WGS: Watertight Geometry Standardization for Scalable 3D Generation

Supplementary Material

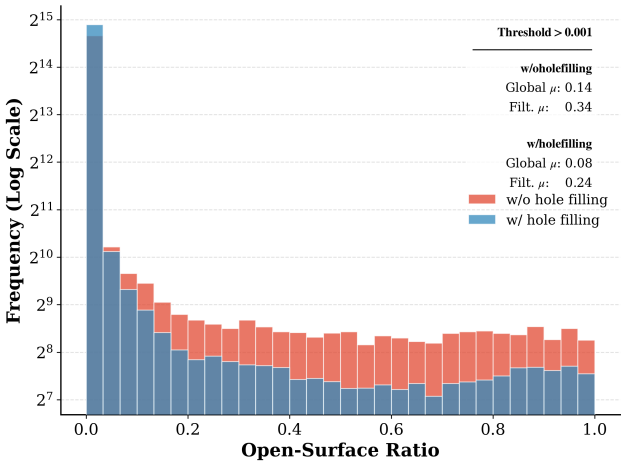

Figure 6. Distribution shifts of open-surface ratio achieved by our WGS pipeline.

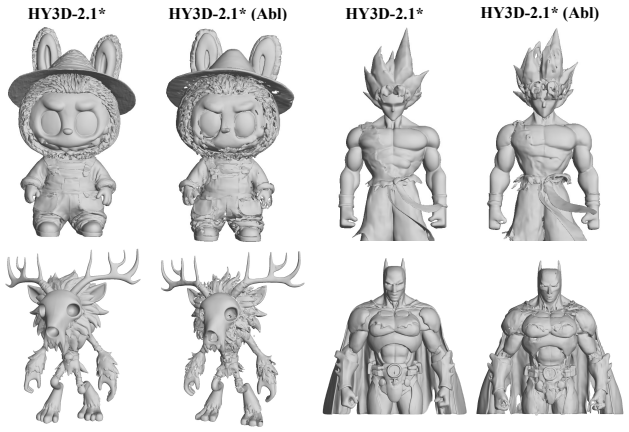

Figure 7. Comparison of generation results from HY3D-2.1 DiT fine-tuned on different datasets. Abl) indicates the models fine-tuned on our processed data without hole filling.

A. Implementation Details

A.1. 3D VAEs

Hunyuan3D 2.1 VAE. We follow the official sampling strategy: uniformly sampled surface and sharp-edge points with normals as inputs, and SDF samples drawn near the surface and uniformly in free space for supervision. We employ a two-stage fine-tuning strategy, first at 1024^3 and subsequently refining at 1536^3 to progressively enhance geometric fidelity. We perform 30k fine-tuning steps on 8 GPUs with a global batch size of 64, 4096 latent tokens per shape, and a learning rate of $1e-5$. During evaluation, meshes are extracted with FlashVDM [18] to accelerate inference.

Sparse-voxel-based VAEs. For sparse-voxel-based VAEs, we extract active voxels and their corresponding SDF values for both input and supervision. For D3D-S2-VAE, we additionally include an auxiliary sharp-edge mask following its original training setting. We fine-tune D3D-S2-VAE for 30k steps on 8 GPUs with a global batch size of 8, using a decoding chunk size of 4 and a learning rate of $5e-5$. For SparC3D-VAE, we train the model from scratch for 40k steps using a global batch size of 8, a decoding chunk size of 2, a compression ratio of $\times 8$, and a learning rate of $5e-5$.

A.2. Image-to-3D DiTs

HY3D-DiT is fine-tuned with 4096 latent tokens per shape obtained from our fine-tuned HY3D-VAE. We train for 25k steps on 16 GPUs with a global batch size of 64.

For D3D-S2, we fine-tune only the SS-DiT-1024 module responsible for generating 128-resolution sparse latents, as this stage dominates high-frequency geometric detail. The first-stage DiT and VAE from the original pipeline are reused during evaluation. We fine-tune their DiT for 30k steps with global batch size of 16.

B. Additional Ablation Studies

B.1. Morphological Closing

To further assess the effectiveness of our hole-filling mechanism, we conduct two complementary ablation experiments.

