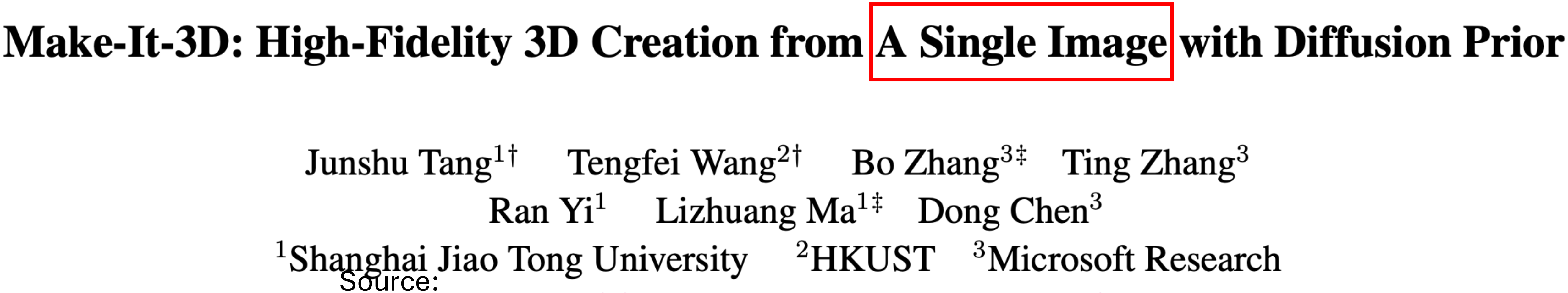


Make-It-3D: High-Fidelity 3D Creation from A Single Image with Diffusion Prior

Junshu Tang[1†] Tengfei Wang[2†] Bo Zhang[3‡] Ting Zhang[3]
Ran Yi[1] Lizhuang Ma[1‡] Dong Chen[3]
[1]Shanghai Jiao Tong University [2]HKUST [3]Microsoft Research

Source: https://make-it-3d.github.io/

ICCV 2023

Citations: 130

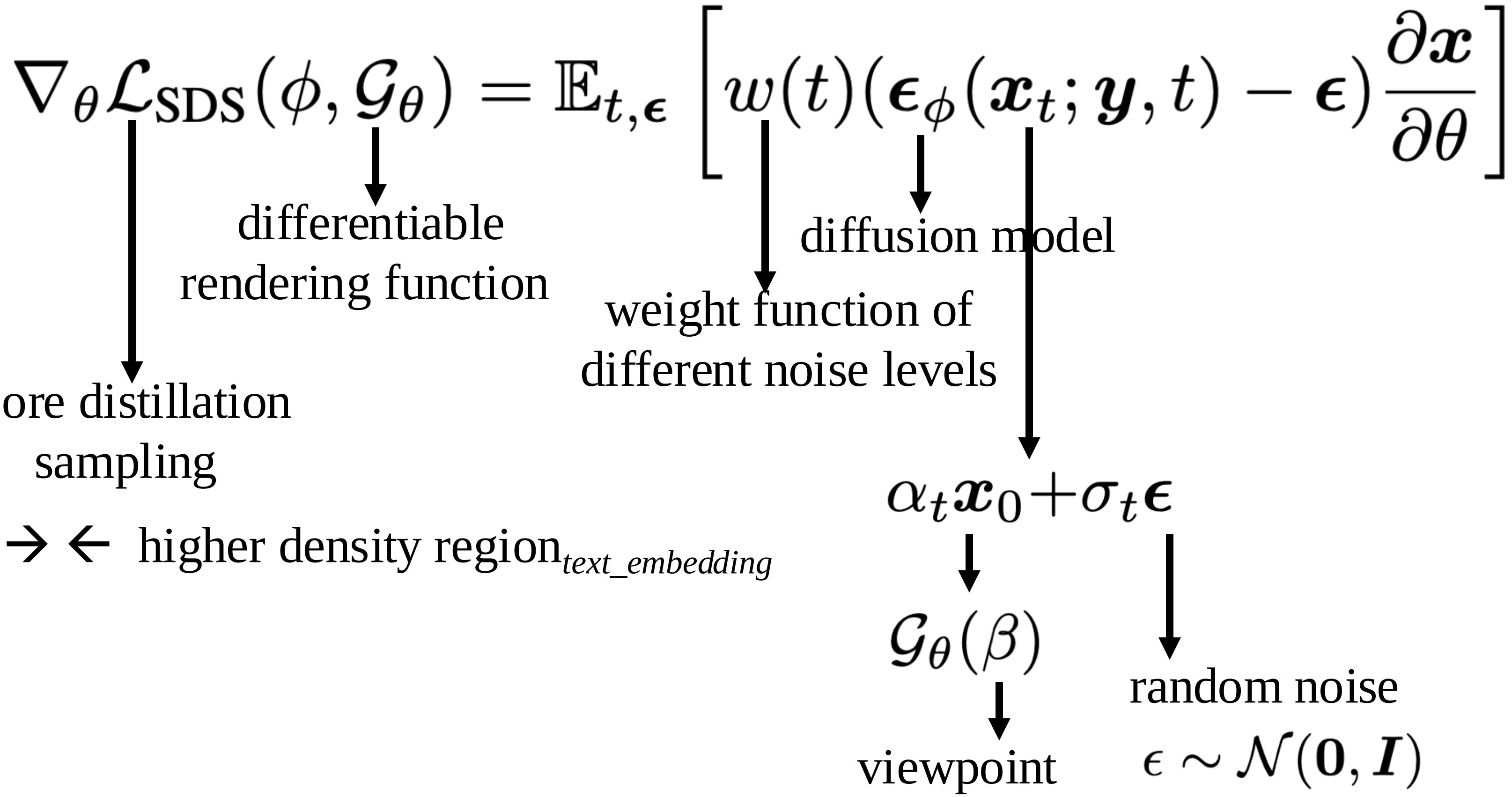
How would they look from a different view?

Let’s Make it 3D !

Challenge: inferring both geometry and missing texture

Preliminaries

measures the similarity (image, text prompt)



$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, \mathcal{G}_\theta) = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[w(t)(\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t;\boldsymbol{y},t)-\boldsymbol{\epsilon})\frac{\partial \boldsymbol{x}}{\partial \theta}\right]$$

differentiable rendering function

diffusion model

weight function of different noise levels

score distillation sampling

$$\alpha_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\epsilon}$$

$$\mathcal{G}_\theta(\beta)$$

viewpoint

random noise

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

push rendered images → ← higher density region$_{\textit{text\_embedding}}$

Coarse Stage: Single-view 3D Reconstruction

1. 3D model → look like the 2D reference picture
2. New views → make sense and look realistic
3. 3D model → realistic shape and depth

Reference view per-pixel loss

$$\mathcal{L}_{\text{ref}} = \|\boldsymbol{x} \odot \boldsymbol{m} - \mathcal{G}_\theta(\beta_{\text{ref}})\|_1$$

foreground
matting mask

Diffusion prior

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, \mathcal{G}_\theta) = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[w(t)(\boldsymbol{\epsilon}_\phi(\boldsymbol{z}_t;\boldsymbol{y},t)-\boldsymbol{\epsilon})\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial \theta}\right]$$

Noisy latent

Coarse Stage: Single-view 3D Reconstruction

→ enforces the generated model to match the reference image

$$\mathcal{L}_{\text{CLIP-D}}(\mathcal{X}, \mathcal{G}_\theta(\beta)) = -\mathcal{E}_{\text{CLIP}}(\mathcal{X}) \cdot \mathcal{E}_{\text{CLIP}}(\hat{\mathcal{X}}_0(\beta, t))$$

CLIP image encoder

Depth prior

$$\mathcal{L}_{\text{depth}} = -\frac{\text{Cov}(d(\beta_{\text{ref}}), d)}{\text{Var}(d(\beta_{\text{ref}}))\text{Var}(d)}$$

Coarse Stage: Single-view 3D Reconstruction

3D model appears visually appealing and plausible

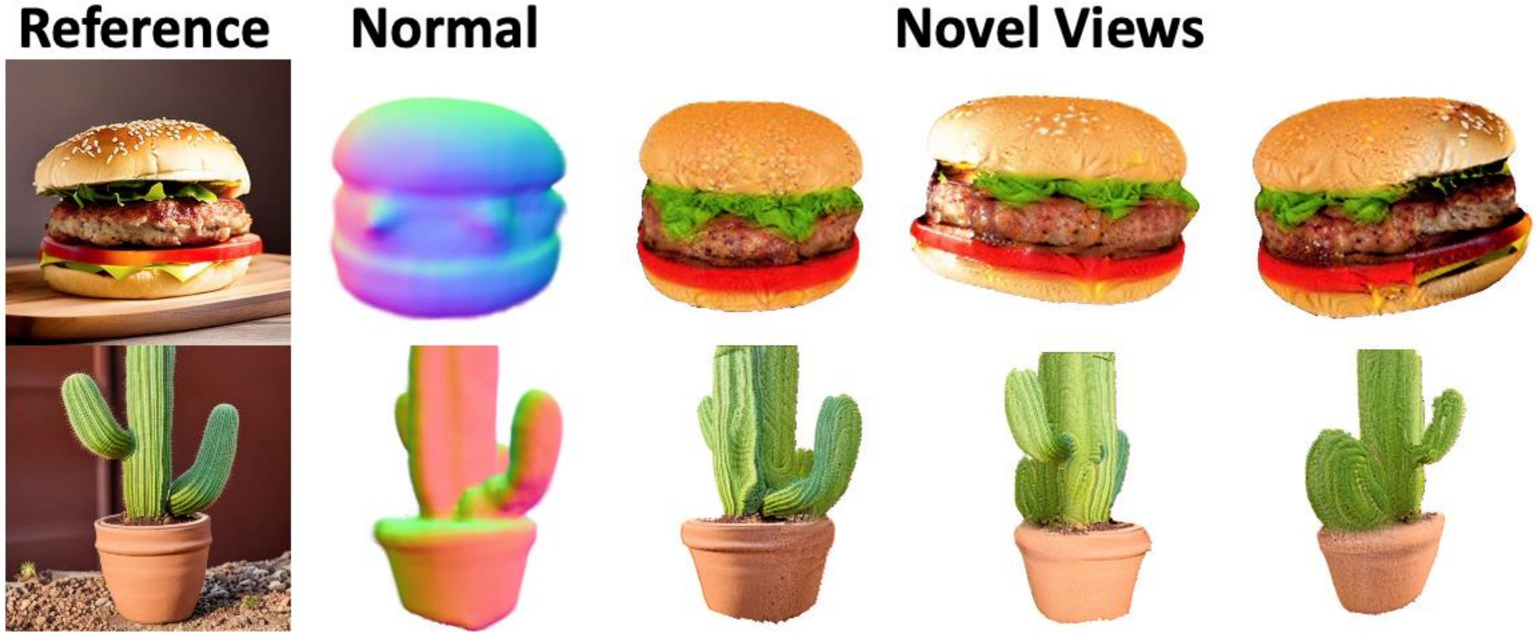
$$\mathcal{L}_{\text{ref}}, \overbrace{\mathcal{L}_{\text{SDS}}, \mathcal{L}_{\text{CLIP-D}}} \text{ and } \mathcal{L}_{\text{depth}}$$

ensure plausible geometry &
resolve most of the shape ambiguity

penalize the pixel-wise difference b/n
the rendering and the input image

encourage the rendering
to align with the reference image

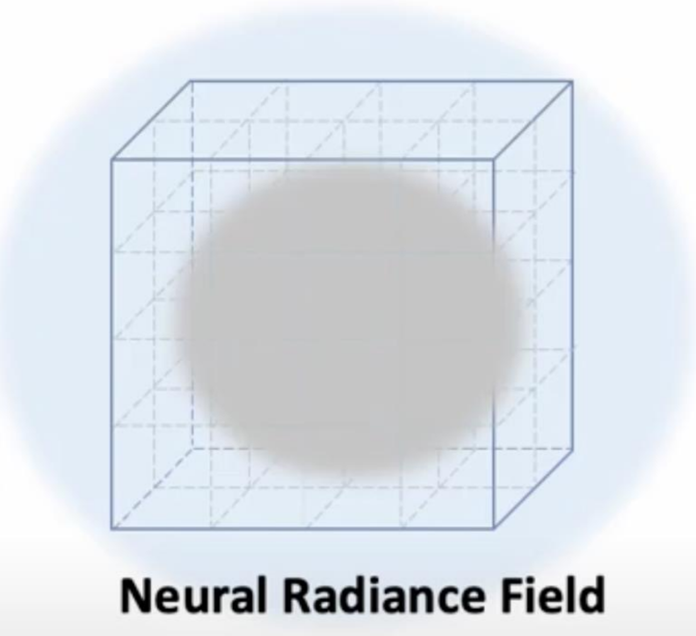
measures the similarity (image, text prompt)



