Fixing Diffusion Models' Limited Understanding of Mirrors and Reflections

Originally published at https://www.unite.ai/fixing-diffusion-models-limited-understanding-of-mirrors-and-reflections/

Since generative AI began to garner public interest, the computer vision research field has deepened its interest in developing AI models capable of understanding and replicating physical laws; however, the challenge of teaching machine learning systems to simulate phenomena such as gravity and liquid dynamics has been a significant focus of research efforts for at least the past five years.

Since latent diffusion models (LDMs) came to dominate the generative AI scene in 2022, researchers have increasingly focused on LDM architecture's limited capacity to understand and reproduce physical phenomena. Now, this issue has gained additional prominence with the landmark development of OpenAI's generative video model Sora, and the (arguably) more consequential recent release of the open source *video* models Hunyuan Video and Wan 2.1.

Reflecting Badly

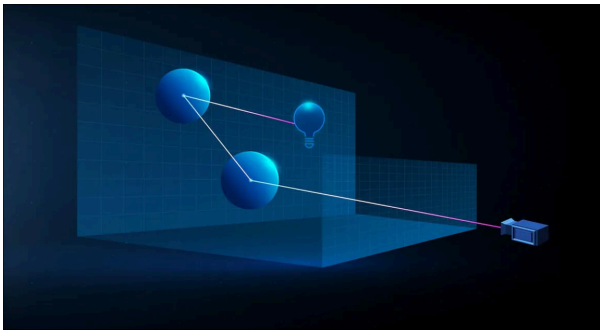
Most research aimed at improving LDM understanding of physics has focused on areas such as gait simulation, particle physics, and other aspects of Newtonian motion. These areas have attracted attention because inaccuracies in basic physical behaviors would immediately undermine the authenticity of AI-generated video.

However, a small but growing strand of research concentrates on one of LDM's biggest weaknesses – it's relative inability to produce accurate *reflections*.

From the January 2025 paper 'Reflecting Reality: Enabling Diffusion Models to Produce Faithful Mirror Reflections', examples of 'reflection failure' versus the researchers' own approach. Source: https://arxiv.org/pdf/2409.14677

This issue was also a challenge during the CGI era and remains so in the field of video gaming, where ray-tracing algorithms simulate the path of light as it interacts with surfaces. Ray-tracing calculates how virtual light rays bounce off or pass through objects to create realistic reflections, refractions, and shadows.

However, because each additional bounce greatly increases computational cost, real-time applications must trade off latency against accuracy by limiting the number of allowed light-ray bounces.



A representation of a virtually-calculated light-beam in a traditional 3D-based (i.e., CGI) scenario, using technologies and principles first developed in the 1960s, and which came to fulmination between 1982-93 (the span between 'Tron' [1982] and 'Jurassic Park' [1993]. Source: https://www.unrealengine.com/en-US/explainers/ray-tracing/what-is-real-time-ray-tracing

For instance, depicting a chrome teapot in front of a mirror could involve a ray-tracing process where light rays bounce repeatedly between reflective surfaces, creating an almost infinite loop with little practical benefit to the final image. In most cases, a reflection depth of two to three bounces already exceeds what the viewer can perceive. A single bounce would result in a black mirror, since the light must complete at least two journeys to form a visible reflection.

Each additional bounce sharply increases computational cost, often doubling render times, making faster handling of reflections one of the most significant opportunities for improving ray-traced rendering quality.

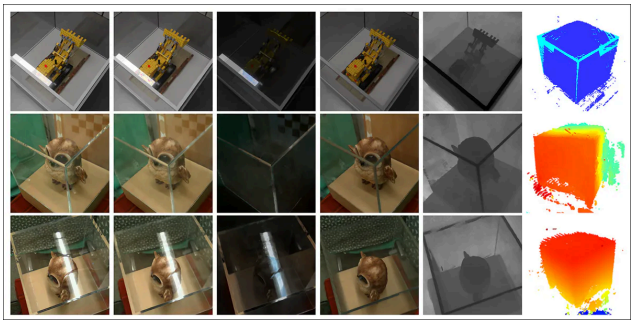
Naturally, reflections occur, and are essential to photorealism, in far less obvious scenarios – such as the reflective surface of a city street or a battlefield after the rain; the reflection of the opposing street in a shop window or glass doorway; or in the glasses of depicted characters, where objects and environments may be required to appear.

A simulated twin-reflection achieved via traditional compositing for an iconic scene in 'The Matrix' (1999).

