Reflecting Reality: Enabling Diffusion Models to Produce Faithful Mirror Reflections

Ankit Dhiman[1,2*] Manan Shah[1*] Rishubh Parihar[1] Yash Bhalgat[3]
Lokesh R Boregowda[2] R Venkatesh Babu[1]

[1]Vision and AI Lab, IISc Bangalore [2]Samsung R & D Institute India - Bangalore
[3]Visual Geometry Group, University of Oxford

Figure 1. We present MirrorFusion, a diffusion-based inpainting model, which generates high-quality geometrically consistent and photo-realistic mirror reflections given an input image and a mask depicting the mirror region. Our method shows superior quality generations as compared to previous state-of-the-art diffusion-based text-to-image and inpainting methods. All the images were generated by prefixing the mirror text prompt: "A perfect plain mirror reflection of " to the input object description.

Abstract

We tackle the problem of generating highly realistic and plausible mirror reflections using diffusion-based generative models. We formulate this problem as an image inpainting task, allowing for more user control over the placement of mirrors during the generation process. To enable this, we create SynMirror, a large-scale dataset of diverse synthetic scenes with objects placed in front of mirrors. SynMirror contains around $198K$ *samples rendered from* $66K$ *unique 3D objects, along with their associated depth maps, normal maps and instance-wise segmentation masks, to capture relevant geometric properties of the scene. Using this dataset, we propose a novel depth-conditioned inpainting method called* MirrorFusion*, which generates high-quality geometrically consistent and photo-realistic mirror reflections given an input image and a mask depicting the mirror region.* MirrorFusion *outperforms state-of-the-art methods on SynMirror, as demonstrated by extensive quantitative and qualitative analysis. To the best of our knowledge, we are the first to successfully tackle the challenging problem of generating controlled and faithful mirror reflections of an object in a scene using diffusion based models. SynMirror and* MirrorFusion *open up new avenues for image editing and augmented reality applications for practitioners and researchers alike. The project page is available at: https://val.cds.iisc.ac.in/reflecting-reality.github.io/.*

*Equal Contribution.

1. Introduction

Recent diffusion-based generative models [11, 12, 39, 43, 47] have achieved remarkable results, producing visually appealing images across various domains. These models can be conditioned using several modalities, such as text [11], depth-image [50], sketch [17], for controlled generation [57, 60, 62], enabling various interesting applications. Despite their success, these models struggle to capture subtle geometric cues such as shadows, lighting and specular reflections, as noted in previous studies [45, 54]. Specifically, the task of generating realistic and controllable mirror reflections remains an unsolved challenge. Existing methods, which tackle perspective issues [51, 62] and address specular reflections for object removal [54] do not address mirror reflections in particular.

To illustrate this limitation, we prompt Stable Diffusion-2.1 [42] with the instruction to generate a scene with a mirror reflection. Fig. 2 shows that Stable Diffusion-2.1 fails to generate plausible and consistent mirror reflections. Further, various state-of-the-art inpainting methods such as Stable Diffusion Inpainting [42] and PowerPaint [65] also fail at the task of generating plausible and controlled reflections, as shown in Fig. 1 (b) & (c).

In this work, we pose the problem of generating mirror reflections as an *Image Inpainting task*. This formulation provides two distinctive advantages: (1) Posing it as an inpainting method aids the reflection generation process to take cues from the input image and (2) allow greater control on the placement of mirrors.

Existing datasets for tasks such as mirror segmentation, detection or novel-view synthesis as shown in Tab. 1 contain mirrors reflecting generic backgrounds and also lack the scale making them unsuitable for the task of training generative models for generating photo-realistic mirror reflections. Therefore, to address this, we introduce SynMirror, a training dataset and MirrorBench, a benchmark dataset designed to train and evaluate the capability of generative models to produce photo-realistic mirror reflections. SynMirror contains $198,204$ samples from rendering $66,068$ unique 3D objects sourced from Objaverse [5] and Amazon Berkeley Objects (ABO) [3].

Samples from SynMirror are created by rendering synthetic scenes with objects placed in front of a mirror using the cycles rendering engine from Blender. To further obtain instance segmentation, depth and normal maps, we utilize `Blenderproc` [6], a procedural Blender pipeline for photorealistic rendering. The generated scenes have diverse mirrors, floor textures and backgrounds.

When recent inpainting methods such as BrushNet [14] are fine-tuned on SynMirror, we observe that they fail in generating correct geometry and depth of an object in the mirror reflection as shown in Fig. 7. We hypothesize that providing additional cues, such as depth maps, could help to alleviate the issue in generating geometrically consistent reflections on the mirror. To this end, we propose *MirrorFusion*, a depth conditioned inpainting method that generates high-quality controlled and photo-realistic mirror reflections. Our method significantly outperforms state-of-the-art diffusion-based inpainting methods on SynMirror, as evidenced by extensive quantitative and qualitative evaluations.

Figure 2. Images generated from Stable Diffusion 2.1 [42]. Text-to-image models, when prompted to generate reflections, struggle to generate consistent and controlled mirror reflections.

We summarize our contributions below:

- To the best of our knowledge, we are the first to address and tackle the challenging problem of generating controlled photo-realistic and geometrically consistent mirror reflections of objects using diffusion models by formulating it as an image inpainting task.
- For this purpose, we present *SynMirror*, a large-scale synthetic dataset of diverse objects with accurate mirror reflections. We also create MirrorBench, a subset of *SynMirror*, for benchmarking the capabilities of generative models in generating photo-realistic mirror reflections of diverse objects.
- We further propose *MirrorFusion*, a novel depth conditioned inpainting method, which produces photo-realistic and controlled mirror reflections in the masked region of an input image, when trained on *SynMirror*.

