EMBEDDING PHYSICAL REASONING INTO DIFFUSION-BASED SHADOW GENERATION UNDER THE SUN AND SKY

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating realistic cast shadows for inserted foreground objects requires reasoning about scene geometry and illumination. However, most learning-based approaches treat shadow generation as an image translation problem and capture these physical relationships only implicitly. This often results in misaligned or implausible shadows. Motivated by the physics of shadow formation, we introduce explicit geometric guidance for outdoor shadow generation. Given a composite image and a foreground object mask, we recover approximate scene geometry and estimate a dominant light direction to derive a coarse shadow estimate via geometric reasoning. While coarse, this estimate provides a spatial anchor for shadow placement. Because illumination cannot always be uniquely inferred from a single image, we predict confidence scores for both lighting and shadow cues and use them to regulate their influence during generation. These cues (shadow mask, light direction, and their confidence scores) condition a diffusion-based generator that refines the estimate into a realistic shadow. Experiments on DESOBAV2 show substantially improved shadow-region fidelity and localization, with an overall 23% lower shadow-region RMSE and 30% lower shadow-mask BER than the prior state-of-the-art method.

1 INTRODUCTION

Shadow formation is a geometric process: given a 3D object and a light direction, light rays from the source are occluded by the object, casting a shadow whose shape and placement are physically determined on receiver surfaces. Yet modern shadow generation methods (Hu et al., 2019; Zhang et al., 2019b; Liu et al., 2020; Hong et al., 2022; Liu et al., 2024; Zhao et al., 2025) primarily operate in 2D pixel space, leaving the model to implicitly discover this physical relationship from data alone. Thus, the result is often inconsistent: shadows that drift, misalign, or hallucinate, especially with complex object geometry or ambiguous lighting.

Our approach instead is motivated by the physics of shadow formation: if approximate scene geometry and a light direction are known, the location and extent of a shadow can be inferred directly through geometric reasoning. Given a composite image and a mask of the foreground object whose cast shadow is absent, as illustrated in Figure 1, we recover approximate scene geometry in the form of a dense 3D point map using MoGe-2 (Wang et al., 2025) and estimate a dominant light direction, which together allow us to compute a physics-guided shadow estimate via geometric ray reasoning (Sato et al., 2003; Panagopoulos et al., 2013). Although coarse, this estimate provides an explicit spatial anchor for generation, encoding the object–receiver–illumination relationship rather than requiring the generator to learn it entirely from data.

However, lighting cannot always be uniquely inferred from an image (Lalonde et al., 2009; Panagopoulos et al., 2013). In many scenes, the visual evidence required to determine the illumination is weak or absent, for example, when no reference object–shadow pairs are visible, making the mapping from appearance to light direction inherently ill-posed. To account for this uncertainty, our framework estimates confidence scores for the predicted light direction and shadow mask and uses them to regulate how strongly each cue influences generation: reliable predictions guide the synthesis, while uncertain ones are down-weighted, allowing the model to rely more on learned image priors.

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.

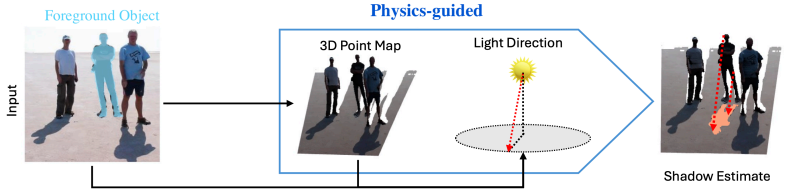


Figure 1: **Shadow estimates from approximate geometry and light direction.** Given a monocular RGB image and a foreground object mask, we recover an approximate 3D point map from MoGe-2 (Wang et al., 2025) and estimate a single dominant light direction, then infer a shadow estimate using object points. This shadow estimate serves as a condition for our shadow generator.

Finally, we incorporate the predicted shadow mask, light direction, and their confidence scores into a diffusion-based shadow generator (Ho et al., 2020; Zhao et al., 2025). The shadow mask provides dense spatial conditioning for shadow placement, while the light direction supplies scene-level cues for global shadow orientation. Confidence scores modulate these signals, and the diffusion model uses the resulting physics-guided conditioning to synthesize a visually coherent shadow, recovering appearance, softness, and scene interaction.

We evaluate our method on DESOBAV2 (Liu et al., 2024). Experiments show that our method improves both image quality and shadow mask accuracy, producing more visually faithful and spatially consistent shadows. Compared with the prior state-of-the-art (Zhao et al., 2025), our method achieves 23% lower shadow-region RMSE and 30% lower shadow-mask BER.

Our contributions are summarized as follows:

- **Physics-guided shadow generation.** We introduce a framework that derives an initial shadow estimate from recovered monocular scene geometry and a predicted dominant light direction, providing a physics-guided spatial anchor for shadow placement.
- **Quality-aware conditioning for uncertain physical cues.** We model the reliability of predicted lighting and shadow cues with confidence scores, allowing unreliable physical guidance to be down-weighted when illumination or shadow estimates are ambiguous.
- **Diffusion-based refinement with geometry and illumination guidance.** We incorporate the predicted shadow mask and light direction into a generator through quality-aware spatial and scene-level conditioning, enabling realistic shadow synthesis with improved shadow localization and geometric consistency.
- **State-of-the-art performance on DESOBAV2.** Experiments demonstrate improved shadow-region fidelity and localization, achieving 23% lower shadow-region RMSE and 30% lower shadow-mask BER than the prior state-of-the-art method.

2 Related Work

Image composition. Image composition integrates a foreground object into a background scene to produce a visually coherent composite image (Niu et al., 2021). Earlier work addressed individual aspects of image composition, including image blending (Pérez et al., 2003; Wu et al., 2019), image harmonization (Cong et al., 2020), and geometric correction (Lin et al., 2018). Recent work has advanced object placement (Zhou et al., 2025) and generative image composition (Song et al., 2023; 2024; Chen et al., 2024). Although such general-purpose methods may synthesize shadows as part of the composite, they typically either do not treat cast-shadow generation as a dedicated task or infer shadow placement only implicitly from image data. Our work focuses specifically on generating the missing cast shadow for a foreground object using explicit geometry and illumination cues.

Shadow generation. While shadow detection (Zhu et al., 2018; Le et al., 2018; Hu et al., 2021; Wang et al., 2024) and shadow removal (Cun et al., 2020; Le & Samaras, 2021; Guo et al., 2023; Hu et al., 2025) operate on observed image content by identifying or correcting shadow regions, shadow generation instead synthesizes a cast shadow that is absent from the input. This requires modeling both shadow appearance and placement while maintaining consistency with the foreground object,

scene geometry, and illumination. Existing shadow generation methods can be broadly categorized into 3D-based rendering approaches and 2D-based generative methods.

3D-based methods use explicit scene representations, including object geometry, lighting, and camera parameters, to simulate shadows through explicit rendering. Recovering these factors from a single image is an ill-posed inverse-rendering problem because geometry, illumination, and material appearance are tightly entangled. Early techniques (Chan & Durand, 2003; Schwarz & Stamminger, 2007) used depth maps and known light sources to produce soft shadows in screen space, while Sheng et al. (2022) used a manually provided pixel height map and user-specified lighting for controllable shadow generation. These methods provide explicit physical control but require detailed scene representations or manual input, limiting their use in image composition.

2D-based generative methods synthesize shadows directly in image space without requiring user-provided 3D scene representations. Early GAN-based methods learned shadow synthesis from object masks and image appearance (Zhang et al., 2019b; Liu et al., 2020). Hong et al. (2022) introduced a two-stage model that predicts a shadow mask before refining its appearance using an image-formation model. SGDiffusion (Liu et al., 2024) uses the composite image and object mask to condition a ControlNet-based diffusion model (Zhang et al., 2023), while GPSDiffusion (Zhao et al., 2025) introduces 2D geometry priors through a predicted rotated shadow bounding box and retrieved shadow shapes. However, because these methods encode physical constraints only indirectly through learned image-space features and priors, shadow shape and placement can remain unreliable.

Depth and lighting estimation. Recent progress in monocular geometry estimation makes it feasible to recover useful approximate scene structure from a single image. Monocular depth models perform well on diverse scenes, with Yang et al. (2024a;b) providing strong relative depth, and Yin et al. (2023); Piccinelli et al. (2024) demonstrating accurate metric depth. In this work, we use the off-the-shelf MoGe-2 (Wang et al., 2025) to recover metric point maps. Illumination can be represented in several forms, including environment maps (Somanath & Kurz, 2021; Verbin et al., 2024), light probes (Bai et al., 2023; Phongthawee et al., 2024), and parameterized lighting directions (Zhang et al., 2019a; Dastjerdi et al., 2023). For outdoor shadow generation, we use a parameterized light direction as a compact cue for geometric shadow estimation.