Image Problems

For this reason, frameworks that were popular prior to the advent of diffusion models, such as Neural Radiance Fields (NeRF), and some more recent challengers such as Gaussian Splatting have maintained their own struggles to enact reflections in a natural way.

The REF^2-NeRF project (pictured below) proposed a NeRF-based modeling method for scenes containing a glass case. In this method, refraction and reflection were modeled using elements that were dependent and independent of the viewer's perspective. This approach allowed the researchers to estimate the surfaces where refraction occurred, specifically glass surfaces, and enabled the separation and modeling of both direct and reflected light components.



Examples from the Ref2Nerf paper. Source: https://arxiv.org/pdf/2311.17116

Other NeRF-facing reflection solutions of the last 4-5 years have included NeRFReN, Reflecting Reality, and Meta's 2024 *Planar Reflection-Aware Neural Radiance Fields* project.

For GSplat, papers such as Mirror-3DGS, Reflective Gaussian Splatting, and RefGaussian have offered solutions regarding the reflection problem, while the 2023 Nero project proposed a bespoke method of incorporating reflective qualities into neural representations.

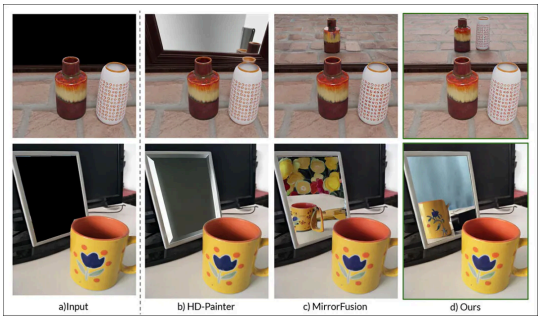
MirrorVerse

Getting a diffusion model to respect reflection logic is arguably more difficult than with explicitly structural, non-semantic approaches such as Gaussian Splatting and NeRF. In diffusion models, a rule of this kind is only likely to become reliably embedded if the training data contains many varied examples across a wide range of scenarios, making it heavily dependent on the distribution and quality of the original dataset.

Traditionally, adding particular behaviors of this kind is the purview of a LoRA or the fine-tuning of the base model; but these are not ideal solutions, since a LoRA tends to skew output towards its own training data, even without prompting, while fine-tunes – besides being expensive – can fork a major model irrevocably away from the mainstream, and engender a host of related custom tools that will never work with any *other* strain of the model, including the original one.

In general, improving diffusion models requires that the training data pay greater attention to the physics of reflection. However, many other areas are also in need of similar special attention. In the context of hyperscale datasets, where custom curation is costly and difficult, addressing every single weakness in this way is impractical.

Nonetheless, solutions to the LDM reflection problem do crop up now and again. One recent such effort, from India, is the *MirrorVerse* project, which offers an improved dataset and training method capable of improving of the state-of-the-art in this particular challenge in diffusion research.



Rightmost, the results from MirrorVerse pitted against two prior approaches (central two columns). Source: https://arxiv.org/pdf/2504.15397

As we can see in the example above (the feature image in the PDF of the new study), MirrorVerse improves on recent offerings tackling the same problem, but is far from perfect.

In the upper right image, we see that the ceramic jars are somewhat to the right of where they should be, and in the image below, which should technically not feature a reflection of the cup at all, an inaccurate reflection has been shoehorned into the right–hand area, against the logic of natural reflective angles.

Therefore we'll take a look at the new method not so much because it may represent the current state-of-the-art in diffusion-based reflection, but equally to illustrate the extent to which this may prove to be an intractable issue for latent diffusion models, static and video alike, since the requisite data examples of reflectivity are most likely to be entangled with particular actions and scenarios.

Therefore this particular function of LDMs may continue to fall short of structure-specific approaches such as NeRF, GSplat, and also traditional CGI.

The new paper is titled *MirrorVerse: Pushing Diffusion Models to Realistically Reflect the World*, and comes from three researchers across Vision and AI Lab, IISc Bangalore, and the Samsung R&D Institute at Bangalore. The paper has an associated project page, as well as a dataset at Hugging Face, with source code released at GitHub.

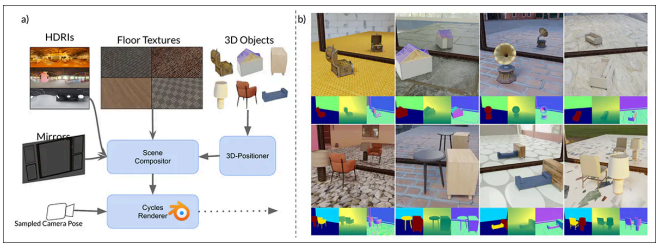
Method

The researchers note from the outset the difficulty that models such as Stable Diffusion and Flux have in respecting reflection-based prompts, illustrating the issue adroitly:

From the paper: Current state-of-the-art text-to-image models, SD3.5 and Flux, exhibiting significant challenges in producing consistent and geometrically accurate reflections when prompted to generate them in a scene.

