PHYSICAL-ANCHOR DIFFUSION: A COARSE-TO-FINE FRAMEWORK FOR SHADOW-AWARE IMAGE HARMONIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Realistic image compositing requires jointly resolving foreground appearance, illumination, and cast-shadow consistency. These factors are mutually constrained: appearance adjustment alone cannot correct lighting mismatch, while unconstrained shadow generation may improve local realism but violate scene geometry. We therefore propose Physical-Anchor Diffusion (PAD), a unified coarse-to-fine framework for shadow-aware image compositing. PAD first harmonizes the foreground in intrinsic albedo space and estimates background illumination using an environment-aware spherical-harmonic representation to relight the foreground consistently with the scene. It then combines the reconstructed scene geometry with the estimated illumination direction to construct a physically grounded coarse shadow prior that encodes shadow direction, contact region, and global layout. Conditioned on this prior, a diffusion model refines the shadow by recovering soft boundaries and realistic local appearance. By constraining global shadow structure with geometry while delegating fine-scale appearance synthesis to diffusion, PAD balances physical plausibility and visual realism. Experiments on the DESOBAv2 benchmark and real-world compositing cases demonstrate that PAD achieves stronger shadow structural fidelity while improving the perceived coherence of foreground appearance, illumination, and cast shadows in the final composites.

1 INTRODUCTION

High-fidelity image compositing (Niu et al., 2021) aims to seamlessly insert a foreground object into a target background while preserving visual realism and physical consistency, with applications in image editing (Pérez et al., 2023; Brooks et al., 2023), digital content creation (Song et al., 2023), virtual advertising (Niu et al., 2021), and augmented reality (Liu et al., 2020).

Realistic compositing requires not only foreground–background appearance matching, but also illumination consistency and physically plausible cast shadows. Cast shadows couple foreground illumination with scene geometry and object placement, making their consistency essential for natural foreground–scene interaction.

Existing methods usually address only part of this problem. Appearance-based approaches (Pérez et al., 2023; Reinhard et al., 2002) lack explicit illumination and shadow modeling, while intrinsic-based methods (Bell et al., 2014; Careaga et al., 2023) enable foreground relighting but pay limited attention to cast-shadow consistency. Diffusion-based methods (Song et al., 2023; Yu et al., 2024; Liu et al., 2024) improve visual realism, and recent geometry-aware approaches introduce geometric cues for more plausible shadow generation (Zhao et al., 2025). However, shadow reasoning is still commonly performed directly from the input composite, where the inserted foreground may retain source-scene shading and illumination cues that conflict with the target background. Moreover, geometry is often used as guidance for learned generation rather than to explicitly determine the global shadow layout.

These limitations motivate two key considerations. First, shadow reasoning should operate on a scene-consistent representation, with foreground appearance and illumination adapted to the target scene beforehand. Second, shadow structure and appearance require different inductive biases:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

global shadow layout should follow geometry and illumination through physical projection, whereas soft boundaries, penumbra, and local details benefit from generative priors.

Based on these observations, we propose **Physical-Anchor Diffusion (PAD)**, a coarse-to-fine framework for shadow-aware image harmonization. PAD first performs intrinsic foreground harmonization and environment-aware relighting to reduce source-scene appearance and illumination discrepancies. It then combines reconstructed scene geometry with the estimated illumination direction to explicitly construct a coarse shadow prior that anchors shadow direction, contact relationship, and global layout. Conditioned on this scene-consistent representation and physical anchor, a diffusion model refines local shadow appearance and details.

Unlike existing diffusion-based shadow generation, including geometry-guided approaches (Liu et al., 2024; Zhao et al., 2025), PAD differs in both *where* and *how* physical reasoning is introduced: it establishes foreground–scene consistency before shadow reasoning and explicitly converts geometry and illumination into a physical shadow hypothesis before diffusion refinement. This separates physically constrained global structure from generative local appearance synthesis.

Our main contributions are:

- We formulate **shadow-aware image harmonization** as a progressive scene-consistency problem, adapting foreground appearance and illumination before modeling cast-shadow interaction with the target scene.
- We propose **Physical-Anchor Diffusion (PAD)**, which explicitly constructs a geometry- and illumination-guided coarse shadow prior through physical projection, anchoring global shadow structure before diffusion refinement.
- Experiments on benchmark and real-world scenarios demonstrate improved shadow structural fidelity and perceptual scene coherence, with controlled studies validating the effectiveness and robustness of physical anchoring.

2 Related Work

2.1 High-Fidelity Image Compositing

High-fidelity image compositing (Niu et al., 2021) aims to insert foreground objects into target scenes with natural visual integration. Early approaches mainly rely on low-level image-space adjustments (Pérez et al., 2023; Reinhard et al., 2002), such as color transfer and tone correction. Although effective for reducing appearance discrepancy, these methods lack explicit modeling of illumination and shadow formation, limiting their reliability under complex lighting conditions (Tsai et al., 2017).