2. Related Work

Diffusion based generative models. Diffusion-based models [47] have revolutionized the field of image synthesis [11, 43]. These generative models are further extended to other modalities such as video [12, 46], audio [19]and text [23]. Further, text-to-image(T2I) models [36, 41, 42, 44] have the capability to generate photo-realistic images with any arbitrary text prompt. However, these methods do not work for generating realistic reflections, as shown in Fig. 2.

Image inpainting methods. Recent advancements in diffusion models [11] have led to significant progress in the inpainting task. Diffusion based inpainting methods [1, 4, 28, 33, 38, 43, 59] have shown tremendous improvement in this

task compared to GAN based methods [26, 64]. A common approach to inpaint with diffusion models involves modifying the standard denoising strategy: sampling masked regions from a pre-trained diffusion model and unmasked areas from the given image. While this method produces satisfactory results, it does not generalize to complex scenes and shapes. Stable Diffusion Inpainting [42] fine-tunes a diffusion model by taking the noisy latents, mask and masked image as inputs to the U-Net architecture. Methods like HD-Painter [30] further enhance this method. A recent method, BrushNet [14], divides the masked image features and noisy latents into separate branches, which increases textual coherence and improves masked image preservation. We show in Fig. 1 that these methods do not perform well for generating reflections on the mirror.

Reflection in vision tasks. Reflection has been extensively explored for image enhancement tasks such as single image reflection removal [13, 21, 53]. This task is ill-posed in nature and requires additional priors to be solved. Other methods [18, 20, 22] use multiple images to solve this task. Specifically, they use polarization cues to remove reflection from the input image. Further, reflection cues are used to detect the glass/reflective surfaces in the real world [25, 31]. Recently, PromptRR [52] uses diffusion models for solving the single image reflection removal task. Further, [27, 34, 58] solves the challenge of reconstructing mirror reflections in a 3D scene for the novel-view synthesis task. In this work, we attempt to solve the challenging task of generating controlled photo-realistic and geometrically consistent mirror reflections for an object in an input image, which has not been addressed in previous works to the best of our knowledge.

3. SynMirror: A synthetic dataset of mirror reflections

We observe that state-of-the-art diffusion models struggle to generate geometrically consistent results for mirror reflections as shown in Fig. 1 and Fig. 2. We hypothesize that the cause of this inferior performance is due to the limited number of samples of images with realistic mirror reflections in various existing datasets used to train these models. Further, we find that existing mirror datasets are inadequate for training generative models as they are primarily designed for reflective mirror detection [56] and lack object diversity [49], which is required to incorporate the priors of mirror reflections in diffusion models. To address this, we propose SynMirror, a first-of-its-kind large-scale synthetic dataset on mirror reflections, with diverse *mirror types, objects, camera poses, HDRI backgrounds* and *floor textures*. We provide a comprehensive characteristic comparison of SynMirror with existing mirror datasets in Tab. 1. Note that SynMirror is more than six times larger than all the existing mirror datasets combined. Further, our data generation pipeline renders color images, instance segmentation masks, depth maps and normal maps as shown in Fig. 3.

Table 1. **A comparison between SynMirror and other mirror datasets.** The proposed dataset has more attributes and is six times larger in size than all other existing datasets combined.

Dataset	Type	Size	Attributes
MSD [56]	Real	4018	RGB, Masks
Mirror-NeRF [58]	Real & Synthetic	9 scenes	RGB, Masks, Multi-View
DLSU-OMRS [9]	Real	454	RGB, Mask
TROSD [48]	Real	11060	RGB, Mask
PMD [24]	Real	6461	RGB, Masks
RGBD-Mirror [32]	Real	3049	RGB, Depth
Mirror3D [49]	Real	7011	RGB, Masks, Depth
SynMirror (Ours)	Synthetic	**198204**	RGB, Depth, Masks, Normals, Multi-View

Additionally, we create MirrorBench, a subset of SynMirror, which serves as a challenging benchmark for generative tasks on mirror reflections. SynMirror can also be leveraged to benchmark other downstream tasks such as monocular depth estimation and novel-view synthesis.

3.1. Dataset Generation and Processing

Object Source. SynMirror consists of 3D assets from two widely used 3D object datasets - Objaverse [5] and Amazon Berkeley Objects (ABO) [3]. Objaverse is a large scale dataset consisting of $800K$ 3D assets with diverse categories and ABO contains catalogued 3D models with complex geometries that correspond to real world household objects. However, some objects from Objaverse are poorly rendered or have low-quality textures. Thus, we use a filtered list of $64K$ 3D objects as filtered by OBJECT 3DIT [35]. Despite this initial filtering, some *"spurious"* objects do not show mirror reflections pertaining to specific shader properties of objects. We elaborate on the filtering method to remove *"spurious"* objects in Appendix A.1. Post filtering, we obtain a subset of $58,115$ high-quality 3D assets from Objaverse. We include all $7,953$ objects from the ABO dataset to cover a wide range of object shapes and appearances, and thus resulting in $66,068$ total objects.

Scene Setting. We use a virtual environment in Blender to compose scenes with realistic reflections. We follow a set of heuristic rules to compose a scene with the 3D asset, a floor, and a mirror; an example is shown in Fig. 3. We model mirrors as rectangular planes of varying sizes and frame textures. The floor is modeled as a plane with a diverse set of textures sampled from 1182 CC-textures [6].

