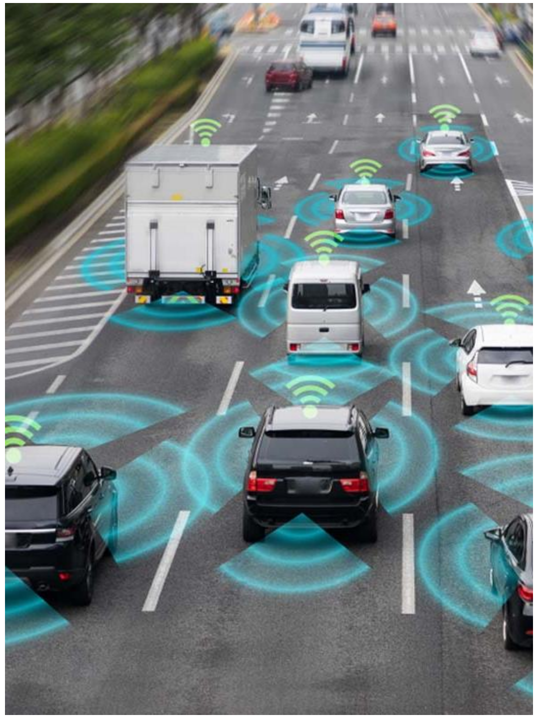
GaussianCube: A Structured and Explicit Radiance Representation for 3D Generative Modeling

Bowen Zhang, Yiji Cheng, Jiaolong Yang, Chunyu Wang,
Feng Zhao, Yansong Tang, Dong Chen, Baining Guo

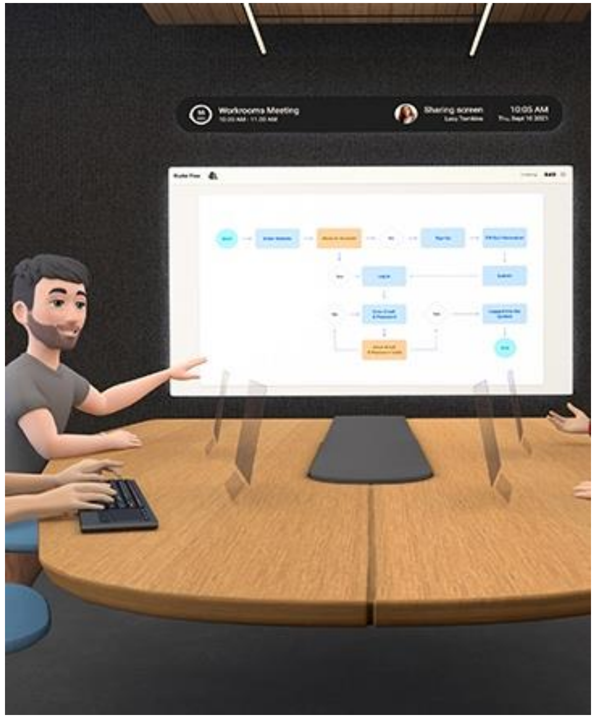
3D Content Creation

Gaming

Film

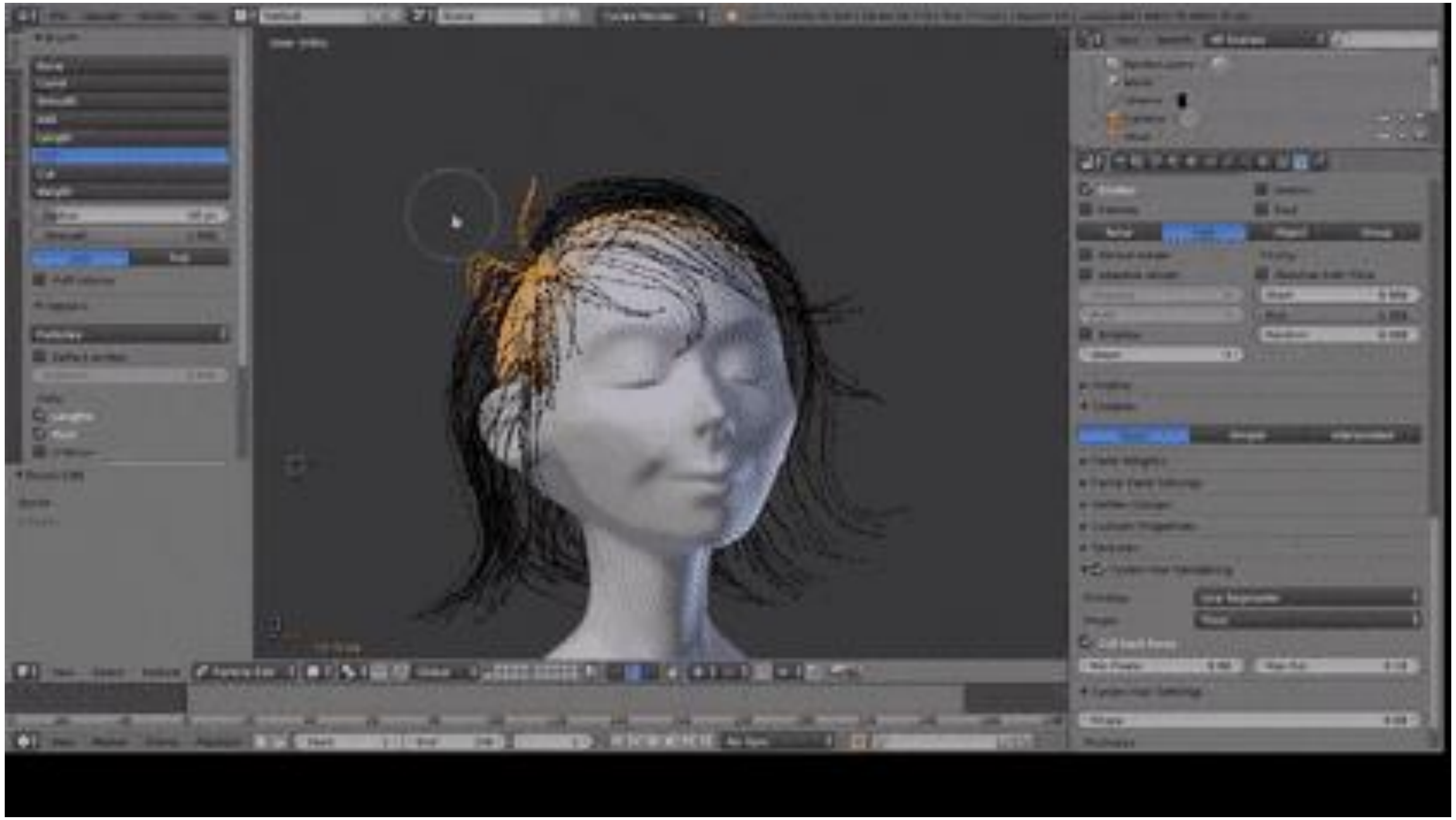


Autonomous Driving



Metaverse

Traditional 3D Content Creation is Expensive

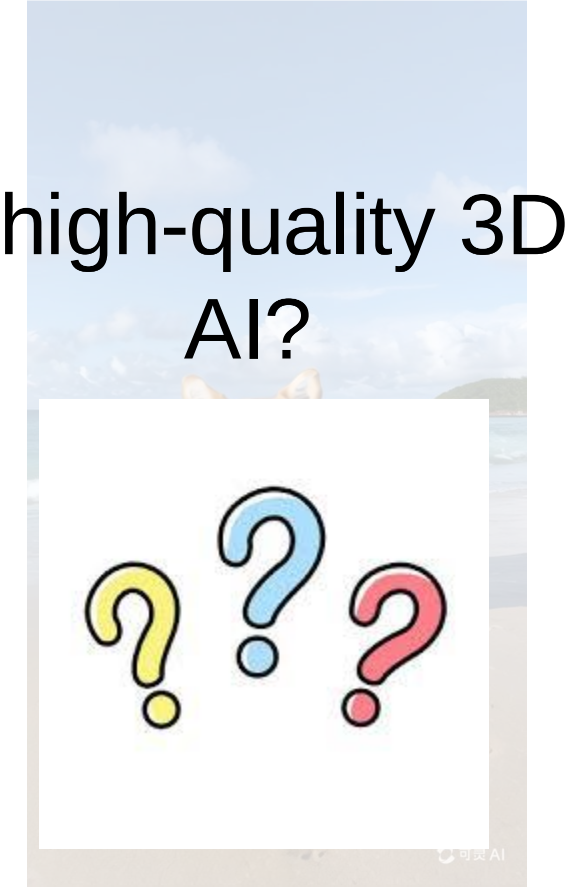


3D Asset Manually designed by 3D artists

Amazing Generation Results from Other Modalities

How to create high-quality 3D content with AI?

