A. Appendix

A.1. Unavailability of Reference Objects

MT3D relies on a high-fidelity 3D object to generate geometrically coherent representations. In this section, we assess the performance of MT3D in situations where an object that directly matches the text prompt is unavailable. We propose two alternative approaches for these scenarios. Firstly, we suggest utilizing existing text-to-3D generators to create initial 3D representations, which can then guide our generation pipeline. Secondly, instead of selecting a high-fidelity object that precisely matches the text prompt, we investigate the feasibility of using 3D objects that approximately belong to the same class as the object of interest.

A.1.1 Utilizing Existing Text-to-3D Generators

To assess the generative capabilities of MT3D, we evaluate its performance using objects generated by existing text-to-3D generators as guiding references. Specifically, we experiment with three types of reference objects: 1) low-fidelity objects generated by Point-E [33], 2) high-fidelity objects generated by HiFA, and 3) objects exhibiting the Janus problem. Objects generated by Point-E typically exhibit low fidelity and limited diversity due to its direct training on 3D datasets. Moreover, Point-E often produces objects with poor shape and structure when handling complex prompts involving out-of-distribution objects. However, even when guided by such low-fidelity reference objects with disoriented geometry and poor structural quality, MT3D successfully generates 3D structures with minimal geometric disorientations. For example, as shown in Figure 8(a), MT3D generates a high-fidelity ceramic lion, preserving a coherent shape and structure despite the deficiencies of the reference object. Although certain generated representations, such as the corgi, display the Janus problem, these outputs demonstrate notable improvement compared to those generated without any reference guidance.

The quality of MT3D generation improves further when it is guided by more geometrically coherent reference objects. For example in Figure 8 (c), when using objects generated by HiFA as references, MT3D produces high-fidelity 3D models with consistent geometry. Even when the reference object has minor geometric inconsistencies, such as the lion and corgi generated by HiFA, MT3D still generates representations with a coherent shape and structure. This robustness can be attributed to the reliance on geometric moments, which focus on the overall shape and structure while being invariant to minor local changes. Additionally, to comprehensively evaluate MT3D's capabilities, we experiment with reference objects exhibiting geometric inconsistencies and the Janus problem (Figure 8 (b)). In these cases, MT3D produces representations with significantly improved shape and structure compared to the reference objects. It is important to note that MT3D leverages both a text-based 2D image generator i.e ControlNet and deep geometric moments (DGM) for generation. Consequently, even when the reference object is of poor quality and the features learned by the DGM module are suboptimal, the text-based ControlNet generator attempts to compensate for these shortcomings and generate a decent 3D representation.

A.1.2 Analogous geometric guidance

Instead of using a high-fidelity 3D reference object that directly matches the text prompt, we explore using 3D objects from a similar category for guidance. For example, in response to a "furry dog" prompt, we provide guidance using a 3D representation of a cat or a lion, rather than a dog. As shown in Figure 9, despite using objects from analogous categories, MT3D successfully generates well-structured and high-fidelity 3D representations. This result stems from the fact that geometric moments capture global shape and structure, which are not significantly affected by fine-grained details. This underscores MT3D's ability to generalize and effectively utilize geometric cues from similar objects, even in the absence of an exact match.

A.2. More Ablation Study

In this section, we present additional ablation studies (Figure 10). Consistent with Section 5.2, models optimized with the SDS and VSD configurations exhibit various geometric inconsistencies. Similarly, the bottom view and tail of the lion are better shaped under $\mathcal{L}_{\text{dgm}}$, while the face of the lion exhibits more detailed features with $\mathcal{L}_{\text{control}}$. Furthermore, the effect of DGM is evident in the case of the parrot, where $\mathcal{L}_{\text{dgm}}$ successfully generates a branch similar to the reference object. In contrast, ControlNet is unable to do so, likely because its training data may not have included many instances of parrots perched on a branch. Thus, ControlNet focuses more on learning the general aspects of shape, while DGM primarily learns shape and structure from the reference object. By leveraging both $\mathcal{L}_{\text{dgm}}$ and $\mathcal{L}_{\text{control}}$, MT3D produces superior representations across the front, back, side, and bottom views, thereby enhancing the overall fidelity of shape and structure.

A.3. Comparison against Fantasia3D, Magic3D and HiFA

In Figure 11 and 12, we qualitatively compare MT3D with state-of-the-art methods—Fantasia3D, Magic3D, and HiFA—using text prompts from their respective original papers. For Fantasia3D and HiFA, we utilize their original source code to generate 3D representations. Since the original code for Magic3D was unavailable at the time of writing, we use the threestudio-based [16] implementation to gen-