Subsequent illumination-aware and intrinsic-based methods (Bell et al., 2014; Careaga et al., 2023) decompose images into albedo and shading to enable more controllable foreground adaptation and relighting (Barron & Malik, 2014; Cong et al., 2020). However, they mainly focus on the foreground itself and pay limited attention to its illumination-dependent interaction with the surrounding scene through cast shadows (Careaga et al., 2023; Yu et al., 2024).

More recently, generative methods have improved visual naturalness and detail synthesis in compositing (Song et al., 2023; Yu et al., 2024). Nevertheless, shadow generation from an unadapted composite may still be affected by source-scene appearance and illumination cues retained in the inserted foreground. In contrast, PAD treats foreground adaptation as a prerequisite for shadow reasoning, establishing scene-consistent appearance and illumination before modeling cast-shadow interaction.

2.2 Shadow-Aware Image Harmonization

Shadows provide essential cues about object placement, scene geometry, and illumination in realistic compositing (Zhou et al., 2024; Zhao et al., 2025). Existing shadow generation methods broadly follow physics-based or data-driven paradigms (Zhang et al., 2019; Liu et al., 2024).

Physics-based approaches (Guennebaud et al., 2006; Jin et al., 2021; Zhou et al., 2024) use geometric and illumination information to determine shadow projection and global structure. However,

simplified physical assumptions and imperfect scene estimation often lead to rigid shadows with limited local appearance realism (Zhou et al., 2024). In contrast, data-driven approaches (Zhang et al., 2019; Hong et al., 2022; Tao et al., 2024; Hu et al., 2019; Liu et al., 2020) provide flexible appearance synthesis but may suffer from inaccurate shadow structure, contact relationships, or spatial alignment due to limited explicit geometric control (Liu et al., 2024; Zhao et al., 2025).

Recent geometry-aware methods introduce geometric cues to improve shadow plausibility (Zhao et al., 2025), but global shadow structure may still be largely resolved through learned generation rather than explicitly determined by physical projection. In contrast, PAD converts reconstructed geometry and estimated illumination into an explicit coarse shadow hypothesis, anchoring shadow direction, contact relationship, and global layout before generative refinement. This combines the structural reliability of physical reasoning with the appearance flexibility of learned synthesis.

2.3 Diffusion-Based Synthesis

Diffusion models have achieved remarkable success in image generation and editing due to their ability to synthesize realistic details and coherent structures (Ho et al., 2020; Rombach et al., 2022; Brooks et al., 2023). In compositing and harmonization, diffusion-based methods further improve visual naturalness, boundary quality, and fine-detail synthesis (Song et al., 2023; Yu et al., 2024).

For shadow generation, diffusion models provide strong priors for realistic appearance synthesis but do not inherently guarantee physically plausible global structure (Liu et al., 2024). Recent methods therefore incorporate geometric or physical cues to improve structural controllability (Zhao et al., 2025). Nevertheless, even with additional geometric guidance, learned generation may still bear substantial responsibility for resolving both global shadow structure and local appearance.

Table 1: Comparison with representative compositing methods. "App. Harm." denotes foreground appearance harmonization, "FG Reshading" denotes foreground reshading, and "Cast Shadow" denotes cast-shadow generation. SA denotes SIGGRAPH Asia.

Method	App. Harm.	FG Reshading	Cast Shadow
DoveNet (CVPR'20)	✓	✗	✗
Intrinsic Harm. (SA'23)	✓	✓	✗
ShadowGAN (CVM'19)	✗	✗	✓
SGRNet (AAAI'22)	✗	✗	✓
ARShadowGAN (CVPR'20)	✗	✗	✓
SGDiffusion (CVPR'24)	✗	✗	✓
GPSDiffusion (CVPR'25)	✗	✗	✓
PAD (Ours)	✓	✓	✓

In contrast, PAD adopts a different division of responsibility: geometry and illumination explicitly determine a coarse shadow hypothesis through physical projection, while diffusion focuses on refining soft boundaries and local appearance under this structural anchor. This physically anchored refinement reduces the burden on diffusion to infer global shadow structure while retaining its flexibility for realistic appearance synthesis. As summarized in Table 1, PAD further connects this shadow refinement with foreground harmonization and illumination-aware reshading in a progressive shadow-aware compositing framework.

3 Method

Given a composite image $I \in \mathbb{R}^{H\times W\times 3}$ and foreground mask $M \in \{0,1\}^{H\times W}$, our goal is to generate a final composite image I^* with consistent foreground appearance, illumination, and cast shadow. Unlike conventional harmonization methods that mainly adjust foreground appearance, we propose **Physical-Anchor Diffusion (PAD)**, a coarse-to-fine framework that jointly models foreground illumination adaptation and shadow formation under scene-consistent constraints, as illustrated in Fig. 1.