Pipeline: Coarse Stage

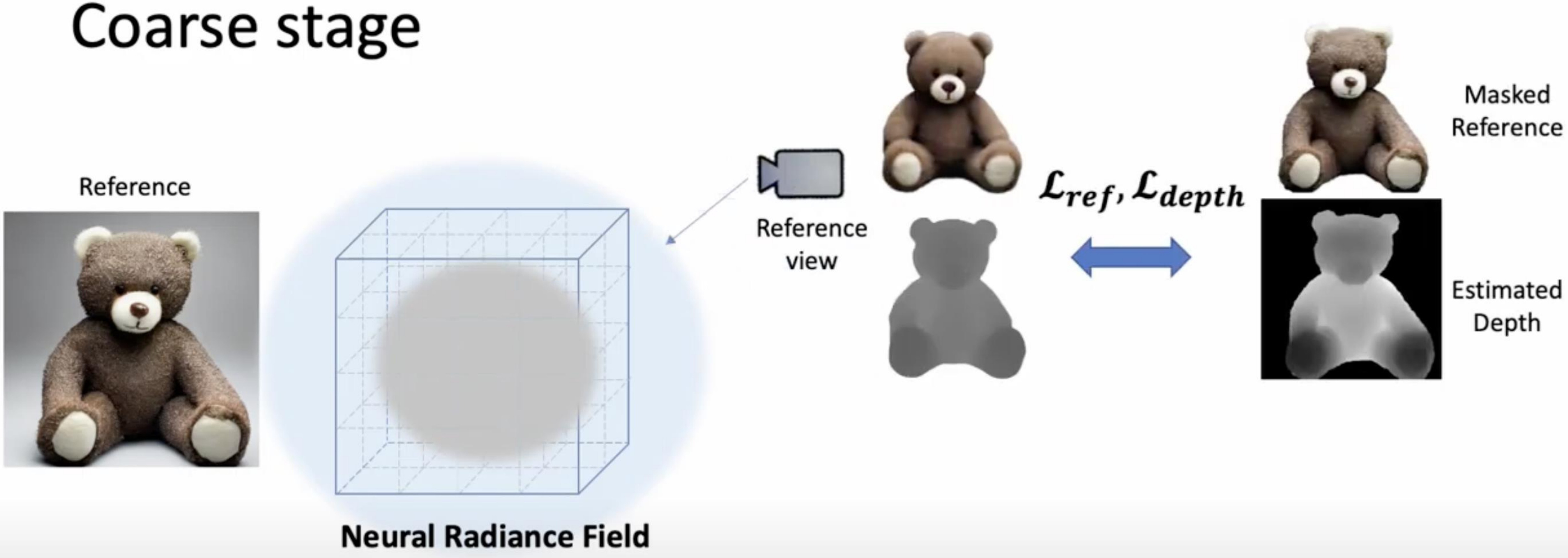
Coarse stage

Reference



Neural Radiance Field

Coarse stage



Reference

Neural Radiance Field

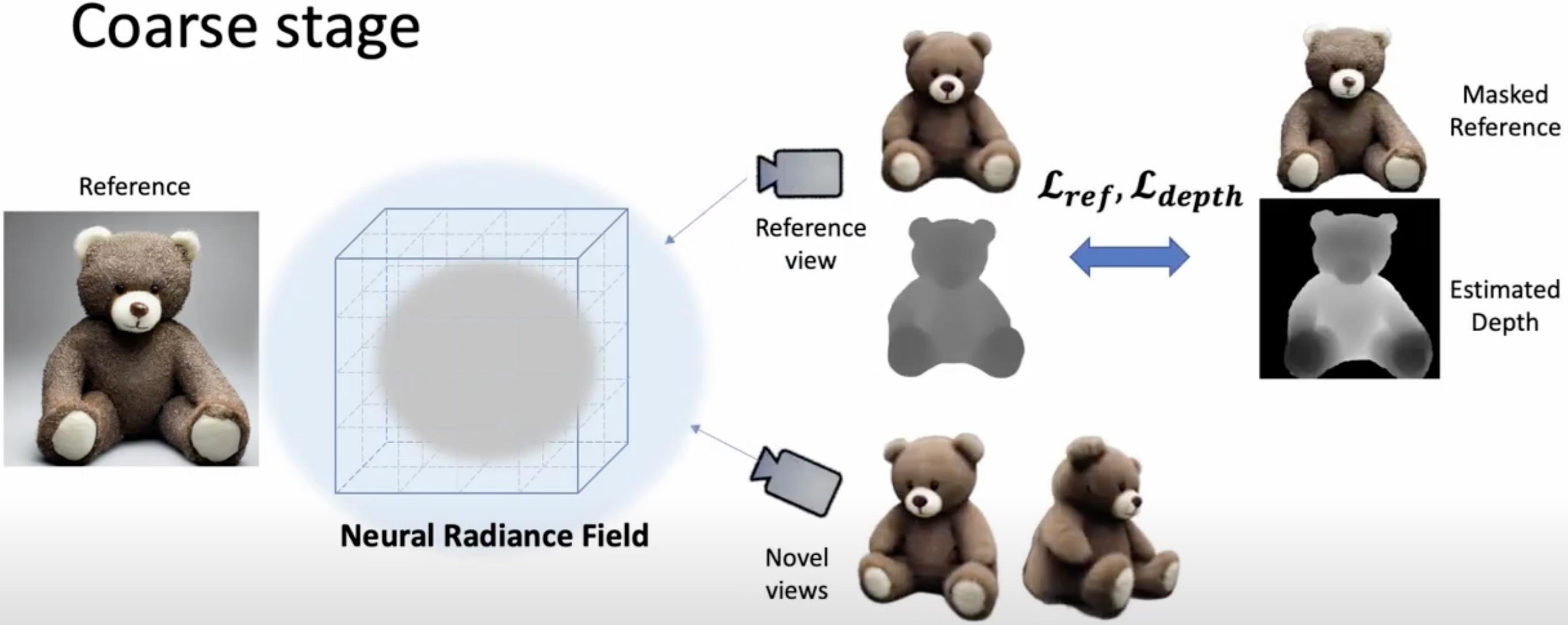
Reference view

$\mathcal{L}_{ref}, \mathcal{L}_{depth}$

Masked Reference

Estimated Depth

Coarse stage



Reference

Neural Radiance Field

Reference view

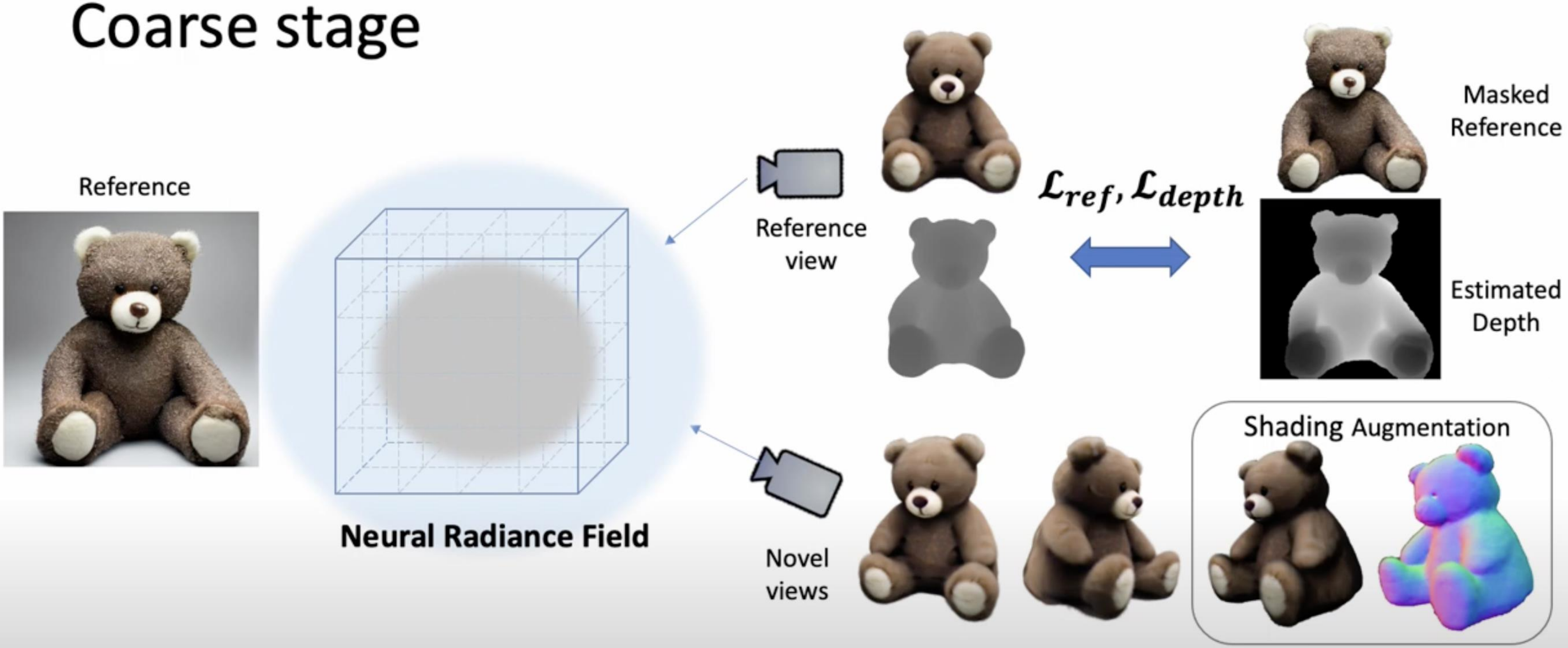
Novel views

$\mathcal{L}_{ref}, \mathcal{L}_{depth}$

Masked Reference

Estimated Depth

Coarse stage



Reference

Neural Radiance Field

Reference view

Novel views

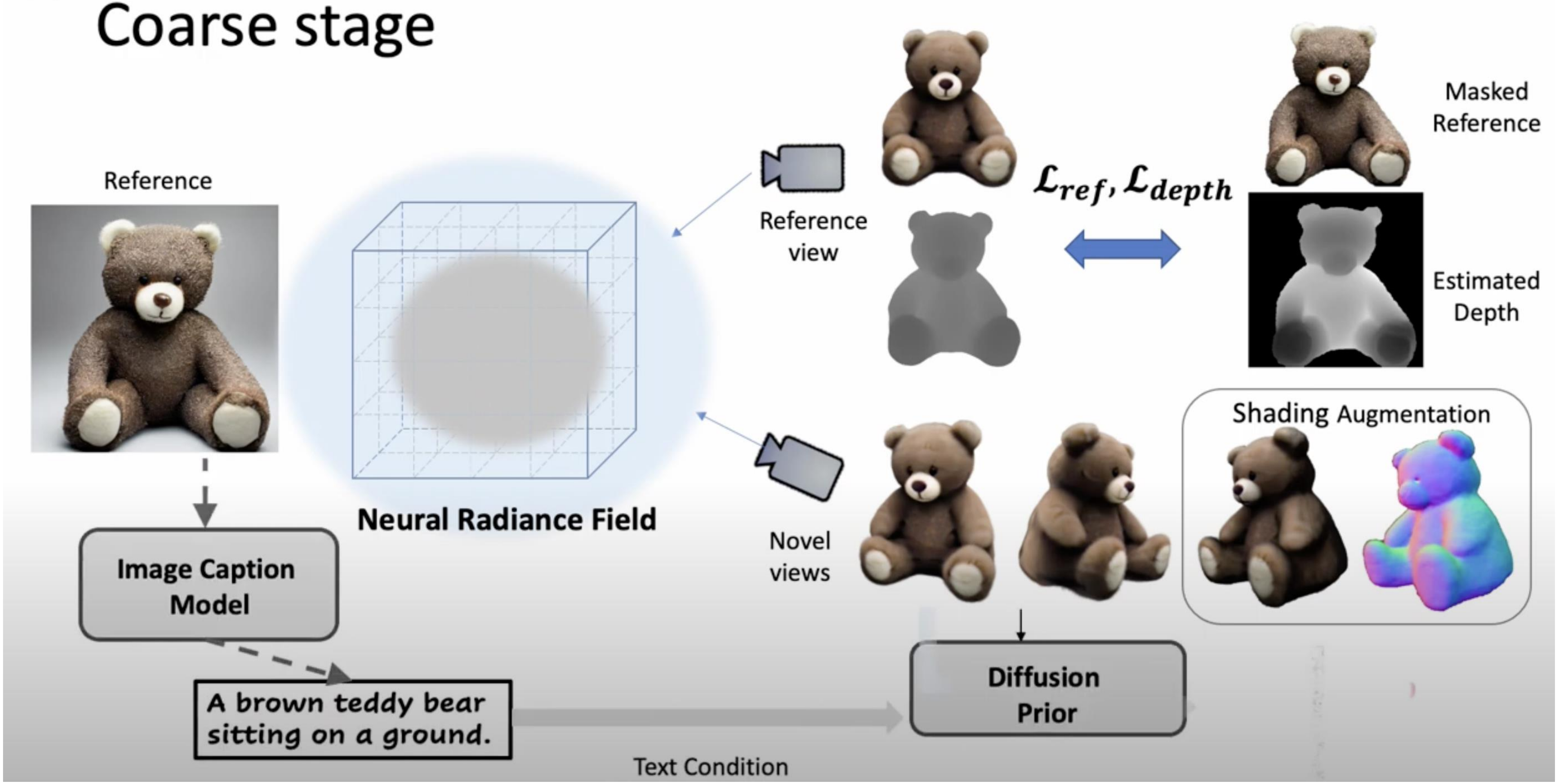
$\mathcal{L}_{ref}, \mathcal{L}_{depth}$

Masked Reference

Estimated Depth

Shading Augmentation

Coarse stage



Reference

Image Caption Model

A brown teddy bear sitting on a ground.

Text Condition

Neural Radiance Field

Reference view

$\mathcal{L}_{ref}, \mathcal{L}_{depth}$

Masked Reference

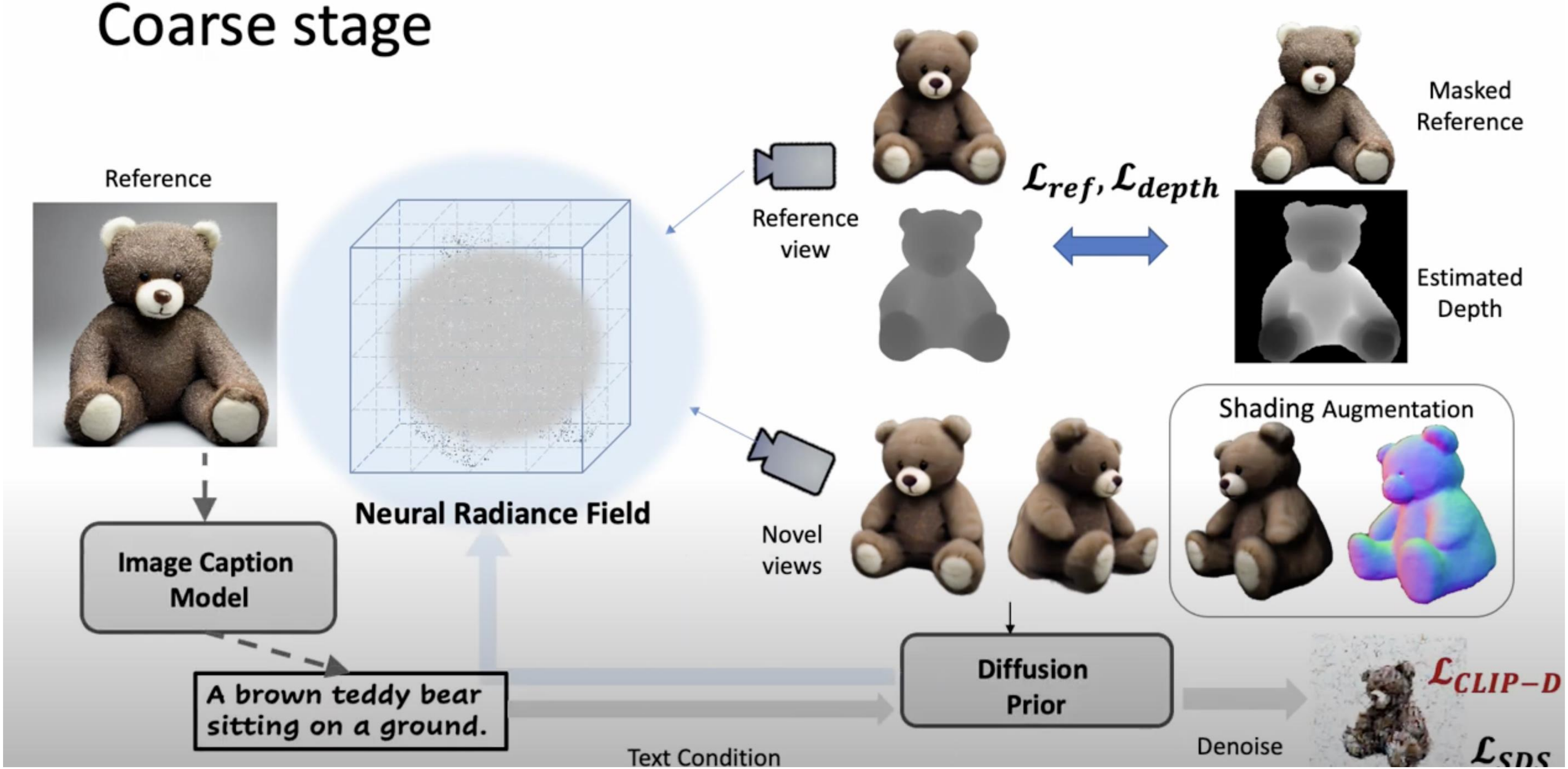
Estimated Depth

Novel views

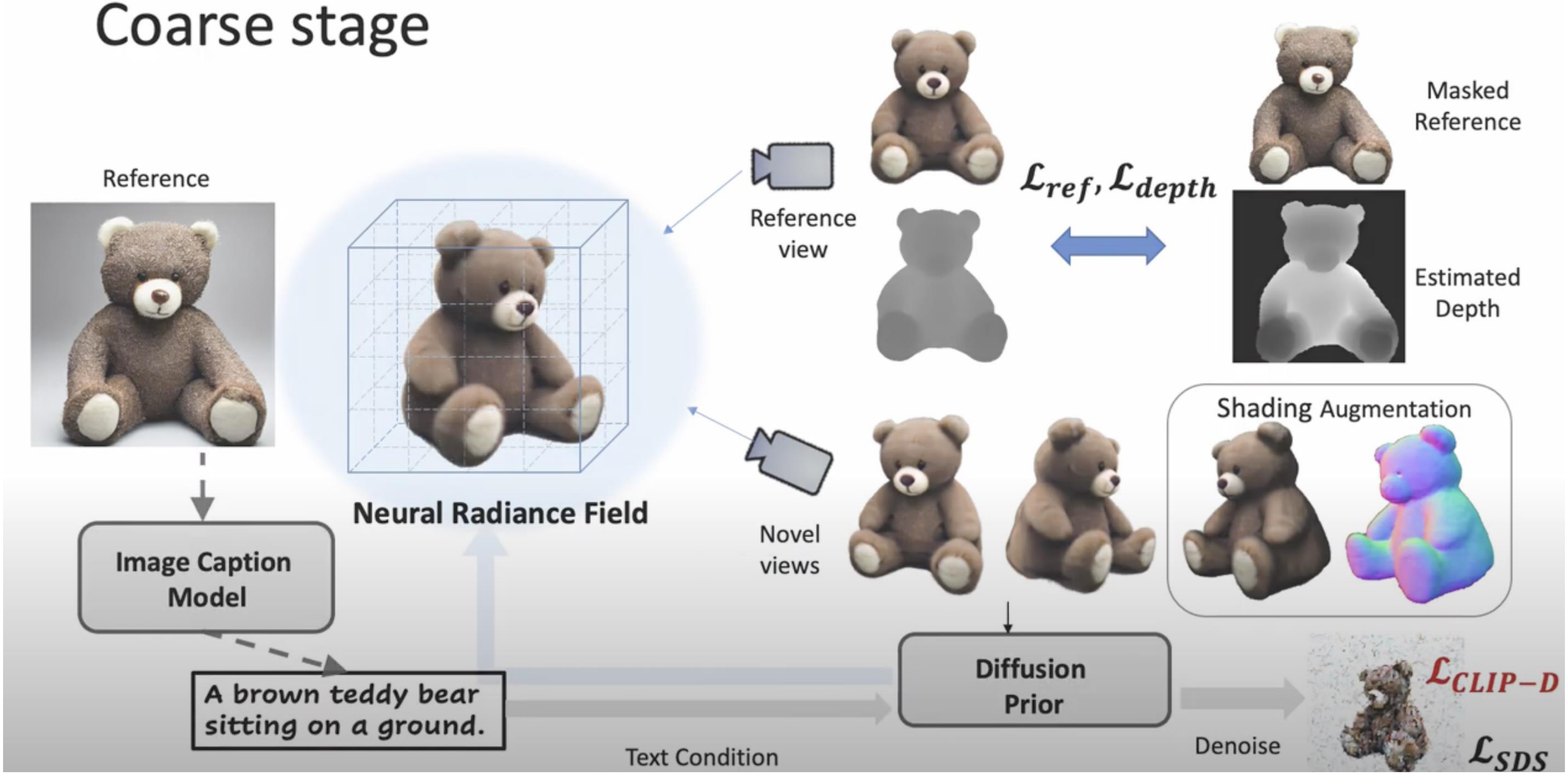
Shading Augmentation

Diffusion Prior

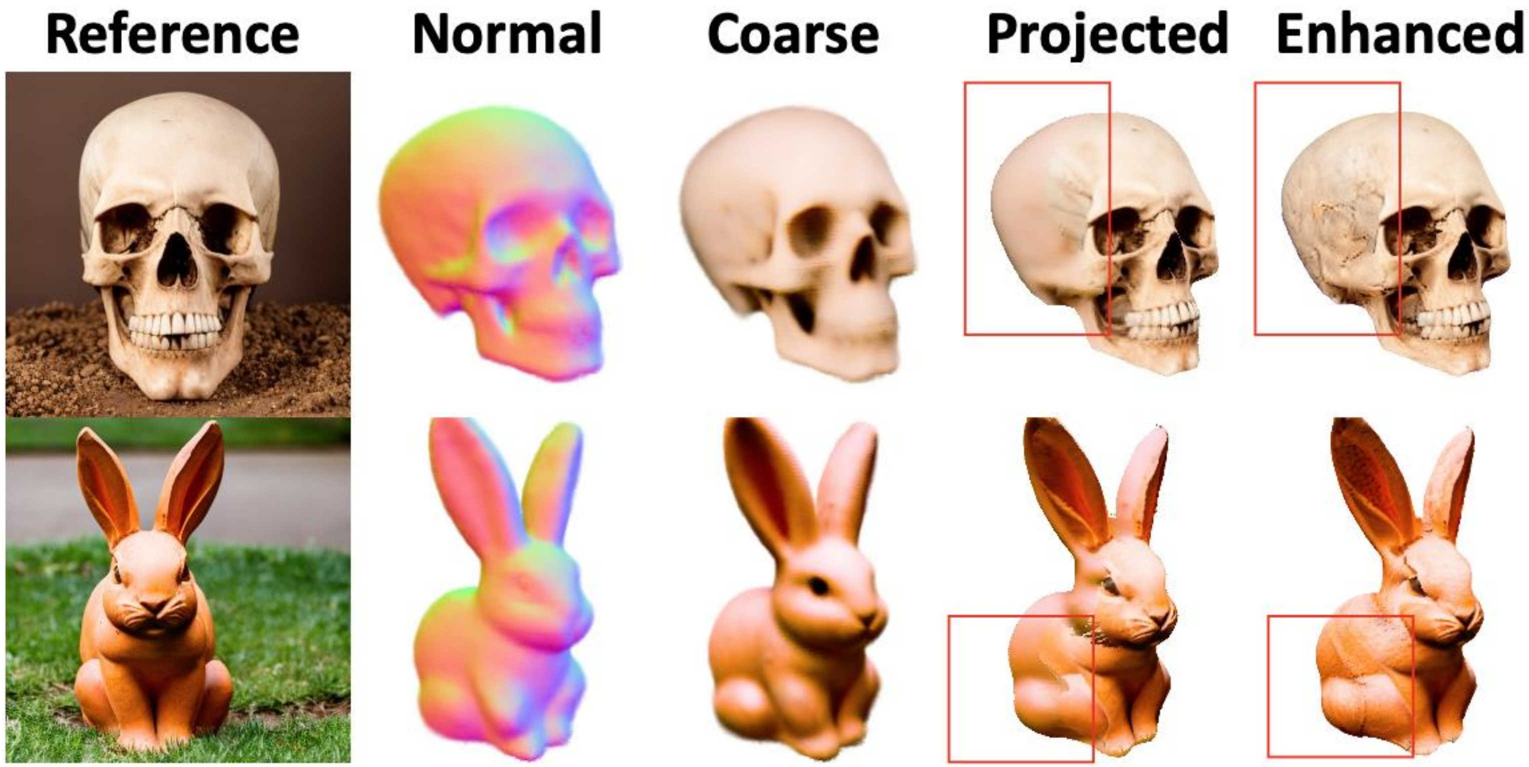
Coarse stage



Coarse stage



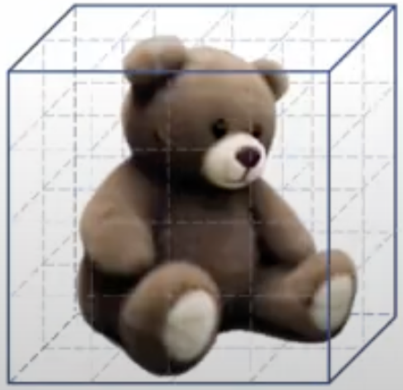
Refine Stage: Neural Texture Enhancement



