

PhysReflect: Geometry and Perception Guided Diffusion for Physically-Plausible Mirror Reflections

SHUHENG GE, Harbin Institute of Technology, China
 HONGWEI REN, Harbin Institute of Technology, China
 LI ZHANG, Harbin Institute of Technology, China
 XIANGQIAN WU, Harbin Institute of Technology, China

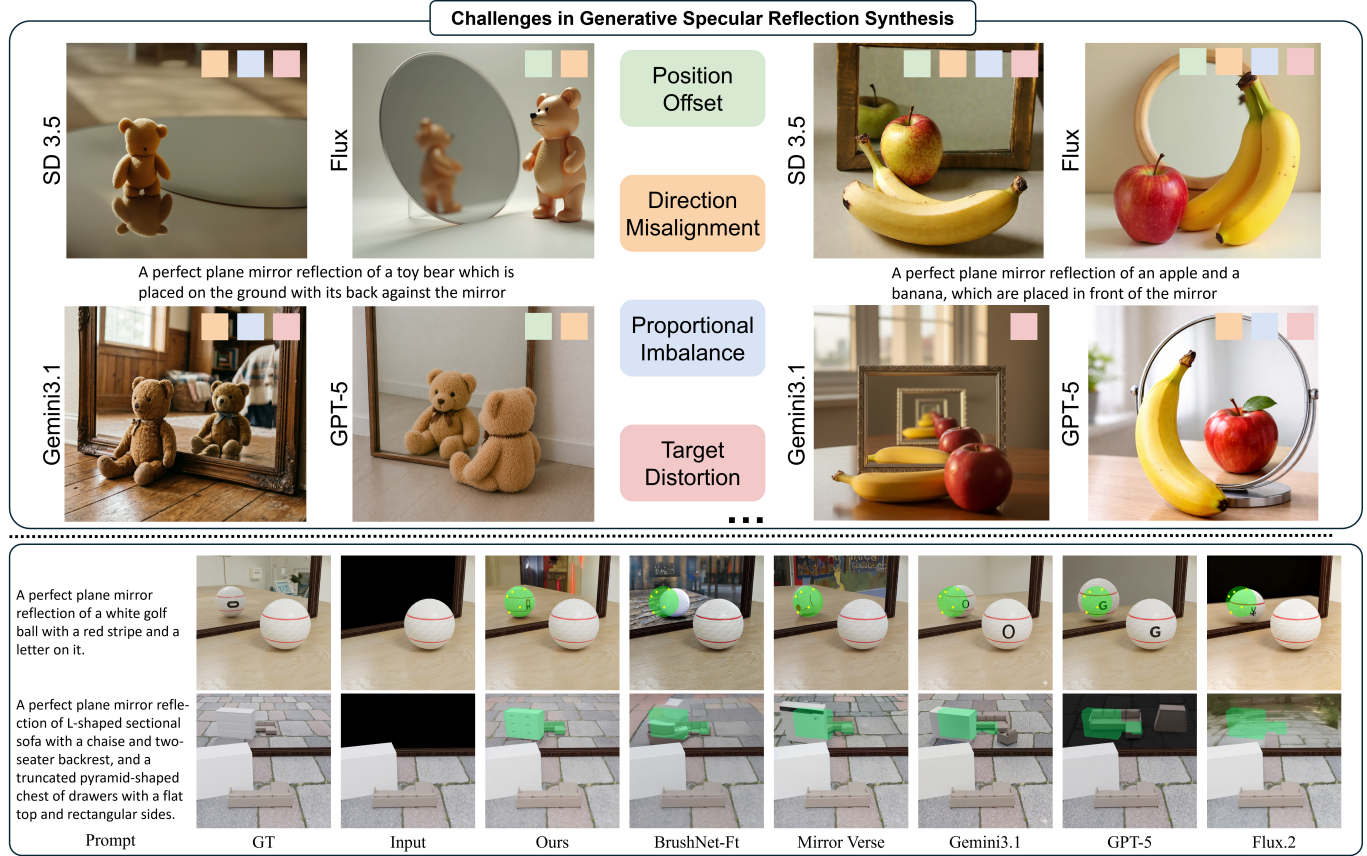


Fig. 1. (Top) **Challenges in Generative Specular Reflection Synthesis.** State-of-the-art open-source models (Stable Diffusion 3.5 [Stability AI 2024], Flux [Black Forest Labs 2024]) and commercial models (GPT-5, Gemini Pro 3.1) consistently produce physically implausible mirror reflections with geometric and perceptual inconsistencies. (Bottom) **Comparison with existing methods and commercial models.** The green mask overlaid on each example indicates the geometrically correct in-mirror object location dictated by the reflection geometry, providing a reference against which the positional and shape deviations of the generated reflections become directly apparent. Our PhysReflect generates reflections that are superior in geometric orientation, spatial positioning, and semantic alignment with the source objects.

Diffusion models generate high-quality images, yet often violate the physical laws governing mirror reflections. Reflections often suffer from geometric aberrations, including positional offsets, directional misalignment, proportional imbalance, and structural distortion. These failures remain evident even in contemporary state-of-the-art generative systems. Existing

methods mitigate this problem through synthetic data scaling or auxiliary depth conditioning, yet their reliance on latent-space denoising objectives leaves reflection-specific geometric and perceptual constraints only indirectly enforced. To bridge this gap, we present **PhysReflect**, a geometry- and perception-guided diffusion framework that decodes the predicted clean latent into image space at each training step, where two complementary mirror-specific objectives are applied under a reliability-aware annealed supervision scheme.

The **Geometric Loss** enforces mirror-induced spatial consistency through sparse epipolar correspondence and dense boundary projection alignment,

Authors' Contact Information: Shuheng Ge, suheng@hit.edu.cn, Harbin Institute of Technology, Harbin, China; Hongwei Ren, Renhongwei@hit.edu.cn; Li Zhang, Harbin Institute of Technology, Harbin, China, zhangli92@hit.edu.cn; Xiangqian Wu, Harbin Institute of Technology, Harbin, China, xqw@hit.edu.cn.

where a SAM2-based *TwinTrack* mechanism provides stable in-mirror localization for boundary-aware supervision. The **Perceptual Loss** preserves reflected appearance by combining **Semantic Consistency Loss**, which maintains reflected identity and appearance via DINOv2 features, and **Lighting Consistency Loss**, which regularizes depth, surface-normal, and illumination coherence under monocular geometry priors. Experiments on synthetic and real-world benchmarks show that PhysReflect outperforms prior mirror-reflection methods in geometric, perceptual, and physical-plausibility metrics, as well as qualitative visual results.

CCS Concepts: • **Computing methodologies** → *Computer vision*.

Additional Key Words and Phrases: Image synthesis, diffusion models, Mirror Reflections, Geometry and Optics

1 Introduction

Diffusion-based generative models have become a standard foundation for visual synthesis and editing, supported by strong progress in image generation [Ho et al. 2020; Podell et al. 2024; Rombach et al. 2022], controllable generation [Mo et al. 2024; Parihara et al. 2024a,b; Zhang et al. 2023], and image inpainting [Ju et al. 2024]. Despite their strong visual realism, prior studies have reported systematic failures in shadows [Sarkar et al. 2024], perspective geometry [Upadhyay et al. 2023], and specular effects [Winter et al. 2024]. Mirror reflection presents a particularly challenging case because a plausible reflected scene must satisfy coupled geometric and perceptual constraints. The reflected content must appear at geometrically correct locations while preserving the identity and appearance of the source object and maintaining coherent illumination with the surrounding scene. As shown in Figure 1, although recent open-source models such as Stable Diffusion 3.5 [Stability AI 2024] and Flux [Black Forest Labs 2024], as well as commercial systems such as GPT-5 and Gemini Pro, can synthesize high-fidelity visual content, their reflections still exhibit positional offsets, directional misalignment, proportional imbalance, and structural distortion.

