Reflecting Reality: Enabling Diffusion Models to Produce Faithful Mirror Reflections

Supplementary Material

Contents

A Dataset 1
- A.1. Filtering out Spurious objects 1
- A.2 Preparation of MirrorBench 1

B Implementation Details 4
- B.1. Training Details: MirrorFusion 4
- B.2. Training Details for Baseline Methods 4
- B.3. Inference Details 4

C Additional Results 5
- C.1. More Results on Google Scanned Objects (GSO) 5
- C.2. Results on real-world scenes. 5
- C.3. Comparison with Commercial Products. . . . 6
- C.4. Robustness to pre-trained monocular depth estimation methods 6
- C.5. More Qualitative Comparisons 6

D Limitations and Social Impact 7

E Additional Details 7
- E.1. Results from recent T2I methods 7
- E.2. Text prompts used in the experiments 7
- E.3. Generation of Segmentation Masks for computing metrics 8

A. Dataset

Our dataset consists of $198,204$ rendered images from $66,068$ objects: $58,115$ objects from Objaverse [5] and $7,953$ from the ABO [3] dataset. We utilize captions provided by Cap3D [29] during training. We provide more details in Sec. 3. To illustrate the diversity in 3D objects, floor textures and HDRI backgrounds, we present more samples in Fig. 10 and 11.

A.1. Filtering out Spurious objects

We discuss how we filter 3D objects from Objaverse [5] and Amazon Berkeley Objects (ABO) [3] in Sec. 3.1. In spite of the initial filtering, we observe some "spurious" objects, for which the reflection is not visible in the mirror. Algorithm 1 illustrates the pseudo-code to identify such "spurious" objects. Specifically, using Blender's Python API, we check the material property of each *child* in the input mesh $\mathcal{M}$ of a 3D object. We expect the 3D objects to be in standard 3D formats: *"*.glb, *.gltf, *.obj, *.fbx"*. If any node in the material property has the attributes: "Mix-Shader", and the name of the input to this node is "Fac," and the name of the linked node is "Light Path", then we observe that the reflection of such a 3D model does not appear in the mirror. We prune out such objects from the initial filtered list. The new filtered list will be made public along with the dataset for future research.

Algorithm 1 Determine if a 3D Object is Spurious

Input: A 3D model $\mathcal{M}$
Output: True, if a 3D model is spurious, else False

```
for C ← child ∈ M do
    for T ← material ∈ C do
        for N ← node ∈ T.material do
            if (N.name == "Mix-Shader")
            and (N.input.name == "Fac")
            and (N.linked.name == "Light Path") then
                RETURN(True)
            end if
        end for
    end for
end for
```

A.2. Preparation of MirrorBench

MirrorBench aims to benchmark various generative models at the task of generating perfect mirror reflections. MirrorBench is created by sampling around $1,000$ objects from SynMirror, with 3 rendered samples per object, totalling to $2,991$ samples. Fig. 12 shows samples of MirrorBench, which consist of two types:

1. **"Unknown"** class objects, referring to categories not present in the training set. We take the first 500 objects from Objaverse in **"Unknown"** category, sorted in the increasing order of category frequency and keep the remaining categories in the training set as **"Known"** categories. There are 1494 samples generated from the objects of **"Unknown"** category.
2. **"Known"** class objects, referring to categories included in the training set. There are 1497 images from this category. This includes renderings from around 250 objects from Objaverse and around 250 objects from ABO.