PAD consists of two stages: *scene-consistent foreground adaptation* and *physics-anchored shadow refinement*. Given (I, M), the first stage constructs an illumination-consistent representation I_r through intrinsic decomposition, albedo harmonization, normal and illumination estimation, and foreground reshading. The second stage reconstructs scene geometry, estimates the dominant shadow projection direction, and constructs a geometry-guided coarse shadow prior C, which anchors the global shadow structure before conditional diffusion refinement. The overall process is formulated as:

$$(I, M) \rightarrow (A, S, N, SH, d) \rightarrow I_r \rightarrow P \rightarrow C \rightarrow I^*. \tag{1}$$

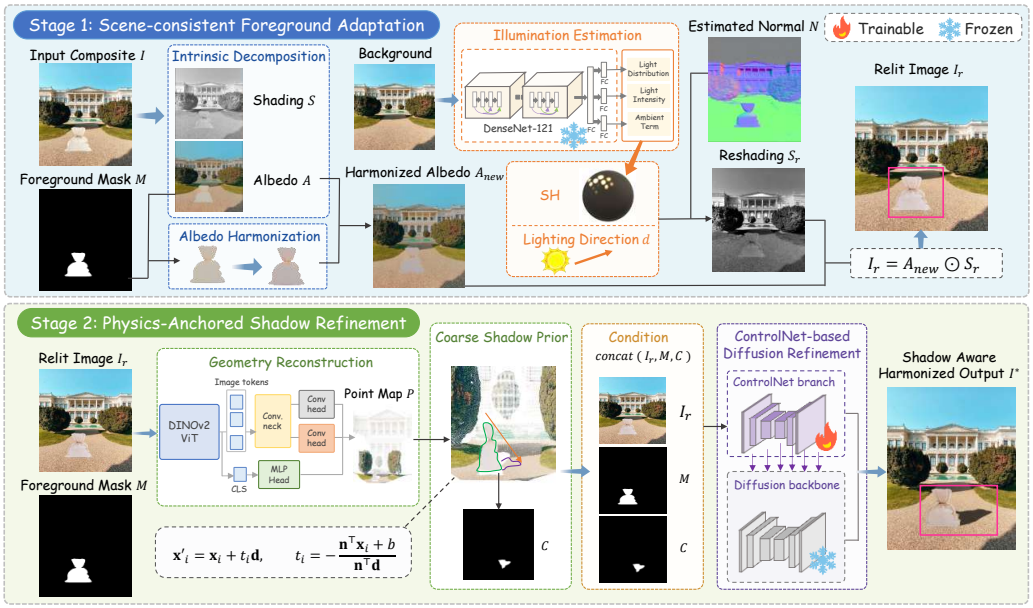

Figure 1: Overview of the proposed Physical-Anchor Diffusion (PAD) framework. The framework first obtains an illumination-consistent representation, constructs a geometry-guided shadow prior, and refines shadows with diffusion.

Here, A and S denote albedo and shading, N is the estimated surface-normal map, SH represents environment illumination using spherical harmonics, and $d \in \mathbb{R}^3$ is the normalized dominant shadow-projection direction derived from SH. I_r denotes the harmonized and relit intermediate image, P is the dense point map reconstructed from I_r, and C is the geometry-guided coarse shadow prior. Finally, the diffusion model generates I^* conditioned on (I_r, M, C), where geometry and illumination constrain global shadow structure while diffusion refines local shadow appearance.

3.1 Scene-consistent Foreground Adaptation

The first stage constructs an illumination-consistent representation by reducing foreground-background appearance and illumination mismatch.

Intrinsic Decomposition. Given the composite image I, we use a pretrained intrinsic decomposition model (Careaga et al., 2023) to decompose it into albedo and shading:

$$I = A \odot S, \tag{2}$$

where A and S denote the albedo and shading components, respectively. The decomposition is performed in linear RGB space to provide a physically meaningful representation for subsequent albedo harmonization and foreground reshading.

Albedo Harmonization. A pretrained albedo harmonization model (Careaga et al., 2023) is applied to the foreground albedo to reduce reflectance inconsistency between the inserted object and target scene. The resulting harmonized albedo is denoted as A_{new}, while the background region remains unchanged.

Normal and Illumination Estimation. We construct the surface-normal representation using complementary geometry cues from pretrained MoGe-2 (Wang et al., 2026) and Omnidata (Eftekhar et al., 2021), where MoGe-2 provides foreground geometry cues and Omnidata provides surface-normal estimates for the surrounding background, yielding the normal map N for foreground reshading. Meanwhile, environment illumination is estimated from the background region $(1-M)\odot I$ using EMLight (Zhan et al., 2021) and represented by spherical harmonic coefficients SH.

Foreground Reshading. Given the estimated normal map N and illumination coefficients SH, foreground shading is reconstructed through spherical harmonic rendering:

$$S_r = (1 - M) \odot S + M \odot R(N, SH), \tag{3}$$

where $R(\cdot)$ denotes the spherical harmonic rendering function. This preserves the original background shading while reconstructing the foreground shading under the estimated scene illumination.

The illumination-consistent intermediate image is obtained as:

$$I_r = A_{\text{new}} \odot S_r, \tag{4}$$

where I_r preserves foreground reflectance while aligning its appearance and shading with the target illumination.

3.2 Physics-Anchored Shadow Refinement

Given the illumination-consistent intermediate image I_r, PAD generates physically plausible shadows through geometry-guided prior construction and conditional diffusion refinement.