To mitigate these reflection inconsistencies, recent methods formulate mirror reflection synthesis as a conditional image inpainting task, relying on large-scale synthetic data, auxiliary geometric conditions, and curriculum-based training strategies to improve reflection realism. MirrorFusion [Dhiman et al. 2025b] constructs a large-scale synthetic dataset with depth, normal, and segmentation annotations, and introduces outside-mirror depth maps as conditional inputs to provide stronger geometric cues, encouraging the model to learn reflection regularities. Building upon this direction, MirrorVerse [Dhiman et al. 2025a] further builds a larger and higher-quality synthetic dataset, and proposes a three-stage curriculum from single-object scenes to multi-object scenes and then to real-world scenes, improving the ability of the model to handle complex reflection scenarios.

Despite these advances, these methods still share a common limitation: they lack explicit physical supervision and rely primarily on latent-space denoising objectives, where reflection geometry and perceptual consistency are only indirectly learned from paired data.

We therefore propose **PhysReflect**, a geometry- and perception-guided diffusion training framework for mirror reflection generation. During training, PhysReflect decodes the predicted clean

latent into image space and applies two complementary mirror-specific objectives for geometric and perceptual consistency under reliability-aware annealed supervision.

PhysReflect organizes this supervision into two complementary objectives: **Geometric Loss** ($\mathcal{L}_{geo}$) and **Perceptual Loss** ($\mathcal{L}_{perc}$). The Geometric Loss constrains mirror-induced spatial consistency through sparse epipolar correspondence and dense boundary projection alignment. Within this geometric objective, *TwinTrack*, a SAM2-based dual-frame propagation mechanism, provides stable in-mirror localization for boundary-aware supervision under generated positional drift and shape deformation. The Perceptual Loss preserves reflected appearance by combining semantic consistency for source identity and appearance preservation with lighting consistency for depth, surface-normal, and illumination coherence.

We validate PhysReflect on both synthetic MirrorBenchV2 [Dhiman et al. 2025a] and real-world MSD [Yang et al. 2019] mirror datasets against the state-of-the-art mirror reflection baseline MirrorVerse. To better evaluate whether generated reflections satisfy geometric and perceptual reflection constraints, we extend the standard image-quality metrics with additional evaluation protocols, including DINO Similarity for perceptual consistency, Object IoU and Boundary F1 for geometric alignment, and SH Err for illumination coherence. Quantitative and qualitative results show that, compared with existing methods, PhysReflect substantially improves the physical plausibility of generated mirror reflections while preserving image quality. Since all geometric and perceptual losses and auxiliary models are used only during training, PhysReflect does not introduce additional inference-time cost.

Our contributions are summarized as follows:

- We present **PhysReflect**, a geometry- and perception-guided training framework for physically plausible mirror reflection generation. Building on established decoded-output supervision, PhysReflect formulates and integrates mirror-specific geometric and perceptual objectives while leaving the baseline inference pipeline unchanged.
- We introduce a **Geometric Loss** that enforces mirror-induced spatial consistency through sparse epipolar correspondence and dense boundary projection alignment. A SAM2-based *TwinTrack* mechanism is integrated into this geometric objective to provide stable in-mirror localization for boundary-aware supervision.
- We adapt pretrained-feature perceptual supervision to mirror reflection using global and mirror-local DINOv2 features to regularize source-to-reflection appearance and semantic consistency, together with geometry-aware low-frequency illumination regularization for illumination coherence.
- Experiments on synthetic and real-world mirror datasets show that PhysReflect improves geometric accuracy, perceptual fidelity, illumination coherence, and physical plausibility over existing mirror-reflection generation methods.

2 Related Work

2.1 Diffusion Inpainting and Conditional Control

Diffusion models and their latent variants have become the dominant backbone for image synthesis and editing [Ho et al. 2020;

Rombach et al. 2022]. For inpainting, prior work either modifies the sampling process of pre-trained models [Avrahami et al. 2023; Lugmayr et al. 2022] or trains task-specific architectures that condition on masks, masked images, text, or spatial cues [Ju et al. 2024; Zhang et al. 2023]. These approaches are effective at improving visual fidelity and controllability, but their constraints are usually injected as inputs or architectural conditions. Such conditioning can guide where and what to generate, yet it does not enforce whether the completed content satisfies task-specific physical relations.

This distinction is central to mirror reflection generation: conditioning provides structural hints, whereas reflection-specific geometric and perceptual consistency must be assessed on the generated pixels. PhysReflect therefore complements conditional control with output-space losses for the decoded prediction.

2.2 Mirror Reflection Generation

Mirror reflection is a long-standing challenge in visual perception and scene understanding. Prior work has mainly studied it from the perspectives of reflective-surface understanding, including mirror or glass region detection and segmentation [Lin et al. 2020; Mei et al. 2022; Yang et al. 2019], and 3D scene reconstruction, where mirror-like effects are modeled with specialized NeRF or 3D Gaussian representations [Liu et al. 2024; Meng et al. 2024; Yin et al. 2023]. These reconstruction-oriented methods are designed for scene-level 3D modeling and usually require multi-view observations, camera poses, and iterative optimization. Recently, with the development of generative models, mirror reflection synthesis has emerged as a challenging setting for studying whether visual generation models can follow physical regularities from lightweight image-level inputs.

MirrorFusion [Dhiman et al. 2025b] formulates mirror reflection synthesis as a conditional inpainting problem, introducing SynMirror and a depth-conditioned BrushNet baseline for paired-data reflection generation. MirrorVerse [Dhiman et al. 2025a] extends this direction with SynMirrorV2, diverse layouts, multi-object scenes, and a curriculum for improved generalization to multi-object and real-world scenes. However, these methods still rely primarily on data-driven reconstruction objectives. Depth conditioning and inpainting masks provide structural cues, but under the standard latent denoising objective, geometric constraints such as source-reflection correspondence and boundary alignment, as well as perceptual constraints such as semantic and illumination consistency remain only indirectly constrained. PhysReflect targets this gap by converting these reflection-specific consistency requirements into differentiable training losses, while keeping the inference-time inputs unchanged.

2.3 Differentiable Supervision and Geometric Priors

Recent studies show that visually realistic generative models can still violate physical and geometric regularities such as shadows, perspective, and specular effects [Sarkar et al. 2024; Upadhyay et al. 2023; Winter et al. 2024]. Since these properties are directly reflected in the generated image, prior work has explored differentiable image- and feature-level supervision beyond the standard latent denoising objective. ReFL [Xu et al. 2023] evaluates a one-step clean-image estimate with a learned reward, while AlignProp [Prabhudesai et al. 2023] and DRaFT [Clark et al. 2024] backpropagate differentiable

rewards through the denoising or sampling process. REPA [Yu et al. 2025] aligns denoiser representations with pretrained visual features, and PixelGen [Ma et al. 2026] applies perceptual and DINO-based losses to clean-image predictions. These works establish decoded-output and pretrained-feature supervision as general techniques; PhysReflect specializes them to mirror reflection through source-reflection geometric constraints and mirror-local appearance and illumination priors.

