MirrorVerse: Pushing Diffusion Models to Realistically Reflect the World

Anonymous CVPR submission

Paper ID 35

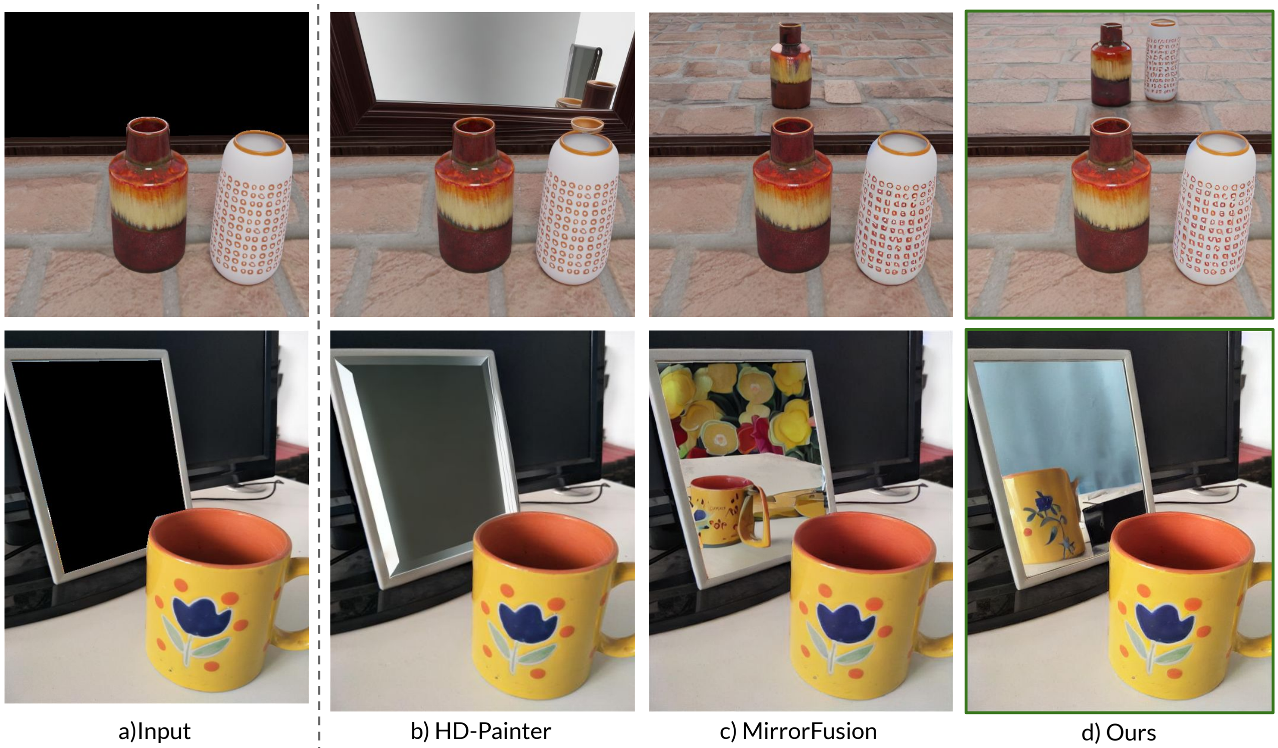

Figure 1. Our model MirrorFusion 2.0, trained on our enhanced dataset SynMirrorV2 surpasses previous state-of-the-art diffusion-based inpainting models at the task of generating mirror reflections. All images were created by appending the prompt: "A perfect plane mirror reflection of " to the object description. All text prompts can be found in the supplementary.

Abstract

Diffusion models have become central to various image editing tasks, yet they often fail to fully adhere to physical laws, particularly with effects like shadows, reflections, and occlusions. In this work, we address the challenge of generating photorealistic mirror reflections using diffusion-based generative models. Despite extensive training data, existing diffusion models frequently overlook the nuanced details crucial to authentic mirror reflections. Recent approaches have attempted to resolve this by creating synthetic datasets and framing reflection generation as an inpainting task; however, they struggle to generalize across different object orientations and positions relative to the mirror. Our method overcomes these limitations by introducing key augmentations into the synthetic data pipeline: (1) random object positioning, (2) randomized rotations, and (3) grounding of objects, significantly enhancing generalization across poses and placements. To further address spatial relationships and occlusions in scenes with multiple objects, we implement a strategy to pair objects during dataset generation, resulting in a dataset robust enough to handle these complex scenarios. Achieving generalization to real-world scenes remains a challenge, so we introduce a three-stage training curriculum to develop the MirrorFusion 2.0 model to improve real-world performance. We provide extensive qualitative and quantitative evaluations to support our approach.

'A perfect plane mirror reflection of a mug which is placed in front of the mirror.'

'A perfect plane mirror reflection of a stuffed toy bear which is placed in front of the mirror.'

'A perfect plane mirror reflection of a chair model which is placed in front of the mirror.'

Figure 2. We observe that current state-of-the-art T2I models, SD3.5 [2] (top row) and Flux [22] (bottom row), face significant challenges in producing consistent and geometrically accurate reflections when prompted to generate reflections in the scene.