Dataset-level geometric consistency. We measure the distribution of the open-surface ratio (defined in Eq. (5)) across the randomly sampled subset from Objaverse[8] dataset, comparing results with and without the proposed morphological closing stage. We also evaluate the mean of this ratio over entire dataset and a filtered dataset whose open surface ratio is larger than a threshold. As shown in Fig. 6, our method substantially reshapes the distribution, dramatically reducing the mean open-surface ratio and yielding topologically more consistent 3D dataset.

Impact on 3D generative modeling. We fine-tune the Hunyuan3D-2.1 DiT on two versions of the dataset—one processed with hole filling and the other left unchanged. As illustrated in Fig. 7, the model trained without hole filling exhibits a failure mode: it generates intended double-layered shells, even when the input image clearly depicts a solid object. This indicates that the model internalizes and reproduces structural inconsistencies present in the training data.

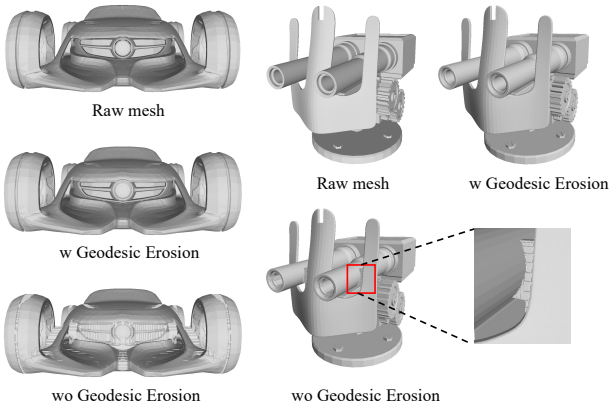

Figure 8. Ablation study of the geodesic erosion stage.

In contrast, fine-tuning on our watertight dataset eliminates this artifact, producing clean, single-layer surfaces that better align with the expected solid geometry.

B.2. Geodesic Erosion.

Figure 8 illustrates the effectiveness of the geodesic erosion stage. Without this mechanism, WGS is able to seal holes but inevitably loses geometric detail, introducing noticeable artifacts in regions of high curvature. With geodesic erosion, these hallucinated voxels are retracted toward the original surface, allowing fine structures to be faithfully restored.

C. More Comparisons of Watertight Remeshing

Figures 9 and 10 present visual comparisons on several challenging 3D models. As shown, visibility-based methods typically produce consistent solid geometry but introduce noisy artifacts in regions with ambiguous visibility. Flood-fill–based approaches yield smooth surfaces, yet often generate double-layered thin shells when the input mesh is non-watertight. ManifoldPlus frequently collapses into zero-volume structures when handling non-watertight inputs, while Dora tends to miss components on highly complex objects. In contrast, our WGS pipeline produces consistent solid geometry while maintaining clean, well-behaved surfaces, and WGS-2048 further delivers near lossless reconstruction quality.

D. More Comparisons of 3D generation

For 3D VAEs, Fig. 11 demonstrates that fine-tuning on our high-resolution 3D data consistently enhances the reconstruction of fine geometric structures. For Image-to-3D DiTs, the generation results in Figs. 12 and 13 further show that our fine-tuned models produce more detailed geometry and significantly sharper structural features.

E. Limitations and Future Directions

Although our WGS pipeline achieves statistically consistent and geometrically accurate state-of-the-art watertight remeshing results, several limitations remain. When applied to 3D VAEs, we observe that VecSet-based VAEs struggle to recover high-fidelity details—a limitation primarily rooted in the representational capacity of current VecSet formulations. Sparse-voxel–based VAEs, on the other hand, can reconstruct substantially finer structures, but their computational cost is significantly higher: the number of latent tokens is nearly an order of magnitude larger than that of VecSet-based models, resulting in increased memory footprint and slower training and inference.

These observations suggest an important direction for future research: the development of more expressive yet computationally efficient 3D representations. Ideally, such representations should preserve the full geometric fidelity of high-resolution meshes while remaining lightweight enough to support large-scale generative modeling. Exploring hybrid representations, adaptive resolution schemes, or compact latent encoders that bridge the gap between sparse voxels and token-efficient structures could provide a promising path toward scalable and high-quality 3D generation.