SD for Image Generation

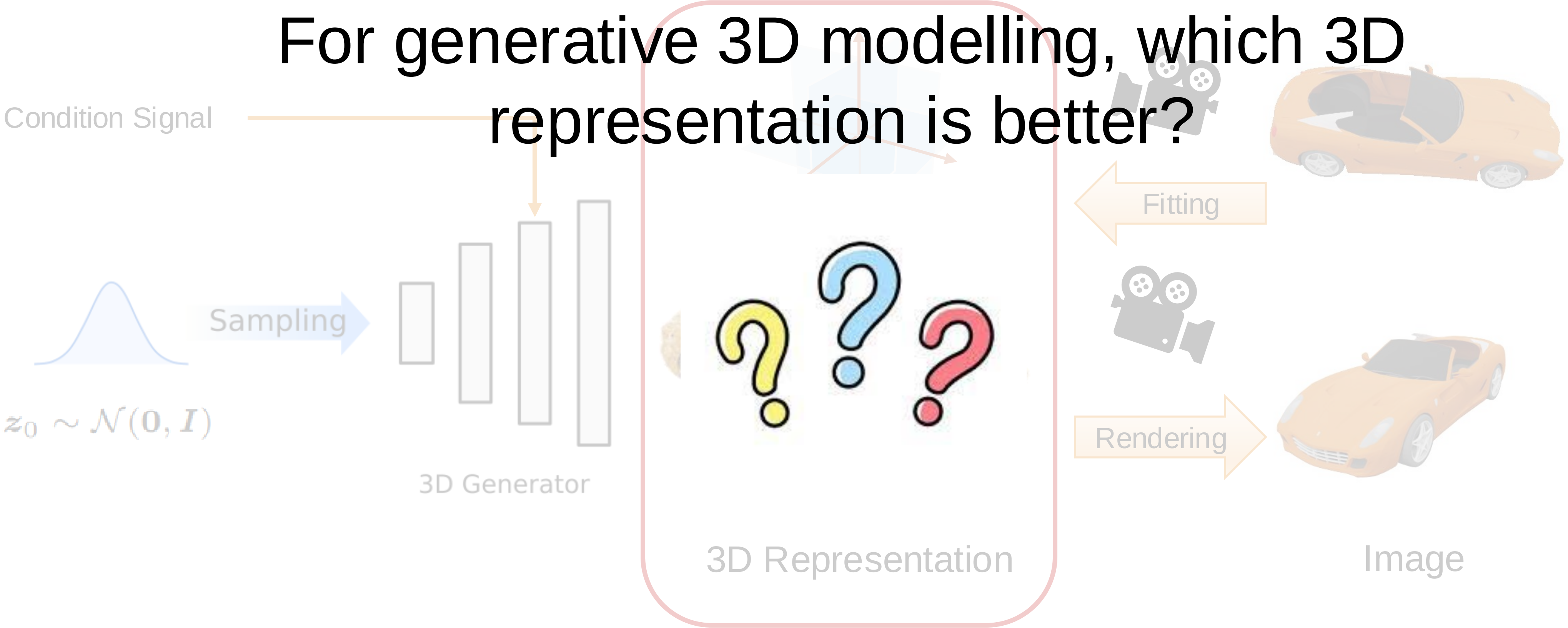


Kling for Video Generation

Suno for Music Generation

3D Generation Pipeline

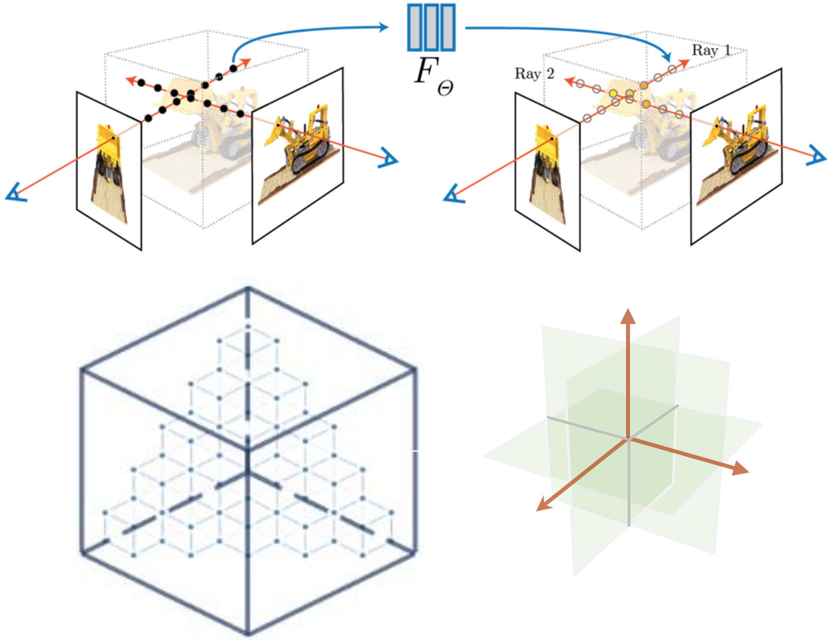
For generative 3D modelling, which 3D representation is better?



3D representation is crucial for high-quality 3D generative modeling!

Current 3D Representations

Neural Radiance Fields

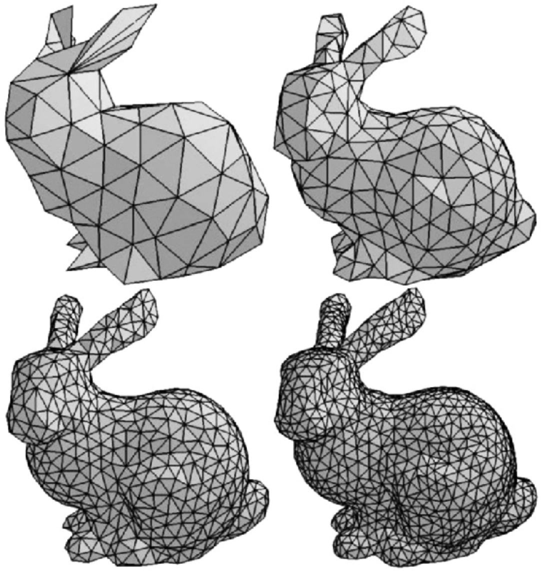


- ✓ Hybrid variants are suitable for deep learning
- ✓ Represent complex object
- ✗ Volumetric rendering is costly
- ✗ Shared decoder in generation tasks limits capability

Point Cloud

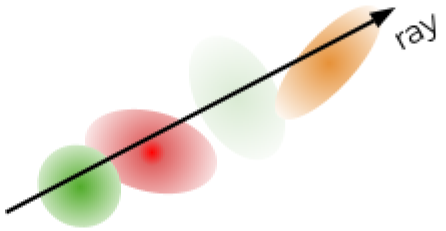
- ✓ Simple
- ✗ Num. of points is not fixed
- ✗ No topological information

Mesh

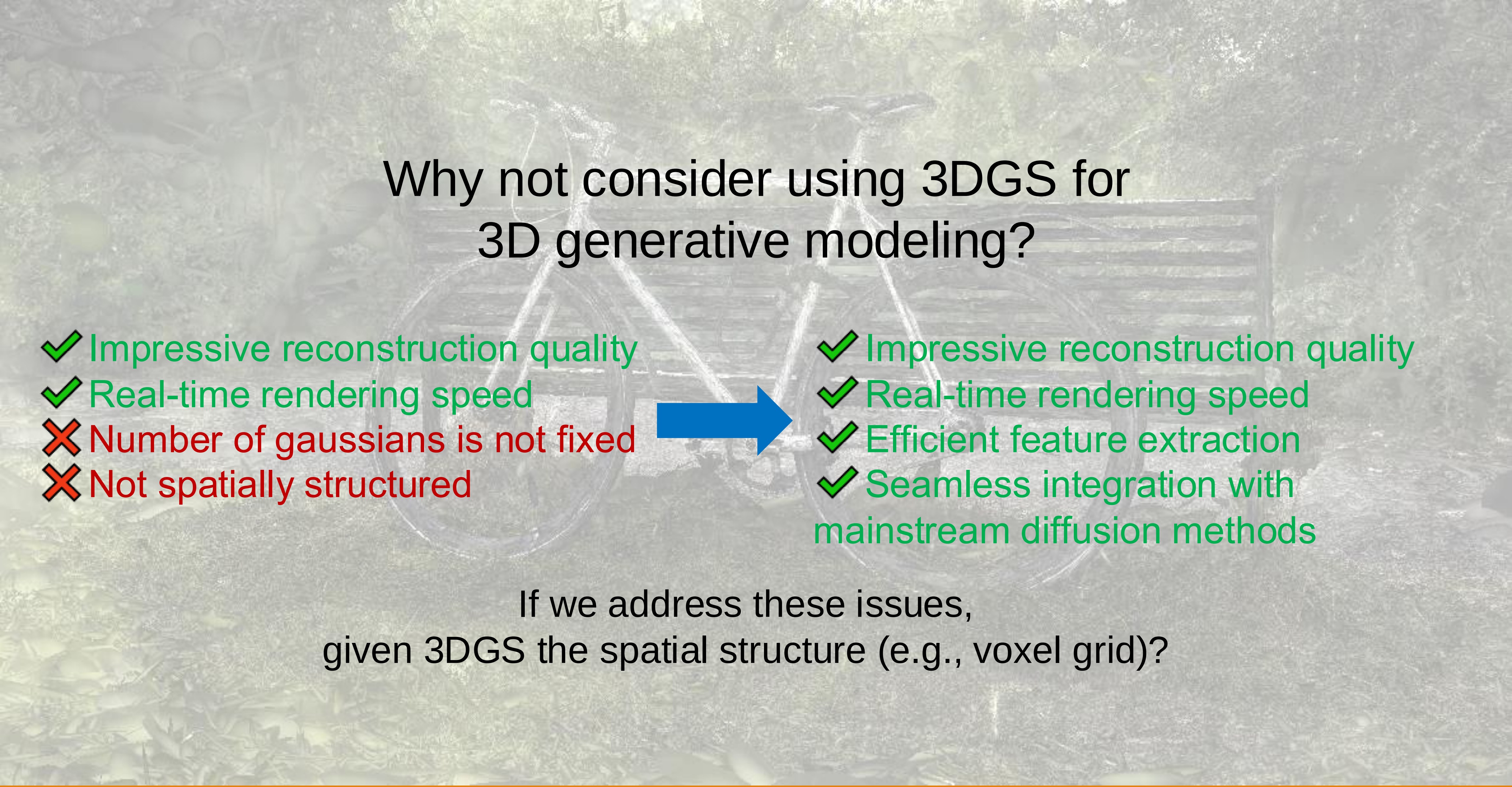


- ✓ Widely adopted in graphics pipeline
- ✓ Fast to render
- ✗ No regular structure
- ✗ Num. of vertexes is not fixed
- ✗ May be complex when representing detailed objects

3D Gaussian Splatting



- ✓ Impressive recon. quality
- ✓ Fast to render
- ✗ Num. of gaussians is not fixed
- ✗ Not spatially structured



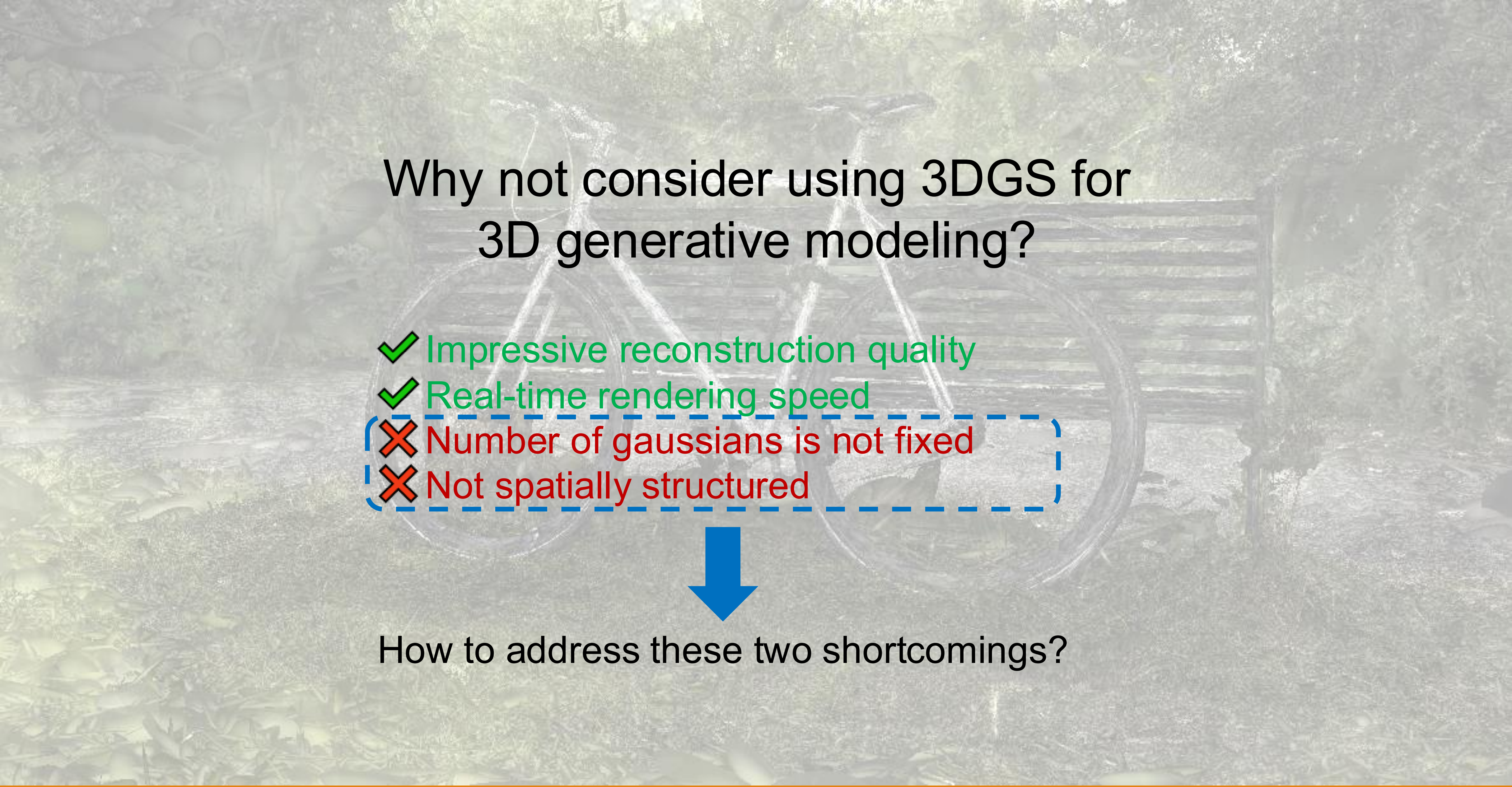
Why not consider using 3DGS for 3D generative modeling?