We compose a scene, by first placing a mirror vertically at a fixed location. Next, we define a region at a fixed distance to the mirror where the 3D object will be placed, represented as a unit cube. We normalize the object to fit in the unit cube. We also rotate the 3D object around its $y-axis$ to increase diversity in object poses. For modeling the background, we use 359 high-quality HDRI environment maps from PolyHeaven [10]. We categorize the floor textures and

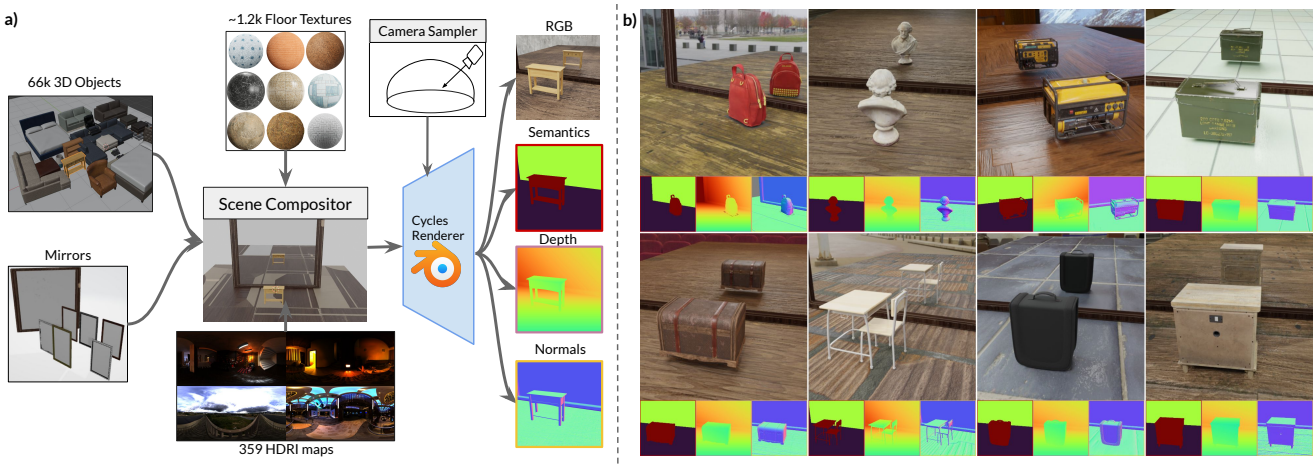

Figure 3. SynMirror: a) **Dataset creation pipeline** - We sample diverse 3D objects, mirrors as 2D planes and diverse floor textures to compose a scene in a blender environment. To enhance realism, we sample high-quality HDRI environment maps as backgrounds. We sample cameras from varied viewpoints, capturing the mirror and the object, and use Blender to render RGB images and dense 2D annotations. b) **Samples from SynMirror** - The generated scenes have complex geometry, textures, and high diversity. The renderings have accurate dense annotations for semantic, depth and normal maps at the original image resolution.

HDRI maps into *indoor* and *outdoor* categories to simulate realistic *indoor*/*outdoor* scenes enhancing photorealism of the renderings. For illuminating the scene, we use an area light placed slightly above and behind the object at a 45 degree angle pointing towards the object and the mirror.

Rendering. We generate a pool of 19 camera poses by interpolating between two extreme camera poses while ensuring that the object and its reflection in the mirror are visible. For each scene iteration, we randomly sample 3 camera poses and render the scene from these virtual cameras by leveraging BlenderProc [6], to obtain RGB, depth maps, surface normal maps, and semantic labels. To obtain high-quality photo-realistic renderings, we render at a 512×512 resolution using 1024 cycles from Blender's Cycles renderer. This allows us to create a rich and comprehensively annotated SynMirror dataset for studying a variety of mirror related tasks. More details about the dataset are provided in Appendix A.

4. Method

We briefly introduce diffusion models in Sec. 4.1. Then, we present our method MirrorFusion in Sec. 4.2. Fig. 4 provides an overview of our method.

4.1. Preliminaries

Diffusion models are a family of generative models that generate images by iterative denoising. In the forward diffusion process, a Gaussian noise $\epsilon \sim \mathcal{N}(0,1)$; is sequentially added for T timesteps to a clean sample x_0 to get a noisy sample x_T. In the backward diffusion process, a clean image x_0 is generated by iterative denoising of noisy image x_T. The iterative denoising process is modeled with a denoising network ϵ_θ conditioned on the timestep $t \in \{1, T\}$ and optional conditioning c (e.g. text prompts, inpainting masks). The denoiser is trained with simple mean square loss L_{DM} as follows:

$$L_{DM} = E_{x_0,\epsilon\sim\mathcal{N}(0,I),t}||\epsilon - \epsilon_\theta\left(z_t, t, c\right)||^2 \quad (1)$$

Training diffusion models directly on the large resolution images x_0 is computationally demanding as it needs several denoising steps to generate a single image. Latent Diffusion Models [42] propose to apply a diffusion process in a compressed latent space of a pre-trained Variational Autoencoder. This enables efficient training and fast inference for generating large-resolution images.

4.2. MirrorFusion

Though trained on large-scale datasets, existing state-of-the-art diffusion models fail to generate consistent reflections with accurate object shape and scene appearance as shown in Fig. 1. We propose MirrorFusion, a novel framework for generating accurate mirror reflections by formulating it as an inpainting problem. Given an input scene image and a mirror mask, MirrorFusion fills the masked region with the consistent reflection of the object and the scene. Generating accurate reflections requires a precise 3D understanding of the scene to reason about the distance of objects from the mirror and the shapes of objects. Hence, modeling reflections with just a 2D inpainting model is suboptimal, and we need to inject explicit 3D cues during the inpainting process. Thus, we propose to condition the inpainting approach with depth maps. The geometric signal from the depth map enables us to generate accurate and consistent reflections that adhere to the 3D structure of the scene and the object. Fig. 4 shows the overview of our method.

