Vivek Madhavaram[1] Shivangana Rawat[2] Chaitanya Devaguptapu[2]
Charu Sharma[1] Manohar Kaul[2]
madhavaram.vardhan@research.iiit.ac.in
[1] Machine Learning Lab, IIIT Hyderabad, India [2] Fujitsu Research India

1. Implementation Details

Our method, FreeEdit occupies a memory of 11 GB from single GPU which will be used by Shap-E [3] and Grounding DINO [4] mentioned in main paper. As most of the modules are pre-trained, we do not train any model in this architecture. The entire procedure for insertion takes a minimum time of 3 minutes and a maximum of 5 minutes. It takes 1 minute for object synthesis and grounding, 1-3 minutes for latent diffusion of images to get the scale and a minute for other procedures. Replacement of objects is carried out in less than a minute. Advancements in 2D diffusion can even reduce this time as most of the time is spent in generating diffusion images. We implemented this method on an NVIDIA RTX A4000 server with RAM size 16 GB and CUDA version 12.0.

2. Replacement method

Inputs for replacing an object in a scene are a text prompt and a 3D scene. Text prompt is processed by GPT - 4 [1] as mentioned in section 3.2. The grounded object is extracted using [6], and the primary object is synthesized if necessary. Otherwise, user-provided mesh is used as the primary object. The generated mesh/input mesh should be scaled according to the dimensions of the grounded object. The grounded object need not always be parallel to the X, Y axes and can be oriented in random directions in the scene. If it is aligned at some angle, it is difficult to extract the exact dimensions using a bounding box. If dimensions are not accurate, the scale of the primary object will be inappropriate, and there will be a chance of intersection. To eliminate this, we find a vector along one edge of the grounding object in XY-Plane. Next, we find the angle it makes with the X–axis. Then, a point at the tail of this vector is considered, and the grounding object is rotated in XY-Plane at this point. The grounded object is rotated in such a way that it becomes parallel to the X-axis, and dimensions are extracted. The grounding object is deleted from the scene along with the other objects placed on it, and inpainting is performed at the deletion site, as mentioned in the main paper. Now, the primary object is scaled according to the grounding object dimensions and shifted to the location of the grounding object in the scene. Finally, the primary object is rotated by angle which initially, the grounding object is rotated, in reverse direction making it a good fit in the scene.

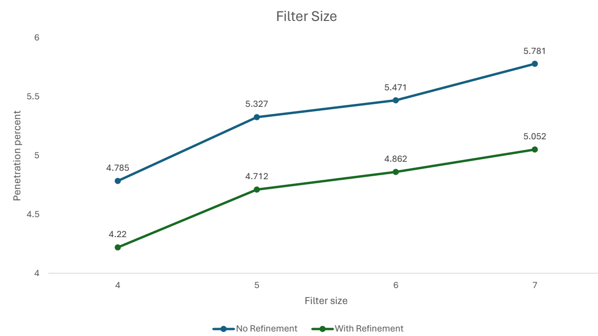


Figure 1. Plot of penetration percent with different filter sizes.

3. Filter size

In FreeEdit architecture, to find the best possible location on the grounding object, we create a voxel grid from surface vertices and perform a convolution operation using a filter of size $n \times n$. Here, n denotes the number of voxels along each axis. This filter is slithered over a voxel grid to determine the location suitable for placing an object on the surface. This n is also used to determine the voxel size, s. Voxel size is inversely proportional to filter size, and as filter size increases, voxel size decreases. With this, the number of cells in the voxel grid increases. As the number of cells increases, vertices become spread out, and the chances of cells being empty increase. As the number of empty cells increases, this might affect the average value in meeting the threshold criteria during convolution operation. In such cases, most of the operations do not meet threshold conditions and fail to find the best suitable location. We conducted experiments with different filter sizes and displayed the results in Table 1. We evaluated these parameters on all the test cases mentioned in the Experiments section using the penetration percent metric and displayed the average values for each parameter. Data sets used for