Dominant Shadow Projection Direction. EMLight (Zhan et al., 2021) represents scene illumination as directional samples on the sphere together with corresponding intensity and ambient components. To obtain a single direction for geometric shadow projection, we evaluate the directional illumination samples and select the direction with the highest non-ambient illumination energy as the dominant incident-light direction. This direction is reversed to obtain the shadow projection direction, transformed into the coordinate convention of the geometry module, and normalized as $d \in \mathbb{R}^3$. In our formulation, d points from the occluder toward potential shadow receivers and is used only for coarse shadow construction.

Geometry Reconstruction. The pretrained MoGe-2 model (Wang et al., 2026) is applied to I_r to obtain a dense point representation $P \in \mathbb{R}^{H\times W\times 3}$, where each pixel corresponds to a 3D scene point. This representation provides the geometric basis for identifying potential shadow receivers under the estimated projection direction.

Coarse Shadow Prior. Given the foreground mask M and point map P, we extract the largest external foreground contour, sample contour points, and lift them into 3D using their corresponding points in P, obtaining a compact geometric representation of the occluder.

We then construct local receiver hypotheses from background points surrounding the foreground. Specifically, nearby background points are partitioned into lower and upper/side regions relative to the bottom of the foreground, yielding floor-like and wall-like receiver candidates. A local plane is fitted to the corresponding 3D points of each candidate using SVD. These planes serve as local geometric approximations rather than semantic floor or wall predictions.

Let $\mathbf{x}_i \in \mathbb{R}^3$ denote the i-th sampled 3D contour point. For a receiver plane with unit normal $\mathbf{n}$ and offset b, the shadow projection is:

$$\mathbf{x}'_i = \mathbf{x}_i + t_i d, \qquad t_i = -\frac{\mathbf{n}^\top \mathbf{x}_i + b}{\mathbf{n}^\top d}. \tag{5}$$

Only intersections satisfying $t_i > 0$ are retained, while degenerate near-parallel ray-plane configurations are discarded. The projected contour is then mapped back to the image plane using the projection associated with P.

To compensate for residual geometric inconsistencies, we first apply a lightweight contact correction to the projected contour and then estimate a RANSAC-based 2D affine deformation from the original contour to the projected contour, which is applied to the foreground mask. This preserves object-dependent shadow shape while following the geometry and illumination constraints defined by P and d.

Applying this procedure to the floor and wall planes produces two candidate masks C_{f} and C_{w}, which are fused as:

$$C = (1 - M) \odot \max(C_{\text{f}}, C_{\text{w}}). \tag{6}$$

The resulting prior C provides structural guidance for shadow direction, contact relationship, and global spatial layout, while diffusion refines boundary transitions and local appearance. Detailed implementation settings are provided in the Appendix.

Table 2: Quantitative comparison with state-of-the-art methods on the DESOBAv2 benchmark. Results are reported on the BOS and BOS-free test subsets. Results of previous methods are taken from VSDiffusion (Li & Zhang, 2026), which uses five generated samples with oracle selection, while PAD reports results from a single sampling output.

Method	BOS Test Images						BOS-Free Test Images					
	GR ↓	LR ↓	GS ↑	LS ↑	GB ↓	LB ↓	GR ↓	LR ↓	GS ↑	LS ↑	GB ↓	LB ↓
ShadowGAN (Zhang et al., 2019)	7.511	67.464	0.961	0.197	0.446	0.890	17.325	76.508	0.901	0.060	0.425	0.842
Mask-SG (Hu et al., 2019)	8.997	79.418	0.951	0.180	0.500	1.000	19.338	94.327	0.906	0.044	0.500	1.000
AR-SG (Liu et al., 2020)	7.335	58.037	0.961	0.241	0.383	0.761	16.067	63.713	0.908	0.104	0.349	0.682
SGRNet (Hong et al., 2022)	7.184	68.255	0.964	0.206	0.301	0.596	15.596	60.350	0.909	0.100	0.271	0.534
DMASNet (Tao et al., 2024)	8.256	59.380	0.961	0.228	0.276	0.547	18.725	86.694	0.913	0.055	0.297	0.574
SGDiffusion (Liu et al., 2024)	6.098	53.611	0.971	0.370	0.245	0.487	15.110	55.874	0.913	0.117	0.233	0.452
GPSDiffusion (Zhao et al., 2025)	**5.896**	46.713	0.966	0.374	0.213	0.423	13.809	55.616	0.917	0.166	0.197	0.384
VSDiffusion (Li & Zhang, 2026)	6.422	46.542	0.964	0.377	**0.182**	**0.360**	14.629	55.377	0.912	0.184	**0.194**	**0.373**
PAD (Ours)	6.397	**38.093**	**0.982**	**0.622**	0.321	0.639	**13.670**	**47.832**	**0.955**	**0.560**	0.329	0.646

Conditional Diffusion Refinement. Following SGDiffusion (Liu et al., 2024), we adopt a ControlNet-based conditional diffusion architecture (Zhang et al., 2023), where the illumination-consistent image I_r, foreground mask M, and coarse shadow prior C are used as conditions to generate the final shadow-aware composite image I^*.