3 METHOD

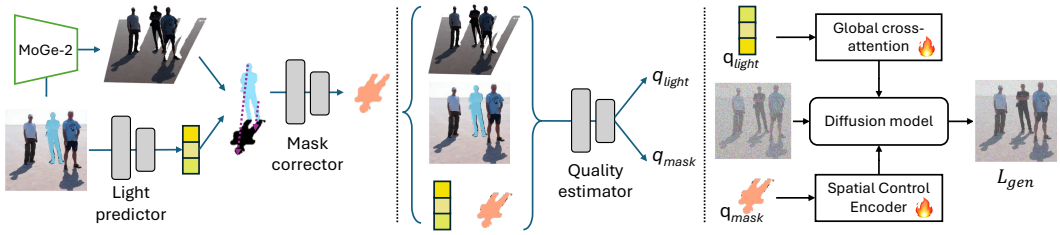
Given a composite image I containing a foreground object whose cast shadow is absent, together with its mask M_{fo}, our goal is to synthesize a physically plausible cast shadow for the object. We use an off-the-shelf monocular geometry estimator (Wang et al., 2025) that provides a dense point map $\mathbf{P}$, and we define the hint stack $\mathbf{H} = \mathrm{concat}(I, M_{fo}, \mathbf{P})$ as the base input to our framework. Our framework, as depicted in Figure 2, has three stages:

- First, a light predictor estimates a dominant 3D light direction $\hat{\mathbf{l}}$ from $\mathbf{H}$ under the directional-light assumption for outdoor scenes. Together with $\mathbf{P}$ and M_{fo}, the predicted direction is used to construct a coarse physics-guided shadow prior through 3D ray-direction alignment. A mask corrector then refines this prior into a full-resolution mask $\tilde{M}_{fs}$.
- Second, a quality estimator predicts two scalar confidence scores q_{light} and q_{mask}, measuring the reliability of $\hat{\mathbf{l}}$ and $\tilde{M}_{fs}$, respectively.
- Third, a frozen Stable Diffusion backbone is augmented with two trainable conditioning pathways: $\tilde{M}_{fs}$, scaled by q_{mask}, conditions a spatial control branch for dense local guidance, while scene features modulated by $\hat{\mathbf{l}}$ and gated by q_{light} are encoded as cross-attention tokens for global scene-level lighting guidance. An empty text prompt is used.

3.1 LIGHT PREDICTOR

We estimate a dominant scene-level 3D light direction from the hint stack $\mathbf{H}$. A ConvNeXt-Small backbone (Liu et al., 2022) processes $\mathbf{H}$, followed by global average pooling and a 2-layer MLP that regresses a vector $\mathbf{l} \in \mathbb{R}^3$, normalized to the unit direction $\hat{\mathbf{l}} = \mathbf{l}/\|\mathbf{l}\|_2$, pointing toward the light source. Training uses a cosine-similarity loss:

$$\mathcal{L}_{\text{lit}} = 1 - \hat{\mathbf{l}} \cdot \hat{\mathbf{l}}_{\text{gt}}, \tag{1}$$



(a) Shadow Mask Estimation (b) Quality Estimation (c) Shadow Generation

Figure 2: **Framework overview. (a) Shadow mask estimation.** A light predictor estimates a dominant 3D light direction from the input image, object mask, and 3D point map. Together with the point map and object mask, this direction is used to construct a coarse physics-guided soft shadow prior through 3D ray-direction alignment. A mask corrector refines the prior into a full-resolution shadow mask. **(b) Quality estimation.** A quality estimator predicts confidence scores q_{light} and q_{mask} for the predicted light direction and shadow mask, respectively. **(c) Shadow generation.** A frozen diffusion backbone is augmented with two trainable conditioning pathways. The confidence-weighted shadow mask provides dense spatial control, while the predicted light direction and its confidence modulate scene features injected through cross-attention for scene-level lighting guidance.

where $\hat{\mathbf{l}}_{\text{gt}}$ is obtained using our offline pseudo-labeling pipeline and used only for training supervision (see Appendix G). The light predictor is trained with the mask corrector and frozen thereafter.

3.2 Shadow Mask Estimation

Given the object mask, point map, and predicted light direction, we predict an auxiliary shadow mask $\tilde{M}_{fs}$ using a geometry-based shadow prior followed by a coarse-to-fine refinement network.

Geometry-based shadow prior. A receiver pixel is shadowed if the object blocks the light ray that would otherwise reach it. We approximate this by checking whether any object pixel lies along the shadow-flow direction toward the receiver. Let $\mathcal{O} = \{o \mid M_{fo}(o) = 1\}$ denote the object pixels and $\mathcal{R} = \{r \mid M_{fo}(r) = 0\}$ denote the candidate receiver pixels. For each $o \in \mathcal{O}$ and $r \in \mathcal{R}$, we compute the unit direction $\hat{\mathbf{v}}$ from the object point to the receiver point and measure its alignment with the shadow flow:

$$\hat{\mathbf{v}} = \frac{\mathbf{P}(r) - \mathbf{P}(o)}{\|\mathbf{P}(r) - \mathbf{P}(o)\|_2}, \qquad a = \hat{\mathbf{v}} \cdot \mathbf{D}, \tag{2}$$

where $\mathbf{D} = -\hat{\mathbf{l}}$ is the shadow flow direction from occluder toward receiver. The alignment score a is converted to a soft probability via a sigmoid with temperature T and angular tolerance τ, and each receiver pixel takes the maximum score across all object pixels:

$$s(o, r) = \sigma\left(\frac{a - \cos\tau}{T}\right), \qquad \text{prior}(r) = \max_{o \in \mathcal{O}} s(o, r). \tag{3}$$

In practice, we compute this prior at 64×64 resolution for efficiency and apply a small Gaussian blur for spatial smoothness.

Coarse-to-fine refinement. The soft prior captures approximate location but is too coarse and noisy for fine shadow shape and boundaries. We refine it in two stages. At 64×64, the soft prior, downsampled object mask, and point map are passed through a lightweight network to produce coarse logits, which are then upsampled to 512×512 as L_c. At full resolution, a residual correction Δ is predicted from L_c, $\mathbf{H}$, and a broadcast light map with each pixel set to $\hat{\mathbf{l}}$, then added in logit space.

The final mask is:

$$\tilde{M}_{fs} = \sigma(L_c + \Delta)\,. \tag{4}$$

The refinement network is trained end-to-end using BCE and Dice losses (Sudre et al., 2017):

$$\mathcal{L}_{\text{mask}} = \text{BCE}(\tilde{M}_{fs}, \hat{M}_{fs}) + \lambda_1 \mathcal{L}_{\text{Dice}}(\tilde{M}_{fs}, \hat{M}_{fs}), \tag{5}$$

where $\hat{M}_{fs}$ is the ground-truth shadow mask. This module is trained jointly with the light predictor and frozen during generator training.

3.3 Quality Estimation

Light and mask predictions can be unreliable because single-image illumination is ill-posed and the recovered point map is approximate. We therefore predict confidence scores $q_{\text{light}},\ q_{\text{mask}} \in [0, 1]$ to gate their influence: reliable cues guide synthesis, while uncertain ones are down-weighted.

Let $\mathbf{H}_{64} = \text{concat}(I_{64},\ M_{fo,64},\ \mathbf{P}_{64})$ and let $\tilde{M}_{fs,64}$ be the downsampled predicted mask. The quality network takes $\mathbf{H}_{64}$ and $\tilde{M}_{fs,64}$ as spatial inputs, along with $\hat{\mathbf{l}}$ and an intermediate light-predictor feature as lighting inputs. A convolutional encoder pools $\text{concat}(\mathbf{H}_{64},\ \tilde{M}_{fs,64})$, and two MLP heads predict q_{light} and q_{mask}; the light head uses the lighting inputs, while the mask head also uses lightweight mask-geometry cues such as centroid offset, elongation, and disagreement with the soft prior.

Light-direction and mask errors are mapped to soft quality targets. Let $\theta = \arccos(\hat{\mathbf{l}} \cdot \hat{\mathbf{l}}_{\text{gt}})$ and $e_{\text{mask}} = 1 - \text{Dice}(\tilde{M}_{fs}, \hat{M}_{fs})$ denote the light and mask errors. We define

$$q^*_{\text{light}} = \exp(-\theta/\sigma_{\text{ang}}), \qquad q^*_{\text{mask}} = \exp(-e_{\text{mask}}/\sigma_{\text{mask}}). \tag{6}$$

The training objective combines Smooth-ℓ_1 regression and pairwise ranking:

$$\mathcal{L}_q = \sum_{c\in\{\text{light,mask}\}} \left(\lambda_{\text{reg}}\mathcal{L}^{\text{reg}}_c + \lambda_{\text{rank}}\mathcal{L}^{\text{rank}}_c\right). \tag{7}$$

The regression loss matches the soft targets, while the ranking loss is applied to pairs whose errors differ by more than a margin m, set separately for each cue, and assigns higher confidence to the lower-error prediction. The ranking term is the primary supervision signal. The quality network is trained with the light predictor and mask corrector frozen, then frozen during generator training.

3.4 Shadow Generator

We follow the recent diffusion-based shadow generator (Zhao et al., 2025) and inject the predicted shadow mask and light direction through confidence-gated spatial and lighting pathways.

Dense spatial conditioning. A ControlNet (Zhang et al., 2023) branch receives the hint stack augmented with the predicted shadow mask. Before concatenation, the mask is scaled by q_{mask}, so unreliable masks contribute weakly while the other hint channels remain unchanged.