- ✓ Impressive reconstruction quality
- ✓ Real-time rendering speed
- ✗ Number of gaussians is not fixed
- ✗ Not spatially structured

→

- ✓ Impressive reconstruction quality
- ✓ Real-time rendering speed
- ✓ Efficient feature extraction
- ✓ Seamless integration with mainstream diffusion methods

If we address these issues,
given 3DGS the spatial structure (e.g., voxel grid)?



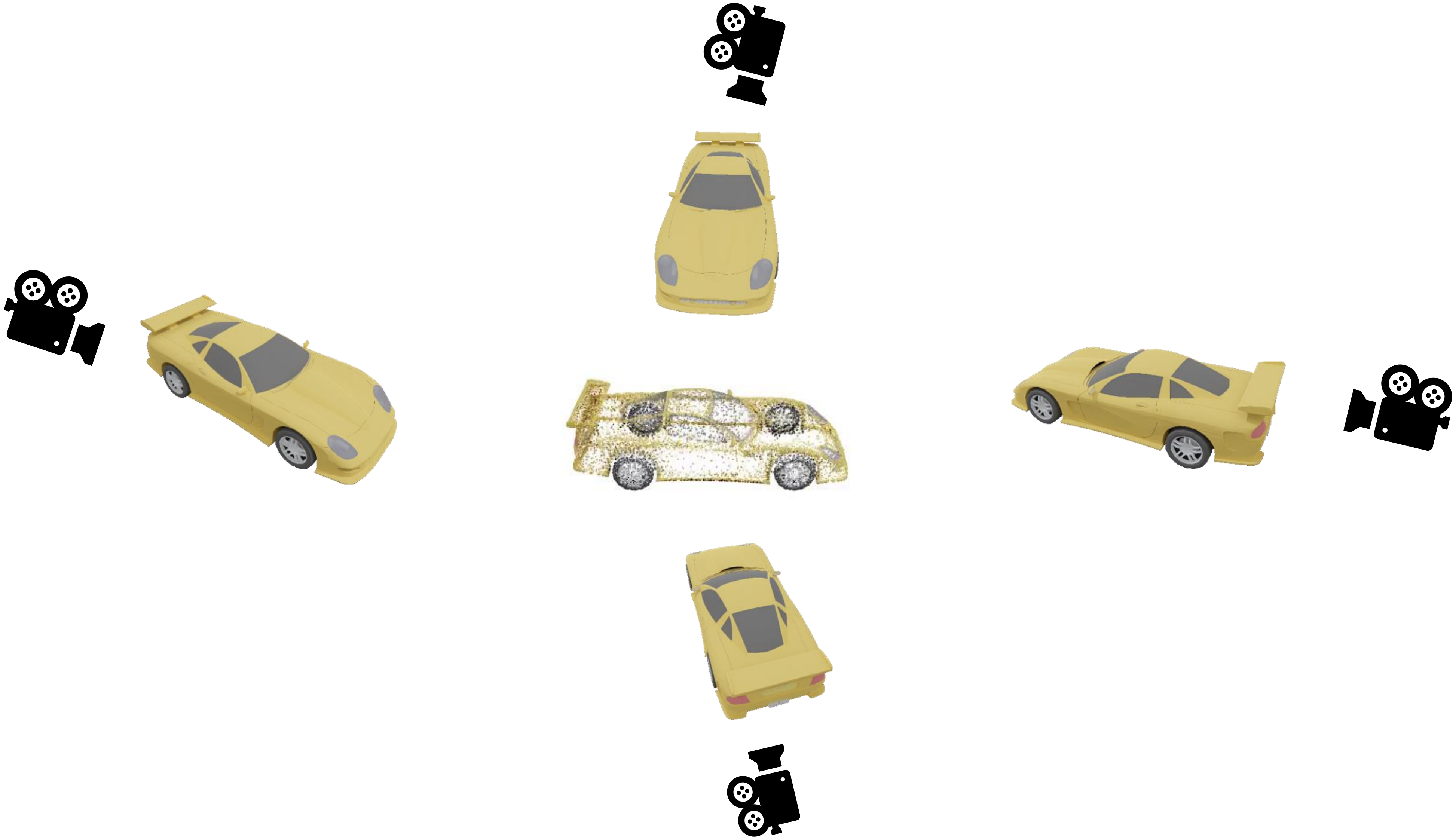
Why not consider using 3DGS for 3D generative modeling?

- ✓ Impressive reconstruction quality
- ✓ Real-time rendering speed
- ✗ Number of gaussians is not fixed
- ✗ Not spatially structured

↓

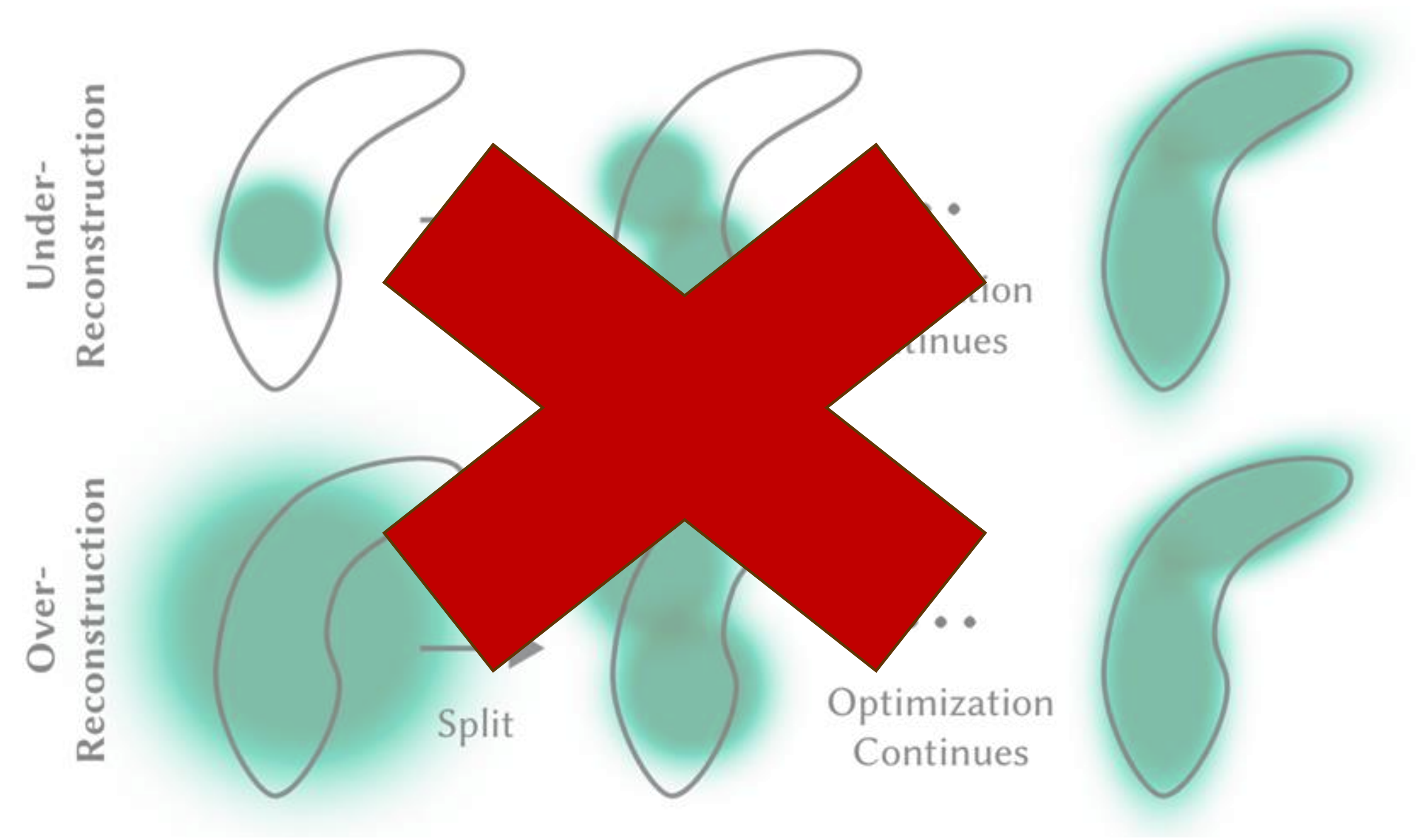
How to address these two shortcomings?

Densification-constrained Gaussian Fitting



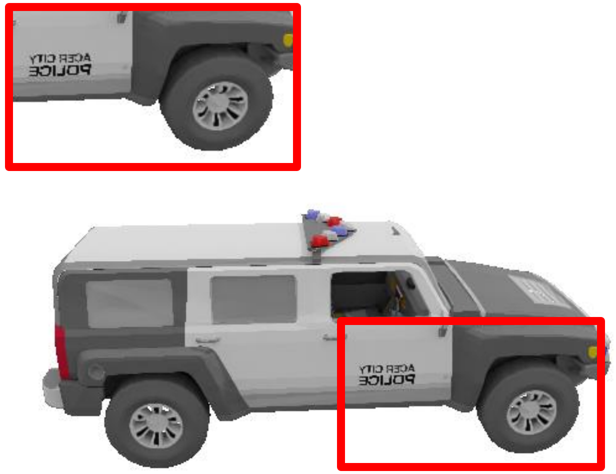
Naïve Approach of Fixing Length

Remove Densification and Pruning in Original GS

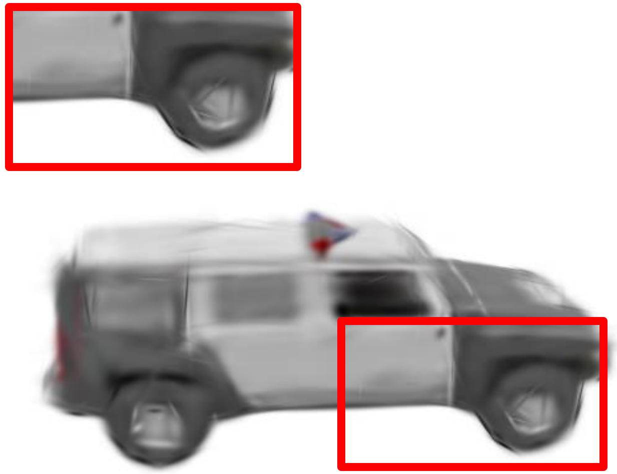


Naïve Approach of Fixing Length

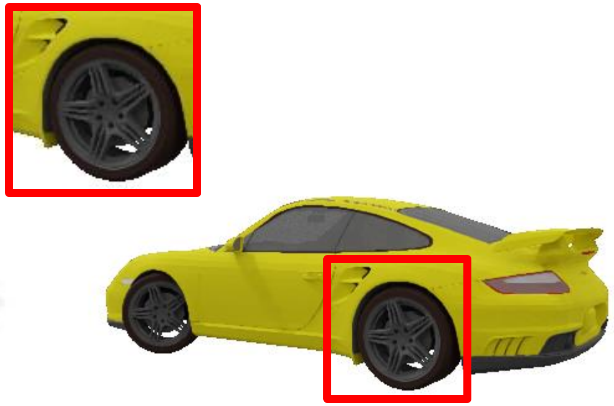
Remove Densification and Pruning in Original GS