Physical and geometric priors have also been explored in task-specific generation. Relighting methods incorporate estimated geometry and illumination to improve physical consistency [Liang et al. 2026; Liu et al. 2026], while novel-view synthesis methods exploit camera geometry. Zero-1-to-3 [Liu et al. 2023] conditions generation on a source image and relative camera pose, whereas EpiDiff [Huang et al. 2024] introduces epipolar-constrained interaction for cross-view feature exchange. Mirror reflection generation, in contrast, requires source-to-reflection correspondence, reflected-boundary alignment, appearance preservation, and illumination coherence within a single image. PhysReflect organizes these requirements into geometric guidance for correspondence and boundary alignment, together with perceptual guidance for semantic and illumination consistency.

3 Preliminaries and Problem Formulation

Following MirrorFusion [Dhiman et al. 2025b], we formulate mirror reflection synthesis as conditional image inpainting, where the mirror mask defines the generation region and the model fills it with plausible reflection content. We adopt a depth-conditioned BrushNet [Ju et al. 2024] backbone, whose Generation Net is a latent diffusion model [Rombach et al. 2022] and whose Conditioning Net injects multi-scale features from the masked image, mirror mask, and outside-mirror depth map through zero-convolution layers. During training, the VAE encoder maps the ground-truth image X_0 to $z_0 = \mathcal{E}(X_0)$, and the model learns the standard noise-prediction objective $\mathcal{L}_{MSE}$ defined in Appendix. The task input is therefore the noised latent z_t , masked image X , mirror mask M_{mirror} , outside-mirror depth D , and text prompt; the model predicts the completed image $\hat{X}_0$, which is supervised by the ground-truth image X_0 .

4 Method

4.1 PhysReflect Overview

As shown in Fig. 2, PhysReflect introduces explicit mirror-specific physical supervision, applied to the decoded prediction in pixel space during training. Unlike conventional latent-space denoising objectives, which only indirectly encourage physical consistency through paired training data, PhysReflect evaluates whether generated reflections satisfy **geometric and perceptual constraints** directly in image space.

To evaluate the proposed mirror-specific supervision on generated content, we decode the predicted clean latent into image space during training, where reflection geometry, object correspondence, boundary alignment, appearance consistency, and lighting coherence become directly measurable.

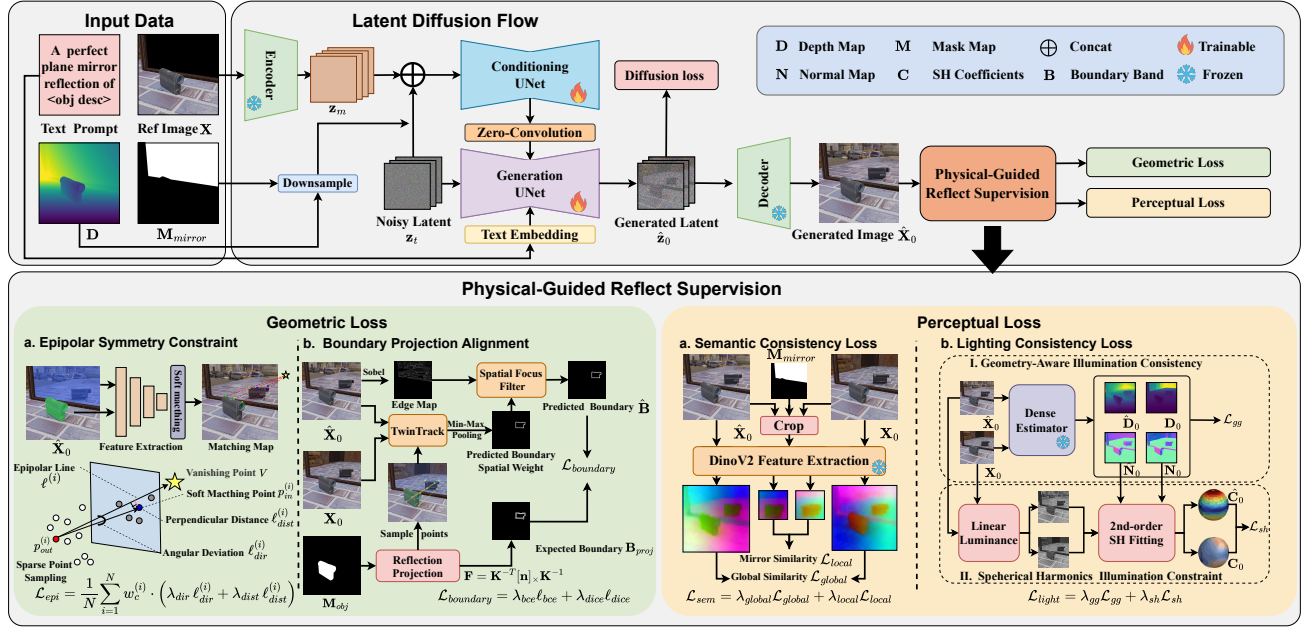


Fig. 2. **Overview of PhysReflect.** PhysReflect builds on a depth-conditioned diffusion inpainting backbone and adds mirror-specific physical supervision applied to the decoded prediction in pixel space during training. Given a noised latent, the model predicts the clean latent, decodes it into image space, and computes two complementary objectives: Geometric Loss and Perceptual Loss. Geometric Loss enforces mirror-induced spatial consistency, while Perceptual Loss preserves reflected appearance through semantic and lighting consistency. Auxiliary priors and frozen models construct supervision only during training, while inference uses the original Baseline inputs and sampling graph.

Given the predicted clean latent representation $\hat{z}_0$, we obtain the decoded image prediction as:

$$\hat{X}_0 = \mathcal{D}(\hat{z}_0), \quad (1)$$

where $\mathcal{D}$ denotes the pretrained VAE decoder. PhysReflect then applies mirror-specific physical supervision to $\hat{X}_0$ in pixel space through two complementary objectives: **Geometric Loss** and **Perceptual Loss**. The Geometric Loss combines sparse epipolar correspondence and dense boundary projection alignment, while the Perceptual Loss combines semantic consistency and lighting consistency.

4.2 Geometric Loss

Mirror reflection generation must place reflected content at physically valid in-mirror locations while preserving geometric symmetry. Existing data-driven objectives provide only implicit spatial supervision, leaving common artifacts such as positional drift and orientation errors. We therefore introduce the Geometric Loss $\mathcal{L}_{geo}$, which combines sparse epipolar correspondence supervision with dense boundary projection alignment.

The mirror plane is defined in the camera coordinate system by a unit normal $\mathbf{n}$ and signed distance d , with $\mathbf{n}^T \mathbf{p} + d = 0$ for any 3D point $\mathbf{p}$ on the plane. The virtual point of a real point $\mathbf{p}$ is:

$$\mathbf{p}_{virt} = \mathbf{p} - 2(\mathbf{n}^T \mathbf{p} + d)\mathbf{n}. \quad (2)$$

This displacement is parallel to $\mathbf{n}$. For homogeneous image coordinates $\tilde{\mathbf{p}}_{out}$ and $\tilde{\mathbf{p}}_{in}$ of a real point and its reflected virtual point, the two camera rays and $\mathbf{n}$ lie in one epipolar plane, yielding:

$$\tilde{\mathbf{p}}_{in}^T \mathbf{F} \tilde{\mathbf{p}}_{out} = 0, \quad \mathbf{F} = \mathbf{K}^{-T} [\mathbf{n}]_{\times} \mathbf{K}^{-1}. \quad (3)$$

Here $[\mathbf{n}]_{\times}$ is the skew-symmetric matrix of $\mathbf{n}$ and $\mathbf{K}$ is the camera intrinsic matrix. The distance d is absent from Eq. 3 because epipolar geometry depends only on the mirror-normal direction, but it is still used for metric projection in Eq. 2 when constructing target boundaries. The derivation is given in Appendix.

