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

Supplementary Material

A. Overview

In this Appendix, we present:

- Section B: Network architectures.
- Section C: Implementation details.
- Section D: Additional Results.
- Section E: Evaluation Details.
- Section F: Detailed Text Prompt.
- Section G: Additional Discussion.

B. Network Architectures

In Fig.11, we present the detailed network architecture of our three key components: the intermediate feature extraction process, the shadow-depth estimator, and the identity estimator. While shadow-depth estimator and the identity estimator are trained separately, they work in conjunction during the inference phase's latent optimization process.

Our design draws inspiration from Readout[30] and Diffusion Hyperfeatures[29]. While these works demonstrated that latent optimization through a compact network can effectively control attributes like human pose and placement, we demonstrate that this control philosophy can be effectively extended to manipulate intrinsic properties like shadows.

B.1. Detail of Intermediate Feature Fetch

The leftmost panel of Fig.11 illustrates our feature extraction process from UNet. We fetch intermediate features at multiple scales in the UNet to capture rich shadow-related information embedded in the model.

B.2. Detailed Architecture of SD Estimator

The middle and right panels in Fig.11 show our shadow-depth estimator architecture. It begins with feature fusion through a series of convolutional layers, followed by weighted sum aggregation. The output layers use multiple convolutions to predict shadow and depth maps.

B.3. Detailed Architecture of ID Estimator

The rightmost panel in Fig.11 depicts our identity estimator, which shares a similar convolutional structure but is specifically designed to extract and maintain identity-related features. Though trained separately from the shadow-depth estimator, it processes the same input features. Unlike the shadow-depth estimator that outputs explicit attribute maps, this network directly produces an identity feature map that guides portrait characteristic preservation during optimization.

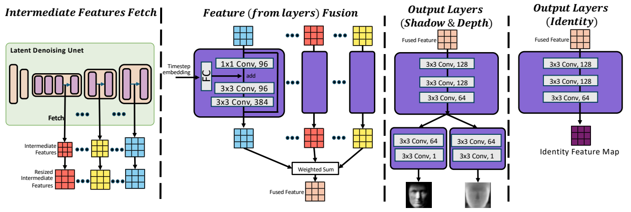


Figure 11. **Details of network architecture.** Both shadow-depth estimator and identity estimator are trained separately but work jointly during inference-time optimization. The pipeline consists of three main steps: (1) Intermediate Features Fetch: extracts multi-scale features from Latent Denoising UNet timesteps as shared input. (2) Feature Fusion: processes features through convolutional layers with weighted sum aggregation. (3) Attributes Output : For Shadow & Depth Output Layers, it generates explicit shadow and depth maps through convolutional layers. For Identity Output Layers, it uses same feature processing structure but outputs identity feature map directly for loss computation.

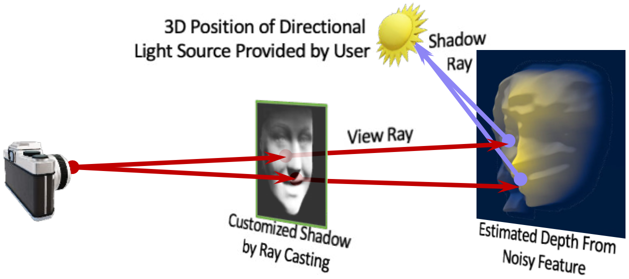

3D Position of Directional Light Source Provided by User
Shadow Ray
View Ray
Customized Shadow by Ray Casting
Estimated Depth From Noisy Feature

Figure 12. Customized shadow acquisition using ray casting. The user specifies the 3D position of the light source, and the shadow is generated based on a depth map estimated by the Shadow Director during inference.