1. Introduction

In recent years, diffusion-based generative models have redefined what is possible in fields spanning from image generation to video synthesis, producing impressive results across various applications [14, 17, 18, 22, 35, 38]. The evolution of these models has been accompanied by a range of methods designed to fine-tune the generation process through conditional inputs, such as edge maps, sketches, depth maps, and segmentation maps [30, 54, 56, 58]. However, there remains a significant gap in their capacity to replicate intricate physical effects—particularly those rooted in the subtlety of real-world physics, including shadows [41], specular reflections [49], and perspective cues [46]. More challenging still, these techniques struggle to authentically generate mirror reflections, a task requiring a nuanced understanding of light, geometry, and realism that current methods do not adequately address. In this work, we address the question: "*Can current methods be fine-tuned to generate plausible mirror reflections?*"

We motivate the problem further by providing generations from current text-to-image (T2I) generation models. We prompt Stable Diffusion 3.5 [2] and FLUX [22] with prompts to generate a scene with a mirror reflection. Fig. 2 shows that these methods fail to generate plausible mirror reflections. Specifically, check the reflection of "teddy-bear" in the generated outputs from both the methods. Further, inpainting methods like HD-Painter [27] also fail for this task, as shown in Fig. 1. A contemporary method called MirrorFusion [12], claiming to generate mirror reflections, falls short on real-world and challenging scenes as apparent in Fig. 1.

Despite their impressive capabilities, powerful diffusion models struggle to generate mirror reflections accurately.

Table 1. Our proposed dataset, **SynMirrorV2**, surpasses existing mirror datasets in terms of attribute diversity and variability. While recent work [12] introduced the synthetic SynMirror dataset, it lacks key augmentations and scenario, limiting its performance in complex and real-world settings (See Fig. 1).

Dataset	Type	Size (#Images)	Attributes
MSD [52]	Real	4,018	RGB, Masks
Mirror-NeRF [55]	Real & Synthetic	9 scenes	RGB, Masks, Multi-View
DLSU-OMRS [15]	Real	454	RGB, Mask
TROSD [44]	Real	11,060	RGB, Mask
PMD [24]	Real	6,461	RGB, Masks
RGBD-Mirror [28]	Real	3,049	RGB, Depth
Mirror3D [45]	Real	7,011	RGB, Masks, Depth
SynMirror [12]	Synthetic	198,204	Single Fixed Objects: RGB, Depth, Masks, Normals, Multi-View
SynMirrorV2 (Ours)	Synthetic with Single & Multiple Objects	**207,610**	Single + **Multiple Objects**: RGB, Depth, Masks, Normals, Multi-View, **Augmentations**

This limitation stems from the models' reliance on poorly learned priors, a consequence of the quality and quantity of their training data. The scarcity of high-quality, real-world images featuring mirrors and their reflections, as evidenced in Tab. 1, poses a significant challenge. While recent work [12] has attempted to address this issue by training on a synthetic dataset, the results, as illustrated in Fig. 1, suggest that the method's performance suffers in complex scenes and real-world settings. We hypothesize that this is due to inherent limitations in the synthetic data generation process and the training dynamics of the model.

To address the shortcomings in the synthetic data generation pipeline, we create an enhanced pipeline incorporating useful augmentations such as randomizing object position and rotation. We also ensure that the objects are anchored to the ground level in the 3D world. We observe that this diverse data improves the generalization of a trained model across the pose and position of objects in the scene. However, it does not generalize to more complex scenes with multiple objects. To address this, we propose a novel pipeline that places multiple objects in the scene based on their semantic categories, further enhancing the quality and utility of the proposed synthetic dataset. Drawing inspiration from previous works, such as those that have improved the generation quality on various tasks, notably image-editing [29], multilingual T2I generation [23, 50, 53] and several others, we aim to leverage the stage-wise training approach that enhanced the results in these methods.

We briefly sum up our contributions as follows:

- We propose SynMirrorV2, a large-scale synthetic dataset with diversity in objects and their relative position and orientation in the scene.
- Further, we create a pipeline to add multiple objects to a scene in SynMirrorV2.
- We show that with a curriculum strategy of training on SynMirrorV2, a generative method can also generalize to real-world scenes. We show this generalization capability on the challenging real-world MSD [52] dataset.

2. Related Work

Image Generative models. Diffusion models [43] have become quite popular for image generation tasks. Diffusion models work by gradually adding noise to data and then learning to reverse this process to generate data from a variety of distributions [8, 11, 17]. Subsequent works have expanded the scope of image generation by incorporating text guidance [37, 40] into the diffusion process, simplifying the reverse process [48], and reformulating diffusion to occur in a latent space [38] for improved speed. [31] explore advancements in diffusion models by addressing bias through distribution-guided debiasing techniques. Further, methods [32, 33] are developed to provide more fine-grained generation control to these models. Building on the success of vision transformers [47], recent approaches [7, 34, 60] have replaced the U-Net architecture in diffusion models with transformer-based designs, leading to high-quality image generation results. Further, there are popular methods [2, 22] for high-quality image generation. However, these methods also fail for the task of generating reflections on the mirror as shown in Fig. 2

Image Inpainting. Building on the advancements in image diffusion models, methods like Palette [39] and Repaint [26] leverage known regions through the denoising process to reconstruct missing parts. Blended Diffusion [3, 4] refines this approach by replacing noise in unmasked areas with known content but struggles with complex scenes and shapes. Stable-Diffusion Inpainting [38] (SDI) enhances results by fine-tuning the denoiser with noisy latents, masks, and masked images. Recent methods, such as HD-Painter [27], PowerPaint [61], SmartBrush [51] build on SDI with additional training. Recently, BrushNet [20] introduces a plug-and-play architecture that preserves unmasked content while improving coherence with textual prompts. However, Fig. 1 highlights the limitations of these methods in generating reflections on the mirror.

Diffusion Models and 3D concepts. Recently, LRM [19] based methods predict 3D model from a single image. Some methods [21] utilize diffusion-based methods to enable editing of these 3D presentations. Other diffusion-based methods [29] use synthetic image pairs for 3D-aware image editing. However, the synthetic-to-real domain gap can limit their applicability. Further, ObjectDrop [49] trains a diffusion model for object insertion/removal using a counterfactual dataset that can handle shadows and specular reflections. Sarkar et al. [41] shows that generated images have different geometric features such as shadows and reflections from the real images. Upadhyay et al. [46] proposed a geometric constraint in the training process to improve the perspective cues in the generated images. Alchemist [42] provides control over the material properties of an object by proposing an object-centric synthetic dataset with physically-based materials.

