Magic3D: High-Resolution Text-to-3D Content Creation

Chen-Hsuan Lin* Jun Gao* Luming Tang* Towaki Takikawa* Xiaohui Zeng*
Xun Huang Karsten Kreis Sanja Fidler† Ming-Yu Liu† Tsung-Yi Lin

NVIDIA Corporation

https://research.nvidia.com/labs/dir/magic3d

Appendix

A. Author Contributions

All authors have significant contributions on ideas, explorations, and paper writing. Specifically, **CHL** and **TYL** led the research, developed fundamental code for experiments and organized team efforts. **JG** led the experiments on generating high-resolution mesh models. **LT** led the experiments on using high-resolution diffusion prior. **TT** led the experiments on sparse scene representations. **XZ** and **KK** led the experiments in controllable generation. **XH** conducted the user study. **SF** and **MYL** advised the research direction and designed the scope of the project.

B. Implementation Details

We follow the implementation details described by Poole *et al*. [7] as closely as possible. We refer readers to the Dreamfusion paper [7] for context and list the major differences below.

Architectural details. As aforementioned in the main paper, we adopt a multi-resolution hash grid encoding architecture from Instant NGP [6] instead of using a large global coordinate-based MLP architecture. We use 16 levels of hash dictionaries of size 2^{19} and dimension 4, spanning 3D gird resolutions from 2^4 to 2^{12} with an exponential growth rate. We use single-layer MLPs with 32 hidden units to predict all of RGB color, volume density, and normal, where the inputs to the MLPs are the concatenated feature vectors from the multi-resolution hash encoding sampled with trilinear interpolation (we refer readers to the Instant NGP paper [6] for more details in this representation). We perform density-based pruning to sparsify the Instant NGP representation with an octree structure every 10 iterations. This allows us to more efficiently render pixels using empty space skipping, even with 3D points as dense as 1024 samples per ray. We do not use the contracting reparametrization of unbounded scenes from Mip-NeRF 360 [2] as it is not supported by our sparse representation.

Scene representation. For the coarse neural field representation, we use a bounding sphere of radius 2 for our experiments. We use the $\mathrm{softplus}$ activation for the density prediction and follow Poole *et al*. [7] to add an initial spatial density bias to encourage the optimization to focus on the object-centric neural field. We empirically found that using a linear form of spatial density bias helps stabilize the optimization, more formally written as

$$\tau_{\text{init}}(\boldsymbol{\mu}) = \lambda_\tau \cdot \left(1 - \frac{\|\boldsymbol{\mu}\|_2}{c}\right) , \tag{1}$$

where $\boldsymbol{\mu}$ is the 3D location, $\lambda_\tau = 10$ is the density bias scale, and $c = 0.5$ is an offset scale. Different from DreamFusion, however, we add this density bias to the *pre*-activation; as a result, the post-activation of the density prediction will vary continuously from $\mathrm{softplus}(\lambda_\tau)$ to 0 as a function of $\|\boldsymbol{\mu}\|_2$.

Camera and light augmentations. We follow Poole *et al*. [7] to add random augmentations to the camera and light sampling for rendering the shaded images. Differently, (a) we sample the point light location such that the angular distance from the random camera center location (w.r.t. the origin) is sampled from $\psi_{\text{cam}} \sim \mathcal{U}(0, \pi/3)$ with a random point light distance $r_{\text{cam}} \sim \mathcal{U}(0.8, 1.5)$, and (b) we use a "soft" version of the textureless and albedo-only augmentation such that various strengths of shading in the rendered images are seen during optimization. (c) we sample the camera distance from $\mathcal{U}(1.5, 2)$, and the focal length $\mathcal{U}(0.7, 1.35)$. When training with high resolution diffusion prior, we increase the focal length and sample from $\mathcal{U}(1.2, 1.8)$.

Optimization. Unless otherwise specified, we optimize the coarse model with batch size 32 using the Adam optimizer [4] with a learning rate of 1×10^{-2} without warmup and decay. Note that the large global coordinate-based MLP architecture in DreamFusion [7] limits its optimization to only an effective batch size of 8. For the coarse model, we add the opacity regularization as suggested by Poole *et*

*†: equal contribution.

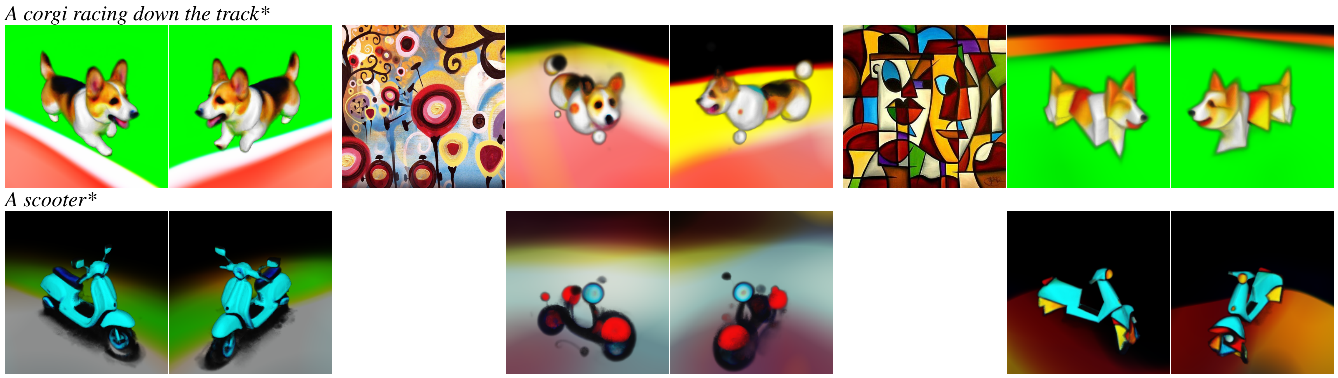


Figure 1. **Magic3D with image style transfer.** The style of the reference image is transferred to the 3D model by providing it to the diffusion model as conditional input. We apply different styles during 3D synthesis with two different text prompts. The first two columns show the 3D model optimized given the text without reference image. Afterwards, we show the reference and the corresponding 3D shapes.