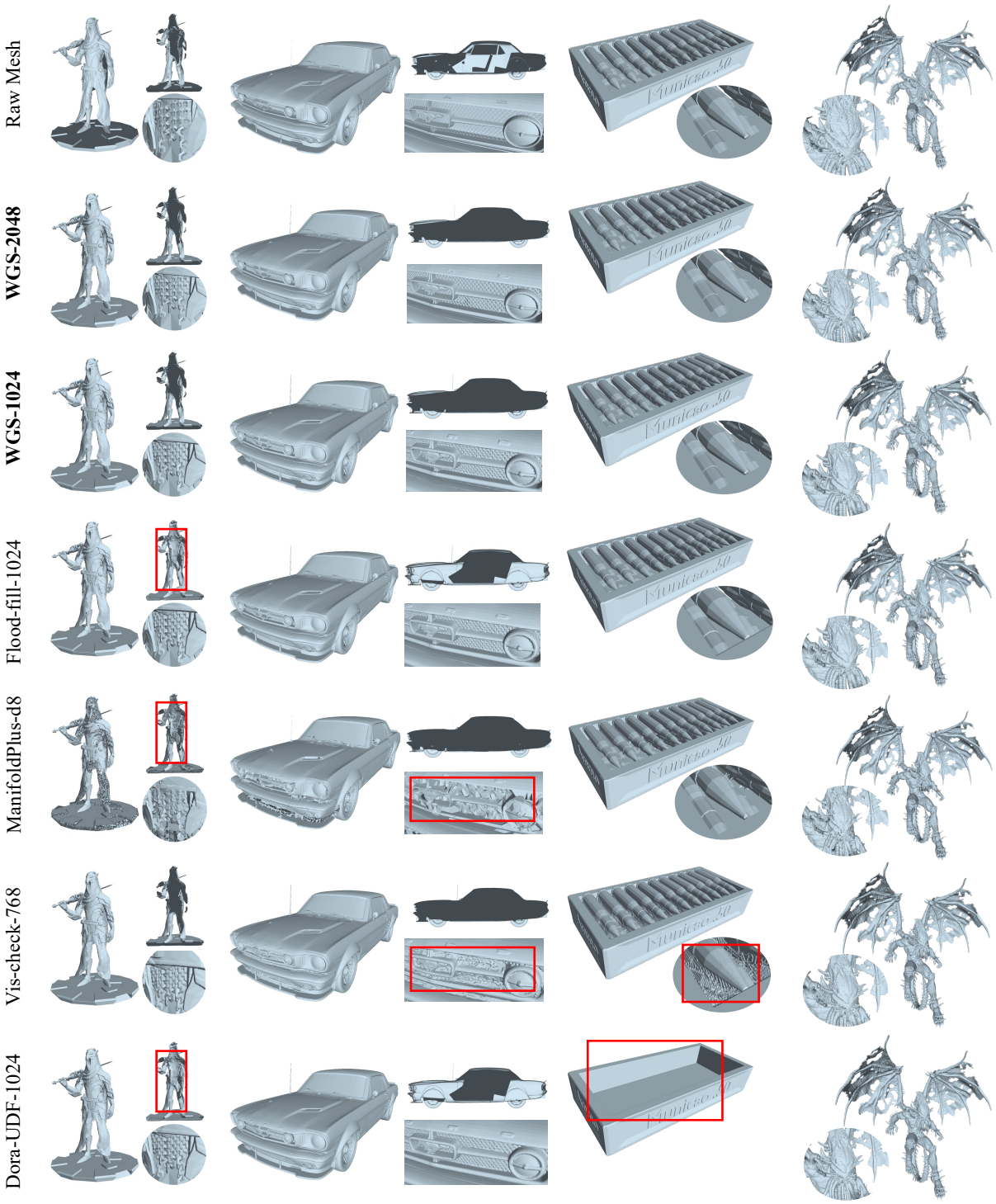

Figure 9. Extended Qualitative Results on Watertight Remeshing. *Best viewed with zoom-in*.