The above geometry provides two complementary ways to supervise generated reflections. First, the epipolar relation in Eq. (3) supervises sparse source-to-reflection correspondences. Second, the metric reflection projection in Eq. (2) provides dense boundary-level alignment by projecting object contours across the mirror plane. We first describe the sparse epipolar constraint, followed by the boundary projection alignment term.

4.2.1 Epipolar Symmetry Constraint. Epipolar Symmetry enforces sparse real-to-reflected correspondences under the mirror-induced geometry. As shown in Fig. 2, we sample N points $\{\mathbf{p}_{out}^{(i)}\}_{i=1}^N$ outside the mirror, find reflected correspondences $\mathbf{p}_{in}^{(i)}$ by DINOv2-based soft matching, and penalize their epipolar distance and reflection-direction inconsistency. The reflection-direction term uses the mirror-normal vanishing point $\mathbf{V}$ as its geometric reference. DINOv2 [Oquab

et al. 2023] patch features are used because mirror reflection weakens hand-crafted descriptors such as SIFT [Lowe 2004] and ORB [Rublee et al. 2011], while DINOv2 provides spatially localized semantic features that are more suitable for dense correspondence than global CLIP embeddings.

To keep retained correspondences differentiable, we replace hard argmax matching with a soft expected position over candidate points $\{c_j\}$ sampled from an epipolar-line band inside the mirror. For each $p_{out}^{(i)}$, we extract L_2 -normalized DINOv2 ViT-S/14 features $f_{out}^{(i)}$ from the real input image and $f_{in,j}$ from $\hat{X}_0$ at candidate position c_j . With temperature $\tau = 0.05$, the matched position is:

$$p_{in}^{(i)} = \sum_j p_{ij} c_j, \quad p_{ij} = \frac{\exp(f_{out}^{(i)} \cdot f_{in,j} / \tau)}{\sum_k \exp(f_{out}^{(i)} \cdot f_{in,k} / \tau)} \quad (4)$$

Gradients pass through this soft expectation and the frozen DINOv2 forward pass to $\hat{X}_0$. Detached confidence and reciprocal filters retain only reliable matches; samples with too few retained matches skip this term. For each retained match, the epipolar loss combines a normalized point-to-line distance $\ell_{dist}^{(i)}$ and a cosine direction penalty $\ell_{dir}^{(i)}$:

$$\mathcal{L}_{epi} = \frac{\sum_{i=1}^N w_c^{(i)} (\lambda_{dist} \ell_{dist}^{(i)} + \lambda_{dir} \ell_{dir}^{(i)})}{\sum_{i=1}^N w_c^{(i)} + \epsilon}. \quad (5)$$

4.2.2 Boundary Projection Alignment with TwinTrack. The epipolar term constrains sparse point correspondences, but it does not fully determine the spatial extent or contour shape of the reflected object. To provide dense geometric supervision, we introduce boundary projection alignment, which compares the predicted reflection boundary with an expected boundary obtained by projecting the outside-mirror object through the mirror geometry.

Given the outside-mirror object mask M_{obj} , we sample foreground pixels and lift them to 3D using the available depth and camera intrinsics. These points are reflected across the mirror plane using Eq. (2) and then reprojected into the image plane. The projected support inside the mirror is rasterized and softly dilated to obtain the expected reflection support. We then extract its soft contour as the projected target boundary $B_{proj} \in [0, 1]$. The predicted boundary is extracted from the decoded image prediction $\hat{X}_0$. Directly segmenting the generated in-mirror object is unreliable during diffusion training, because the reflection may exhibit positional drift, shape distortion, or incomplete appearance. As shown in Fig. 3, to obtain a stable supervision region, we introduce *TwinTrack*, which treats the ground-truth image X_0 and the decoded prediction $\hat{X}_0$ as a two-frame pseudo video. A frozen SAM2 [Ravi et al. 2025] propagates the ground-truth contour prompt from X_0 to $\hat{X}_0$, producing an in-mirror mask M_{sam2} that localizes the generated reflection.

Importantly, TwinTrack is used only as a detached spatial selector rather than as a shape target. We compute a differentiable edge map from the generated image:

$$\hat{E} = \sqrt{(S_x * \hat{X}_{0,gray})^2 + (S_y * \hat{X}_{0,gray})^2}, \quad (6)$$

where S_x and S_y denote Sobel kernels. The TwinTrack mask is converted into a detached soft weight $W_{sam} \in [0, 1]$, and the predicted

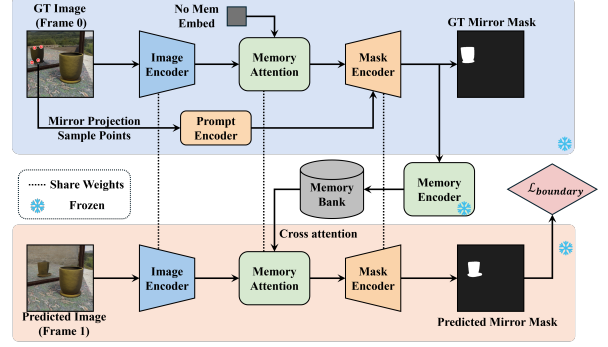


Fig. 3. **TwinTrack for source-boundary extraction.** The ground-truth image X_0 and decoded prediction $\hat{X}_0$ form a two-frame pseudo video. A frozen SAM2 propagates the ground-truth contour prompt to the prediction frame, producing a detached mask M_{sam2} that gates Sobel edges from $\hat{X}_0$ and yields the differentiable source boundary $\hat{B}$.

boundary is obtained as:

$$\hat{B} = \hat{E} \odot W_{sam}. \quad (7)$$

Thus, gradients flow through the image-derived edge map $\hat{E}$, while W_{sam} only determines where the boundary loss is evaluated.

We align the predicted boundary $\hat{B}$ with the projected target boundary B_{proj} inside the mirror region. The loss combines a BCE term for local boundary activation and a Soft-Dice term for sparse contour overlap:

$$\mathcal{L}_{boundary} = \lambda_{bce} \mathcal{L}_{bce} + \lambda_{dice} \mathcal{L}_{dice}. \quad (8)$$

Combining sparse epipolar correspondence and dense boundary projection alignment, the final geometric loss is defined as:

$$\mathcal{L}_{geo} = \lambda_{epi} \mathcal{L}_{epi} + \lambda_{boundary} \mathcal{L}_{boundary}. \quad (9)$$

4.3 Perceptual Loss

Geometric supervision constrains where the reflection appears and how its contour aligns with the mirror-induced projection, but it does not guarantee that the reflected content remains perceptually faithful. A reflection may be geometrically well aligned while still suffering from identity drift, color shifts, texture changes, or inconsistent illumination. At the same time, exact pixel-level matching can be overly restrictive because reflected appearance may differ locally in shading, contrast, and texture while remaining perceptually consistent with the source. We therefore introduce a **Perceptual Loss** $\mathcal{L}_{perc}$ that combines semantic consistency and lighting consistency to regularize the identity and appearance of generated reflections. These terms serve as perceptual and low-frequency illumination priors rather than exact optical constraints.