Scene-level lighting conditioning. The light direction provides a global cue for shadow orientation. A lightweight encoder maps the image and object mask to features, which are modulated by $\hat{\mathbf{l}}$ via FiLM (Perez et al., 2018) with strength gated by q_{light}. The features are projected into tokens and injected into the denoising U-Net via IP-Adapter (Ye et al., 2023) cross-attention.

Training. Given timestep t and Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the noised latent is:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \tag{8}$$

and the denoising network is trained to predict the added noise:

$$\mathcal{L}_{\text{gen}} = \mathbb{E}\left[\|\boldsymbol{\epsilon} - f(\mathbf{z}_t, t, \textit{cond})\|_2^2\right], \tag{9}$$

where *cond* denotes all conditioning inputs. During training, we upweight the noise-prediction loss within a dilated ground-truth shadow region. The denoising U-Net is frozen; only the ControlNet branch and the lighting-conditioning pathway, including its encoder, FiLM layers, and cross-attention adapter, are updated during generator training.

4 Experiments and Results

Dataset. We conduct experiments on DESOBAV2 (Liu et al., 2024), which contains 21,575 scenes with 28,573 image tuples. The public release is split into 27,823 training tuples and 750 test tuples. The test set covers two conditions: background-object-shadow (BOS), where reference background-object–shadow pairs exist in the image; and BOS-free, where there is only a single object–shadow pair. We obtain per-image point maps from MoGe-2 (Wang et al., 2025) and derive 22,364 approximate

Table 1: Comparison with state-of-the-art methods on DESOBAV2 under BOS and BOS-free settings. GR/LR, GS/LS, and GB/LB denote global/local RMSE, SSIM, and BER, respectively. GAN-based methods are pretrained on DESOBA. Diffusion methods use best-of-five selection; *italics* denote single-sample results averaged over five independent draws. Best scores are in **bold**.

Method	BOS						BOS-free					
	GR↓	LR↓	GS↑	LS↑	GB↓	LB↓	GR↓	LR↓	GS↑	LS↑	GB↓	LB↓
ShadowGAN (Zhang et al., 2019b)	8.681	70.459	0.961	0.174	0.470	0.938	19.146	87.149	0.903	0.052	0.483	0.961
Mask-SG (Hu et al., 2019)	10.450	73.776	0.938	0.186	0.485	0.966	19.662	89.637	0.895	0.054	0.488	0.972
AR-SG (Liu et al., 2020)	8.873	69.336	0.957	0.190	0.463	0.922	19.594	84.939	0.896	0.057	0.468	0.925
SGRNet (Hong et al., 2022)	9.017	71.582	0.961	0.189	0.446	0.887	20.883	85.841	0.894	0.056	0.450	0.881
SGDiffusion (Liu et al., 2024)	7.366	51.830	0.962	0.332	0.218	0.433	14.664	54.931	0.914	0.168	0.179	0.348
GPSDiffusion (Zhao et al., 2025)	6.694	41.266	0.967	0.449	0.140	0.274	18.207	52.602	0.906	0.178	0.147	0.271
Ours	**4.493**	**33.530**	**0.974**	**0.533**	**0.107**	**0.211**	**9.785**	**36.126**	**0.934**	**0.322**	**0.079**	**0.151**
GPSDiffusion (average of 5)	7.980	50.290	0.964	0.318	0.197	0.387	21.560	61.920	0.897	0.111	0.202	0.376
Ours (average of 5)	5.230	39.910	0.972	0.428	0.143	0.283	11.780	43.330	0.927	0.241	0.113	0.217

light-direction pseudo-labels using our offline pipeline. Light losses use only training tuples with valid directions, while all training tuples are retained for mask and generator training.

Baselines and Metrics. We compare against four GAN-based methods, ShadowGAN (Zhang et al., 2019b), MaskShadowGAN (Hu et al., 2019), ARShadowGAN (Liu et al., 2020), and SGRNet (Hong et al., 2022), pretrained on DESOBA (Hong et al., 2022). For SGDiffusion (Liu et al., 2024) and GPSDiffusion (Zhao et al., 2025), we re-evaluate their released checkpoints on the public DESOBAV2 test split using identical preprocessing. Following prior work (Zhao et al., 2025), we report RMSE and SSIM for image quality, and BER for shadow mask accuracy, evaluated on both the global image (G) and the local foreground-shadow region (L). For all diffusion-based methods, we generate five samples per test case and select the sample with the highest local SSIM against the ground truth.

Implementation Details. We train the shadow generator with AdamW (Loshchilov & Hutter, 2019) at a learning rate of 1×10^{-5} for 30 epochs on two NVIDIA RTX A6000 GPUs with a batch size of 4. Inference uses 50-step DDIM (Song et al., 2021), followed by the post-processing network of Liu et al. (2024); remaining settings are provided in Appendix F.

4.1 Comparisons with SOTA

Quantitative Results. Our method achieves the best performance across all metrics in both BOS and BOS-free settings (Table 1). Under the best-of-five protocol, compared with GPSDiffusion (Zhao et al., 2025), our method reduces LRMSE by 31% in BOS-free ($52.602 \rightarrow 36.126$) and 19% in BOS ($41.266 \rightarrow 33.530$), with consistent global and local SSIM gains. GBER/LBER decrease by 24%/23% in BOS and 46%/44% in BOS-free.

For single-sample evaluation, we compare the methods using one sample per test case in each of five draws and average the results, without selecting the best sample. Under this protocol, our method outperforms GPSDiffusion across every metric, reducing LRMSE/LBER by 21%/27% in BOS and 30%/42% in BOS-free. These results show that the gains persist under single-sample inference and do not rely on ground-truth-based best-of-five selection. Across both protocols, the improvements in LRMSE and LBER indicate greater shadow-region fidelity and more accurate localization.

Qualitative Results. Figure 3 compares generated images and predicted shadow masks from SGDiffusion (Liu et al., 2024), GPSDiffusion (Zhao et al., 2025), and our method. In the shown cases, SGDiffusion produces incomplete shadow shapes or inconsistent intensity (rows 4, 6, and 8), while GPSDiffusion can drift or produce inaccurate shadow extent and shape (rows 1, 3, and 7). In contrast, our method anchors shadow location and shape using a physics-guided shadow estimate and dominant lighting cue, producing plausible results on varied receivers, including uneven and textured surfaces (rows 2 and 5). Additional qualitative comparisons are provided in Appendix H.

Challenging Cases. Reliable light-direction pseudo-labels cannot always be recovered from object–shadow pairs, particularly under spatially varying illumination or out-of-frame shadow truncation. Such cases are excluded from light-predictor supervision. At inference, the predicted confidence

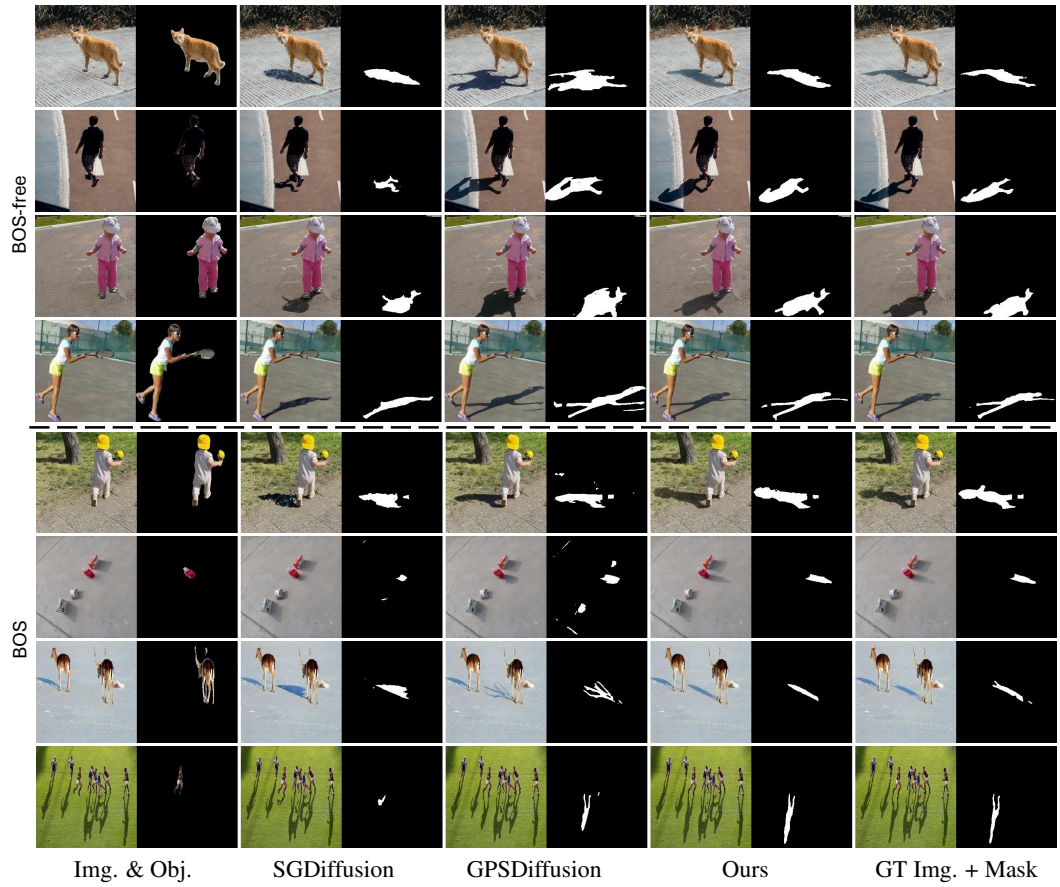


Figure 3: **Qualitative comparison with state-of-the-art methods.** Results in both BOS (with reference background object–shadow pairs) and BOS-free (with a single object–shadow pair) settings. We compare generated images and predicted shadow masks from SGDiffusion (Liu et al., 2024), GPSDiffusion (Zhao et al., 2025), and our method with the ground truth. Our method produces more accurately localized shadow masks and visually coherent shadows across both settings.