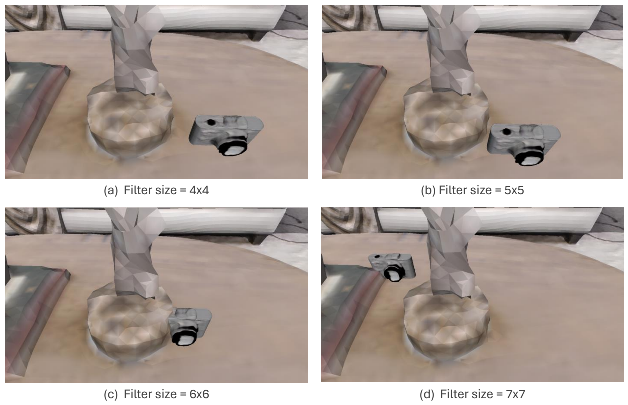


Figure 2. **Different filter sizes:** (a), (b), (c) and (d) Insertion of object using different filter sizes for prompt ”Insert camera on the table”.

evaluation are ScanNet [2] and Replica [5]. Same values are used to plot a linear graph in Figure 1. We observe that

Refinement	Voxels=4	Voxels=5	Voxels=6	Voxels=7
No	4.785	5.327	5.471	5.781
Yes	4.22	4.712	4.862	5.052

Table 1. Comparison of FreeEdit with different filter sizes

as the filter size increases, the penetration percent also increases. Also, the penetration percentage is less when automated refinement is applied compared to no refinement. This reveals that as the number of voxels increases, the intersection of the primary object with the scene increases.We also show an example of object insertion with different filter sizes in Figure 2. For this experiment, the primary object generated and its scale with respect to the grounding object is kept constant. From the figure, it is quite evident that the placement of the primary object is better with a small filter size compared to bigger numbers.

4. Role of threshold value

During convolution operation, the other important parameter is the threshold value. On applying convolution operation followed by average operation on the voxel grid, the output is determined by this threshold value. If the average of all cells within a filter is greater than this threshold, the output is 1, else 0. As the threshold value decreases, the strictness in eliminating intersections decreases. We conducted a few experiments with different threshold values and the results are displayed in Table 2. Values displayed in the table are the average value of the penetration percent metric of all test cases mentioned in the Experiment section in the main paper. A linear plot is displayed in Figure 3 with

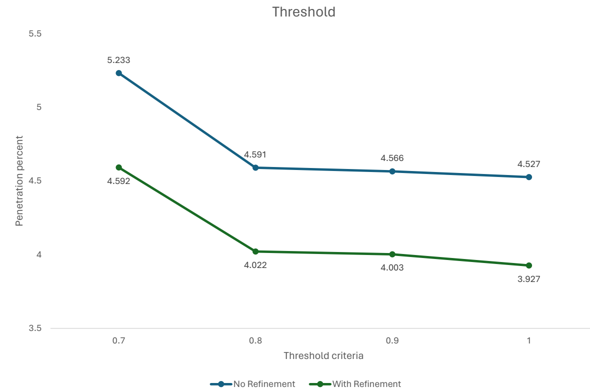


Figure 3. Plot of penetration percent with different threshold criteria.

Refinement	Threshold = 0.7	Threshold = 0.8	Threshold = 0.9	Threshold = 1
No	5.233	4.591	4.566	4.527
Yes	4.592	4.022	4.003	3.927

Table 2. Comparison of FreeEdit with different threshold values

values from the table.

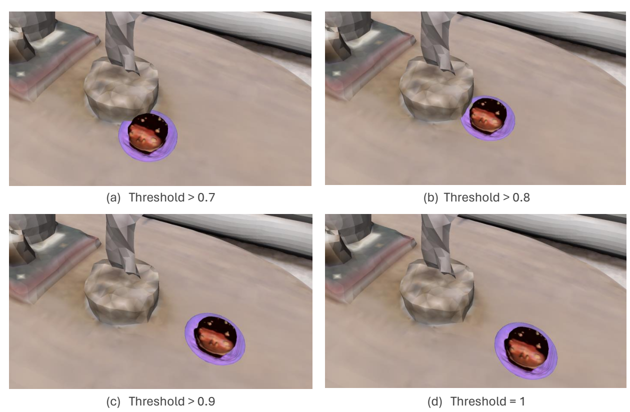


Figure 4. **Different thresholds:** (a), (b), (c) and (d) Insertion of object using different threshold criteria for prompt ”Insert dessert on the table”.