We follow the standard SGDiffusion denoising objective:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t; I_r, M, C)\|_2^2\right], \tag{7}$$

where x_0 denotes the ground-truth shadow-aware composite image and x_t is its noisy version at timestep t.

Following SGDiffusion (Liu et al., 2024), an auxiliary binary cross-entropy loss $\mathcal{L}_{\text{aux}}$ is introduced for shadow localization supervision:

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}}, \tag{8}$$

where $\lambda_{\text{aux}} = 0.1$ balances the diffusion objective and auxiliary shadow supervision.

4 Experiments

4.1 Experimental Settings

Implementation Details. We implement PAD in PyTorch and train it on a single NVIDIA RTX 4090 GPU. The diffusion model is initialized from pretrained SGDiffusion weights (Liu et al., 2024) and fine-tuned on DESOBAv2 using AdamW with a learning rate of 1×10^{-5} and a batch size of 1. The BOS background shadow hint is not used during training to reflect practical scenarios.

Dataset and Metrics. We evaluate all methods on DESOBAv2 (Liu et al., 2023) under the standard train/test split. The dataset contains two test subsets: **BOS**, which provides paired foreground–shadow annotations, and **BOS-free**, which does not contain such references and is therefore more challenging for generalization. Unless otherwise specified, results are reported on both subsets.

We report six standard metrics (Liu et al., 2024): global/local RMSE (GRMSE/LRMSE) for reconstruction error, global/local SSIM (GSSIM/LSSIM) for structural similarity, and global/local BER (GBER/LBER) for binary shadow-mask accuracy. Global metrics are computed over the entire image, while local metrics are computed within the ground-truth shadow region. Lower RMSE and BER and higher SSIM indicate better performance.

Computational Overhead. We analyze the inference cost of PAD on a single NVIDIA RTX 4090 GPU. For a representative image, the full pipeline requires 32.38 s with a peak memory usage of 21.97 GB. Diffusion-based shadow refinement is the main bottleneck, requiring 19.11 s. Therefore, the current implementation is more suitable for offline compositing rather than real-time applications.



Figure 2: Qualitative comparison with existing methods. Each row corresponds to a different scene, and each column shows the result of a different method.

4.2 Comparison with State-of-the-Art Methods

4.2.1 Quantitative Comparison

Tab. 2 presents the quantitative comparison on DESOBAv2. PAD achieves the best **LRMSE**, **GSSIM**, and **LSSIM** on both BOS and BOS-free subsets, and obtains the best **GRMSE** on BOS-free. Compared with GPSDiffusion (Zhao et al., 2025), PAD reduces LRMSE from 46.713 to 38.093 and improves LSSIM from 0.374 to 0.622 on BOS. Similar improvements are observed on BOS-free, where LRMSE decreases from 55.616 to 47.832 and LSSIM increases from 0.166 to 0.560. These results demonstrate that the proposed coarse-to-fine design effectively preserves shadow structure through physical anchoring while improving local realism through diffusion refinement.

PAD does not achieve the best **GBER** and **LBER**, since it prioritizes structurally coherent soft shadows over strict binary overlap. Moreover, PAD does not use BOS shadow hints during training, making the evaluation closer to practical scenarios without auxiliary references. Therefore, PAD achieves stronger structural metrics, especially in challenging BOS-free settings.

4.2.2 Qualitative Comparison

Fig. 2 compares representative composite scenes with large foreground–background discrepancies in illumination, color tone, and scene context. PAD generates shadows with more consistent lighting, more stable global structure, and smoother local transitions.

IHIC mainly improves foreground appearance adaptation but does not explicitly model cast shadows (Careaga et al., 2023), resulting in weaker scene interaction. SGDiffusion and GPSDiffusion generate plausible shadows (Liu et al., 2024; Zhao et al., 2025), but may suffer from inaccurate projection direction, structural artifacts, or insufficient adaptation to scene geometry. In contrast, PAD benefits from the illumination-consistent relit representation and geometry-guided coarse prior, which jointly provide guidance for shadow layout and appearance refinement.

As shown in Fig. 2, PAD generates more accurate shadow direction, support region, and contact relationship while achieving smoother boundaries and better foreground–background coherence.

4.3 Ablation Studies

To understand the contribution of each component in PAD, we analyze three key modules: *(1)* illumination harmonization and reshading, *(2)* geometry-guided coarse shadow prior, and *(3)* diffusion



(a) Input Composite | **(b) Ours** | **(c) w/o Diffusion Refine** | **(d) w/o Coarse Shadow Prior** | **(e) w/o Reshading** | **(f) w/o Harmonization**

Figure 3: Component-wise qualitative ablation of PAD. The first column shows the input composite image and its foreground mask. The remaining columns present the generated results and corresponding shadow masks under different component settings.

Table 3: Controlled comparison of coarse shadow priors on DESOBAv2. All variants use the same diffusion backbone, training data, optimization settings, and illumination-consistent conditional inputs, and differ only in the construction of the coarse shadow prior.