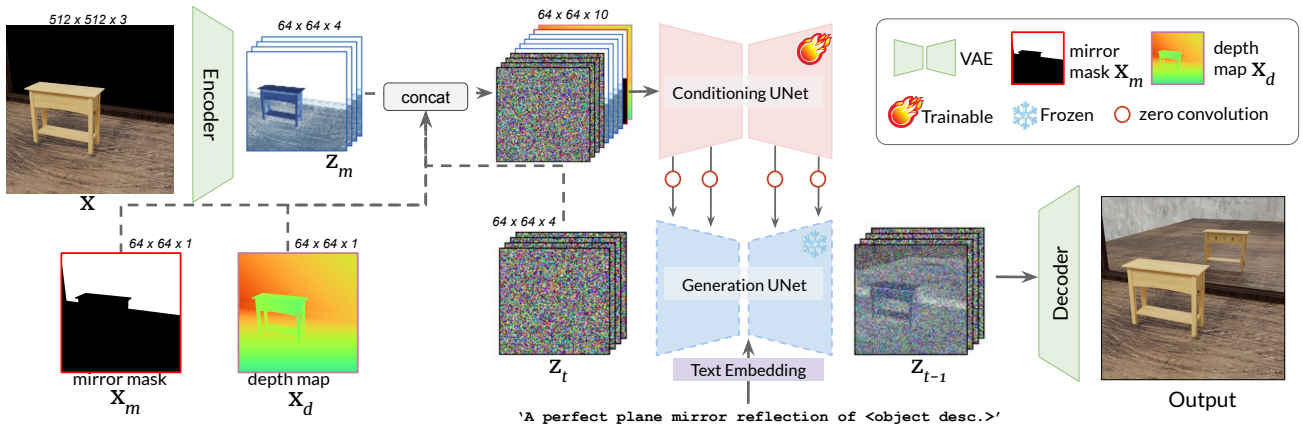

Figure 4. **Overview of the architecture.** We encode the input image x using a pre-trained image encoder from Stable Diffusion to get z_m. Subsequently, we resize the mirror mask m and depth map d to obtain resized mask x_m and depth x_d. Then, we concatenate noisy latents z_t, z_m, x_m and x_d which are fed into the Conditioning U-Net $\epsilon_\theta^{'}$. Each layer of the Generation U-Net ϵ_θ is conditioned via zero convolutions with corresponding layers of $\epsilon_\theta^{'}$. Additionally, ϵ_θ is conditioned by text embeddings. The pre-trained decoder then decodes the denoised latent to produce an image with mirror reflections. Detailed information can be found in Sec. 4.2

4.2.1 Model architecture

MirrorFusion is a diffusion-based inpainting model conditioned on the input mirror mask and depth map. We use a base dual branch architecture for inpainting following BrushNet [14] as shown in Fig. 4. The core idea is to clone a pretrained diffusion model ϵ_θ without cross-attention layers to $\epsilon_\theta^{'}$. Subsequently, the features from the conditioning model $\epsilon_\theta^{'}$ are inserted into the generation model ϵ_θ using zero-convolutional layers. During training, *only the conditioning model is updated, keeping the generation model frozen*. This conditioning mechanism provides a strong hierarchical conditioning for generation without altering the original generation model.

4.2.2 Depth conditioning

Geometric information about objects and scenes is crucial for generating 3D consistent reflections. Recent works [2, 37] show that injecting depth maps enables 3D geometric control in the diffusion models. Inspired by this, we utilize depth-conditioning for our inpainting architecture. Specifically, the noisy latent z_t, masked image latent z_m, inpainting mask x_m, and the depth map x_d are all concatenated and passed as input to the conditioning U-Net $\epsilon^{'}$. The generation U-Net ϵ is an unaltered text-to-image diffusion model, which takes a noisy latent z_t and predicts a cleaner version z_{t-1}. Each layer of generation U-Net ϵ_i is conditioned with the corresponding layer of conditioning U-Net using zero-convolutions ($\mathcal{Z}$) as follows:

$$\epsilon_\theta\left(z_t, t, c\right)_i = \epsilon_\theta\left(z_t, t, c\right)_i + w \cdot \mathcal{Z}\left(\epsilon_\theta^{'}\left(\left[z_t, z_m, x_m, x_d\right], t\right)_i\right) \tag{2}$$

w is the preservation scale to adjust the influence of conditioning. We set w to be 1.0 for all our experiments.

Impact of Depth Conditioning. We demonstrate the importance of depth conditioning for the reflection generation task as shown in Fig. 5. From Fig. 5 (a), it can be clearly seen that BrushNet [14] fine-tuned on SynMirror fails to generate accurate mirror reflection of the object in the input image with high fidelity. For a simple object like a "baseball ball", the *"w/o depth"* BrushNet-FT model generates a ball in which the shape is not preserved. Similarly, in Fig. 5 (b), the shape of "chair" is asymmetrical. These examples show that depth information provided with the proposed normalization scheme generates better reflections on the mirror.

4.2.3 Depth Normalization

The range of input depth is between $[0, \infty)$. The encoder of the U-Net expects the input to be in the range $[-1, 1]$. Hence, we need to normalize the input depth. As discussed in [15], using affine-invariant depth scaling can bring the input depth into the desired range. Image reflection tasks only require the relative distance between the mirror and the scene it reflects. Depth values behind the mirror will not be critical for the reflection generation task. Hence, we use a specifically tailored normalization for our task which is computed as:

$$\hat{d} = \left(\frac{d_{clipped}}{d_{max} + \Delta_{depth}} - 0.5\right) \times 2, \tag{3}$$

where $d_{clipped}$ is input depth clipped between range $[0, d_{max} + \Delta_{depth}]$. d_{max} represents the maximum depth on the mirror mask m. We set Δ_{depth} to the value of 0.5. This normalization aids us in adapting to any of the pre-trained monocular depth estimation methods [15, 55].