The researchers have developed *MirrorFusion 2.0*, a diffusion-based generative model aimed at improving the photorealism and geometric accuracy of mirror reflections in synthetic imagery. Training for the model was based on the researchers' own newly-curated dataset, titled *MirrorGen2*, designed to address the generalization weaknesses observed in previous approaches.

MirrorGen2 expands on earlier methodologies by introducing *random object positioning*, *randomized rotations*, and *explicit object grounding*, with the goal of ensuring that reflections remain plausible across a wider range of object poses and placements relative to the mirror surface.



Schema for the generation of synthetic data in MirrorVerse: the dataset generation pipeline applied key augmentations by randomly positioning, rotating, and grounding objects within the scene using the 3D-Positioner. Objects are also paired in semantically consistent combinations to simulate complex spatial relationships and occlusions, allowing the dataset to capture more realistic interactions in multi-object scenes.

To further strengthen the model's ability to handle complex spatial arrangements, the MirrorGen2 pipeline incorporates *paired* object scenes, enabling the system to better represent occlusions and interactions between multiple elements in reflective settings.

The paper states:

> *'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], an additional [object] from the paired category is sampled and arranged to prevent overlap, ensuring distinct spatial regions within the scene.'*

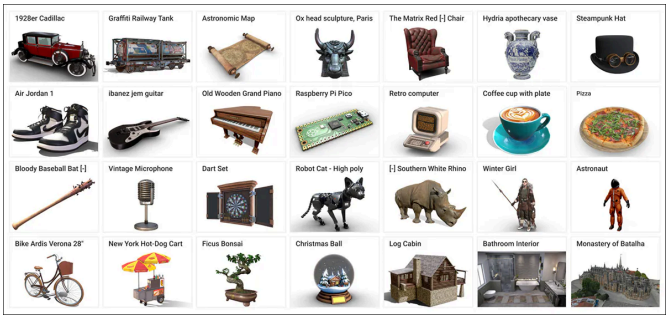
In regard to explicit object grounding, here the authors ensured that the generated objects were 'anchored' to the ground in the output synthetic data, rather than 'hovering' inappropriately, which can occur when synthetic data is generated at scale, or with highly automated methods.

Since dataset innovation is central to the novelty of the paper, we will proceed earlier than usual to this section of the coverage.

Data and Tests

SynMirrorV2

The researchers' SynMirrorV2 dataset was conceived to improve the diversity and realism of mirror reflection training data, featuring 3D objects sourced from the Objaverse and Amazon Berkeley Objects (ABO) datasets, with these selections subsequently refined through OBJECT 3DIT, as well as the filtering process from the V1 MirrorFusion project, to eliminate low-quality asset. This resulted in a refined pool of 66,062 objects.



Examples from the Objaverse dataset, used in the creation of the curated dataset for the new system. Source: https://arxiv.org/pdf/2212.08051

Scene construction involved placing these objects onto textured floors from CC-Textures and HDRI backgrounds from the PolyHaven CGI repository, using either full-wall or tall rectangular mirrors. Lighting was standardized with an area-light positioned above and behind the objects, at a forty-five degree angle. Objects were scaled to fit within a unit cube and positioned using a precomputed intersection of the mirror and camera viewing frustums, ensuring visibility.

Randomized rotations were applied around the y-axis, and a grounding technique used to prevent 'floating artifacts'.

To simulate more complex scenes, the dataset also incorporated multiple objects arranged according to semantically coherent pairings based on ABO categories. Secondary objects were placed to avoid overlap, creating 3,140 multi-object scenes designed to capture varied occlusions and depth relationships.

Examples of rendered views from the authors' dataset containing multiple (more than two) objects, with illustrations of object segmentation and depth map visualizations seen below.

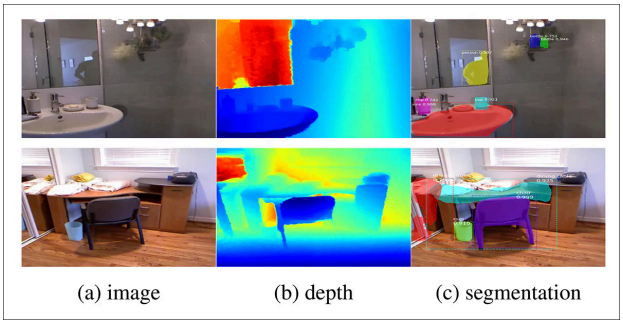
Training Process

Acknowledging that synthetic realism alone was insufficient for robust generalization to real-world data, the researchers developed a three-stage curriculum learning process for training MirrorFusion 2.0.