Figure 10. **Samples from SynMirror.**

Figure 11. **Samples from SynMirror.**

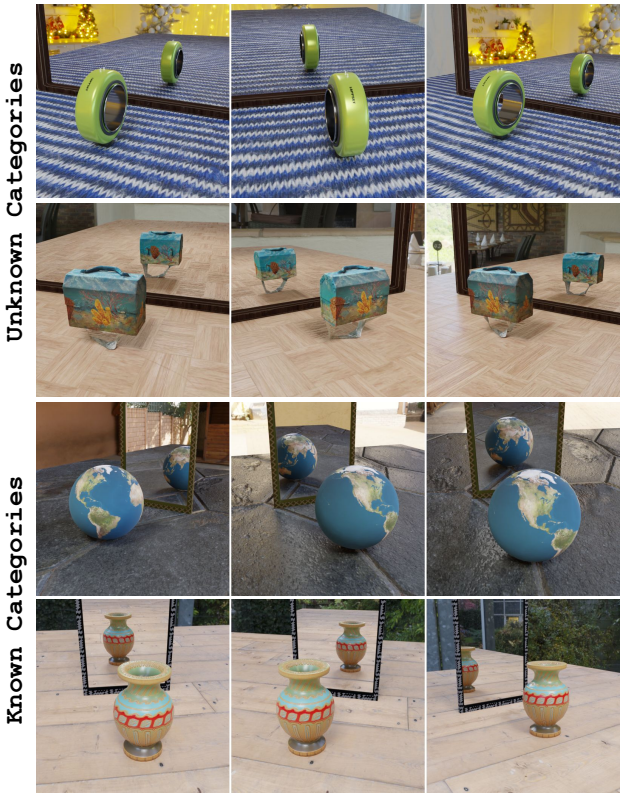

Figure 12. **Samples from MirrorBench.** The first two rows contain samples from **"Unknown"** categories and the bottom two rows contain samples from **"Known"** categories. Notice the challenging nature of MirrorBench. We provide more details in Appendix A.2

B. Implementation Details

B.1. Training Details: MirrorFusion

We follow the BrushNet [14] architecture for MirrorFusion and provide depth conditioning as discussed in Sec. 4.2. The Generation and Conditional U-Net weights are initialized from the Stable Diffusion v1.5 [42] checkpoint. During training, the weights of the generation U-Net are kept frozen, while the weights of the conditioning network are updated. The extra channels processing the down-sampled depth and mask images in the first convolution layer of the conditioning U-Net are initialized to zero. We train MirrorFusion on SynMirror, using the original input image resolution of 512×512. We utilize the AdamW optimizer with a learning rate $1e-5$. We train our model for $20,000$ steps on 8 NVIDIA A6000 GPUs with an effective batch size of 16, which takes around 12 hours. During training, we randomly drop text prompts 20% of the time to allow the model to take cues from the input depth map. We find the checkpoint at $15,000$ to produce the best qualitative results and use it for further inference. We also run an additional experiment where we make the generation U-Net trainable. We call this model **MirrorFusion***. We use the same training hyper-parameters and consider the checkpoint at $17,000$ steps. From Fig. 17 and Fig. 18, we can see improved results compared to the frozen generation U-Net. However, the VRAM requirements and training time almost double, due to the increase in the number of trainable parameters.

B.2. Training Details for Baseline Methods

Fine-tuning of BrushNet [14]. Keeping the generation U-Net frozen, we fine-tune BrushNet using the input mask and masked image using the same hyperparameters used to train MirrorFusion. We do not randomly drop text prompts and select the checkpoint at $17,000$ steps for evaluation. We refer to this model as ***"BrushNet-FT"*** in Sec. 5 of the main paper and compare our results against it. We found that initializing the weights from the Stable Diffusion v1.5 [42] checkpoint was superior as compared to initializing from the pre-trained BrushNet [14] checkpoint.

B.3. Inference Details

During inference, we set the classifier free guidance scale (CFG) to 7.5 and use the UniPC scheduler [63] for 50 timesteps across all experiments.

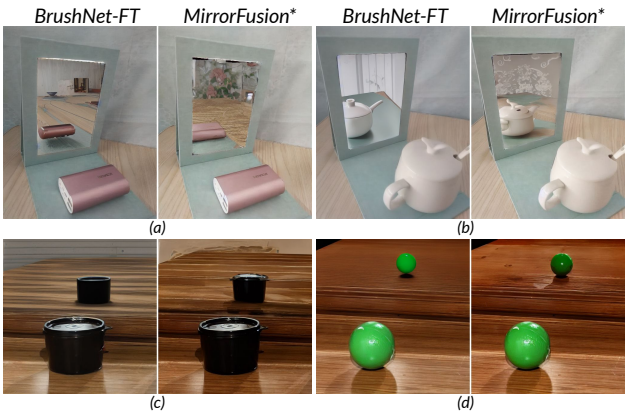

Figure 13. **Performance on Real-world scenes** We show results on images from MSD [56] dataset (a) & (b) and examples from images captured using a smartphone device (c) & (d). Appendix C.2 describes the experimental details and text prompts used for the inference. We observe that "BrushNet-FT" does not generate accurate reflections, whereas our method is able to generate plausible reflections on the mirror.

