

VoxHammer: Training-Free Precise and Coherent 3D Editing in Native 3D Space

Supplementary Material

1. Text-Condition 3D Editing

Benefiting from the versatility of TRELIS [1], our framework also supports text-condition 3D editing by injecting textual conditions into the inversion and denoising stages of the base model for masked assets, as illustrated in Fig. 2. Leveraging this capability, we evaluate *VoxHammer* on text-condition 3D editing tasks, where it achieves competitive performance in preserving unedited regions and maintaining overall 3D quality, as presented in Fig. 1. However, condition alignment is not always reliable, as the model may deviate from textual instructions. As shown in Tab. 1, this underscores the need to further enhance the fidelity of text-conditioned guidance.

2. Explanation of Evaluation Metrics

In terms of evaluating unedited region preservation, Chamfer Distance assesses the geometry consistency, while masked PSNR, SSIM and LPIPS of rendered multi-view images evaluate the consistency of structures and appearance. In terms of evaluating editing quality, FID assesses the overall visual similarity between the edited results and the original object. FVD evaluates the temporal continuity and stability across multi-view images. In terms of edit controllability, the text-asset alignment score from CLIP-T measures the similarity between the editing results and the editing text, while DINO-I measures the similarity between the editing results and the original object. Since our task focuses on 3D local editing, DINO-I can reflect the accuracy of the edits to some extent. Overall, these metrics provide a comprehensive quantitative evaluation of unedited region preservation, overall editing quality, and editing accuracy from different perspectives, collectively reflecting the overall performance of the 3D editing method.

3. More Results

More results of image-condition 3D editing are shown in Fig. 3, which demonstrates the ability to achieve precise and coherent 3D editing.

4. Limitation

Although *VoxHammer* preserves unedited regions and maintains overall 3D quality, several limitations remain. First, textual alignment is not yet optimal, partly due to the scarcity of large-scale captioned 3D datasets, making text condition less robust than image-based guidance. Second, editing fidelity is bounded by the resolution of the TRELIS [1] backbone, limiting precision for high-resolution as-

sets. Finally, our pipeline comprises of 3D encoding, inversion, denoising and decoding. Due to the time-consuming rendering phase in the 3D encoding stage (> 1 min), *VoxHammer* takes about 2 minutes to edit one 3D asset, indicating room for efficiency improvements toward interactive use.

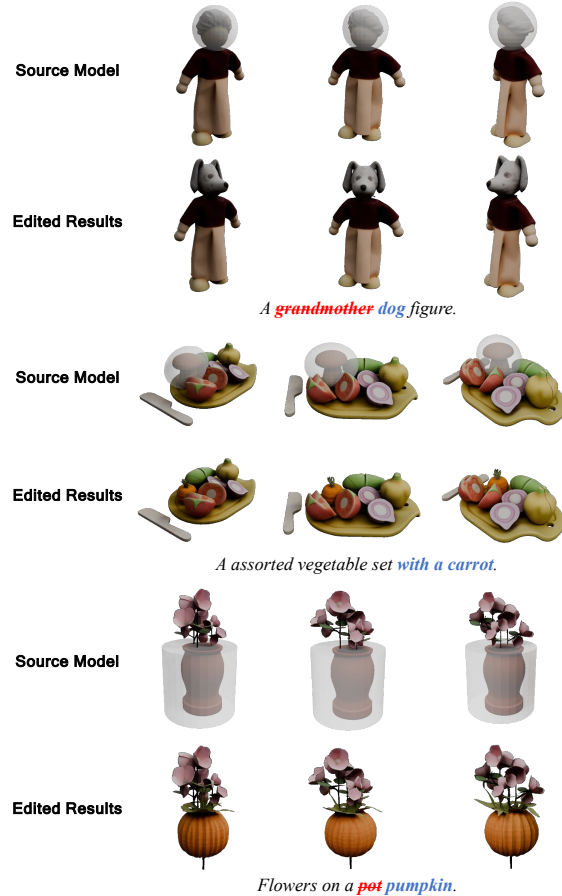


Figure 1. Visualization results of text-condition 3D editing.

References

- [1] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation, 2025. 1

Table 1. Quantitative comparison on text-condition and image-condition 3D editing.