C. Implementation Details

C.1. Shadow Acquisition Details

For the ray-casting option, we use the shadow generation algorithm from [19]. This algorithm takes a depth map and a user-defined lighting position in 3D space as inputs. In brief, a shadow ray is cast toward the light source for each point on the estimated depth map (Figure 12). This depth map is generated by the Shadow Director during inference. If the ray intersects another part of the 3D structure, the

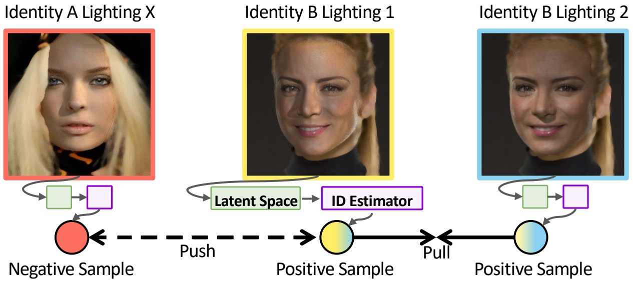


Figure 13. **Training and Mechanism of the Identity Estimator.** The Identity Estimator ensures identity consistency during shadow manipulation. In the training phase, we use three images: two with the same identity but different lighting conditions, and a third with a random identity. Identity feature maps (represented by colored spheres) are independently extracted for each image, following a similar process to the Shadow-Depth Estimator. The triplet loss minimizes the distance between features of the same identity (positive samples) and maximizes the distance from features of different identities (negative samples), enabling the Identity Estimator to effectively distinguish identities. During latent optimization in the inference phase, a reference feature is first generated from a text prompt to guide subsequent shadow manipulations for that prompt.

point is marked as shadowed. This method ensures that shadows align accurately with the 3D geometry.

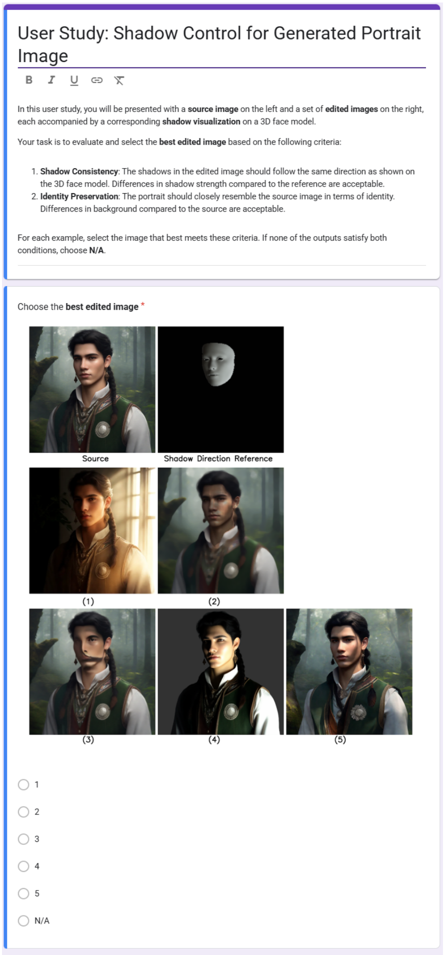
C.2. Training Phase Details

Our framework is built on the stable-diffusion-xl-base-1.0 model. In training phase, the maximum time step is 1000. In each training iteration, we randomly add noise to clean latent features through the scheduler. Shadow-depth estimator and identity estimator are trained separately but share identical training settings: learning rate of 1×10^{-3}, zero weight decay. For the shadow-depth estimator, we use L1 loss to supervise both depth and shadow map predictions against their ground truth (synthetic dataset). And the batch size is 8. For the identity estimator, we employ a hinge loss with 0.5 margin to ensure positive samples remain close in feature space. Moreover, the batch size is 3. 2 positive samples and 1 negative sample. Figure 13 straightforwardly illustrates the training mechanism of ID-Estimator.

C.3. Inference Phase Details

Our shadow manipulation pipeline provides three user control parameters: input type (either binar y mask or 3D light position), and shadow strength (ranging from 0 to 1). A shadow strength of 0 maintains current shadows, while 1 triggers maximum manipulation with 30 optimization iterations. To reduce shadow strength, users can add light to desired regions using 3D light positioning.

The manipulation process involves one generation round with 100 time steps. Shadow manipulation occurs specifically at time step 40. At this step, we first generate the depth map and create a customized shadow map based on user's setting. We then optimize the unconditional branch latent features using Adam optimizer with learning rate 5e-2. The number of optimization iterations is determined by the user-specified shadow strength (strength × 30). During optimization, we combine identity and shadow losses with weights of 3 and 1 respectively. The CFG scale is set to 6 throughout the process.