In Stage 1, the authors initialized the weights of both the conditioning and generation branches with the Stable Diffusion v1.5 checkpoint, and fine-tuned the model on the single-object training split of the SynMirrorV2 dataset. Unlike the above-mentioned *Reflecting Reality* project, the researchers did not freeze the generation branch. They then trained the model for 40,000 iterations.

In Stage 2, the model was fine-tuned for an additional 10,000 iterations, on the multiple-object training split of SynMirrorV2, in order to teach the system to handle occlusions, and the more complex spatial arrangements found in realistic scenes.

Finally, In Stage 3, an additional 10,000 iterations of finetuning were conducted using real-world data from the MSD dataset, using depth maps generated by the Matterport3D monocular depth estimator.



Examples from the MSD dataset, with real-world scenes analyzed into depth and segmentation maps. Source: https://arxiv.org/pdf/1908.09101

During training, text prompts were omitted for 20 percent of the training time in order to encourage the model to make optimum use of the available depth information (i.e., a 'masked' approach).

Training took place on four NVIDIA A100 GPUs for all stages (the VRAM spec is not supplied, though it would have been 40GB or 80GB per card). A learning rate of $1e^{-5}$ was used on a batch size of 4 per GPU, under the AdamW optimizer.

This training scheme progressively increased the difficulty of tasks presented to the model, beginning with simpler synthetic scenes and advancing toward more challenging compositions, with the intention of developing robust real-world transferability.

Testing

The authors evaluated MirrorFusion 2.0 against the previous state-of-the-art, MirrorFusion, which served as the baseline, and conducted experiments on the MirrorBenchV2 dataset, covering both single and multi-object scenes.

Additional qualitative tests were conducted on samples from the MSD dataset, and the Google Scanned Objects (GSO) dataset.

The evaluation used 2,991 single-object images from seen and unseen categories, and 300 two-object scenes from ABO. Performance was measured using Peak Signal-to-Noise Ratio (PSNR); Structural Similarity Index (SSIM); and Learned Perceptual Image Patch Similarity (LPIPS) scores, to assess reflection quality on the masked mirror region. CLIP similarity was used to evaluate textual alignment with the input prompts.

In quantitative tests, the authors generated images using four seeds for a specific prompt, and selecting the resulting image with the best SSIM score. The two reported tables of results for the quantitative tests are shown below.

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

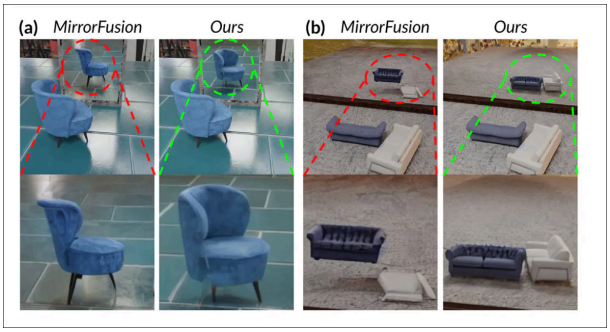
Metrics	Reflection Generation Quality			Text Alignment
Models	PSNR ↑	SSIM ↑	LPIPS ↓	CLIP Sim ↑
Ours 40k	17.77	0.743	0.126	**26.17**
Ours 50k	**18.00**	**0.744**	**0.119**	26.09

Left, Quantitative results for single object reflection generation quality on the MirrorBenchV2 single object split. MirrorFusion 2.0 outperformed the baseline, with the best results shown in bold. Right, quantitative results for multiple object reflection generation quality on the MirrorBenchV2 multiple object split. MirrorFusion 2.0 trained with multiple objects outperformed the version trained without them, with the best results shown in bold.

The authors comment:

> *'[The results] show that our method outperforms the baseline method and finetuning on multiple objects improves the results on complex scenes.'*

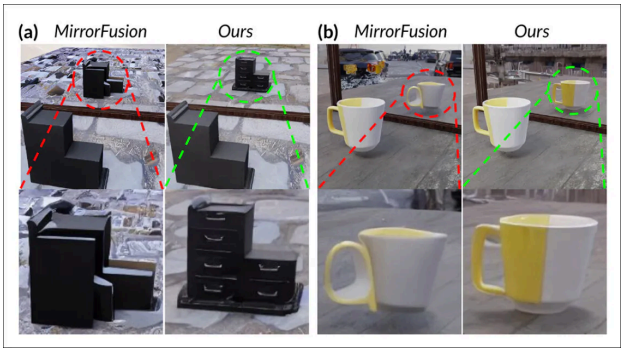
The bulk of results, and those emphasized by the authors, regard qualitative testing. Due to the dimensions of these illustrations, we can only partially reproduce the paper's examples.