Coarse Prior	BOS				BOS-Free			
	GRMSE ↓	LRMSE ↓	GSSIM ↑	LSSIM ↑	GRMSE ↓	LRMSE ↓	GSSIM ↑	LSSIM ↑
No coarse prior	6.285	51.742	0.973	0.401	14.982	54.936	0.922	0.194
Learned box prior	**5.971**	45.684	0.975	0.447	13.944	53.281	0.929	0.236
Geometry-guided physical prior	6.397	**38.093**	**0.982**	**0.622**	**13.670**	**47.832**	**0.955**	**0.560**

refinement, which respectively improve illumination consistency, global shadow structure, and local appearance.

Component-wise Qualitative Ablation. Fig. 3 illustrates the effect of removing each component. The full model produces composites with both structurally plausible shadows and consistent foreground appearance. Removing diffusion refinement produces rigid shadows with less natural boundaries, while removing the coarse prior degrades shadow direction and global layout. Removing illumination harmonization or reshading increases foreground–background inconsistency due to illumination mismatch. These results indicate that the three components are complementary: foreground adaptation establishes illumination consistency, the physical prior constrains global shadow formation, and diffusion refinement recovers realistic local appearance.

Controlled Comparison of Coarse Shadow Priors. To isolate the effect of coarse shadow guidance, we compare three variants under the same diffusion backbone, training data, optimization settings, and illumination-consistent conditional inputs: *(1) no coarse prior*, where shadow structure is inferred only from the relit image and foreground mask; *(2) learned box prior*, where a coarse shadow region is predicted from data; and *(3)* our *geometry-guided physical prior*, constructed from the estimated scene geometry and shadow projection direction.

As shown in Table 3, explicit coarse priors consistently improve shadow structure. On BOS, the learned box prior reduces LRMSE from 51.742 to 45.684 compared with no prior, while our geometry-guided physical prior further reduces it to 38.093 and improves LSSIM from 0.401 and 0.447 to 0.622, respectively. A similar trend is observed on BOS-free, where the physical prior reduces LRMSE from 54.936 and 53.281 to 47.832 and improves LSSIM from 0.194 and 0.236 to 0.560. These results demonstrate that geometry- and illumination-based projection provides more reliable shadow support and spatial layout than purely learned localization, particularly when appearance cues alone are insufficient.

The geometry-guided prior does not achieve the lowest GRMSE on BOS (6.397 vs. 5.971 for the learned box prior). Since GRMSE is computed over the entire image and can be dominated by large non-shadow regions, it is less sensitive to local shadow alignment; the substantially improved LRMSE and structural metrics therefore provide complementary evidence for the effectiveness of the proposed prior.

Table 4: Robustness to illumination-direction perturbations on DESOBAv2.

Direction Perturbation	BOS				BOS-Free			
	GRMSE ↓	LRMSE ↓	GSSIM ↑	LSSIM ↑	GRMSE ↓	LRMSE ↓	GSSIM ↑	LSSIM ↑
0° (PAD)	**6.397**	**38.093**	**0.982**	**0.622**	**13.670**	**47.832**	**0.955**	**0.560**
5°	6.560	39.420	0.980	0.610	13.890	48.950	0.952	0.546
10°	6.910	41.760	0.975	0.587	14.280	51.120	0.946	0.518
20°	7.680	46.810	0.962	0.528	15.140	55.480	0.933	0.455

Finally, DESOBAv2 primarily evaluates shadow generation, while its foreground–background pairs are relatively consistent in appearance and illumination. Consequently, the benefits of illumination harmonization and reshading are not fully captured by shadow-oriented metrics but are qualitatively reflected in Fig. 3.

Sensitivity to Illumination Direction. We further evaluate robustness to illumination-estimation errors by perturbing the estimated shadow-projection direction by $5°$, $10°$, and $20°$ while keeping all other components unchanged. As shown in Table 4, PAD remains stable under small perturbations and gradually degrades as the angular error increases, demonstrating tolerance to moderate direction-estimation errors while confirming that the estimated illumination direction provides meaningful guidance for global shadow structure.

4.4 User Study

We conduct a user study with 35 participants on 20 real-world foreground–background pairs. Participants compare anonymized and randomly ordered results from SGDiffusion, GPSDiffusion, and PAD according to two criteria: *(1)* physical plausibility of the cast shadow and *(2)* overall visual naturalness of the composite. Each result is additionally scored using a five-point scale, where 5 indicates the best quality. As shown in Table 5, PAD achieves the highest preference rates and average scores on both criteria, demonstrating improved physical plausibility and visual coherence compared with existing methods.

Table 5: User study results on physical plausibility and visual naturalness. Pref. denotes the percentage of trials in which a method is selected as the best result. Scores are averaged over five-point ratings.