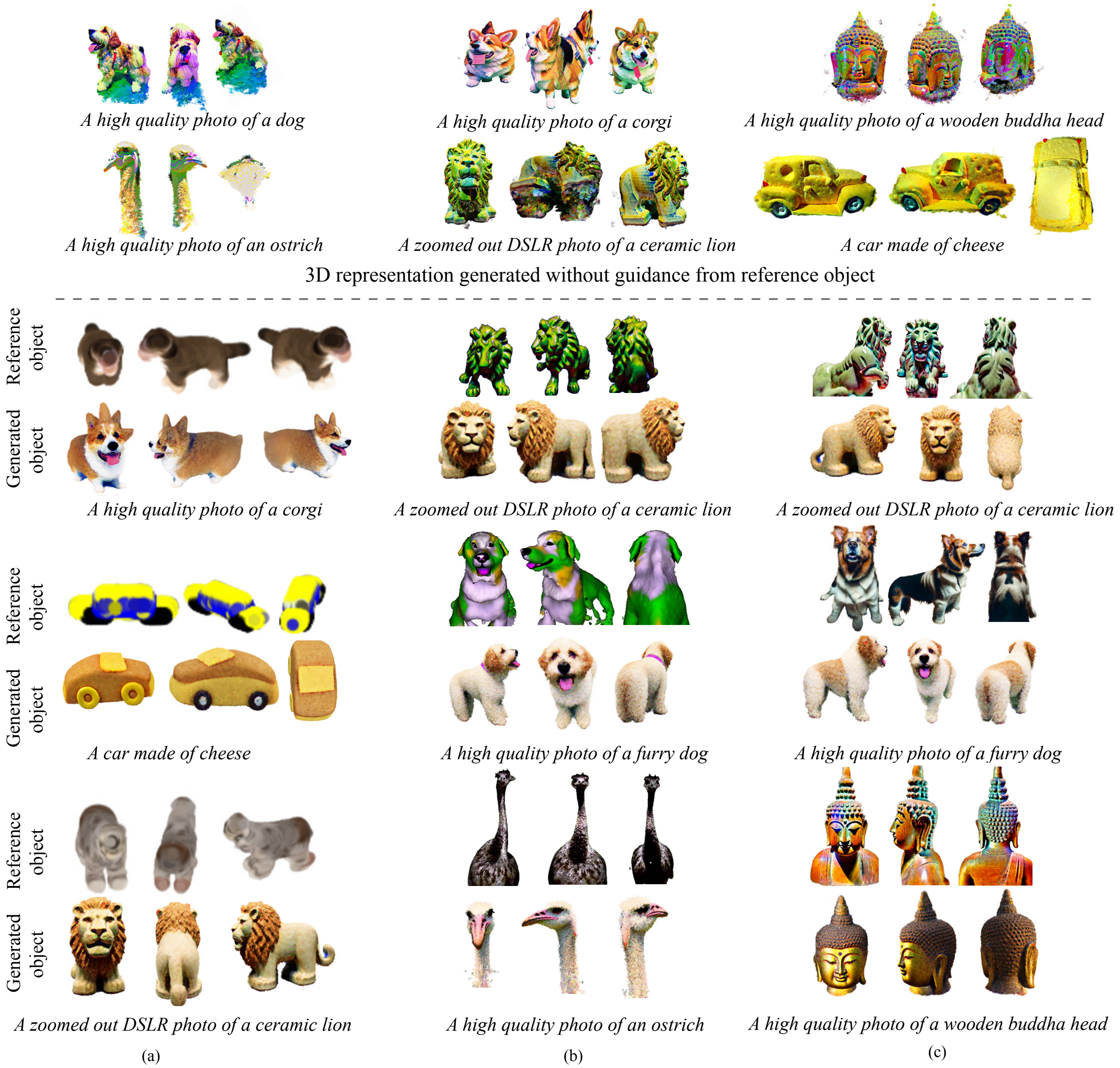
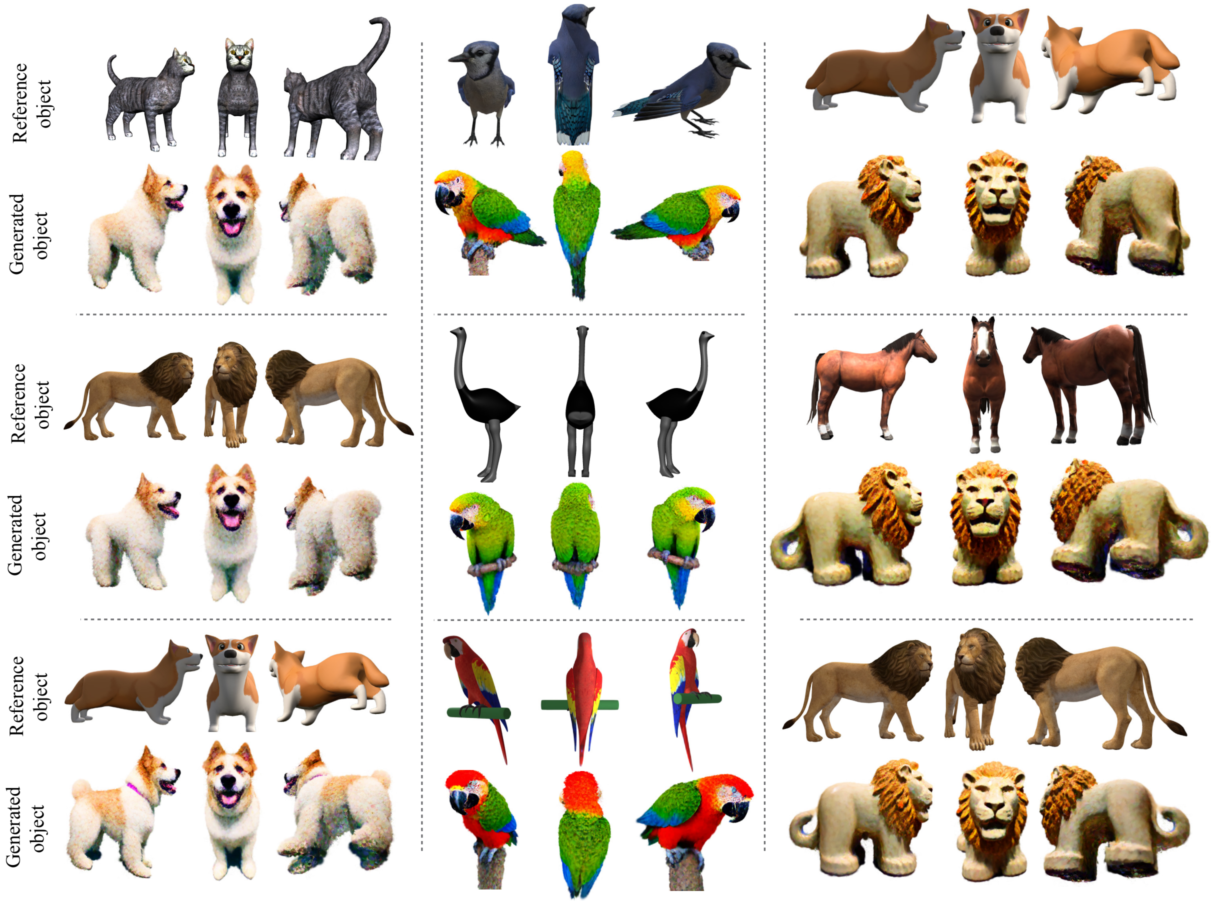

