Direct3D-S2: Gigascale 3D Generation Made Easy with Spatial Sparse Attention

Shuang Wu[1,2]*, Youtian Lin[1,2]*, Feihu Zhang[2], Yifei Zeng[1,2], Yikang Yang[1],
Yajie Bao[2], Jiachen Qian[2], Siyu Zhu[3], Xun Cao[1], Philip Torr[4], Yao Yao[1]✉

[1]Nanjing University [2]DreamTech [3]Fudan University [4]University of Oxford

Motivation

- Image and video generation models employ **symmetric** VAE structure:

  image / video -> latent representation -> image / video

  Current 3D generation models typically use **asymmetric** VAE structure:

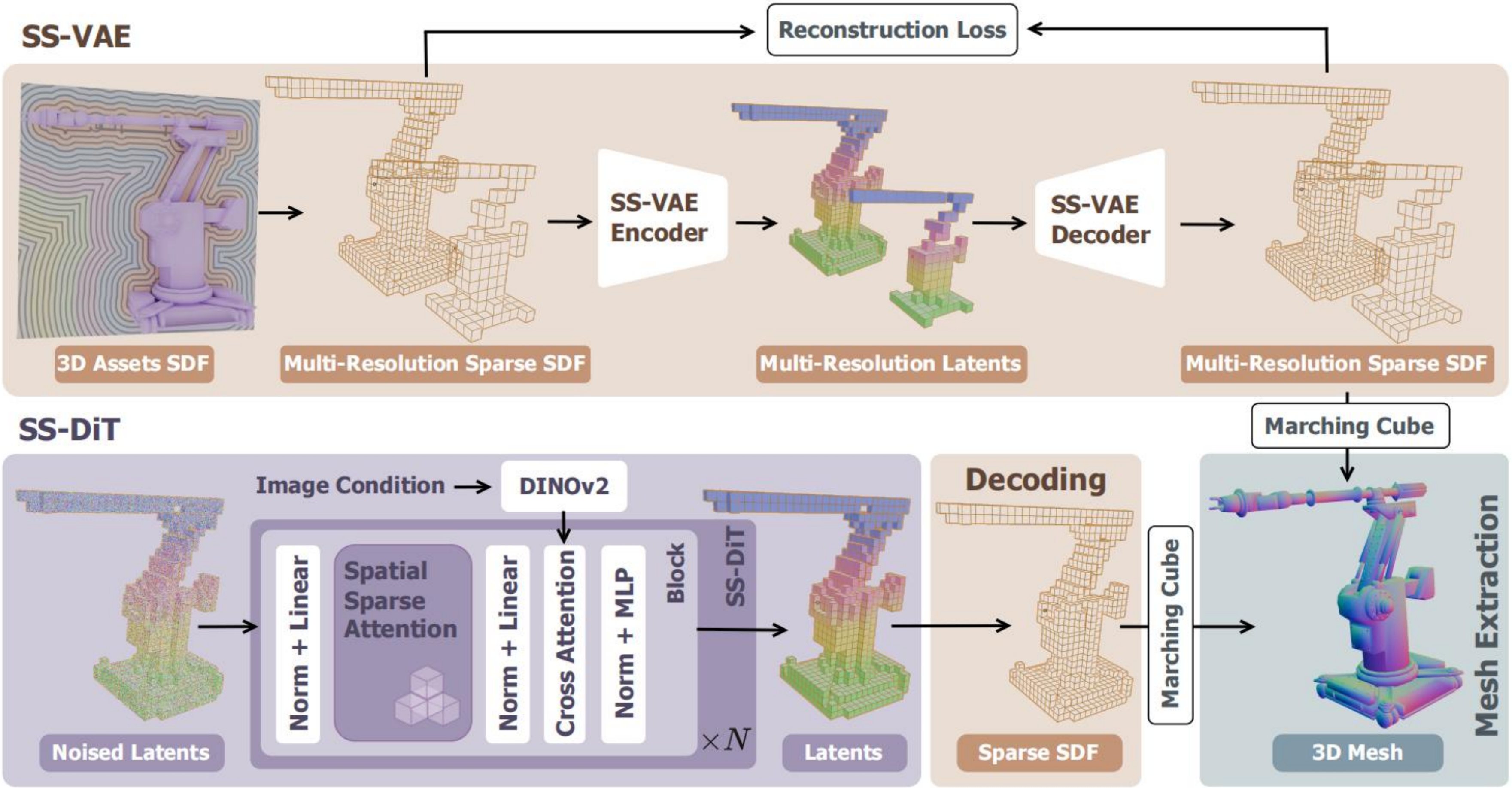
  point clouds -> latent representation -> SDF (3Dshape2Vecset)