For semantic consistency, we use frozen DINOv2 [Oquab et al. 2023] patch features to preserve the identity and appearance of reflected content. Let $\Phi(\cdot)$ denote the DINOv2 feature extractor. We compute semantic consistency at two scales: the mirror-local term $\mathcal{L}_{local}$ compares normalized patch features within the mirror region and provides the main constraint for reflected-content fidelity, while the global term $\mathcal{L}_{global}$ compares full-image features and acts as a

scene-level regularizer. The semantic consistency loss is:

$$\mathcal{L}_{sem} = \lambda_{global} \mathcal{L}_{global} + \lambda_{local} \mathcal{L}_{local}. \quad (10)$$

For lighting consistency, we regularize geometry-aware photometric coherence in the reflected region. A frozen dense prediction model, Lotus [He et al. 2025], estimates depth and surface normals from the decoded prediction $\hat{X}_0$. The depth-normal term $\mathcal{L}_{dn}$ encourages geometry-compatible local structure, while the spherical-harmonics term $\mathcal{L}_{sh}$ captures low-frequency illumination consistency under the estimated normals. Specifically, we fit second-order spherical harmonics in the mirror region and use the resulting normal-conditioned shading representation to define $\mathcal{L}_{sh}$, without requiring exact pixel-level lighting reproduction. Here, spherical harmonics are used as a compact descriptor of low-frequency scene illumination rather than as a reflectance model of the mirror surface. The lighting consistency loss is:

$$\mathcal{L}_{light} = \lambda_{dn} \mathcal{L}_{dn} + \lambda_{sh} \mathcal{L}_{sh}. \quad (11)$$

The final perceptual loss combines semantic and lighting consistency:

$$\mathcal{L}_{perc} = \lambda_{sem} \mathcal{L}_{sem} + \lambda_{light} \mathcal{L}_{light}. \quad (12)$$

4.4 Training Strategy

PhysReflect augments the standard diffusion denoising objective with training-time physical supervision. The overall training loss is:

$$\mathcal{L}_{total} = \mathcal{L}_{MSE} + w(t) (\lambda_{geo} \mathcal{L}_{geo} + \lambda_{perc} \mathcal{L}_{perc}), \quad (13)$$

where $\mathcal{L}_{MSE}$ denotes the standard noise-prediction loss, and $w(t)$ is a timestep-dependent weight for pixel-space physical supervision. We follow the three-stage training protocol of MirrorVerse [Dhiman et al. 2025a] and apply these losses on top of this base training procedure.

Both geometric and perceptual losses are computed on the decoded prediction $\hat{X}_0$ or structures derived from it. Their reliability therefore depends on the quality of the decoded image prediction at each diffusion timestep. At high-noise timesteps, $\hat{X}_0$ may contain unstable structures and inaccurate photometric cues, making geometric matching, boundary extraction, semantic feature comparison, and lighting estimation unreliable. We therefore introduce an annealed supervision scheme, where $w(t)$ down-weights losses at unreliable high-noise timesteps and emphasizes low-noise predictions where pixel-space physical measurements are more stable.

All auxiliary estimators, matchers, and cached targets are used only during training. At inference time, PhysReflect keeps the original BrushNet inputs and sampling procedure unchanged, introducing no additional inference overhead. Details of $w(t)$, single-step decoding, and caching are provided in Appendix.

5 Experiments

5.1 Experimental Setup

5.1.1 Datasets and Evaluation Metrics. We evaluate PhysReflect on the synthetic MirrorBenchV2 benchmark derived from SynMirrorV2 [Dhiman et al. 2025a] and the real-world MSD dataset [Yang et al. 2019]. SynMirrorV2 contains 201,610 Blender-rendered mirror-scene samples built from 66,062 target objects, and MirrorBenchV2 is the evaluation split derived from SynMirrorV2. We report results

on its single-object and multi-object splits. MSD contains 4,018 real indoor mirror images with cluttered layouts, diverse mirror shapes, and natural illumination, and is used to test real-world generalization.

We evaluate mirror reflections using three categories of metrics:

- **Visual Quality:** Measures pixel-level fidelity within the mirror region using PSNR $\uparrow$, SSIM $\uparrow$, and LPIPS $\downarrow$ [Zhang et al. 2018].
- **Semantic Consistency:** Assesses perceptual preservation of object identity and appearance via CLIP Similarity $\uparrow$ [Radford et al. 2021] and DINO Similarity $\uparrow$ [Oquab et al. 2023].
- **Physical Plausibility:** Evaluates geometry and illumination consistency. IoU $\uparrow$ and Boundary F1 $\uparrow$ measure mask and contour alignment, m_IoU $\uparrow$ and m_BF1 $\uparrow$ average multi-object results, and SH_Err $\downarrow$ measures low-frequency illumination consistency using second-order spherical harmonics.

5.1.2 Implementation Details. For a fair comparison, we follow the method settings of prior work and fine-tune the pretrained Stable Diffusion v1.5 [Rombach et al. 2022] model with the Adam optimizer, using a learning rate of 1×10^{-5} and a batch size of 4. Training is conducted in FP32 precision on four NVIDIA H100 GPUs, with 60,000, 30,000, and 20,000 steps for the single-object, multi-object, and real-world data stages, respectively. All methods are evaluated at 512×512 resolution using the same sampling strategy, prompt inputs, four random seeds per sample, and the best unmasked-region SSIM selection protocol of MirrorVerse [Dhiman et al. 2025a]. All ablation experiments are run on two NVIDIA H100 GPUs. Additional experimental Setup details are provided in Appendix.

5.2 Main Results on Synthetic Benchmarks

We first evaluate PhysReflect on the synthetic MirrorBenchV2 benchmark, where paired ground truth reflections and geometric annotations are available. This setting allows us to jointly measure image fidelity, semantic preservation, and physical plausibility under controlled reflection geometry. We report results separately on the single-object split (Table 1) and the more challenging multi-object split (Table 2) to expose how each method behaves as geometric complexity increases.

5.2.1 Single-Object Scenes. Table 1 shows that physical supervision improves reflection correctness without sacrificing image fidelity. The gains over MirrorVerse are modest on reconstruction metrics but much larger on IoU, Boundary F1, and SH_Err, which are the metrics most directly tied to our geometry and lighting losses. Together with the improvements in reconstruction and perceptual metrics and the human evaluation in Section 5.5, these results provide complementary evidence of improved reflection quality.

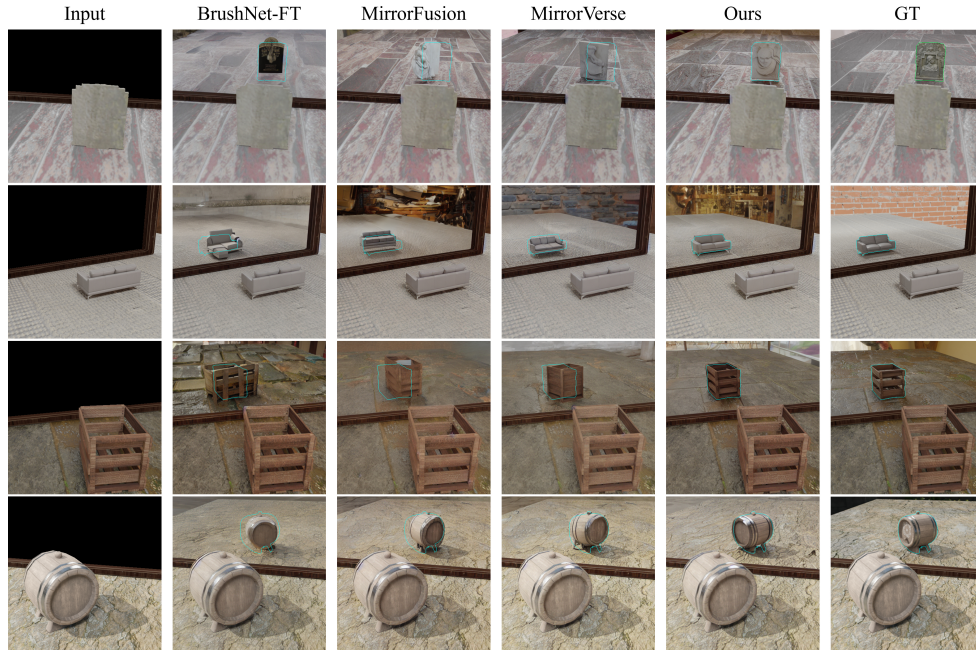
5.2.2 Multi-Object Scenes. The multi-object split in Table 2 stresses the method with occlusion, depth variation, and multiple reflected instances. The advantage of PhysReflect becomes stronger in this setting, especially on per-object geometry metrics. This supports the role of explicit physical constraints: when several objects compete for plausible mirror content, a data-only objective can produce locally convincing textures but loses instance-level placement, whereas our losses keep each reflection tied to its corresponding

Table 1. **Single-object results on MirrorBenchV2.** Best results are in **bold**, second best in underlined.