Method	Physical Plaus. Pref.	Visual Nat. Pref.	Physical Plaus. Score	Visual Nat. Score
SGDiffusion	22.14%	19.86%	3.31	2.84
GPSDiffusion	22.86%	21.43%	3.68	3.97
PAD (Ours)	**55.00%**	**58.71%**	**4.13**	**4.17**

5 Conclusion and Limitations

We presented **Physical-Anchor Diffusion (PAD)**, a coarse-to-fine framework for shadow-aware image compositing. PAD first establishes scene-consistent foreground appearance and illumination, then constructs a coarse shadow hypothesis from scene geometry and estimated illumination for diffusion refinement. Experiments on DESOBAv2 demonstrate improved shadow structural fidelity and local reconstruction quality, particularly on the challenging BOS-free subset, while controlled studies validate the effectiveness and robustness of physical anchoring.

PAD has several limitations. First, the physical anchor assumes a dominant illumination direction, which may be insufficient for complex multi-light environments. Second, the framework depends on estimated geometry and illumination, and errors in these intermediate predictions can propagate to the coarse shadow prior and subsequent refinement. Although diffusion can tolerate moderate perturbations in the physical anchor, substantial estimation errors may still lead to inaccurate shadow structures. Finally, diffusion refinement introduces considerable computational and memory overhead, limiting the applicability of the current framework to real-time scenarios.

PAD is modular with respect to intrinsic decomposition, illumination estimation, geometry reconstruction, and diffusion refinement. Future work will explore uncertainty-aware physical projection, richer illumination modeling, more efficient refinement, and extensions to video and multi-object compositing.

AI USE STATEMENT

In this work, we have not used generative AI tools for any research task that requires disclosure, including generating synthetic data, developing theoretical models or conceptual frameworks, formulating or proving mathematical claims, proposing hypotheses, designing research methodology or experiments, implementing methods, and interpreting experimental results. We used generative AI tools solely for grammar checking and correcting minor grammatical errors in the manuscript. All AI-assisted edits were carefully reviewed by the authors to ensure that they did not alter the technical content, scientific claims, or reported results. We take responsibility for the final content of this work, including all text, claims, and results.

ETHICS STATEMENT

This work uses the publicly available DESOBAv2 benchmark for training and evaluation, and no new dataset or personally identifiable information is collected. Our user study involved 35 participants who evaluated anonymized and randomly ordered results in terms of physical plausibility and visual naturalness. Participation was voluntary, and only aggregate preference rates and five-point ratings are reported. No personally identifiable information is collected or reported. The generated images are intended solely for research on realistic image compositing and harmonization.

REPRODUCIBILITY STATEMENT

All experiments are conducted on the publicly available DESOBAv2 benchmark using its standard train/test split. The model formulation, physical shadow projection, training objective, datasets, evaluation metrics, and experimental protocol are described in the main paper. We also report the optimization settings and computing infrastructure used in our experiments. All pretrained models and existing methods used in our framework are appropriately cited. We will release the source code and implementation details upon publication to facilitate reproduction of the reported results.

REFERENCES

Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1670–1687, 2014.

Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)*, 33(4):1–12, 2014.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Chris Careaga, S. Mahdi H. Miangoleh, and Yağız Aksoy. Intrinsic harmonization for illumination-aware compositing. In *Proc. SIGGRAPH Asia*, 2023.

Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8394–8403, 2020.

Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10786–10796, 2021.

Gaël Guennebaud, Loïc Barthe, and Mathias Paulin. Real-time soft shadow mapping by backprojection. *Rendering techniques*, 2006:227–234, 2006.

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

Yan Hong, Li Niu, and Jianfu Zhang. Shadow generation for composite image in real-world scenes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 914–922, 2022.

Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. Mask-shadowgan: Learning to remove shadows from unpaired data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2472–2481, 2019.

Yeying Jin, Aashish Sharma, and Robby T Tan. Dc-shadownet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5027–5036, 2021.

Jing Li and Jing Zhang. Vsdiffusion: Taming ill-posed shadow generation via visibility-constrained diffusion. *arXiv preprint arXiv:2603.08020*, 2026.

Daquan Liu, Chengjiang Long, Hongpan Zhang, Hanning Yu, Xinzhi Dong, and Chunxia Xiao. Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8139–8148, 2020.

Qingyang Liu, Jianting Wang, and Li Niu. Desobav2: Towards large-scale real-world dataset for shadow generation. *arXiv preprint arXiv:2308.09972*, 2023.

Qingyang Liu, Junqi You, Jianting Wang, Xinhao Tao, Bo Zhang, and Li Niu. Shadow generation for composite image using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8121–8130, 2024.

Li Niu, Wenyan Cong, Liu Liu, Yan Hong, Bo Zhang, Jing Liang, and Liqing Zhang. Making images real again: A comprehensive survey on deep image composition. *arXiv preprint arXiv:2106.14490*, 2021.

Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 577–582. 2023.

Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2002.

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*, pp. 10684–10695, 2022.

Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18310–18319, 2023.

Xinhao Tao, Junyan Cao, Yan Hong, and Li Niu. Shadow generation with decomposed mask prediction and attentive shadow filling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5198–5206, 2024.

Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3789–3797, 2017.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *Advances in Neural Information Processing Systems*, 38:35928–35959, 2026.