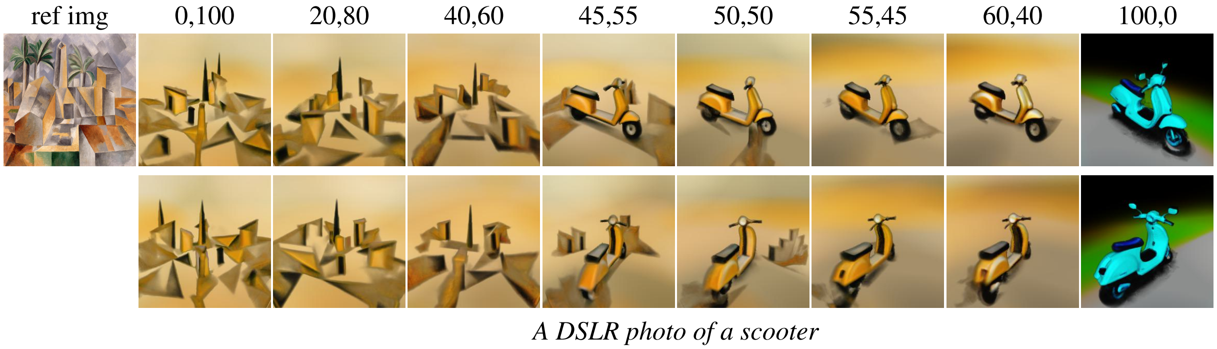


Figure 2. **Magic3D with image style transfer with different guidance weights.** Each column uses a different combination of guidance weights ω_{text}, ω_{joint}. All results are optimized with noise level threshold $t = 1.0$. When $\omega_{\text{text}} = 0, \omega_{\text{joint}} = 100$, the setting is equivalent to using a single guidance weight. The style image dominates the resulting scene if the ω_{joint} is too large. We generally find that using a guidance weight combination around $50, 50$ results in the best performance.

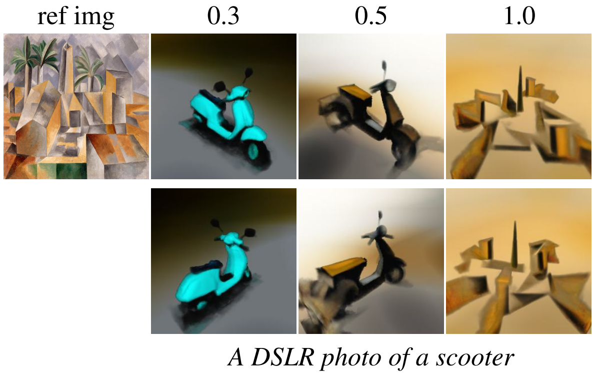


Figure 3. **Magic3D with image style transfer with different noise level thresholds.** Each column uses a different noise level threshold t. All results are optimized with guidance weights ω_{text}, $\omega_{\text{joint}} = 40, 60$. When the noise level threshold $t = 0$, the setup is equivalent to using no style image guidance. We generally find that setting the threshold around 0.5 provides the best performance.

al. [7] to encourage sparsity in the volume density field, but we do not add the orientation regularization as we empirically found it to hurt optimization.

Score Distillation Sampling. In the first stage, we sample the timestep $t \sim \mathcal{U}(0, 1)$ and set $w(t) = 1$. In the second stage, we find the range of timestep t in SDS affects the quality. We sample $t \sim \mathcal{U}(0.02, 0.5)$ in our experiments. In general, setting t_{max} in the range of $[0.5, 0.7]$ works well. We set $w(t) = \sigma_t^2$ in this stage.

C. Alternative High-Resolution Prior

In addition to LDM, we also consider using Super Resolution (SR) diffusion prior [1, 8] for increasing the resolution of a coarse model. This diffusion model is trained to generate a high-resolution image conditioning on a low-resolution input image. In SDS, this model predicts noises added in high resolution, i.e., $\epsilon_\phi(x_t; y, t, x_{\text{low}})$, where x_{low} denotes a 64×64 low-resolution image. We render x_{low} with a frozen coarse model to optimize the second-stage fine model. Fig. 5

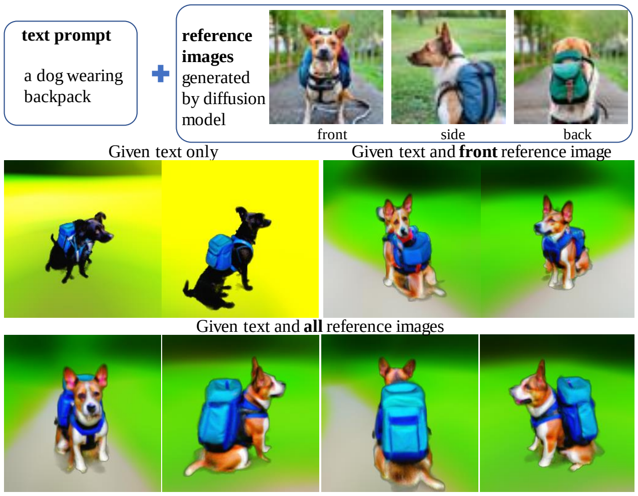

text prompt
a dog wearing backpack
reference images generated by diffusion model
front
side
back
Given text only
Given text and **front** reference image
Given text and **all** reference images

Figure 4. **Magic3D with image content transfer.** Given the reference images generated by the diffusion prior, we optimize 3D models that look similar to the object in the images. We show the generated 3D models given *(i)* text only, *(ii)* text and front view's reference image only and *(iii)* text and different view's reference images. *(ii)* and *(iii)* preserve the look of the dog in the image. With multiple reference images, *(iii)* yields higher quality and more 3D-consistent outputs.