3. Dataset

3.1. Data Generation Pipeline

Fig. 2 highlights the failure of state-of-the-art models in handling the reflection generation task. MirrorFusion [12] addresses this challenge by proposing a synthetic dataset but struggles in complex scenarios involving multiple objects and real-world scenes (Fig. 1). We attribute this limitation to the lack of diversity in their dataset. To mitigate these shortcomings, we introduce SynMirrorV2, a large-scale dataset which significantly expands diversity with varied backgrounds, floor textures, objects, camera poses, mirror orientation, object positions, and rotations. Tab. 1 compares existing mirror datasets, while Fig. 3 showcases samples from SynMirrorV2.

Object Sources. We source objects from Objaverse [9] and Amazon Berkeley Objects (ABO) [6] datasets. Objaverse, a large-scale dataset, contains $800K$ diverse 3D assets, while ABO contributes $7,953$ common household objects. To ensure quality, we refine our selection using a curated list of $64K$ objects from OBJECT 3DIT [29] and the filtering procedure discussed in [12], eliminating low-quality textures and sub-par renderings. After filtering, we get $58,109$ objects from Objaverse. In total, we utilize $66,062$ objects.

Scene Resources. To create a realistic scene, we require assets such as a mirror, floor and background. We create a plane for the floor and apply diverse textures sourced from CC-textures [10]. We use HDRI samples provided by PolyHaven [16] to represent the background. In our experiments, we use different kinds of mirrors: full-wall mirrors and tall rectangular mirrors. For lighting, we position an area-light slightly above and behind the object at a $45°$ angle, directing it towards both the object and the mirror.

Object Placement in the scene. To begin, we fix the mirror's position within the scene as a fixed reference point. The sampled object is then scaled to fit within a unit cube, ensuring uniformity in size across all objects. We proceed by sampling the object's x-y position from a pre-computed region that guarantees both visibility of the object in the mirror and camera. This pre-computed region is determined by identifying the intersection between the mirror's viewing frustum and the camera's viewing frustum. Once the position is set, we randomly sample an angle for the object's rotation around the y-axis to introduce variability. However, even with these steps, there may be instances where the object appears to float in the air, which can undermine the dataset's utility. To address this, we apply a straightforward grounding technique, detailed in the supplementary material. Together, these strategies contribute to the diversity and overall quality of the proposed dataset.

Multiple Objects. A typical scene includes multiple objects arranged in varied layouts, producing a range of depth and occlusion scenarios that enhance scene realism. To

CVPR #35

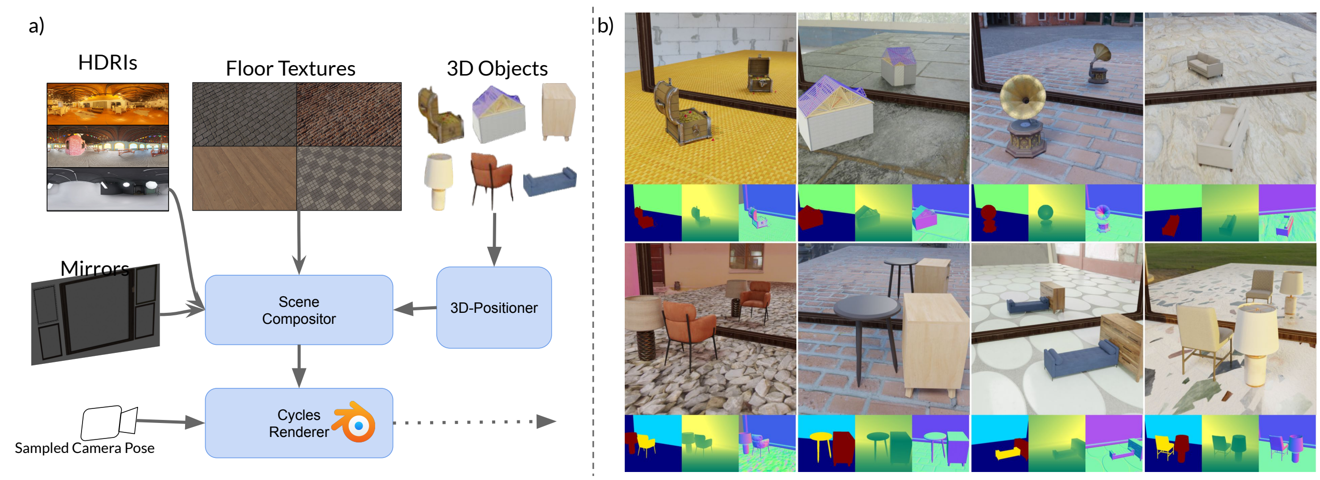

Figure 3. **Dataset Generation Pipeline.** Our dataset generation pipeline introduces key augmentations such as **random positioning, rotation, and grounding** of objects within the scene using the 3D-Positioner. Additionally, we pair objects in semantically consistent combinations to simulate complex spatial relationships and occlusions, capturing realistic interactions for **multi-object** scenes.