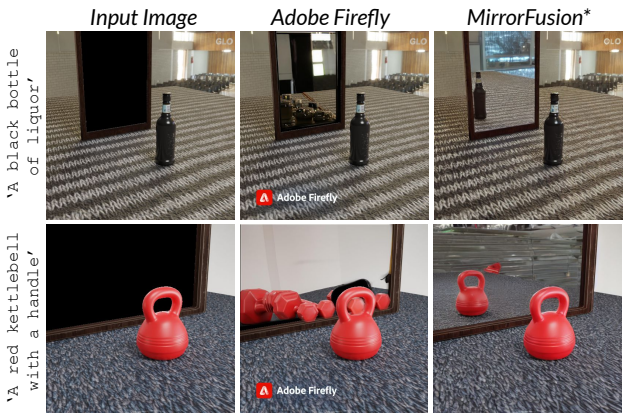

Figure 14. **Qualitative Comparison with Commercial Products** We compare our results with Adobe Firefly. Our method is significantly better than the existing commercial product. This highlights the challenging nature of the task and the effectiveness of our proposed method in addressing it.

C. Additional Results

C.1. More Results on Google Scanned Objects (GSO)

We provide additional results on 3D models from Google Scanned Objects (GSO) [7] in Fig. 15. GSO contains real-world scanned objects. We create renderings using these objects with the pipeline discussed in Sec. 3. We notice that our method MirrorFusion* consistently generates accurate reflections of objects and floors in the mirror. However, *"BrushNet-FT"*, is not able to generate the reflection of the floor correctly in image with blue ball (Fig. 15 (o), and (p)) and carton (Fig. 15 (l)) Further, it does not get the appearance of the pencil-box right, as shown in Fig. 15 (g) and (h). Additionally, it generates the reflection with the wrong structure in Fig. 15 (c) and (d). These results further substantiate the generalization capabilities of our method.

Text prompts used for results in Fig. 15 are as follows:

- **(a) & (b).** *"A perfect plane mirror reflection of a sofa with purple cushioning."*
- **(c) & (d).** *"A perfect plane mirror reflection of a yellow chair."*
- **(e) & (f).** *"A perfect plane mirror reflection of a white stool with a purple top."*
- **(g) & (h).** *"A perfect plane mirror reflection of a purple bag with bluish circular patterns."*
- **(i) & (j).** *"A perfect plane mirror reflection of a camouflaged military-style bag."*
- **(k) & (l).** *"A perfect plane mirror reflection of a cardboard box on a patterned floor."*
- **(m) & (n).** *"A perfect plane mirror reflection of a yellow and white mug on a grey surface."*
- **(o) & (p).** *"A perfect plane mirror reflection of a blue ball with an orange cover."*

C.2. Results on real-world scenes.

We present real-world examples from the MSD [56] dataset in Fig. 13 (a) and (b), utilizing the ground truth (GT) masks provided within the dataset as the corresponding mirror masks. Since our method requires depth, we infer it from Marigold and normalize it as described in Sec. 4.2.1. We observe that the baseline method fails to position the object accurately and produces incorrect color in Fig. 13 (a). In contrast, our method generates better reflections on the mirror.

We also capture more examples from a hand-held smartphone device in Fig. 13 (c) & (d). We manually annotate the mask corresponding to the mirror location and infer the depth from Marigold [15] as described above. Similar to the previous observation, our method preserves the shape of the object. Check the lid in Fig. 13 (c) and the roundness of the ball in Fig. 13 (d). These results show that our method generates better reflections than the baselines on real-world settings. Our method shows promising results on real-world settings, but still has scope for improvement, showing the challenging nature of this task.

Text prompts used for generating the results in Fig. 13 are as follows:

- **(a).** *"A perfect plane mirror reflection of a rose gold colored portable power-bank."*
- **(b).** *"A perfect plane mirror reflection of a white ceramic teapot."*
- **(c).** *"A perfect plane mirror reflection of a black round box with a black lid on it."*
- **(d).** *"A perfect plane mirror reflection of a green color round ball."*