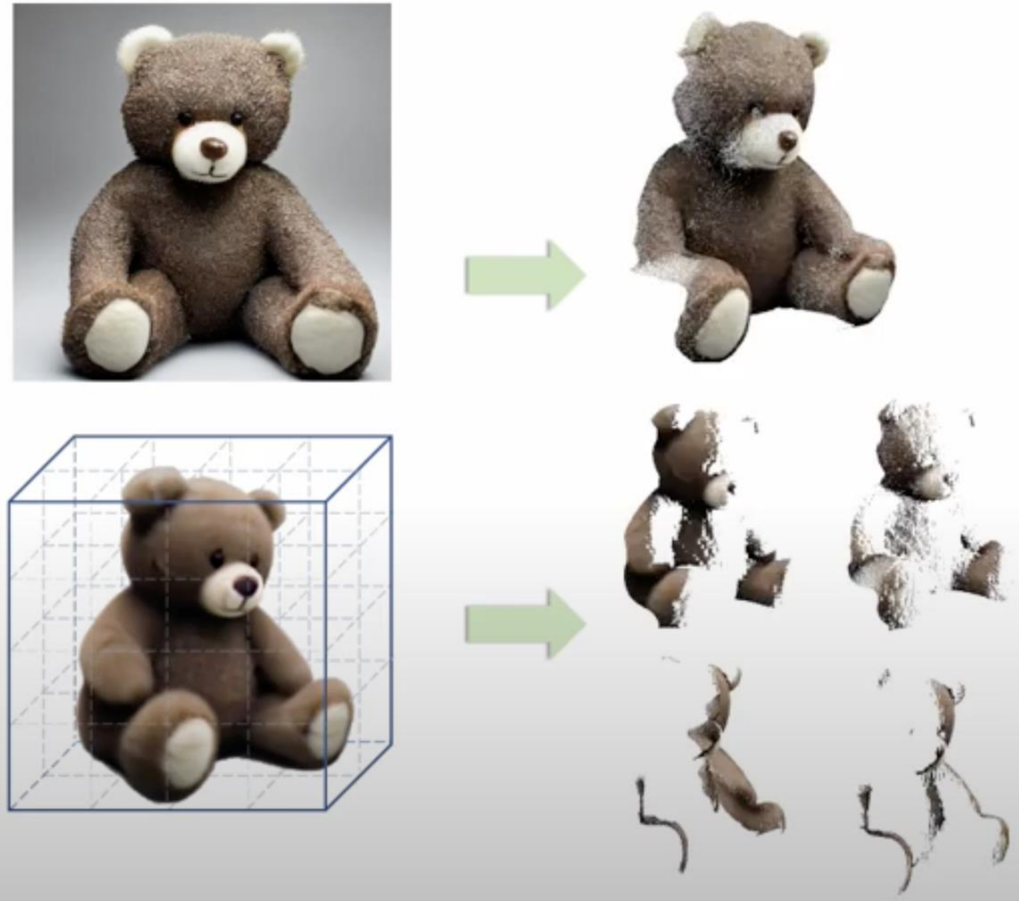
Key: certain pixels can be observable (novel & reference views)

Pipeline: Refine Stage

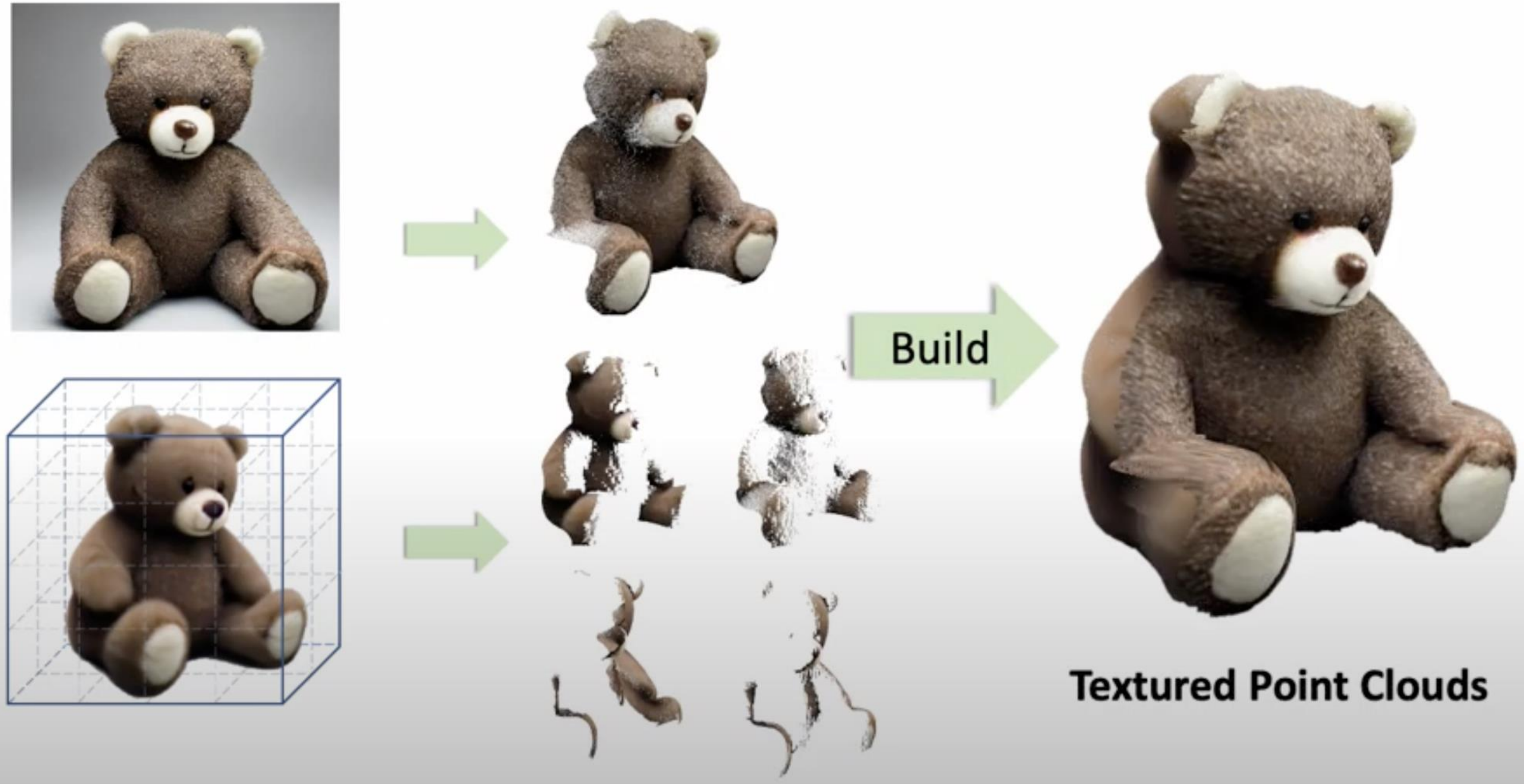
Refine stage

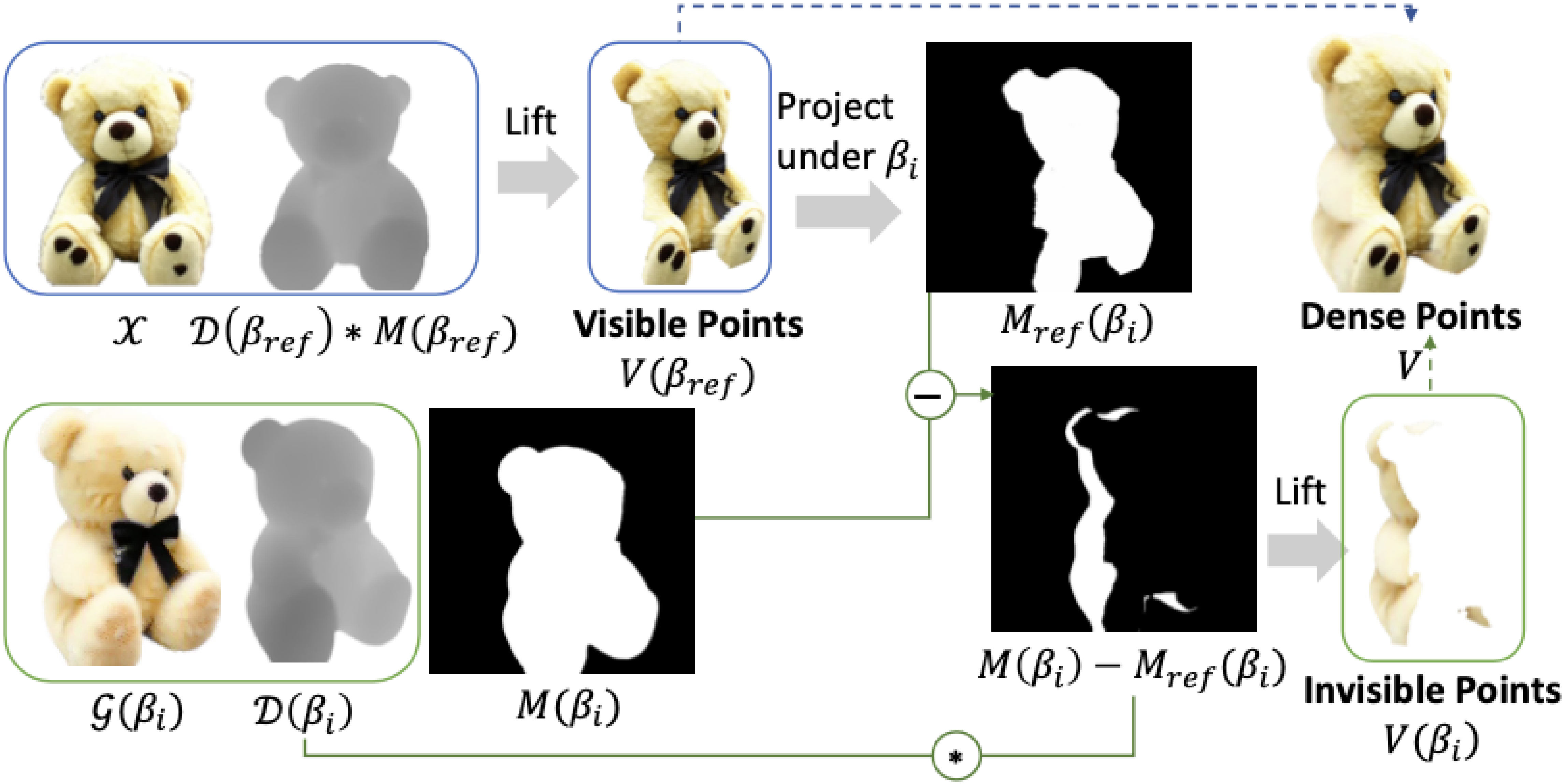


Refine stage



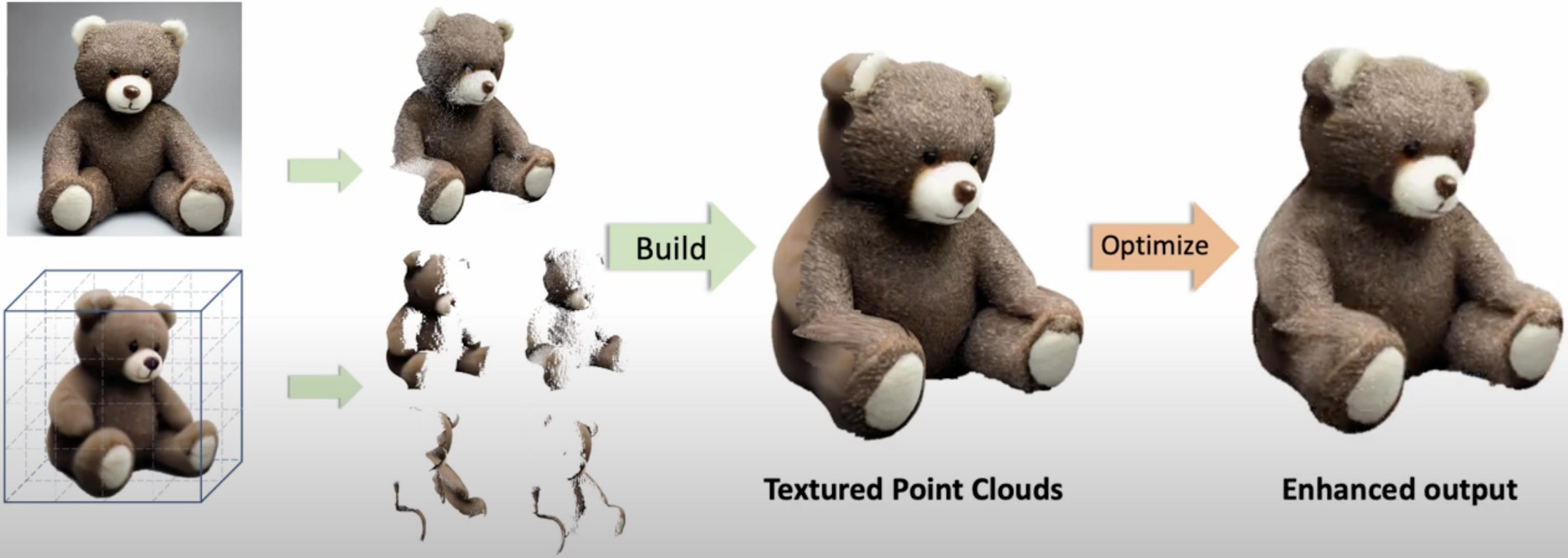
Refine stage





$$V(\beta_{\text{ref}}) = R_{\text{ref}} K^{-1} \mathcal{P}(\mathcal{D}(\beta_{\text{ref}}) * M(\beta_{\text{ref}}))),$$