A high quality photo of a dog
A high quality photo of a corgi
A high quality photo of a wooden buddha head
A high quality photo of an ostrich
A zoomed out DSLR photo of a ceramic lion
A car made of cheese
3D representation generated without guidance from reference object
Reference object
Generated object
A high quality photo of a corgi
A zoomed out DSLR photo of a ceramic lion
A zoomed out DSLR photo of a ceramic lion
Reference object
Generated object
A car made of cheese
A high quality photo of a furry dog
A high quality photo of a furry dog
Reference object
Generated object
A zoomed out DSLR photo of a ceramic lion
A high quality photo of an ostrich
A high quality photo of a wooden buddha head
(a)
(b)
(c)

Figure 8. (Top) Illustration of 3D objects generated by SDS, as described in Figure 6). (Bottom) Illustration of 3D objects generated by MT3D using various reference objects for different input text prompts. The 3D reference objects are generated from (a) Point-E, (b) Magic3D, and (c) HiFA.

erate its representations. All experiments were conducted on 4 × A100 GPUs with a batch size of 32 and random seeds. We observed that the 3D representations generated by these state-of-the-art methods closely match those reported in the original papers. Across all prompts, MT3D consistently produces high-fidelity and geometrically coherent 3D representations.

A.4. Complex Prompts

In Figure 13, we conduct experiments with more complex text prompts including multiple objects and in different scenes. MT3D utilizes a single 3D representation to guide the generation process, which limits its ability to generate multiple objects or place objects within a particular scene. Across all complex prompts, the generated 3D representations exhibit a bias toward the geometry of the reference object. For example, in text prompt corresponding to ‘blue



(a) *A high quality photo of a furry dog* (b) *A high quality photo of a parrot* (c) *A zoomed out DSLR photo of a ceramic lion*

Figure 9. Illustration of 3D objects generated by our model using various high-fidelity objects corresponding to different input text prompts. The top row in each block represents the input high-fidelity object, and the bottom row shows the generated 3D asset.

jay standing on a large basket of rainbow macarons', MT3D generates a blue jay and a macaron but no basket. Similarly, for 'hamburger in restaurant', no restaurant was generated. Utilizing multiple reference objects may help address this issue. Extending MT3D to incorporate multiple reference objects, thereby enabling the generation of multiple objects and complex scenes, would be a valuable direction for future work.

A.5. Additional Results

In this section we provide additional qualitative comparisons between the proposed MT3D and state-of-the-art generators, including Fantasia3D, Magic3D, and HiFA. Our proposed MT3D generates geometrically consistent renderings in most cases. For instance, for the prompt "A car made of cheese," other state-of-the-art generators fail to maintain the geometric features of the car and instead render pieces of cheese. In contrast, MT3D preserves the car's geometry while applying a cheese texture. Additionally, several other examples in Figures 14,15 and 16, further validate the superior performance of our proposed method compared to other state-of-the-art generators.

A.6. High-fidelity 3D objects used to guide MT3D

In Figures 17, we present the high-fidelity 3D objects from Objaverse used to guide the generation with MT3D.