  multi-view images -> latent representation -> rgb / normal / depth (TRELLIS)

  Inefficient or the reconstruction loss is large

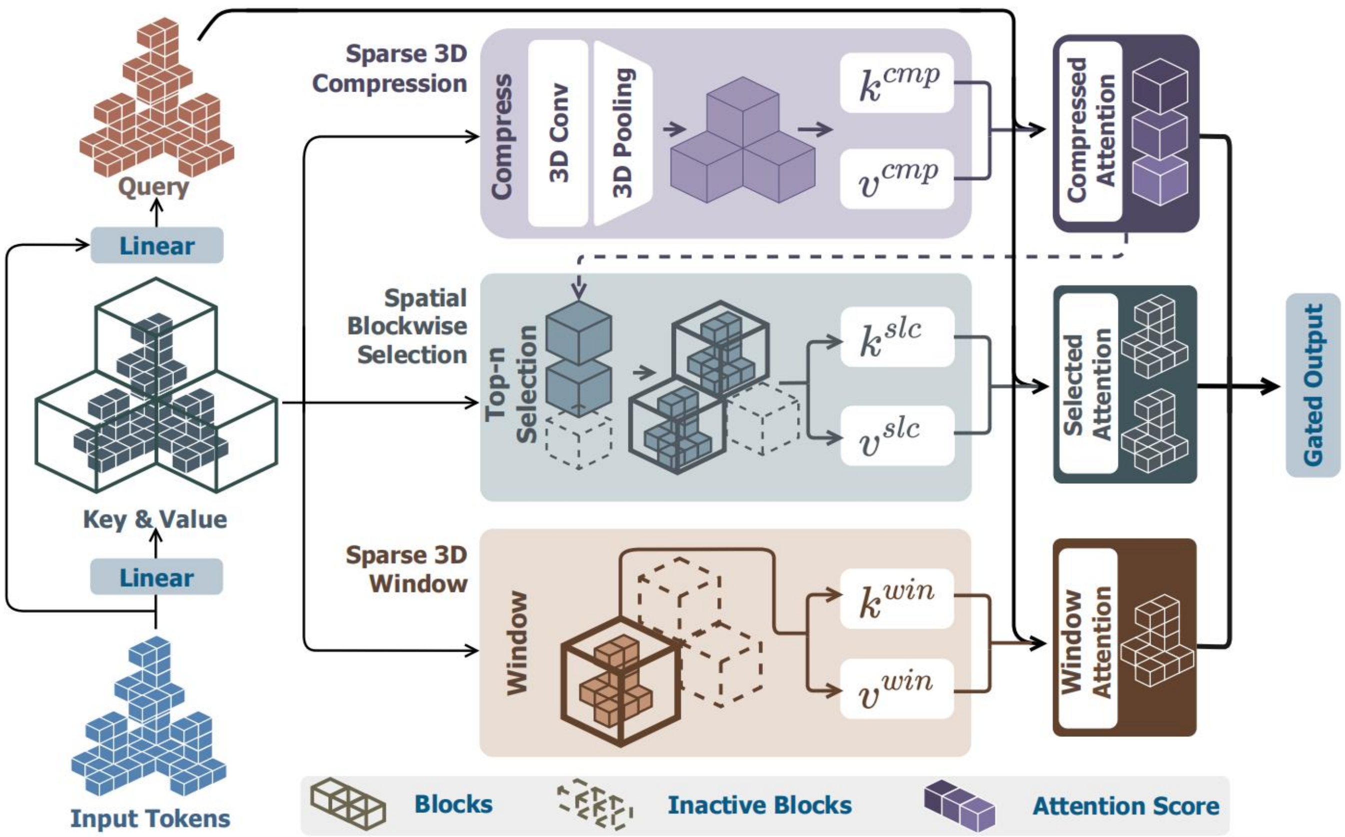
- Current 3D generation models are difficult to **scale to high resolution** as the quadratic cost of full attention in DiT

