Parametric Shadow Control for Portrait Generation in Text-to-Image Diffusion Models

Haoming Cai[1], Tsung-Wei Huang[2], Shiv Gehlot[2],
Brandon Y. Feng[3], Sachin Shah[1], Guan-Ming Su[2], Christopher Metzler[1]
[1]University of Maryland [2]Dolby Labs [3]MIT

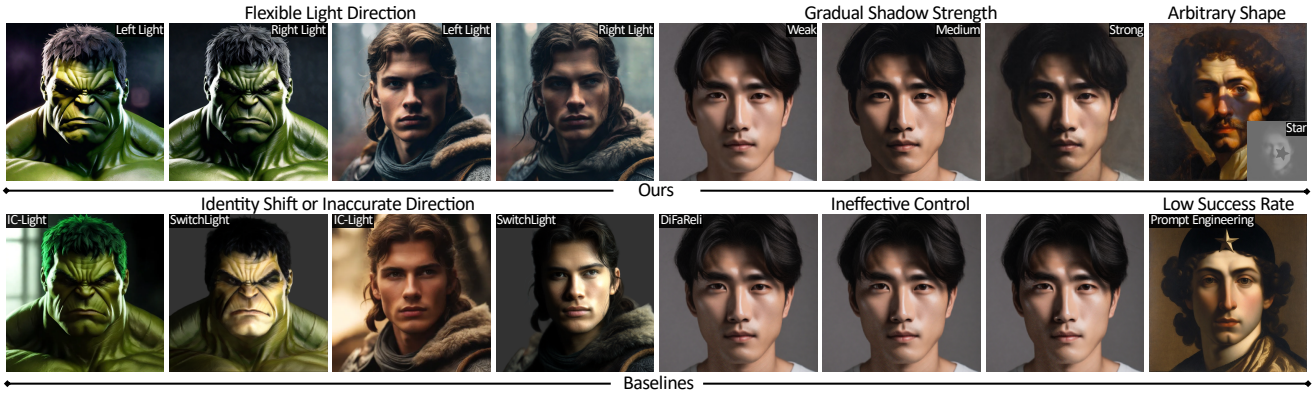

Figure 1. **Shadow manipulation on generated portrait images with diverse styles challenges existing editing methods.** To address this limitation, we present Shadow Director, which enables intuitive and parametric shadow control during diffusion-based portrait generation, instead of post-processing, across diverse artistic styles. Top: Our method provides parametric control over (left) directional light position, (middle) progressive shadow strength, and (right) arbitrary shadow shapes, while maintaining identity and style consistency. Bottom: Comparison with baseline methods reveals their limitations: SwitchLight [24] and IC-Light [60] suffer from identity shifts or inaccurate lighting direction despite requiring substantial training resources, DiFaReli [35] fails at effectively providing strong variation in shadow strength. Prompt engineering results in unpredictable outcomes.

Abstract

Text-to-image diffusion models excel at generating diverse portraits, but lack intuitive shadow control. Existing editing approaches, as post-processing, struggle to offer effective manipulation across diverse styles. Additionally, these methods either rely on expensive real-world light-stage data collection or require extensive computational resources for training. To address these limitations, we introduce Shadow Director, a method that extracts and manipulates hidden shadow attributes within well-trained diffusion models. Our approach uses a small estimation network that requires only a few thousand synthetic images and hours of training—no costly real-world light-stage data needed. Shadow Director enables parametric and intuitive control over shadow shape, placement, and intensity during portrait generation while preserving artistic integrity and identity across diverse styles. Despite training only on synthetic data built on real-world identities, it generalizes effectively to generated portraits with diverse styles, making it a more accessible and resource-friendly solution. Homepage: https://www.hm-cai.com/ShadowDirector/

1. Introduction

Control over visual attributes in AI-generated imagery has advanced significantly [30, 58], yet shadow control for portraits remains unexplored in diffusion-based generation. While current diffusion models excel at generating portraits with consistent lighting based on text prompts, prompt engineering lacks intuitive controls for shadow effects, forcing digital artists to rely on time-consuming trial-and-error prompt engineering.

1.1. Generated Portraits Challenge Editing Method

Conventional editing-based methods struggle with shadow manipulation in generated images across diverse artistic styles. As shown in Figure 1, these methods not only fail

to achieve effective shadow control but also struggle to preserve non-realistic artistic elements such as unique tones in stylized artwork, vibrant color palettes, and subject identity.

This limitation partially arises from restricted diversity and quality in training data. Real-world OLAT (one-light-at-a-time) data by light stage system is prohibitively expensive to collect and inherently limited in diversity. Real-world OLAT datasets primarily capture a small group of real-world subjects, restricting variations in skin tones and facial features while lacking the necessary variety to handle artistic styles. Synthetic data, on the other hand, typically lacks the quality compared to real-world OLAT data.

These limitations in training data have led researchers to incorporate diffusion models with their rich prior information, as seen in methods like DiFaReli [35] and IC-Light [60]. DiFaReli attempts to fine-tune diffusion models on synthetic data but fails to unlock the models' full capabilities for diverse artistic styles. IC-Light achieves better performance by scaling up to 10 million training examples (combining real and synthetic data), but struggles with identity preservation and requires substantial computational resources (8 H100 GPUs for a week), making its training pipeline impractical to many researchers. Table 1 summarizes the key aspects and computational resources required for these editing methods.

These challenges motivate us to ask a **question**: *Can we achieve effective shadow manipulation on generated portraits with diverse styles using few synthetic data and computational resources?* Recent studies have found that well-trained diffusion models implicitly encode rich information about various visual attributes [12] like depth, normal [14], Shading [26], and human pose [30]. This suggests that shadow information for portraits might similarly be embedded within these models, raising a **further question**: *is there an efficient way to access and utilize this shadow-related prior knowledge without extensive retraining?*

1.2. Key Technical Insight and Approach

To address the question of efficiently accessing shadow information in diffusion models, we present Shadow Director: a method for extracting and manipulating shadow information embedded within diffusion models during generation.