Refine stage

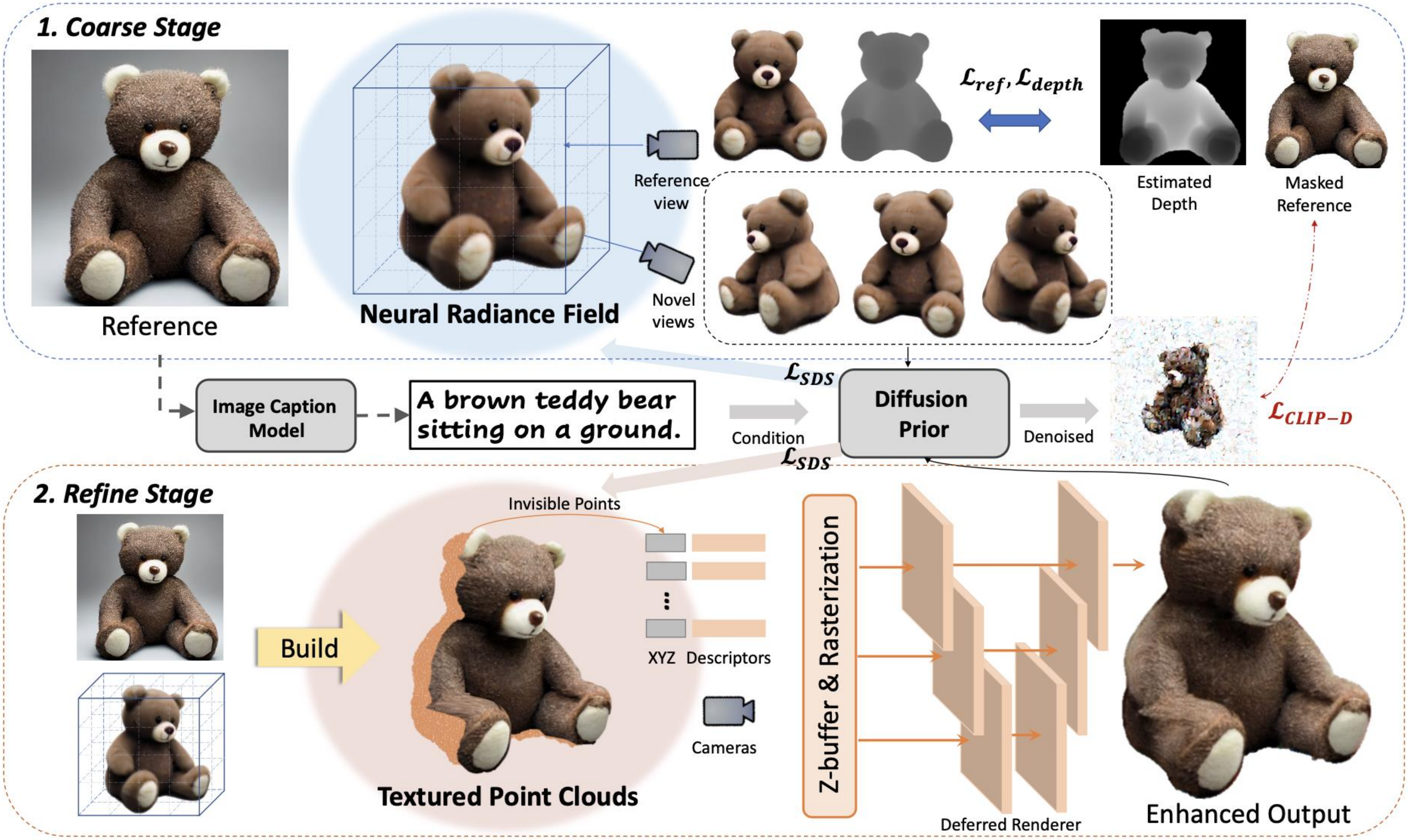


Build

Optimize

Textured Point Clouds

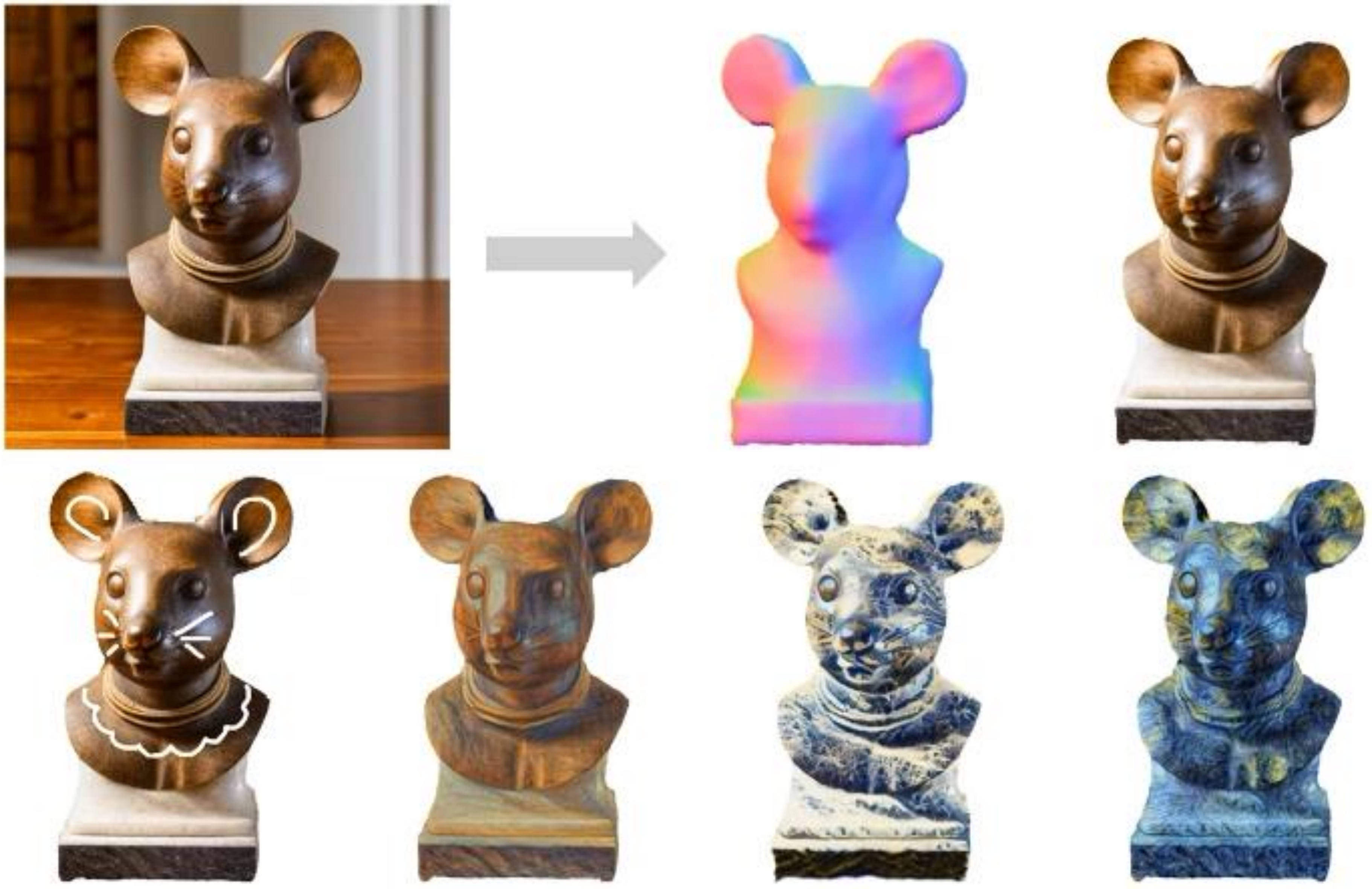
Enhanced output


1. Coarse Stage
Reference
Neural Radiance Field
Reference view
Novel views
$\mathcal{L}_{ref}, \mathcal{L}_{depth}$
Estimated Depth
Masked Reference
$\mathcal{L}_{SDS}$
Image Caption Model
A brown teddy bear sitting on a ground.
Condition
Diffusion Prior
Denoised
$\mathcal{L}_{CLIP-D}$
$\mathcal{L}_{SDS}$
2. Refine Stage
Build
Invisible Points
XYZ
Descriptors
Cameras
Textured Point Clouds
Z-buffer & Rasterization
Deferred Renderer
Enhanced Output

Applications

Diverse Text to 3D

Texture Modification

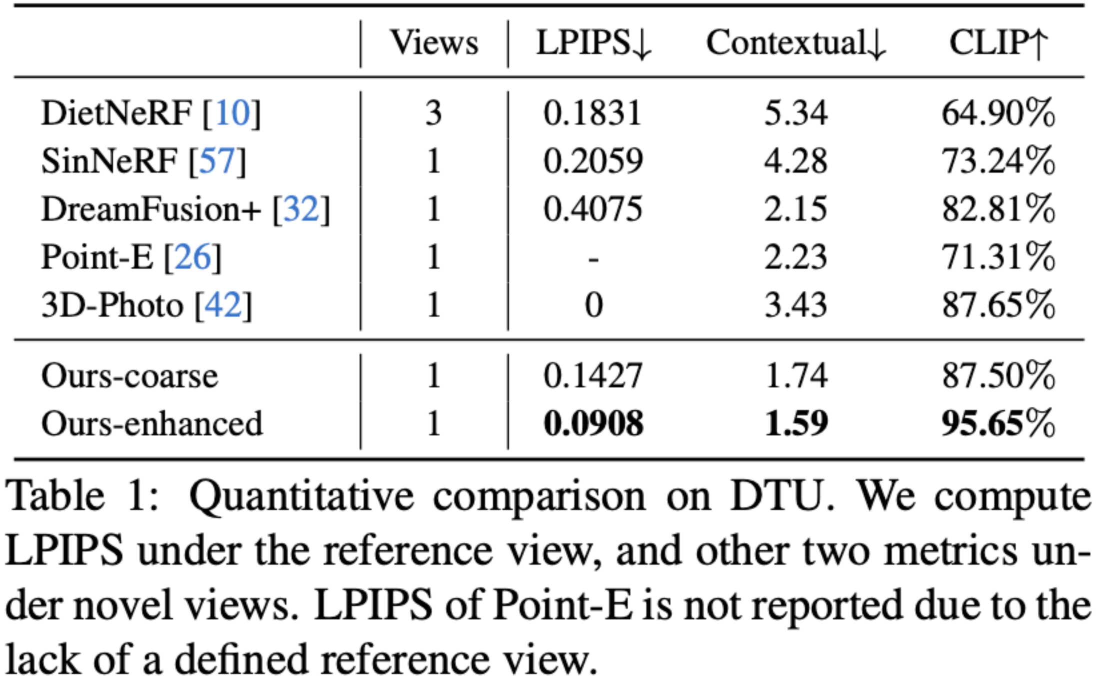


Experimental Results

reconstruction quality at the reference view

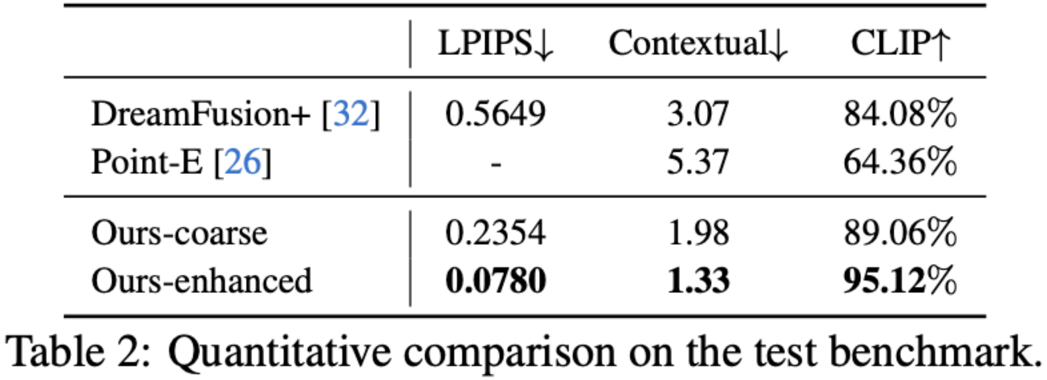
pixel-level similarity between novel-view rendering and the reference

semantic similarity between the novel view and the reference



	Views	LPIPS↓	Contextual↓	CLIP↑
DietNeRF [10]	3	0.1831	5.34	64.90%
SinNeRF [57]	1	0.2059	4.28	73.24%
DreamFusion+ [32]	1	0.4075	2.15	82.81%
Point-E [26]	1	-	2.23	71.31%
3D-Photo [42]	1	0	3.43	87.65%
Ours-coarse	1	0.1427	1.74	87.50%
Ours-enhanced	1	**0.0908**	**1.59**	**95.65%**

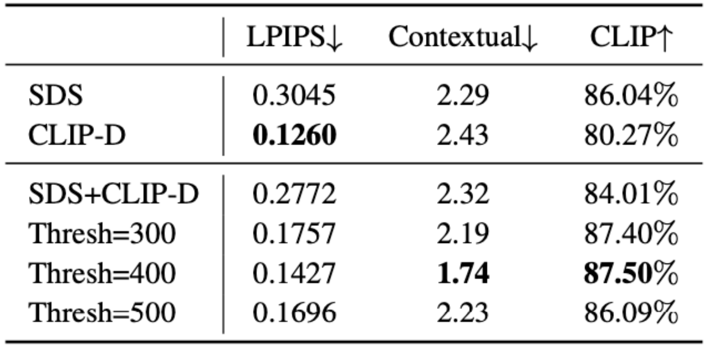
Table 1: Quantitative comparison on DTU. We compute LPIPS under the reference view, and other two metrics under novel views. LPIPS of Point-E is not reported due to the lack of a defined reference view.



	LPIPS↓	Contextual↓	CLIP↑
DreamFusion+ [32]	0.5649	3.07	84.08%
Point-E [26]	-	5.37	64.36%
Ours-coarse	0.2354	1.98	89.06%
Ours-enhanced	**0.0780**	**1.33**	**95.12%**

Table 2: Quantitative comparison on the test benchmark.

Experimental Results



	LPIPS↓	Contextual↓	CLIP↑
SDS	0.3045	2.29	86.04%
CLIP-D	**0.1260**	2.43	80.27%
SDS+CLIP-D	0.2772	2.32	84.01%
Thresh=300	0.1757	2.19	87.40%
Thresh=400	0.1427	**1.74**	**87.50%**
Thresh=500	0.1696	2.23	86.09%

Table 3: Ablation study on SDS and CLIP-D loss on the test benchmark. We compute LPIPS under the reference view, and the other two metrics under novel views. “Thresh” denotes the boundary of time steps using SDS or CLIP-D in the denoising process.

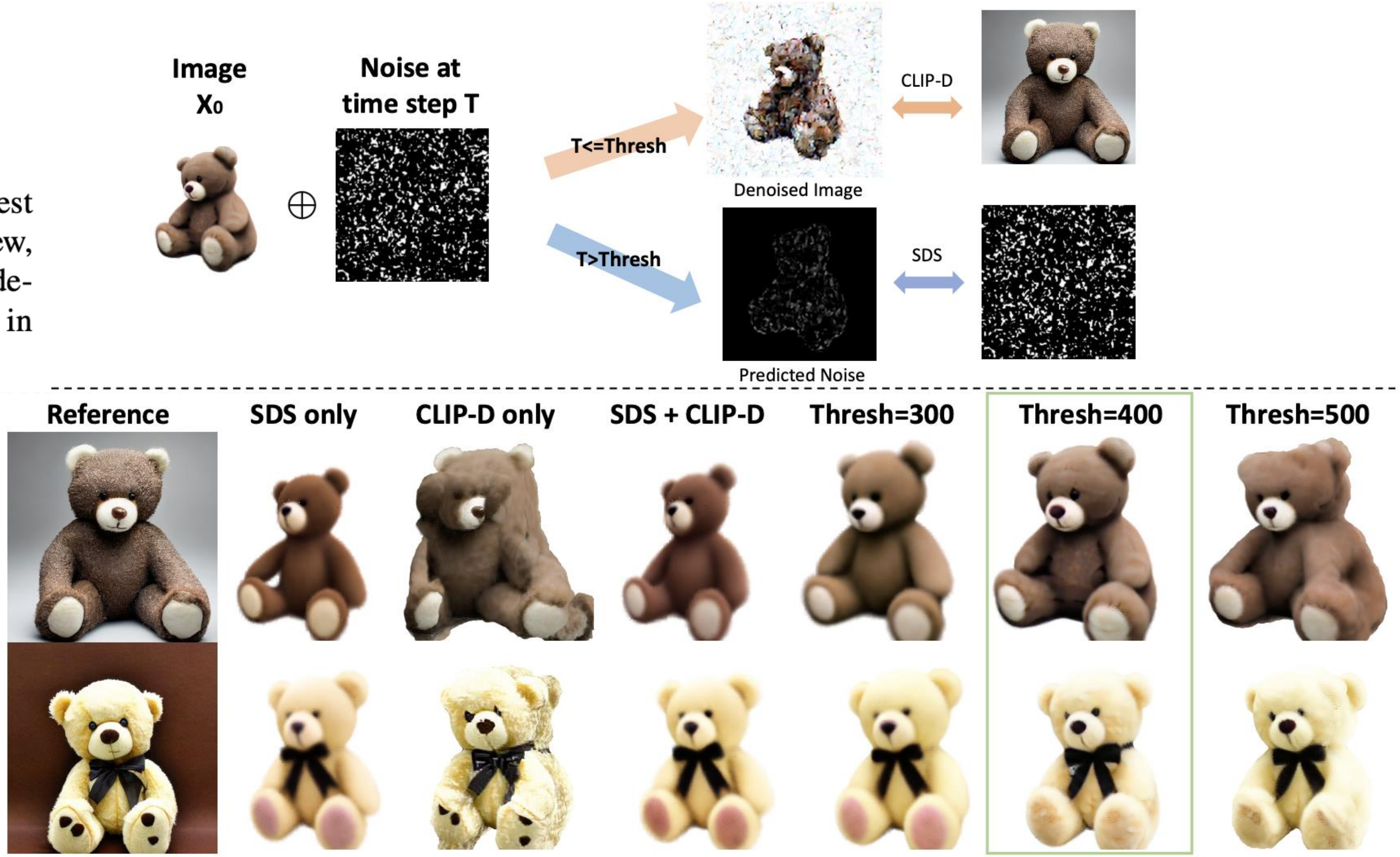


Figure 12: Analysis of SDS and CLIP-D loss.

Experimental Results

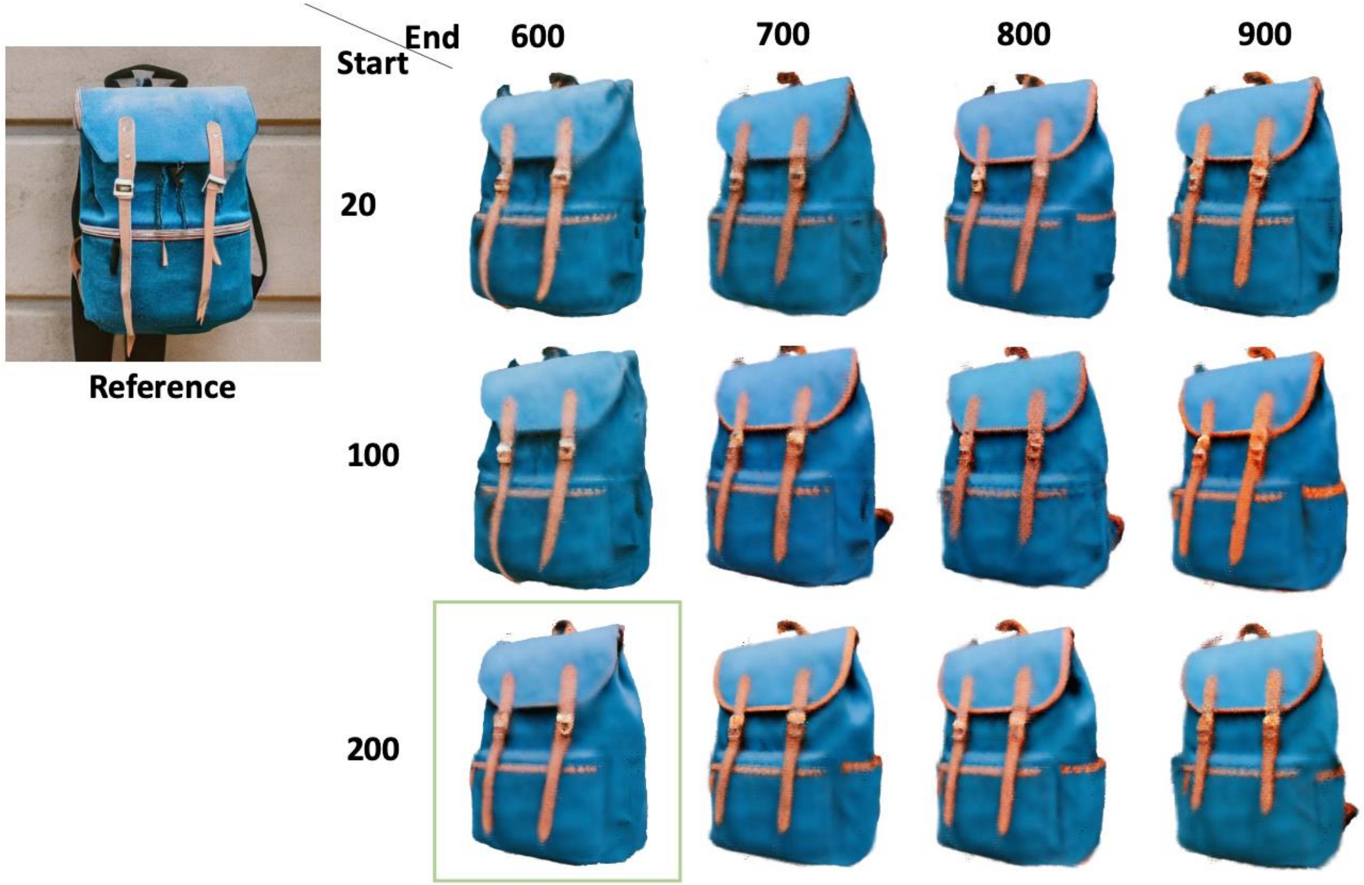


Figure 13: Analysis of the time step range in SDS process. We visualize novel view results in the coarse stage that are trained with different time step ranges (from start to end).