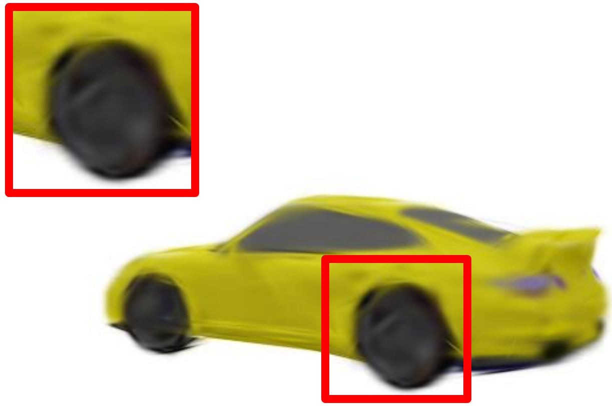
Ground-truth



Fitting Results



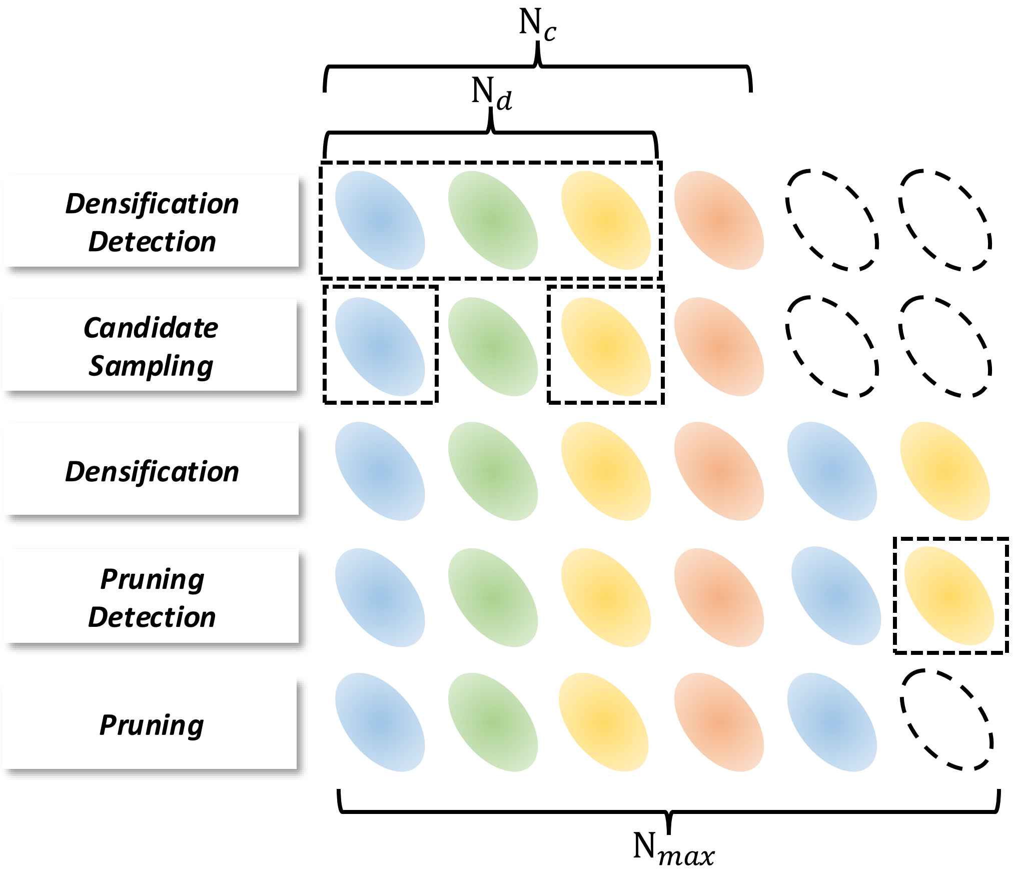
Ground-truth



Fitting Results

Densification-constrained Gaussian Fitting

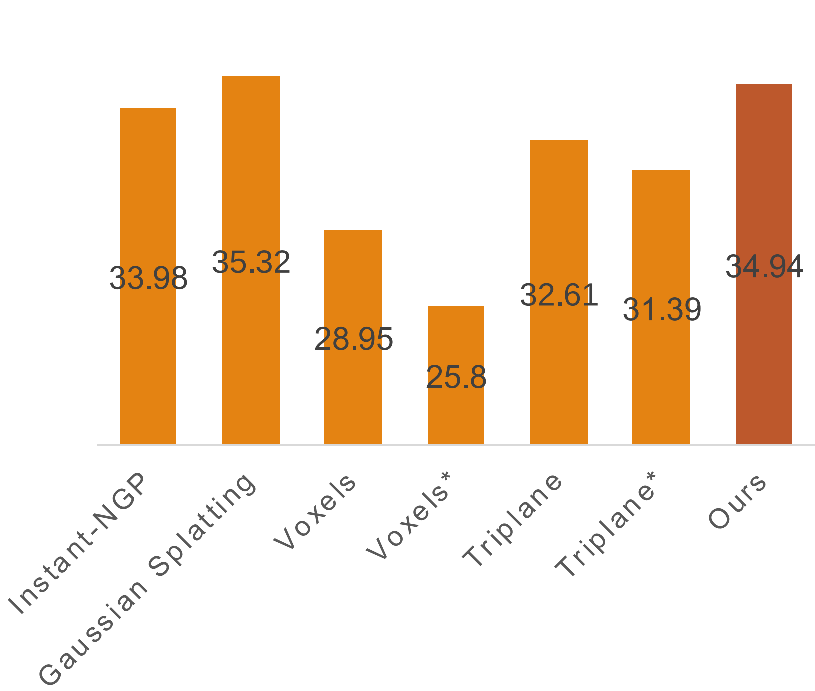
N_{max}: Predefined Maximum Number of Gaussians Used During Fitting (32,768 in this work)
N_c: Number of Gaussians in Current Iteration
N_d: Number of candidates to Perform Densification



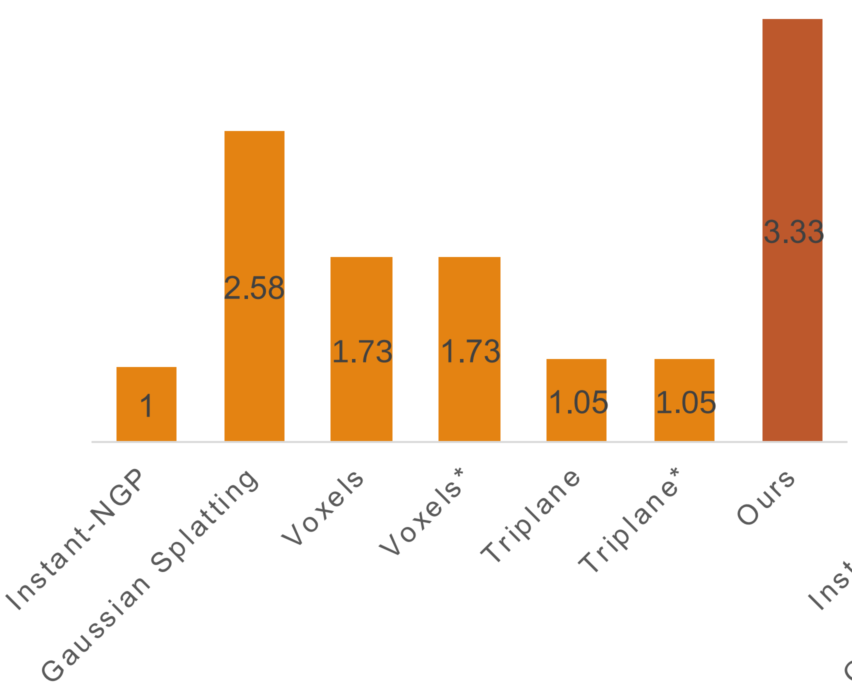
After fitting, we pad Gaussians with α = 0 to N_{max} without affecting the rendering results.

Fitting Results Evaluation

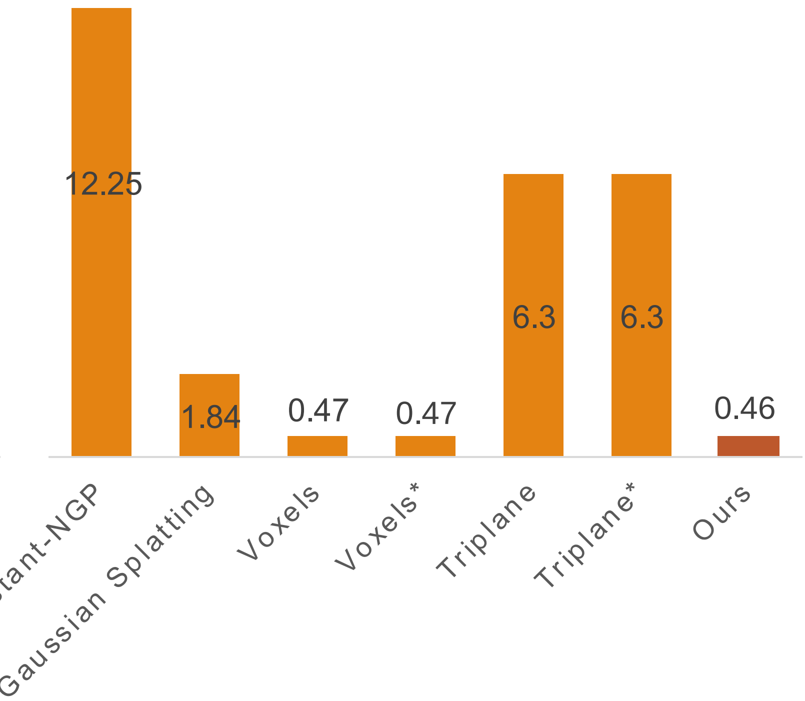
PSNR



RELATIVE SPEED

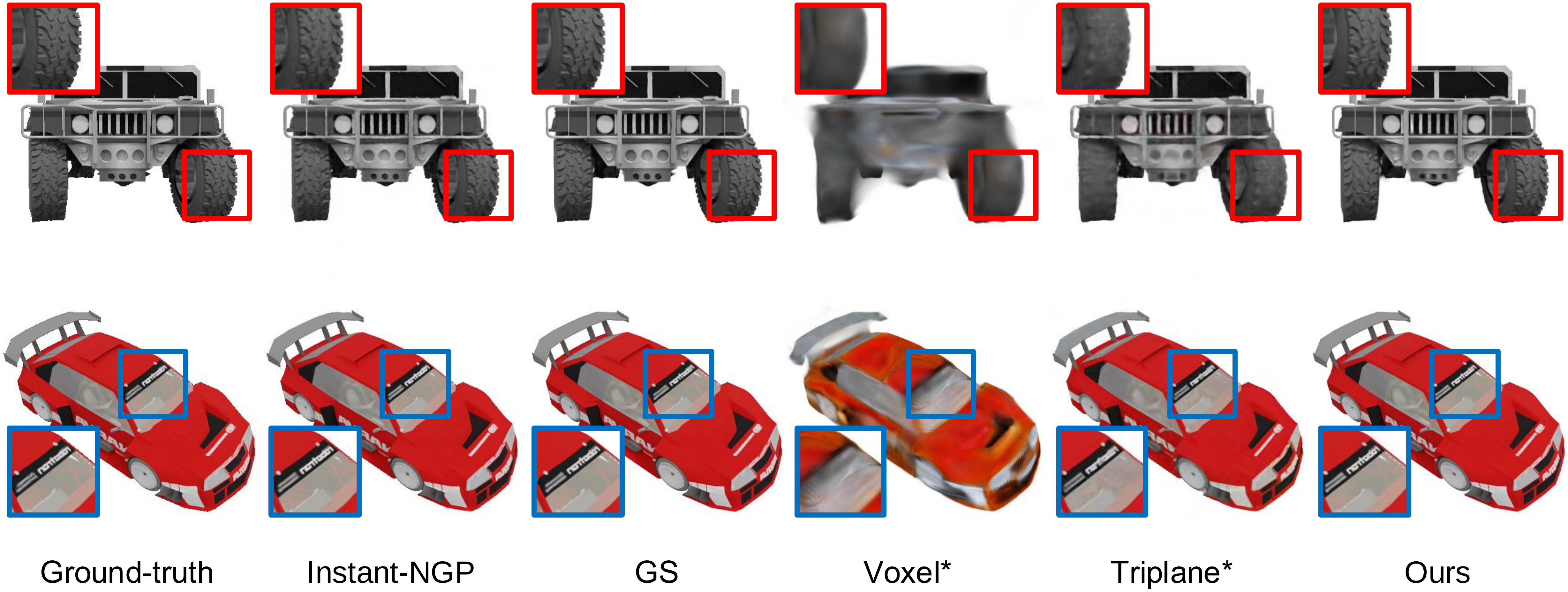


PARAMETERS



∗ denotes that the implicit feature decoder is shared across different objects.