Our approach is based on the hypothesis that well-trained diffusion models already encode shadow information within their intermediate latent features. Although this information exists implicitly, we found that a trained shadow estimation network can effectively access it. Specifically, we train a small shadow estimation network that extracts shadow maps from the diffusion model's noisy intermediate features during denoising. Shadow control is implemented through an optimization process during image generation: we calculate the difference between the current estimated shadow (from noisy latent features) and our target shadow, then optimize the latent features to achieve the desired shadow while preserving identity.

	SwitchLight[24]	IC-Light[60]	DiFaReli[35]	Ours
Type	editing	editing	editing	conditional generation
Philosophy	neural rendering	scaled dataset	diffusion model	attribute reveal
Synthetic Data involved	✓	✓	✓	✓
NO Real-world OLAT			✓	✓
Diffusion Model		✓	✓	✓
Data Quantities (k)	30+	10,000+	60	5.7
GPU	32-A6000	8-H100	1-V100s	2-A6000
Training Time	1 week	140 hours	8 days	8 hours

Table 1. Comparison with accessible shadow manipulation methods. Key advantages of our approach: (1) Better adaptation to generated images through conditional generation instead of post-processing, (2) No real-world OLAT data required, and (3) Minimal data quantity needed (only 5.7k samples). Such resource-efficient design proves that our design based on attribute reveal is promising. It enables researchers with limited data and computing resources to more easily develop better shadow manipulation methods in the future.

Unlike editing methods that work as post-processing, Shadow Director enables parametric and intuitive shadow control directly during the portrait generation process, preserving artistic integrity across diverse styles. Our experiments show that Shadow Director effectively extracts shadow information from diffusion models while requiring only a few thousand synthetic examples of limited quality and just two A6000 GPUs with a few hours of training. This resource-efficient design philosophy demonstrates that leveraging information already present in diffusion models can significantly reduce dependence on expensive light-stage data collection, making it feasible for researchers with limited data and computing resources to develop better shadow manipulation methods.

In summary, our contributions include:

- We propose Shadow Director, a novel approach that accesses and manipulates implicit shadow information embedded in the latent space during diffusion model denoising, fundamentally different from editing methods.
- Our method achieves effective shadow control using only a small synthetic dataset and a few hours of training, demonstrating remarkable data efficiency without depending on costly light-stage data.
- Shadow Director enables intuitive and parametric control over multiple shadow attributes while preserving identity. Despite training only on synthetic data built on real-world identities, it generalizes to diverse artistic styles where editing methods fail.

2. Related Work

2.1. Portrait Relighting via Neural Networks

Portrait relighting has been extensively explored in 2D [18, 19, 22, 24, 33, 34, 38, 41, 42, 51, 52, 56, 59, 61], and 3D [4, 5, 32, 37, 44, 46, 51, 62], often relying on inverse

rendering [3, 39] or light-stage data [10]. Physics-based methods [24, 33] decompose albedo, normal, and shading but fail to capture complex effects like subsurface scattering [11, 25, 31]. Diffusion-based relighting approaches [6, 20, 22, 38, 56, 59] use HDR priors or synthetic datasets but remain constrained by domain gaps [42, 43, 55]. Models like DiffRelight [15] require per-subject fine-tuning for 3D. While SwitchLight [24], IC-Light [59], and Relightful Harmonization [38] mitigate gaps through real-image pretraining, pseudo-labeling, and large-scale augmentation.

2.2. Diffusion Models for Lighting Manipulation

With the rapid advancement of diffusion models, a wide range of image and video tasks have seen significant progress [8, 9, 17, 21, 40, 49, 50]. While primarily designed for generative tasks, diffusion models implicitly encode a wealth of structural and semantic information within their intermediate representations[29, 30, 47, 57], which can be leveraged for downstream applications. Moreover, numerous studies have shown that diffusion models exhibit strong capabilities in estimating and recovering intrinsic attributes such as depth [23], surface normals[14], and 3D structure[27], which are valuable for lighting manipulation. Recently, researchers have begun exploring how lighting-related information embedded within diffusion models can be utilized [2, 12]. Works such as [26, 53] have demonstrated the feasibility of training diffusion models for illumination manipulation in both indoor and outdoor scenes. These studies collectively suggest that well-trained diffusion models inherently encode shadow information. This raises the possibility of accessing and utilizing this hidden shadow information to enable shadow control for generated images across diverse artistic styles.

3. Method

The key challenge in our approach is accessing and manipulating shadow information embedded within diffusion models during the denoising process, while preserving identity and artistic style. Our solution, Shadow Director, extracts shadow information from UNet's features and enables intuitive user control through test-time optimization on diffusion model's latent feature maps. What distinguishes our approach is the ability to reveal and manipulate implicit shadow information during generation with minimal training resources. Sec. 3.1 details the design of Shadow Director for accessing latent shadow information. Sec. 3.2 explains our synthetic training dataset.

3.1. Shadow Director

Module Overview. As shown in Figure 2, Shadow Director comprises two key components: a Shadow-Depth (SD) Estimator and an Identity (ID) Estimator. Both operate on the same UNet's feature maps during denoising. The SD Estimator extracts 2D shadow and depth maps to enable shadow manipulation, while the ID Estimator produces feature embeddings (ID embeddings) to enable identity preservation. Both estimators are implemented as compact neural networks, with architecture details in Appendix B.