Experimental Results

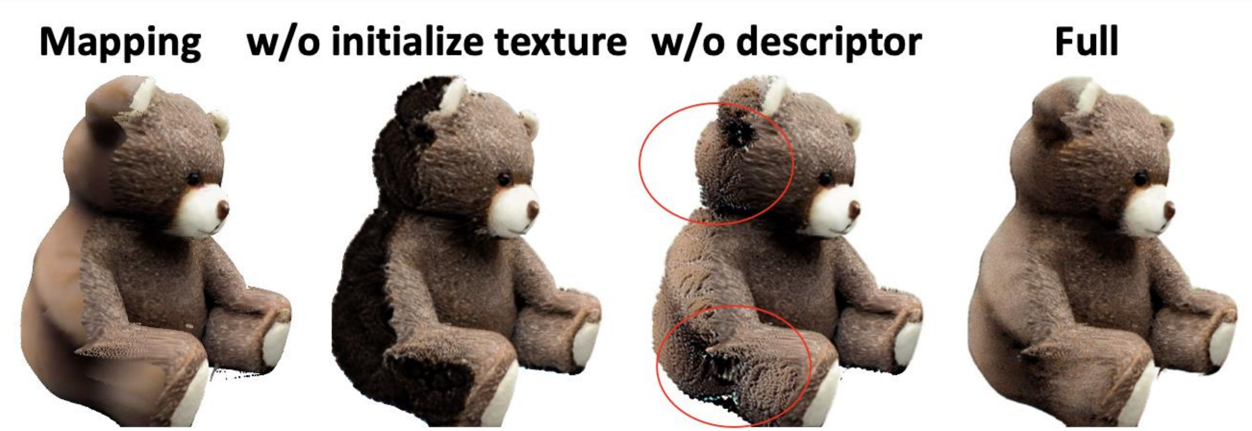


Figure 14: Analysis of texture initialization and point descriptors.

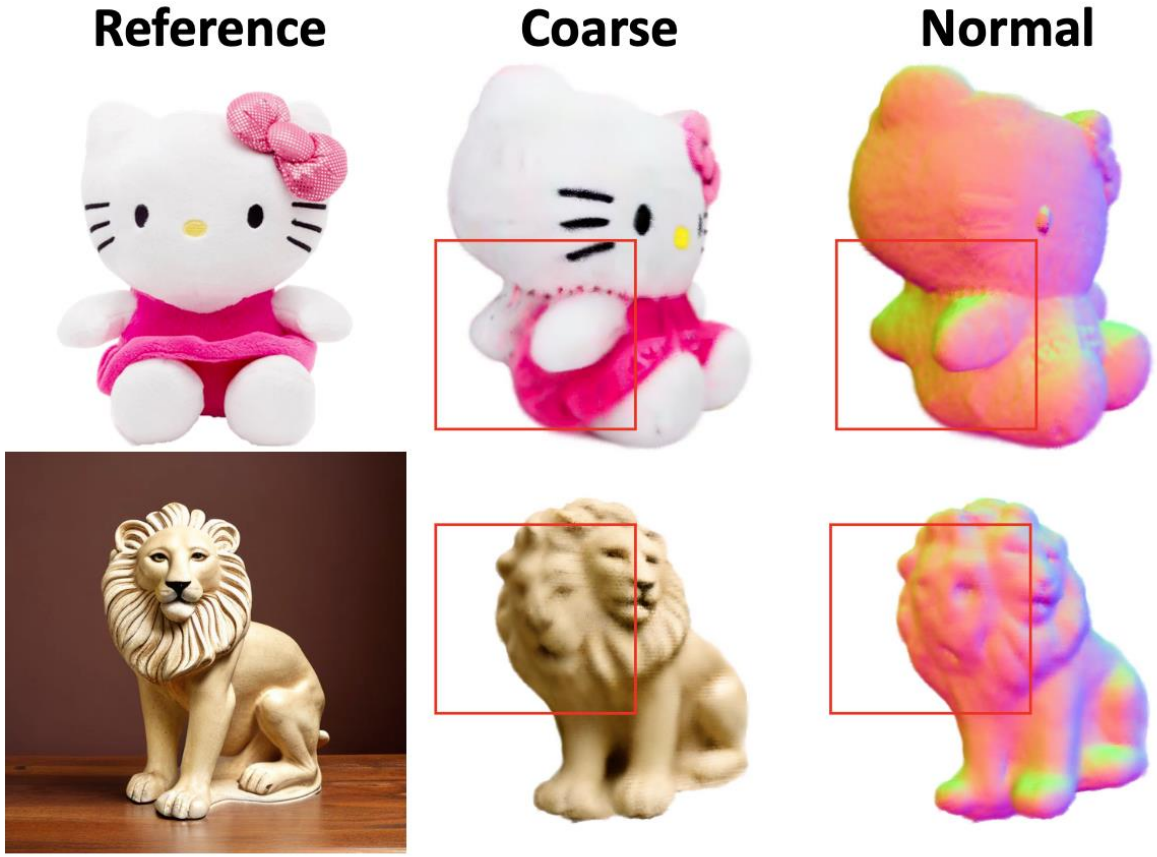


Figure 15: Failure cases due to the geometry ambiguity.

Experimental Results

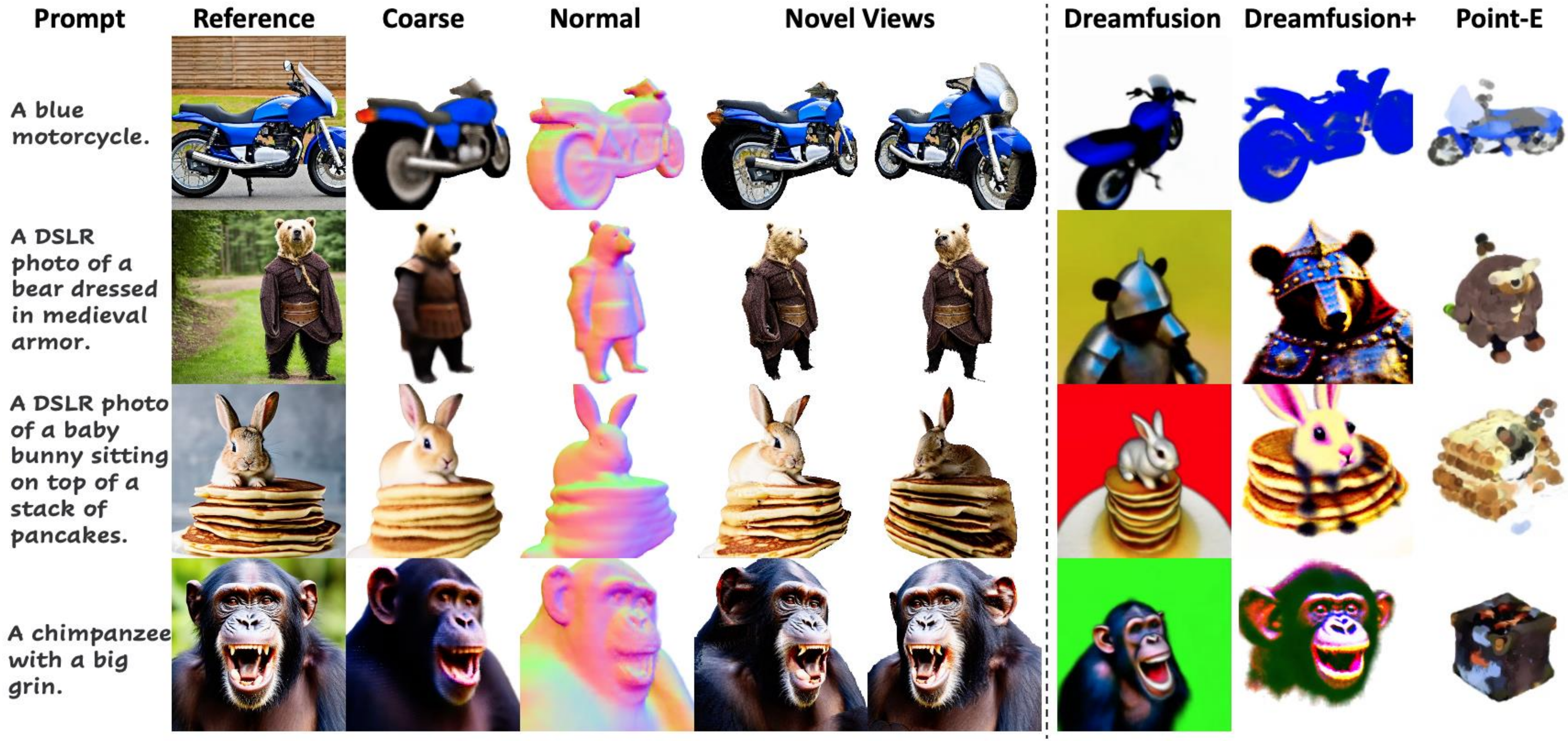

Prompt
Reference
Coarse
Normal
Novel Views
Dreamfusion
Dreamfusion+
Point-E
A blue motorcycle.
A DSLR photo of a bear dressed in medieval armor.
A DSLR photo of a baby bunny sitting on top of a stack of pancakes.
A chimpanzee with a big grin.

Figure 7: Qualitative comparison on the test benchmark with two diffusion-based 3D content creation models, Dreamfusion and Point-E. We show our results with high-fidelity geometry and texture. The results of Dreamfusion are from its website.

Experimental Results

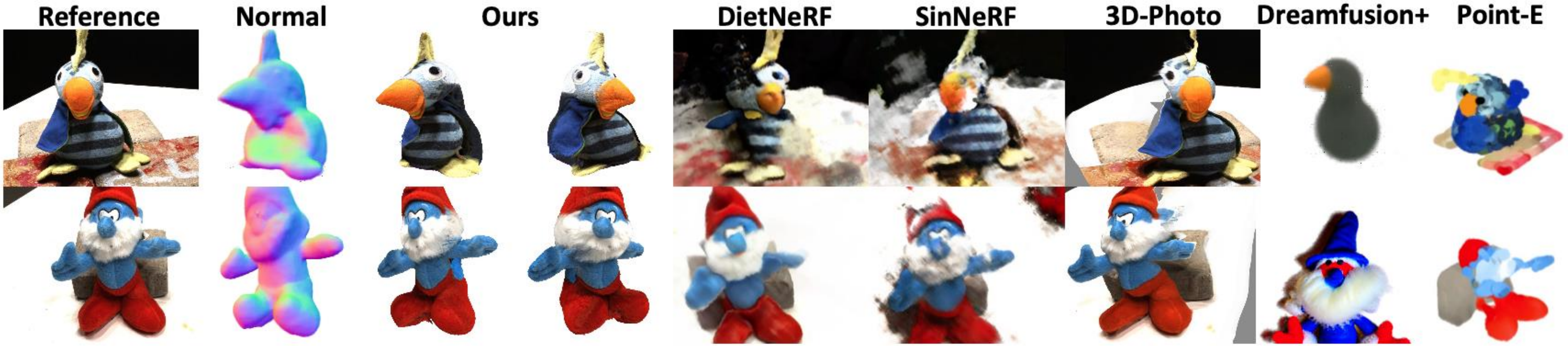


Figure 8: Qualitative comparison of novel view synthesis on DTU with state of the arts. Our method generates sharper and more plausible details in both geometry and texture.

Experimental Results

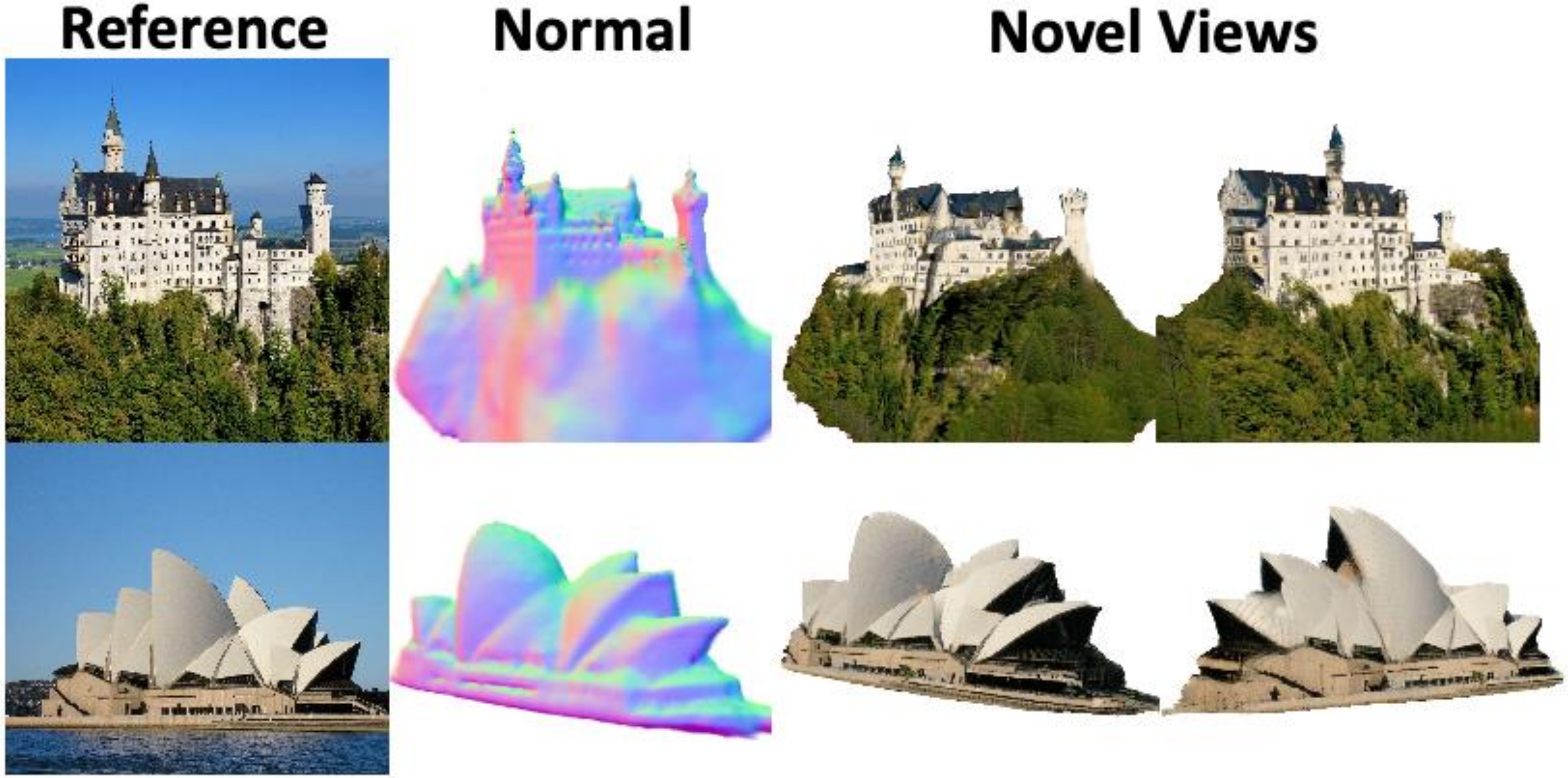


Figure 9: *Make-It-3D* enables high-fidelity 3D creation on real complex scenes.