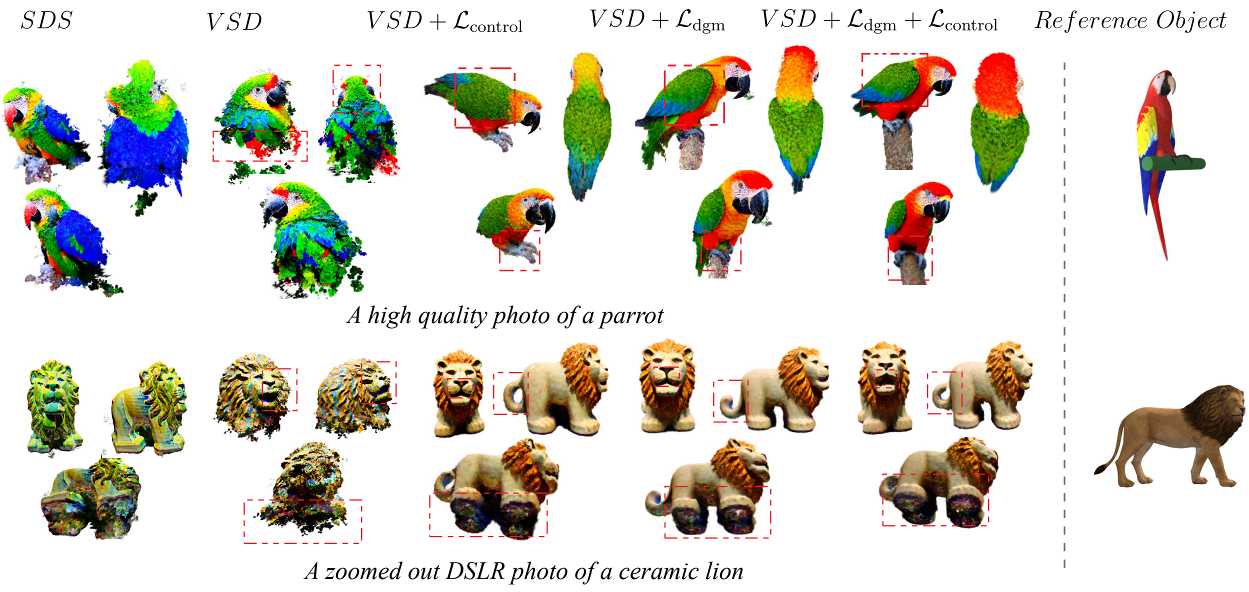

SDS
VSD
$VSD + \mathcal{L}_{\text{control}}$
$VSD + \mathcal{L}_{\text{dgm}}$
$VSD + \mathcal{L}_{\text{dgm}} + \mathcal{L}_{\text{control}}$
$Reference\ Object$
A high quality photo of a parrot
A zoomed out DSLR photo of a ceramic lion

Figure 10. The results of an ablation study of MT3D.

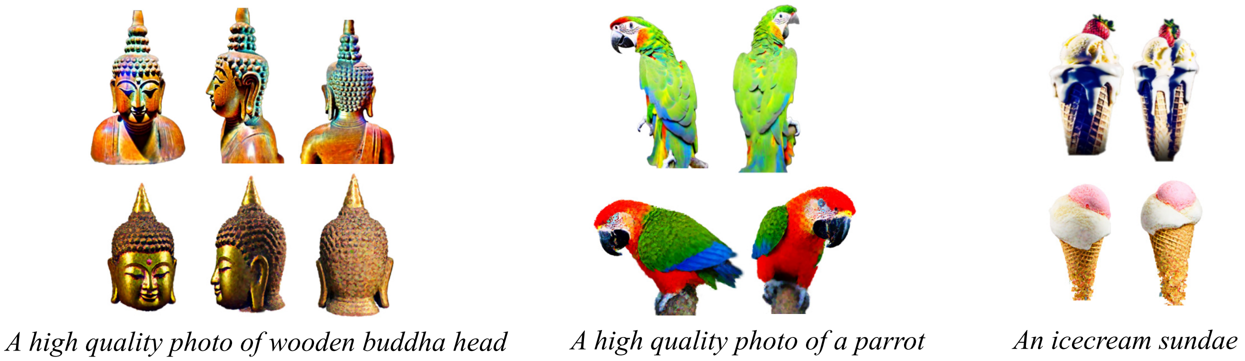
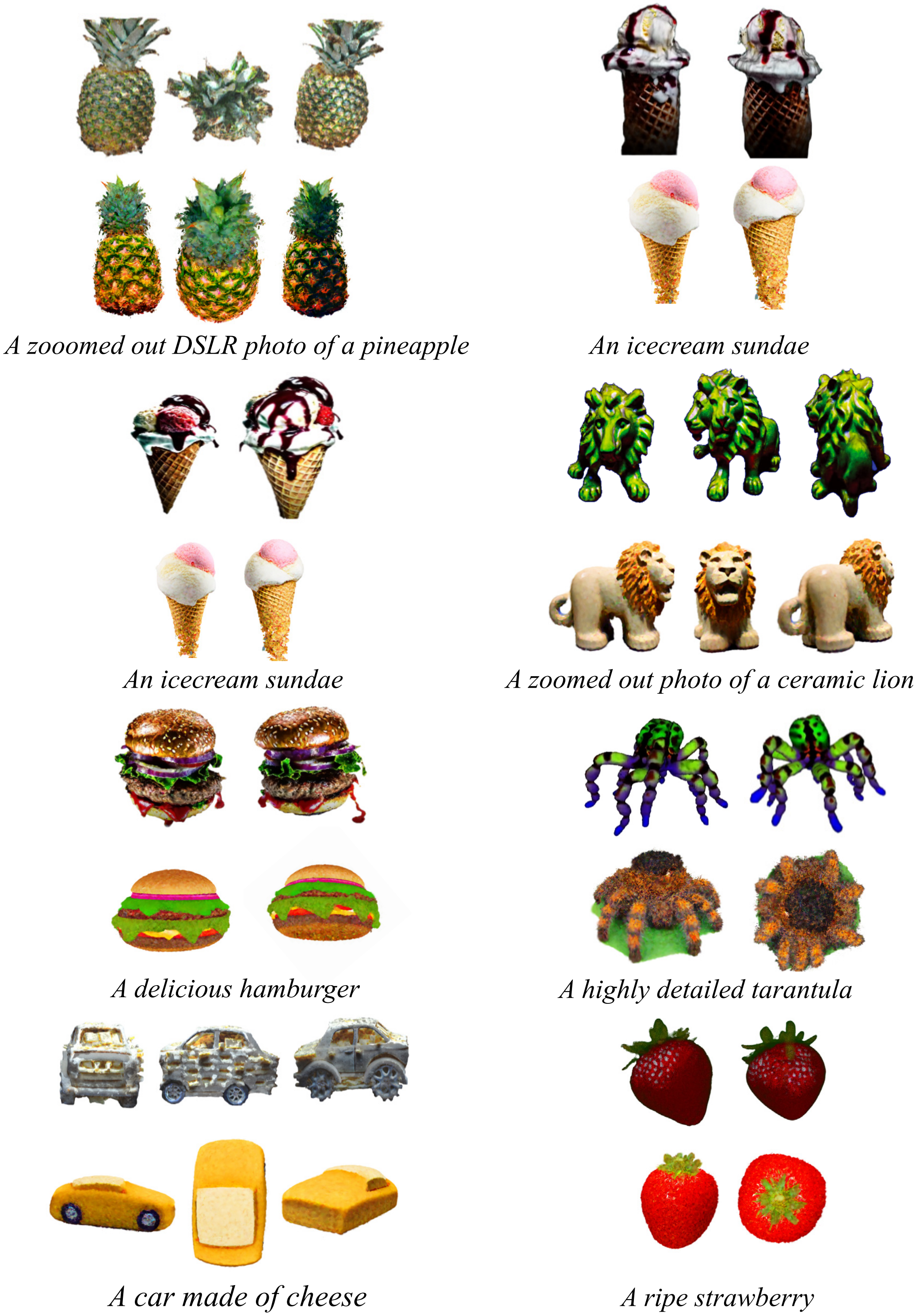