We observe that as the threshold value increases, the penetration percent decreases. This reveals that as the threshold increases, the strictness in eliminating the intersection of the primary object with the scene increases, reducing the penetration percent. Figure 4 showcases an experiment carried out with different threshold criteria, keeping the primary object, scale and filter size constant. We can infer from the figure that with lesser threshold value, the intersection is huge compared to higher values.

(a) min scale (0.15) (b) average scale (0.39) (c) max scale (0.66)

Figure 5. **Different scales:** Insertion of water bottle on platform with different scales (a)Minimum Scale (scale = 0.15), (b) Average scale (scale = 0.39) and (c) Maximum scale (scale = 0.66).

5. Scale

In FreeEdit, we considered multiple images to determine the scale of the primary object with respect to the grounding object. We calculated scales in all these images and considered the minimum value out of all generated scales. This is because the results are more realistic compared to average and max values. The max value can be 1, which tells that the width of the primary object is the same as the width of the grounding object. Consider an example of placing a candle on a table. In this case, utilizing the max value will generate a candle of a size table that doesn't seem accustomed. We cannot even consider the average value as it may be right skewed if most of the values in the generated group are close to the maximum scale. This skewness results in unnatural scales in some cases. One such example is visualized in Figure 5. This example depicts placing a water bottle on a platform. Considering the minimum scale, the object seems realistic compared to the average and maximum scale. Primary objects in average and maximum scale are of huge size and practically, water bottle of such dimensions do not exist.

6. Visualization

We present additional results of our method FreeEdit in this section other than those displayed in the main paper. A few examples of object insertion are shown in Figure 6, and a few examples of object replacement are displayed in Figure 7.

6.1. Insertion

Few examples on object insertion in a scene are shown in Figure 6,

6.2. Replacement

Few examples on object replacement in a scene are shown in Figure 7,

7. FreeEdit vs Baseline

In this section we show the how our proposed method, FreeEdit, works better compared with the baseline mentioned in the main paper in terms of iterative addition. For this comparison, let us consider a common prompt "I started decorating by placing the wooden tray in the center of the table. I carefully arranged the vase with dried flowers on the table. I also added the scented candles." containing multiple objects in this narration. Figure 8 displays the results of our proposed method and baseline. It is quite evident that when there are prior things on the grounding object, our method tries to insert new item at a vacant place whereas baseline ignores the prior conditions and inserts all the objects at single location.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 2

[3] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 1

[4] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1

[5] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica

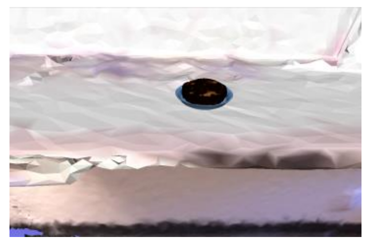
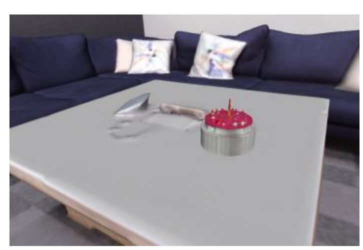
Input Prompt	LLM Response	Object Detection	Output
Suggest one thing to be placed on dining table to make my room look like Japanese. Make sure it is luminescent.	**You can place a luminescent sakura bonsai on your dining table to give your room a Japanese touch.**	Primary object: **Luminescent Sakura bonsai** Grounding object: ***Table***	
Which is the best suitable object to put flower vase among table, stool, platform?	**The best suitable object to put a flower vase among a table, stool, platform is the table.**	Primary object: ***Flower Vase*** Grounding object: ***Table***	
Recommend a yummy dessert to place on the counter.	**Sure, you can place a Chocolate Lava Cake on the counter.**	Primary object: ***Chocolate Lava Cake*** Grounding object: ***Counter***	
I relax by drinking the soda placed on the stool. So put it there.	**The soda needs to be placed on the stool for relaxation.**	Primary object: ***Soda*** Grounding object: ***Stool***	
You can drop your bag on the bed till we finish our work	**The person is instructed to place their bag on the bed until work is completed.**	Primary object: ***Bag*** Grounding object: ***Bed***	
What accessory can I place on the kitchen counter to enhance its aesthetic?	**You can place a fruit bowl on the kitchen counter to enhance its aesthetic.**	Primary object: ***Fruit Bowl*** Grounding object: ***Kitchen Counter***	
Its getting late for the birthday party. Please arrange a cake on the table.	**The task is to arrange a cake on the table for a birthday party.**	Primary object: ***Cake*** Grounding object: ***Table***	

Figure 6. Examples of object insertion in the scene.