Training of Shadow Director. Both estimators are trained independently on our synthetic dataset. For the SD Estimator, at each training iteration, we generate a noisy latent feature by adding random noise to the clean latent feature of a relit image. This noisy latent feature is fed into the UNet, and internal features are extracted for shadow and depth estimation. The SD Estimator is trained to minimize $\mathcal{L}_{\text{SD}} = \mathcal{L}_{\text{L1}}(S_{\text{pred}}, S_{\text{gt}}) + \mathcal{L}_{\text{L1}}(D_{\text{pred}}, D_{\text{gt}})$, where S and D represent shadow and depth maps respectively. For the ID Estimator, each training iteration requires three images: an original image with its original lighting, a same-identity image with different lighting, and a different-identity image. Similar to SD-Estimator training, we generate noisy latent features by adding random noise to the clean latent features of these three images. Their corresponding ID embeddings are then extracted independently by the ID-Estimator. We apply a triplet loss (details in Appendix Fig.13) [7, 13, 30] to enforce that embeddings of the same identity become closer while embeddings of different identities are pushed apart. This training approach enables the ID Estimator to maintain consistent identity representations despite lighting variations, ensuring our Shadow Director preserves the subject's identity. We train this ID-Estimator on our synthetic dataset.

Enable Shadow Control. Shadow control is achieved through test-time optimization at a selected early denoising step. We pause denoising and optimize only the UNet's input tensor (noisy latent feature, fire icon in Fig. 2). Before optimization, a single forward pass extracts (i) reference identity embeddings I_{ref}. Also, (ii) the user-defined target shadow map S_{target}—via a binary mask (Fig. 3a) or depth-based ray casting (Fig. 3b). Users can adjust three aspects of the shadow: intensity (via latent optimization strength), shape & placement (via mask), and light position (via ray casting).

Optimization Details. During latent optimization, only the UNet's input tensor x_e (noisy latent feature map) is optimizable; all network parameters remain frozen. At the selected early denoising step, we first extract the internal feature

$$h = fetch(\text{UNet}(x_e)),$$

and then obtain the shadow and identity estimates

$$S = \text{SD}(h), \quad I = \text{ID}(h).$$

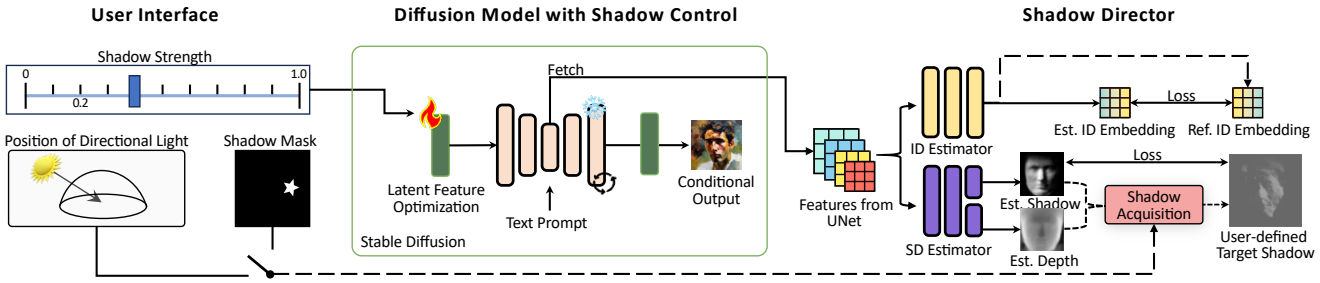

Figure 2. **Shadow control pipeline for image generation.** Our approach consists of three main components: (1) *User Interface* (left), which provides intuitive controls for shadow strength, directional light position, and shadow shape; (2) *Diffusion Model with Shadow Control* (middle), where latent features are optimized at selected denoising steps; and (3) *Shadow Director* (right), which extracts shadow and identity attributes from UNet internal features using two estimators. Shadow Director is trained to infer these attributes from Unet's noisy feature maps. Shadow control is achieved through test-time optimization on the noisy latent features (the U-Net input marked with a fire symbol). By applying this optimization at an early denoising step, we gain greater freedom in manipulating shadows during image generation. Before latent optimization begins, both estimators perform an initial forward pass once (dashed lines) to obtain the user-defined target shadow and reference identity embedding. The shadow acquisition process is detailed in Fig. 3. During latent optimization, the noisy latent features are guided to match the user-defined target shadow while minimizing changes to the identity embeddings. Notably, the only optimizable component in the pipeline is the noisy latent feature at the selected denoising step. Further architectural details of the estimators and feature extraction are provided in Appendix B.

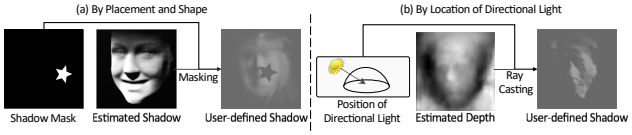

Figure 3. **Two customization options for shadow maps, applied before latent optimization. (a)** Shadow placement and shape: A user-defined binary mask is applied to the estimated shadow map. Masked regions become darker in the customized shadow, explicitly defining shadow areas. **(b)** Shadow synthesis through directional lighting: Using the estimated depth map and user-specified directional light position, we implement ray casting to generate geometrically consistent shadows. Users select only one of these two methods. Detail of ray casting is presented in Appendix C.1.

We perform few iterations of gradient descent on x_e: at each iteration, we compute

$$\mathcal{L}_{\text{shadow}} = \|S - S_{\text{target}}\|_1, \quad \mathcal{L}_{\text{identity}} = 1 - \cos(I, I_{\text{ref}}),$$

and update

$$x_e \leftarrow x_e - \alpha \nabla_{x_e} \left(\lambda_{\text{shadow}} \mathcal{L}_{\text{shadow}} + \lambda_{\text{identity}} \mathcal{L}_{\text{identity}}\right).$$

After latent optimization, we resume denoising to produce the final image with the desired shadow. Notably, we never decode or use intermediate noisy RGB images via the VAE; all optimization remains in the latent feature space.