A high quality photo of wooden buddha head
A high quality photo of a parrot
An icecream sundae

Figure 11. Qualitative comparison of 3D assets generated by MT3D and HiFA. The top row for each prompt displays HiFA-generated objects, while the bottom row shows MT3D-generated assets.



A zooomed out DSLR photo of a pineapple

An icecream sundae

A delicious hamburger

A car made of cheese

An icecream sundae

A zoomed out photo of a ceramic lion

A highly detailed tarantula

A ripe strawberry

Figure 12. (Left) Qualitative comparison of 3D assets generated by MT3D and Fantasia3D. (Right) Qualitative comparison of 3D assets generated by MT3D and Magic3D. For each prompt, the top row shows assets from Fantasia3D (Left) and Magic3D (Right), while the bottom row displays those generated by MT3D.

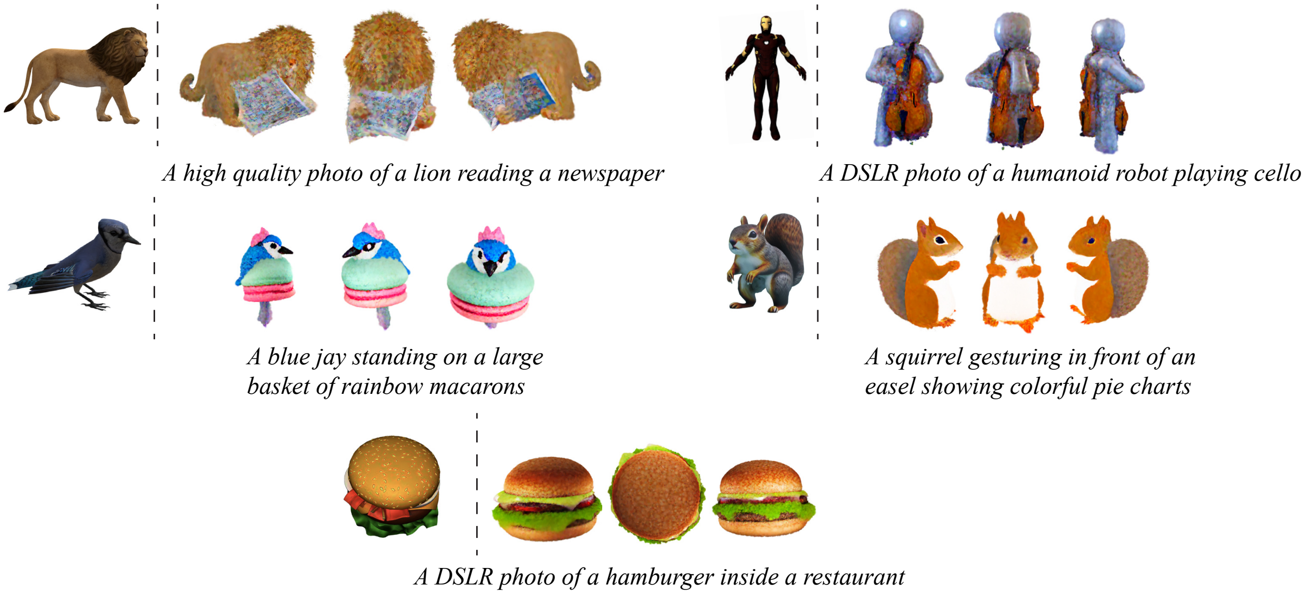

A high quality photo of a lion reading a newspaper
A DSLR photo of a humanoid robot playing cello
A blue jay standing on a large basket of rainbow macarons
A squirrel gesturing in front of an easel showing colorful pie charts
A DSLR photo of a hamburger inside a restaurant

Figure 13. Illustrations of 3D objects generated by MT3D using various complex prompts. For each text prompt, the left shows the high-fidelity reference object, while the right represents the 3D asset generated by MT3D using the reference object.