Experimental Results

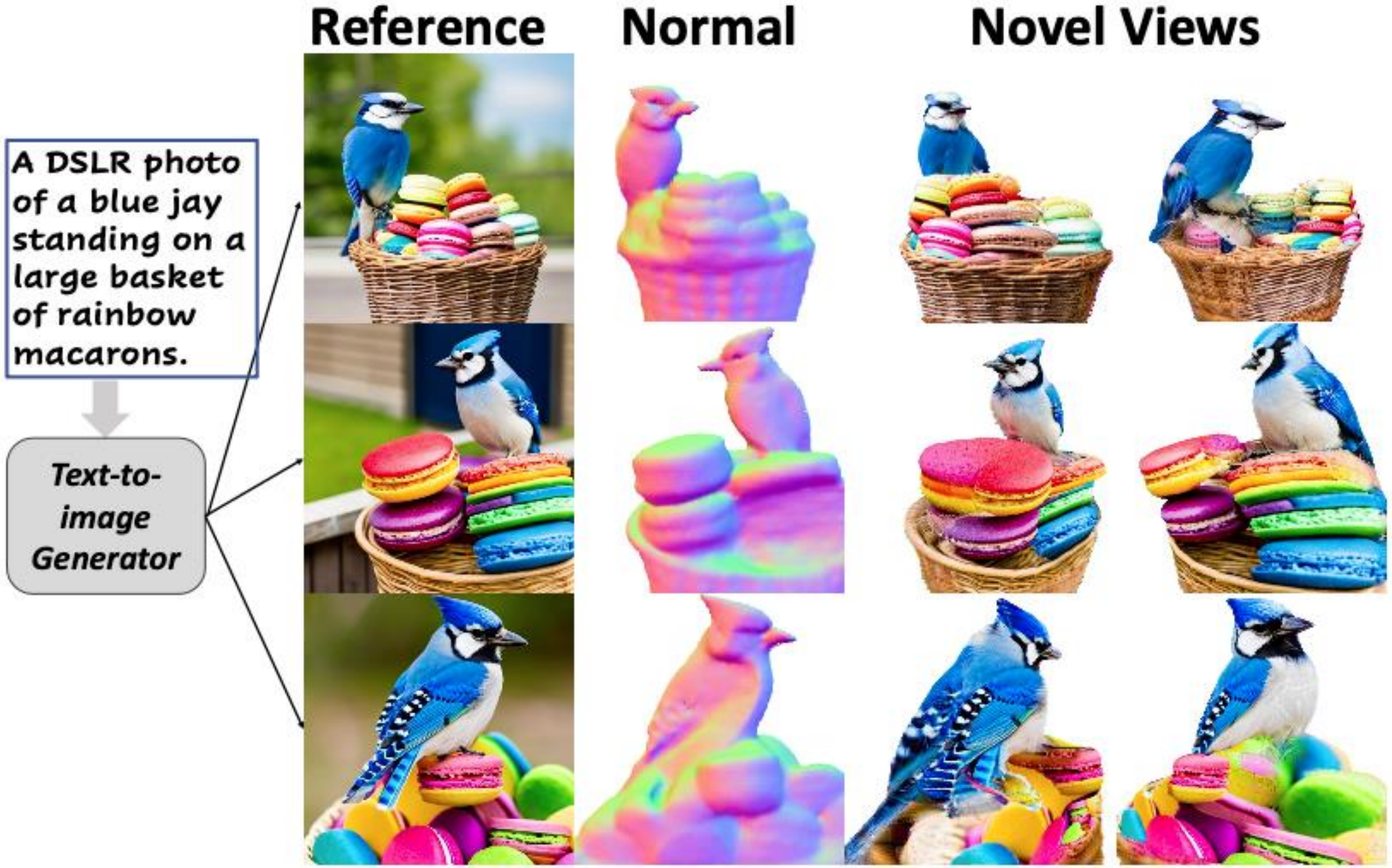

Reference
Normal
Novel Views
A DSLR photo of a blue jay standing on a large basket of rainbow macarons.
Text-to-image Generator

Figure 10: *Make-It-3D* generates diverse and visually stunning 3D models given a text description.

Experimental Results

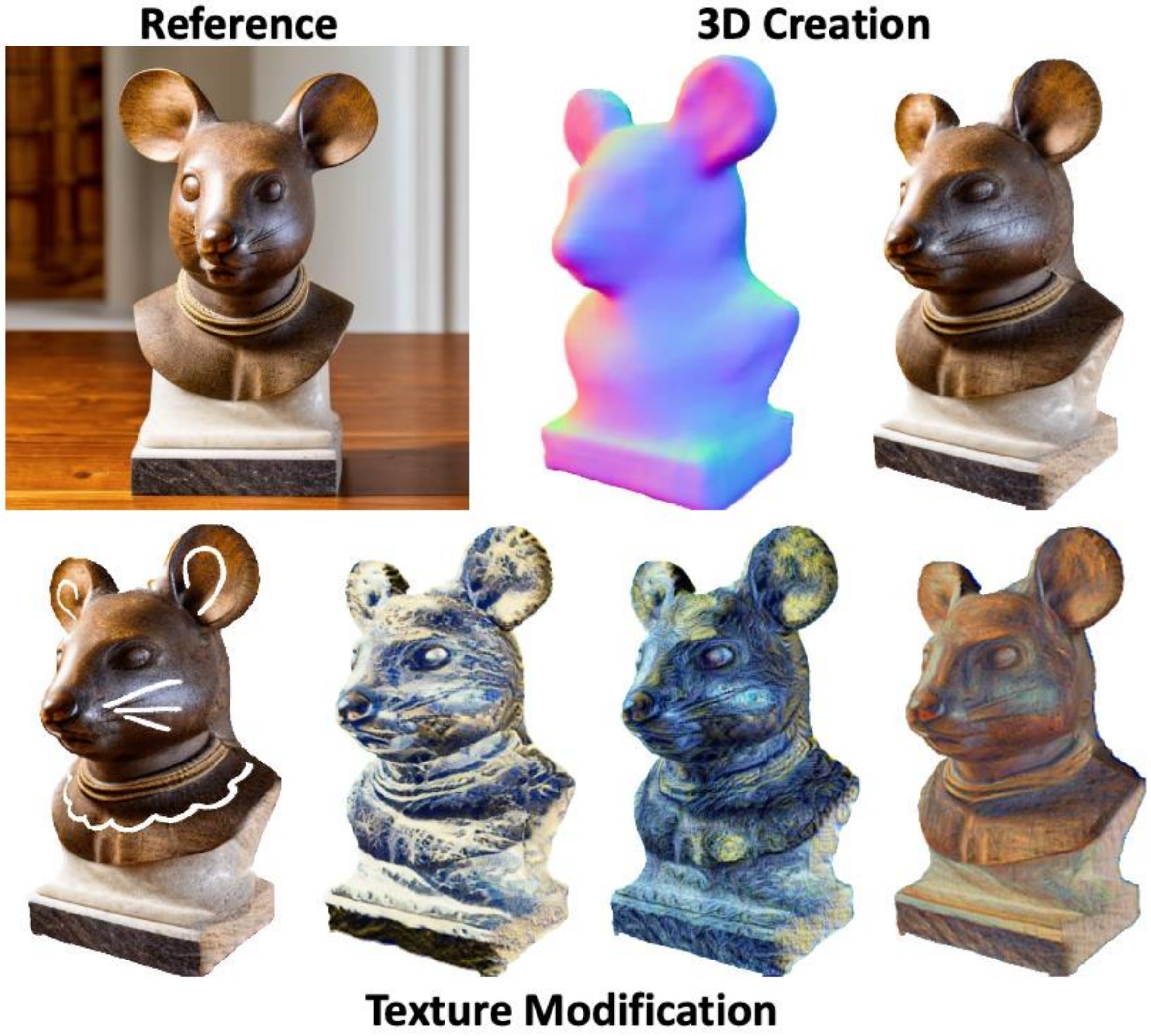
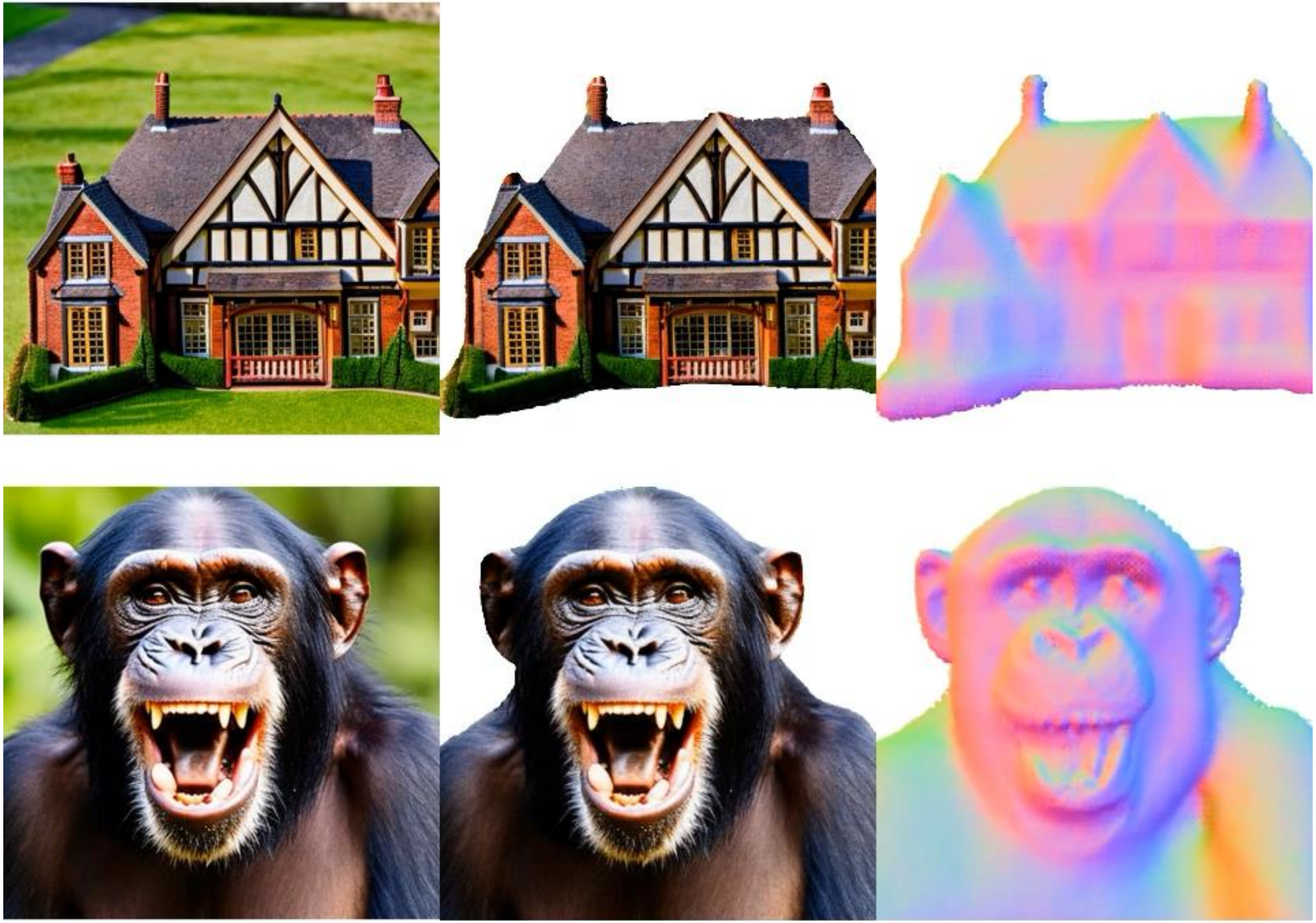
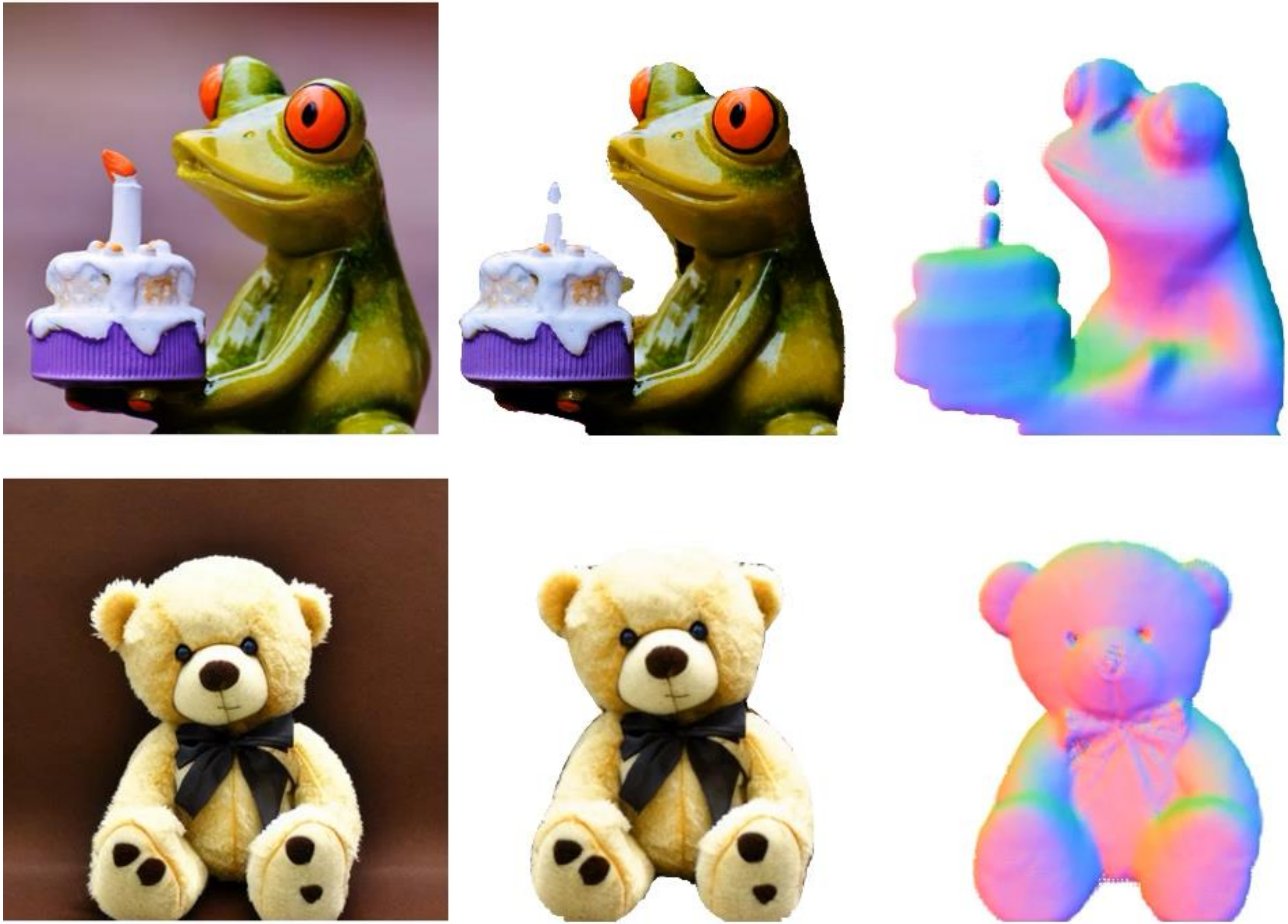
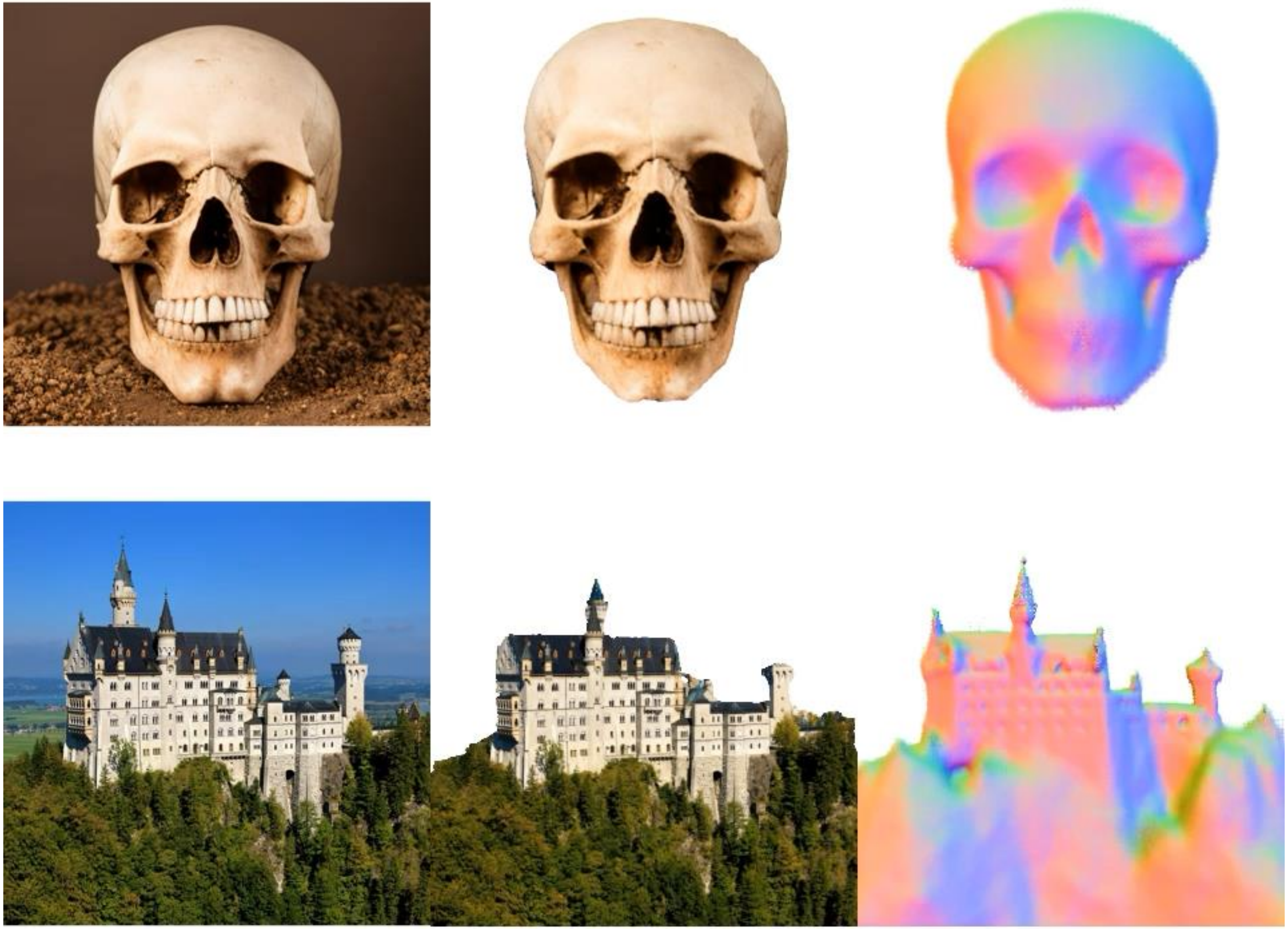
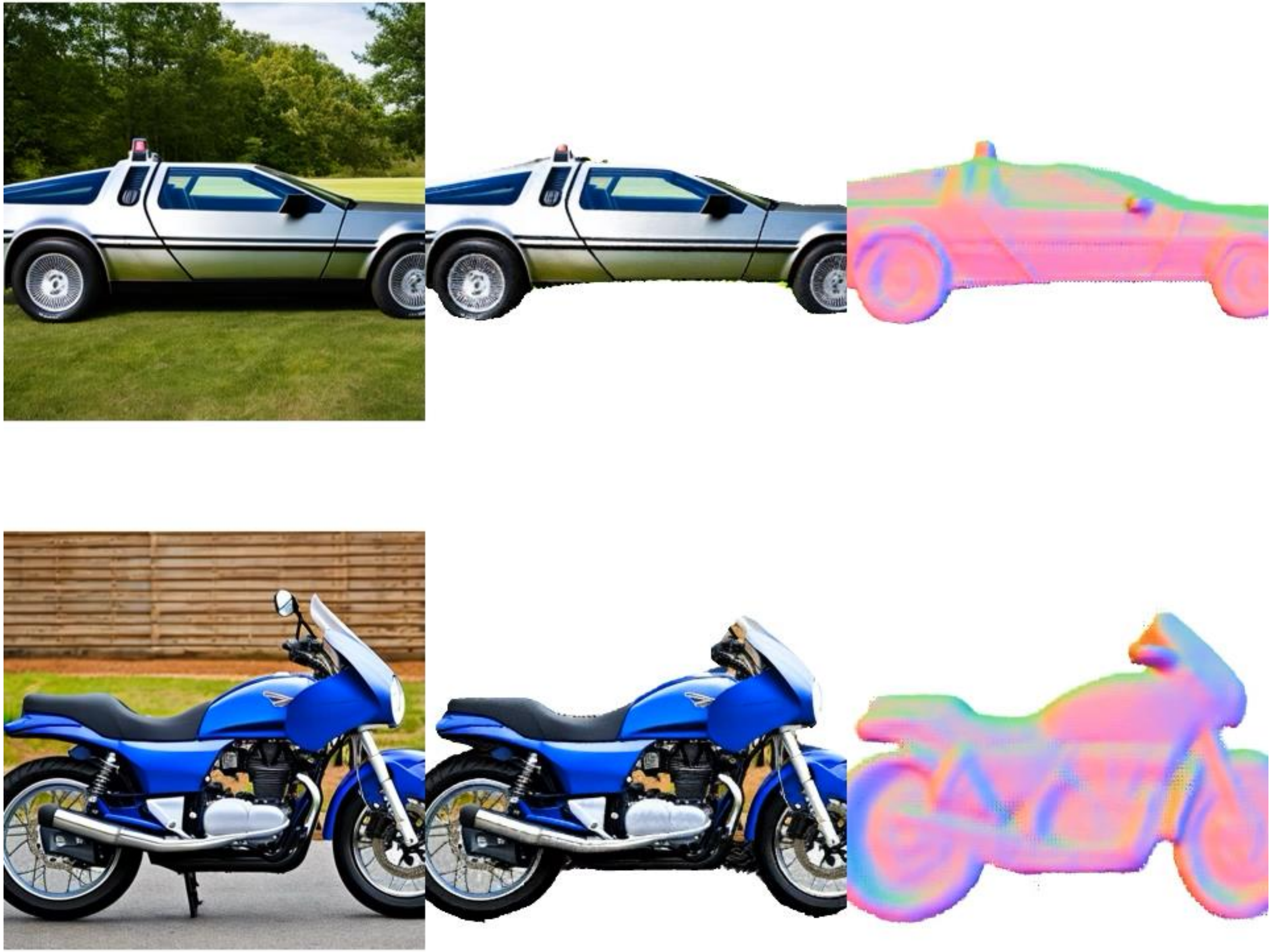