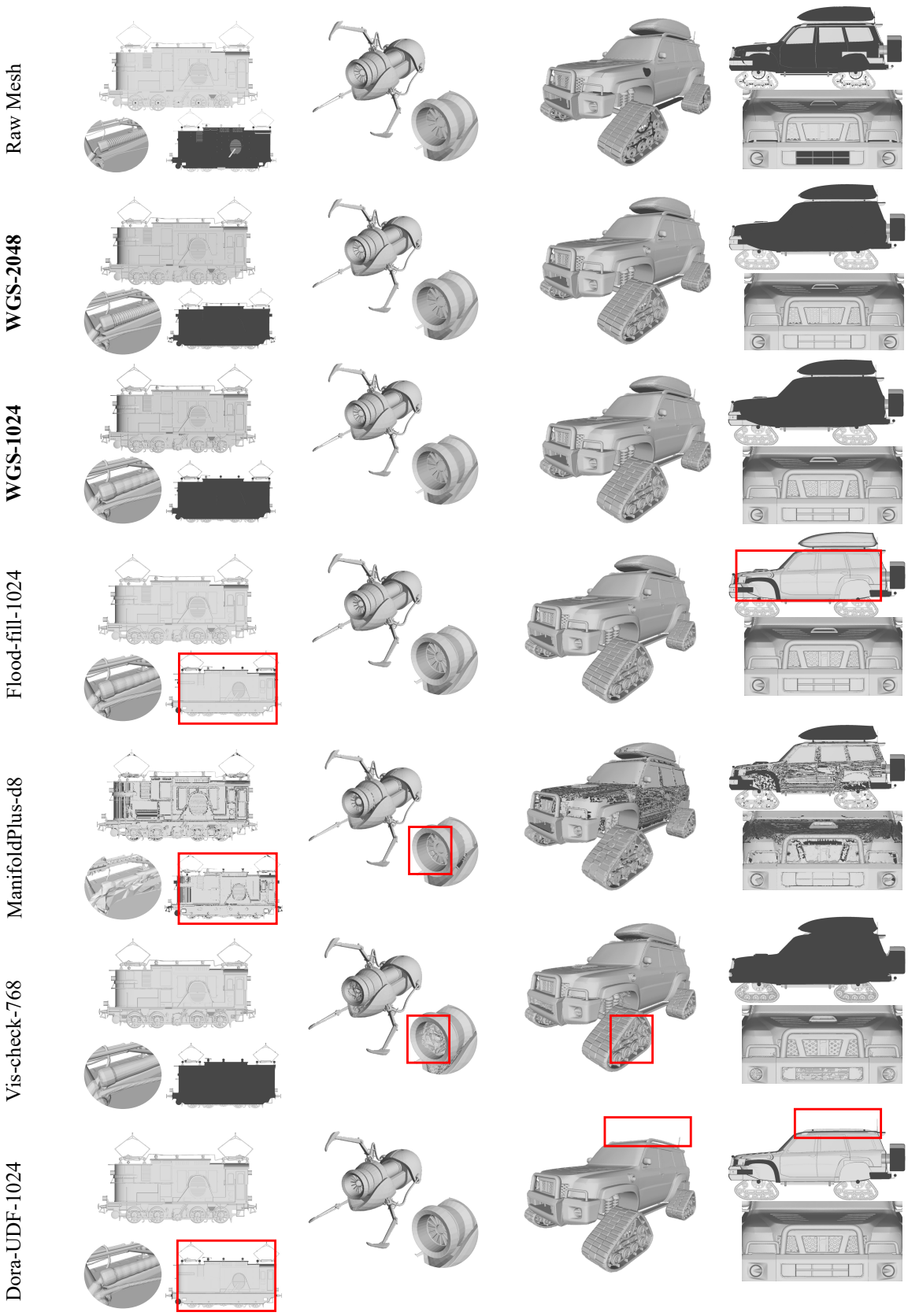

Figure 10. Extended Qualitative Results on Watertight Remeshing. *Best viewed with zoom-in*.