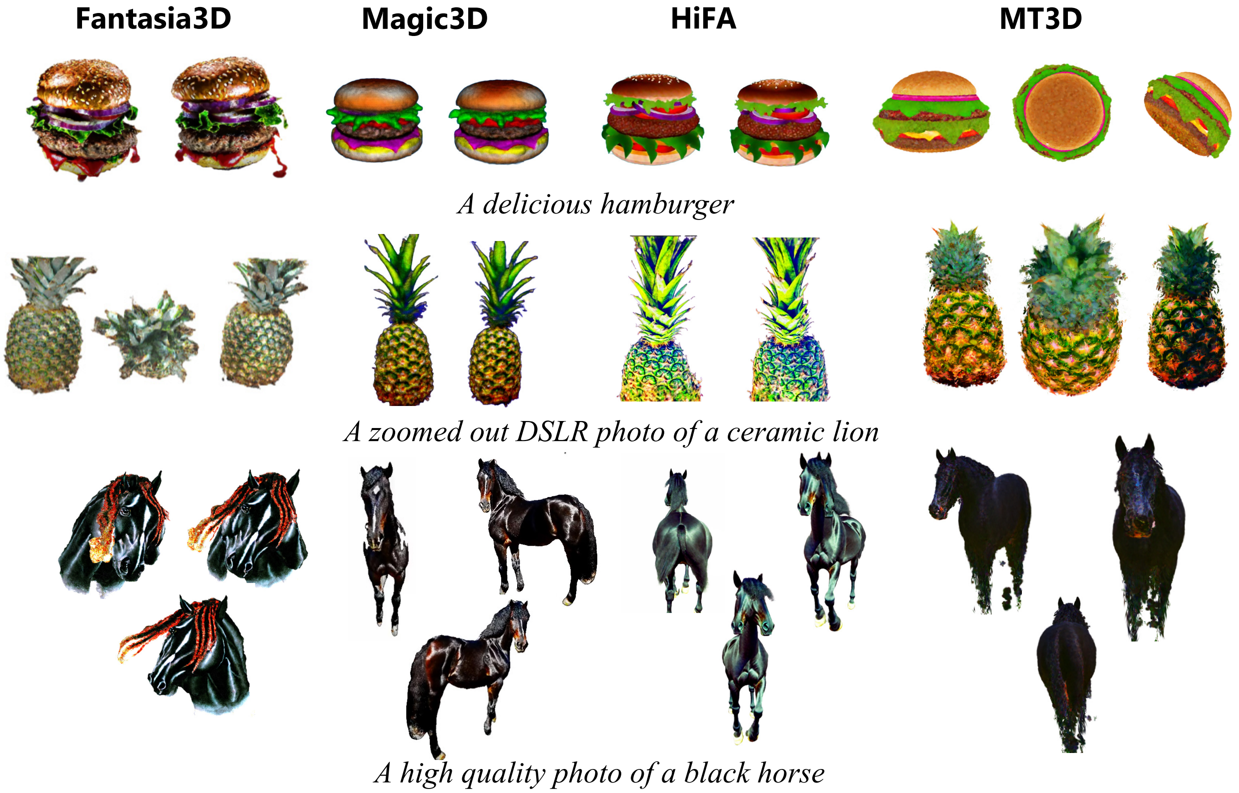

Fantasia3D
Magic3D
HiFA
MT3D
A delicious hamburger
A zoomed out DSLR photo of a ceramic lion
A high quality photo of a black horse

Figure 14. Additional qualitative comparison between the proposed MT3D and state-of-the-art generators, including Magic3D, Fantasia3D, and HiFA