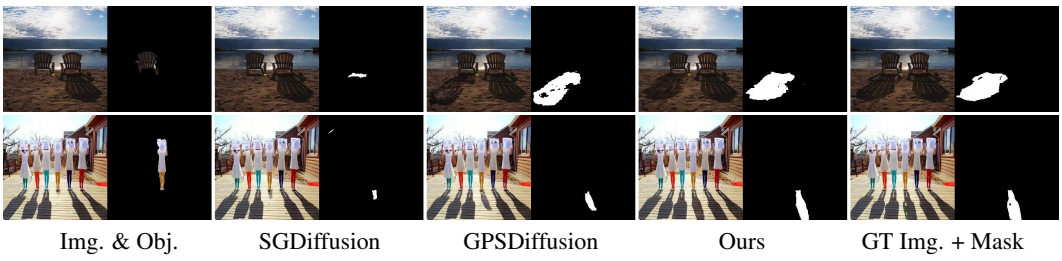


Figure 4: **Challenging cases.** (*Top*) Spatially varying illumination makes a single dominant light direction ambiguous. (*Bottom*) An out-of-frame shadow truncates the reference object–shadow pair.

scores regulate cue strength during generation. Figure 4 shows that our method produces coherent, well-placed shadows, while the baselines produce incomplete or displaced shadows.

4.2 Performance of the Light and Mask Predictors

Table 2 summarizes predictor performance on DESOBAV2. We report light-direction angular error on samples with valid light directions and full-resolution shadow-mask Dice on all test samples. Light prediction is more accurate in BOS than BOS-free, with mean angular errors of 6.78° and 10.11°,

Table 2: Predictor statistics on the DESOBAV2 test set. N denotes the number of evaluated samples. Light-direction error is reported in degrees on samples with valid directions, while mask quality is reported as full-resolution Dice on all samples.

Light	N	Mean	p50 / p90
BOS	375	6.780	5.90 / 11.40
BOS-free	235	10.110	7.86 / 20.94

Mask	N	Mean	p50 / p90
BOS	500	0.502	0.534 / 0.712
BOS-free	250	0.486	0.513 / 0.649

Table 3: Reliability of predicted quality scores on DESOBAV2 test set. We report Pearson correlation with the corresponding cue error and mean errors among the highest- and lowest-scoring 20% of samples. Mask error is $1 - \mathrm{Dice}$, and light error is angular error in degrees. Negative correlations and lower top-group errors indicate reliable ranking.

Score	Split	N	Pearson r	Top 20% Error	Bottom 20% Error
q_{mask}	BOS	500	-0.751	0.368	0.743
	BOS-free	250	-0.556	0.430	0.649
q_{light}	BOS	375	-0.355	4.583	10.461
	BOS-free	235	-0.518	6.308	16.240



Img. & Obj. Shadow Est. Forced $q = 1$ Predicted q GT

Figure 5: **Quality-guided generation.** We compare forcing $q_{\text{mask}} = q_{\text{light}} = 1$ with using their predicted values. Full-strength conditioning can propagate errors from inaccurate cues, whereas the predicted scores attenuate unreliable guidance and produce better-aligned shadows and masks.

respectively. This gap is consistent with the richer scene-level evidence in BOS, where reference background object–shadow pairs constrain illumination. Without such pairs, light-direction estimation is more ambiguous. Mask quality remains similar across the two settings, with mean Dice scores of 0.502 and 0.486. Overall, the predictors provide useful but imperfect cues, motivating confidence-aware conditioning rather than treating their outputs as fixed constraints. See Appendices A and B for predictor visualizations.

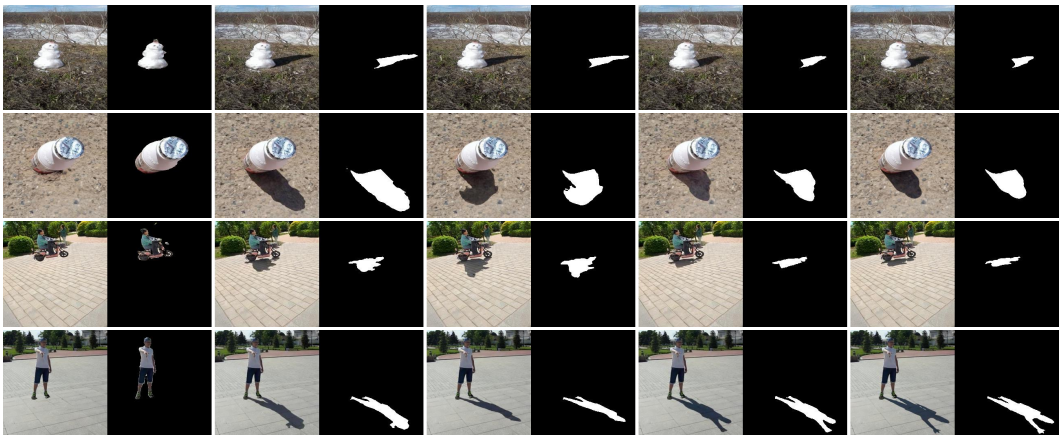
4.3 Performance of the Quality Estimation Network

The quality network predicts $q_{\text{mask}}, q_{\text{light}} \in [0, 1]$ to represent the relative reliability of the shadow-mask and light-direction cues. Table 3 evaluates whether these scores track their corresponding prediction errors. We report Pearson correlation with error and the mean errors of the highest- and lowest-scoring 20% of samples. For both cues and both test settings, the correlations are negative, and the highest-scoring samples have substantially lower errors than the lowest-scoring samples. Thus, the predicted scores provide meaningful reliability signals for confidence-aware conditioning. Beyond tracking cue error, the predicted scores outperform fixed $q_{\text{mask}} = q_{\text{light}} \in \{0, 0.5, 1\}$ on all six BOS-free metrics, supporting the benefit of input-adaptive conditioning (Appendix C).

Figure 5 further illustrates how the quality estimates affect generation when the predicted conditioning cues are inaccurate. Setting $q_{\text{mask}} = q_{\text{light}} = 1$ applies both cues at full strength and can propagate their errors. In contrast, the predicted scores attenuate unreliable conditioning, allowing the generator to rely more on learned image priors and yield better-aligned shadows and masks.

Table 4: Ablation of the predictor and quality-network design on the DESOBAV2 BOS-free setting. Configuration 1 is GPSDiffusion. † marks predictors updated during generator training; Configuration 3 applies their outputs without quality gating. Best displayed scores are in **bold**.

Config.	PointMap	Light	Mask	Quality	GRMSE↓	LRMSE↓	GSSIM↑	LSSIM↑	GBER↓	LBER↓
1	–	–	–	–	18.207	52.602	0.906	0.178	0.147	0.271
2	✓	–	–	–	10.041	37.010	0.933	0.317	0.082	0.157
3	✓	†	†	–	10.295	37.976	0.932	0.316	0.089	0.170
4	✓	✓	–	✓	9.873	37.651	**0.934**	0.319	0.087	0.168
5	✓	✓	✓	✓	**9.785**	**36.126**	**0.934**	**0.322**	**0.079**	**0.151**



Img. & Obj. PointMap (PM) PM+Predictors Ours GT

Figure 6: **Ablations.** Point-map guidance provides coarse shadow placement, while ungated predictor conditioning can introduce errors in shadow extent or direction. Our full model modulates these cues using their predicted quality, improving shadow shape and localization.

4.4 Ablation Studies

Table 4 evaluates point-map guidance, predictor-based conditioning, and quality-aware conditioning on DESOBAV2 (BOS-free). The point-map-only configuration substantially outperforms GPSDiffusion (Configuration 2 vs. 1), supporting the value of monocular geometry for shadow placement. Applying jointly trained light and mask predictors without quality gating slightly degrades all metrics (Configuration 3 vs. 2), indicating that direct conditioning on imperfect predictions can be counterproductive. Quality-gated light conditioning alone improves global RMSE and SSIM but remains weaker in BER (Configuration 4), showing that the global lighting cue cannot replace dense mask guidance for accurate localization. The full configuration achieves the best overall performance, supporting the combined use of geometric, lighting, mask, and quality-aware guidance.