Fitting Results Evaluation



Ground-truth Instant-NGP GS Voxel* Triplane* Ours

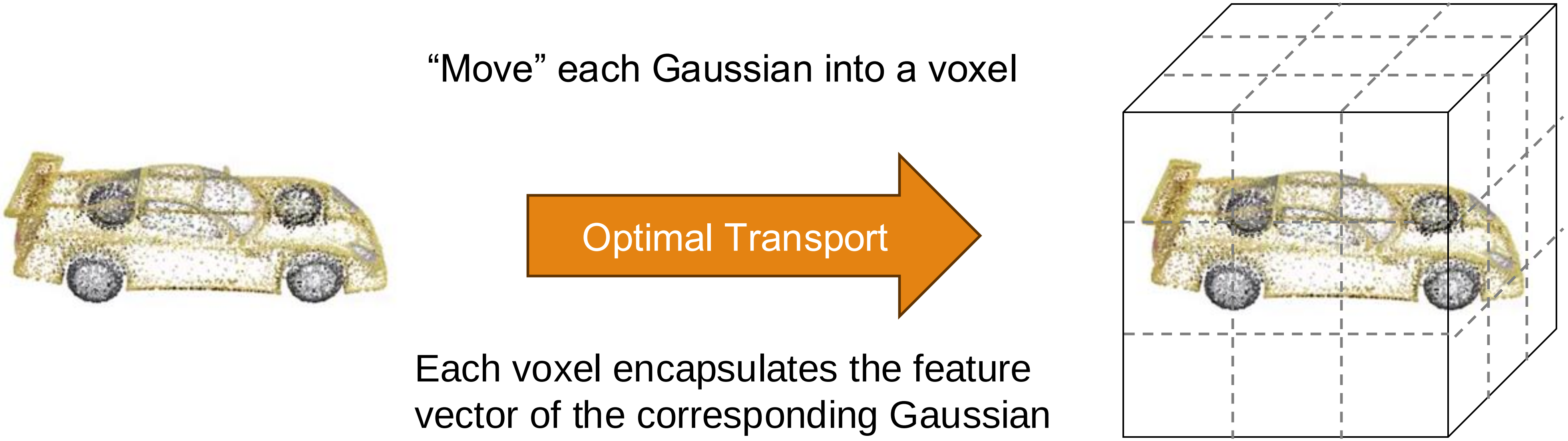
∗ denotes that the implicit feature decoder is shared across different objects.

What are the advantages of Densification-constrained Gaussian Fitting?

Next, structure the 3D Gaussians into voxel grids!

- ✓ Strong capacity: impressive reconstruction quality using only 32,768 Gaussians
- ✓ Fast: fitting converge faster than other methods
- ✓ Compact: orders of magnitude fewer parameters compared to existing works of similar quality, also reduces the modeling difficulty for the diffusion models

Gaussian Structuralization via Optimal Transport



“Move” each Gaussian into a voxel

Optimal Transport

Each voxel encapsulates the feature vector of the corresponding Gaussian

Gaussian Structuralization via Optimal Transport

What are the advantages of OT-based structuralization?

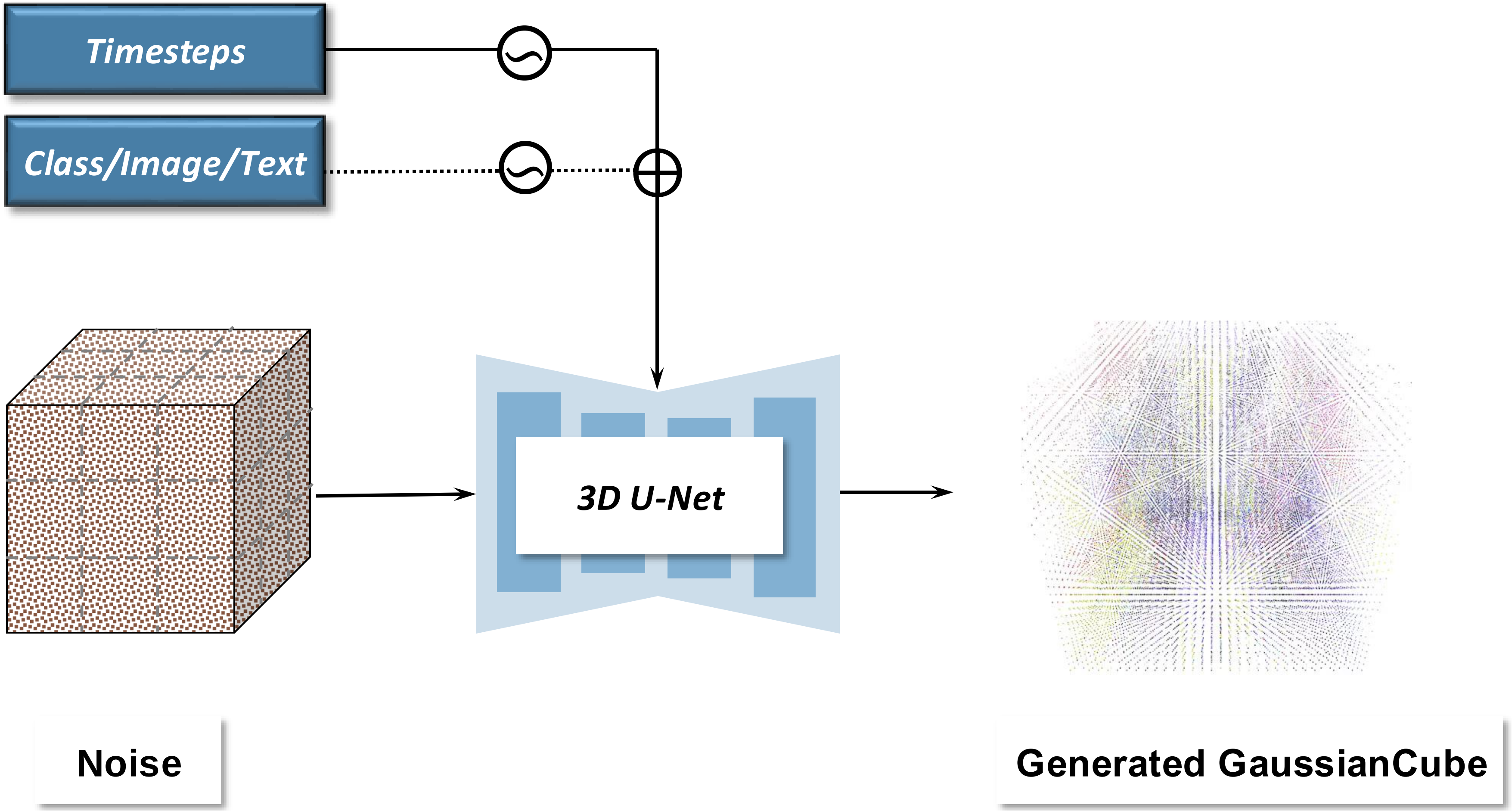
✓ Allow the use of standard 3D U-Net as backbone for diffusion without elaborate designs

✓ Achieve maximal spatial correspondence, characterized by minimal total transport distances

✓ Standard 3D convolution can capture the correlations among neighboring Gaussians, facilitating efficient feature extraction