Method	Reflection Gen. Quality			Semantic Consist.		Physical Plausibility		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	CLIP Sim $\uparrow$	DINO Sim $\uparrow$	IoU $\uparrow$	Bound. F1 $\uparrow$	SH_Err $\downarrow$
BrushNet-FT [Ju et al. 2024]	18.44	0.76	0.121	25.75	<u>0.64</u>	0.474	0.207	107.42
MirrorFusion [Dhiman et al. 2025b]	18.38	0.76	0.122	25.89	0.63	0.409	0.182	51.48
MirrorVerse [Dhiman et al. 2025a]	<u>18.74</u>	<u>0.77</u>	<u>0.108</u>	25.69	0.64	0.598	<u>0.316</u>	<u>49.93</u>
PhysReflect (Ours)	19.14	0.78	0.103	<u>25.87</u>	0.68	0.671	0.406	15.33

Table 2. **Multi-object results on MirrorBenchV2.** This split includes larger depth variation, occlusion, and complex layouts. Best results are in **bold**, second best in underlined.

Method	Reflection Gen. Quality			Semantic Consist.		Physical Plausibility				
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	CLIP Sim $\uparrow$	DINO Sim $\uparrow$	IoU $\uparrow$	Bound. F1 $\uparrow$	m_IoU $\uparrow$	m_BF1 $\uparrow$	SH_Err $\downarrow$
BrushNet-FT [Ju et al. 2024]	17.37	<u>0.752</u>	0.143	27.22	<u>0.63</u>	0.534	0.223	0.442	0.198	8.19
MirrorFusion [Dhiman et al. 2025b]	17.15	0.734	0.144	26.64	0.59	0.541	0.249	0.457	0.225	5.28
MirrorVerse [Dhiman et al. 2025a]	17.90	0.745	<u>0.119</u>	26.04	0.61	0.638	<u>0.324</u>	0.534	<u>0.289</u>	5.04
PhysReflect (Ours)	18.44	0.764	0.113	<u>26.76</u>	0.64	0.734	0.467	0.621	0.417	4.51

Fig. 4. **Single-object qualitative comparison on MirrorBenchV2.** PhysReflect improves projected-target alignment, identity preservation, and boundary quality. The final column is rendered GT.

projected geometry. The high CLIP score of BrushNet-FT further illustrates that text-image alignment alone is not a reliable proxy for physically correct reflection layout.

5.3 Generalization to Real-World Scenes

We next evaluate on MSD to test whether the learned physical constraints transfer to real mirror scenes with natural illumination, clutter, diverse mirror shapes, and capture noise. Since MSD is a real-world dataset, depth maps and camera intrinsics are estimated rather than rendered, and metrics are computed against the captured

target image within the mirror region. We therefore treat this image as a reference target, not as deterministic rendered GT. We compare direct evaluation and MSD-adapted settings for both PhysReflect and MirrorVerse.

Table 3 shows the same trend on real scenes. PhysReflect does not maximize CLIP similarity, but it improves image fidelity, DINO-based visual consistency, and SH_Err, indicating better preservation of object appearance and lighting structure. The non-finetuned PhysReflect row already approaches or exceeds the MSD-finetuned MirrorVerse baseline on several metrics, which suggests that the

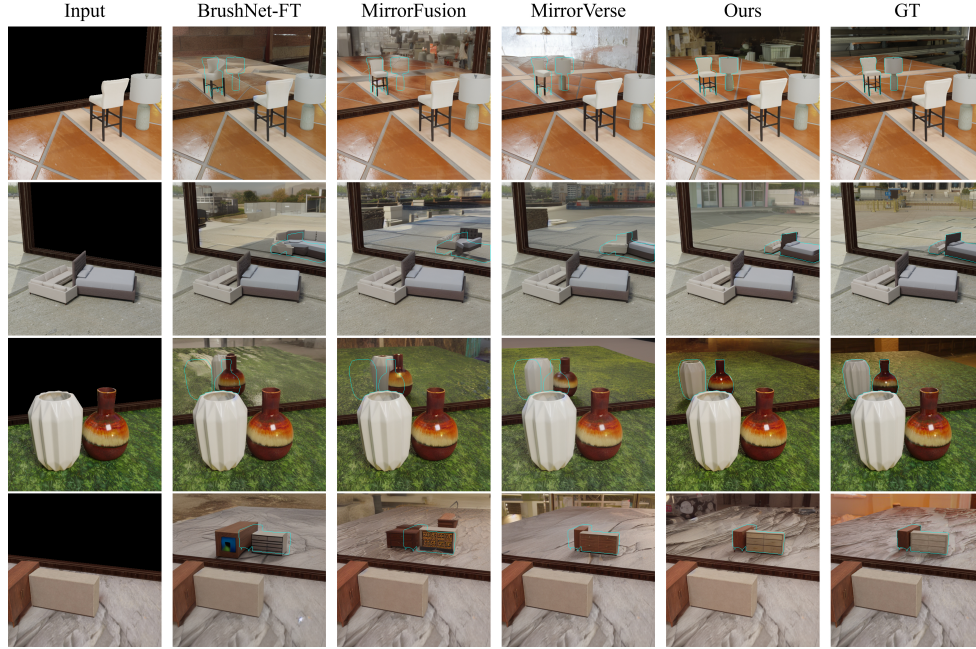


Fig. 5. **Multi-object qualitative comparison on MirrorBenchV2.** PhysReflect reduces missing, duplicated, or misplaced reflected objects under occlusion and depth variation. The final column is rendered GT.

Table 3. **Real-world generalization on MSD.** “FT” denotes MSD fine-tuning; methods without “FT” are evaluated directly on MSD. Best results are in **bold**, second best in underlined.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	CLIP $\uparrow$	DINO $\uparrow$	SH_Err $\downarrow$
MirrorVerse	16.76	0.843	0.132	26.90	0.48	128.74
MirrorVerse-FT	16.61	0.850	<u>0.119</u>	27.67	0.51	96.47
PhysReflect	<u>16.84</u>	<u>0.851</u>	0.129	26.41	<u>0.52</u>	<u>75.95</u>
PhysReflect-FT	17.11	0.860	0.114	26.72	0.58	65.69

proposed supervision transfers beyond the synthetic distribution rather than only memorizing dataset-specific reflection patterns.