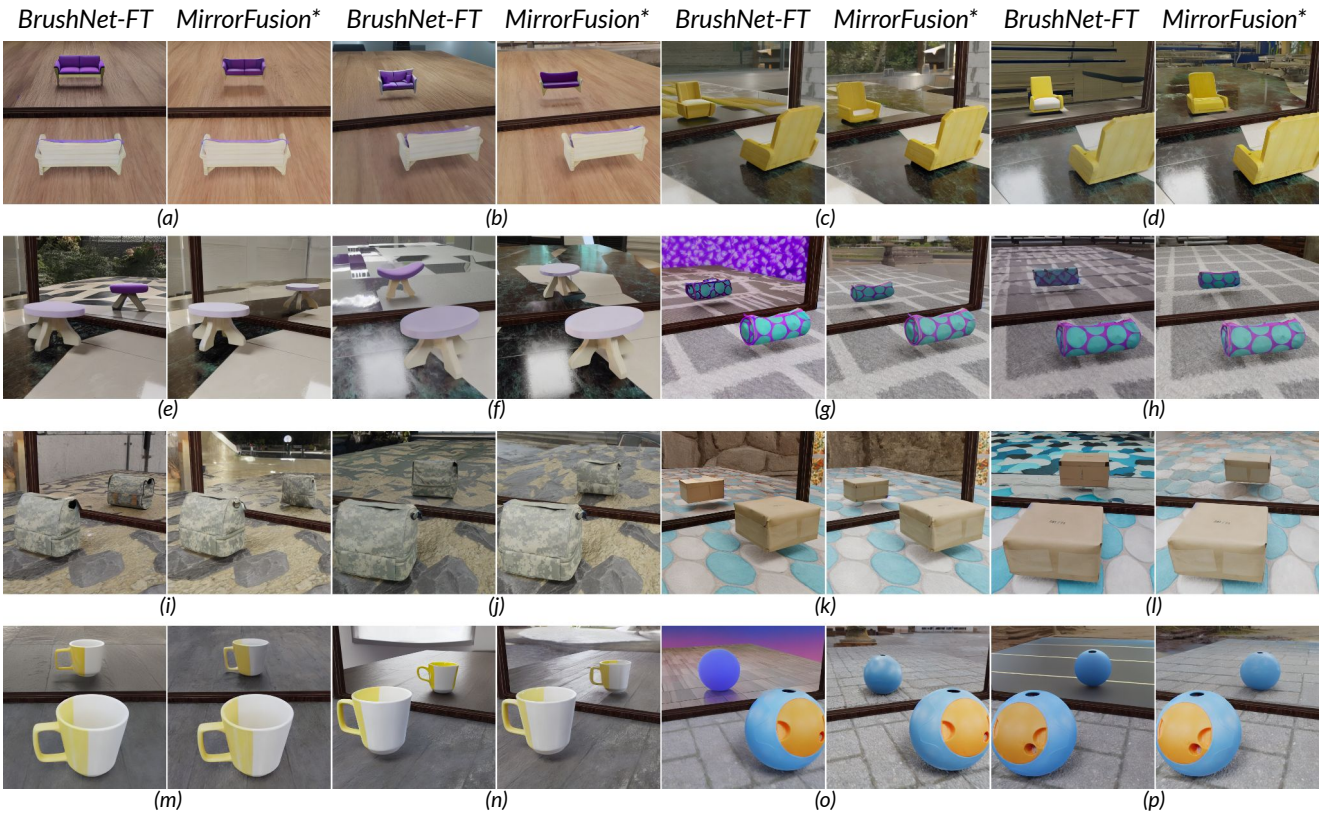

Figure 15. **Qualitative Comparison on unseen 3D assets from GSO.** We show results from (a) & (b) "3D Dollhouse Sofa", (c) & (d) "3D Dollhouse Swing", (e) & (f) "3D Dollhouse TablePurple", (g) & (h) "Big Dot Aqua Pencil Case", (i) & (j) "Digital Camo Double Decker Lunch Bag", (k) & (l) "INTERNATIONAL PAPER Willamette 4 Brown Bag" , (m) & (n) "Room Essentials Mug White Yellow" and (o) & (p) "Toys R Us Treat Dispenser Smart Puzzle Foobler". Appendix C.1 describes how images are generated and text-prompts used for the inference. We observe that "BrushNet-FT" does not generate accurate reflections in (c),(d),(f),(g),(h) whereas our method is able to generate correct reflections on the mirror.

C.3. Comparison with Commercial Products.

We compare our method with commercial products such as Adobe Firefly in Fig. 14. Our method significantly outperforms existing commercial solutions. Results from Fig. 14 highlight the challenging nature of the task of generating plausible mirror reflections and the critical gap that exists in current state-of-the-art methods. Text prompts used in Fig. 14 are as follows:

- 1^{st} **row.** *"A perfect plane mirror reflection of a black bottle of liquor."*
- 2^{nd} **row.** *"A perfect plane mirror reflection of a red kettle-ball with a handle."*

C.4. Robustness to pre-trained monocular depth estimation methods

Our method is invariant to the choice of the pre-trained monocular depth estimation method. We present results from two state-of-the-art methods, Marigold [15] and DepthAnythingV2 [55], in Fig. 16. We observe minimal variation in the generation of reflections between both options, thereby confirming the robustness of our approach to the preference of the pre-trained monocular depth estimation method.