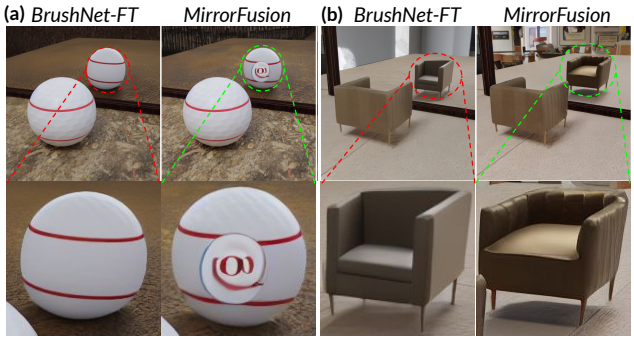

Figure 5. **Impact of depth conditioning** on the reflection generation quality. Notice the irregular shape of the "baseball" and "chair" marked in red. In comparison, our method preserves the structure of the object (marked in green).

Inference. During inference, we assume a predefined mask is available to indicate where the mirror reflection should be generated. The user can easily create this mask, giving better control over the generation of the reflections. We leverage Marigold [15], a monocular depth estimation method to generate the scene depth map. Then, we feed these inputs to our pipeline as shown in Fig. 4. We show in Appendix C.4 that we can utilize alternative methods, such as DepthAnything [55] as a preferred monocular depth estimation technique, demonstrating the robustness of our method to the choice of different monocular depth estimation techniques.

5. Experiments & Results

In this section, we discuss the dataset, baseline comparisons, and extensive experiments used to evaluate our model. We provide additional training and implementation details in Appendix B.

Dataset. As discussed in Sec. 3.1, SynMirror consists of $66,068$ objects and $198,204$ rendered images. We sample 1000 objects from the full dataset to create MirrorBench, a benchmark to evaluate our method and various other baselines. In this benchmark, we have 1497 *"known"* class samples, i.e., these object categories were seen during training, and 1494 *"unknown"* class samples, i.e., these object categories were unseen during training. MirrorBench thus comprises of a total of 2991 images. We also show the generalization capabilities of our method on a few samples from the Google Scanned Objects(GSO) [7] dataset.

Baselines. As discussed in Sec. 4, we formulate generating reflection of an object as an image-inpainting problem. We evaluate various state-of-the-art inpainting methods on MirrorBench. We compare our method with pretrained Stable-Diffusion-Inpainting [42], PowerPaint [65] and BrushNet [14]. We denote zero-shot methods by appending "**-ZS**" to their names. We fine-tune BrushNet on SynMirror and refer to this fine-tuned version as **"BrushNet-FT"** hereafter.

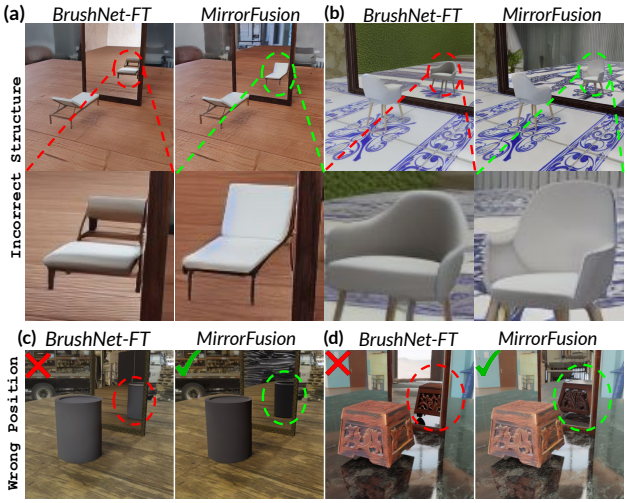

Figure 6. **Additional Results.** Our method effectively preserves the shape of objects, as demonstrated in (a) the lawn chair and (b) the swivel chair. Check in the zoomed-in regions. Additionally, our method accurately positions the objects within the mirror (c) and (d), corroborating the effectiveness of depth-conditioning in our method. Text-prompts used are described in Appendix E.2.

Metrics. We benchmark based on four aspects: masked region preservation, reflection generation quality, Reflection Geometry and text alignment.

- ***Masked Region Preservation.*** We report Peak-Signal-to-Noise ratio (PSNR), Structural Similarity (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) [61] in the unmasked region between generated and the real image. This shows how much original image content is preserved by an inpainting method.
- ***Reflection Generation Quality.*** For measuring the quality of the reflection of the object and the scene, we compute PSNR, SSIM and LPIPS on the masked region containing the object and floor of the ground truth image.
- ***Reflection Geometry.*** We measure the geometric accuracy of the generated reflection using Intersection over Union (IoU) between the segmentation mask of the ground-truth object and the generated object in the reflection region specified by the input mirror mask. We utilize SAM [16] to get the mask of an object in the reflection region. More details are provided in Appendix E.3.
- ***Text Alignment.*** To evaluate the text-image consistency between the generated image and the text prompts, we use CLIP [40] Similarity.