5.4 Qualitative Results

The synthetic comparisons in Figures 4 and 5 show that PhysReflect reduces spatial drift, missing or duplicated reflected objects, and boundary mismatch in both single-object and multi-object scenes. The MSD visualization in Figure 6 further illustrates the same behavior on real images, where the final column is denoted as Reference because real captured reflections are used for visual comparison rather than deterministic rendered ground truth.

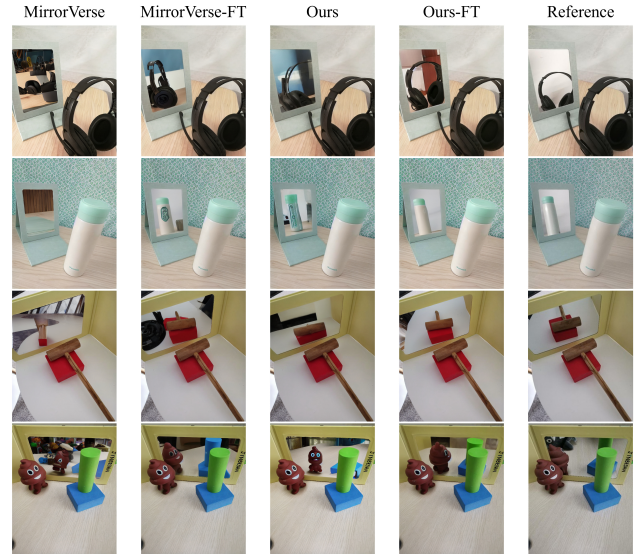


Fig. 6. **Qualitative comparison on MSD.** Reference denotes the captured target image used for visual comparison; MSD depth and camera intrinsics are estimated, so this target is not deterministic rendered GT.

5.5 User Study

To complement the automated metrics, we conducted a blind human perceptual study with nine evaluators. For each evaluated sample

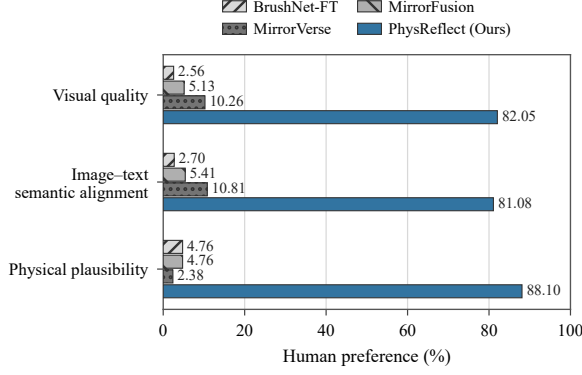


Fig. 7. **User study.** Distribution of recorded preferences (%) among consensus samples in a four-way comparison by nine evaluators. Higher is better.

and criterion, participants performed a four-way forced-choice comparison among PhysReflect, MirrorVerse [Dhiman et al. 2025a], MirrorFusion [Dhiman et al. 2025b], and BrushNet-FT [Ju et al. 2024]. A strict-majority preference was recorded when one method received at least five of the nine votes. Strict-majority consensus was reached for 58.5%, 55.5%, and 84.0% of the 200 samples for visual quality, image-text semantic alignment, and perceived physical plausibility, respectively, with corresponding no-consensus rates of 41.5%, 44.5%, and 16.0%. Figure 7 reports the preference distribution among consensus samples, while sample-level and individual-vote statistics are provided in the supplementary material. Among consensus samples, PhysReflect receives the largest preference share across all three criteria, supporting that its improvements are perceptually observable rather than limited to automated metrics.

5.6 Ablation Study

We distinguish between *components*, which control where and when supervision is applied, and *supervisions*, which define what physical signal is enforced. The components are TwinTrack and *Annealed Supervision*; the supervisions are epipolar supervision $\mathcal{L}_{epi}$, boundary supervision $\mathcal{L}_{boundary}$, semantic supervision $\mathcal{L}_{sem}$, and lighting supervision $\mathcal{L}_{light}$. Alternative Annealed Supervision variants are reported in the appendix.

Table 4 separates the effect of supervision scheduling from that of supervision content. Annealed Supervision improves the balance between feature consistency, reconstruction, and geometry, indicating that pixel-space supervision is most effective at reliable denoising stages. TwinTrack further strengthens boundary supervision by providing a cleaner detached selector for the generated reflection contour.

The supervision terms are complementary: epipolar and boundary losses improve geometric plausibility, semantic supervision preserves reflected content, and lighting supervision regularizes shading coherence. No single term dominates every metric, whereas the full model provides the strongest overall balance between physical plausibility and visual quality.

The ablation visualizations in Figures 8 and 9 provide complementary evidence for the design choices. The geometry and semantic

comparison shows that geometry supervision mainly anchors reflected content to the projected support, while semantic supervision improves object identity and texture preservation. As shown in the lighting ablation, the lighting loss improves highlights and the transition of reflection shading without changing the inference-time input.

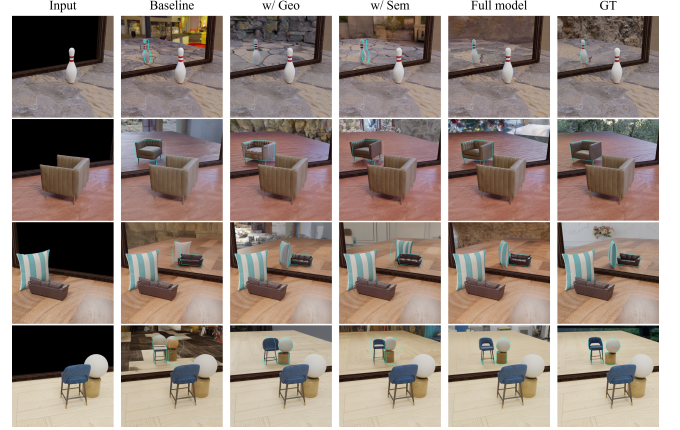


Fig. 8. **Geometry and semantic ablation.** Columns are Input, Baseline, w/ Geo, w/ Sem, Full, and GT, where w/ Geo and w/ Sem denote adding $\mathcal{L}_{epi} + \mathcal{L}_{boundary}$ and $\mathcal{L}_{sem}$ to the Baseline, respectively. Full denotes the complete model with all proposed components and losses.

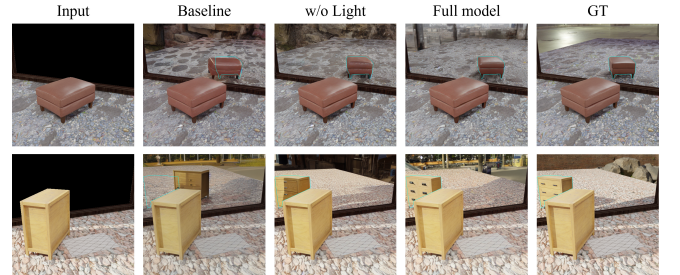


Fig. 9. **Lighting ablation.** The w/o lighting result denotes the Full model without the lighting loss, while the Full model uses all proposed losses. The lighting loss improves highlights, shading transitions, and mirror-region illumination coherence.

5.7 Efficiency

PhysReflect is designed to shift additional cost to training rather than inference. Table 5 confirms this behavior: because TwinTrack, DINOv2, and Lotus are frozen training-time modules, the deployed model keeps the same inputs and nearly the same computation graph as the BrushNet-based generator. The measured latency and memory remain essentially unchanged under both DDIM and UniPC sampling, so the physical gains above do not come from extra inference-time computation.

Table 4. **Ablation of components and supervisions on the single-object split of MirrorBenchV2.** Checkmarks indicate enabled components or losses. Annealed Supervision denotes the final inverse-timestep reweighting strategy. Best results are in **bold**.