Figure 6 visualizes this progression. Point-map guidance provides coarse shadow placement, while ungated predictor conditioning can introduce errors in shadow extent or direction. The full model modulates these cues using their predicted quality, improving shadow shape and localization.

5 Conclusion

We present a physics-guided framework for generating cast shadows for foreground objects in outdoor composite images. Our framework explicitly links approximate monocular scene geometry and a predicted dominant light direction to cast-shadow placement through a coarse geometric estimate. Rather than treating these uncertain physical estimates as exact constraints, the framework predicts quality scores that regulate the influence of the resulting shadow-mask and lighting cues during diffusion refinement. This combination provides an explicit spatial anchor while retaining learned flexibility for realistic shadow appearance. Experiments on DESOBAV2 demonstrate state-of-the-art shadow-region fidelity and localization.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on the design of the ray-alignment and quality-estimation components, assist with their implementation, and support the interpretation and presentation of experimental results. We did not use generative AI tools for other tasks requiring disclosure, and tasks unrelated to this work are not applicable. Additionally, we used generative AI tools to create or edit software code and to edit the paper for readability, clarity, and concision. The authors reviewed all AI-assisted work, tested the AI-assisted implementations and code, verified interpretations and numerical claims against the experimental results, and checked manuscript edits for accuracy. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The main paper describes the model, training objectives, dataset, baselines, evaluation metrics, and inference protocol. Additional details on light-direction pseudo-label acquisition and architectural, optimization, and computational settings are provided in Appendices G and F. We will release the pseudo-label construction, training, and evaluation code, together with the derived supervision targets and trained checkpoints, subject to applicable licenses.

REFERENCES

Jiayang Bai, Zhen He, Shan Yang, Jie Guo, Zhenyu Chen, Yan Zhang, and Yanwen Guo. Local-to-global panorama inpainting for locale-aware indoor lighting prediction. *IEEE Transactions on Visualization and Computer Graphics*, 29(11):4405–4416, 2023.

Eric Chan and Frédo Durand. Rendering fake soft shadows with smoothies. In *Proceedings of the Eurographics Symposium on Rendering*, pp. 208–218, 2003.

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. AnyDoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6593–6602, 2024.

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.

Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10680–10687, 2020.

Mohammad Reza Karimi Dastjerdi, Jonathan Eisenmann, Yannick Hold-Geoffroy, and Jean-François Lalonde. Everlight: Indoor-outdoor editable hdr lighting estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7420–7429, 2023.

Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. 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.

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.

Shilin Hu, Hieu Le, ShahRukh Athar, Sagnik Das, and Dimitris Samaras. Shadow removal refinement via material-consistent shadow edges. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2631–2641. IEEE, 2025.

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.

Xiaowei Hu, Tianyu Wang, Chi-Wing Fu, Yitong Jiang, Qiong Wang, and Pheng-Ann Heng. Revisiting shadow detection: A new benchmark dataset for complex world. *IEEE Transactions on Image Processing*, 30:1925–1934, 2021.

Jean-François Lalonde, Alexei A. Efros, and Srinivasa G. Narasimhan. Estimating natural illumination from a single outdoor image. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 183–190, 2009.

Hieu Le and Dimitris Samaras. Physics-based shadow image decomposition for shadow removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9088–9101, 2021.

Hieu Le, Tomas F Yago Vicente, Vu Nguyen, Minh Hoai, and Dimitris Samaras. A+ d net: Training a shadow detector with adversarial shadow attenuation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 662–678, 2018.

Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. ST-GAN: Spatial transformer generative adversarial networks for image compositing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9455–9464, 2018.

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. 8136–8145, 2020.

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.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

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.

Alexandros Panagopoulos, Chaohui Wang, Dimitris Samaras, and Nikos Paragios. Simultaneous cast shadows, illumination and geometry inference using hypergraphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2):437–449, 2013. doi: 10.1109/TPAMI.2012.110.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM Transactions on Graphics*, 22(3):313–318, 2003. doi: 10.1145/882262.882269.

Pakkapon Phongthawee, Worameth Chinchuthakun, Nontaphat Sinsunthithet, Varun Jampani, Amit Raj, Pramook Khungurn, and Supasorn Suwajanakorn. Diffusionlight: Light probes for free by painting a chrome ball. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 98–108, 2024.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10106–10116, 2024.

Imari Sato, Yoichi Sato, and Katsushi Ikeuchi. Illumination from shadows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(3):290–300, 2003.

Michael Schwarz and Marc Stamminger. Bitmask soft shadows. In *Computer Graphics Forum*, volume 26, pp. 515–524. Wiley Online Library, 2007.

Yichen Sheng, Yifan Liu, Jianming Zhang, Wei Yin, A Cengiz Oztireli, He Zhang, Zhe Lin, Eli Shechtman, and Bedrich Benes. Controllable shadow generation using pixel height maps. In *European Conference on Computer Vision*, pp. 240–256. Springer, 2022.

Gowri Somanath and Daniel Kurz. Hdr environment map estimation for real-time augmented reality. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11298–11306, 2021.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

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.

Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, He Zhang, Wei Xiong, and Daniel Aliaga. IMPRINT: Generative object compositing by learning identity-preserving representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8048–8058, 2024.

Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *International Workshop on Deep Learning in Medical Image Analysis*, pp. 240–248. Springer, 2017.

Dor Verbin, Ben Mildenhall, Peter Hedman, Jonathan T Barron, Todd Zickler, and Pratul P Srinivasan. Eclipse: Disambiguating illumination and materials using unintended shadows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 77–86, 2024.

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. *arXiv preprint arXiv:2507.02546*, 2025.

Yonghui Wang, Shaokai Liu, Li Li, Wengang Zhou, and Houqiang Li. SwinShadow: Shifted window for ambiguous adjacent shadow detection. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 20(11):360:1–360:20, 2024. doi: 10.1145/3688803.

Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. GP-GAN: Towards realistic high-resolution image blending. In *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 2487–2495, 2019. doi: 10.1145/3343031.3350944.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10371–10381, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024b.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9043–9053, 2023.

Jinsong Zhang, Kalyan Sunkavalli, Yannick Hold-Geoffroy, Sunil Hadap, Jonathan Eisenman, and Jean-François Lalonde. All-weather deep outdoor lighting estimation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 10158–10166, 2019a.

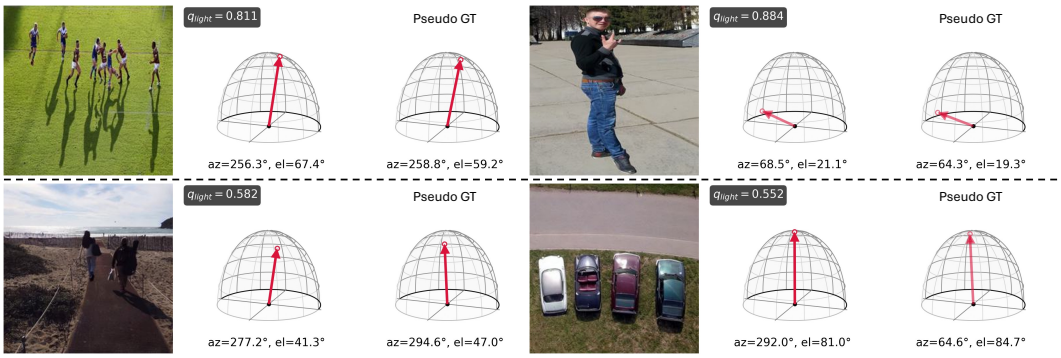


Figure 7: **Light-direction predictions.** Predicted and pseudo-labeled light directions are shown as arrows pointing toward the light source on a hemisphere defined in the input-view coordinate frame. Azimuth and elevation are shown for each direction, together with q_{light} for the prediction.

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, 2019b.

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.

Hang Zhou, Xinxin Zuo, Rui Ma, and Li Cheng. BOOTPLACE: Bootstrapped object placement with detection transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19294–19303, 2025.

Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 121–136, 2018.

A Visualization of Light-Direction Predictions

Figure 7 compares our predicted light directions with the light-direction pseudo-labels produced by the offline labeling pipeline (see Appendix G). We visualize each direction as an arrow on a hemisphere, together with its azimuth and elevation, and report the predicted light-quality score q_{light}. The hemisphere is defined in the input-view coordinate frame and is therefore aligned with the camera viewpoint rather than a world coordinate frame. Accordingly, the directions and angles should be interpreted relative to the image view. For readability, the front half of the hemisphere is rendered more lightly. The visualizations are consistent with the quantitative results in Table 3: predictions closer to their pseudo-labels generally receive higher q_{light} values, while predictions with larger angular errors receive lower scores.