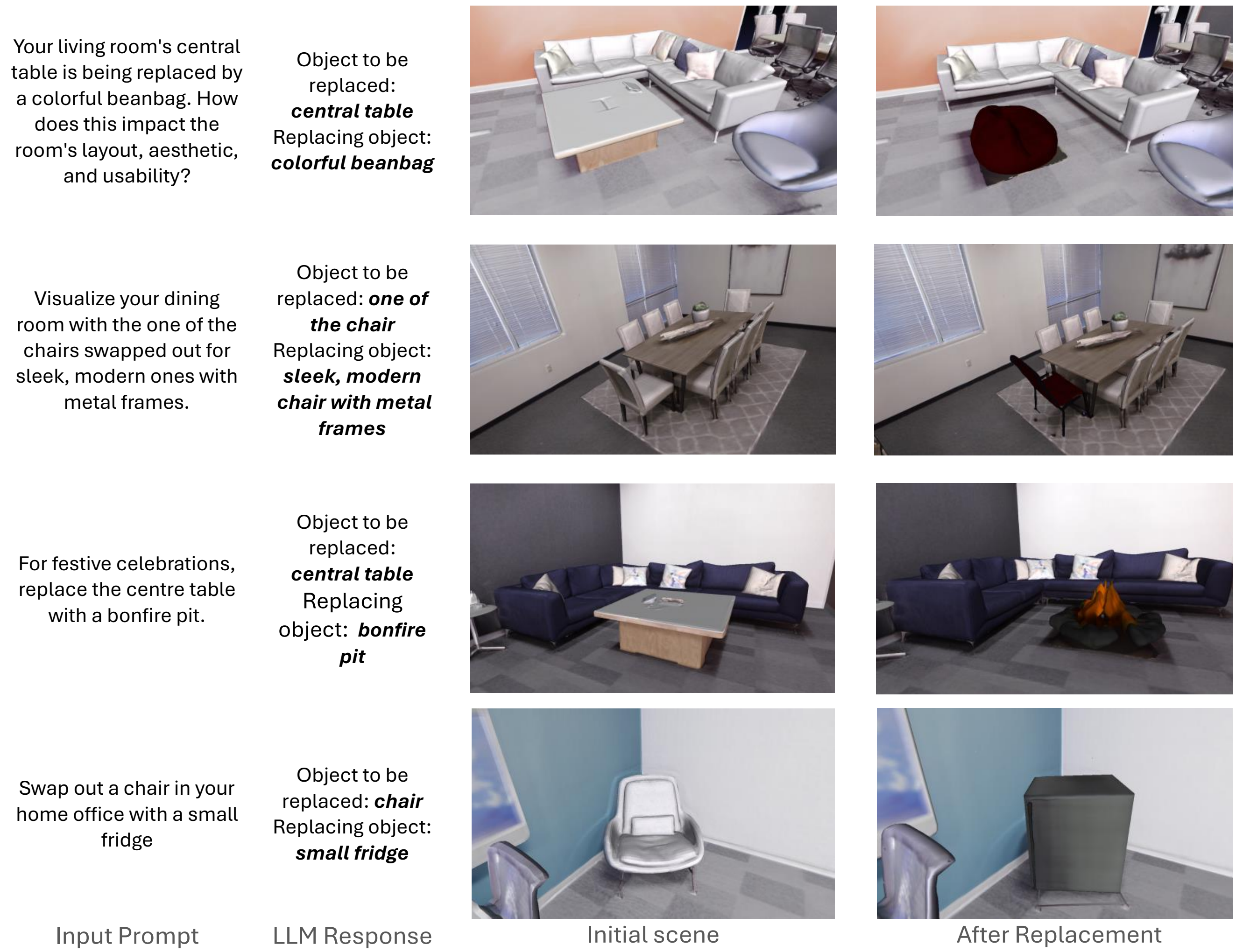


Figure 7. Examples of object replacement in the scene.

Figure 8. Traditional vs Proposed methods

dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2

[6] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1