3.2. Synthetic Training Dataset

As shown in Figure 4, our synthetic dataset makes Shadow Director a practical and accessible solution for broader researchers. We construct this dataset using the CelebA

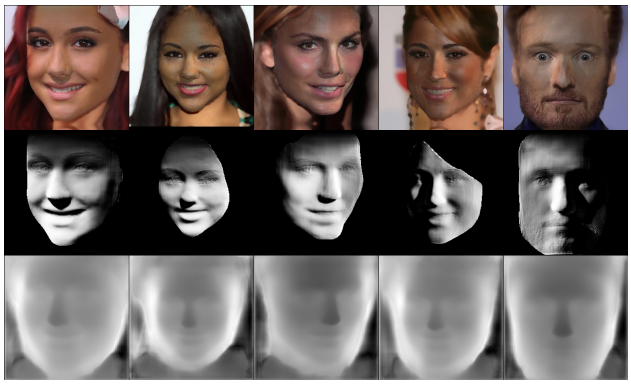

Figure 4. **Samples of Our Synthetic Dataset.** Data pairs: relit images (top row), shadow maps (middle row), and depth maps (bottom row). Unlike IC-Light requiring 10 million samples, our approach needs only a few thousand paired synthetic examples. These are generated using GeomConsistentFR [19]. Despite limited photorealism in shadow, this dataset proves sufficient for accessing shadow information embedded within diffusion models. This demonstrates a promising research direction: reducing dependency on expensive light-stage data while focusing on revealing lighting information already hidden in diffusion models.

dataset [28], generating six lighting variations per identity with GeomConsistentFR [19]. Specifically, for each subject, we randomly sample six light source positions within a circular frontal region of the face, with each light directed towards the facial center, and synthesize the corresponding relit images. This design enables training our ID-Estimator to focus solely on identity features while disregarding lighting variations. Since GeomConsistentFR itself is trained

Figure 5. **Shadow intensity control on generated portraits across diverse styles.** The top two rows demonstrate gradual shadow intensity control, while the bottom four rows highlight strong and weak shadow variations for better visual comparison. Shadow Director enables parametric control over shadow strength, ranging from weak to strong, while preserving both identity and artistic integrity.

on synthetic data, the generated relit images exhibit limited photorealism. However, our experiments show that this dataset is sufficient for Shadow Director to extract and utilize the rich shadow information already encoded in diffusion models. In this work, we focus on maximizing the potential of limited but accessible synthetic data in conjunction with a well-trained diffusion model.

4. Experiment

Additional qualitative results demonstrating Shadow Director's control are provided in Appendix D. Detailed implementation is provided in Appendix C.

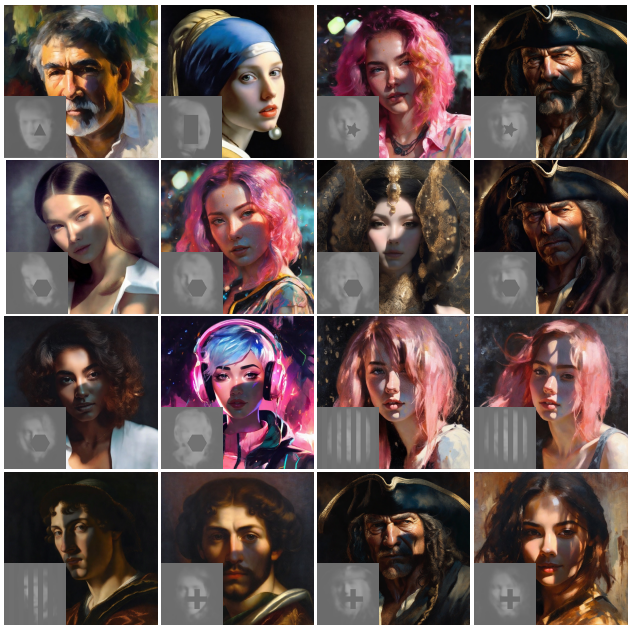

Figure 6. **Shadow shape control with user-defined masks on diverse portrait images.** Shadow Director enables precise control over shadow shapes and placement using user-defined masks (shown as gray overlays).

4.1. Settings of Training and Inference

We implement our pipeline based on Stable Diffusion XL (SDXL) 1.0. For the synthetic dataset, we randomly select 1,000 identities from the CelebA dataset, rendering each with 6 different lighting conditions using GeomConstantFR. This generates a total of 6,000 synthetic images, which we split into 5,700 for training and 300 for validation to evaluate the SD Estimator's prediction accuracy. Both the SD Estimator and ID Estimator are trained on a single NVIDIA A6000 GPU, with each estimator taking approximately 8 hours to train. We use the Adam optimizer with a learning rate of 1e-4 and a batch size of 8. For inference-time shadow control, we begin latent feature optimization at denoising step 40 out of 100 (corresponding to t=0.6). We use the Adam optimizer with a learning rate of 2e-4 for latent feature optimization. Our method adds only about 20 percent additional computation time compared to standard SDXL inference.