Method	Unedited Region Preservation				Overall 3D Quality		Condition Alignment
	CD, ↓	PSNR (M) ↑	SSIM (M) ↑	LPIPS (M) ↓	FID ↓	FVD ↓	CLIP-T ↑
Text-condition 3D editing	0.010	38.61	0.992	0.024	25.93	150.4	0.277
Image-condition 3D editing	0.012	41.68	0.994	0.027	23.05	187.8	0.287

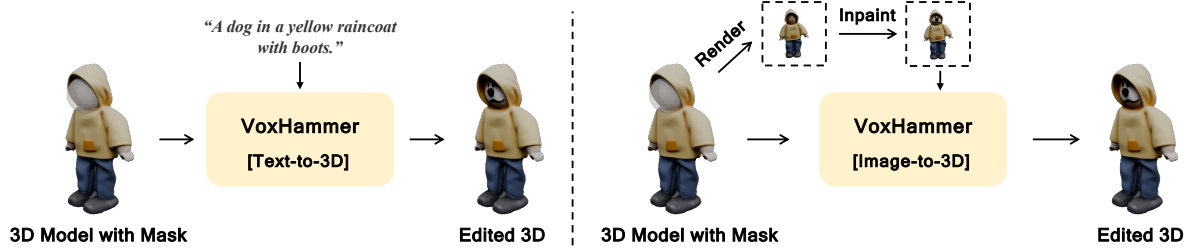


Figure 2. Pipeline of text-condition (left) and image-condition (right) 3D editing.

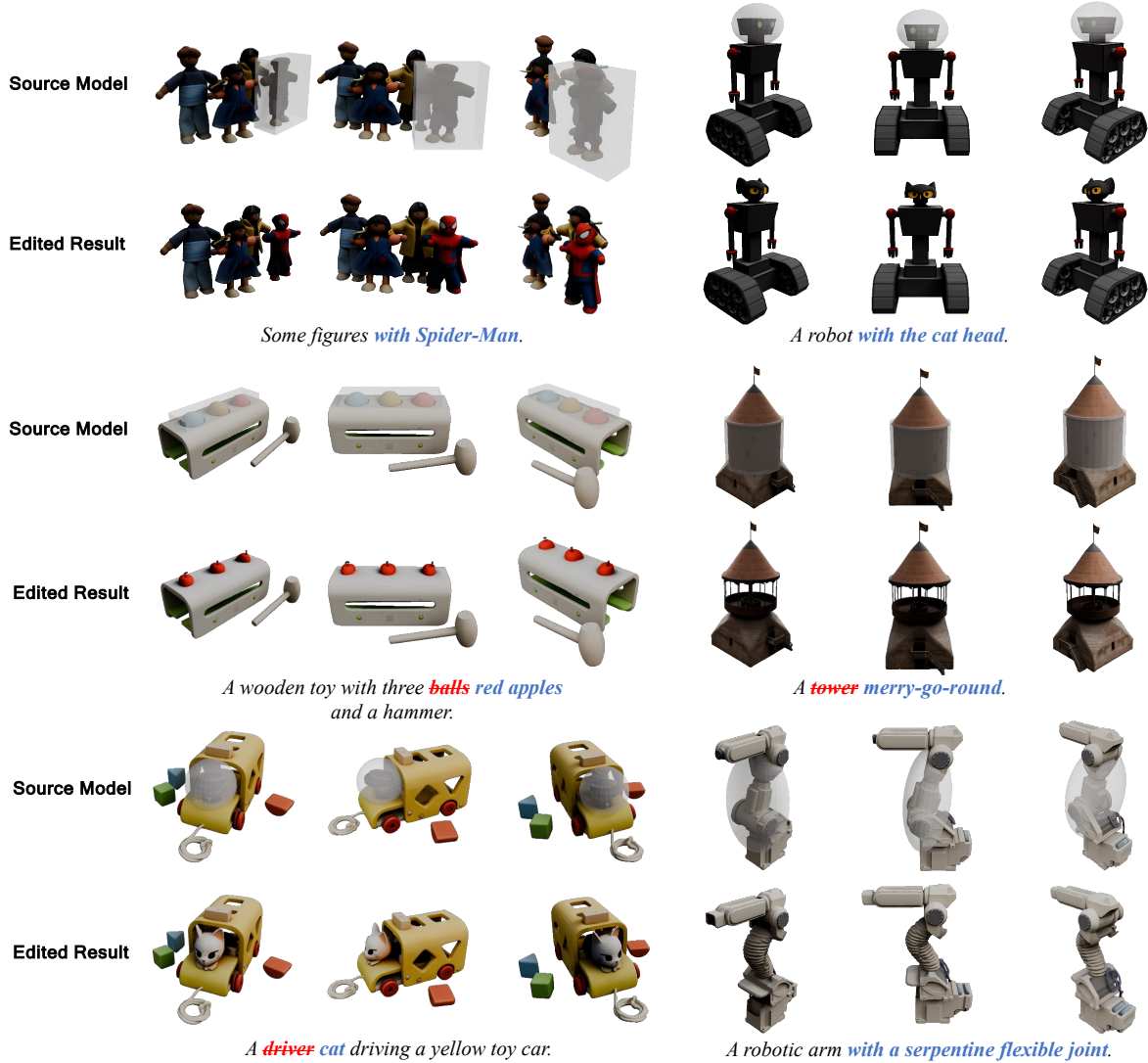


Figure 3. More visualization results of image-condition 3D editing.