B Visualization of Shadow-Mask Predictions

Figure 8 compares the geometry-derived soft shadow prior, the refined shadow-mask prediction, and the ground-truth shadow mask. The prior and refined prediction are visualized as probability heatmaps, while the ground-truth mask is binary. The soft prior provides a coarse estimate of shadow orientation and spatial support from the recovered geometry and predicted light direction. The mask predictor refines this estimate to better capture the ground-truth shadow shape and extent. We additionally report the scalar mask-quality score q_{mask} for each refined prediction. The visualizations

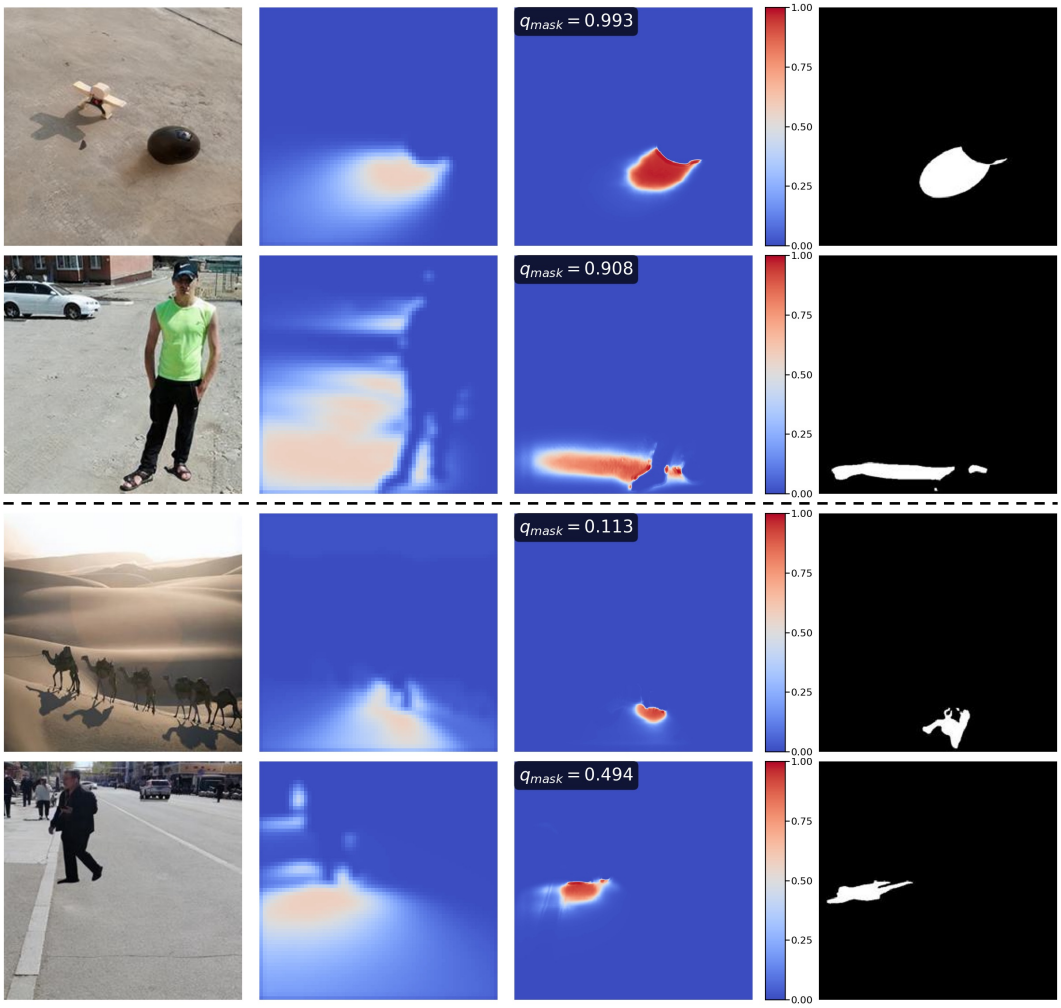


Figure 8: **Shadow-mask predictions.** The geometry-derived soft shadow prior and refined shadow-mask prediction are shown as probability heatmaps, while the ground-truth mask is binary. We also report the scalar quality score q_{mask} for each refined prediction.

follow the trend quantified in Table 3: higher q_{mask} values generally correspond to better agreement with the ground-truth mask, while lower values are associated with larger mask errors.

C Sensitivity to Quality Conditioning

We further evaluate whether per-input quality prediction is more effective than applying a fixed conditioning strength. At inference, we replace both predicted quality scores with a shared constant $q_{\text{mask}} = q_{\text{light}} \in \{0, 0.5, 1\}$ while keeping the predicted cues and all model parameters unchanged. Table 5 reports results on the DESOBAV2 BOS-free setting.

The predicted quality scores achieve the best performance across all six metrics. Among the fixed settings, $q = 0.5$ generally performs best, but it remains consistently below per-input prediction, while fully trusting both cues with $q = 1$ produces the weakest results on five of the six metrics. These results support adapting the conditioning strength to the estimated reliability of each input. Because the fixed values are applied to a model trained with predicted quality scores, this experiment is interpreted as an inference-time sensitivity analysis rather than a comparison with separately trained fixed-gating models.

Table 5: Sensitivity to quality conditioning on DESOBAV2 BOS-free. Fixed values are applied to both q_{mask} and q_{light} at inference. All results use the best-of-five protocol. Best scores are in **bold**.

q setting	GRMSE↓	LRMSE↓	GSSIM↑	LSSIM↑	GBER↓	LBER↓
Fixed 0	11.00	38.84	0.9244	0.2939	0.0803	0.1535
Fixed 0.5	10.97	38.69	0.9246	0.2964	0.0799	0.1528
Fixed 1	11.24	39.26	0.9240	0.2950	0.0824	0.1573
Predicted	**9.78**	**36.13**	**0.9335**	**0.3225**	**0.0793**	**0.1514**

Table 6: Additional conditioning ablations on the DESOBAV2 BOS-free setting. Rows 1–6 are evaluated on all 250 samples. Row 6^* reevaluates the full model on the 235 samples with valid light-direction pseudo-labels, providing a matched comparison with the oracle in Row 7. `bg` denotes background-mask conditioning, `PL` denotes the offline pseudo-labeled light direction, and `GT` denotes the ground-truth shadow mask. Best learned-setting scores on the full test set are in **bold**. † indicates that the module is not frozen during training.

Config.	PointMap	Light	Mask	Quality	GRMSE↓	LRMSE↓	GSSIM↑	LSSIM↑	GBER↓	LBER↓
1	–	–	–	–	18.207	52.602	0.906	0.178	0.147	0.271
2	✓	–	–	–	10.041	37.010	0.933	0.317	0.082	0.157
3	✓	†	†	–	10.295	37.976	0.932	0.316	0.089	0.170
4	✓	✓	–	✓	9.873	37.651	**0.934**	0.319	0.087	0.168
5	✓	`bg`	✓	✓	9.916	36.996	0.933	0.316	0.083	0.160
6	✓	✓	✓	✓	**9.785**	**36.126**	**0.934**	**0.322**	**0.079**	**0.151**
6^*	✓	✓	✓	✓	9.812	36.296	0.934	0.322	0.079	0.152
7	✓	`PL`	`GT`	–	4.345	18.362	0.957	0.583	0.016	0.032

D Additional Ablations with Oracle Conditioning

Table 6 extends the main ablation with background-mask and oracle-conditioning variants. In Row 5, the light predictor receives a background mask instead of the foreground object mask. This variant remains competitive but consistently underperforms Row 6, supporting foreground-object conditioning for estimating the relevant light direction.

For a matched oracle comparison, Rows 6^* and 7 are evaluated on the 235 BOS-free test samples with valid light-direction pseudo-labels. Row 6^* evaluates the full learned model on this subset. In Row 7, the predicted light direction and shadow mask are replaced by the offline pseudo-labeled direction and ground-truth mask, respectively, while the quality estimator is bypassed. Oracle conditioning substantially improves all metrics, showing that the generator can effectively use more accurate conditioning and that improving light and mask estimation offers substantial additional headroom.

E Limitations

Figure 9 illustrates two representative limitations of our method. First, when the foreground object is truncated or occluded, the recovered geometry does not capture its unseen portions. The resulting geometric shadow estimate can therefore be incomplete, leaving the generator to infer the missing shadow structure from learned priors and sometimes producing an incorrect shadow shape. Second, the dominant light-direction cue constrains shadow orientation but does not explicitly represent attenuation strength. Consequently, the generated shadow can have plausible placement but differ from the ground truth in intensity. Future work could incorporate object-shape completion and explicit attenuation estimation within the reliability-aware conditioning framework.

F Additional Implementation Details

This section complements the main paper with additional architectural and optimization details to support reproducibility.

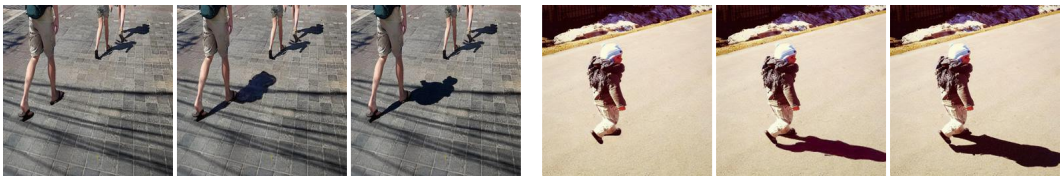


Figure 9: **Representative limitations.** Each group shows the input, our result, and ground truth. (*Left*) Incomplete foreground geometry can lead to an incorrect shadow shape. (*Right*) Without an explicit attenuation cue, the generated shadow can have plausible placement but inaccurate intensity.