4.2. Controlling Shadow Strength

Our approach enables users to gradually adjust shadow intensity while preserving surrounding lighting conditions. Unlike text-prompt engineering, which lacks precise control over intensity changes, our provides fine-intuitive control over shadow strength. As shown in Fig.5, our method achieves modulation of shadow strength through latent optimization during inference time. Shadow intensity correlates

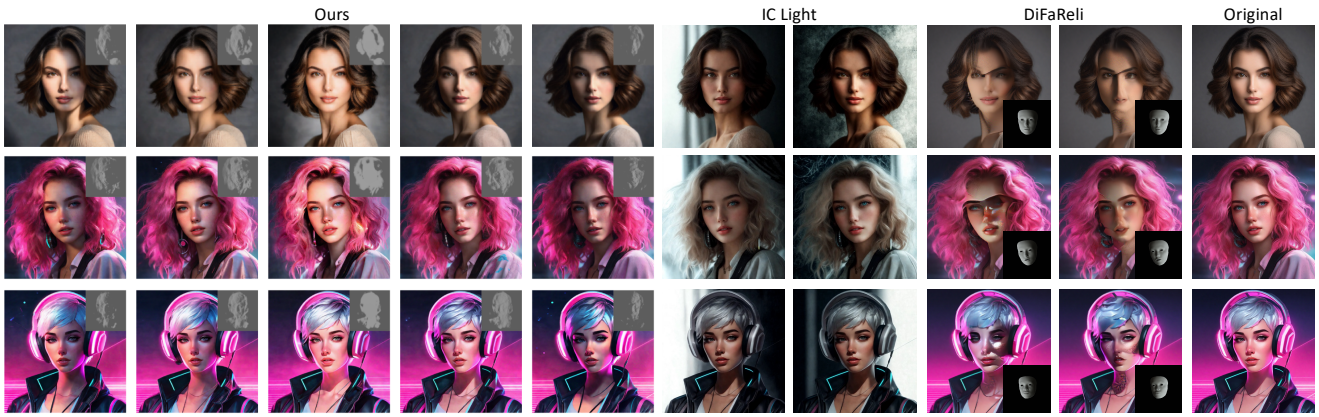

Figure 7. **Shadow Synthesis by Controlling Lighting Direction: Comparison with Diffusion-Based Methods.** Our approach (leftmost 5 columns) consistently preserves identity and artistic style while providing varied shadow control based on specified lighting positions. In contrast, diffusion-based editing methods IC-Light (columns 6-7) and DiFaReli (columns 8-9) show significant identity shifts and style inconsistencies compared to the original portraits (rightmost column). Results are shown across three different portrait styles: realistic (top row), super-realistic (middle row), and cartoon (bottom row). Our method maintains vibrant colors, facial features, and artistic elements while providing intuitive shadow control by directional lighting position. Please see Appendix D for more lighting positions.

directly with the number of optimization iterations applied, allowing for gradual adjustments based on user preference.

4.3. Controlling Shadow Placement and Shape

Our method offers intuitive control over shadow placement and shape through user-defined shadow mask. In contrast to existing methods that generate unrealistic blended synthetic shadows, our approach ensures natural shadow placement and integration. Results in Fig.10 demonstrate controlled shadow placement using custom shadow masks.

4.4. Controlling Directional Light Position

Our method allows users to specify the 3D position of the directional light source for portrait relighting. As shown in Fig. 7, Shadow Director generates geometrically consistent shadows while preserving identity and artistic style. In contrast, other diffusion-based methods such as IC-Light and DiFaReli struggle with effective shadow control and often introduce identity shifts.

To demonstrate our method's performance across a broad range of portraits, we randomly generate 800 portraits using text prompts (detailed in Appendix E) without specifying lighting conditions. For each portrait, all methods generate two relit images under left and right directional lighting correspondingly.

For qualitative evaluation, we present results in Fig. 8. Comparisons are made against IC-Light [60], Switch-Light [24], Hou et al. [19], and DiFaReli [35]. For quantitative evaluation, we partially follow the protocols in [6] and [38], assessing both objective metrics (Table 2) and user studies (Table 3). Since ground-truth relit images are unavailable, non-reference image quality assessment methods are used to measure realism and image fidelity. CLIP-IQA+[48], NIMA[45], and ARNIQA [1] assess image authenticity to ensure that shadow manipulation does not degrade perceptual quality. Following [6], CLIP Vision Score (CVS) is used to measure identity preservation through cosine similarity between the clip embedding [36] of source and relit images. CLIP Vision-Text Alignment (CVTA) evaluates how well relit images align with the augmented text prompts specifying "left light" or "right light." To further validate the results, we conduct user studies (Table 3) to assess shadow consistency and identity preservation through subjective evaluation. 40 users and 20 samples involved. For each sample, order of methods is random.

4.5. Ablation Study

We further investigate the necessity of each component in Shadow Director through a series of ablation studies. These studies are summarized in Table 4, with corresponding quantitative results presented in Table 5.

Optimal Denoising Step for Shadow Control. We investigate whether shadow control is most effective during early/middle denoising steps, later steps, or only at the final step. Our experiments confirm that applying shadow control during early and middle denoising steps (t=0.5-0.7) provides the optimal balance between shadow manipulation flexibility and image quality preservation, outperforming later-stage or final-step interventions.

Input of Shadow Estimator. We examine which feature source provides the most effective shadow information by comparing three options: UNet input, UNet internal features, and UNet output. Our experiments confirm that UNet

Figure 8. **Shadow Synthesis by Controlling Lighting Direction.** Comparison of portrait relighting across different editing methods generated portrait. Each row corresponds to a different method, while each column maintains left lighting direction for easy comparison.

Table 2. Quantitative results on 800 unseen portrait image generated from text prompts.

Metric	Ori	Hou 2022 [19]	DiFaReLi [35]	IC-Light [60]	SwitchLight [24]	Ours
CVTA ↑	0.3307	0.2629	0.2466	0.2606	0.2520	**0.2692**
CVS ↑	1.0000	0.9386	0.9049	0.9140	0.8945	**0.9460**
CLIPIQA+ ↑	0.7366	0.3747	0.3906	0.4185	0.3876	**0.4484**
NIMA ↑	6.3738	5.0195	5.2370	5.7193	5.3219	**5.7735**
ARNIQA ↑	0.6895	0.4603	0.4831	0.5236	0.4442	**0.5480**

Table 3. User study for preference. 2.44% selected 'None'

	Ours	IC-Light	SwitchLight	DiFaReli	Hou
Preference Rate	45.69%	27.44%	20.46%	1.86%	2.09%

internal features yield better performance due to their richer information. Consistent with findings in recent studies [30]

Latent-space vs. Noisy RGB-space Estimation. This study compares our latent-space shadow estimation approach against an alternative RGB-space pipeline. The alternative first generates noisy RGB images from the UNet's output at early denoising steps, then applies an RGB-based shadow estimator (like our data generation method Geom-ConstantFR) to obtain noisy shadow and depth maps. After acquiring the user-defined shadow through a similar process, the loss between this target and the estimated noisy shadow in RGB space is backpropagated through both the RGB estimator and UNet to optimize the latent feature map. Although conceptually similar to our approach, this RGB-space alternative introduces reliability issues during latent optimization. Our experiments demonstrate that latent-space estimation achieves better shadow quality and identity preservation, confirming our design choice to operate directly on UNet's internal feature maps rather than noisy RGB outputs from denoising step.