5.1. Qualitative Results

Comparison with Zero-shot Baselines. We observe that all zero-shot baselines fail in generating realistic reflections on the mirror. PowerPaint generates the "lipstick" at the incorrect position, whereas Stable Diffusion 1.5 Inpainting

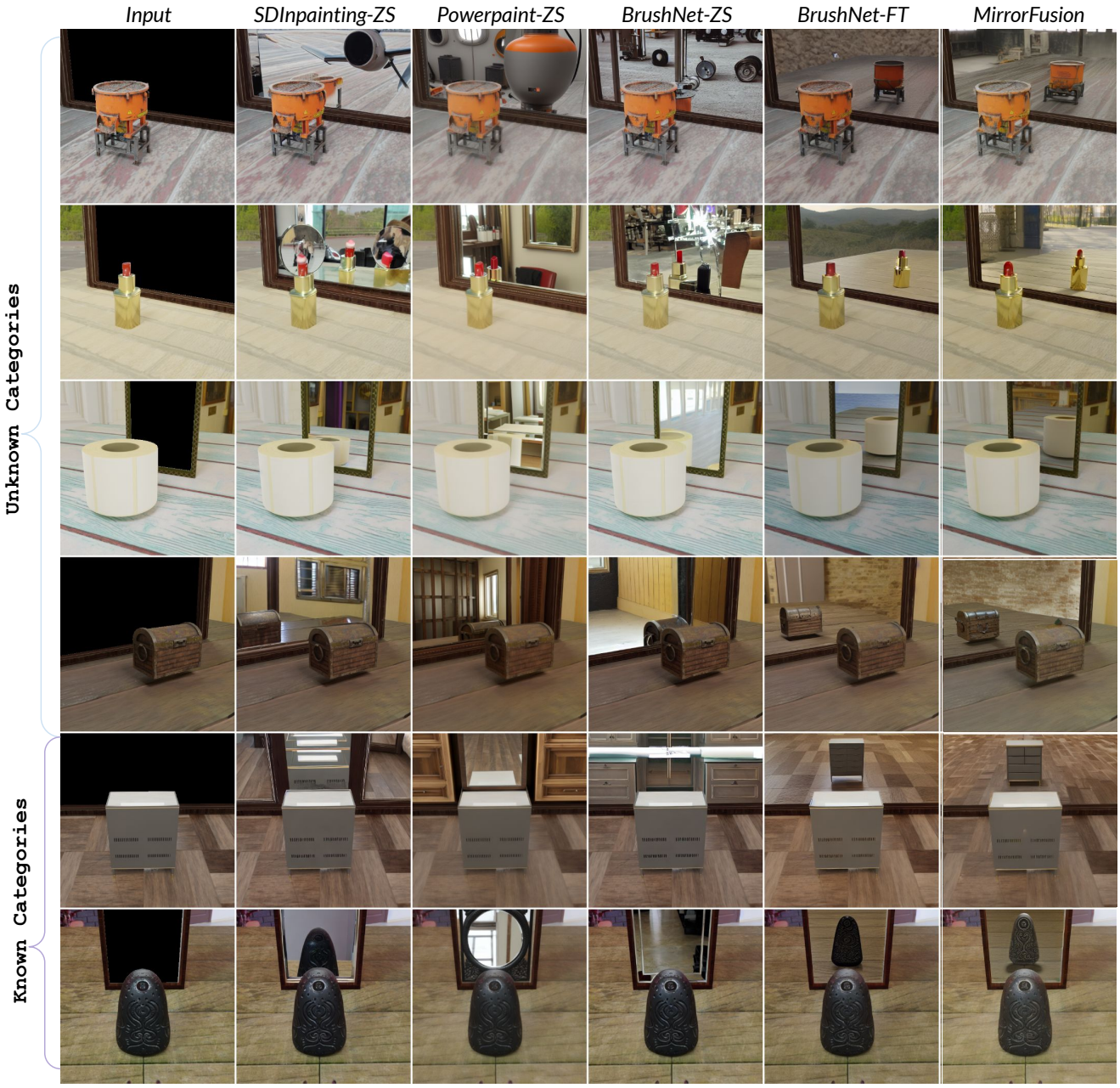

Figure 7. **Reflection generation comparison with general inpainting methods.** We compare our results with zero-shot baselines Stable Diffusion 1.5 Inpainting-ZS, PowerPaint-ZS and BrushNet-ZS. Further, we finetune BrushNet on SynMirror and refer to it as BrushNet-FT. The top four rows compare results on *"unknown"* categories, and the bottom two rows show results on *"known"* categories from Mirror-Bench. Zero-shot methods either fail to generate a reflection on the mirror or generate a reflection at an incorrect position. In comparison, BrushNet-FT generates plausible reflections, but with geometric inaccuracies. Our method improves on shape preservation of the object, floor texture and correct placement of the object in the mirror reflection.

generates two reflections when only one is present. No zero-shot method is able to provide a plausible reflection for a "cement mixer" as shown in Fig. 7 (top row). BrushNet-FT performs better than the zero-shot methods, which shows the utility of the proposed dataset. However, it has issues such as the incorrect size of the object in the generated reflections for "cement mixer" and "lipstick" as shown in the first two rows of Fig. 7. In comparison, our method is able to generate realistic reflections of the objects.

Figure 8. **Change of Viewpoints for mirror and object.** Our method preserves the shape of the object from different viewpoints. This illustrates our method's ability to utilize 3D cues and generate accurate reflection of the object.

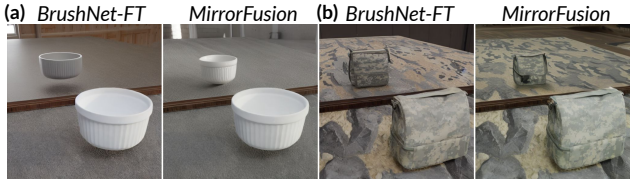

Figure 9. **Generalization on GSO [7].** Our method generates accurate reflections for unseen real-world scanned objects. This substantiates the generalization capabilities of our method.

Additional Results. Fig. 6 shows more comparisons between BrushNet-FT and our method. First row shows that there are some structural inaccuracies in the generated reflection. Check how BrushNet-FT is not able to get the structure of the "lawn chair" in Fig. 6 (a) and swivel chair in Fig. 6 (b). However, our method is able to generate the reflection of the object in a geometrically accurate position. Further in Fig. 6 (c) and (d), notice that the reflection is generated at wrong position in the mirror by BrushNet-FT.