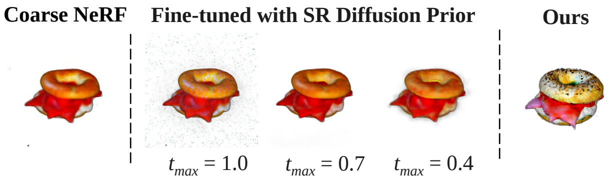


Figure 5. Fine-tuning NeRF with SR prior fails to add high-resolution details. t_{max} is the maximum timestep sampled in SDS.

shows this approach fails to add high-quality details to the input coarse model.

D. Style-Guided Text-to-3D Synthesis

We also explore controlling the 3D generation with multi-modal conditioning. The eDiff-I diffusion prior [1] is designed such that it can condition on a reference image when performing text-to-image generation. Such an image conditioning design makes it easy to change the style of the generated output. However, we find that naïvely feeding the style image as input to the model when computing the SDS gradient can result in a poor 3D model that is essentially overfitting to the input image. We hypothesize that the conditioning signal by the image is significantly stronger than the text prompt during optimization. To better balance the guidance strength between image and text conditioning, we extend our model's classifier-free guidance scheme [3] and compute the final $\tilde{\epsilon}_\phi(x_t; y_{\text{text}}, y_{\text{image}}, t)$:

$$\begin{aligned}\tilde{\epsilon}_\phi(x_t; y_{\text{text}}, y_{\text{image}}, t) &= \epsilon_\phi(x_t; t)\\ &+ \omega_{\text{text}}[\epsilon_\phi(x_t; y_{\text{text}}, t) - \epsilon_\phi(x_t; t)]\\ &+ \omega_{\text{joint}}[\epsilon_\phi(x_t; y_{\text{text}}, y_{\text{image}}, t) - \epsilon_\phi(x_t; t)]\ ,\end{aligned} \tag{2}$$

where y_{text} and y_{image} are text and image conditioning respectively, and ω_{text} and ω_{joint} are the guidance weights for text and joint text-and-image conditioning respectively. Note that for $\omega_{\text{joint}} = 0$, the scheme is equivalent to standard classifier-free guidance with respect to text conditioning only.

Fig. 1 shows our style-guided text-to-3D generation results. When optimizing the 3D model, we feed the reference image to the eDiff-I model. We set $\omega_{\text{text}}, \omega_{\text{joint}} = 50, 50$ or $40, 60$ and apply the image guidance when $t < 0.5$ only. We do not provide high-resolution results for this experiment because LDM does not support reference image conditioning.

Guidance weight and noise level threshold. We ablate different combinations of guidance weights and noise level thresholds in Figs. 2 and 3, respectively. The guidance weights ω_{text} and ω_{joint} balance the guidance strength during optimization (see Eq. 2). A similar guidance formulation has also been used by Liu *et al.* for compositional text-to-image generation [5]. We also find that applying the image conditioning only below a certain noise level threshold can help control style transfer. The intuition is that image-based style guidance is most relevant for optimizing the generated 3D object's details, which are modeled at lower noise levels. Notice that we do not provide high-resolution results for this experiment because LDM does not support image conditioning inputs.

Content image as reference. We also explore using multiple images as inputs during 3D synthesis to transfer the content in the images to the 3D model, as shown in Fig. 4: Given a text prompt, we first ask the eDiff-I model to generate the front view, side view and back view images. When optimizing the 3D model for the same text prompt from different views, we then feed the corresponding generated view image as input to guide the 3D synthesis. This approach requires some degree of consistency with respect to subject identity across the different view images, which can be achieved by generating a set of different view images first and choosing accordingly. Overall, the experiment shows that we can apply the text-to-image diffusion model to generate images that can be used for guidance during 3D model optimization. As we see, this does not only provide enhanced control by preserving the identity of the subjects in the images, but also improves output quality and 3D consistency. Generally, depending on image type, image conditioning can be used either for object-centric content transfer to 3D (Fig. 4) or for abstract 3D stylization (Figs. 1, 2, and 3).