Text prompts for Fig. 16 are as follows, each row uses the same text prompt:

- 1^{st} **row.** *"A perfect plane mirror reflection of a rectangular cabinet with a door, two drawers, a truncated triangular base, and a triangular top."*
- 2^{nd} **row.** *"A perfect plane mirror reflection of a swivel chair with curved backrest, slanted seat, curved armrests, and a triangular top."*

C.5. More Qualitative Comparisons

As discussed in Sec. 5, we compare our method with zero-shot baselines, denoted by *"-ZS"* and baselines trained on SynMirror, denoted by *"-FT"*. We provide additional results in Fig. 17 and 18. Consistent with the findings in the main paper, our method generates better mirror reflections while preserving the fidelity of both the object's appearance

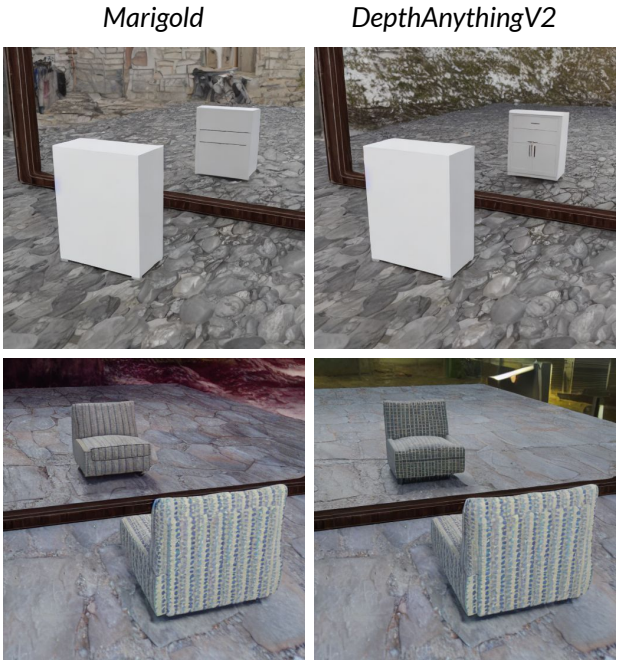

Figure 16. **Choice of pre-trained monocular depth estimation method during inference.** We observe negligible differences in the reflection generation across both choices, Marigold [15] and DepthAnythingV2 [55], supporting the stability of our method regardless of the chosen pre-trained monocular depth estimation technique. We use "Marigold" in all our experiments.

and the floor.

Fig. 17 Each row in this figure uses the same text prompt. Text prompts are as follows:

- 1^{st} **row.** *"A perfect plane mirror reflection of a multifunctional electronic device, including HDMI Blu-ray player, stereo receiver, amplifier, CD, and DVD player."*
- 2^{nd} **row.** *"A perfect plane mirror reflection of a red flashlight with a metal pipe."*
- 3^{rd} **row.** *"A perfect plane mirror reflection of a red kettlebell with a handle."*
- 4^{th} **row.** *"A perfect plane mirror reflection of a concrete block."*
- 5^{th} **row.** *"A perfect plane mirror reflection of a wooden barrel."*

Fig. 18 Each row in this figure uses the same text prompt. Text prompts are as follows:

- 1^{st} **row.** *"A perfect plane mirror reflection of a large, red, rusty metal barrel."*
- 2^{nd} **row.** *"A perfect plane mirror reflection of a small stuffed animal toy."*
- 3^{rd} **row.** *"A perfect plane mirror reflection of a modern office chair with a blue upholstered seat, back, and headrest."*
- 4^{th} **row.** *"A perfect plane mirror reflection of a Gaft Shower Gel Box."*
- 5^{th} **row.** *"A perfect plane mirror reflection of a black cowboy hat."*

D. Limitations and Social Impact

Limitations. As our method leverages synthetic data to train a model capable of producing realistic mirror reflections, the model still has scope for improvement in generating reflections for highly complex objects and scenarios. Although our model generates plausible results on real-world images, there is significant scope for improvement, which can be achieved by using more advanced photo-realistic simulators or collecting large-scale real-world images. We aim to address these issues in our future work.

