cushion and the large boxes (left) and the box (right). However, it fails to *reconstruct* layout changes, especially when they are only observed in a few views.