Algorithm 1 Procedure to Render Multiple Objects

Require: Input 3D model $\mathcal{M}_1$

Function $\textsc{GetPairedObjectCategory}(M)$
  $c \leftarrow \text{GetSemanticCategory}(M)$
  $L \leftarrow \text{GetPairedCategoriesList}(c)$
  $c_{paired} \leftarrow \text{SampleCategory}(L)$
  return c_{paired}
Function $\textsc{SampleObject}(c)$
  $L_{obj} \leftarrow \text{GetListObjects}(c)$
  $\mathcal{M} \leftarrow \text{Sample3DObject}(L_{obj})$
  return $\mathcal{M}$
Main Algorithm
  $c_{paired} \leftarrow \textsc{GetPairedObjectCategory}(\mathcal{M}_1)$
  $\mathcal{M}_2 \leftarrow \textsc{SampleObject}(c_{paired})$
  Initialize position of $\mathcal{M}_2$ at X
  while $\mathcal{M}_2$ collides with $\mathcal{M}_1$ **do**
    $T_r \leftarrow \text{SampleRandomPosition}()$
    $X \leftarrow T_r$
  end while

capture this complexity, our dataset incorporates scenes with multiple objects, as described in Algorithm 1. We start by sampling K objects from the original ABO dataset and identifying each object's class from [6]. Categories are manually paired to ensure semantic coherence—for instance, pairing a chair with a table. During rendering, after positioning and rotating the primary object K_1, an additional object K_2 from the paired category is sampled and arranged to prevent overlap, ensuring distinct spatial regions within the scene. This process yields 3,140 scenes featuring diverse object configurations and spatial relationships, providing a robust foundation for realistic scene representation.

Rendering. Following scene composition, we randomly sample three camera poses from a predefined list of 19 camera positions and render each scene using BlenderProc [10] to obtain RGB, depth, normal, and semantic label outputs. All renderings are produced at a resolution of 512×512 pixels. We set the "cycles rendering" parameter to 1024, which is necessary for accurately capturing reflections. Representative samples are provided in Fig. 3 and additional examples are available in the supplementary material.

4. Method

Preliminaries Diffusion models are generative models that can construct data samples by progressively removing noise. In the forward diffusion process, Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$ is incrementally added to an initial clean sample x_0 over T timesteps to create a noisy sample x_T. In the reverse process, a clean image x_0 is reconstructed by iteratively denoising x_T. This denoising process is carried out by a denoising network ϵ_θ which is conditioned on the timestep $t \in \{1, T\}$ and optional additional conditioning c (e.g. text prompts, inpainting masks). Training loss of the denoiser is as follows:

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

Model Architecture. Building upon MirrorFusion [12], we also formulate this task as an inpainting task. Our model employs a base dual branch network similar to BrushNet [20] and additionally uses depth map conditioning for the condition branch of BrushNet. In particular, we concatenate the noisy latent z_t, masked image z_m, inpainting mask x_m and depth map x_d, and provide this as an input to

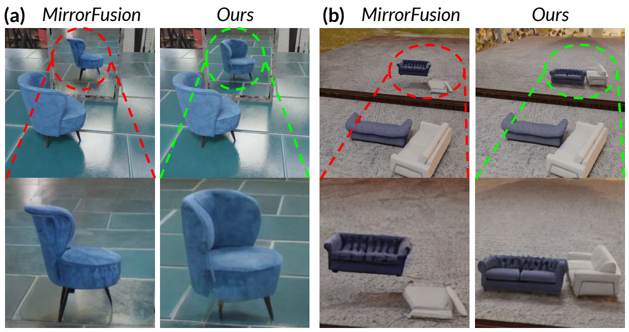

Figure 4. **Comparison on MirrorBenchV2.** The baseline fails to maintain accurate reflections and spatial consistency, showing (a) incorrect chair orientation and (b) distorted reflections of multiple objects. In contrast, our method correctly renders (a) the chair and (b) the sofas with accurate position, orientation, and structure, demonstrating superior performance.

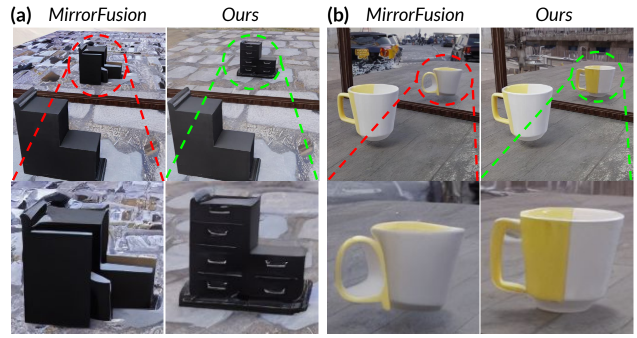

Figure 5. **Comparison on GSO [13] dataset.** In (a), the baseline method misrepresents object structure, while our method preserves spatial integrity and produces realistic reflections. In (b), the baseline yields incomplete and distorted reflections of the mug, whereas our approach generates accurate geometry, color, and detail, showing superior performance on out-of-distribution objects.

the conditioning U-Net branch. Each layer of the generation U-Net ϵ_i is conditioned with the corresponding layer of the conditioning U-Net $\epsilon^{'}$ with the help of 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. We, train the model with the loss in Eq. (1).

Training details. We follow a 3 stage training curriculum to improve the generalization of the model on real-world scenes. We utilize the AdamW [25] optimizer with a learning rate of $1e^{-5}$ and a batch size of 4 per GPU. We train on 4 NVIDIA A100 GPUs in all stages.

- ***Stage 1.*** In the first stage, we initialize the weights of both the conditioning and generation branch with the Stable Diffusion v1.5 checkpoint and finetune the model on the single object train split of our proposed SynMirrorV2. In contrast to [12], we do not keep the generation branch frozen and train the model till $40,000$ iterations. The variation in the position and rotation in the SynMirrorV2 compared to SynMirror allows us to train the model for longer iterations without any degradation in the generation quality compared to [12].
- ***Stage 2.*** In the second stage, we finetune the model for $10,000$ iterations on the multiple objects train split of SynMirrorV2 to incorporate the concepts of occlusions as present in realistic scenes.
- ***Stage 3.*** We propose a third stage training on real-world data from the MSD [52] dataset for another $10,000$ iterations to bridge the domain gap between synthetic and real-world image inpainting.