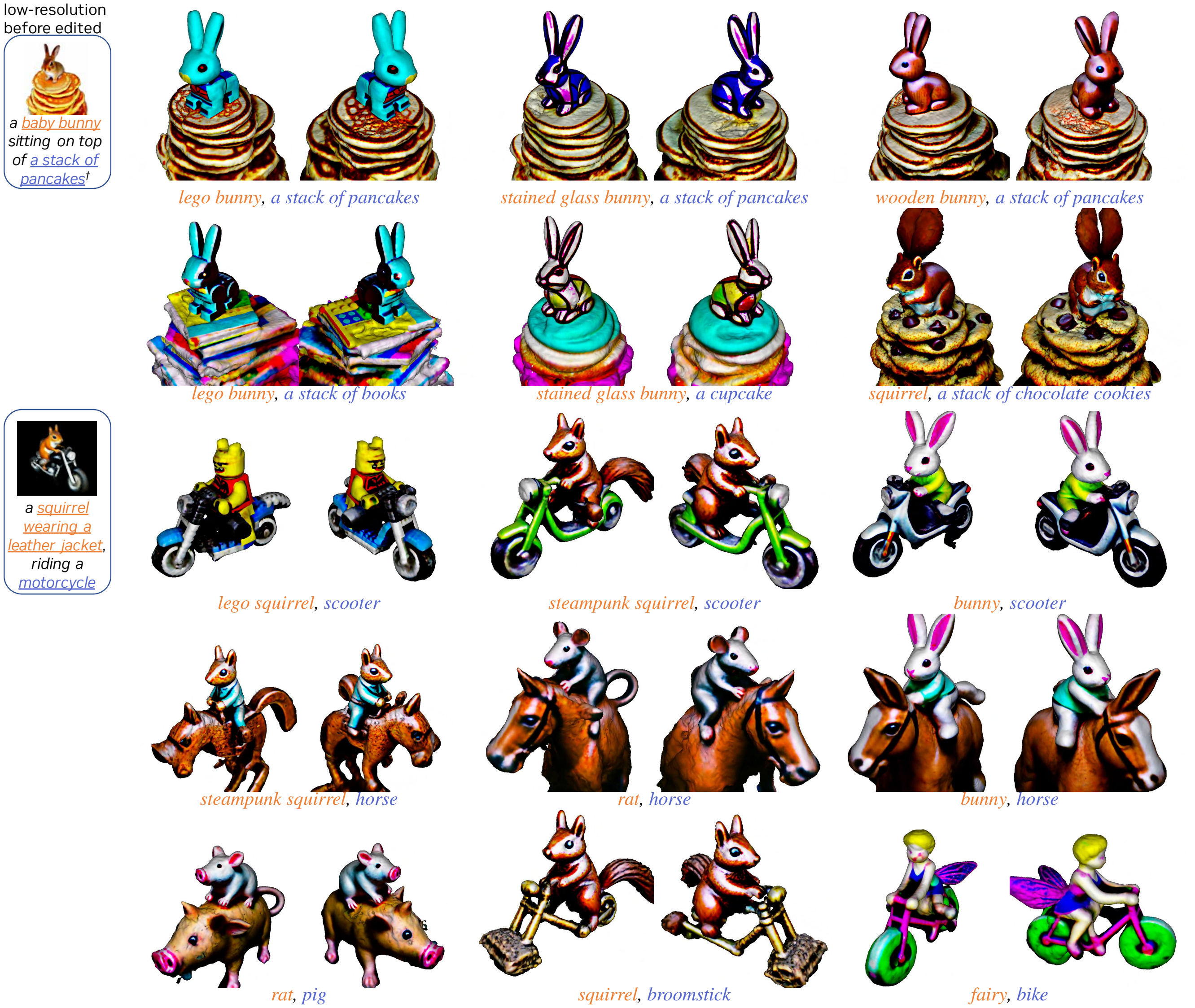

low-resolution before edited
a baby bunny sitting on top of a stack of pancakes†
lego bunny, a stack of pancakes
stained glass bunny, a stack of pancakes
wooden bunny, a stack of pancakes
lego bunny, a stack of books
stained glass bunny, a cupcake
squirrel, a stack of chocolate cookies
a squirrel wearing a leather jacket, riding a motorcycle
lego squirrel, scooter
steampunk squirrel, scooter
bunny, scooter
steampunk squirrel, horse
rat, horse
bunny, horse
rat, pig
squirrel, broomstick
fairy, bike

Figure 6. **Magic3D with prompt-based editing**. Given a coarse model (first column) generated with a base prompt, we replace the underscored text with new text and fine-tune the NeRF to get a high-resolution NeRF model with LDM. We further fine-tune the high-resolution mesh on top of the NeRF model. Such a prompt-based editing technique gives artists better control over the 3D generation output.

E. Additional Results

We provide more qualitative comparisons with Dreamfusion [7] in Figs. 7, 8, 9, 10, 11. Our Magic3D achieved much higher quality in terms 3D geometry and texture.

We also show more results on prompt-based editing in Fig. 6. Our Magic3D enable high-quality editing of the 3D content through simple text prompt modification.