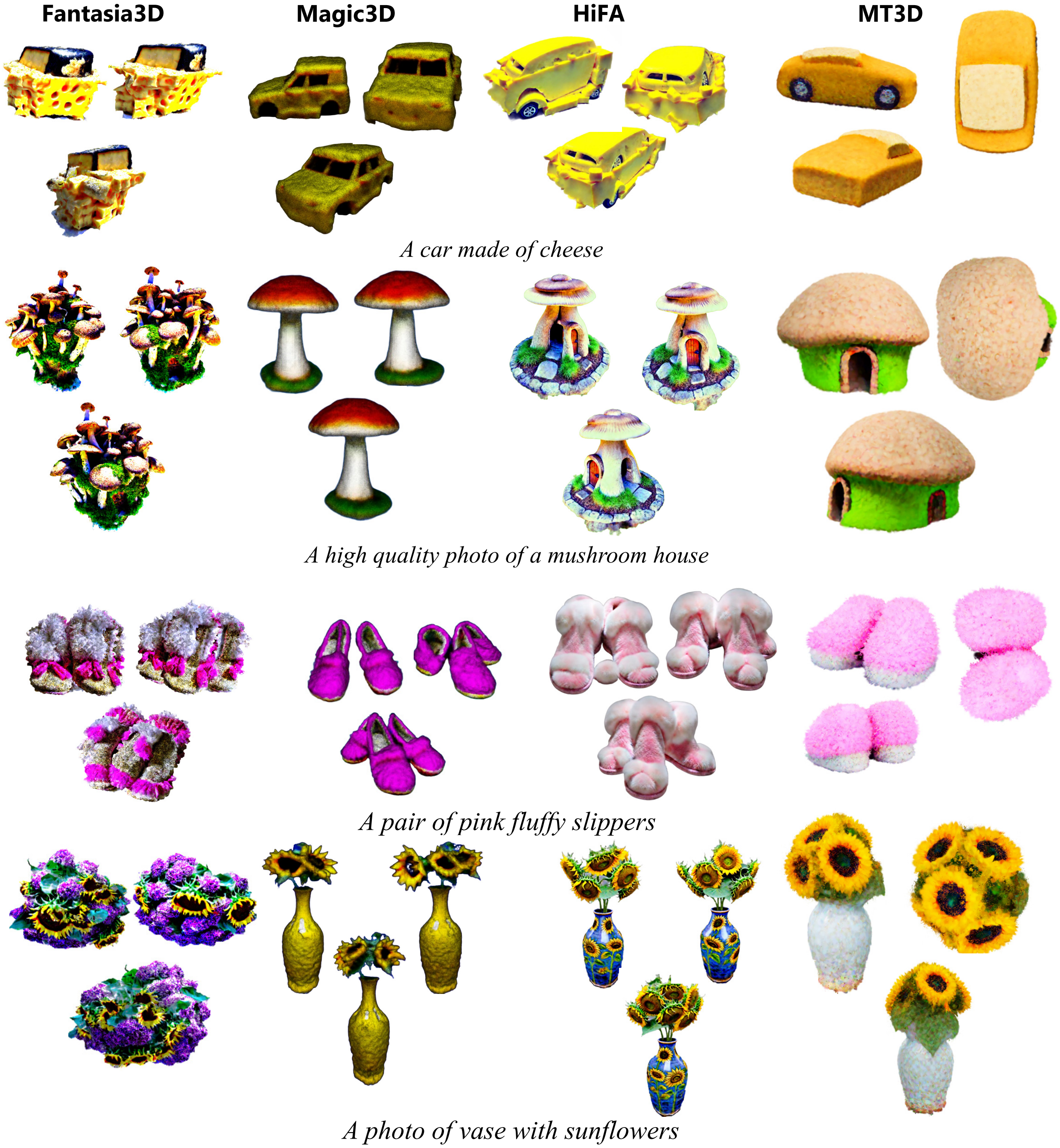
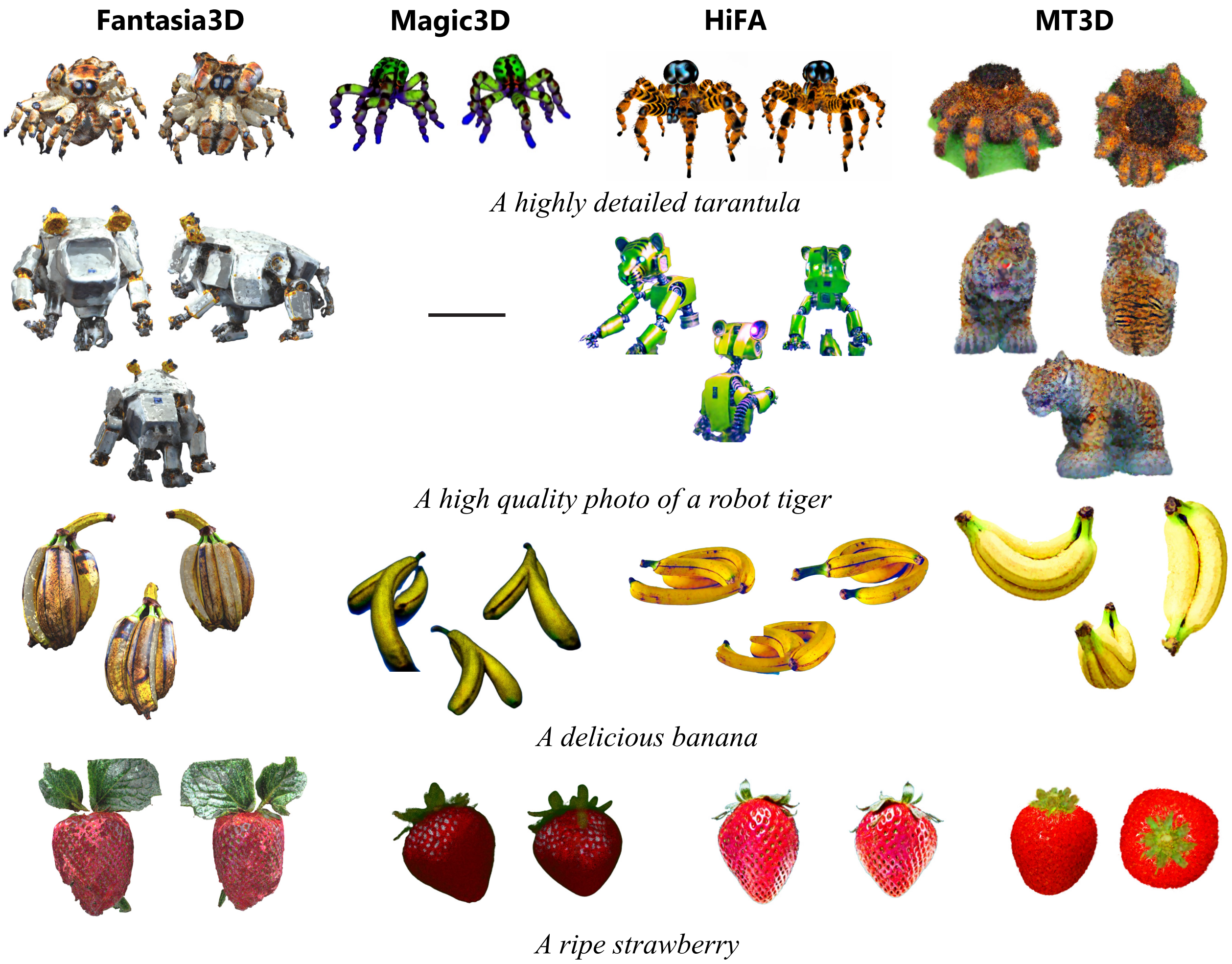