The Pipeline of Direct3D-S2



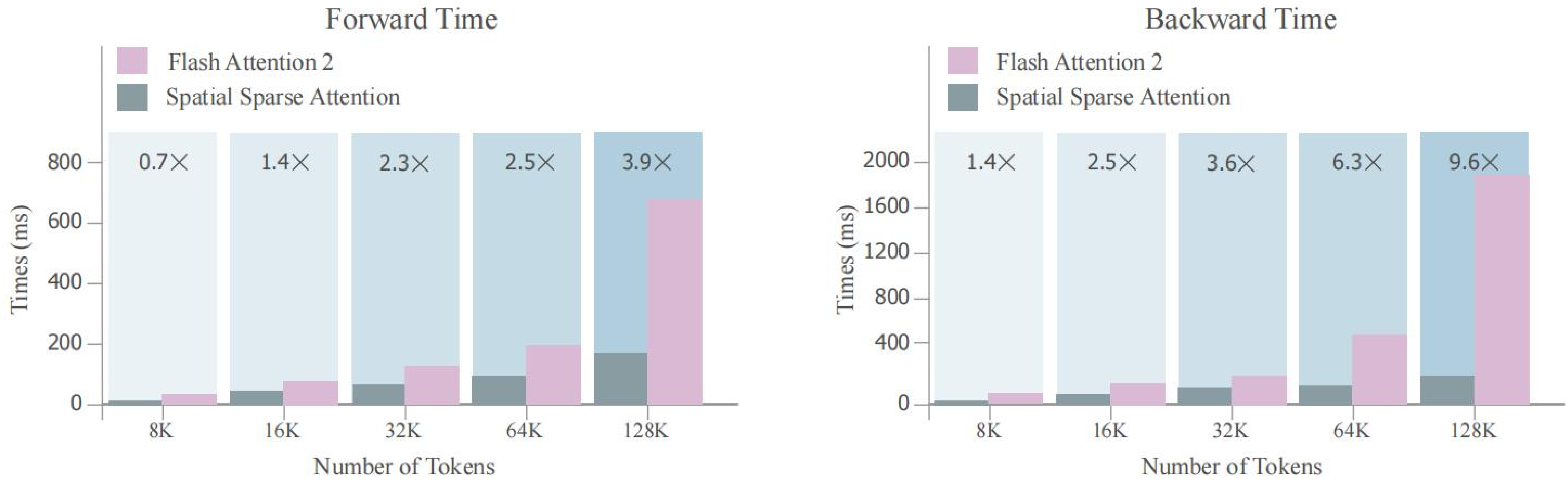
- A fully end-to-end sparse SDF VAE, which employs a **symmetric encoder-decoder network** to efficiently encode high-resolution sparse SDF volumes into sparse latent representations **z**
- An image-conditioned diffusion transformer (SS-DiT) based on **z**, with a novel **Spatial Sparse Attention (SSA)** mechanism that significantly improves the training and inference efficiency.

Spatial Sparse Attention



- Blockwise selection attention, motivated by Deepseek's Native Sparse Attention (NSA)
- **Sparse 3D Compression**: Employ sparse Conv3D to compress the block into a token, and compute attention between query tokens and compressed key/value tokens
- **Spatial Blockwise Selection**: For each query token, select the top-k blocks based on the attention score, and compute attention with all tokens contained in those blocks
- **Sparse 3D Window**: Compute attention between the query token and all local tokens within its window
- The final output of SSA are aggregated from the three modules using predicted gate scores