References

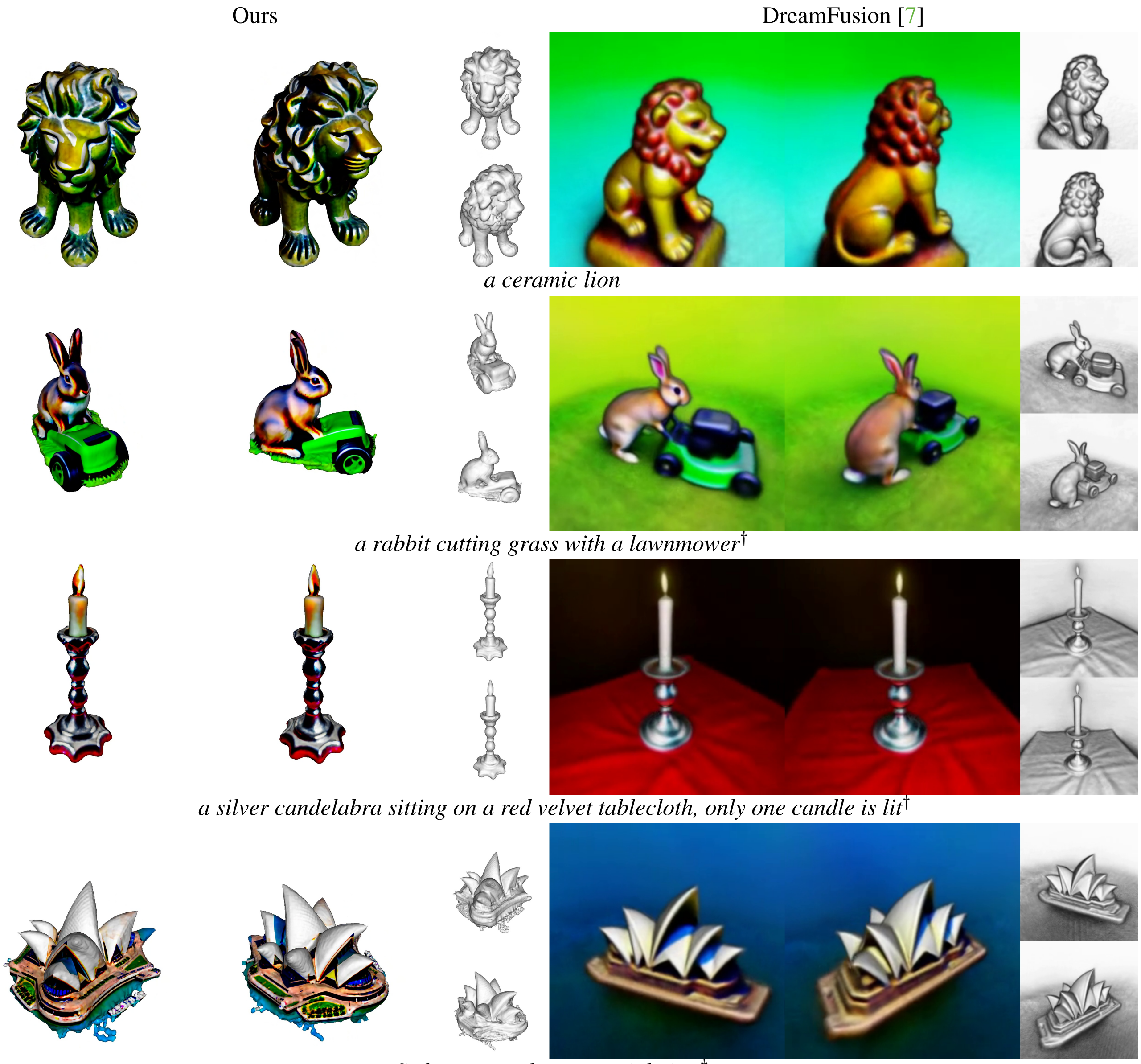
[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2, 3

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022. 1

[3] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 3

[4] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 1

[5] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B



a ceramic lion

a rabbit cutting grass with a lawnmower†

a silver candelabra sitting on a red velvet tablecloth, only one candle is lit†

Sydney opera house, aerial view†

Figure 7. **Qualitative comparison with DreamFusion [7].** We use the same text prompt as in DreamFusion. For each 3D model, we render it from two views with a textureless rendering for each view and remove the background to focus on the 3D shape. For the DreamFusion results, we take frames from the videos published on the official webpage. Magic3D generates much higher quality 3D shapes on both geometry and texture compared with DreamFusion. ∗ *a DSLR photo of...* † *a zoomed out DSLR photo of...*

Tenenbaum. Compositional visual generation with composable diffusion models. *arXiv preprint arXiv:2206.01714*, 2022. 3