Ziqi Yu, Jing Zhou, Zhongyun Bao, Gang Fu, Weilei He, Chao Liang, and Chunxia Xiao. Cfdiffusion: Controllable foreground relighting in image compositing via diffusion model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3647–3656, 2024.

Fangneng Zhan, Changgong Zhang, Yingchen Yu, Yuan Chang, Shijian Lu, Feiying Ma, and Xuansong Xie. Emlight: Lighting estimation via spherical distribution approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

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*, pp. 3836–3847, 2023.

Shuyang Zhang, Runze Liang, and Miao Wang. Shadowgan: Shadow synthesis for virtual objects with conditional adversarial networks. *Computational Visual Media*, 5(1):105–115, 2019.

Haonan Zhao, Qingyang Liu, Xinhao Tao, Li Niu, and Guangtao Zhai. Shadow generation using diffusion model with geometry prior. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7603–7612, 2025.

Jing Zhou, Ziqi Yu, Zhongyun Bao, Gang Fu, Weilei He, Chao Liang, and Chunxia Xiao. Foreground harmonization and shadow generation for composite image. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8267–8276, 2024.

A Failure Cases and Practical Limitations

Although PAD performs robustly in most scenes, several failure modes remain, as illustrated in Fig. 4.

Ambiguous Contact. Inaccurate object–receiver contact or non-planar surfaces can lead to errors in the estimated shadow support and contact location.

Complex Illumination. Scenes with multiple strong lights or spatially varying illumination may violate the dominant-direction assumption, resulting in inaccurate shadow layout or intensity.

Reflection and Transparency. Reflective, transparent, or highly specular surfaces may introduce appearance and shadow effects that are not captured by the current physical prior.

Intermediate Estimation Errors. Errors in geometry or illumination estimation can propagate to the coarse shadow prior and final output. Diffusion refinement can tolerate moderate errors, but severe estimation failures may still cause incorrect shadow structure.

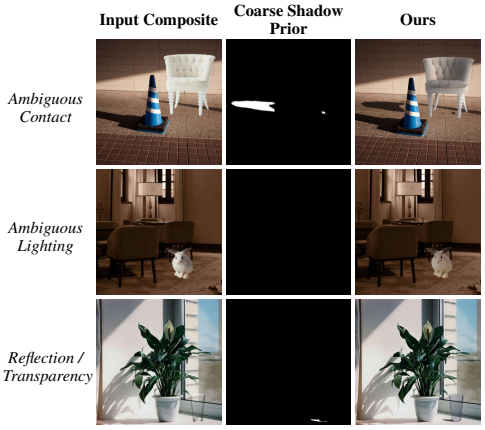

Figure 4: Representative failure cases of PAD under ambiguous contact geometry, complex illumination, and reflective or transparent surfaces.

B Implementation Details

B.1 Physical Shadow Prior Construction

For all experiments, input images are resized such that the maximum spatial dimension does not exceed 1024 pixels. We extract the largest external foreground contour and uniformly sample 240 contour points. The foreground mask is dilated using an 81×81 kernel, and background points in the resulting ring are divided into lower and upper/side regions relative to the foreground bottom, forming floor-like and wall-like receiver hypotheses. A local plane is fitted to each region using SVD when at least 80 valid 3D points are available.

Contour points are projected onto each receiver plane along the estimated shadow direction. Intersections with $t \leq 0$ are discarded, and configurations satisfying $|\mathbf{n}^\top \mathbf{d}| < 10^{-6}$ are treated as near-parallel and ignored.

To compensate for residual geometric misalignment, we align the bottom anchor of the projected contour with that of the foreground using an 8-pixel bottom band, with translation limited to 35% of the larger image dimension. A 2D affine transformation is then estimated using RANSAC with an 8-pixel reprojection threshold, 2000 maximum iterations, confidence 0.99, and 10 refinement iterations. The transformation is applied to the original foreground mask, followed by a Gaussian edge blur with kernel size 5.

The floor-like and wall-like shadow candidates are fused using a pixel-wise maximum. The deformation strength is fixed to 1.0, and all experiments use a random seed of 0.

B.2 Geometry and Projection Details

Pixel-wise 2D–3D Correspondence. MoGe-2 produces a dense point map $P \in \mathbb{R}^{H \times W \times 3}$ with pixel-wise correspondence to the input image. Therefore, a foreground contour location (u_i, v_i) is lifted to 3D by directly retrieving its corresponding point $P(v_i, u_i)$. Non-finite points are excluded from subsequent geometric operations.

3D-to-2D Projection. After ray–plane intersection, the projected contour points lie in the same 3D coordinate system as the MoGe point map. For each scene, we fit a lightweight affine 3D-to-2D projector from the dense point map and use it to map the projected 3D contour back to image

coordinates. The resulting 2D contour is used only as a geometric deformation guide; the final coarse shadow prior is obtained by warping the original foreground mask rather than directly rasterizing the projected contour.

Normal Representation for Reshading. For foreground reshading, we construct the normal representation using MoGe-based foreground geometry cues together with Omnidata background normal estimates. This representation is used with the estimated spherical-harmonic illumination to reconstruct foreground shading under the target-scene illumination.