Method						Metric								
Component		Supervision				Reflection Gen. Quality			Semantic Consist.		Physical Plausibility			
Annealed Supervision	TwinTrack	$\mathcal{L}_{epi}$	$\mathcal{L}_{boundary}$	$\mathcal{L}_{sem}$	$\mathcal{L}_{light}$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	CLIP Sim $\uparrow$	DINO Sim $\uparrow$	IoU $\uparrow$	Bound. F1 $\uparrow$	SH_Err $\downarrow$	
✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓	✓				18.33	0.76	0.109	25.68	0.61	0.554	0.296	58.41	
			✓				18.24	0.77	0.112	25.47	0.65	0.581	0.347	59.36
				✓			18.44	0.77	0.110	25.82	0.62	0.541	0.237	49.24
					✓		18.41	0.78	0.108	25.83	0.66	0.607	0.351	58.03
			✓				18.27	0.81	0.111	25.47	0.67	0.611	0.357	56.33
				✓			15.28	0.69	0.153	24.66	0.51	0.352	0.189	114.42
					✓		19.04	0.78	0.108	25.91	0.68	0.593	0.318	20.51
						✓	18.85	0.78	0.109	25.76	0.66	0.599	0.316	16.84
						✓	18.97	0.78	0.108	25.88	0.68	0.604	0.322	16.97
						✓	18.21	0.74	0.114	25.85	0.69	0.515	0.289	87.24
		✓	✓	✓	✓	✓	18.93	0.78	0.107	25.89	0.67	0.621	0.351	17.25

Table 5. **Inference cost comparison.** Average time over 100 images and peak memory for one 512×512 image on a single H100 GPU using FP32 precision. Lower is better.

Method	DDIM 50			UniPC 15		
	S-Time $\downarrow$	M-Time $\downarrow$	Mem. $\downarrow$	S-Time $\downarrow$	M-Time $\downarrow$	Mem. $\downarrow$
BrushNet-FT	7.12	7.06	9.76	1.28	1.25	9.97
MirrorVerse	7.31	7.29	9.89	1.45	1.42	10.21
PhysReflect	7.31	7.23	9.82	1.45	1.38	10.03

6 Conclusion

We presented PhysReflect, a geometry- and perception-guided diffusion framework that applies mirror-specific physical supervision to decoded clean-image predictions in pixel space during training. Through complementary geometric and perceptual objectives with reliability-aware annealed supervision, PhysReflect improves physical plausibility on synthetic and real-world benchmarks while leaving the inference procedure unchanged and preserving perceptual quality and efficiency.

Limitations. Our method is currently designed for predominantly planar-mirror scenes and relies on estimated geometric priors, whose errors may weaken supervision in challenging real-world cases. The synthetic-to-real domain gap also remains a limitation. The frozen training-time priors introduce additional training cost but do not affect inference efficiency. Representative failure cases are provided in the Appendix.

References

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

Black Forest Labs. 2024. FLUX. <https://blackforestlabs.ai/announcing-black-forest-labs/>.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. 2024. Directly Fine-Tuning Diffusion Models on Differentiable Rewards. In *International Conference on Learning Representations*.

Ankit Dhiman, Manan Shah, and R Venkatesh Babu. 2025a. Mirrorverse: Pushing diffusion models to realistically reflect the world. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 11239–11249.

Ankit Dhiman, Manan Shah, Rishubh Parihar, Yash Bhalgat, Lokesh R Boregowda, and R Venkatesh Babu. 2025b. Reflecting reality: Enabling diffusion models to produce faithful mirror reflections. In *2025 International Conference on 3D Vision (3DV)*. IEEE, 824–834.

Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and Yingcong Chen. 2025. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In *International Conference on Learning Representations*, Vol. 2025. 89454–89467.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, and Lu Sheng. 2024. EpiDiff: Enhancing Multi-View Synthesis via Localized Epipolar-Constrained Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9784–9794.

Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. 2024. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *European Conference on Computer Vision*. Springer, 150–168.

Zhexin Liang, Zhaoxi Chen, Yongwei Chen, Tianyi Wei, Tengfei Wang, and Xingang Pan. 2026. PI-Light: Physics-Inspired Diffusion for Full-Image Relighting. *arXiv preprint arXiv:2601.22135* (2026).

Jiaying Lin, Guodong Wang, and Rynson WH Lau. 2020. Progressive mirror detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3697–3705.

Jiayue Liu, Xiao Tang, Freeman Cheng, Roy Yang, Zhihao Li, Jianzhuang Liu, Yi Huang, Jiaqi Lin, Shiyong Liu, Xiaofei Wu, et al. 2024. MirrorGaussian: Reflecting 3d gaussians for reconstructing mirror reflections. In *European Conference on Computer Vision*. Springer, 377–393.

Pengwei Liu, Hangjie Yuan, Bo Dong, Jiazheng Xing, Jinwang Wang, Rui Zhao, Weihua Chen, and Fan Wang. 2026. UniLumos: Fast and Unified Image and Video Relighting with Physics-Plausible Feedback. *Advances in Neural Information Processing Systems* 38 (2026), 82052–82080.

Ruoshi Liu, Rundui Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023. Zero-1-to-3: Zero-shot One Image to 3D Object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9298–9309.

David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.

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

Zehong Ma, Ruihan Xu, and Shiliang Zhang. 2026. PixelGen: Improving Pixel Diffusion with Perceptual Supervision. (2026).

Haiyang Mei, Bo Dong, Wen Dong, Jiayi Yang, Seung-Hwan Baek, Felix Heide, Pieter Peers, Xiaopeng Wei, and Xin Yang. 2022. Glass segmentation using intensity and spectral polarization cues. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12622–12631.

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

Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. 2024. 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*. 7465–7475.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).

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

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

- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, Vol. 2024. 1862–1874.
- Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. 2023. Aligning Text-to-Image Diffusion Models with Reward Backpropagation. (2023).
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmlR, 8748–8763.
- Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. 2025. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, Vol. 2025. 28085–28128.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.
- Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. 2011. ORB: An efficient alternative to SIFT or SURF. In *2011 International conference on computer vision*. Ieee, 2564–2571.
- Ayush Sarkar, Hanlin Mai, Amitabh Mahapatra, Svetlana Lazebnik, David A Forsyth, and Anand Bhattad. 2024. 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*. 28140–28149.
- Stability AI. 2024. Stable Diffusion 3.5. <https://stability.ai/news/introducing-stable-diffusion-3-5>.
- Rishi Upadhyay, Howard Zhang, Yunhao Ba, Ethan Yang, Blake Gella, Sicheng Jiang, Alex Wong, and Achuta Kadambi. 2023. Enhancing diffusion models with 3d perspective geometry constraints. *ACM Transactions on Graphics (TOG)* 42, 6 (2023), 1–15.
- Daniel Winter, Matan Cohen, Shlomi Fruchter, Yael Pritch, Alex Rav-Acha, and Yedid Hoshen. 2024. Objectdrop: Bootstrapping counterfactuals for photorealistic object removal and insertion. In *European Conference on Computer Vision*. Springer, 112–129.
- Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. In *Advances in Neural Information Processing Systems*, Vol. 36. 15903–15935.
- Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson WH Lau. 2019. Where is my mirror?. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8809–8818.
- Ze-Xin Yin, Jiaxiong Qiu, Ming-Ming Cheng, and Bo Ren. 2023. Multi-space neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12407–12416.
- Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. 2025. Representation Alignment for Generation: Training Diffusion Transformers Is Easier Than You Think. In *International Conference on Learning Representations*.
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3836–3847.
- Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.