[6] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. 1

[7] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1, 2, 4, 5, 6, 7, 8, 9

[8] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2

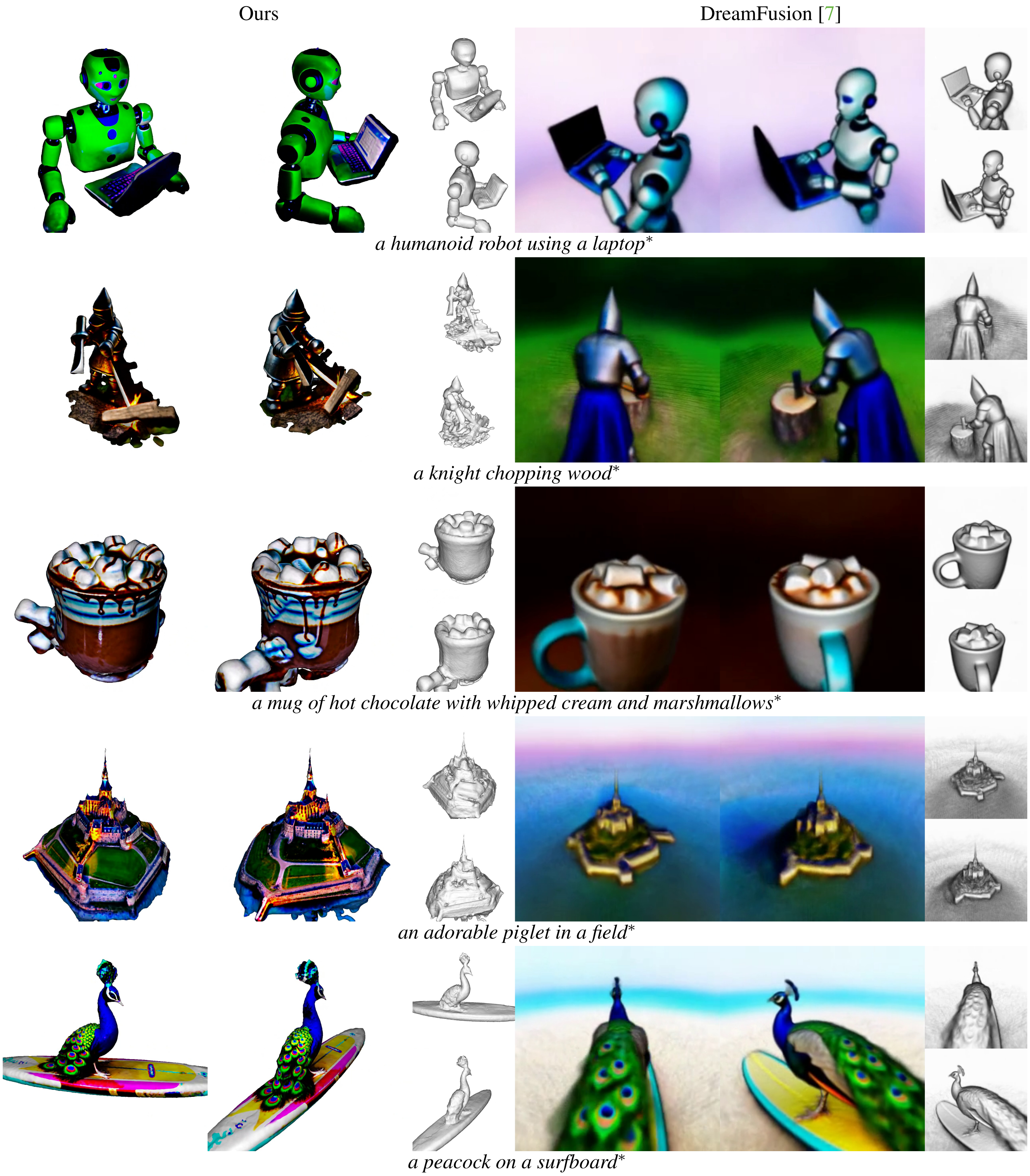
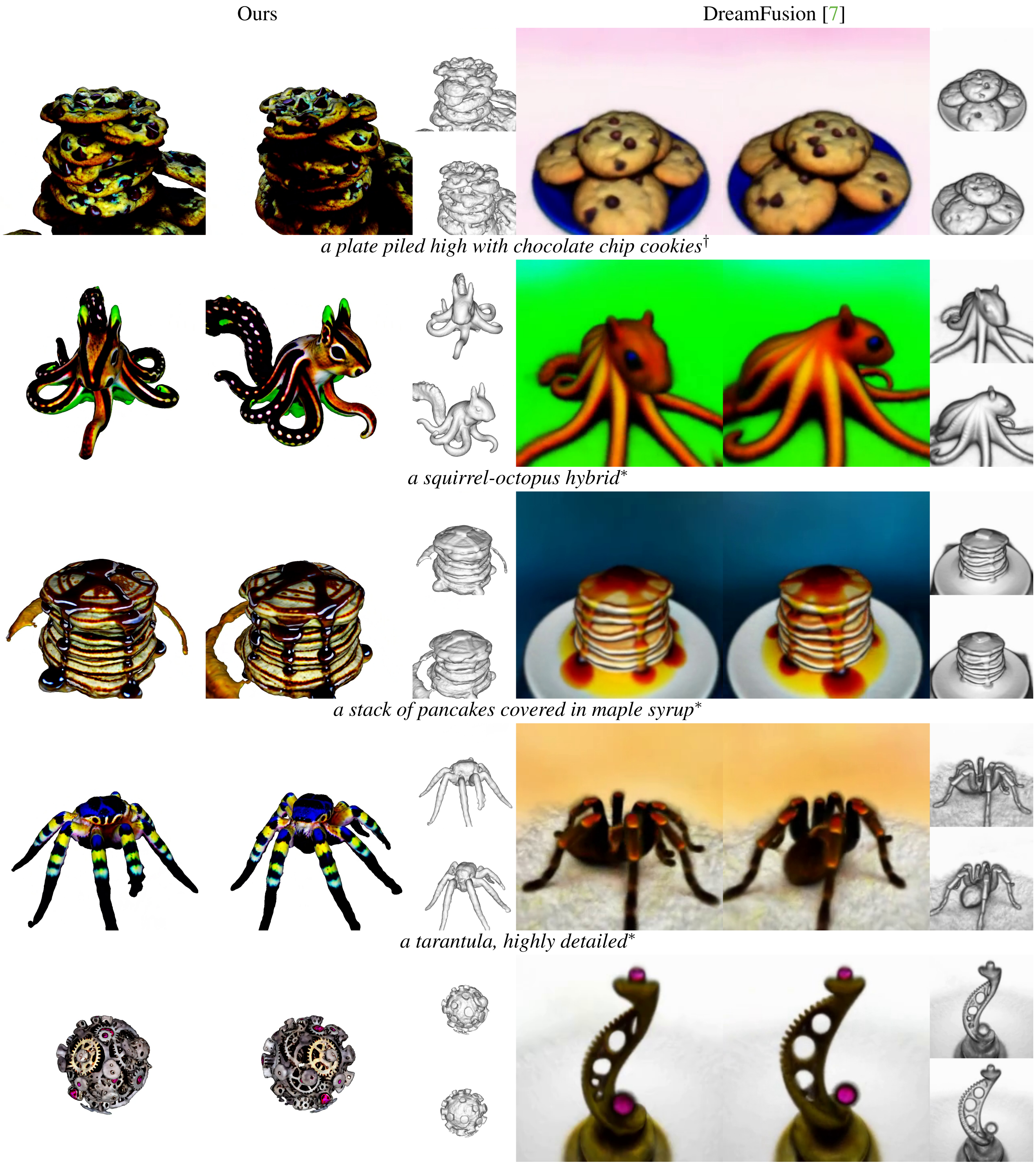