Social Impact. Our method uses diffusion-based generative models, which, despite their potential, can be exploited for spreading misinformation. Therefore, it is crucial to use these models responsibly.

E. Additional Details

E.1. Results from recent T2I methods

We present additional results from the recent Stable Diffusion 3 [8] model in Fig. 19. Text prompts are generated by using the prefix: *"A perfect plane mirror reflection of"* and suffix: *"in front of the mirror positioned at an angle with respect to the mirror."* to the object description of the input image. We observe that standalone text-to-image methods are inadequate in generating controlled and realistic mirror reflections.

E.2. Text prompts used in the experiments

This section provides the text prompts for the image generations in the main paper.

Figure 1. Each row in this figure uses the same text prompt. Text prompts are as follows:

- **First row.** *"A perfect plane mirror reflection of a swivel chair with a curved backrest, slanted seat, slender metal frame, and padded seat and backrest."*
- **Second row.** *"A perfect plane mirror reflection of a large red, yellow, and black industrial cement mixer."*

Figure 2. Text prompts are already mentioned in the Figure of the main paper.

Figure 5. Text prompts are as follows:

- **(a).** *"A perfect plane mirror reflection of a white golf ball with a red stripe and the letter O on it."*
- **(b).** *"A perfect plane mirror reflection of a chair with a curved slatted frame, tufted backrest, and curved seat."*

Figure 6. Text prompts are as follows:

- **(a).** *"A perfect plane mirror reflection of a modern wooden chaise lounge with a white cushion."*

- **(b).** *"A perfect plane mirror reflection of a swivel chair with a curved backrest, slender armrest, and swivel base."*
- **(c).** *"A perfect plane mirror reflection of a black cylindrical with a lid."*
- **(d).** *"A perfect plane mirror reflection of a wooden box with intricate floral and heart-shaped carvings on each side, featuring a dark brown hue with visible wood grain texture."*

Figure 7. Each row in this figure uses the same text prompt. Text prompts are as follows:

- 1^{st} **row.** *"A perfect plane mirror reflection of a large red, yellow, and black industrial cement mixer."*
- 2^{nd} **row.** *"A perfect plane mirror reflection of a gold lipstick container."*
- 3^{rd} **row.** *"A perfect plane mirror reflection of a cylindrical object with a cream-colored exterior and a central hollow core; vertical seams divide the outer surface."*
- 4^{th} **row.** *"A perfect plane mirror reflection of a weathered wooden treasure chest with metal reinforcements, large metal ring on the side, and mossy accents."*
- 5^{th} **row.** *"A perfect plane mirror reflection of a grey cabinet with gold legs and chest of drawers."*
- 6^{th} **row.** *"A perfect plane mirror reflection of a black stone with intricate swirl designs on it."*

Figure 8. Each row in this figure uses the same text prompt. Text prompts are as follows:

- 1^{st} **row.** *"A perfect plane mirror reflection of a slanted-top cuboid footstool."*
- 2^{nd} **row.** *"A perfect plane mirror reflection of a footstool with a cuboid base, spherical top, seat, and backrest."*

Figure 9. Text prompts are as follows:

- **(a).** *"A perfect plane mirror reflection of a white ceramic bowl on a textured gray surface.."*
- **(b).** *"A perfect plane mirror reflection of a camouflaged military-style bag"*

E.3. Generation of Segmentation Masks for computing metrics

We compare the accuracy of the geometry of the generated reflection by comparing IoU between the segmentation mask of the reflection in the ground-truth object and the segmentation mask of the reflection in the generated object in Sec. 5. We utilize SAM to generate these segmentation masks. We provide initial seed points to SAM [16] along with a rough bounding box. SAM then segments out the reflection of the object in ground truth as well as the generated image. Camera viewpoint variations within our dataset pose a challenge for reliable seed point initialization. We address this by manually creating a mapping to select seed points based on the camera pose. To accelerate the evaluation, we cache the segmentation masks of the ground-truth images.