Qualitative Results on GSO. We further benchmark the performance of our method on a completely held-out set of GSO objects. Fig. 9 compares our method with the fine-tuned baseline: BrushNet-FT. Notice that the bowl is floating in the air for BrushNet-FT and the size of the bag in the reflection is large and unnatural. In comparison, our method generates the reflection with better photo-realism and geometric accuracy.

Change of Viewpoints. To evaluate the consistency of our method in generating reflections across varying viewpoints, we designed a continuous trajectory for testing. The results inferred from our method, as depicted in Fig. 8, demonstrate that the reflection of the "sofa-seat" remains consistent as the viewpoint shifts. Additionally, our method preserves high-fidelity reflections for the floor.

Table 2. **Image Generation Quality.** We compare the quality of the inpainted image with fine-tuned baseline method. The best results are shown in **bold**. Our method outperforms the baseline across all metrics, proving its effectiveness.

Metrics	Masked Image Preservation			Text Alignment
Models	**PSNR** ↑	**SSIM** ↑	**LPIPS** ↓	**CLIP Sim** ↑
Brushnet-FT [14]	23.06	**0.84**	0.058	24.90
Ours	**24.22**	**0.84**	**0.051**	**25.23**

Table 3. **Reflection Quality.** We compare the quality of the generated reflection image with the baseline method. We observe that our method has better object quality metrics. Best results are shown in **bold**.

Metrics	Reflection Generation Quality			Reflection Geometry
Models	**PSNR** ↑	**SSIM** ↑	**LPIPS** ↓	**IoU** ↑
Brushnet-FT [14]	19.15	**0.84**	0.082	0.566
Ours	**20.35**	**0.84**	**0.075**	**0.567**

5.2. Ablation studies

Tab. 2 & 3 quantitatively compare BrushNet-FT and MirrorFusion on the image quality and the generated reflection quality. These values are reported on MirrorBench. We generate 4 outputs for each test sample using different random seeds. We then select the image with the best mask SSIM score as the representative image out of the four images. The reported value for any metric is the average of that metric for all representative images across the dataset. MirrorFusion (Ours) with depth cues outperforms BrushNet-FT, which doesn't take depth as an input. This corroborates the necessity of adding depth as an input to the model. We discuss limitations and societal impact in Appendix D.

6. Conclusion

In this work, we propose SynMirror, a large-scale scale challenging and diverse dataset to train generative models for the task of generating realistic mirror reflections. We identify shortcomings in current models and propose MirrorFusion, a novel inpainting method conditioned on depth maps for generating geometrically accurate mirror reflections. Extensive qualitative and quantitative results on MirrorBench shows the superior performance of MirrorFusion in comparison to other methods. Our work is the first step towards generating geometrically accurate and plausible mirror reflections using diffusion based generative models. We believe that SynMirror and MirrorBench will pave way for research in several mirror-related tasks.

Acknowledgement. We are grateful to Kotak IISc AI-ML Centre for providing the compute resources. We thank Om Rastogi for setting up the metric evaluation framework.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18208–18218, 2022. 2

[2] Shariq Farooq Bhat, Niloy Mitra, and Peter Wonka. Loosecontrol: Lifting controlnet for generalized depth conditioning. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 5

[3] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21126–21136, 2022. 2, 3, 1

[4] Ciprian Corneanu, Raghudeep Gadde, and Aleix M Martinez. Latentpaint: Image inpainting in latent space with diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4334–4343, 2024. 2

[5] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2, 3, 1

[6] Maximilian Denninger, Dominik Winkelbauer, Martin Sundermeyer, Wout Boerdijk, Markus Knauer, Klaus H. Strobl, Matthias Humt, and Rudolph Triebel. Blenderproc2: A procedural pipeline for photorealistic rendering. *Journal of Open Source Software*, 8(82):4901, 2023. 2, 3, 4

[7] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 6, 8, 5

[8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. *arXiv*, 2024. 7, 11

[9] Mark Edward M. Gonzales, Lorene C. Uy, and Joel P. Ilao. Designing a lightweight edge-guided convolutional neural network for segmenting mirrors and reflective surfaces. *Computer Science Research Notes*, 3301:107–116, 2023. 3

[10] Poly Haven. Poly haven : The public 3d asset library, 2024. Accessed: 2024-08-10. 3

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[12] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2

[13] Qiming Hu and Xiaojie Guo. Single image reflection separation via component synergy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13138–13147, 2023. 3

[14] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. *arXiv preprint arXiv:2403.06976*, 2024. 2, 3, 5, 6, 8, 4

[15] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 5, 6, 7

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 6, 8

[17] Subhadeep Koley, Ayan Kumar Bhunia, Deeptanshu Sekhri, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. It's all about your sketch: Democratising sketch control in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7204–7214, 2024. 2

[18] Naejin Kong, Yu-Wing Tai, and Sung Yong Shin. High-quality reflection separation using polarized images. *IEEE Transactions on Image Processing*, 20(12):3393–3405, 2011. 3

[19] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020. 2

[20] Chenyang Lei, Xuhua Huang, Mengdi Zhang, Qiong Yan, Wenxiu Sun, and Qifeng Chen. Polarized reflection removal with perfect alignment in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1750–1758, 2020. 3

[21] Anat Levin, Assaf Zomet, and Yair Weiss. Separating reflections from a single image using local features. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, pages I–I. IEEE, 2004. 3

[22] Rui Li, Simeng Qiu, Guangming Zang, and Wolfgang Heidrich. Reflection separation via multi-bounce polarization state tracing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 781–796. Springer, 2020. 3

[23] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022. 2

[24] Jiaying Lin, Guodong Wang, and Rynson W.H. Lau. Progressive mirror detection. In *Proc. CVPR*, 2020. 3

[25] Jiaying Lin, Zebang He, and Rynson WH Lau. Rich context aggregation with reflection prior for glass surface detection.