Ours
DreamFusion [7]
a humanoid robot using a laptop*
a knight chopping wood*
a mug of hot chocolate with whipped cream and marshmallows*
an adorable piglet in a field*
a peacock on a surfboard*

Figure 8. **Qualitative comparison with DreamFusion [7].** We use the same text prompt as in DreamFusion. For each 3D model, we render it from two views with a textureless rendering for each view and remove the background to focus on the 3D shape. For the DreamFusion results, we take frames from the videos published on the official webpage. Magic3D generates much higher quality 3D shapes on both geometry and texture compared with DreamFusion. ∗ *a DSLR photo of...* † *a zoomed out DSLR photo of...*



a plate piled high with chocolate chip cookies†

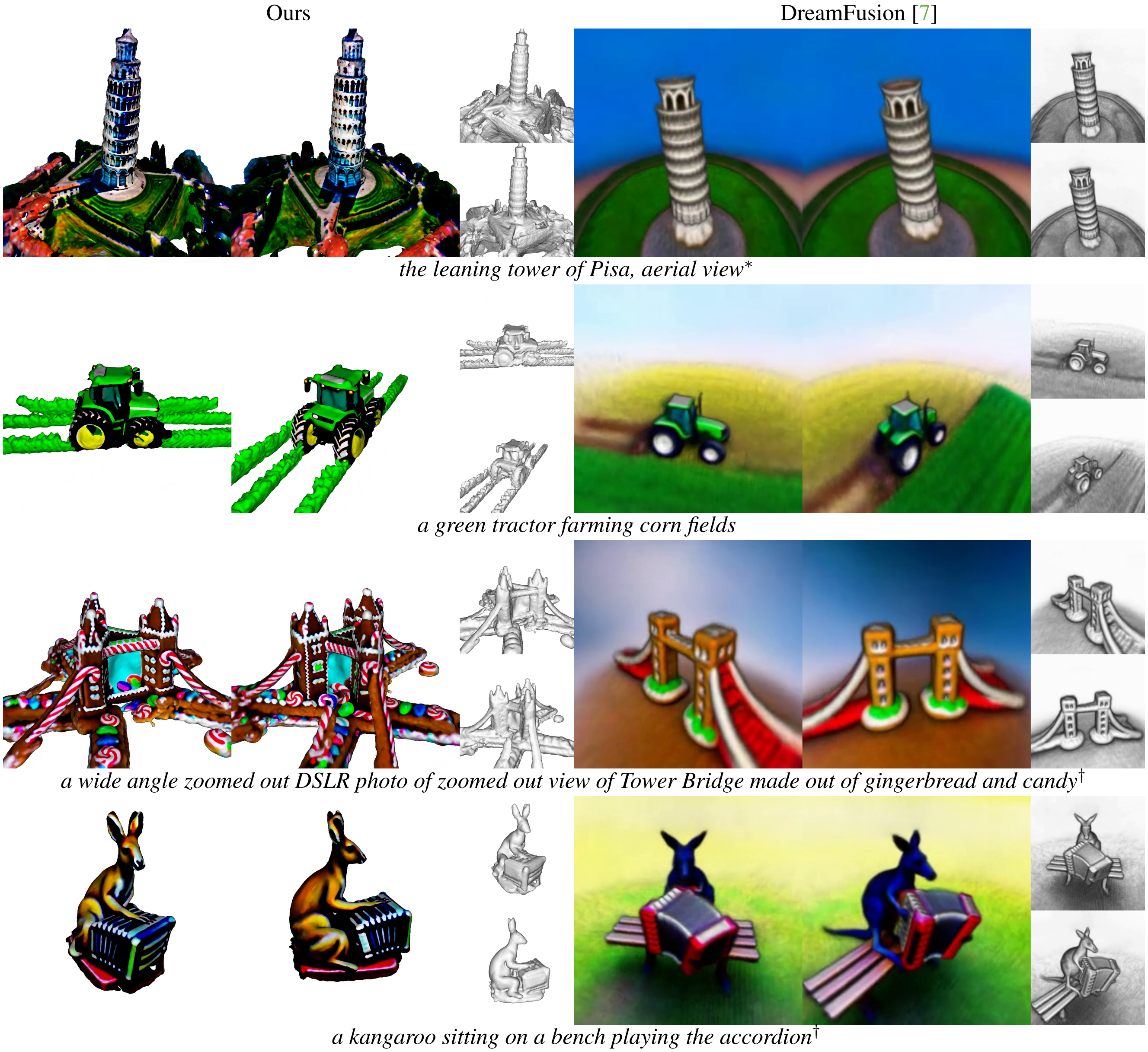
*a squirrel-octopus hybrid**

*a stack of pancakes covered in maple syrup**

*a tarantula, highly detailed**

*a very beautiful small organic sculpture made of fine clockwork and gears with tiny ruby bearings, very intricate, caved, curved. Studio lighting, High resolution, white background**