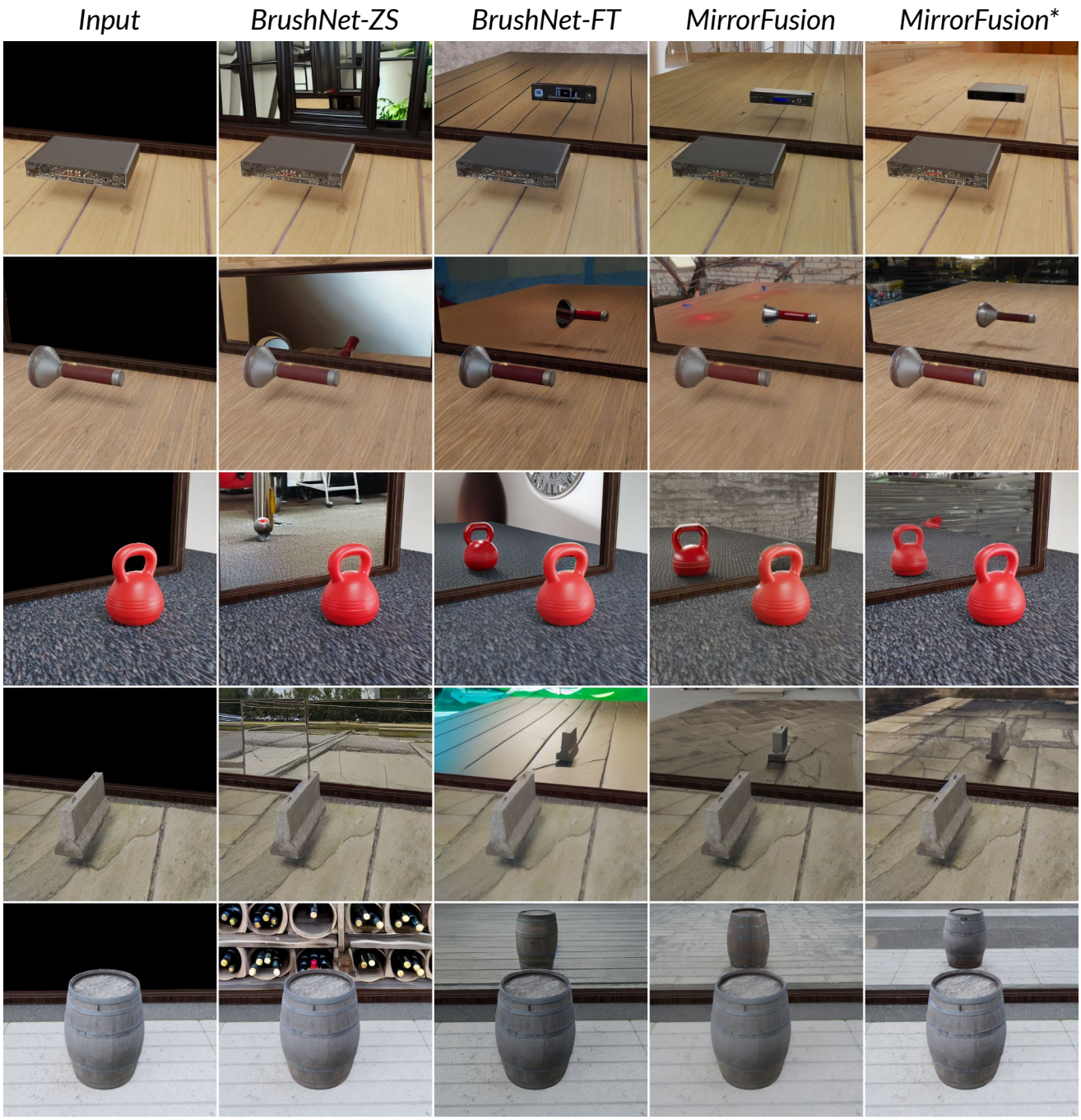

Figure 17. **Qualitative Comparison.** We observe that the state-of-the-art inpainting method "BrushNet-ZS" is not able to generate plausible reflections (2^{nd} **column**). "BrushNet-FT" which is fine-tuned on SynMirror is able to generate plausible reflections, 3^{rd} **column**, but fails to accurately get the shape of the object. For example, the top surface of "dvd-player" in 1^{st} row is completely missing. The "flashlight" reflection's structure and appearance do not correspond with the object (2^{nd} row). Compared to these baselines MirrorFusion generates plausible reflections. Still there is issue in the shape of the "flashlight" in 2^{nd} row. These issues are mitigated by MirrorFusion*, which generates realistic, plausible and geometrically accurate reflections on the mirror.

Figure 18. **Qualitative Comparison.** Similar to the observation in Fig. 17, we observe that the state-of-the-art inpainting method "BrushNet-ZS" is not able to generate plausible reflections (2^{nd} **column**). "BrushNet-FT" which is fine-tuned on SynMirror is able to generate plausible reflections, 3^{rd} **column** but fails to get shape of the object in the reflection. For example, observe the "chair" in 3^{rd} row, the head of the chair is missing. The pose of the toy in 2^{nd} row does not correspond to that of the real object. Compared to this MirrorFusion and MirrorFusion* generates plausible reflections on the mirror.