✓ Post-processing does not affect fitting quality

3D Diffusion on GaussianCube



Timesteps

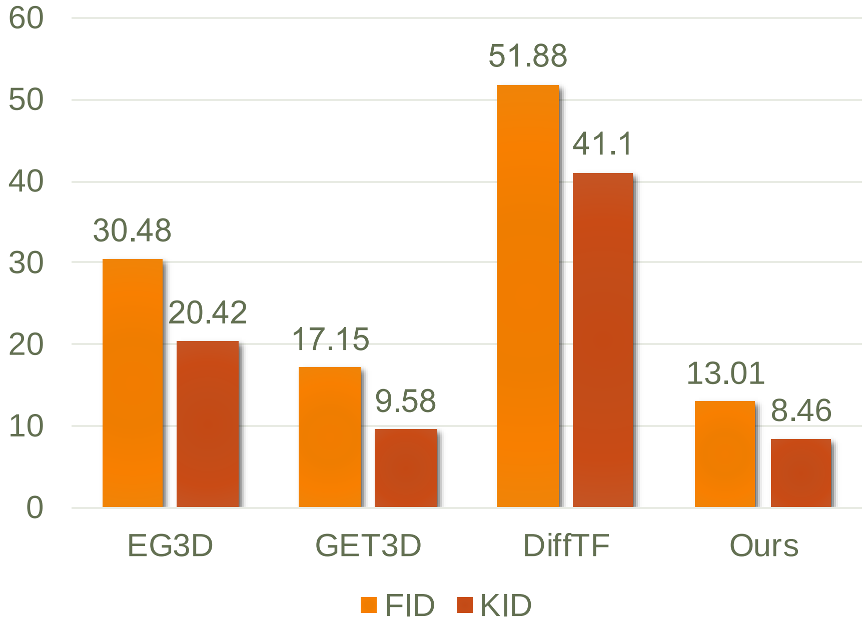
Class/Image/Text

3D U-Net

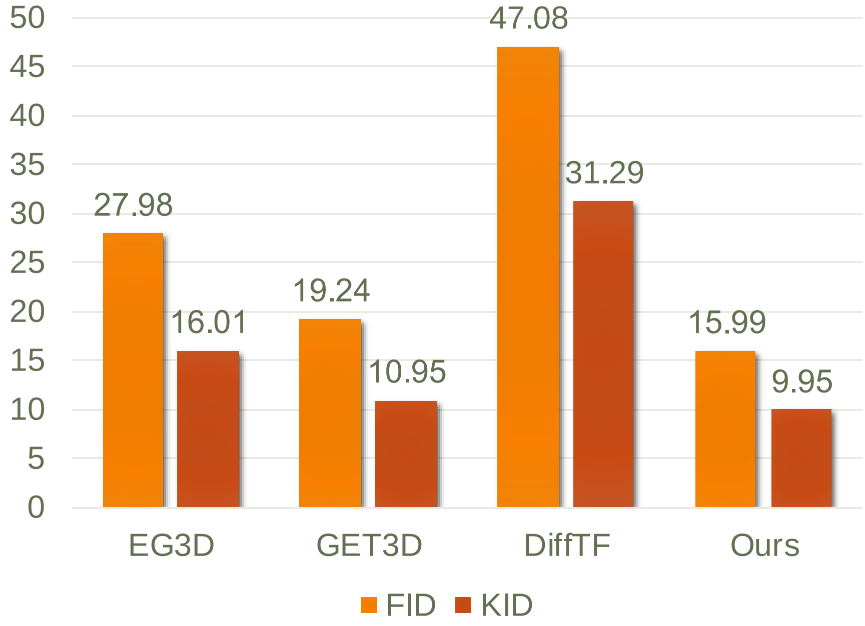
Noise

Generated GaussianCube

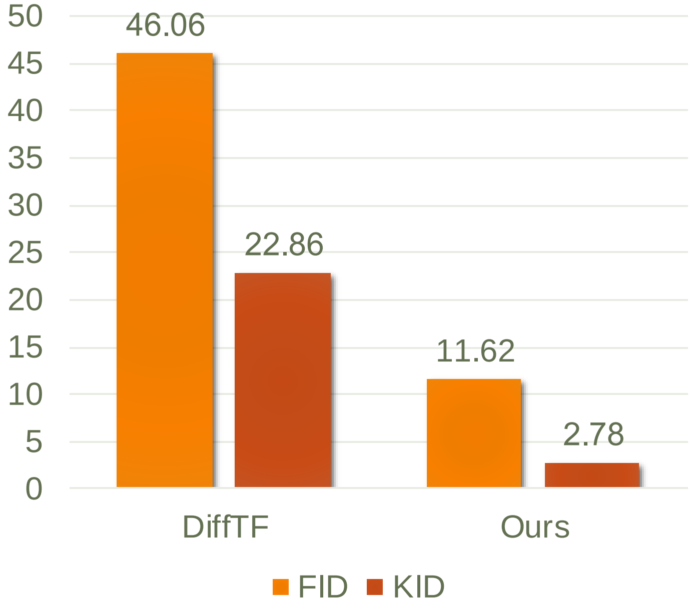
Generation Results Evaluation



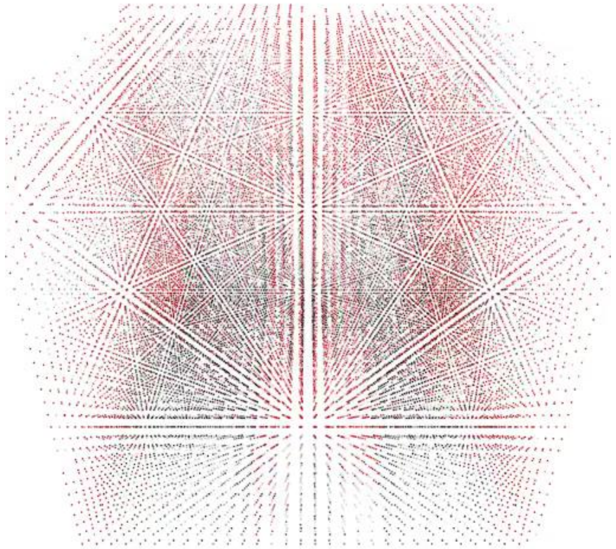
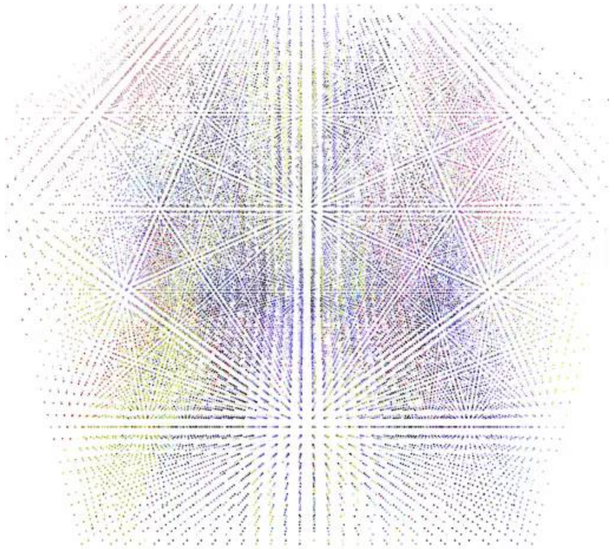
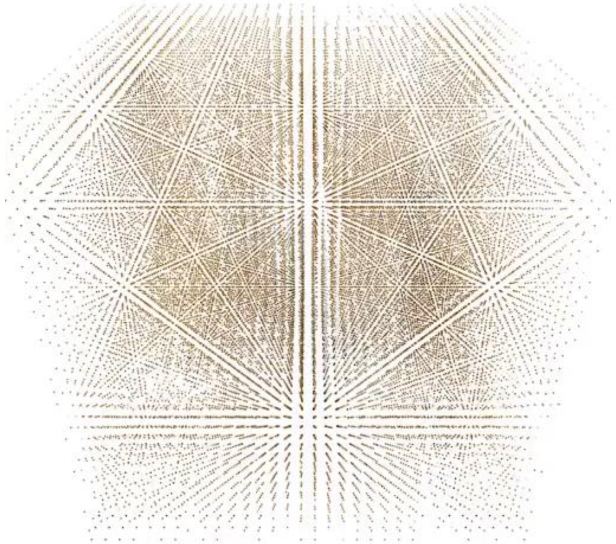
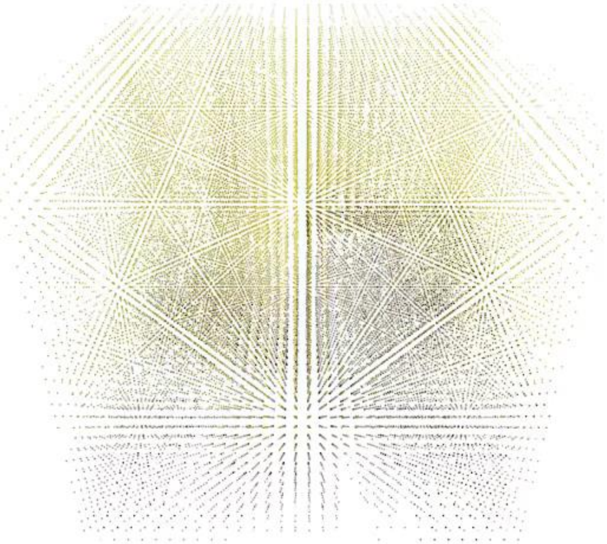
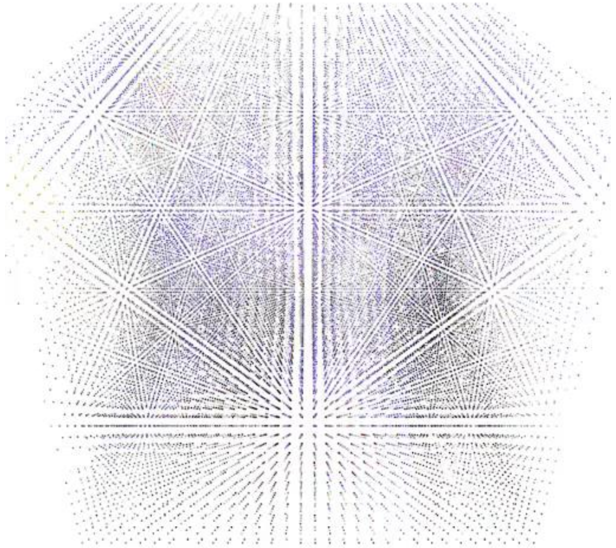
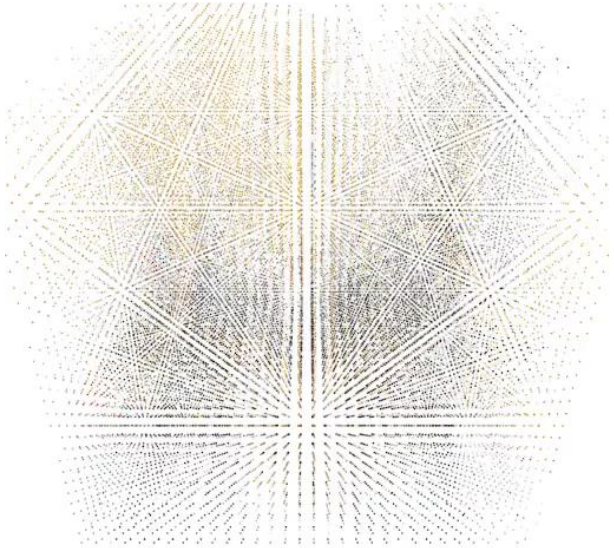
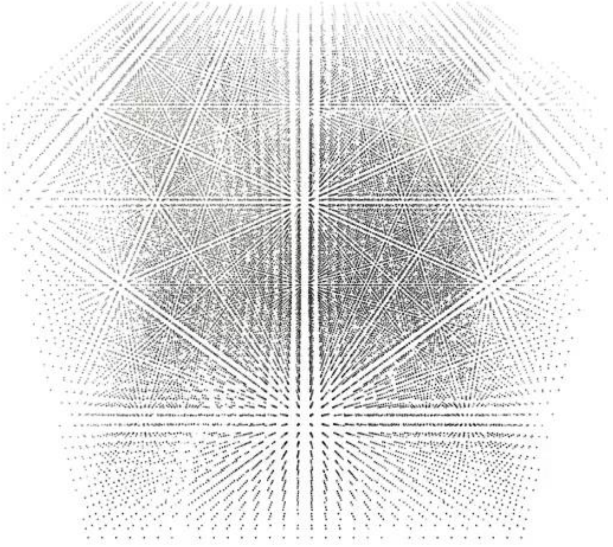
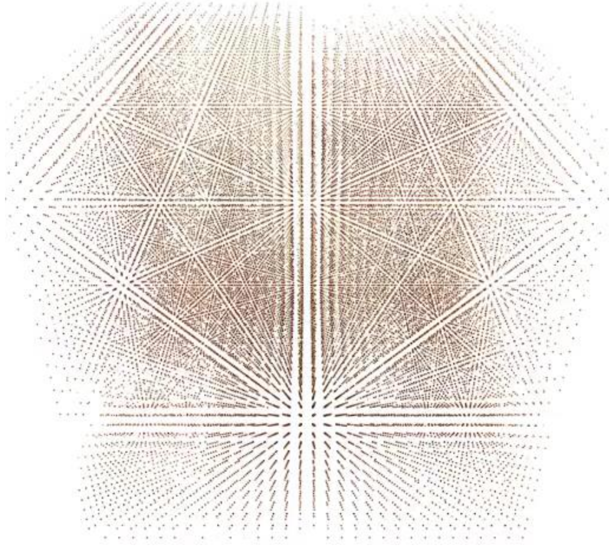
Unconditional Generation on ShapeNet Car



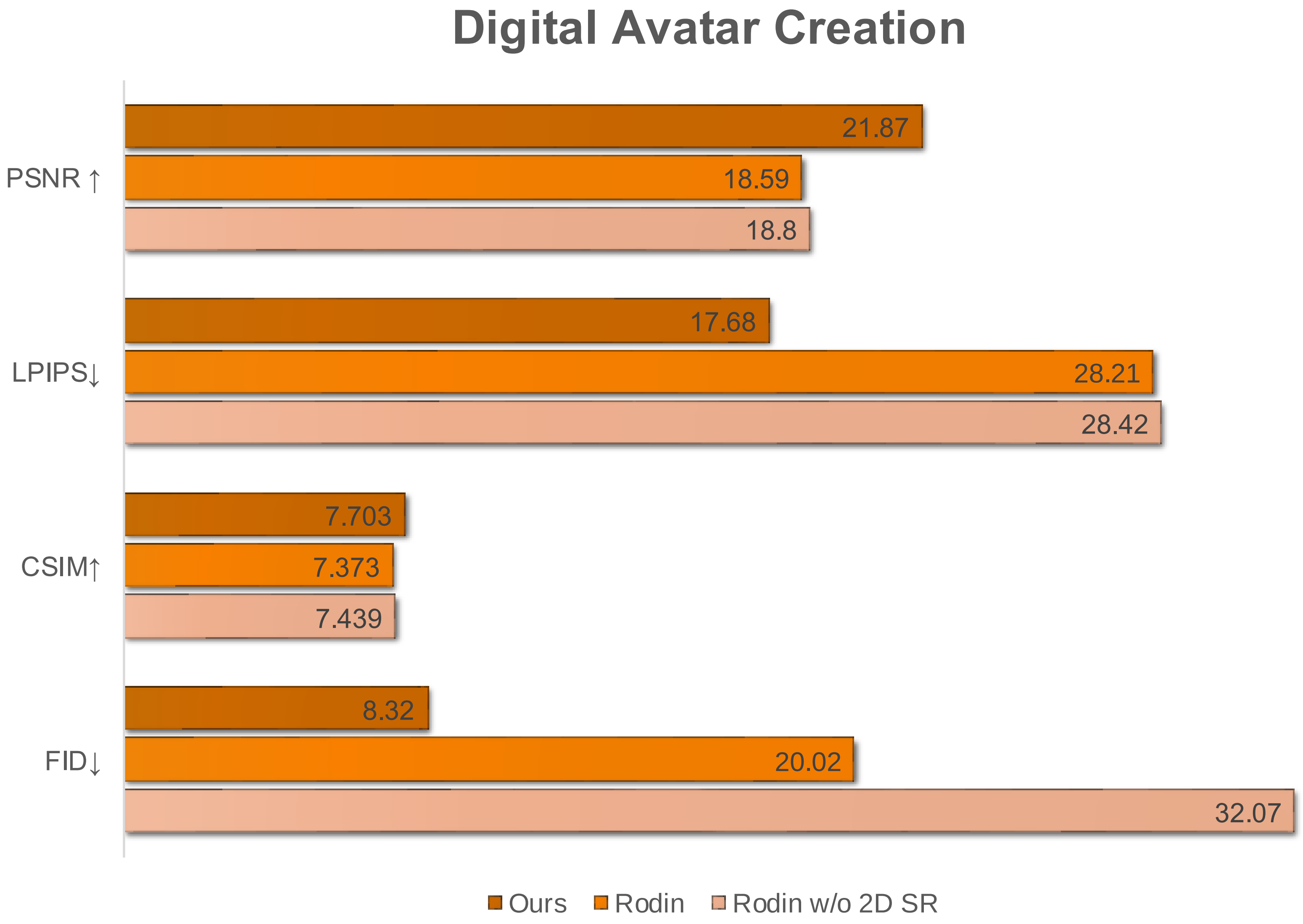
Unconditional Generation on ShapeNet Chair