Necessity of ID Embedding for ID Preservation. We demonstrate why our dedicated ID embedding approach is superior to the simpler alternative of applying L1 loss directly on latent features. In this alternative approach,

Table 4. **Necessity of Design in Shadow Director.**

	Feature Extraction Location		Denoising Step for Latent Optim			Constraint Choice	
	Unet's Internal	Unet's Output	Early and Middle	Latter	Last One	SD	ID
(a) Ours	✓	-	✓	-	-	SD-Estimator	ID-Estimator
Ablation Study 1: Optimal Denoising Step for Shadow Control							
(b)	✓	-	-	✓	-	SD-Estimator	ID-Estimator
(c)	✓	-	-	-	✓	SD-Estimator	ID-Estimator
Ablation Study 2: Input of Shadow Estimator							
(d)	-	✓	✓	-	-	SD-Estimator	ID-Estimator
Ablation Study 3: Latent-space vs. Noisy RGB-space Estimation							
(e)	✓ (for Ours-ID)	✓ (for RGB-SD)	-	-	-	RGB-Estimator	ID-Estimator
Ablation Study 4: Necessity of ID Embedding for ID Preservation							
(f)	✓	-	-	-	-	SD-Estimator	L1 Loss
(g)	✓ (for Ours-SD)	✓ (for L1-ID)	-	-	-	SD-Estimator	L1 Loss

Table 5. **Necessity of Design in Shadow Director.**

Metric	**(a) Ours**	Method **(b)**	Method **(c)**	Method **(d)**	Method **(e)**	Method **(f)**	Method **(g)**
CVTA ↑	**0.2692**	0.2410	0.2356	0.2197	0.1975	0.1556	0.1485
CVS ↑	**0.9460**	0.9325	0.9320	0.8945	0.8173	0.7685	0.7679
CLIPIQA+ ↑	**0.4484**	0.4325	0.4392	0.3639	0.3220	0.2873	0.2682

the SD-Estimator remains, but identity preservation is attempted by constraining the optimized UNet input/output features to remain close to their original values via L1 distance. Unlike this direct feature constraint, our ID embedding network specifically captures meaningful identity characteristics while allowing shadow-relevant features to change. Our experiments confirm that specialized identity embeddings are essential for maintaining consistent identity during shadow manipulation, as direct feature constraints either limit shadow flexibility or fail to preserve key identity elements.

Necessity of ID-Estimator. Without the ID Estimator to constrain latent optimization, the generated image tends to overemphasize shadow optimization, often resulting in unexpected textures and identity distortions. As shown in Fig.9, optimizing for strong shadows without identity preservation loss leads to significant unintended artifacts.

5. Discussion

Relation to Shadow Removal. Shadow Director does not perform shadow removal on the original image, as it does not reconstruct occluded details. Instead, it is a conditional generation framework that guides the diffusion model to generate or suppress shadows in specific region during synthesis. Unlike shadow removal methods that recover lost information, Shadow Director conditions the model to produce lighting-consistent images. For example, in the fifth column of Fig. 8, the left cheek shadow disappears not due to removal but because the model, conditioned on new lighting, generates an image where the shadow no longer forms. This distinction highlights Shadow Director's role in controlling shadow generation rather than removing it.

Occlusion handling and object relighting? Yes. Our method handles occlusions (hair, hats, glasses) without retraining. It also generalizes to simple object relighting without retraining, though complex materials and geometry remain challenging. For full details and results, please check Appendix D.1 and D.2.

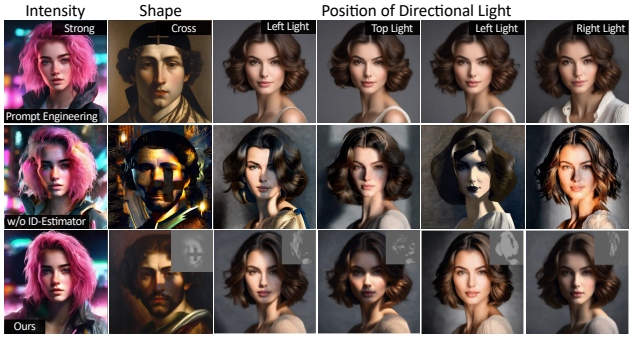

Figure 9. **Necessity of our ID-Estimator and failure of naive prompt engineering.** The first row shows Prompt Engineering fails to produce desired shadows. The second row highlights artifacts and identity inconsistencies without our ID-Estimator. The third row (Ours) demonstrates that our full method achieves effective shadow manipulation while preserving identity and style.

Figure 10. **Limitations and Future Work.** (a) Lighting effects in prompts (neon lighting) may override Shadow Director's control. (b) Training on CelebA, which consists mostly of simple, straight-on portraits, makes it difficult to preserve intricate clothing details in more complex outfits. (c) The model does not handle lighting temperature, which may lead to color tone inconsistencies.