User Study: Shadow Control for Generated Portrait Image

In this user study, you will be presented with a **source image** on the left and a set of **edited images** on the right, each accompanied by a corresponding **shadow visualization** on a 3D face model.

Your task is to evaluate and select the **best edited image** based on the following criteria:

1. **Shadow Consistency**: The shadows in the edited image should follow the same direction as shown on the 3D face model. Differences in shadow strength compared to the reference are acceptable.
2. **Identity Preservation**: The portrait should closely resemble the source image in terms of identity. Differences in background compared to the source are acceptable.

For each example, select the image that best meets these criteria. If none of the outputs satisfy both conditions, choose **N/A**.

Choose the **best edited image** *

○ 1

○ 2

○ 3

○ 4

○ 5

○ N/A

Figure 14. Example of the user study.

D. Additional Experimental Results

D.1. Handle Occlusions

Yes. While no occluded examples appear in our training set, our method handles occlusions without retraining. In Fig.15, our results cast reasonable shadows for hair, hats, and glasses. Supp file includes 10+ cases (Fig. 15-20) with pre-existing shadow. Moreover, when a user requests an external shadow via text prompt, our pipeline faithfully preserves it. This zero-shot ability arises from the diffusion model's implicit geometric priors to make generated image

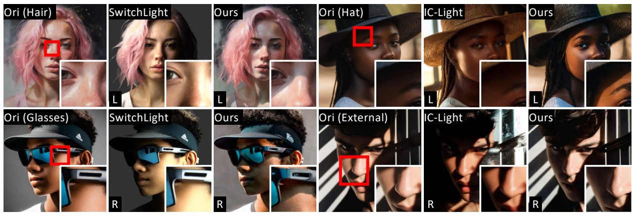


Figure 15. Occlusion handling without retraining. L: left-relit, R: right-relit.

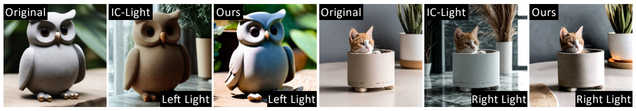


Figure 16. Shadow control on generated object without retraining.

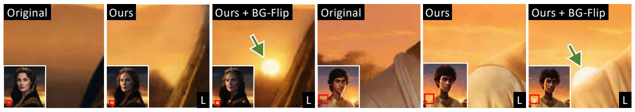


Figure 17. Background flip in latent space harmonizes the scene lighting with our left-relit portraits. After performing latent optimization on the foreground, we apply a simple flip to the background using the portrait mask. This realigns the sunset (green arrow) to match the new facial shading, demonstrating our pipeline's ease of background-aware harmonization.

visually plausible.

D.2. Generalize to Object Relighting

Yes. Without any retraining, our portrait relighting method can control shadows on simple geometric objects (Fig.16), demonstrating generalization potential beyond faces. However, complex materials and geometry remain challenging, as our estimators has not been specialized trained on objects. Fine-tuning on a small object-centric dataset should extend our method to broader range of object efficiently.

D.3. Decoupling Foreground and Background Relighting

Shadow Director aims to focus on foreground portrait relit, resulting in foreground and background solation. Isolating foreground subject shadow from the background allows users to generate stylized lighting effect not restricted by the background. We empower users to generate physically implausible editing for artistic expression if they wish, while still providing the option to adjust lighting in a physically plausible manner. This decoupled design gives users more freedom for control in generation and aligns with prior directional-light editing methods. Meanwhile, our method leaves room for background-aware optimization: just adding a background estimator to harmonize the scene without altering our core pipeline. Fig.17 shows how easily a background tweak fits into our pipeline, even without a dedicated estimator. After we finish the latent optimization (left-relit), we apply a flip on the background (via portrait-mask) in the latent space. This flip realigns the sunset with the new facial lighting. This demonstrates that finer, scene-wide harmonization is feasible within our pipeline design.

D.4. Controlling Directional Light Position

We provide additional results about Shadow Synthesis via Lighting Position Control in Figures 23, Figures 24, Figures 25, Figures 26, Figures 27, Figures 28.