Predictor training and shadow-prior settings. The light predictor and mask corrector are jointly optimized for 10 epochs using AdamW with a learning rate of 1×10^{-4}, weight decay 1×10^{-2}, and batch size 8. The light and mask losses have equal weights, $w_{\text{lit}} = w_{\text{mask}} = 1$. For mask refinement, we use

$$\mathcal{L}_{\text{mask}} = \text{BCE} + 0.2\,\mathcal{L}_{\text{Dice}}. \tag{10}$$

The geometry-derived soft shadow prior is computed at 64×64 resolution with angular tolerance $\tau = 10^\circ$ and sigmoid temperature $T = 0.05$. We then apply a 5×5 Gaussian filter with standard deviation 1.0 to smooth the prior before mask refinement.

Quality-estimator architecture. The refined shadow mask is area-downsampled to 64×64, normalized by its mean value, and clipped to $[0, 5]$ before being concatenated with $\mathbf{H}_{64}$. The spatial encoder consists of three stride-2 convolutional layers with 32, 64, and 64 output channels. Mean-and-standard-deviation pooling converts the resulting feature map into a 128D descriptor.

The light-quality head concatenates this descriptor with the predicted 3D light direction and the 256D feature from the light predictor. The mask-quality head concatenates the descriptor with the predicted direction and a 4D geometric feature comprising the 2D centroid offset, log elongation, and mean absolute disagreement between the refined mask and soft shadow prior. Both heads use two hidden layers of width 256 followed by a scalar output.

Quality-estimator training. The light and mask predictors remain frozen while training the quality estimator. We train it for 10 epochs with AdamW using a learning rate of 1×10^{-4}, weight decay 1×10^{-2}, and batch size 8. The quality-target scales are

$$\sigma_{\text{ang}} = 10^\circ, \qquad \sigma_{\text{mask}} = 0.70. \tag{11}$$

We use $\lambda_{\text{reg}} = 0.2$ and $\lambda_{\text{rank}} = 1.0$ for both quality heads. The pairwise ranking margins are $m_{\text{light}} = 2^\circ$ for angular error and $m_{\text{mask}} = 0.05$ for Dice error, with ranking temperature 1.0. Mask-quality targets and ranking errors are computed after area-downsampling the predicted and ground-truth masks to 64×64.

In addition to the objective in the main paper, we use a small auxiliary correlation regularizer:

$$\mathcal{L}_q^{\text{total}} = \mathcal{L}_q + \lambda_{\text{area}}\,\text{Corr}\Big(A(\tilde{M}_{fs,64}), z_{\text{mask}}\Big)^2, \tag{12}$$

where Corr denotes batchwise Pearson correlation, $A(\tilde{M}_{fs,64})$ is the mean predicted-mask area, and z_{mask} is the pre-sigmoid mask-quality logit. We set $\lambda_{\text{area}} = 0.1$.

Generator-specific settings. Training uses 512×512 inputs and 64×64 diffusion latents. To construct the shadow-region loss weights, we resize the ground-truth shadow mask to 64×64 and dilate it once using a 6×6 kernel. Pixels whose resulting mask value is at least 0.6 receive denoising-loss weight 10, while all remaining pixels receive weight 1. At inference, classifier-free guidance uses a scale of 9.0 and DDIM sampling uses $\eta = 0$.

Computational cost. Generator training uses approximately 41 GiB of memory per GPU and takes approximately 80 minutes per epoch; the training server additionally has dual AMD EPYC 7352 CPUs. Under the five-sample, 50-step evaluation protocol, our method requires approximately 50



Figure 10: **Valid approximate light.** Examples where the automatically estimated light direction produces a rendered shadow density that closely matches the observed foreground shadow mask. The cast shadows from the object points align well with the observed shadow location and shape, indicating a plausible approximate light direction.

seconds per test case, compared with 44 seconds for SGDiffusion (Liu et al., 2024) and 47 seconds for GPSDiffusion (Zhao et al., 2025). Peak GPU memory is 32.5 GiB for our method, compared with 18.4 GiB and 27.6 GiB for SGDiffusion and GPSDiffusion, respectively.

G Light-Direction Pseudo-Label Acquisition

We estimate approximate dominant light directions offline from the object–shadow pairs available in DESOBAV2. For each scene, MoGe-2 (Wang et al., 2025) predicts a dense 3D point map $\mathbf{p}(x, y) \in \mathbb{R}^3$. We select one foreground object–shadow pair and use its object mask M_{fo} and observed shadow mask M_{fs} to sample two sets of points: occluder points on the object and receiver points on the shadowed surface. The resulting direction is shared by all tuples from that scene.

Foreground pair selection. For each scene, we select the object–shadow pair whose shadow mask has the largest area. If the mask is fragmented, only its largest connected component is retained. Cases with extremely small shadow regions, unreliable geometry, or failed optimization are excluded from light-direction supervision.

Receiver surface estimation. We approximate the receiver surface by fitting a plane to 3D points sampled inside the shadow mask. Given receiver points $\{\mathbf{r}_k\}$, the plane normal is obtained via PCA on the covariance matrix

$$\mathbf{\Sigma} = \frac{1}{K} \sum_k (\mathbf{r}_k - \bar{\mathbf{r}})(\mathbf{r}_k - \bar{\mathbf{r}})^\top, \tag{13}$$

where $\bar{\mathbf{r}}$ is the centroid. The eigenvector corresponding to the smallest eigenvalue defines the plane normal $\mathbf{n}$.

Shadow casting. Assuming a single directional light, we parameterize the light direction, pointing toward the source, by azimuth ϕ and elevation θ:

$$\boldsymbol{\ell}(\phi, \theta) = [-\cos\phi\cos\theta, -\sin\theta, -\sin\phi\cos\theta]. \tag{14}$$

For each occluder point $\mathbf{x}_i$, we cast a ray along $\mathbf{d} = -\boldsymbol{\ell}$ and intersect it with the receiver plane:

$$\mathbf{y}_i = \mathbf{x}_i + t_i \mathbf{d}, \quad t_i = \frac{(\bar{\mathbf{r}} - \mathbf{x}_i) \cdot \mathbf{n}}{\mathbf{d} \cdot \mathbf{n}}. \tag{15}$$



Figure 11: **Invalid approximate light.** Examples rejected by our manual check. Here, the approximate light either yields cast shadows that noticeably deviate from the observed shadow region (in location or shape), or the observed shadow is too partial to support a reliable elevation estimate. Such cases are discarded from the supervision.

Rays with $|\mathbf{d} \cdot \mathbf{n}|$ near zero or $t_i \leq 0$ are treated as invalid. Valid intersections are projected to 2D receiver-grid coordinates, denoted by $\mathbf{q}_i$.

Shadow density rendering. The cast points are rendered onto a fixed grid using Gaussian splatting:

$$D_{\boldsymbol{\ell}}(\mathbf{p}) = \sum_i w_i \, G_\sigma(\mathbf{p} - \mathbf{q}_i), \tag{16}$$

producing a soft shadow density map. The density is then converted into a soft binary shadow estimate by applying a percentile-based threshold τ and logistic flattening:

$$R_{\boldsymbol{\ell}} = \sigma(\gamma(D_{\boldsymbol{\ell}} - \tau)), \tag{17}$$

where γ controls the transition sharpness. The observed shadow mask is processed in the same way to obtain a target map R_S.

Light optimization. We estimate the light direction by minimizing the discrepancy between rendered and observed shadows using a soft Dice loss together with an out-of-bounds penalty $\mathcal{L}_{oob}$ that discourages cast points from falling outside the receiver ROI:

$$\mathcal{L} = 1 - \frac{2\langle R_{\boldsymbol{\ell}}, R_S \rangle}{\|R_{\boldsymbol{\ell}}\|_1 + \|R_S\|_1 + \varepsilon} + \lambda_{oob}\mathcal{L}_{oob}. \tag{18}$$

We first perform a coarse grid search over (ϕ, θ) and then refine the best candidates with gradient-based optimization. The final estimate

$$\hat{\boldsymbol{\ell}} = \boldsymbol{\ell}(\hat{\phi}, \hat{\theta}) \tag{19}$$

is used as pseudo supervision for training the light predictor.

Manual verification. After automatic estimation, we visualize the rendered shadow density and compare it with the observed shadow mask (Figure 10). Estimates that do not adequately explain the observed shadow location and shape are rejected, as are cases with observed shadows too partial for reliable estimation (Figure 11). After filtering, we retain valid directions for 17,104 of the 21,575 scenes, covering 22,364 image tuples: 21,754 training tuples and 610 test tuples. Training-split labels are used only to supervise light-related losses, while test-split labels are used only for evaluation and appendix diagnostics. The pipeline runs offline; standard inference requires no pseudo-labels.

Table 7: Performance on the 140 DESOBAV2 test tuples without valid light-direction pseudo-labels. GR/LR, GS/LS, and GB/LB denote global/local RMSE, SSIM, and BER, respectively. All results use best-of-five selection. Best scores are in **bold**.