Spatial Sparse Attention



Comparison of the forward and backward time of SSA and FlashAttention-2

Image-to-3D Results

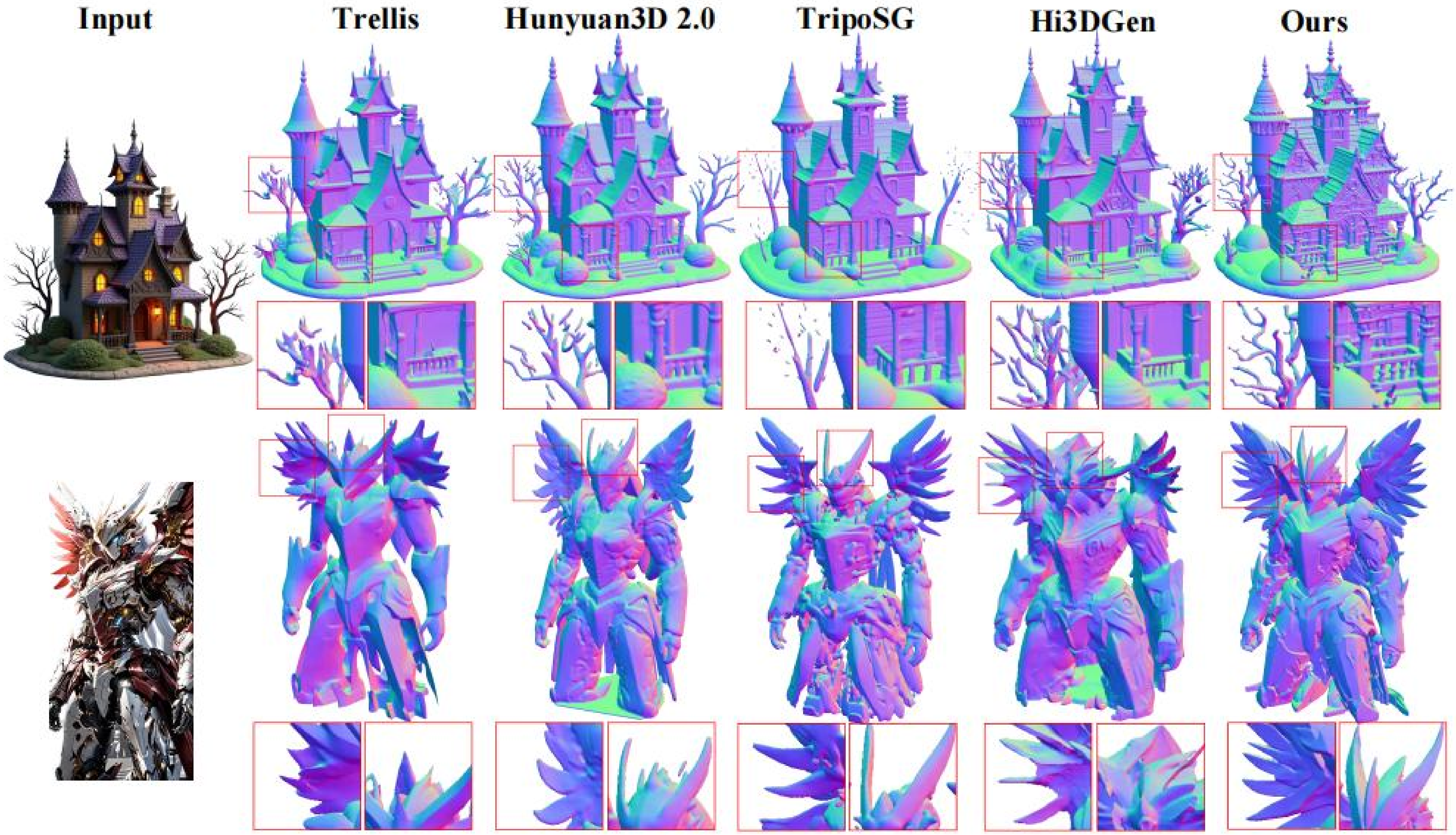
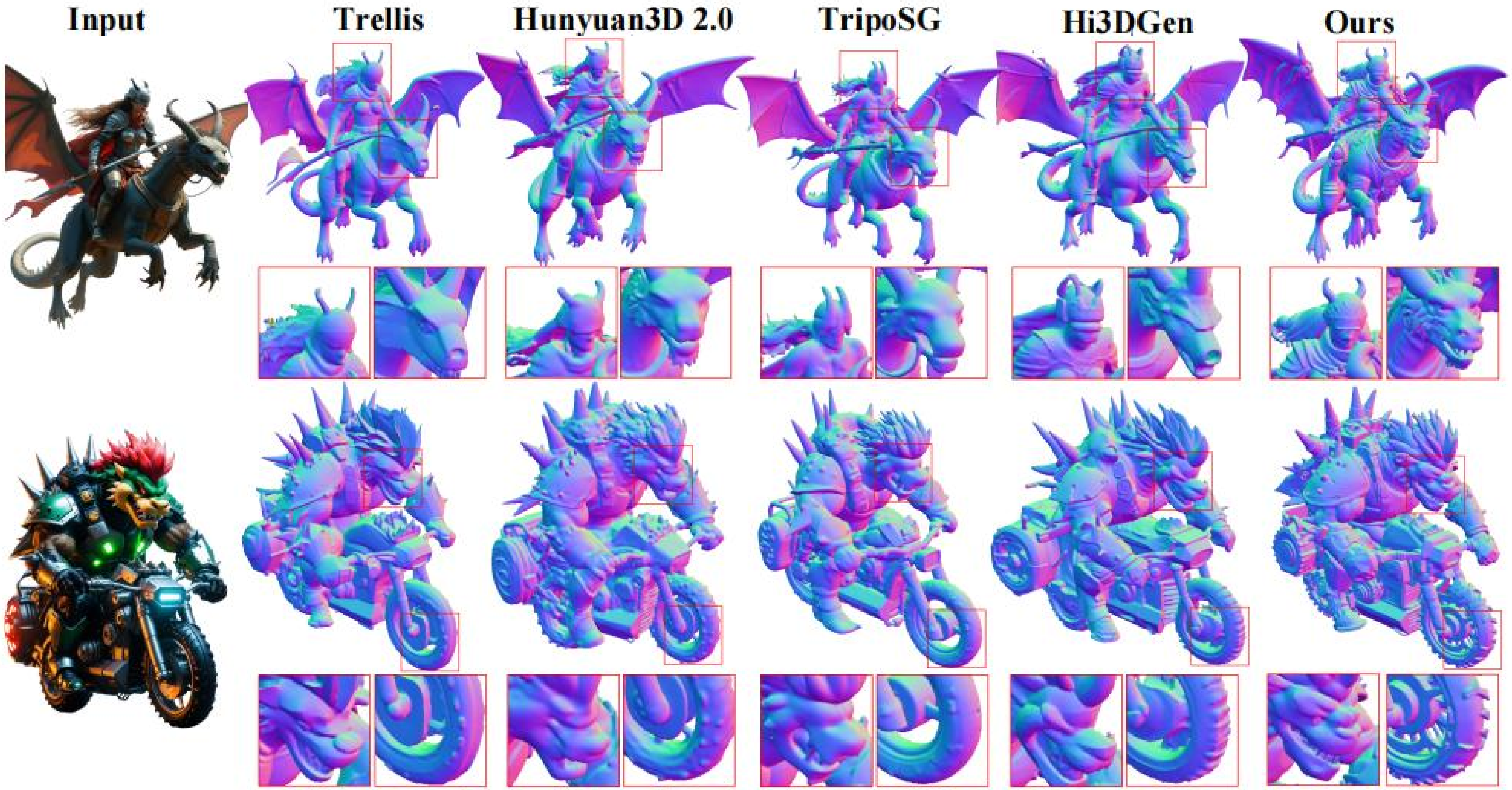
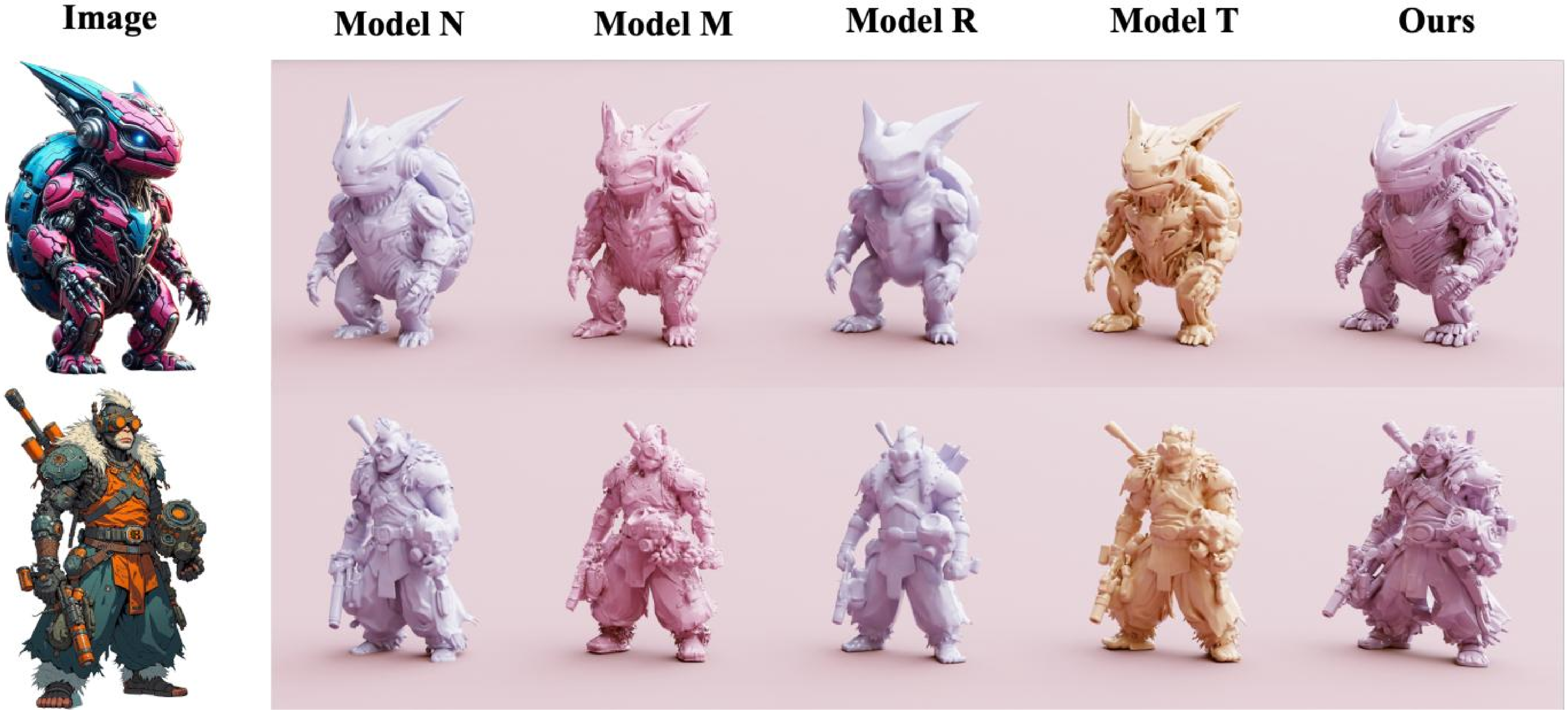


Image-to-3D Results



Comparison with Closed-Source Models



Direct3D-S2: Gigascale 3D Generation Made Easy with Spatial Sparse Attention

Shuang Wu[1,2]*, Youtian Lin[1,2]*, Feihu Zhang[2], Yifei Zeng[1,2], Yikang Yang[1], Yajie Bao[2], Jiachen Qian[2], Siyu Zhu[3], Xun Cao[1], Philip Torr[4], Yao Yao[1]✉

[1]Nanjing University [2]DreamTech [3]Fudan University [4]University of Oxford

Thanks!

https://www.neural4d.com/research-page/direct3d-s2/index.html