

A APPENDIX

A.1 CODE

Our code and dataset are available at this anonymous external link: https://anonymous.4open.science/r/Interactive_Action-E033

A.2 MORE SYNTHETIC DATA

In the main text, we only shows several sampled input images, here we show more images by SAPIEN and images from Unity3D which are used for ablation study.

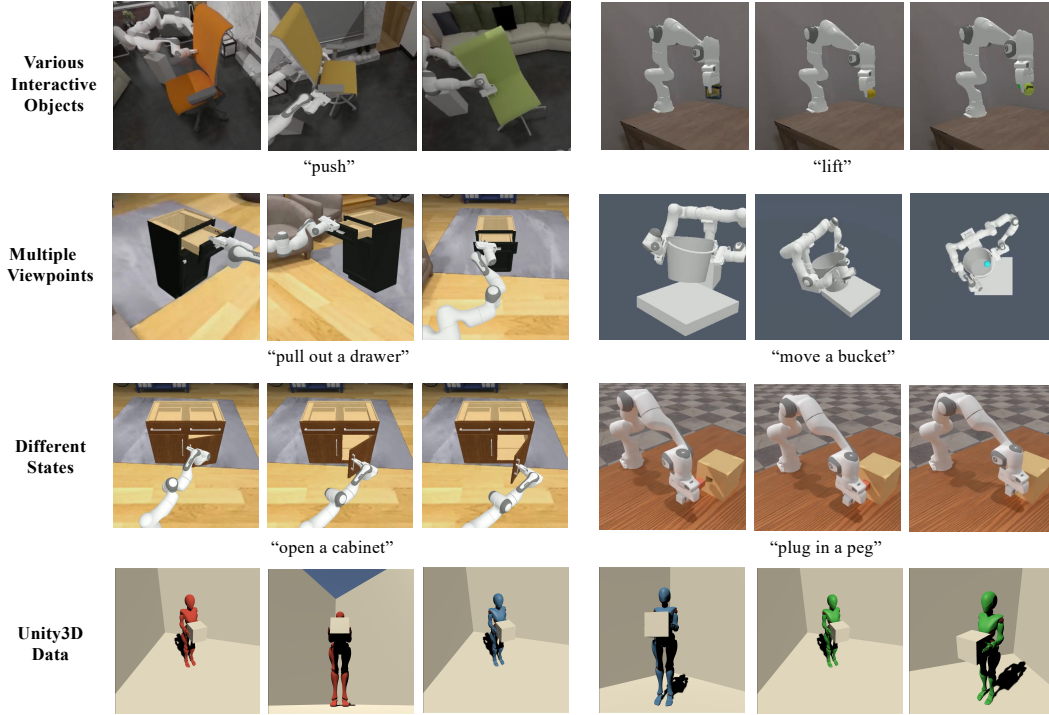


Figure 9: Examples of synthetic dataset built with SAPIEN engine. This dataset contains various interactive actions, various objects, multiple rendering viewpoints and different states which offers sufficient physical knowledge. Note that here the name of the action is only for representation convenience since one action can correspond to several synonymous words the action "pick up the cube" can be replaced by "lift the cube", "open the door" can be replaced by "pull out the door".

A.3 DATA GENERATION PIPELINE

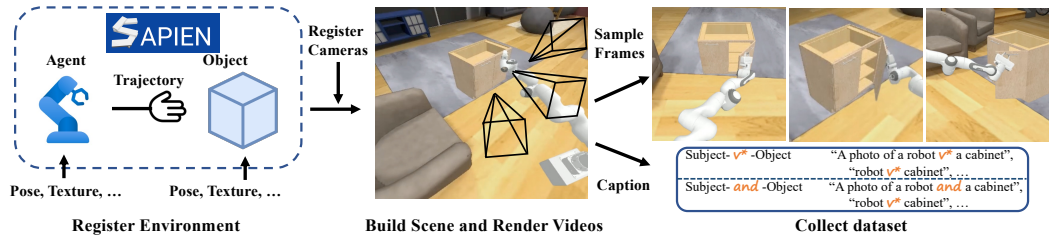


Figure 10: Synthetic data generation pipeline based on SAPIEN.

As can be seen from Figure 10, SAPIEN is a 3D realistic and physics-rich simulation platform. The platform provides an extensive library of pre-designed articulated robot agents, objects, and

habitat assets, enabling users to simulate realistic interactive environments. These robot agents are programmed to manipulate objects along trajectories sampled from a predefined set. Under the guidance of these sampled trajectories, the agents can efficiently navigate and interact with objects.

Users have the flexibility to select specific objects for the agents to interact with, enhancing the customization of the simulation. The habitat scenes are meticulously crafted to offer a realistic background, adding depth and authenticity to the simulations. This realism is further bolstered by user-defined cameras, which allow for the capture of the interactive actions from multiple view-points. To optimize the observational quality of these interactions, users can finely adjust the coordinates of both the agents and objects, as well as the camera angles. This adjustability ensures that users can obtain clear and precise observations of the interactive actions.

After the image data obtained, we caption them with two kinds of textual prompts: “Subject- v -Object” structure and “Subject-and-Object” structure. Thus we obtain paired training data.

A.4 LIMITATIONS AND BROADER IMPACT



Figure 11: Examples of our failure cases. The leftmost represents that some animals are hard to finish some interactive tasks, the middle shows that sometimes subject may not match with the textual prompts, the rightmost indicates that although generated images can satisfy the interaction requirement, there are still some counterfactual results.

Limitation. As shown in Figure 11, our work can only activate the inherent knowledge within stable diffusion models but struggles to incorporate new knowledge. This limitation is influenced by the pretrained stable diffusion models, which vary in effectiveness depending on the subject. For instance, it is challenging for the model to depict “an otter picking up objects”, likely because it has not been exposed to many images of otters performing such actions. In contrast, depicting a monkey performing similar tasks is much easier due to the model’s frequent exposure to such images. This reflects a bias in stable diffusion.

Moreover, the output is occasionally not well-aligned with the text, introducing irrelevant words and objects. Additionally, the precision, especially concerning angles, is insufficient.

Potential Negative Societal Impacts. The entity relational composition capabilities of our method could be applied maliciously on real human figures.