Method	GR↓	LR↓	GS↑	LS↑	GB↓	LB↓
GPSDiffusion (Zhao et al., 2025)	6.16	39.50	0.9692	0.5029	0.1513	0.2969
Ours	**4.28**	**33.83**	**0.9749**	**0.5637**	**0.1261**	**0.2497**

Evaluation without valid pseudo-labels. The remaining 140 test tuples have no valid light-direction pseudo-label. We use label availability only to define a diagnostic subset and evaluate our method and GPSDiffusion (Zhao et al., 2025) on these tuples under the same best-of-five protocol. As shown in Table 7, our method outperforms GPSDiffusion across all six metrics. This indicates that the gains are not confined to cases for which the offline pipeline can recover a valid light direction.

H More Qualitative Results

Figures 12 and 13 provide additional qualitative comparisons in the BOS-free and BOS settings, respectively. For each scene, we show the input image and foreground object mask, the generated images and predicted shadow masks from SGDiffusion (Liu et al., 2024), GPSDiffusion (Zhao et al., 2025), and our method, and the corresponding ground truth.

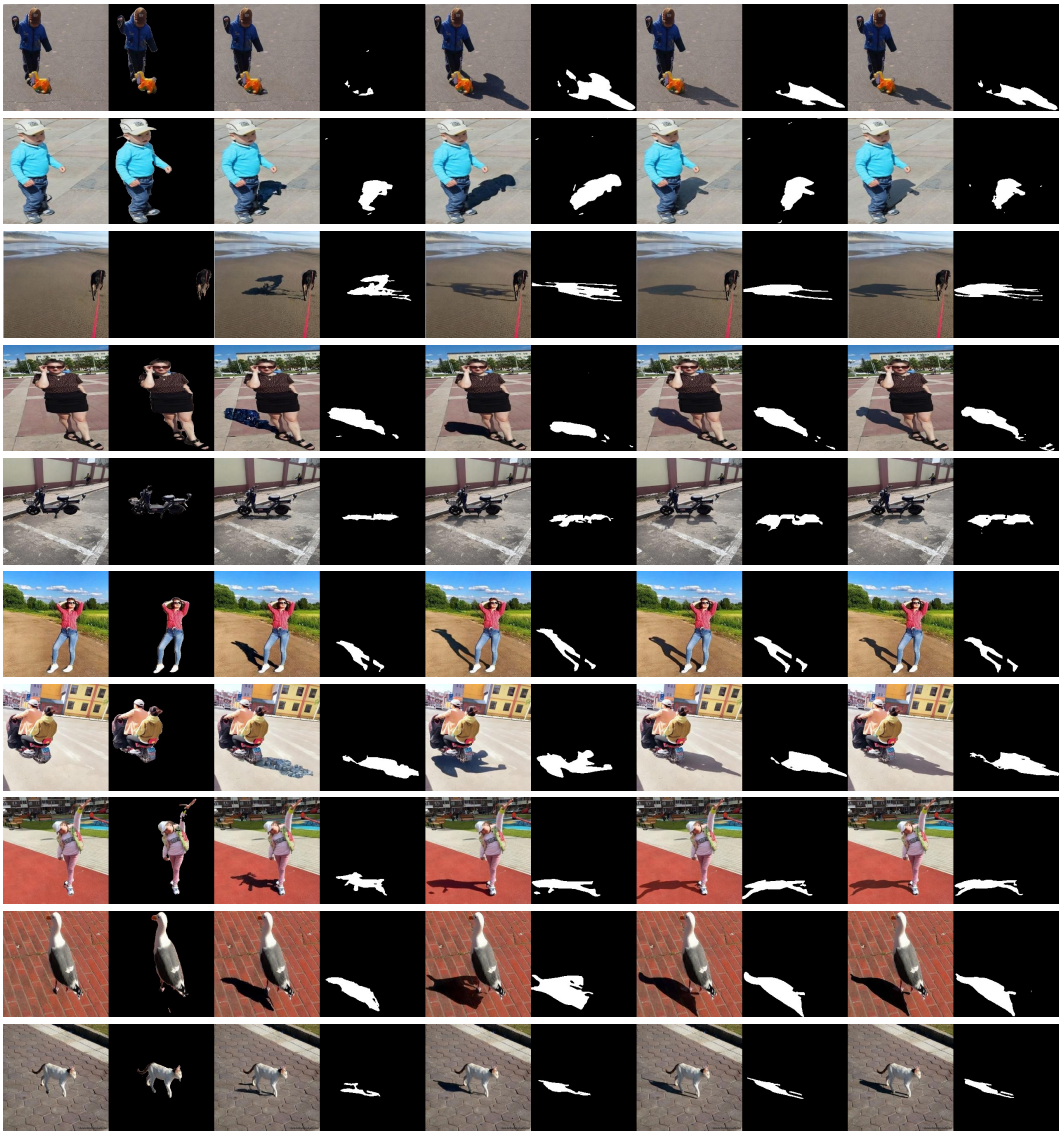


Figure 12: **Additional qualitative results in the BOS-free setting.** We compare the generated images and predicted shadow masks from SGDiffusion (Liu et al., 2024), GPSDiffusion (Zhao et al., 2025), and our method with the ground truth. Across these examples, our method produces shadows with more accurate placement, shape, and extent.

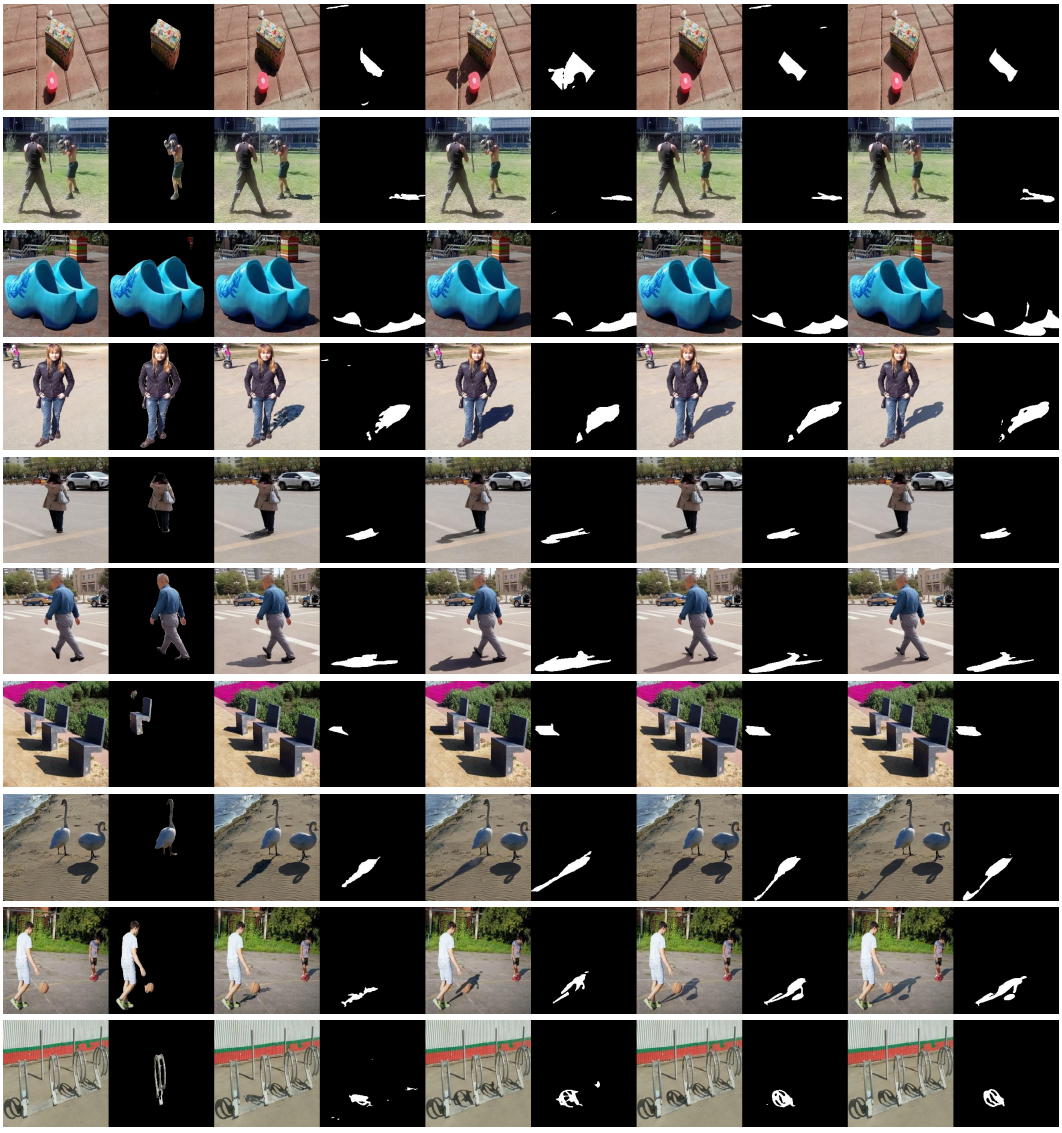


Figure 13: **Additional qualitative results in the BOS setting.** We compare the generated images and predicted shadow masks from SGDiffusion (Liu et al., 2024), GPSDiffusion (Zhao et al., 2025), and our method with the ground truth. Across these examples, our method produces shadows with more accurate placement, shape, and extent.