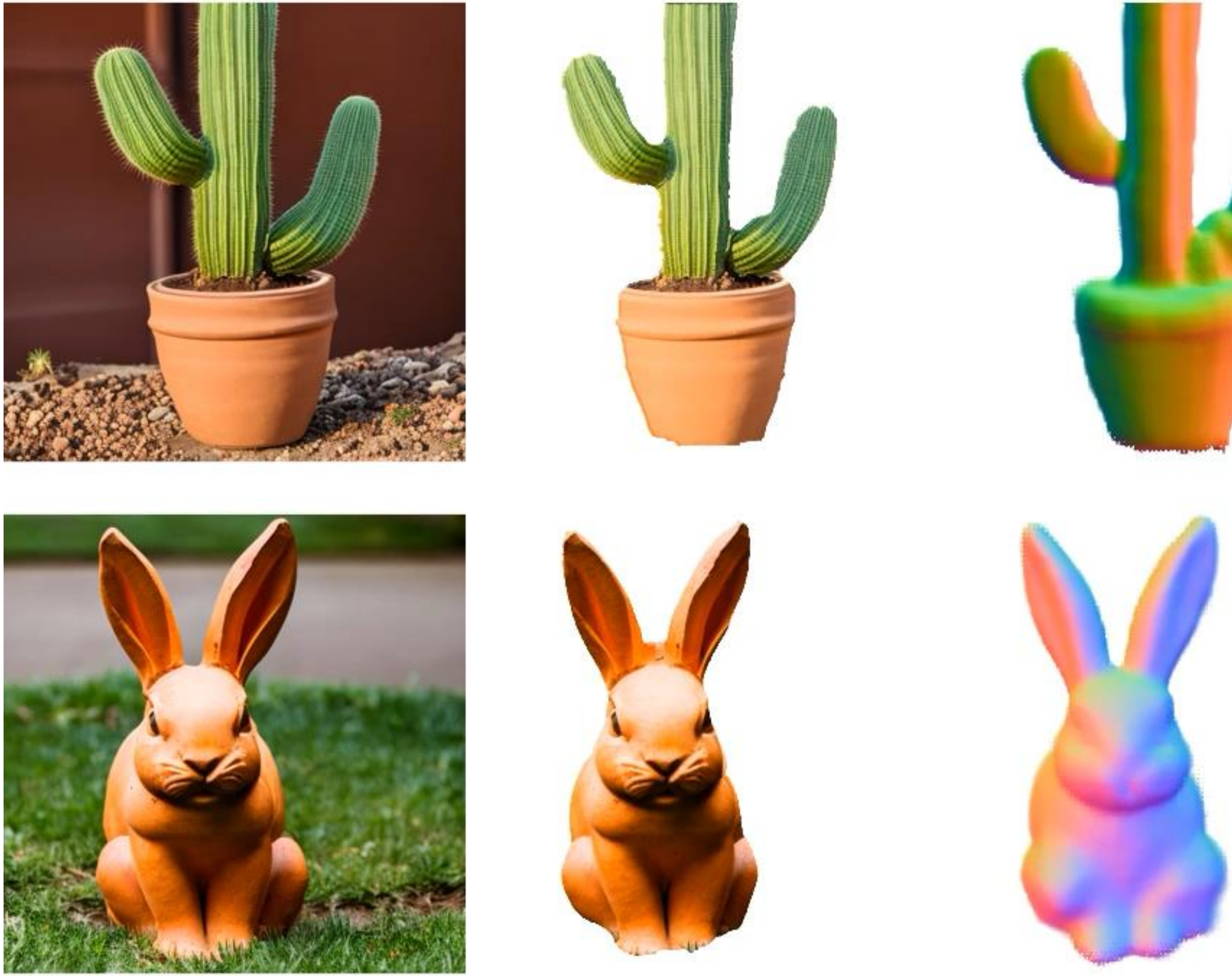
Figure 11: *Make-It-3D* achieves 3D-aware texture modification such as tattoo drawing and stylization.