Figure 9. **Qualitative comparison with DreamFusion [7].** We use the same text prompt as in DreamFusion. For each 3D model, we render it from two views with a textureless rendering for each view and remove the background to focus on the 3D shape. For the DreamFusion results, we take frames from the videos published on the official webpage. Magic3D generates much higher quality 3D shapes on both geometry and texture compared with DreamFusion. ∗ *a DSLR photo of...* † *a zoomed out DSLR photo of...*



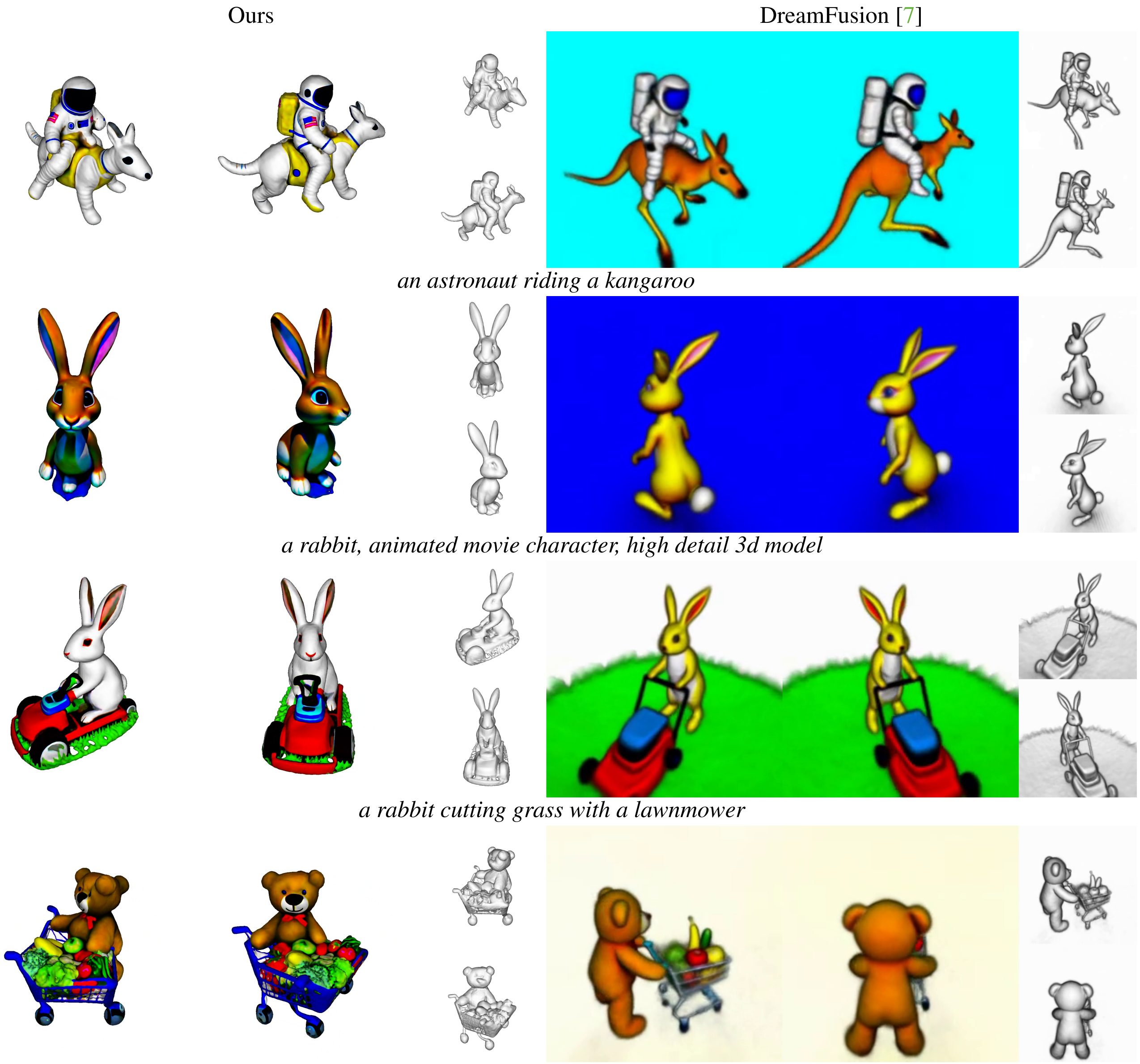
the leaning tower of Pisa, aerial view*

a green tractor farming corn fields

a wide angle zoomed out DSLR photo of zoomed out view of Tower Bridge made out of gingerbread and candy†

a kangaroo sitting on a bench playing the accordion†

Figure 10. **Qualitative comparison with DreamFusion [7].** We use the same text prompt as in DreamFusion. For each 3D model, we render it from two views with a textureless rendering for each view and remove the background to focus on the 3D shape. For the DreamFusion results, we take frames from the videos published on the official webpage. Magic3D generates much higher quality 3D shapes on both geometry and texture compared with DreamFusion. $*$ *a DSLR photo of...* † *a zoomed out DSLR photo of...*



an astronaut riding a kangaroo

a rabbit, animated movie character, high detail 3d model

a rabbit cutting grass with a lawnmower

a teddy bear pushing a shopping cart full of fruits and vegetables

Figure 11. **Qualitative comparison with DreamFusion [7].** We use the same text prompt as in DreamFusion. For each 3D model, we render it from two views with a textureless rendering for each view and remove the background to focus on the 3D shape. For the DreamFusion results, we take frames from the videos published on the official webpage. Magic3D generates much higher quality 3D shapes on both geometry and texture compared with DreamFusion. $*$ *a DSLR photo of...* † *a zoomed out DSLR photo of...*