In the first two stages, we use ground truth depth maps and for the third stage, we generate depth maps using a monocular depth estimator [5]. To enhance learning and reduce reliance on text prompts, we randomly drop them 20% of the time during training, enabling the model to utilize depth information better.

Inference. During inference, we use a CFG value of 7.5 and utilize the UniPC scheduler [59] for 50 time steps. During inference, we allow the user to provide the mask depicting the mirror and estimate the input depth map using Depth-Pro [5] by passing the masked image as input.

5. Experiments & Results

We discuss the evaluation strategy and compare our current method with the previous state-of-the-art method, MirrorFusion [12], referring to this as the baseline. Additionally, we also provide ablation studies on different design choices in Sec. 5.1.

Dataset. Compared to MirrorBench, MirrorBenchV2 consists of renderings of single and multiple objects in a scene. Additionally, we qualitatively test our method on several images from the MSD dataset and renderings from the Google Scanned Objects(GSO) [13] dataset. For single object renderings, we have a total of $2,991$ images, which come from categories that are both seen and unseen during training. We create 300 images that contain two objects from the ABO dataset in the same scene to test the model on generating reflections for multiple objects.

Metrics. We benchmark various methods on the quality of the generated reflection and textual alignment of the generated image with the input prompt.

- ***Reflection Generation Quality.*** We evaluate reflection quality using Peak-Signal-to-Noise ratio (PSNR), Structural Similarity (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) [57] on the masked mirror region.

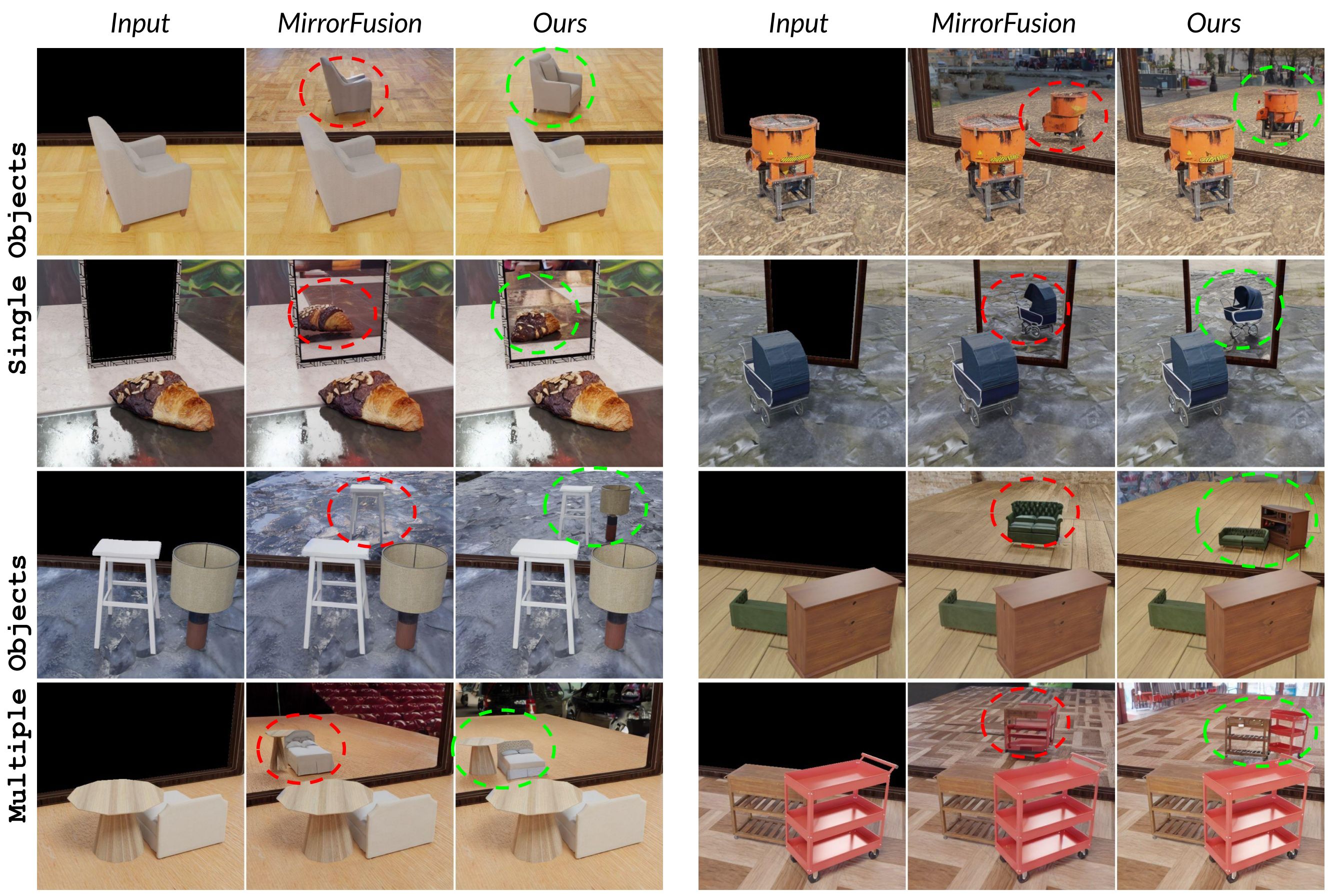

Figure 6. **Results on MirrorBenchV2.** We compare our method with the baseline MirrorFusion [12] on MirrorBenchV2. The baseline method shown struggles with pose variations, even in single-object scenes, and fails to produce accurate reflections for multiple objects. In contrast, our method handles variations in the object orientation effectively and generates geometrically accurate reflections, even in complex, multi-object scenarios.