E. Evaluation Details

E.1. Baseline Implementation Details

In Hou [19] method, one may find there is distortion around face. We found that paper, like DiFaReLi, when they reimplement Hou's method. Also has similar distortion. Therefore, it comfirm our implemention correctness. For DiFaReli, after installation, we test author's demo to obtain visually exact same result. However, we found that DiFaReli cannot push the shadow strength be harder. The range of enable shadow strength is relatively limited, indicating this is a challenging task. Meanwhile, we find that the all those required element as needed input for DiFaReli looks reasonable, comfirming our re-implementation is correct. In specific, to do relighting with DiFaReli, there should be a source img to provide refernece shadow. We found that this target shadow is correctly transfer to the relit image's related input, as we shown in the main paper experiment figure. Notably, we use very simple portrait and official demo portrait image as target shadow. However, the DiFaReLi still fails on generated image, indicating this taks is a challenge one.

E.2. User Study

We conduct a user study as shown in Figure 14. We ask the users to choose the outputs based on specific criteria: content preservation and text reflection. We show 20 samples and the outputs of four models: ours. IC-Light. SwitchLight, Hou, and DiFaReLi. The image used in the user study can be seen in Figures 23, Figures 24, Figures 25, Figures 26, Figures 27, Figures 28

F. Text Prompt Templates and Examples

To ensure controlled portrait generation while maintaining consistency in identity and lighting, we designed structured text prompts tailored for our diffusion model. These prompts balance diversity in artistic styles while minimizing interference from excessive accessories, complex clothing, or elaborate backgrounds, which could affect shadow manipulation.

F.1. Prompt Template

Our prompts follow a structured format to enforce consistency in composition and lighting conditions. The template is as follows:

A [STYLE] close-up portrait of a [AGE] [GENDER] with [ETHNICITY] and [FACE_SHAPE] features. Wearing [SIMPLE_CLOTHING].

Where:

- **STYLE**: Specifies the artistic style (e.g., oil painting, cinematic, gothic, fantasy).
- **AGE**: Defines the subject's age category (e.g., young, middle-aged, elderly).
- **GENDER**: Indicates gender identity (e.g., man, woman).
- **ETHNICITY**: Ensures diversity in generated subjects (e.g., Asian, African, Nordic, Mediterranean).
- **FACE_SHAPE**: Controls facial structure (e.g., angular, round, chiseled).
- **SIMPLE_CLOTHING**: Limits clothing complexity (e.g., dark tunic, plain robe, leather vest) to preserve identity consistency.
- **BACKGROUND**.

F.2. Example Prompts

To illustrate the variety of generated portraits, we provide a few example prompts:

- *A gothic close-up portrait of a young man with Nordic and chiseled features. Wearing a dark tunic. A blurred studio background.*
- *A cinematic close-up portrait of an elderly woman with African and angular features. Wearing a simple robe. A blurred studio background.*
- *A fantasy close-up portrait of a middle-aged warrior with Eastern European and strong features. Wearing a leather vest. A blurred studio background.*
- *A Renaissance-inspired close-up portrait of a young queen with Mediterranean and delicate features. Wearing an embroidered cloak. A blurred studio background.*

These structured prompts allow controlled generation of diverse portraits while ensuring identity preservation, shadow consistency, and style variety.

G. Additional Discussion

G.1. Why can Identity and Shadow be Decoupled?

Decoupling shadow and identity is well-established by prior work on style and lighting transfer: StyleAligned [16] shows that swapping style elements (lighting and texture) between generated images preserves identity, and LumiNet [54] shows a diffusion model can separate lighting from scene. Meanwhile, our ID Estimator is trained to be illumination-invariant, focusing solely on identity and ignoring shadow variations. These prior findings and our illumination-invariant design explain why we can adjust lighting without significantly affecting identity.

G.2. Extend to Environment maps?

While our design did not originally target environment-map relighting, our lightweight modular architecture makes it plausible. One can replace ray casting with a neural shadow synthesis module, which takes an environment map along with outputs from new normals and shading estimators to produce shadow map. Then, joint optimize them with ID estimator with rendering constraints. Since this extension would fit within our pipeline, we can preserve our method's strengths.