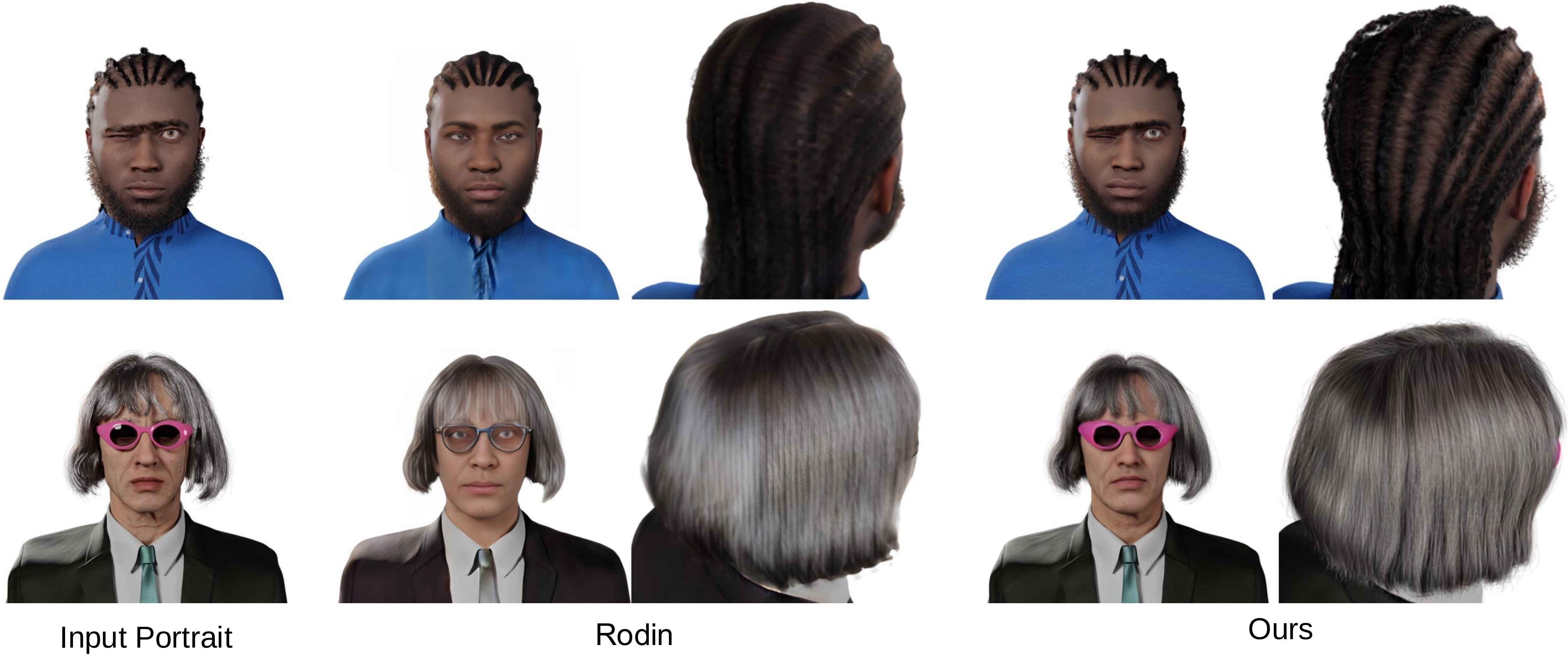
Class-conditioned Generation on OmniObject3D

Generation Results Evaluation

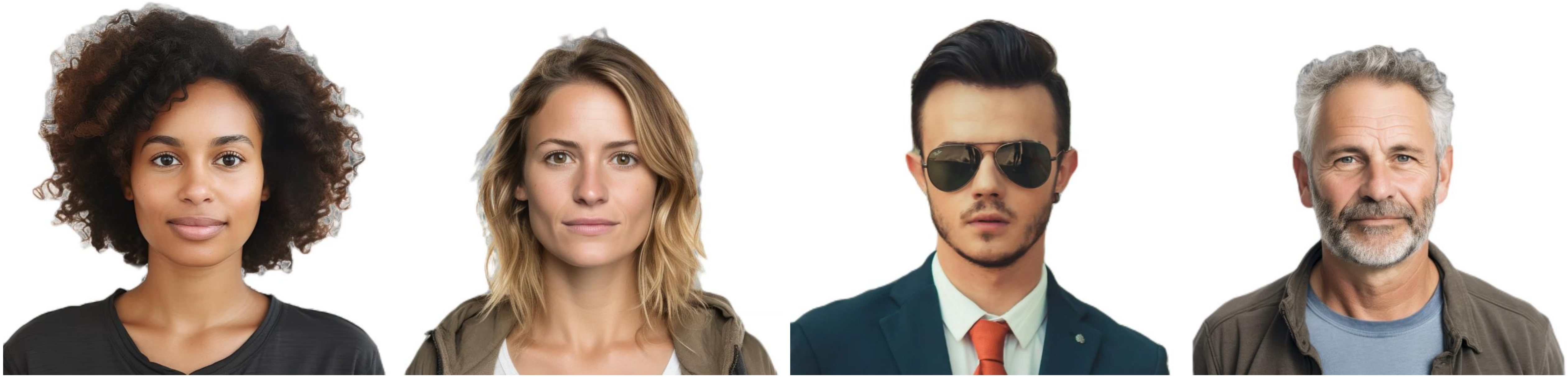


Generation Results Evaluation



Generation Results Evaluation

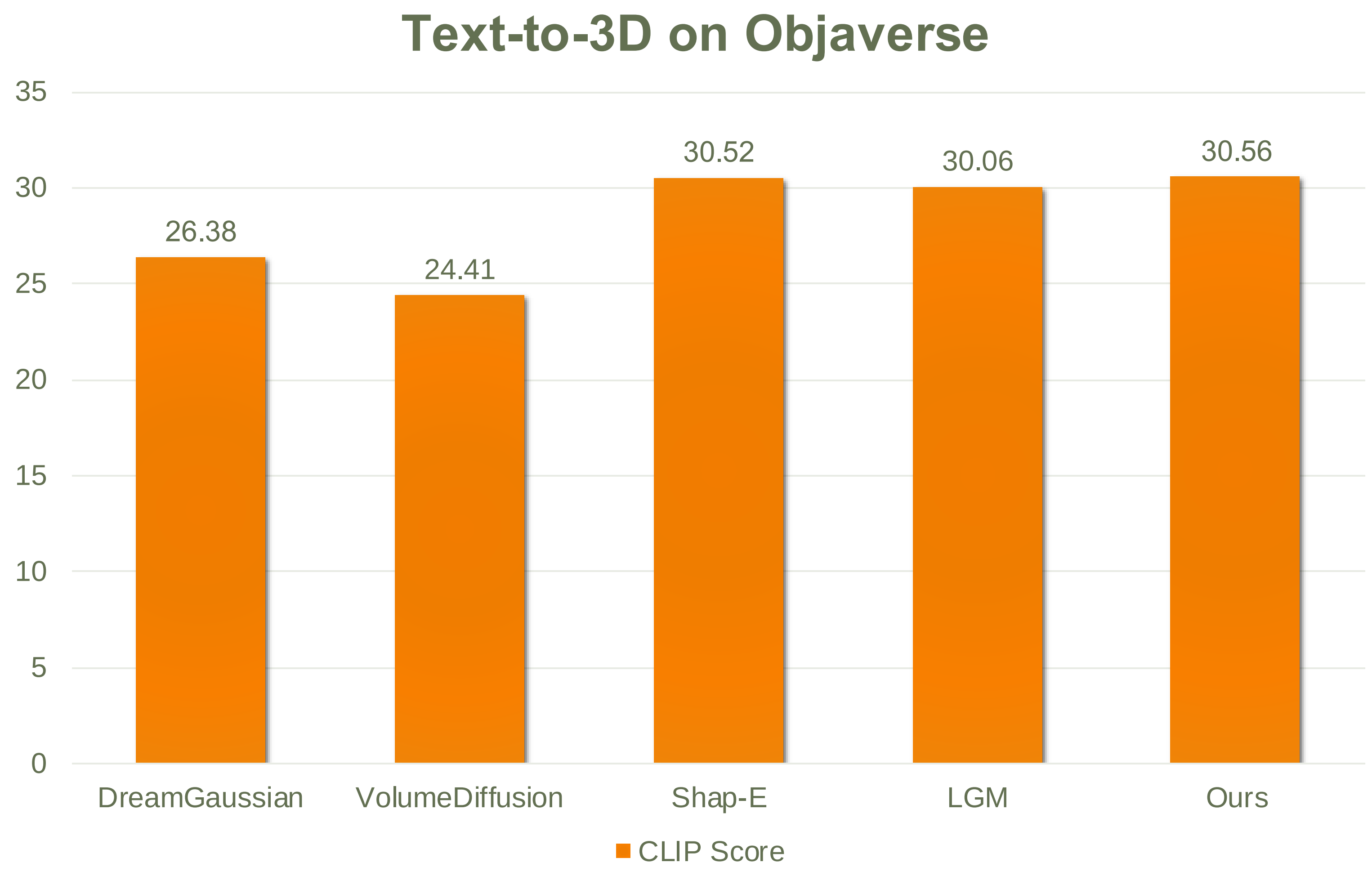
3D avatar creation from in-the-wild portrait