- ***Text Alignment.*** We use CLIP [36] Similarity for assessing textual alignment.

Qualitative results on MirrorBenchV2. In Fig. 4 (a), a single chair that is slightly rotated is placed in front of a mirror. We observe that the baseline method completely misrepresents the chair's orientation in the generated reflection as seen in the mirror. Notice the zoomed-in region where the reflection appears as if the object was cut and pasted onto the mirror. In contrast, MirrorFusion 2.0 trained on SynMirrorV2 accurately captures the chair's orientation in the reflection, as shown in the zoomed-in region highlighted by the green circle.

Fig. 4 (b), shows a scene with a white sofa rotated and placed to the right of a gray sofa. The baseline method produces two artifacts in the reflection: 1) the gray sofa appears to be floating in the air, and 2) the generated reflection of the white sofa is completely incorrect. In contrast, our method accurately generates the scene in the reflection. These results demonstrate the effectiveness of our augmentation strategies, as described in Sec. 3. We show more examples with both single and multiple objects in Fig. 6.

Qualitative results on GSO [13]. We further evaluate the generalization ability of MirrorFusion 2.0 on real-world scanned objects from GSO, shown in Fig. 5. MirrorFusion 2.0 generates significantly more accurate and realistic reflections. For instance, in Fig. 5 (a), MirrorFusion 2.0 correctly reflects the drawer handles (highlighted in green), while the baseline model produces an implausible reflection (highlighted in red). Likewise, for the "White-Yellow mug" in Fig. 5 (b), MirrorFusion 2.0 delivers a convincing geometry with minimal artifacts, unlike the baseline, which fails to accurately capture the object's geometry and appearance.

Qualitative results on the Real-World MSD dataset. MirrorFusion 2.0 performs well on MirrorBenchV2 and real-world objects from GSO but struggles with complex scenes, such as cluttered cables on a table and reflections across multiple mirrors (see Fig. 7). To improve coherence, we fine-tune it on a subset of the MSD dataset and

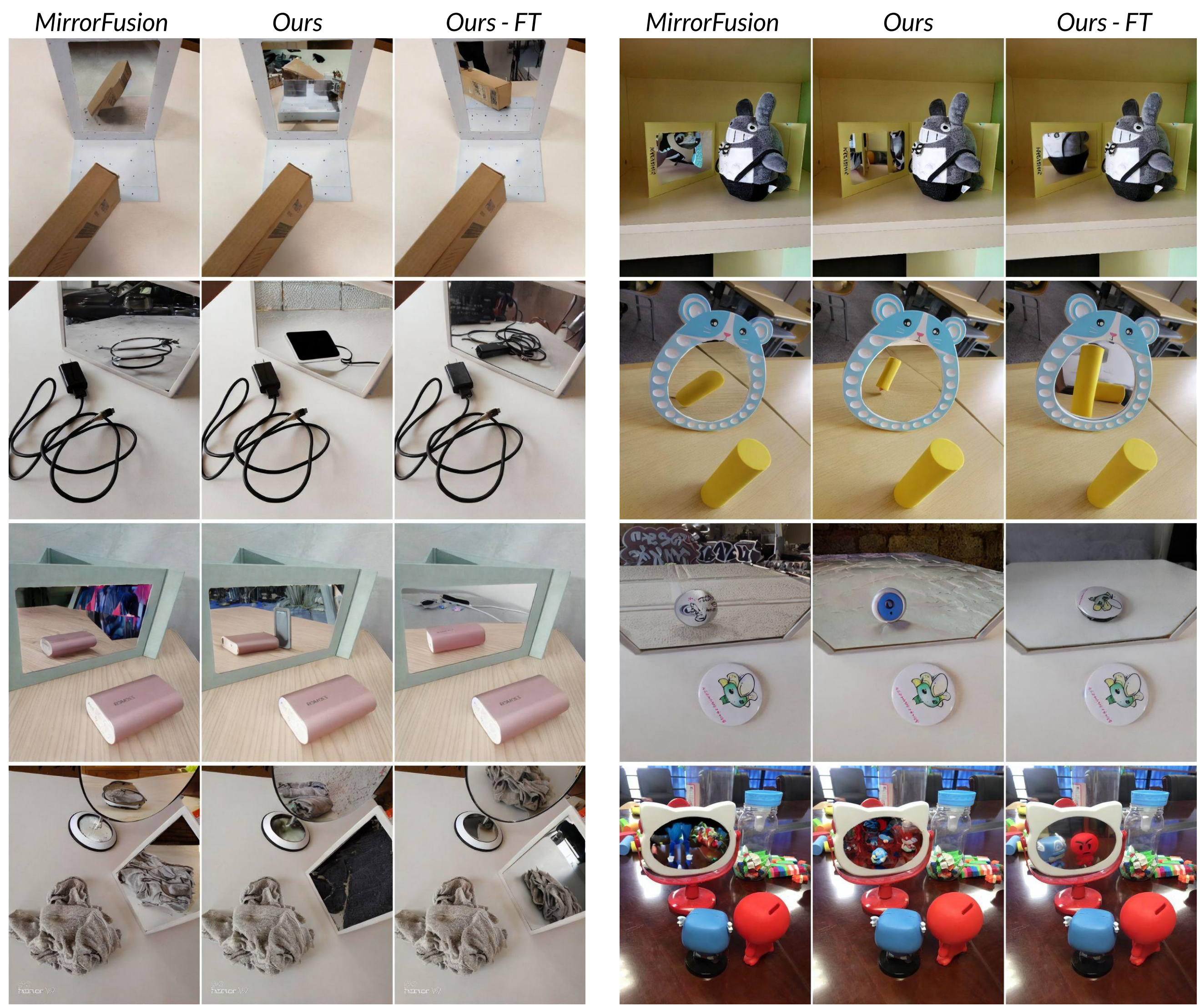

Figure 7. **Real World Scenes.** We show results for MirrorFusion [12], our method and our method fine-tuned on the MSD [52] dataset. We observe that our method can generate reflections capturing the intricacies of complex scenes, such as a cluttered cable on the table and the presence of two mirrors in a 3D scene.

test it on a held-out split, enhancing its ability to handle real-world scenarios. As shown in Fig. 7, this fine-tuning enables high-fidelity reflections, accurately capturing details like the "black cable" on the table and the "towel" in both mirrors. These results demonstrate how our dataset improves diffusion models, enabling more realistic reflections in challenging settings. Fig. 7 illustrates further examples on the real-world MSD dataset.