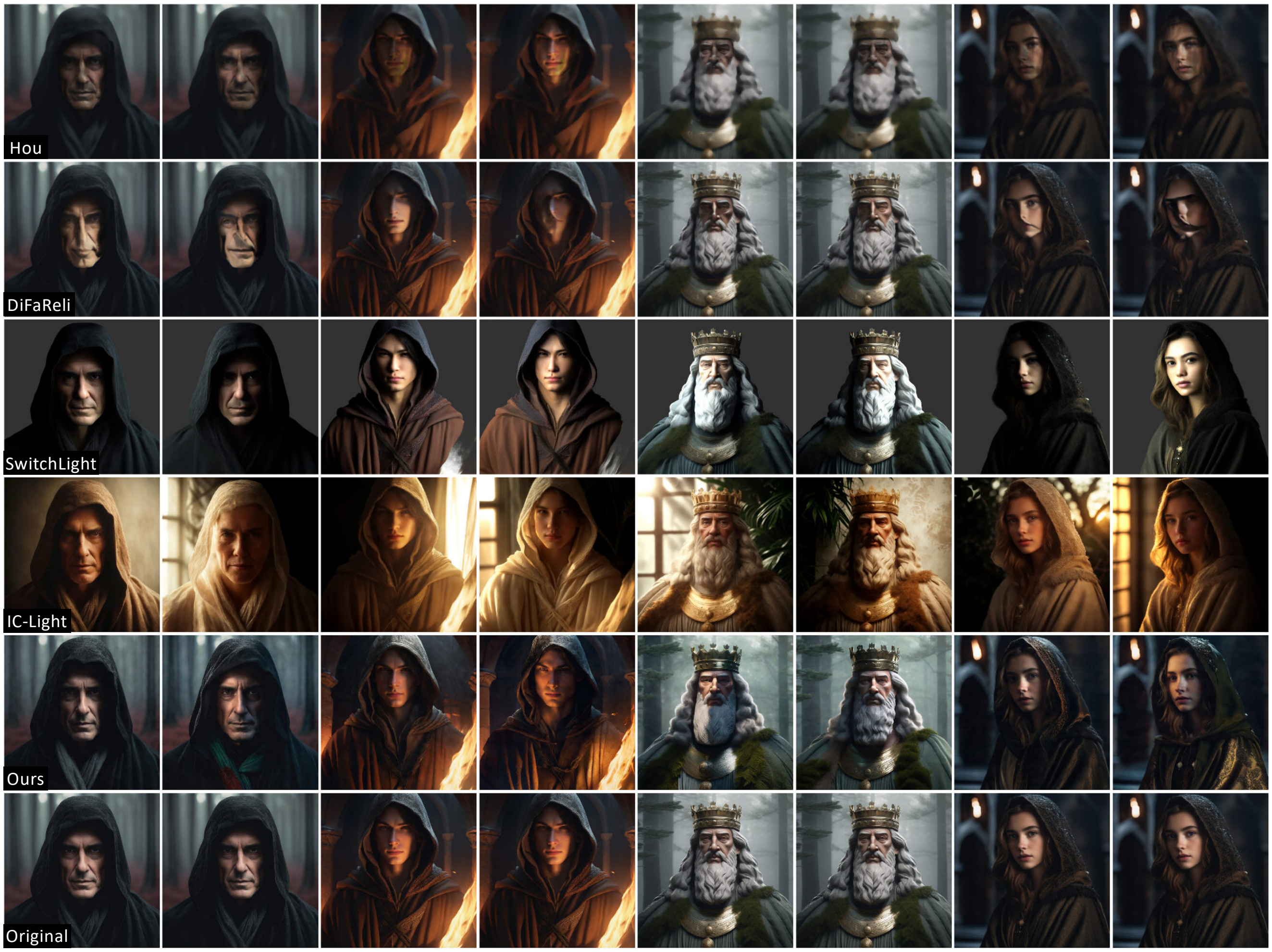


Figure 18. Comparison of portrait relighting across different editing methods on diverse artistic styles. Each row corresponds to a different method, while each column maintains left and right lighting direction for convenient visual comparison.

