Supplementary Material for OBJECT 3DIT: Language-guided 3D-aware Image Editing

1 Training and Inference Details

We closely follow the training procedure established by [1], with a few modifications. Our approach uses an effective batch size of 1024, which is smaller than the batch size of 1536 used by Zero-1-to-3. This adjustment was necessary because of the additional memory requirements caused by the reintroduction of the CLIP text encoder. This batch size is achieved by using a local batch size of 64 across 40GB NVIDIA RTX A6000 GPUs, along with two gradient accumulation steps. Similar to Zero123, we train on images with a resolution of 256×256, resulting in a latent spatial dimension of 32×32. Following their protocol, we utilize the AdamW optimizer, with a learning rate of 1e-4 for all parameters of the model except for those of the concatenation MLP, which uses a learning rate of $1e-3$. Our training process runs for a total of 20,000 steps. We then select the best checkpoint based on our metrics computed from an unseen object validation set. As was the case in StableDiffusion, we freeze the CLIP text encoder during training. For inference, we generate images with the DDIM [2] sampler using 200 steps. We do not use classifier-free guidance, i.e. the cfg term is set to 1.0.

2 Robustness to Severity of Transformation

We analyze the robustness of our method by measuring the performance of the single task rotation model as the complexity of the scene and severity of transformation changes. In Figure 1, we show the average of our Mask PSNR metric as the number of objects in the scene varies from 1 to 4, where a slight drop in performance occurs as the number of objects increases. In Figure 2, we show average Mask PSNR for rotations in a given angle range on a pie chart, where it can be seen that the model does better with smaller angle deviations.

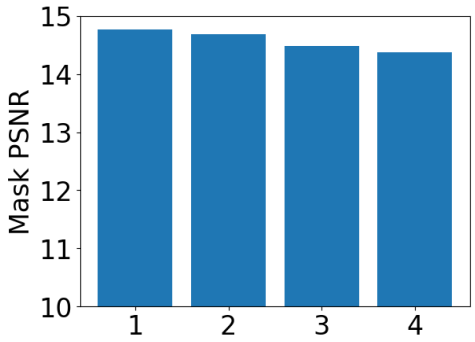

Figure 1: Average Mask PSNR of the single-task rotation model as the number of objects in the scene varies.

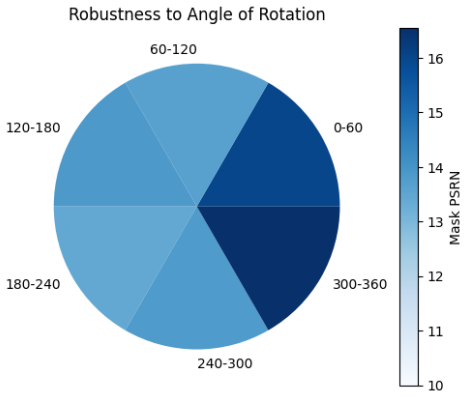

Figure 2: Average Mask PSNR of the single-task rotation model for rotation angles falling within a slice.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Table 1: Summary of key statistics of the OBJECT dataset.

Total objects	62950
Total categories	1613
Object per category median	6.0
Object per category mean	39.03
Object per category std	138.00

3 Additional Results

In Figure 3, we show qualitative results from our multitask model on each of the four editing tasks. We also attach three gifs showing our model manipulating a single CLEVR scene. In the rotation gif, we prompt the single task model to rotate the "blue box" by increments of 30 degrees. In the removal gif, we prompt the model to remove each object with the following descriptions along with a visually estimated location: "blue cylinder", "purple sphere", "blue cube", "green cylinder", "brown cylinder". For translation, we prompt the model to translate the "blue cube" alone the line ranging from $(0.2, 0.8)$ to $(0.8, 0.2)$ in increments of $(0.05, 0.05)$.

4 Dataset Analysis

In this seciton, we provide some details about the composition and statistical makeup of our dataset. In Table 1, we show a statistical overview of the dataset, including total number of objects and categories, as well as the mean, median, and standard deviation of objects per category. We also visualize the distribution of objects across categories, as can be show in the histogram in Figure 4. Finally, we visualize the frequency of category names in the wordcloud in Figure 5.

References

[1] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. *arXiv preprint arXiv:2303.11328*, 2023.