Comparison on MirrorBenchV2: the baseline failed to maintain accurate reflections and spatial consistency, showing incorrect chair orientation and distorted reflections of multiple objects, whereas (the authors contend) MirrorFusion 2.0 correctly renders the chair and the sofas, with accurate position, orientation, and structure.

Of these subjective results, the researchers opine that the baseline model failed to accurately render object orientation and spatial relationships in reflections, often producing artifacts such as incorrect rotation and floating objects. MirrorFusion 2.0, trained on SynMirrorV2, the authors contend, preserves correct object orientation and positioning in both single-object and multi-object scenes, resulting in more realistic and coherent reflections.

Below we see qualitative results on the aforementioned GSO dataset:

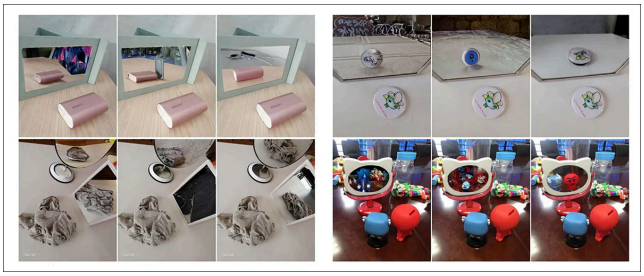


Comparison on the GSO dataset. The baseline misrepresents object structure and produced incomplete, distorted reflections, while MirrorFusion 2.0, the authors contend, preserves spatial integrity and generates accurate geometry, color, and detail, even on out-of-distribution objects.

Here the authors comment:

> *'MirrorFusion 2.0 generates significantly more accurate and realistic reflections. For instance, in Fig. 5 (a – above), 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.'*

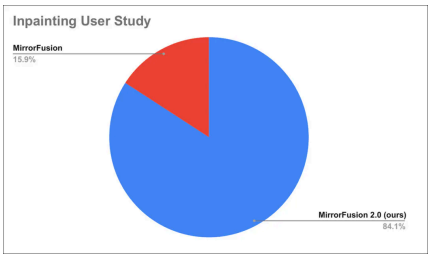
The final qualitative test was against the aforementioned real-world MSD dataset (partial results shown below):



Real-world scene results comparing MirrorFusion, MirrorFusion 2.0, and MirrorFusion 2.0, fine-tuned on the MSD dataset. MirrorFusion 2.0, the authors contend, captures complex scene details more accurately, including cluttered objects on a table, and the presence of multiple mirrors within a three-dimensional environment. Only partial results are shown here, due to the dimensions of the results in the original paper, to which we refer the reader for full results and better resolution.

Here the authors observe that while MirrorFusion 2.0 performed well on MirrorBenchV2 and GSO data, it initially struggled with complex real-world scenes in the MSD dataset. Fine-tuning the model on a subset of MSD improved its ability to handle cluttered environments and multiple mirrors, resulting in more coherent and detailed reflections on the held-out test split.

Additionally, a user study was conducted, where 84% of users are reported to have preferred generations from MirrorFusion 2.0 over the baseline method.



Results of the user study.

Since details of the user study have been relegated to the appendix of the paper, we refer the reader to that for the specifics of the study.

Conclusion

Although several of the results shown in the paper are impressive improvements on the state-of-the-art, the state-of-the-art for this particular pursuit is so abysmal that even an unconvincing aggregate solution can win out with a modicum of effort. The fundamental architecture of a diffusion model is inimical to the reliable learning and demonstration of consistent physics, so that the problem is ill-posed, and apparently not disposed toward an elegant solution.

Further, adding data to existing models is already the standard method of remedying shortfalls in LDM performance, with all the disadvantages listed earlier. It is reasonable to assume that if future high-scale datasets were to pay more attention to the distribution (and annotation) of reflection-related data points, we could expect that the resulting models would handle this scenario better.

Yet the same is true of multiple other bugbears in LDM output – who can say which of them most deserves the effort and money involved in the kind of solution that the authors of the new paper propose here?

First published Monday, April 28, 2025. Tuesday April 29th: made grammar correction in final paras.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai