VSDiffusion: Taming Ill-Posed Shadow Generation via Visibility-Constrained Diffusion

Jing Li[1] and Jing Zhang[1⋆]

East China University of Science and Technology, Shanghai, China
lij1@mail.ecust.edu.cn, jingzhang@ecust.edu.cn

Abstract. Generating realistic cast shadows for inserted foreground objects is a crucial yet challenging problem in image composition, where maintaining geometric consistency between objects and their shadows in complex scenes remains difficult due to the ill-posed nature of shadow formation. To address this challenge, we propose VSDiffusion, a visibility-constrained two-stage framework that narrows the solution space by incorporating visibility priors. In Stage I, a coarse shadow mask is predicted to localize plausible shadow regions. In Stage II, conditional diffusion guided by lighting and depth cues estimated from the composite image is used to generate accurate shadows. Within VSDiffusion, visibility priors are injected through two complementary pathways: (1) a visibility control branch with shadow-gated cross-attention that provides multi-scale structural guidance, and (2) a learned soft prior map that reweights the training loss in error-prone regions to encourage geometric correction. In addition, we introduce a high-frequency guided enhancement module to sharpen shadow boundaries and improve texture interaction with the background. Extensive experiments on the widely used DESOBAv2 benchmark demonstrate that VSDiffusion produces geometrically consistent shadows in complex scenes and establishes new state-of-the-art results across most evaluation metrics. The code is available at https://github.com/Jadelingli/VSDiffusion.

Keywords: Shadow Generation · Image Composition · VSDiffusion · Visibility-Constrained

1 Introduction

"The beginnings and ends of shadow lie between light and darkness."–da Vinci.

In recent years, diffusion models [14, 32] and their conditional control techniques [46] have greatly advanced image synthesis, enabling high-resolution, realistic, and controllable generation. However, these methods do not explicitly enforce physical consistency when compositing foreground objects into specific backgrounds, which is crucial in real-world applications such as film production and e-commerce design. In particular, if the foreground shadow is missing, has

⋆ Corresponding author.

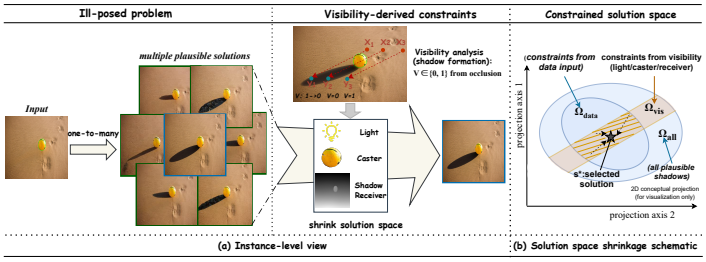

Fig. 1: (a) Shadow generation is an ill-posed problem with multiple plausible solutions for a single input. Visibility-aware geometric constraints are introduced to narrow the solution space. (b) Progressive solution space narrowing. The optimal solution S^* lies in the intersection of Ω_{vis}, Ω_{data}, and Ω_{all}, achieving a balance between data fidelity and geometric plausibility.

an incorrect direction, or exhibits an implausible shape, the composite can appear unnatural even when the object itself is well aligned. Therefore, generating shadows that are geometrically correct and perceptually realistic is essential for improving composited image quality.

Early methods formulate shadow generation as a physics-driven light transport problem, relying on geometric reconstruction and illumination estimation to render physically consistent shadows [7,8,38]. In contrast, recent learning-based approaches shift toward data-driven synthesis, where deep models learn implicit shadow appearance priors directly from training data rather than explicitly modeling physical processes [15,25,48]. However, these methods are typically trained on composited images with weak shadow supervision (*e.g.*, binary masks), without explicitly modeling physical factors such as illumination distribution and scene geometry. This under-constrained formulation leads to an ill-posed synthesis problem, where multiple plausible shadow appearances may correspond to the same input. As a result, models are prone to learning superficial texture correlations rather than physically meaningful shadow formation, limiting their ability to generate geometrically consistent shadows.

From a mathematical perspective, shadow generation is inherently ill-posed [18], as multiple plausible shadows can correspond to the same scene, and small changes in scene configuration can lead to large variations in shadow appearance. From a physical standpoint, a shadow is an interpretable phenomenon caused by occlusion: when light rays are blocked, visibility changes [5,9]. Furthermore, visibility editing techniques in computer graphics [28] suggest that shadow geometry can be precisely controlled by explicitly modeling visibility, independent of illumination and material properties.

Motivated by the above insight, we try to revisit shadow generation from a visibility-centric perspective by reducing the ill-posedness of the problem through visibility-aware constraints. Instead of relying on full physical simulation, we exploit available physical cues to approximately infer scene visibility changes. These cues serve as informative constraints that can effectively shrink the fea-

sible solution space. From this perspective, we prioritize geometric plausibility, especially layout consistency and lighting direction, before refining appearance details, enabling a more stable and robust shadow generation framework.

Based on the above analysis, we formalize the shadow generation problem as illustrated in Fig. 1. Because the input (a composite image) lacks precise physical quantities, shadow generation is a typical one-to-many mapping with a large solution space, which is a classical ill-posed problem. We then analyze shadow formation through visibility: a shadow arises when the visibility between a light point x and a receiver point y is blocked by intermediate geometry [21]. If the path is unobstructed, $V(x, y) = 1$ and y is not in shadow; if it is fully blocked, $V(x, y) = 0$ and y lies in shadow. The transition of V from 1 to 0 defines the shadow boundary. It follows that shadow geometry is mainly determined by the spatial relationship among the light source, the caster, and the shadow receiver [13, 41].

To fully exploit the above visibility factors for narrowing the solution space, we propose visibility-constrained shadow diffusion model (VSDiffusion), in which visibility-derived priors are injected into the diffusion process in two complementary ways. First, we extract prior features with a Visibility Control Branch (VCB) and design a Shadow-Gated Cross Attention (SGCA) module, which injects conditional features into denoising in a multi-scale and sparsely localized manner to provide controllable structural guidance. Second, we adopt a lightweight U-Net [33] to predict a visibility-guided spatial weight map, which reweights the training loss to emphasize error-prone shadow regions and improve geometric alignment. Moreover, to address blurry shadow contours and weak texture interactions, we introduce the High-Frequency Guided Enhancement (HFGE) module to refine boundary transitions and improve texture fusion with the background, further enhancing perceptual realism. The main contributions of this paper are listed as follows:

- We formulate shadow generation as an ill-posed problem and propose a two-stage framework that shifts shadow generation from a purely data-driven paradigm toward a visibility-prior-guided formulation, effectively narrowing the feasible solution space and enhancing geometric consistency.
- We introduce two complementary prior injection mechanisms, including structured guidance via the SGCA module during the denoising process and a spatially weighted optimization that emphasizes geometrically sensitive areas, thereby improving physical plausibility.
- We design the HFGE module to enhance high-frequency detail modeling for improved shadow boundary sharpness and more realistic interactions between shadows and background textures.

2 Related Work

2.1 Image Composition

Image composition [26] is the task of constructing a coherent scene representation by jointly modeling foreground-background interaction, illumination effects, and

structural compatibility. Early approaches treated composition as a signal processing problem, focusing primarily on suppressing low-level discontinuities and boundary artifacts. The Porter–Duff operator [30] introduced an alpha-based linear blending formulation to standardize image overlay, while Poisson image editing [29] further enforced seamless transitions through gradient-domain optimization. These methods mainly rely on local smoothing and lack the ability to model high-level semantic structure and illumination effects.

With the advancement of deep learning, image composition has evolved from pixel-level optimization to data-driven semantic consistency modeling. In particular, encoder–decoder architectures such as U-Net [33] and conditional generative adversarial networks [17] have been widely employed to improve structural coherence and textural realism. This line of research has expanded into learning-based blending [43, 45], object placement evaluation [23, 27], and image harmonization [2, 3, 6]. Despite improved global statistics and color consistency, these methods may still produce semantic inconsistencies and structural artifacts around occlusion boundaries, largely due to the difficulty of capturing 3D spatial interactions using 2D representations.

Recently, diffusion-based generative models [4, 20, 34] have redefined image composition by replacing direct pixel editing with iterative content regeneration under strong generative priors. Composition is therefore formulated as a conditional denoising process guided by background context and initial layout constraints. Although these methods improve perceptual realism, their learned priors tend to favor plausible appearance synthesis, making it challenging to simultaneously guarantee physical consistency and fine-grained structural fidelity in complex scenes.

2.2 Shadow Generation

Shadow generation aims to synthesize geometrically and perceptually consistent cast shadows conditioned on scene layout and illumination. Existing shadow generation methods can mainly be divided into two categories: rendering-based methods and non-rendering methods.

Rendering-based methods synthesize shadows using explicit geometric or physically interpretable representations, often requiring reliable structural or lighting information. Classical techniques such as shadow mapping [24, 40] and its variants [7, 8, 38] enable efficient real-time rendering but may suffer from discretization artifacts. Recent studies incorporate neural rendering strategies to improve controllability and visual realism. For example, SSN [36] generates soft shadows from 2D masks by modeling ambient illumination effects. Methods based on Pixel Height Maps [35] and PixHt-Lab [37] further introduce per-pixel structural representations to approximate 3D projection geometry.

In contrast, non-rendering methods directly learn shadow synthesis from data, leveraging deep learning to capture implicit shadow appearance priors for realistic generation. Early GAN-based models, including ShadowGAN [47] and ARShadowGAN [22], employ conditional adversarial learning to capture shadow-context relationships. Some shadow removal frameworks also provide

complementary cues for shadow synthesis, such as Mask-ShadowGAN [16]. Two-stage architectures, such as SGRNet [15], separate shadow region localization from shading refinement to improve robustness. Similarly, DMASNet [39] decomposes mask prediction into coarse localization and shape refinement to enhance boundary quality in complex scenes. Recent advances in diffusion-based shadow generation have introduced diverse strategies. SGDiffusion [25] mainly adapts ControlNet for shadow generation and improves shadow appearance through intensity modulation. GPSDiffusion [48] relies on rotated bounding box prediction and discrete shape prototype retrieval to constrain shadow shapes, instead of localizing them based on physical logic.

Despite these developments, data-driven shadow synthesis methods relying on implicit physical modeling may still struggle to guarantee physical plausibility and structural fidelity in cluttered or complex environments.

In summary. General image composition methods mainly enforce appearance consistency but seldom explicitly model cast-shadow geometry, which restricts their ability to model physically plausible illumination effects. Meanwhile, dedicated shadow generation methods face a fundamental trade-off: rendering-based approaches rely on accurate geometric or lighting inputs and thus may suffer from limited generalization, whereas data-driven approaches often learn implicit shadow priors without explicit physical constraints. Motivated by the above analysis, we propose to incorporate visibility-aware physical cues into a diffusion-based framework to reduce the ill-posed nature of shadow generation while maintaining high-quality generative capability.

3 Method

We first present an overview of the proposed VSDiffusion framework, and then describe each major module in detail.

3.1 Overview

Our method aims to generate physically consistent shadows for foreground objects in composite images. Given a composite image I_c where the inserted foreground object lacks its corresponding shadow, along with a foreground mask M_{fo} and a background shadow mask M_{bs}, the goal is to infer a plausible shadow. This problem is ill-posed [18], leading to non-unique and ambiguous solutions. To mitigate this ambiguity, we propose a two-stage framework that progressively constrains the solution space as shown in Fig. 2.

In the first stage, we predict a coarse foreground shadow mask $M_{fs}^{(1)}$, which serves as a spatial prior and reduces geometric uncertainty. Specifically, the background encoder E_b and foreground encoder E_f process (I_c, M_{bs}) and (I_c, M_{fo}), respectively. Their features interact via cross-attention integration [15] to produce $M_{fs}^{(1)}$, which is then combined with M_{fo} to form M_{comb}. In the second

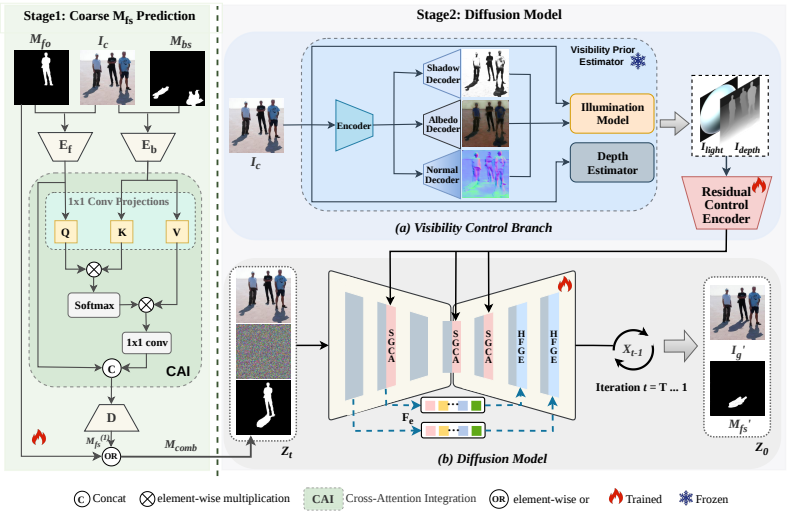

Fig. 2: Framework of VSDiffusion. It consists of two stages. Stage 1 predicts a coarse foreground shadow mask $M_{fs}^{(1)}$. Stage 2 includes two sub-modules: (a) Visibility Control Branch, which utilizes visibility prior estimator to extract I_{light} and I_{depth} from I_{c}, subsequently encoded by RCE to provide structural guidance. (b) Diffusion Model. The U-Net progressively denoises Z_t into Z_0.

stage, it includes two sub-modules: Visibility Control Branch (VCB) and diffusion model. We incorporate visibility priors into the diffusion process to further regularize the shadow generation and alleviate the inherent ill-posedness. These priors guide the denoising U-Net. At each timestep, the U-Net denoises $(I_c, M_{\text{comb}}, x_t)$ and outputs the final shadowed image I'_g.

In this work, these visibility priors are injected through two complementary pathways: (i) As structurally controllable conditional guidance, VCB injects priors via Shadow-Gated Cross Attention (SGCA), which sparsely injects guidance at three U-Net scales to align geometry, preserve global layout, and refine boundaries. (ii) As a training-level spatial supervision signal, we introduce S_{prior}-Weighted Loss (SWL), which uses a soft prior map to spatially reweight pixel-level supervision to focus on error-prone regions. These two components are complementary, providing both architectural constraints and adaptive optimization guidance. In addition, High-Frequency Guided Enhancement (HFGE) extracts high-frequency cues from shallow encoder layers and residually injects them during high-resolution decoding to enhance shadow boundaries and texture details. Next, we describe VCB, SGCA, HFGE and SWL in detail.

3.2 Visibility Control Branch

Predicting shadows from a single input image I_{c} remains highly ambiguous without explicit constraints on scene structure and illumination, often yielding mul-

tiple visually plausible yet non-unique solutions. Therefore, we propose the Visibility Control Branch (VCB) to guide the denoising network by incorporating visibility-derived constraints, thereby narrowing the solution space. As shown in Fig. 2 (a), the VCB comprises two components: visibility prior estimator [31,44] and Residual Control Encoder [19].

The visibility prior estimator is designed to extract visibility-related priors (light and depth) from I_{c}. For illumination estimation, it employs a shared encoder followed by three lightweight decoders to predict shadow, albedo, and normal maps. These estimated attributes, together with I_{c}, are then fed into an illumination model based on Lambertian reflectance [1,11]. This model recovers the second-order spherical harmonics coefficients representing global illumination, which are visualized as an illumination map I_{light} to depict light direction and intensity. Since the illumination I can be derived analytically from the input image and the estimated albedo α, shadow weight s, and surface normal n, no separate network branch is required for lighting prediction. For a pixel (x, y), the image formation process is formulated as:

$$i(x,y) = \alpha(x,y) \odot s(x,y)\, B\big(n(x,y)\big)\, I, \tag{1}$$

where $i(x, y)$ is the pixel value in I_{c}, $\odot$ denotes element-wise multiplication, $B(n)$ denotes the second-order Spherical Harmonics (SH) basis functions evaluated at the surface normal n, and I denotes the unknown second-order SH coefficients, *i.e.* I_{light}. By formulating this as a least-squares problem, I can be solved via inversion. Overall, I_{light} is obtained following an inverse-rendering framework [44]. Additionally, the visibility prior estimator utilizes a pre-trained MiDaS model [31] as a monocular depth estimator to produce a depth map I_{depth} from I_{c}.

To stably integrate these visibility priors, we employ a Residual Control Encoder [19] for feature extraction. This encoder adopts a lightweight residual design (7 blocks) with zero convolutions, in contrast to the encoder in ControlNet [46], which may ignore control signals during early training and cause instability. This design ensures that the diffusion process gradually incorporates I_{light} and I_{depth}, improving geometric and illumination alignment.

3.3 Geometry-Detail Denoiser

Shadow-Gated Cross Attention Module. To effectively exploit visibility priors, we propose the Shadow-Gated Cross-Attention (SGCA) module. Unlike conventional dense conditioning across many layers [46], SGCA injects these priors ($I_{\mathrm{light}}, I_{\mathrm{depth}}$) into three strategically selected U-Net anchors (early, mid, and late) for geometric alignment, global layout, and refinement. Dense injection may restrict the generation space, compromise consistency, and cause over-conditioning. By injecting guidance at these anchors, SGCA maintains controllability while avoiding over-conditioning and redundant computation, offering a better trade-off between conditioning strength and generative fidelity. Here, sparse refers to the sparsity of injection locations rather than token-level attention sparsity. As shown in Fig. 3, let $X^{(s)} \in \mathbb{R}^{B\times C_s\times H_s\times W_s}$ denote the U-Net

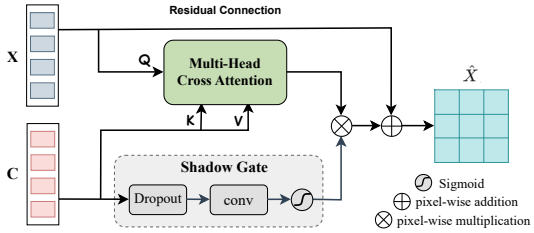

Fig. 3: Architecture of the Shadow-Gated Cross-Attention (SGCA) module.

feature at scale s, and $C^{(s)}$ represent the aligned conditional feature from the Residual Control Encoder. SGCA computes cross-attention where the queries Q are projected from $X^{(s)}$, and keys K and values V are derived from $C^{(s)}$. To adaptively regulate the influence of external priors, a Shadow Gate $G^{(s)}$ is predicted from $C^{(s)}$ by a lightweight head $\psi(\cdot)$:

$$G^{(s)} = \sigma\Big(\psi\Big(C^{(s)}\Big)\Big), \tag{2}$$

$$\hat{X}^{(s)} = X^{(s)} + G^{(s)} \odot \mathrm{unflat}\Big(A^{(s)}\Big), \tag{3}$$

where σ is the sigmoid function, $\mathrm{unflat}(\cdot)$ restores the attention tokens to spatial dimensions, and $A^{(s)} = MHA(Q, K, V)$. The gate strengthens the injection when conditional features are beneficial for shadow inference and suppresses them otherwise, preventing texture degradation and artifact amplification.

High-Frequency Guided Enhancement. Shadow generation often suffers from blurry boundaries, jagged artifacts, and over-smoothed background textures. To address these issues, we introduce High-Frequency Guided Enhancement (HFGE). HFGE extracts robust high-frequency cues from shallow U-Net encoder features and injects them as residual guidance into the late decoder stages, ensuring sharp shadow edges without compromising background fidelity.

We extract high-frequency information from shallow encoder features F_e based on two key insights. First, shallow features contain mixed-frequency signals, including local gradients and texture responses, which are suitable for describing edges and fine details. Second, in diffusion models, high-frequency details are typically refined in later steps. As observed by Falck *et al.* [10], high-frequency components enter the low Signal-to-Noise Ratio (SNR) regime earlier and faster during the DDPM forward process. Low-frequency structures tend to appear first, while high-frequency details emerge gradually at later stages. Thus, injecting guidance at late high-resolution decoding stages aligns with this natural refinement order. Conversely, injecting too early, *e.g.*, in the encoder or bottleneck, risks feature dilution through downsampling or introducing noise that degrades global shadow formation.

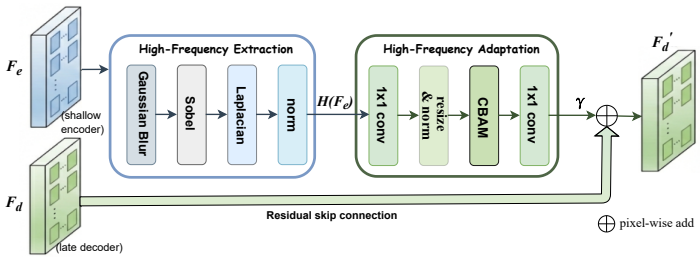

Fig. 4: Architecture of the High-Frequency Guided Enhancement (HFGE) module.

As shown in Fig. 4, the HFGE module consists of two main stages: High-Frequency Extraction (blue region) and High-Frequency Adaptation (green region). For High-Frequency Extraction, we extract robust high-frequency cues from shallow encoder features F_e through four sequential steps. First, we apply spatial Gaussian smoothing to F_e to filter out local pseudo-textures, producing $\tilde{F}_e$. Second, we use Sobel filters to compute gradients g_x and g_y, obtaining the gradient magnitude G to emphasize structural boundaries:

$$G = \sqrt{g_x^2 + g_y^2 + \epsilon}. \tag{4}$$

Third, to further capture fine-grained structures like thin edges, we incorporate the response L from the Laplacian operator. Finally, we normalize the magnitude over spatial dimensions to obtain the high-frequency feature $H(F_e)$:

$$H(F_e) = \mathrm{Norm}(G + \alpha L), \tag{5}$$

where α is a weighting factor and $\mathrm{Norm}(\cdot)$ denotes spatial magnitude normalization. For High-Frequency Adaptation, since $H(F_e)$ and decoder features F_d reside in different feature spaces, we introduce an adaptation branch to align the signals. Specifically, we employ a 1×1 convolution and bilinear interpolation to match the channel dimensions and spatial resolution of F_d. Then, a lightweight CBAM [42] is applied to adaptively reweight the responses, focusing on critical shadow contours. The calibrated guidance is then residually added to F_d:

$$F_d' = F_d + \gamma \cdot \mathrm{CBAM}\big(\phi(H(F_e))\big), \tag{6}$$

where $\phi(\cdot)$ denotes projection and scale alignment, and γ controls the injection strength. This design allows high-frequency details to refine boundaries while preserving the underlying background texture.

3.4 Training Losses

We adopt a three-step training strategy: (i) pretraining Stage I alone to learn basic geometric localization; (ii) training Stage II for shadow synthesis with Stage I frozen; and (iii) jointly fine-tune the full framework at a lower learning rate. We optimize the model with the following loss functions.

First Stage Loss. The coarse foreground shadow mask $M_{\text{fs}}^{(1)}$ is supervised by a combination of Binary Cross-Entropy (BCE) and Dice loss against the ground truth M_{fs}:

$$\mathcal{L}_{s1} = \mathcal{L}_{\text{bce}}\big(M_{fs}^{(1)}, M_{fs}\big) + \mathcal{L}_{\text{dice}}\big(M_{fs}^{(1)}, M_{fs}\big). \tag{7}$$

Second Stage Loss. Stage II focuses on high-fidelity shadow reconstruction and mask refinement via two terms.
1) Base Loss: To ensure pixel-level consistency for the generated image $\tilde{I}$ and refined mask $\tilde{M}_{fs}$, we define the base loss as:

$$\mathcal{L}_{\text{base}} = \frac{1}{N}\sum_{x,y}\left(\left\|\tilde{I}(x,y) - I(x,y)\right\|_1 + \lambda_1 \left\|\tilde{M}_{fs}(x,y) - M_{fs}(x,y)\right\|_2^2\right), \tag{8}$$

where I is the ground truth shadowed image, and N is the pixel count.
2) S_{prior}-Weighted Loss (SWL): Shadow errors tend to focus on small but critical regions, such as thin shadow edges or misaligned boundaries. Standard global losses often dilute gradient in these critical areas. Thus, we introduce a soft prior map S_{prior} to spatially re-weight the optimization process. This map is generated by a lightweight U-Net network G_{p}, which aggregates visibility-related cues and maps the output to [0,1] via a sigmoid function:

$$S_{\text{prior}} = \sigma\Big(G_p\Big(I_{\text{light}}, I_{\text{depth}}, M_{fo}, M_{fs}^{(1)}\Big)\Big), \quad S_{\text{prior}} \in [0,1], \tag{9}$$

where G_{p} is the lightweight S_{prior} predictor, $\sigma(\cdot)$ is the Sigmoid function, I_{light} and I_{depth} provide visibility-related information, and M_{fo} is the foreground object mask. Moreover, G_{p} is jointly optimized with the second stage in an end-to-end manner, without additional explicit supervision, allowing it to adaptively identify challenging areas during the reconstruction process. However, the model might minimize $\mathcal{L}_{s2}$ by simply driving S_{prior} toward a near-zero map. To avoid this gradient collapse, we apply mean normalization to enforce budget-preserving gradient reallocation:

$$\hat{S}_{\text{prior}}(x,y) = \frac{S_{\text{prior}}(x,y)}{\frac{1}{N}\sum_{i,j} S_{\text{prior}}(i,j) + \epsilon}. \tag{10}$$

We retain the global base supervision and add the spatially weighted correction term within S_{prior}:

$$\mathcal{L}_{swl} = \frac{1}{N}\sum_{x,y}\Big(\hat{S}_{\text{prior}}(x,y)\cdot \ell_{\text{base}}(x,y)\Big). \tag{11}$$

SWL allocates more gradient budget to the error-prone shadow regions, thus improving boundary alignment and reducing shadow leakage. Thus, the entire loss for the second stage is:

$$\mathcal{L}_{s2} = \mathcal{L}_{base} + \lambda_2 \mathcal{L}_{swl}. \tag{12}$$

Joint Loss. Finally, we jointly fine-tune both stages using the following loss function, where λ_1, λ_2, and γ are hyperparameters used to balance the different loss terms.

$$\mathcal{L}_{joint} = \mathcal{L}_{s2} + \gamma \mathcal{L}_{s1}. \tag{13}$$

4 Experiments

To systematically assess the performance of VSDiffusion in shadow generation, we perform extensive quantitative comparisons, qualitative evaluations, and ablation analyses.

4.1 Dataset and Implementation Details

We conduct experiments on DESOBAv2, using the train/test split from [25], which contains 27,823 training images and 750 test images. The test set includes 500 BOS samples (with background object–shadow pairs) and 250 BOS-free samples. We evaluate both image quality and mask quality. For image quality, we report RMSE and SSIM over the full image, denoted as global RMSE (GR) and global SSIM (GS), respectively. We also compute local RMSE (LR) and local SSIM (LS) within the ground-truth foreground shadow region to better reflect shadow fidelity. For mask quality, we measure the balanced error rate (BER) between the ground-truth binary foreground shadow mask and the predicted mask binarized with a threshold of 0.5. We report global BER (GB) on the full image and local BER (LB) within the ground-truth foreground shadow region.

The VSDiffusion is implemented in PyTorch and trained on two NVIDIA RTX 3090 GPUs with batch size 18 per GPU. We use Adam (β_1=0.9, β_2=0.999) for 700 epochs with an initial learning rate of 3×10^{-5}. We apply Kaiming initialization and EMA with decay 0.9999. All images are resized to 256×256 for both training and testing. Following [12], we adopt an image-space conditional diffusion architecture to better preserve background textures for image-to-image shadow generation. Loss weights are set to λ_1=0.2, λ_2=0.5, and γ=0.1.

4.2 Comparative experiments

In this section, we mainly compared with representative non-rendering shadow generation methods, including ShadowGAN [47], Mask-SG [16], AR-SG [22], SGRNet [15], DMASNet [39], SGDiffusion [25], and GPSDiffusion [48]. These baselines covered both one-stage and two-stage pipelines, as well as GAN-based and diffusion-based generators.

Quantitative comparison Tab. 1 summarizes the quantitative results on DESOBAv2. Under both BOS and BOS-free settings, our method achieves the best or consistently competitive performance on the local metrics and BER, which are more sensitive to shadow geometry and boundary alignment. **BOS setting.**

Table 1: Quantitative comparisons on DESOBAv2. We evaluate shadow generation with RMSE, SSIM, and BER. BOS denotes test images with background shadow references. ↑/↓ indicate higher/lower is better. Best results are highlighted in bold.

Method	BOS Test Images						BOS-free Test Images					
	GR↓	LR↓	GS↑	LS↑	GB↓	LB↓	GR↓	LR↓	GS↑	LS↑	GB↓	LB↓
ShadowGAN [47]	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 [16]	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 [22]	7.335	58.037	0.961	0.241	0.383	0.682	16.067	63.713	0.908	0.104	0.349	0.682
SGRNet [15]	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 [39]	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 [25]	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 [48]	**5.896**	46.713	0.966	0.374	0.213	0.423	**13.809**	55.616	**0.917**	0.166	0.197	0.384
Ours	6.422	**46.542**	0.964	**0.377**	**0.182**	**0.360**	14.629	**55.377**	0.912	**0.184**	**0.194**	**0.373**

Compared with the current state-of-the-art method GPSDiffusion [48], our approach reduces global BER (GB) and local BER (LB) by approximately 0.03 and 0.06, respectively, while achieving slight improvements in local RMSE (LR) and local SSIM (LS). **BOS-free setting.** When background object–shadow references are removed, most methods exhibit noticeable performance degradation, indicating increased ambiguity in the solution space. Despite this challenge, our method maintains stable performance, improving LS by about 0.02 and further reducing GB and LB by approximately 0.003 and 0.01, compared with the SOTA method GPSDiffusion. These results demonstrate that the proposed visibility-aware constraints and prior-guided supervision generalize effectively, enabling plausible shadow geometry and well-aligned boundaries even in the absence of explicit references.

Qualitative comparison We evaluate VSDiffusion against other methods under both BOS-free (rows 1-4) and BOS (rows 5-7) settings in Fig. 5. While existing diffusion-based models achieve basic darkening, they often struggle with shadow direction, shape fidelity, and boundary alignment. For instance, in row 2, SGDiffusion [25] and GPSDiffusion [48] show direction errors, where the generated shadows are inconsistent with the scene lighting direction. In row 5, AR-SG [22] produces distorted shadow shapes, and the contours do not match the object geometry. In addition, SGRNet [15] often produces blurred edges and halo-like transitions across multiple examples, which reduces local boundary accuracy. In contrast, VSDiffusion generates shadows with more consistent geometry and sharper boundaries across diverse scenes, and the advantage is more evident under BOS-free. This work introduces visibility priors to constrain shadow direction and contact relationships, which narrows the feasible solution space even without reference cues. It further applies the S_{prior}-Weighted Loss (SWL) to emphasize boundary-critical regions and uses High-Frequency Guided Enhancement to enhance high-frequency details, leading to more accurate contours and boundaries.

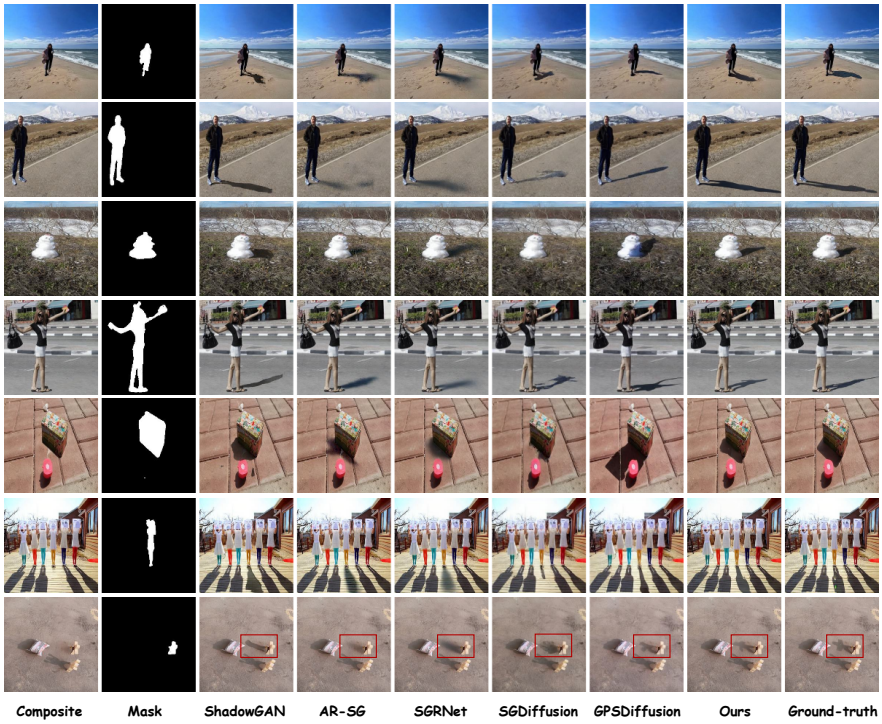

Fig. 5: Qualitative comparison on the DESOBAv2. From left to right: the composite image (without foreground shadows), the foreground object mask, results of other SOTA methods, VSDiffusion, and the ground truth.

Efficiency Analysis We further report the computational efficiency comparison in Tab. 2. All methods are evaluated at the same resolution (256×256) with a batch size of 1 on the same GPU. The inference time is averaged over all test images and includes all inference modules required by each method. As shown in Tab. 2, VSDiffusion achieves the lowest computational cost among the compared diffusion-based shadow generation methods. It is worth noting that the total parameters of VSDiffusion include not only the diffusion backbone,

Table 2: Computational efficiency comparison of diffusion-based shadow generation methods.

Method	Total Params	Trainable Params	Inference Time
SGDiffusion	1.458B	391.6M	8.0791 s/image
GPSDiffusion	1.479B	382.0M	6.9049 s/image
VSDiffusion	**547.182M**	**300.4M**	**1.6986 s/image**

but also the external prior estimation modules, and the coarse shadow prediction module. Even with these auxiliary modules counted, VSDiffusion still uses fewer total parameters and achieves significantly faster inference. This indicates that our VSDiffusion improves shadow generation without introducing excessive model complexity.

4.3 Ablation Study

As shown in Tab. 3, we evaluate the effectiveness of each component on the DESOBAv2 dataset. Overall, the results indicate that both visibility-aware priors and high-frequency refinement play essential roles in improving shadow geometry consistency and boundary alignment.
1) Visibility Priors Injection. The Visibility Control Branch (VCB) and SWL inject visibility guidance through structural conditioning and adaptive supervision, respectively. Specifically, VCB imposes architectural constraints to regularize shadow placement, thereby improving geometry-sensitive metrics such as BER and LR. SWL reshapes the training signal in a spatially adaptive manner. As visualized in Fig. 6 (col. 3), we show the effective supervision map $\hat{S}_{\text{prior}} \cdot \ell_{\text{base}}$. Here, red regions correspond to *high weight* $\times$ *high error*, indicating that SWL emphasizes supervision on difficult regions, typically around shadow boundaries

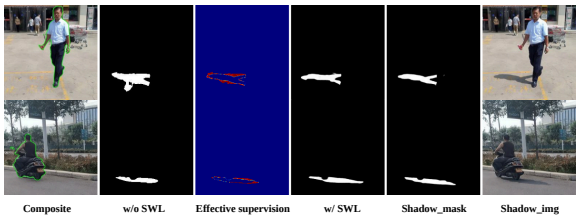

Fig. 6: Qualitative ablation of VSDiffusion. Visual comparisons demonstrate that our SWL effectively concentrates supervision on critical shadow boundaries and penumbras. In Effective supervision, the red regions *i.e.*, high weight $\times$ high error.

Table 3: Ablation study on DESOBAv2. VCB and SWL are two visibility-inspired ways to inject priors into the diffusion model. SWL is short for the S_{prior}-Weighted Loss. HFGE denotes the High-Frequency Guided Enhancement module. ↑/↓ indicate higher/lower is better, and best results are in bold.

VCB	HFGE	SWL	GR↓	LR↓	GS↑	LS↑	GB↓	LB↓
			6.655	50.598	0.962	0.364	0.221	0.438
✓			6.592	48.265	0.963	0.366	0.203	0.403
	✓		6.621	49.214	0.963	0.371	0.211	0.407
		✓	6.580	48.876	0.963	0.369	0.204	0.398
✓	✓		6.571	48.147	**0.964**	0.374	0.193	0.381
✓	✓	✓	**6.422**	**46.542**	**0.964**	**0.377**	**0.182**	**0.360**

Table 4: Detailed ablation study on visibility prior injection.

Method	GR↓	LR↓	GS↑	LS↑	GB↓	LB↓
Baseline	6.655	50.598	0.962	0.364	0.221	0.438
w/o gated attention	6.532	47.672	0.963	0.371	0.186	0.378
w/o sparse injection	6.481	48.087	**0.964**	0.374	0.190	0.383
Ours	**6.422**	**46.542**	**0.964**	**0.377**	**0.182**	**0.360**

and penumbras. Consequently, with SWL (col. 4), the refined shadow mask exhibits better boundary alignment with the ground-truth mask (col. 5) than the setting without SWL (col. 2). Quantitatively, adding SWL on top of VCB and HFGE further reduces LR by 1.605 and LB by 0.021 compared with the fifth row in Tab. 3. Moreover, the ablation results in Tab. 4 further validate the effectiveness of the proposed SGCA module, including both sparse injection and gated attention.

2) High-Frequency Refinement (HFGE): The HFGE module improves local structural fidelity by leveraging high-frequency encoder representations. By strengthening fine-grained geometric cues, HFGE alleviates texture-level distortions and improves LS by 0.008 when added to the VCB-only setting.

5 Conclusion

In this paper, we present VSDiffusion for physically consistent shadow generation. We formulate shadow generation as an inherently ill-posed problem and reduce its ambiguity by explicitly modeling the visibility formation process to progressively constrain the solution space. Our two-stage framework combines coarse shadow localization with diffusion refinement guided by visibility priors derived from light, caster, and shadow receiver modeling. Structurally controllable conditioning and adaptive spatial supervision serve as complementary mechanisms that enable effective prior integration, while a high-frequency enhancement module improves boundary sharpness and texture fidelity. Extensive experiments demonstrate stable improvements in geometric alignment and visual realism, particularly under reference-free settings. These results highlight the effectiveness of visibility-derived constraints in reducing ambiguity in shadow generation. Future work will explore extending this work to photorealistic subject-driven image editing with adaptive, reference-free shadow intensity calibration.

References

1. Basri, R., Jacobs, D.W.: Lambertian reflectance and linear subspaces. IEEE transactions on pattern analysis and machine intelligence **25**(2), 218–233 (2003)
2. Chen, H., Gu, Z., Li, Y., Lan, J., Meng, C., Wang, W., Li, H.: Hierarchical dynamic image harmonization. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 1422–1430 (2023)

3. Chen, J., Zhang, Y., Zou, Z., Chen, K., Shi, Z.: Dense pixel-to-pixel harmonization via continuous image representation. IEEE Transactions on Circuits and Systems for Video Technology **34**(5), 3876–3890 (2023)
4. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6593–6602 (2024)
5. Cohen, M.F., Wallace, J.R.: Radiosity and realistic image synthesis. Morgan Kaufmann (1993)
6. Cong, W., Zhang, J., Niu, L., Liu, L., Ling, Z., Li, W., Zhang, L.: Dovenet: Deep image harmonization via domain verification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8394–8403 (2020)
7. Donnelly, W., Lauritzen, A.: Variance shadow maps. In: Proceedings of the 2006 symposium on Interactive 3D graphics and games. pp. 161–165 (2006)
8. Dou, H., Yan, Y., Kerzner, E., Dai, Z., Wyman, C.: Adaptive depth bias for shadow maps. In: Proceedings of the 18th meeting of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games. pp. 97–102 (2014)
9. Dutré, P.: Global illumination compendium. Computer Graphics, Department of Computer Science, Katholieke Universiteit Leuven (2003)
10. Falck, F., Pandeva, T., Zahirnia, K., Lawrence, R., Turner, R., Meeds, E., Zazo, J., Karmalkar, S.: A fourier space perspective on diffusion models. arXiv preprint arXiv:2505.11278 (2025)
11. Forsyth, D.A., Ponce, J.: A modern approach. Computer vision: a modern approach **17**(21-48), 1 (2003)
12. Guo, L., Wang, C., Yang, W., Huang, S., Wang, Y., Pfister, H., Wen, B.: Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14049–14058 (2023)
13. Haines, E., Akenine-Möller, T.: Ray Tracing Gems: High-Quality and Real-Time Rendering with DXR and Other APIs. Apress (2019)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
15. Hong, Y., Niu, L., Zhang, J.: Shadow generation for composite image in real-world scenes. In: Proceedings of the AAAI conference on artificial intelligence. pp. 914–922 (2022)
16. Hu, X., Jiang, Y., Fu, C.W., Heng, P.A.: Mask-shadowgan: Learning to remove shadows from unpaired data. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2472–2481 (2019)
17. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
18. Kabanikhin, S.I.: Inverse and Ill-Posed Problems: Theory and Applications. De Gruyter, Berlin, Boston (2011). https://doi.org/10.1515/9783110224016
19. Kocsis, P., Philip, J., Sunkavalli, K., Nießner, M., Hold-Geoffroy, Y.: Lightit: Illumination modeling and control for diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9359–9369 (2024)
20. Kulal, S., Brooks, T., Aiken, A., Wu, J., Yang, J., Lu, J., Efros, A.A., Singh, K.K.: Putting people in their place: Affordance-aware human insertion into scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17089–17099 (2023)
21. Laine, S.: An incremental shaft subdivision algorithm for computing shadows and visibility. Master's thesis, Helsinki University of Technology (2006)

22. Liu, D., Long, C., Zhang, H., Yu, H., Dong, X., Xiao, C.: 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)
23. Liu, L., Liu, Z., Zhang, B., Li, J., Niu, L., Liu, Q., Zhang, L.: Opa: object placement assessment dataset. arXiv preprint arXiv:2107.01889 (2021)
24. Liu, N., Pang, M.Y.: Shadow mapping algorithms: A complete survey. In: 2009 International Symposium on Computer Network and Multimedia Technology. pp. 1–5. IEEE (2009)
25. Liu, Q., You, J., Wang, J., Tao, X., Zhang, B., Niu, L.: 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)
26. Niu, L., Cong, W., Liu, L., Hong, Y., Zhang, B., Liang, J., Zhang, L.: Making images real again: A comprehensive survey on deep image composition. arXiv preprint arXiv:2106.14490 (2021)
27. Niu, L., Liu, Q., Liu, Z., Li, J.: Fast object placement assessment. arXiv preprint arXiv:2205.14280 (2022)
28. Obert, J., Pellacini, F., Pattanaik, S.: Visibility editing for all-frequency shadow design. Computer Graphics Forum **29**(4), 1441–1449 (2010)
29. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. ACM Transactions on Graphics **22**(3), 313–318 (2003)
30. Porter, T., Duff, T.: Compositing digital images. In: Proceedings of the 11th annual conference on Computer graphics and interactive techniques. pp. 253–259 (1984)
31. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence **44**(3), 1623–1637 (2022)
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
33. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
34. Sarukkai, V., Li, L., Ma, A., Ré, C., Fatahalian, K.: Collage diffusion. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 4208–4217 (2024)
35. Sheng, Y., Liu, Y., Zhang, J., Yin, W., Oztireli, A.C., Zhang, H., Lin, Z., Shechtman, E., Benes, B.: Controllable shadow generation using pixel height maps. In: European Conference on Computer Vision. pp. 240–256. Springer (2022)
36. Sheng, Y., Zhang, J., Benes, B.: Ssn: Soft shadow network for image compositing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4380–4390 (2021)
37. Sheng, Y., Zhang, J., Philip, J., Hold-Geoffroy, Y., Sun, X., Zhang, H., Ling, L., Benes, B.: Pixht-lab: Pixel height based light effect generation for image compositing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16643–16653 (2023)
38. Stamminger, M., Drettakis, G.: Perspective shadow maps. In: Proceedings of the 29th annual conference on Computer graphics and interactive techniques. pp. 557–562 (2002)

39. Tao, X., Cao, J., Hong, Y., Niu, L.: Shadow generation with decomposed mask prediction and attentive shadow filling. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 5198–5206 (2024)
40. Williams, L.: Casting curved shadows on curved surfaces. In: Proceedings of the 5th annual conference on Computer graphics and interactive techniques. pp. 270–274 (1978)
41. Woo, A., Poulin, P., Fournier, A.: A survey of shadow algorithms. IEEE Computer Graphics and Applications **10**(6), 13–32 (2002)
42. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
43. Wu, H., Zheng, S., Zhang, J., Huang, K.: Gp-gan: Towards realistic high-resolution image blending. In: Proceedings of the 27th ACM international conference on multimedia. pp. 2487–2495 (2019)
44. Yu, Y., Smith, W.A.: Outdoor inverse rendering from a single image using multiview self-supervision. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(7), 3659–3675 (2021)
45. Zhang, J., Sunkavalli, K., Hold-Geoffroy, Y., Hadap, S., Eisenman, J., Lalonde, J.F.: All-weather deep outdoor lighting estimation. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 10158–10166 (2019)
46. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
47. Zhang, S., Liang, R., Wang, M.: Shadowgan: Shadow synthesis for virtual objects with conditional adversarial networks. Computational Visual Media **5**(1), 105–115 (2019)
48. Zhao, H., Liu, Q., Tao, X., Niu, L., Zhai, G.: Shadow generation using diffusion model with geometry prior. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7603–7612 (2025)