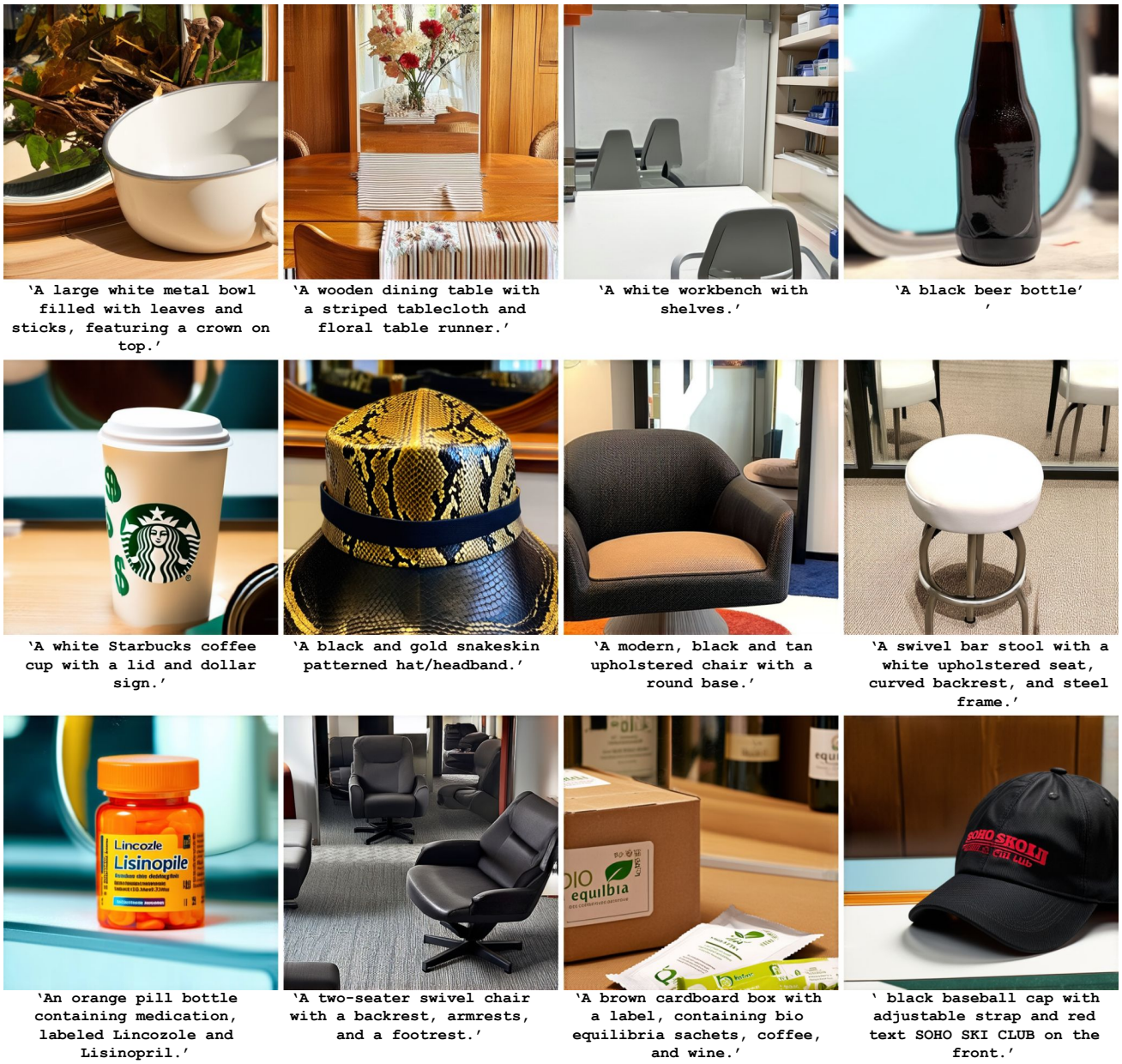

Figure 19. **Additional results of images generated from Stable Diffusion 3 [8].** Text-to-image models struggle to produce consistent and controlled mirror reflections when prompted to generate them. We use the prefix *"A perfect plane mirror reflection of"* and suffix *"in front of the mirror positioned at an angle with respect to the mirror."* along with the object description.