Quantitative results with baselines. For evaluating the metrics, we generate images using four seeds for a particular prompt and select the image that has the best SSIM score on the unmasked region. For a particular metric, we report the average value across MirrorBenchV2 by averaging the metric for all the selected images. Tabs. 2 and 3 show that our method outperforms the baseline method and finetuning on multiple objects improves the results on complex scenes.

Table 2. **Single Object Reflection Generation Quality.** We compare the quantitative results between the baseline and MirrorFusion 2.0 on the **single object** split of MirrorBenchV2. The best results are shown in **bold**. This shows the effectiveness of the dataset by achieving improved scores.

Metrics	Reflection Generation Quality			Text Alignment
Models	**PSNR** ↑	**SSIM** ↑	**LPIPS** ↓	**CLIP Sim** ↑
baseline [12]	18.31	0.76	0.122	**26.00**
Ours 40k	**18.79**	**0.77**	**0.108**	25.96

CVPR
#35

Table 3. **Multiple Object Reflection Generation Quality.** We compare the quantitative results between MirrorFusion 2.0 trained without multiple objects and MirrorFusion 2.0 trained with multiple objects on the **multiple object** split of MirrorBenchV2. The best results are shown in **bold**. This shows the effectiveness of finetuning further on multiple objects.

Metrics	Reflection Generation Quality			Text Alignment
Models	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**	**CLIP Sim ↑**
Ours 40k	17.77	0.743	0.126	**26.17**
Ours 50k	**18.00**	**0.744**	**0.119**	26.09

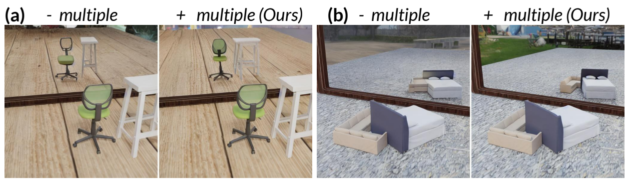

Figure 8. **Impact of adding multiple objects.** We observe that training without multiple objects leads to (a) poor reflection generation and (b) artifacts like object blending, supporting the need for finetuning the model on such scenarios.

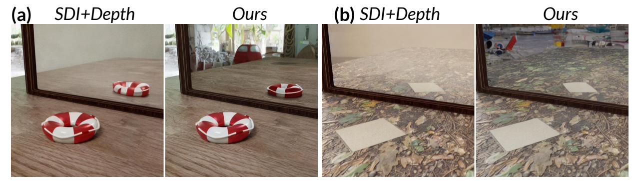

Figure 9. **Comparison with SDI+Depth baseline.** We observe color leakage issues in "SDI+Depth" generations. A dual-branch architecture proves to be a better choice, yielding superior outcomes.

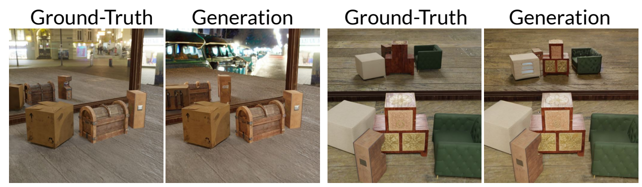

Figure 10. **Limitations.** Our method performs well in multi-object scenes (more than two objects) but retains some artifacts, which can be reduced by synthesizing the dataset through the proposed data-generation pipeline and further increasing the diversity and scale.

User study. To evaluate the effectiveness of our proposed strategy, we also conducted a user study where we provided users with 40 different samples containing single, multiple, GSO objects, and real-world generations from the baseline and MirrorFusion 2.0. 84% **of users preferred generations from MirrorFusion 2.0 over the baseline method**. We provide more details in Appendix D.5.

Limitations. Fig. 10 illustrates examples where our method accurately captures overall geometry but introduces minor artifacts which can be easily addressed by synthesizing additional training data and fine-tuning the model.

5.1. Ablation Studies

Impact of multiple objects dataset. To evaluate the impact of adding multiple objects to our dataset, we compare MirrorFusion 2.0 with ("+ multiple") and without ("-multiple") object training in Fig. 8." MirrorFusion 2.0-w/o multiple" struggles to generate plausible mirror reflections, as evident in Fig. 8 (b), where the bed and sofa appear to blend together. In contrast," MirrorFusion 2.0-with multiple" accurately captures the spatial relationships between objects within the mirror reflection. These results highlight the importance of including multiple objects in the dataset, enabling the model to learn spatial relationships and effectively handle occlusions.

Ablation on architecture. To further validate our architectural choice, we adapt Stable Diffusion Inpainting to accept depth maps as input similar to the changes made for MirrorFusion 2.0 and train this modified model on our proposed dataset referring to it as "SDI+Depth". We compare "SDI+Depth" with MirrorFusion 2.0 in Fig. 9. While "SDI+Depth" accurately positions objects in the mirror, it suffers from significant artifacts, including color leakage in contrast to MirrorFusion 2.0. We suspect that this happens due to the early combination of the noisy latent features, mask, and conditioning information in the initial convolution layer, restricting later layers from accessing clean features. These findings suggest that a dual branch architecture to provide the conditioning information separately as done in MirrorFusion 2.0 is a better choice.