[2] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

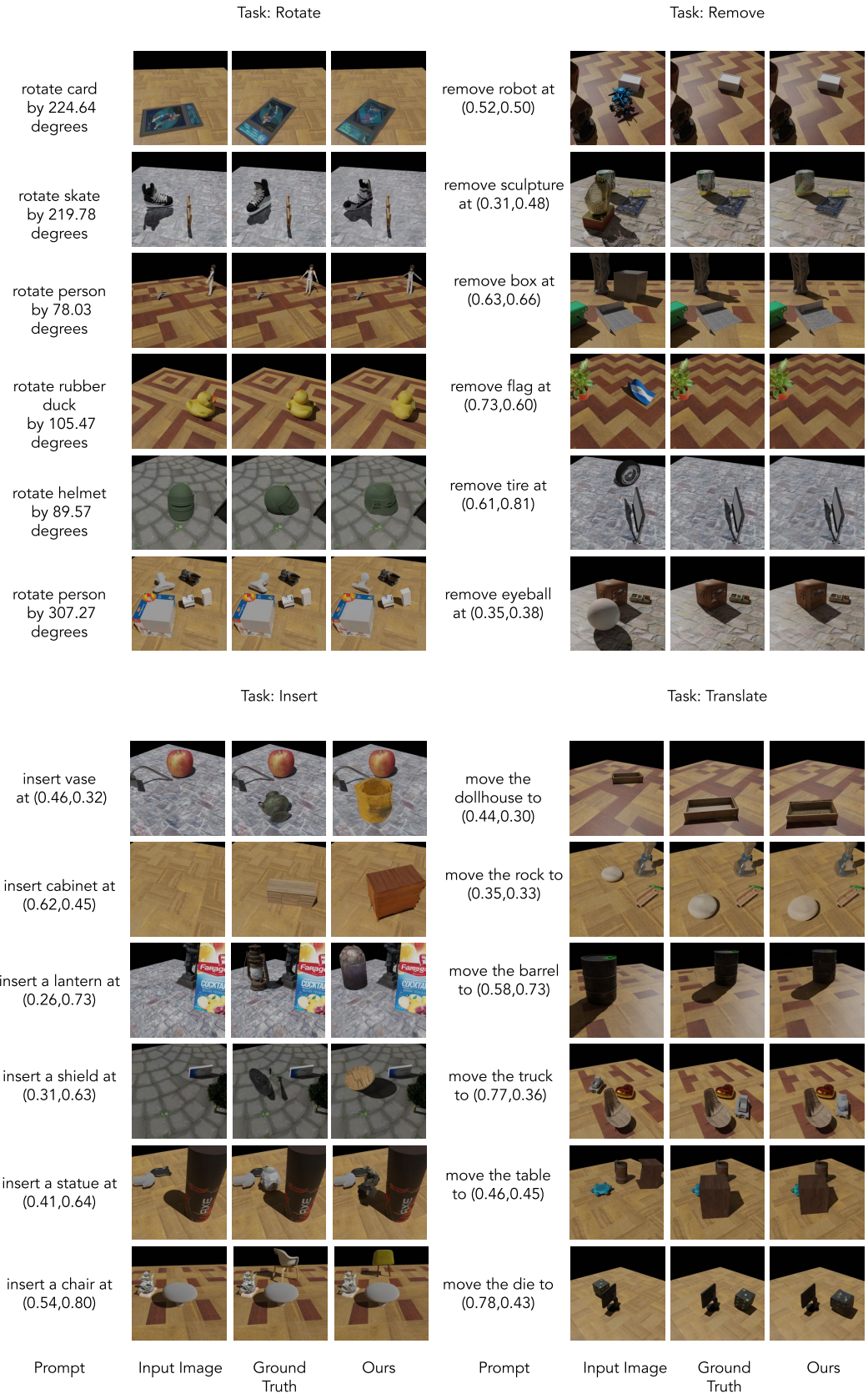

Figure 3: Generated examples by the 3DIT multitask model.

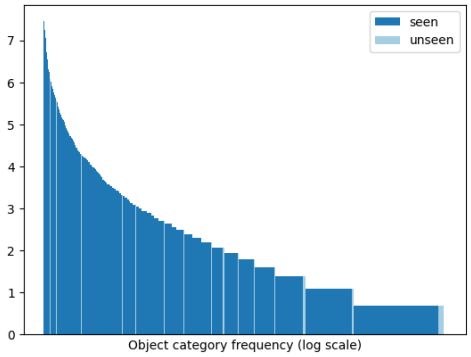

Figure 4: Object categories from seen and unseen splits sorted by frequency, in log scale.

Figure 5: A wordcloud visualizing the frequency of various object category names.