Limitation. Shadow Director is trained on CelebA, which mainly consists of close-up, straight-on portraits with simple clothing, making intricate garment preservation in full-body images challenging. Expanding dataset diversity or constraining latent optimization could help mitigate this. Additionally, strong lighting effects in text prompts, such as neon lighting, may override Shadow Director's control, which could be improved through refined hyperparameter tuning. The model also lacks explicit lighting temperature control, sometimes causing color tone inconsistencies. Despite these challenges, Shadow Director demonstrates that effective shadow control can be achieved with accessible training resources, paving the way for improvements with broader datasets and enhanced lighting models.

6. Conclusion

We introduce Shadow Director, a diffusion-model control framework for shadow manipulation in portrait generation. By leveraging latent shadow information, our method enables parametric control while preserving identity across diverse styles. Shadow Director achieves this with minimal training data and computational resources, demonstrating a resource-efficient direction for shadow control in diffusion models.

Acknowledgment. H.C. and C.A.M. were supported in part by gift funds from Dolby and a UMD Grand Challenges Seed Grant.

References

[1] Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. Arniqa: Learning distortion manifold for image quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 189–198, 2024. 6

[2] Hadi Alzayer, Philipp Henzler, Jonathan T Barron, Jia-Bin Huang, Pratul P Srinivasan, and Dor Verbin. Generative multiview relighting for 3d reconstruction under extreme illumination variation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10933–10942, 2025. 3

[3] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1670–1687, 2014. 3

[4] Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn Mcphail, Ravi Ramamoorthi, Yaser Sheikh, and Jason Saragih. Deep relightable appearance models for animatable faces. *ACM Transactions on Graphics (ToG)*, 40(4):1–15, 2021. 2

[5] Ziqi Cai, Kaiwen Jiang, Shu-Yu Chen, Yu-Kun Lai, Hongbo Fu, Boxin Shi, and Lin Gao. Real-time 3d-aware portrait video relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6221–6231, 2024. 2

[6] Junuk Cha, Mengwei Ren, Krishna Kumar Singh, He Zhang, Yannick Hold-Geoffroy, Seunghyun Yoon, HyunJoon Jung, Jae Shin Yoon, and Seungryul Baek. Text2relight: Creative portrait relighting with text guidance. *arXiv preprint arXiv:2412.13734*, 2024. 3, 6

[7] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11(3), 2010. 3

[8] Haoyu Chen, Xiaojie Xu, Wenbo Li, Jingjing Ren, Tian Ye, Songhua Liu, Ying-Cong Chen, Lei Zhu, and Xinchao Wang. Posta. 3

[9] Jingxi Chen, Brandon Y Feng, Haoming Cai, Tianfu Wang, Levi Burner, Dehao Yuan, Cornelia Fermuller, Christopher A Metzler, and Yiannis Aloimonos. Repurposing pre-trained video diffusion models for event-based video interpolation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12456–12466, 2025. 3

[10] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 145–156, 2000. 3

[11] Craig Donner and Henrik Wann Jensen. A spectral bssrdf for shading human skin. *Rendering techniques*, 2006:409–418, 2006. 3

[12] Xiaodan Du, Nicholas Kolkin, Greg Shakhnarovich, and Anand Bhattad. Generative models: What do they know? do they know things? let's find out! *arXiv preprint arXiv:2311.17137*, 2023. 2, 3

[13] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. 3

[14] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pages 241–258. Springer, 2024. 2, 3

[15] Mingming He, Pascal Clausen, Ahmet Levent Taşel, Li Ma, Oliver Pilarski, Wenqi Xian, Laszlo Rikker, Xueming Yu, Ryan Burgert, Ning Yu, et al. Diffrelight: Diffusion-based facial performance relighting. *arXiv preprint arXiv:2410.08188*, 2024. 3

[16] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4775–4785, 2024. 4

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[18] Andrew Hou, Ze Zhang, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Towards high fidelity face relighting with realistic shadows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14719–14728, 2021. 2

[19] Andrew Hou, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Face relighting with geometrically consistent shadows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4217–4226, 2022. 2, 4, 6, 7, 1, 3

[20] Andrew Hou, Zhixin Shu, Xuaner Zhang, He Zhang, Yannick Hold-Geoffroy, Jae Shin Yoon, and Xiaoming Liu. Compose: Comprehensive portrait shadow editing. In *European Conference on Computer Vision*, pages 356–373. Springer, 2024. 3

[21] Yuru Jia, Lukas Hoyer, Shengyu Huang, Tianfu Wang, Luc Van Gool, Konrad Schindler, and Anton Obukhov. Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control. In *European Conference on Computer Vision*, pages 91–109. Springer, 2024. 3

[22] Haian Jin, Yuan Li, Fujun Luan, Yuanbo Xiangli, Sai Bi, Kai Zhang, Zexiang Xu, Jin Sun, and Noah Snavely. Neural gaffer: Relighting any object via diffusion, 2024. 2, 3

[23] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 3

[24] Hoon Kim, Minje Jang, Wonjun Yoon, Jisoo Lee, Donghyun Na, and Sanghyun Woo. Switchlight: Co-design of physics-driven architecture and pre-training framework for human

portrait relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25096–25106, 2024. 1, 2, 3, 6, 7

[25] Theodore Kim, Holly Rushmeier, Julie Dorsey, Derek Nowrouzezahrai, Raqi Syed, Wojciech Jarosz, and AM Darke. Countering racial bias in computer graphics research. In *ACM SIGGRAPH 2022 Talks*, pages 1–2. 2022. 3

[26] Peter Kocsis, Julien Philip, Kalyan Sunkavalli, Matthias Nießner, and Yannick Hold-Geoffroy. Lightit: Illumination modeling and control for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9359–9369, 2024. 2, 3

[27] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023. 3

[28] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018. 4

[29] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36: 47500–47510, 2023. 3, 1