In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13415–13424, 2021. 3

[26] Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, and Jing Liao. Pd-gan: Probabilistic diverse gan for image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9371–9381, 2021. 3

[27] Jiayue Liu, Xiao Tang, Freeman Cheng, Roy Yang, Zhihao Li, Jianzhuang Liu, Yi Huang, Jiaqi Lin, Shiyong Liu, Xiaofei Wu, et al. Mirrorgaussian: Reflecting 3d gaussians for reconstructing mirror reflections. *arXiv preprint arXiv:2405.11921*, 2024. 3

[28] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022. 2

[29] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[30] Hayk Manukyan, Andranik Sargsyan, Barsegh Atanyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Hd-painter: high-resolution and prompt-faithful text-guided image inpainting with diffusion models. *arXiv preprint arXiv:2312.14091*, 2023. 3

[31] Haiyang Mei, Xin Yang, Yang Wang, Yuanyuan Liu, Shengfeng He, Qiang Zhang, Xiaopeng Wei, and Rynson WH Lau. Don't hit me! glass detection in real-world scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3687–3696, 2020. 3

[32] Haiyang Mei, Bo Dong, Wen Dong, Pieter Peers, Xin Yang, Qiang Zhang, and Xiaopeng Wei. Depth-aware mirror segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3044–3053, 2021. 3

[33] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2

[34] Jiarui Meng, Haijie Li, Yanmin Wu, Qiankun Gao, Shuzhou Yang, Jian Zhang, and Siwei Ma. Mirror-3dgs: Incorporating mirror reflections into 3d gaussian splatting. *arXiv preprint arXiv:2404.01168*, 2024. 3

[35] Oscar Michel, Anand Bhattad, Eli VanderBilt, Ranjay Krishna, Aniruddha Kembhavi, and Tanmay Gupta. Object 3dit: Language-guided 3d-aware image editing. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[36] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[37] Karran Pandey, Paul Guerrero, Matheus Gadelha, Yannick Hold-Geoffroy, Karan Singh, and Niloy J Mitra. Diffusion handles enabling 3d edits for diffusion models by lifting activations to 3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7695–7704, 2024. 5

[38] Rishubh Parihar, Harsh Gupta, Sachidanand VS, and R Venkatesh Babu. Text2place: Affordance-aware text guided human placement. In *European Conference on Computer Vision*, pages 57–77, 2024. 2

[39] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[41] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 4, 6

[43] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022. 2

[44] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[45] Ayush Sarkar, Hanlin Mai, Amitabh Mahapatra, Svetlana Lazebnik, David A Forsyth, and Anand Bhattad. Shadows don't lie and lines can't bend! generative models don't know projective geometry... for now. *arXiv preprint arXiv:2311.17138*, 2023. 2

[46] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[47] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[48] Tianyu Sun, Guodong Zhang, Wenming Yang, Jing-Hao Xue, and Guijin Wang. Trosd: A new rgb-d dataset for transparent and reflective object segmentation in practice. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(10):5721–5733, 2023. 3

[49] Jiaqi Tan, Weijie Lin, Angel X Chang, and Manolis Savva. Mirror3D: Depth refinement for mirror surfaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[50] Michał J Tyszkiewicz, Pascal Fua, and Eduard Trulls. Gecco: Geometrically-conditioned point diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2128–2138, 2023. 2

[51] Rishi Upadhyay, Howard Zhang, Yunhao Ba, Ethan Yang, Blake Gella, Sicheng Jiang, Alex Wong, and Achuta Kadambi. Enhancing diffusion models with 3d perspective geometry constraints. *ACM Transactions on Graphics (TOG)*, 42(6):1–15, 2023. 2

[52] Tao Wang, Wanglong Lu, Kaihao Zhang, Wenhan Luo, Tae-Kyun Kim, Tong Lu, Hongdong Li, and Ming-Hsuan Yang. Promptrr: Diffusion models as prompt generators for single image reflection removal. *arXiv preprint arXiv:2402.02374*, 2024. 3

[53] Kaixuan Wei, Jiaolong Yang, Ying Fu, David Wipf, and Hua Huang. Single image reflection removal exploiting misaligned training data and network enhancements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8178–8187, 2019. 3

[54] Daniel Winter, Matan Cohen, Shlomi Fruchter, Yael Pritch, Alex Rav-Acha, and Yedid Hoshen. Objectdrop: Bootstrapping counterfactuals for photorealistic object removal and insertion. *arXiv preprint arXiv:2403.18818*, 2024. 2

[55] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 5, 6, 7

[56] Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson W.H. Lau. Where is my mirror? In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 3, 5

[57] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arxiv:2308.06721*, 2023. 2

[58] Junyi Zeng, Chong Bao, Rui Chen, Zilong Dong, Guofeng Zhang, Hujun Bao, and Zhaopeng Cui. Mirror-nerf: Learning neural radiance fields for mirrors with whitted-style ray tracing. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4606–4615, 2023. 3

[59] Guanhua Zhang, Jiabao Ji, Yang Zhang, Mo Yu, T. Jaakkola, and Shiyu Chang. Towards coherent image inpainting using denoising diffusion implicit models. In *International Conference on Machine Learning*, 2023. 2

[60] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2

[61] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[62] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[63] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 4

[64] Haitian Zheng, Zhe Lin, Jingwan Lu, Scott Cohen, Eli Shechtman, Connelly Barnes, Jianming Zhang, Ning Xu, Sohrab Amirghodsi, and Jiebo Luo. Image inpainting with cascaded modulation gan and object-aware training. In *European Conference on Computer Vision*, pages 277–296. Springer, 2022. 3

[65] Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. *arXiv preprint arXiv:2312.03594*, 2023. 2, 6