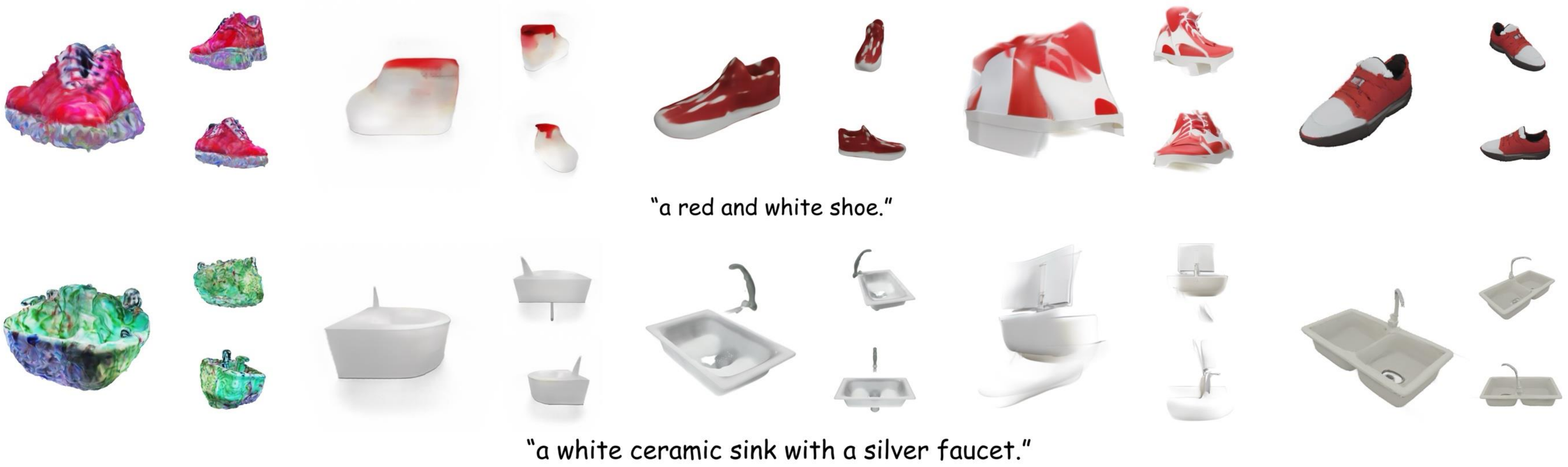
Generated Avatars

Input Portraits

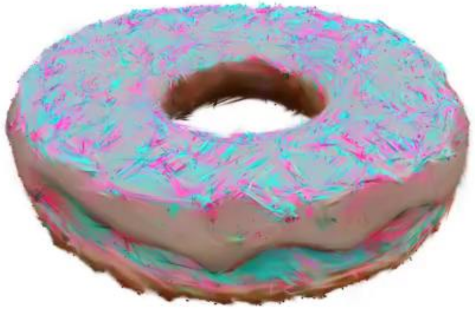
Generation Results Evaluation



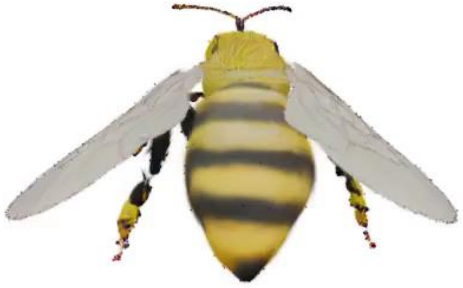
Generation Results Evaluation



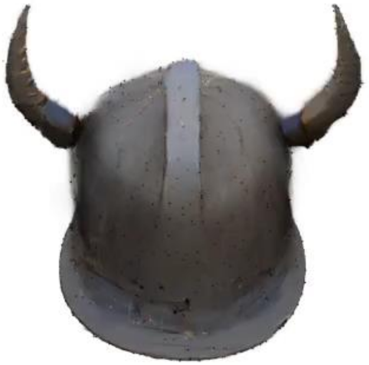
DreamGaussian VolumeDiffusion Shap-E LGM Ours



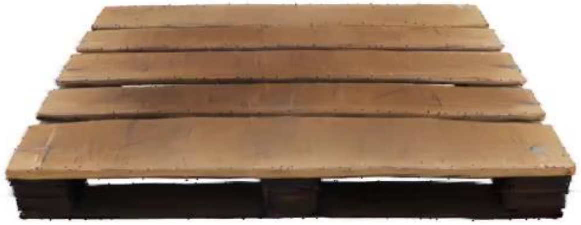
"a donut with blue frosting and sprinkles."



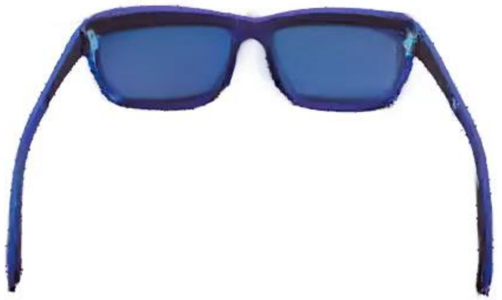
"a yellow and black bee with a white wing."



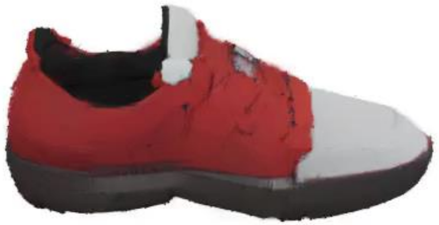
"a silver helmet with horns on top."



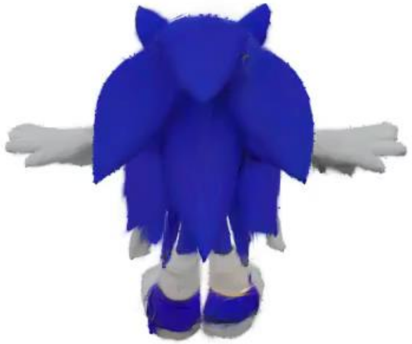
"a wooden pallet."



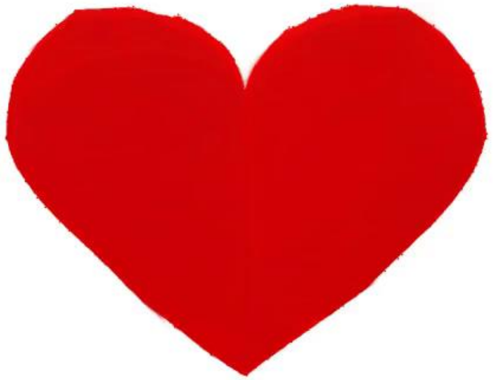
"a pair of sunglasses with blue lenses."



"a red and white shoe."



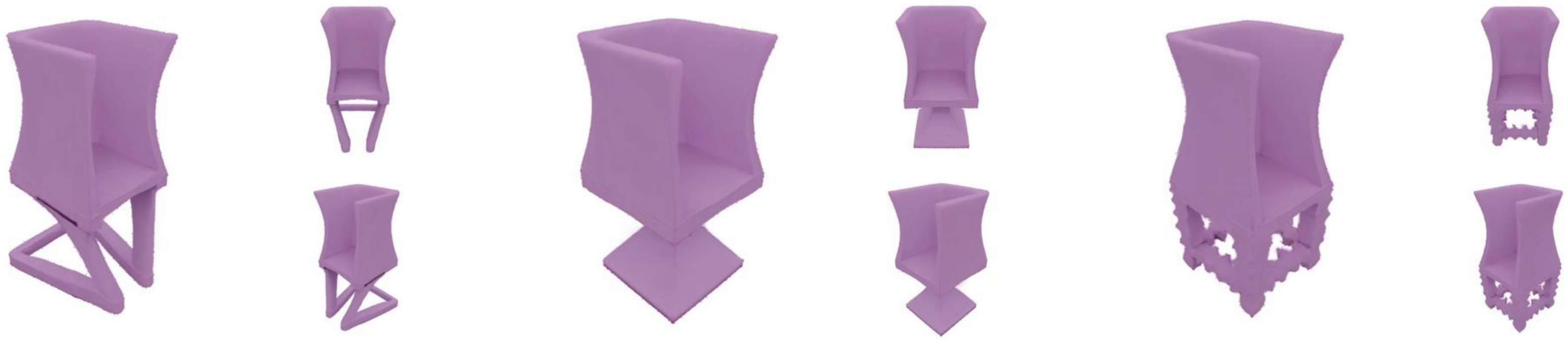
"a blue and white cartoon character of Sonic the Hedgehog."



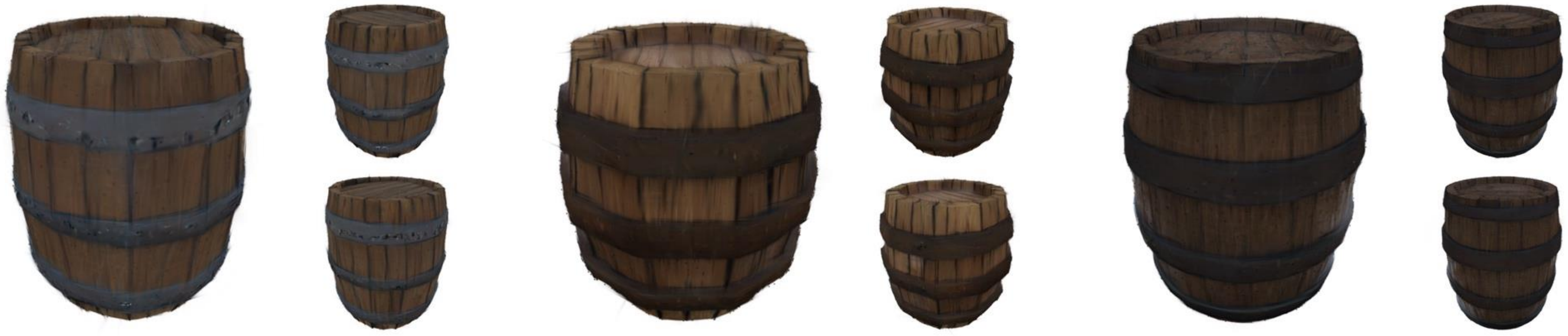
"a red heart."

Diverse Results from the Same Text Input

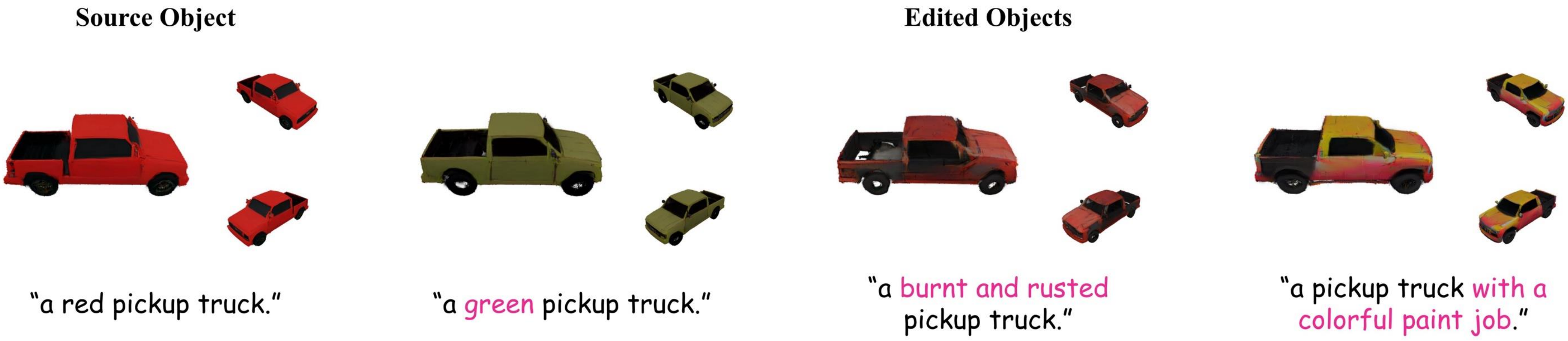
“a purple chair with a curved back and a curved seat.”



“a wooden barrel with a metal band around the middle.”



Text-guided 3D Editing


Source Object
Edited Objects
"a red pickup truck."
"a green pickup truck."
"a burnt and rusted pickup truck."
"a pickup truck with a colorful paint job."

Thanks!

Some slides inspired from Tengfei Wang and Jun Gao, thanks for their nice works.