[30] Grace Luo, Trevor Darrell, Oliver Wang, Dan B Goldman, and Aleksander Holynski. Readout guidance: Learning control from diffusion features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8217–8227, 2024. 1, 2, 3, 7

[31] Tomohiro Mashita, Yasuhiro Mukaigawa, and Yasushi Yagi. Measuring and modeling of multi-layered subsurface scattering for human skin. In *Virtual and Mixed Reality-New Trends: International Conference, Virtual and Mixed Reality 2011, Held as Part of HCI International 2011, Orlando, FL, USA, July 9-14, 2011, Proceedings, Part I 4*, pages 335–344. Springer, 2011. 3

[32] Yiqun Mei, Yu Zeng, He Zhang, Zhixin Shu, Xuaner Zhang, Sai Bi, Jianming Zhang, HyunJoon Jung, and Vishal M Patel. Holo-relighting: Controllable volumetric portrait relighting from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4263–4273, 2024. 2

[33] Rohit Pandey, Sergio Orts-Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul E Debevec, and Sean Ryan Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Trans. Graph.*, 40(4):43–1, 2021. 2, 3

[34] Sylvain Paris, François X Sillion, and Long Quan. Lightweight face relighting. In *11th Pacific Conference on-Computer Graphics and Applications, 2003. Proceedings.*, pages 41–50. IEEE, 2003. 2

[35] Puntawat Ponglertnapakorn, Nontawat Tritrong, and Supasorn Suwajanakorn. Difareli: Diffusion face relighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22646–22657, 2023. 1, 2, 6, 7

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 6

[37] Pramod Rao, Gereon Fox, Abhimitra Meka, Mallikarjun BR, Fangneng Zhan, Tim Weyrich, Bernd Bickel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, et al. Lite2relight: 3d-aware single image portrait relighting. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 2

[38] Mengwei Ren, Wei Xiong, Jae Shin Yoon, Zhixin Shu, Jianming Zhang, HyunJoon Jung, Guido Gerig, and He Zhang. Relightful harmonization: Lighting-aware portrait background replacement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6452–6462, 2024. 2, 3, 6

[39] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6296–6305, 2018. 3

[40] Shuwei Shi, Biao Gong, Xi Chen, Dandan Zheng, Shuai Tan, Zizheng Yang, Yuyuan Li, Jingwen He, Kecheng Zheng, Jingdong Chen, et al. Motionstone: Decoupled motion intensity modulation with diffusion transformer for image-to-video generation. *arXiv preprint arXiv:2412.05848*, 2024. 3

[41] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5541–5550, 2017. 2

[42] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 2, 3

[43] Tiancheng Sun, Zexiang Xu, Xiuming Zhang, Sean Fanello, Christoph Rhemann, Paul Debevec, Yun-Ta Tsai, Jonathan T Barron, and Ravi Ramamoorthi. Light stage super-resolution: continuous high-frequency relighting. *ACM Transactions on Graphics (TOG)*, 39(6):1–12, 2020. 3

[44] Tiancheng Sun, Kai-En Lin, Sai Bi, Zexiang Xu, and Ravi Ramamoorthi. Nelf: Neural light-transport field for portrait view synthesis and relighting. *arXiv preprint arXiv:2107.12351*, 2021. 2

[45] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8): 3998–4011, 2018. 6

[46] Feitong Tan, Sean Fanello, Abhimitra Meka, Sergio Orts-Escolano, Danhang Tang, Rohit Pandey, Jonathan Taylor, Ping Tan, and Yinda Zhang. Volux-gan: A generative model for 3d face synthesis with hdri relighting. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 2

[47] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 3

[48] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2555–2563, 2023. 6

[49] Tianfu Wang, Menelaos Kanakis, Konrad Schindler, Luc Van Gool, and Anton Obukhov. Breathing new life into 3d assets with generative repainting. In *Proceedings of the British Machine Vision Conference (BMVC)*. BMVA Press, 2023. 3

[50] Tianfu Wang, Mingyang Xie, Haoming Cai, Sachin Shah, and Christopher A Metzler. Flash-split: 2d reflection removal with flash cues and latent diffusion separation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5688–5698, 2025. 3

[51] Yifan Wang, Aleksander Holynski, Xiuming Zhang, and Xuaner Zhang. Sunstage: Portrait reconstruction and relighting using the sun as a light stage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20792–20802, 2023. 2

[52] Zhibo Wang, Xin Yu, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. Single image portrait relighting via explicit multiple reflectance channel modeling. *ACM Transactions on Graphics (ToG)*, 39(6):1–13, 2020. 2

[53] Xiaoyan Xing, Konrad Groh, Sezer Karagolu, Theo Gevers, and Anand Bhattad. Luminet: Latent intrinsics meets diffusion models for indoor scene relighting. *arXiv preprint arXiv:2412.00177*, 2024. 3

[54] Xiaoyan Xing, Konrad Groh, Sezer Karaoglu, Theo Gevers, and Anand Bhattad. Luminet: Latent intrinsics meets diffusion models for indoor scene relighting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 442–452, 2025. 4

[55] Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. Learning to relight portrait images via a virtual light stage and synthetic-to-real adaptation. *ACM Transactions on Graphics (TOG)*, 41(6): 1–21, 2022. 3

[56] Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. Dilightnet: Fine-grained lighting control for diffusion-based image generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 2, 3

[57] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36:45533–45547, 2023. 3

[58] 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*, pages 3836–3847, 2023. 1

[59] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Ic-light github page, 2024. 2, 3

[60] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2, 6, 7

[61] Xuaner Zhang, Jonathan T Barron, Yun-Ta Tsai, Rohit Pandey, Xiuming Zhang, Ren Ng, and David E Jacobs. Portrait shadow manipulation. *ACM Transactions on Graphics (TOG)*, 39(4):78–1, 2020. 2

[62] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7194–7202, 2019. 2