6. Conclusion

We introduce SynMirrorV2, a novel large-scale synthetic dataset designed to advance mirror reflection generation significantly. By employing targeted data augmentations, we achieved robust variability in object pose, position, and occlusion, alongside the ability to handle multi-object scenes. Our qualitative and quantitative evaluations demonstrate SynMirrorV2's efficacy in reflection generation, with promising generalization to real-world scenes using curriculum training. This dataset holds substantial potential for driving progress in various mirror-related tasks. Future research will explore advanced data augmentation techniques to enhance real-world performance further.

References

[1] Adobe. Adobe firefly. https://www.adobe.com/products/firefly.html, 2025. Accessed: 2025-03-23. 8, 9

[2] Stability AI. Stable diffusion 3.5. https://huggingface.co/stabilityai/stable-diffusion-3.5-large, 2025. Accessed: 2025-03-23. 2, 3

[3] 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. 3

[4] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM transactions on graphics (TOG)*, 42 (4):1–11, 2023. 3

[5] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv: 2410.02073*, 2024. 5

[6] 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. 3, 4

[7] Katherine Crowson, Stefan Andreas Baumann, Alex Birch, Tanishq Mathew Abraham, Daniel Z Kaplan, and Enrico Shippole. Scalable high-resolution pixel-space image synthesis with hourglass diffusion transformers. In *Forty-first International Conference on Machine Learning*, 2024. 3

[8] Giannis Daras, Mauricio Delbracio, Hossein Talebi, Alexandros G. Dimakis, and Peyman Milanfar. Soft diffusion: Score matching with general corruptions. *Trans. Mach. Learn. Res.*, 2023, 2023. 3

[9] 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. 3

[10] Maximilian Denninger, Dominik Winkelbauer, Martin Sundermeyer, Wout Boerdijk, Markus Wendelin 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. 3, 4

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3

[12] Ankit Dhiman, Manan Shah, Rishubh Parihar, Yash Bhalgat, Lokesh R Boregowda, and R Venkatesh Babu. Reflecting reality: Enabling diffusion models to produce faithful mirror reflections, 2024. 2, 3, 4, 5, 6, 7

[13] 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. 5, 6, 8

[14] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2

[15] 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. 2

[16] Poly Haven. Poly haven : The public 3d asset library, 2025. Accessed: 2025-03-23. 3

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[18] 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

[19] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *ArXiv*, abs/2311.04400, 2023. 3

[20] Xu 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. In *European Conference on Computer Vision*, 2024. 3, 4

[21] Kunal Kathare, Ankit Dhiman, K Vikas Gowda, Siddharth Aravindan, Shubham Monga, Basavaraja Shanthappa Vandrotti, and Lokesh R Boregowda. Instructive3d: Editing large reconstruction models with text instructions. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 3246–3256, 2025. 3

[22] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2025. Accessed: 2025-03-23. 2, 3

[23] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and Qinglin Lu. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding, 2024. 2

[24] Jiaying Lin, Guodong Wang, and Rynson W.H. Lau. Progressive mirror detection. In *Proc. CVPR*, 2020. 2

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 5

[26] 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. 3

[27] 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. 2, 3

[28] 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. 2

[29] 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. 2, 3

[30] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7465–7475, 2024. 2

[31] Rishubh Parihar, Abhijnya Bhat, Abhipsa Basu, Saswat Mallick, Jogendra Nath Kundu, and R Venkatesh Babu. Balancing act: distribution-guided debiasing in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6668–6678, 2024. 3

[32] 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. Springer, 2024. 3

[33] Rishubh Parihar, VS Sachidanand, Sabariswaran Mani, Tejan Karmali, and R Venkatesh Babu. Precisecontrol: Enhancing text-to-image diffusion models with fine-grained attribute control. In *European Conference on Computer Vision*, pages 469–487. Springer, 2024. 3

[34] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3

[35] 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

[36] 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

[37] 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. 3

[38] 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

[39] 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. 3

[40] 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. 3

[41] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28140–28149, 2024. 2, 3

[42] Prafull Sharma, Varun Jampani, Yuanzhen Li, Xuhui Jia, Dmitry Lagun, Fredo Durand, Bill Freeman, and Mark Matthews. Alchemist: Parametric control of material properties with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24130–24141, 2024. 3

[43] 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. 3

[44] 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. 2

[45] 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. 2

[46] 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, 3

[47] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3

[48] Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22532–22541, 2023. 3

[49] Daniel Winter, Matan Cohen, Shlomi Fruchter, Yael Pritch, Alex Rav-Acha, and Yedid Hoshen. Objectdrop: Bootstrapping counterfactuals for photorealistic object removal and insertion, 2024. 2, 3

[50] Xiaojun Wu, Dixiang Zhang, Ruyi Gan, Junyu Lu, Ziwei Wu, Renliang Sun, Jiaxing Zhang, Pingjian Zhang, and Yan Song. Taiyi-diffusion-xl: Advancing bilingual text-to-image generation with large vision-language model support, 2024. 2

[51] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22428–22437, 2023. 3

[52] 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. 2, 5, 7, 8, 9

[53] Fulong Ye, Guangyi Liu, Xinya Wu, and Ledell Yu Wu. Alt-diffusion: A multilingual text-to-image diffusion model. In *AAAI Conference on Artificial Intelligence*, 2023. 2

[54] 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

[55] 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. 2

[56] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2

[57] 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. 5

[58] 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

[59] 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. 5

[60] Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *Trans. Mach. Learn. Res.*, 2024, 2023. 3

[61] 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. In *European Conference on Computer Vision*, 2023. 3