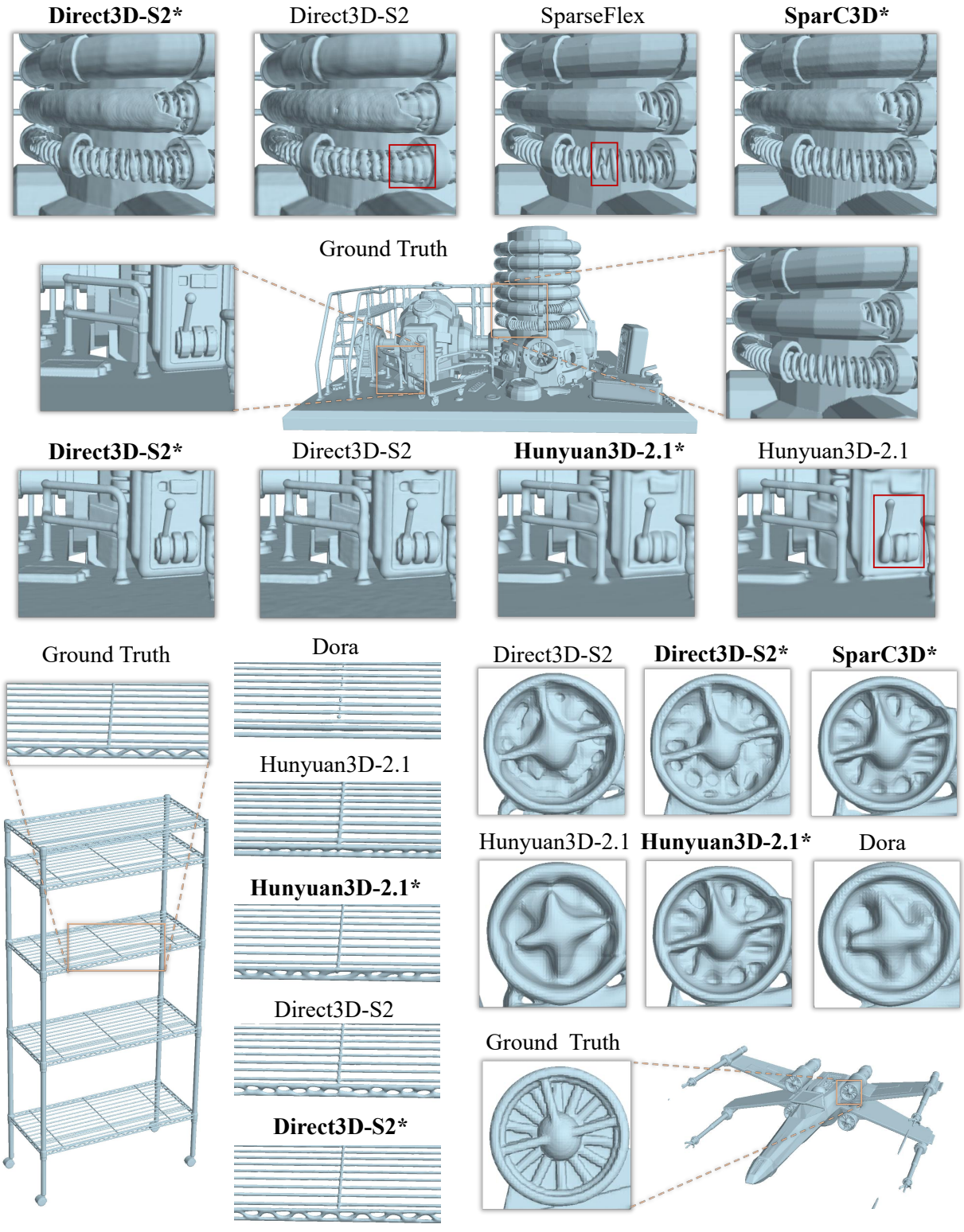

Figure 11. Extended Qualitative Results on VAE Reconstructions. *Best viewed with zoom-in*.

Figure 12. Extended Qualitative Results on Image-to-3D Generation. *Best viewed with zoom-in.*

Figure 13. Extended Qualitative Results on Image-to-3D Generation. *Best viewed with zoom-in*.