Fantasia3D
Magic3D
HiFA
MT3D
A car made of cheese
A high quality photo of a mushroom house
A pair of pink fluffy slippers
A photo of vase with sunflowers

Figure 15. More comparisons between MT3D and state-of-the-art generators



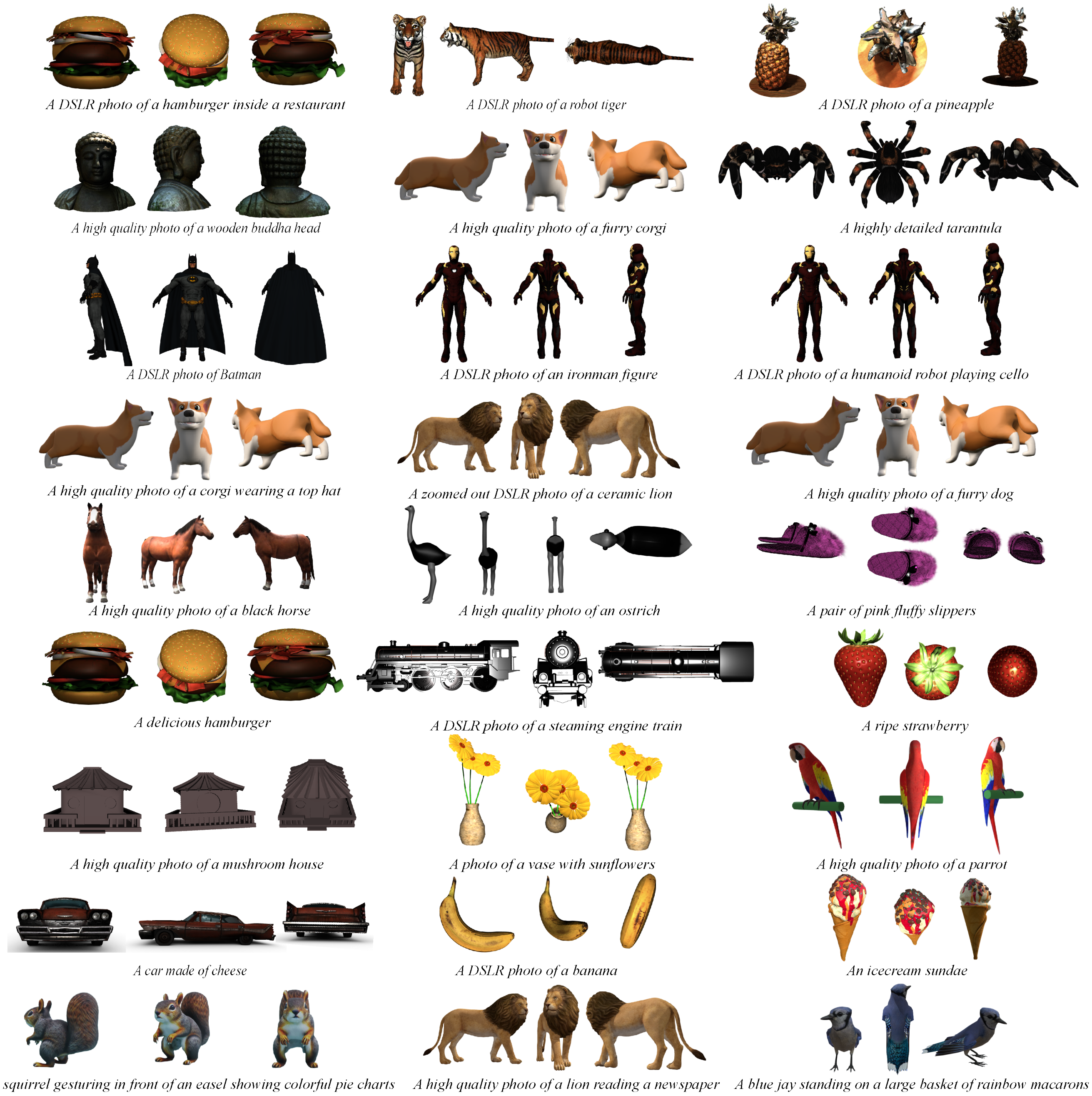
A highly detailed tarantula

A high quality photo of a robot tiger

A delicious banana

A ripe strawberry

Figure 16. More comparisons between MT3D and state-of-the-art generators



A DSLR photo of a hamburger inside a restaurant

A DSLR photo of a robot tiger

A DSLR photo of a pineapple

A high quality photo of a wooden buddha head

A high quality photo of a furry corgi

A highly detailed tarantula

A DSLR photo of Batman

A DSLR photo of an ironman figure

A DSLR photo of a humanoid robot playing cello

A high quality photo of a corgi wearing a top hat

A zoomed out DSLR photo of a ceramic lion

A high quality photo of a furry dog

A high quality photo of a black horse

A high quality photo of an ostrich

A pair of pink fluffy slippers

A delicious hamburger

A DSLR photo of a steaming engine train

A ripe strawberry

A high quality photo of a mushroom house

A photo of a vase with sunflowers

A high quality photo of a parrot

A car made of cheese

A DSLR photo of a banana

An icecream sundae

A squirrel gesturing in front of an easel showing colorful pie charts

A high quality photo of a lion reading a newspaper

A blue jay standing on a large basket of rainbow macarons

Figure 17. Illustrates high-fidelity 3D objects corresponding to various text prompts used for guiding MT3D throughout the paper.