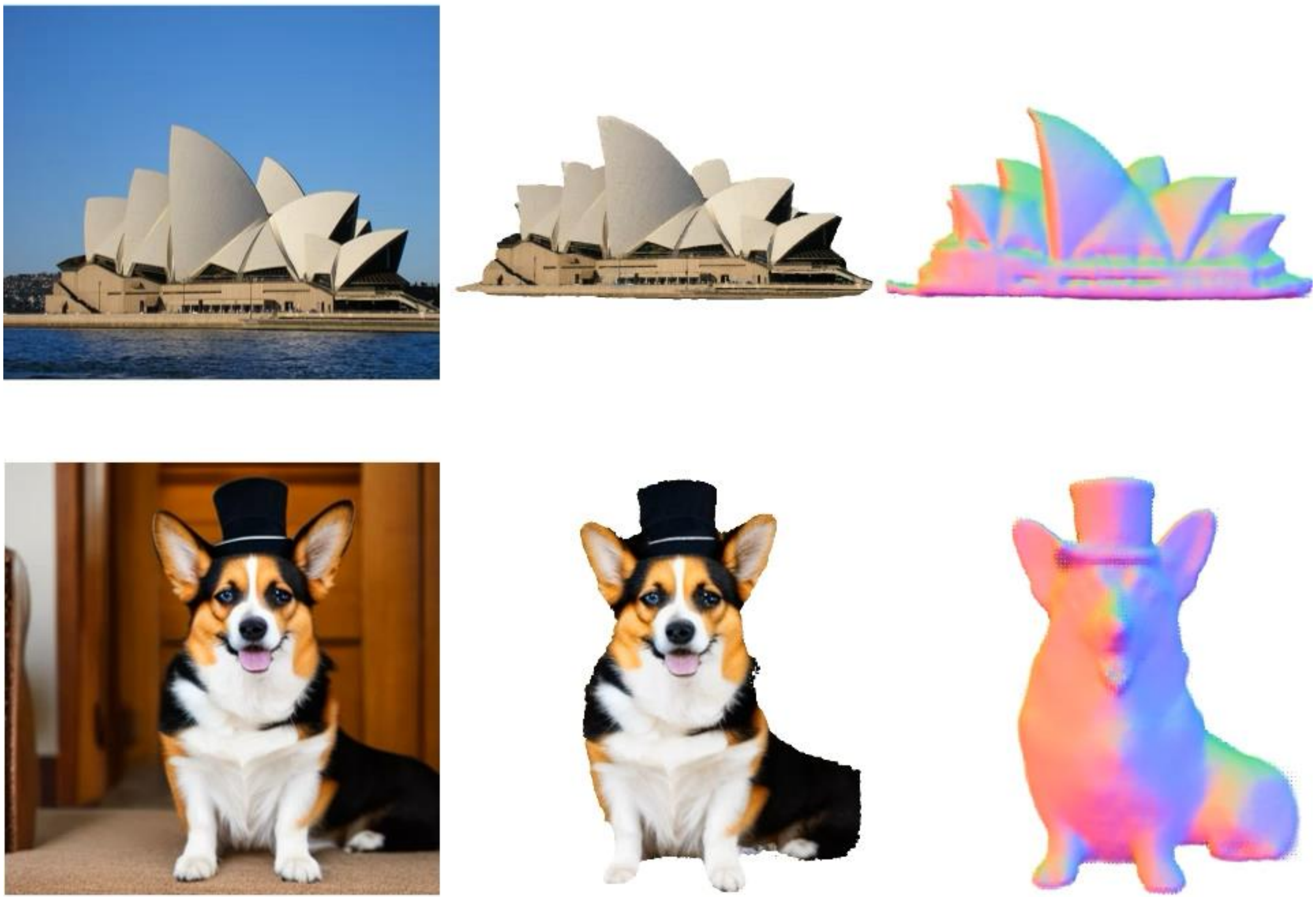


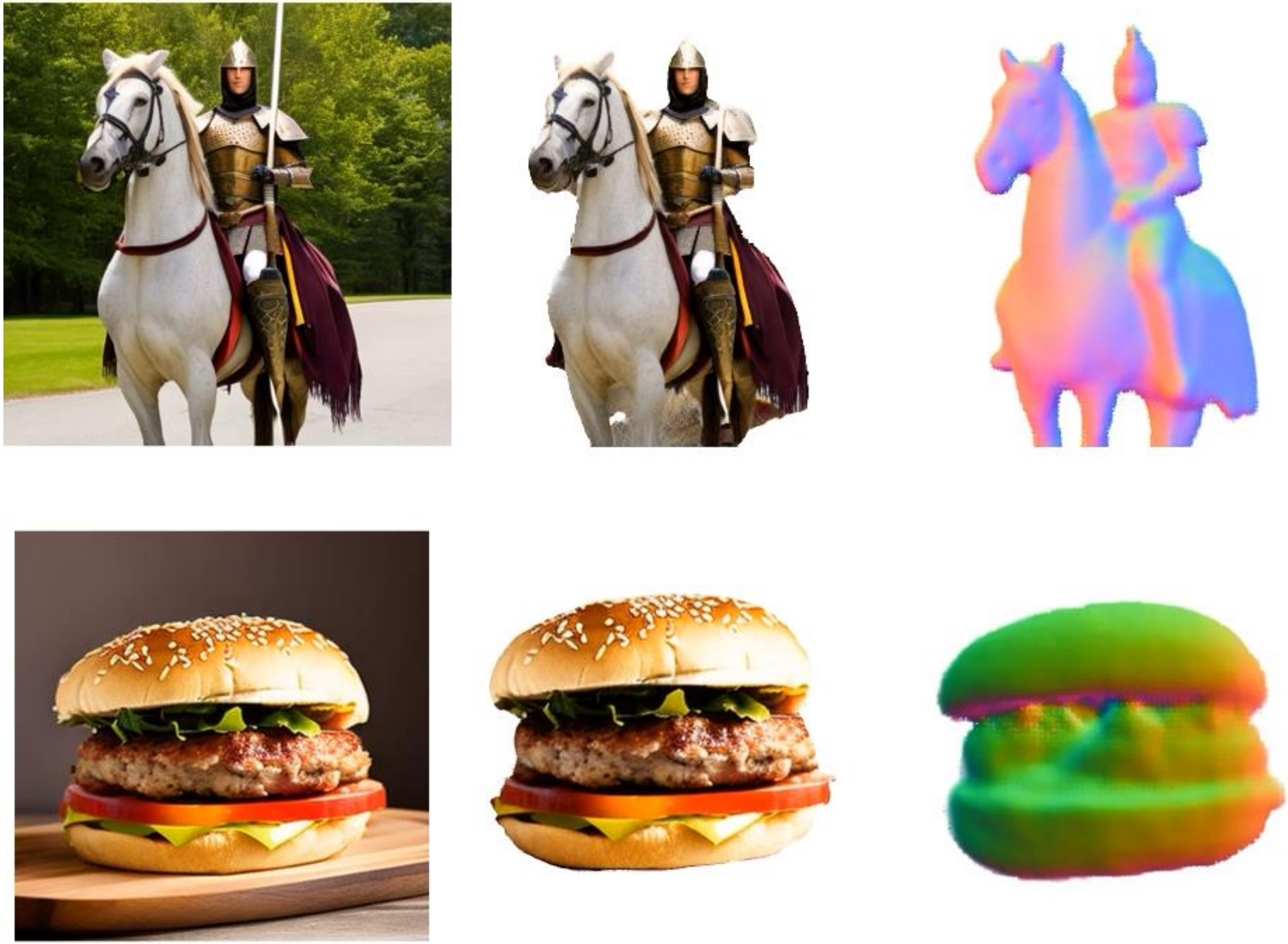


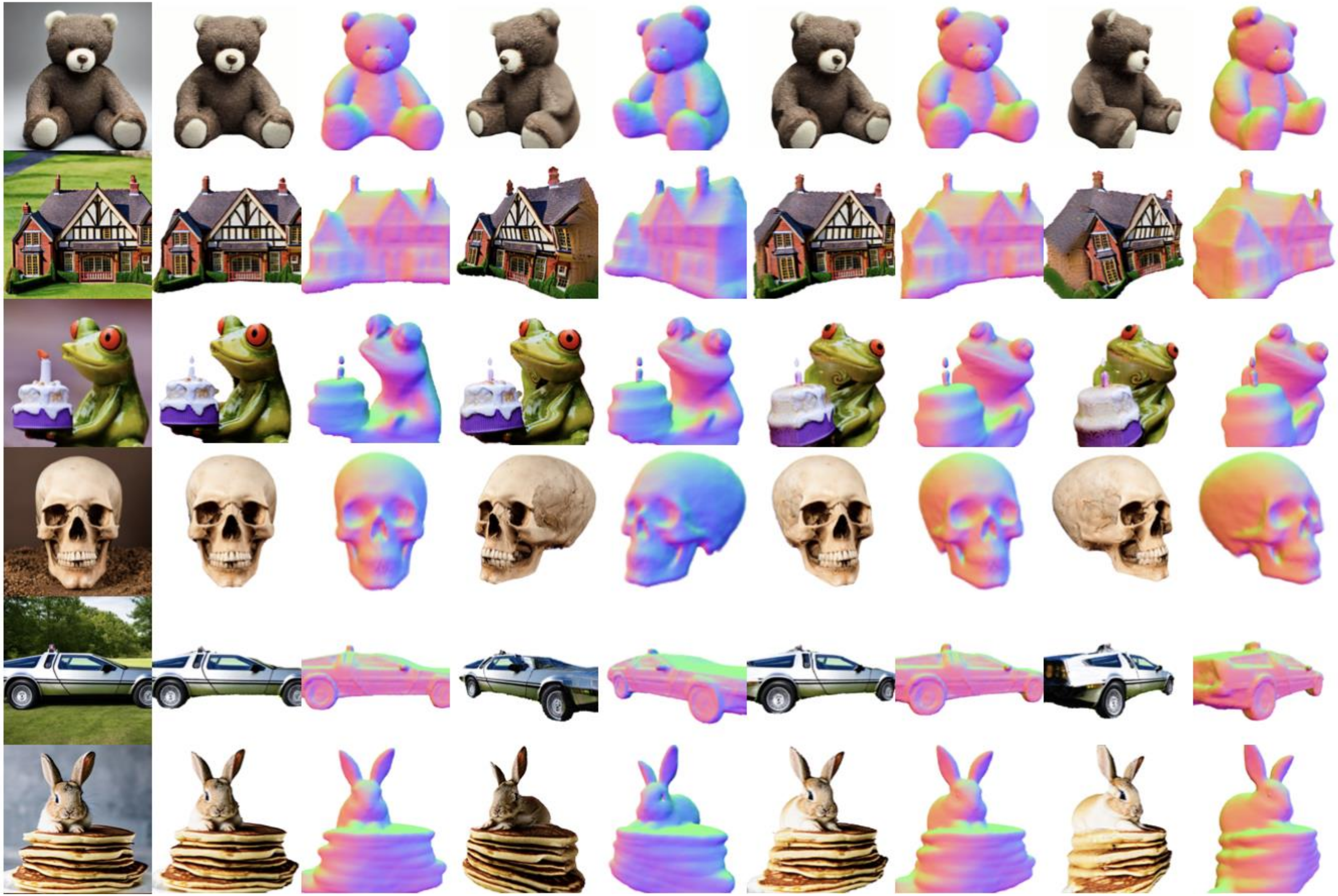


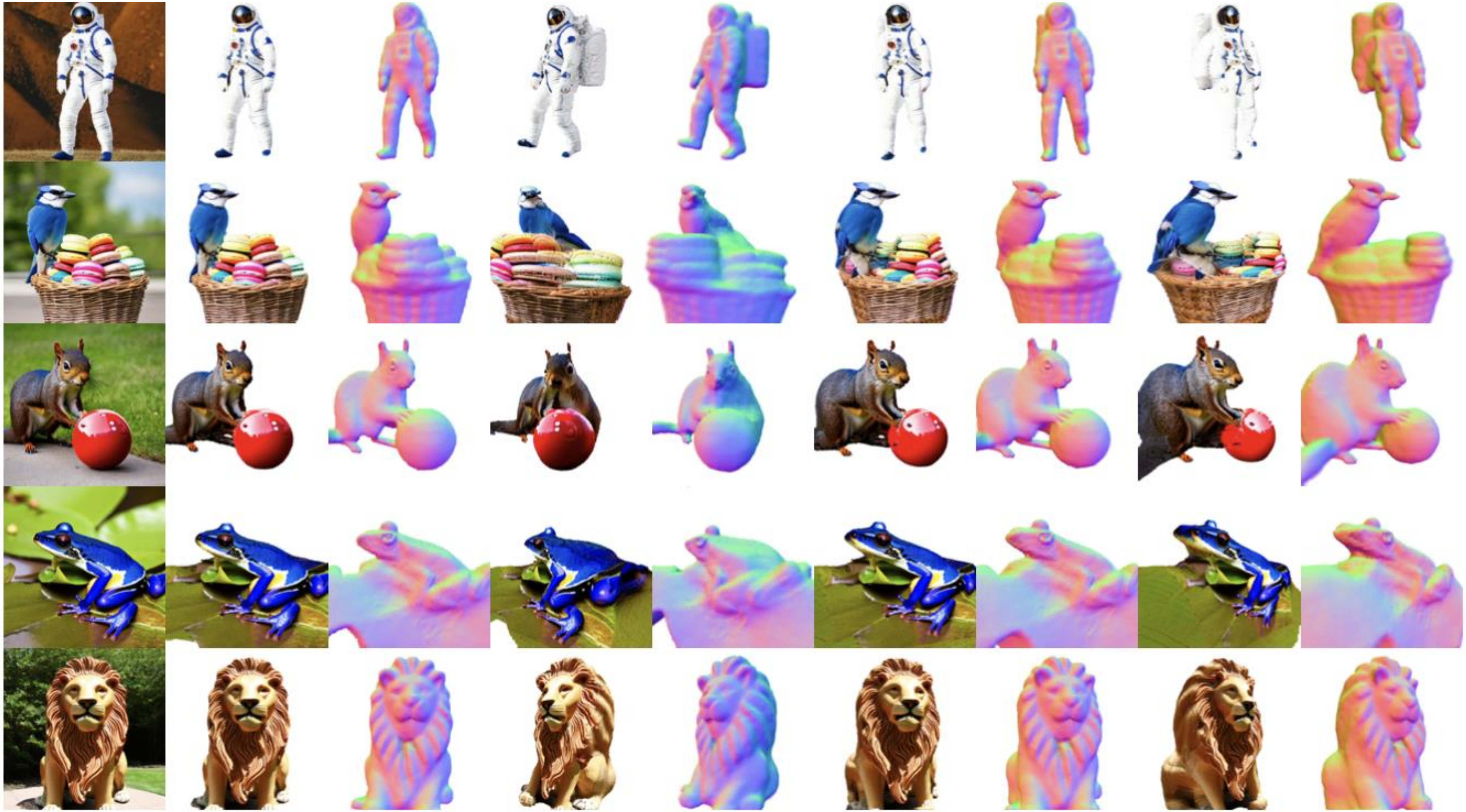


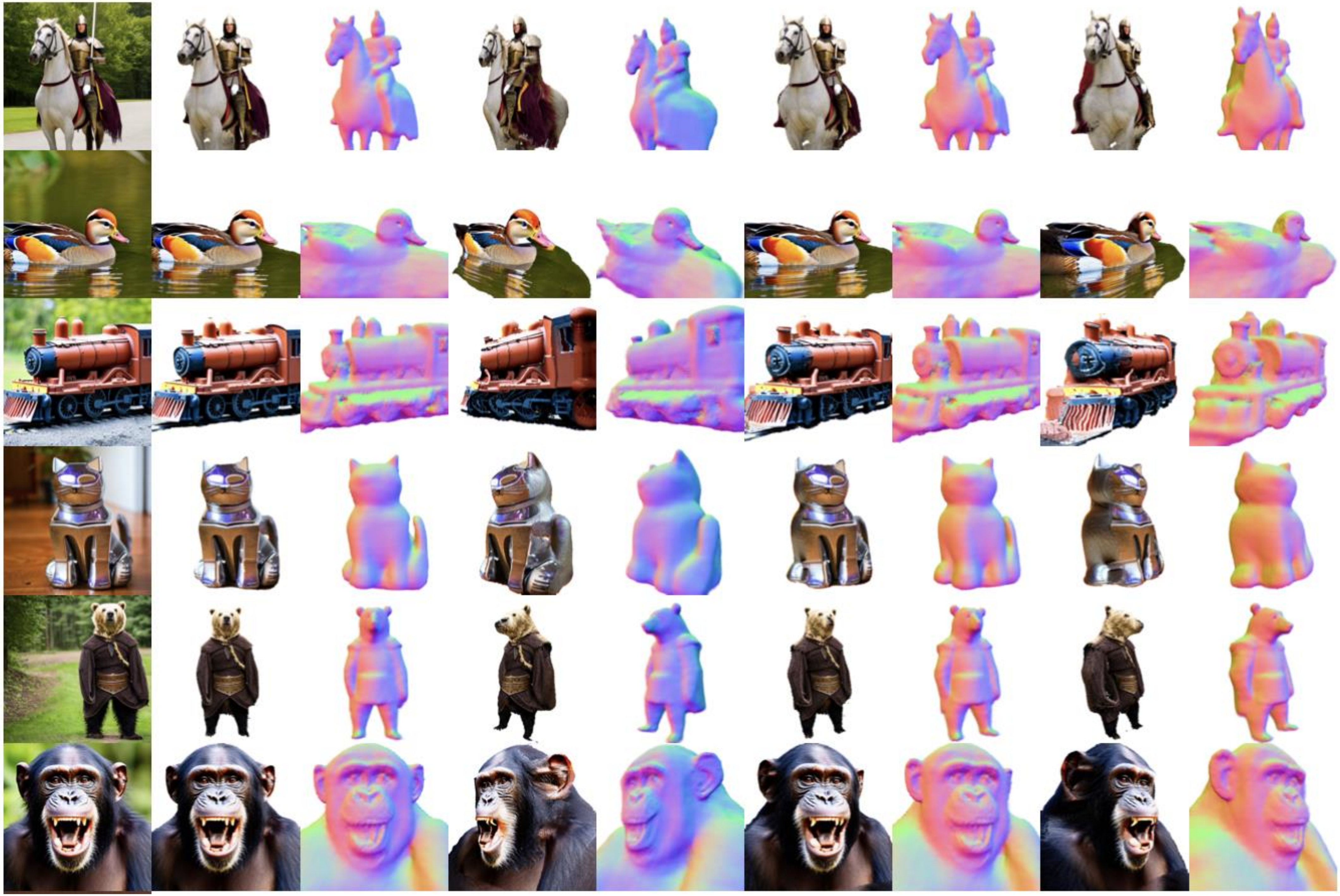


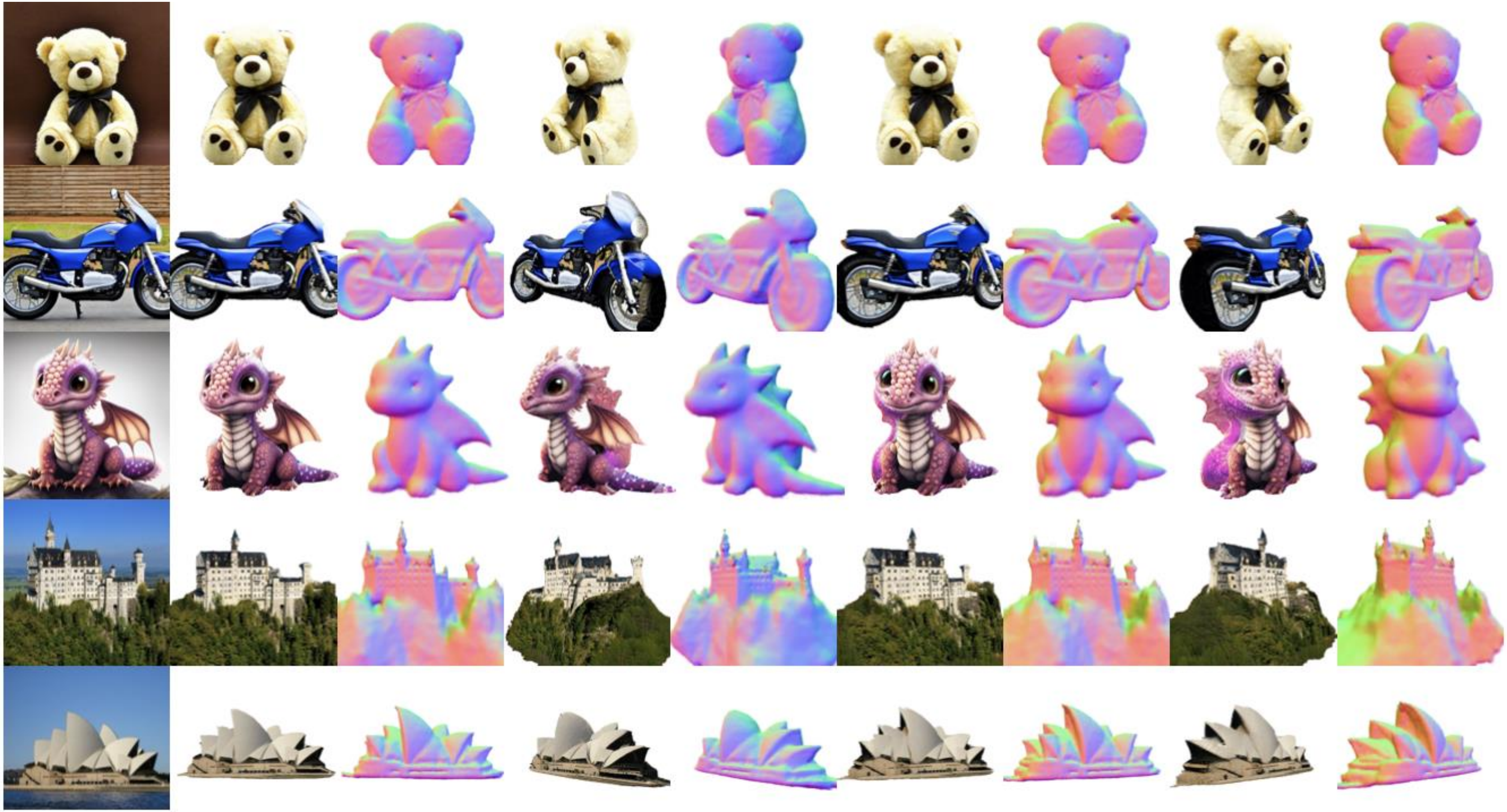


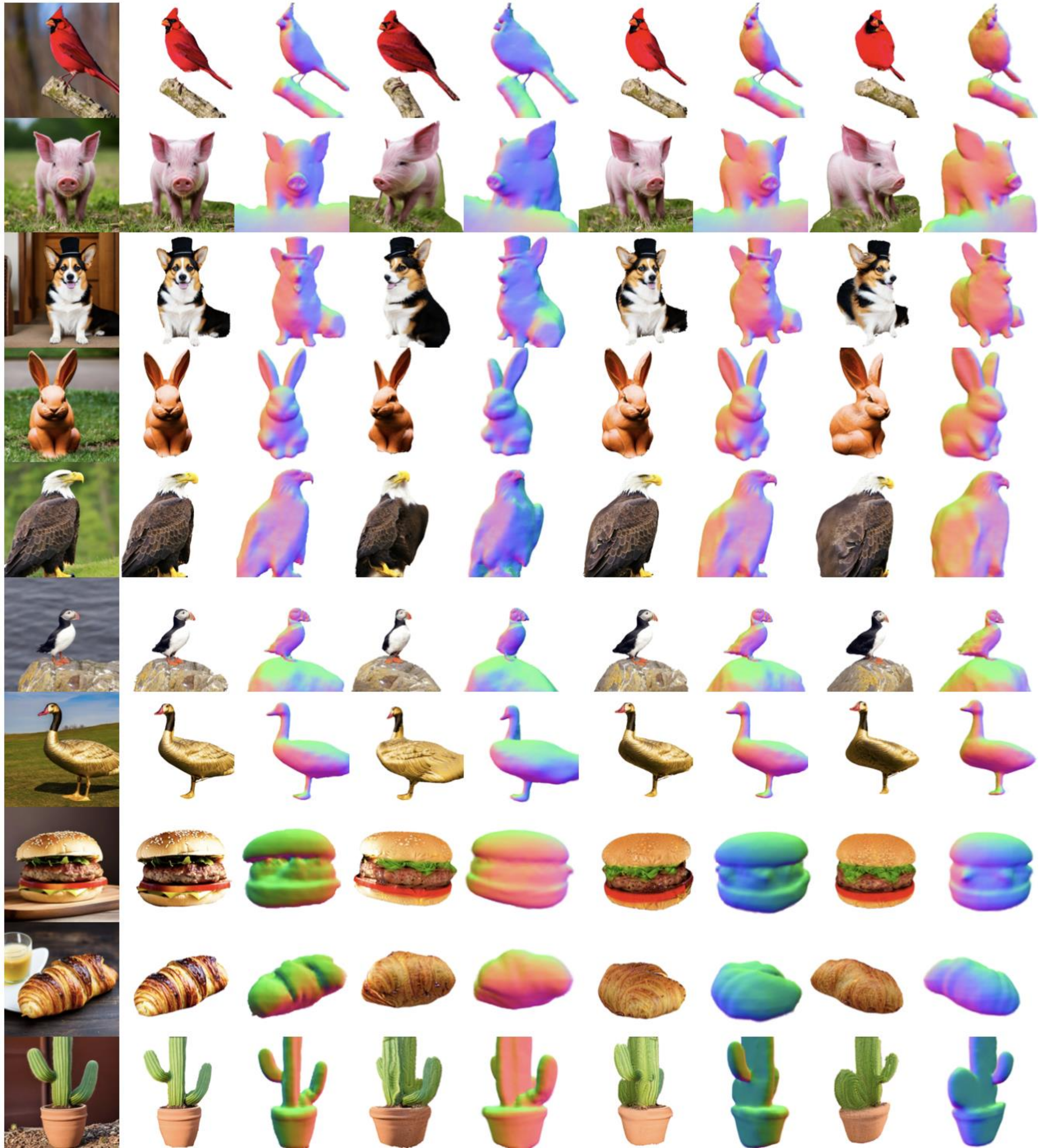


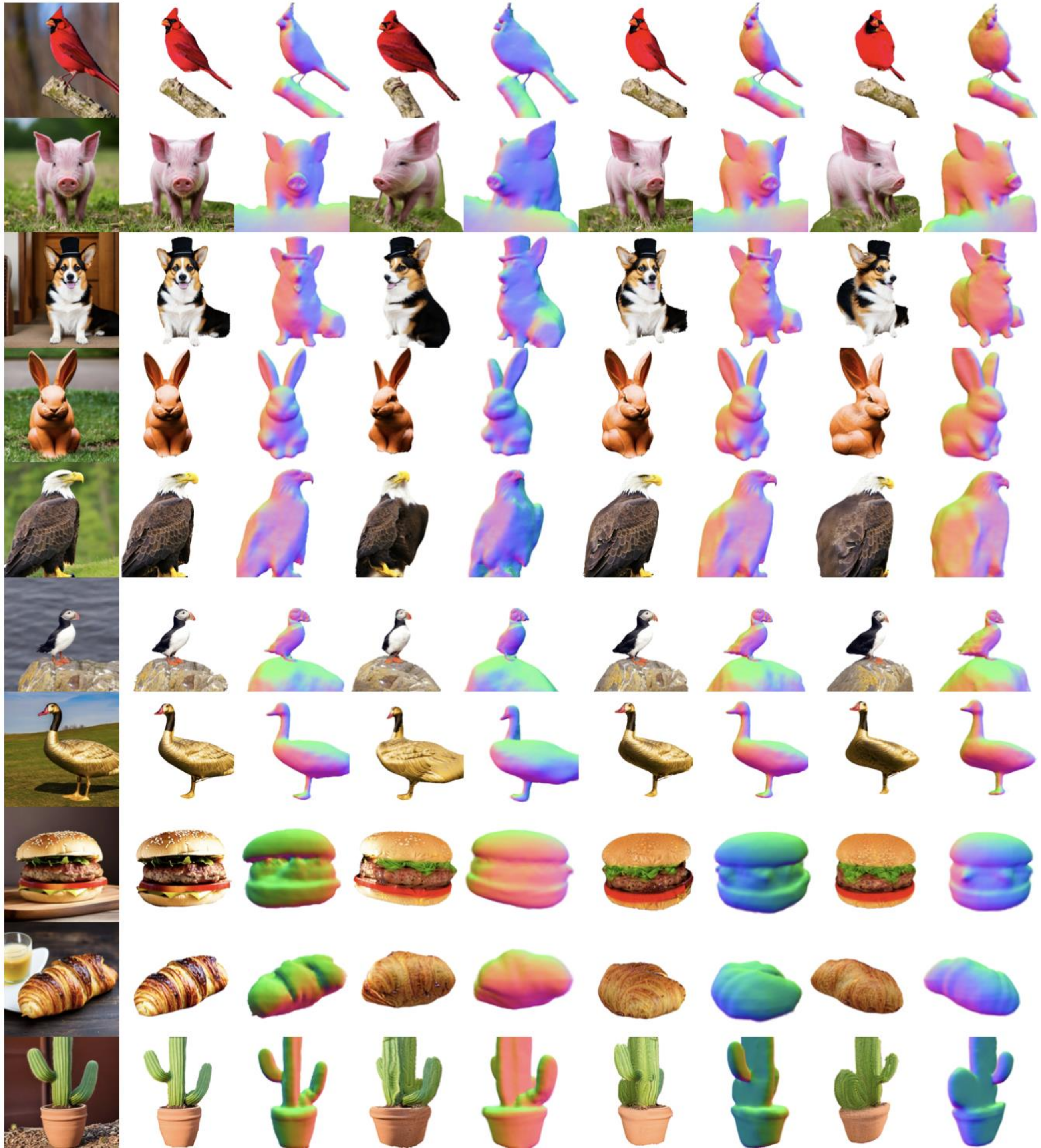












Q&A