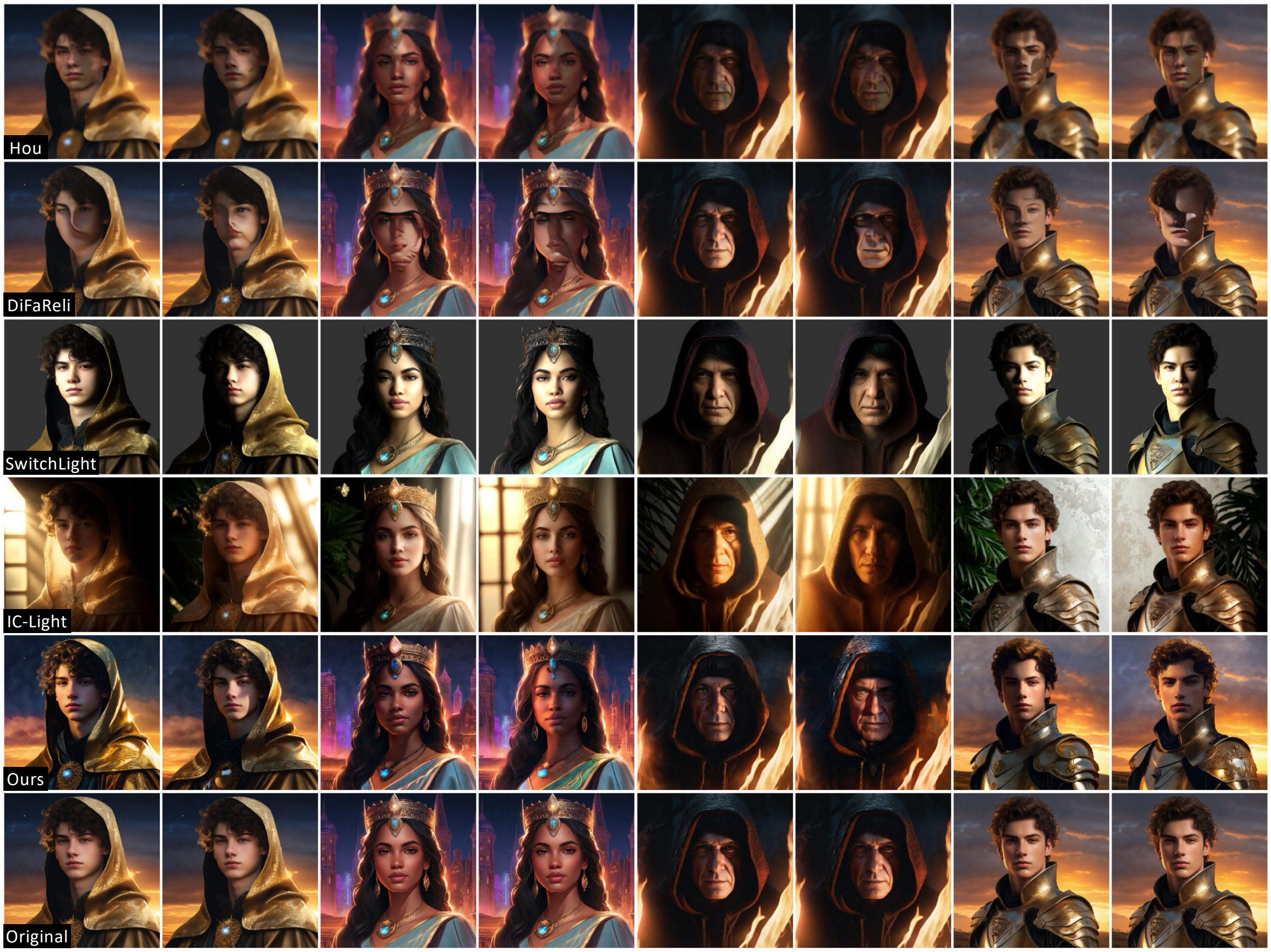


Figure 19. Comparison of portrait relighting across different editing methods on diverse artistic styles. Each row corresponds to a different method, while each column maintains left and right lighting direction for convenient visual comparison.

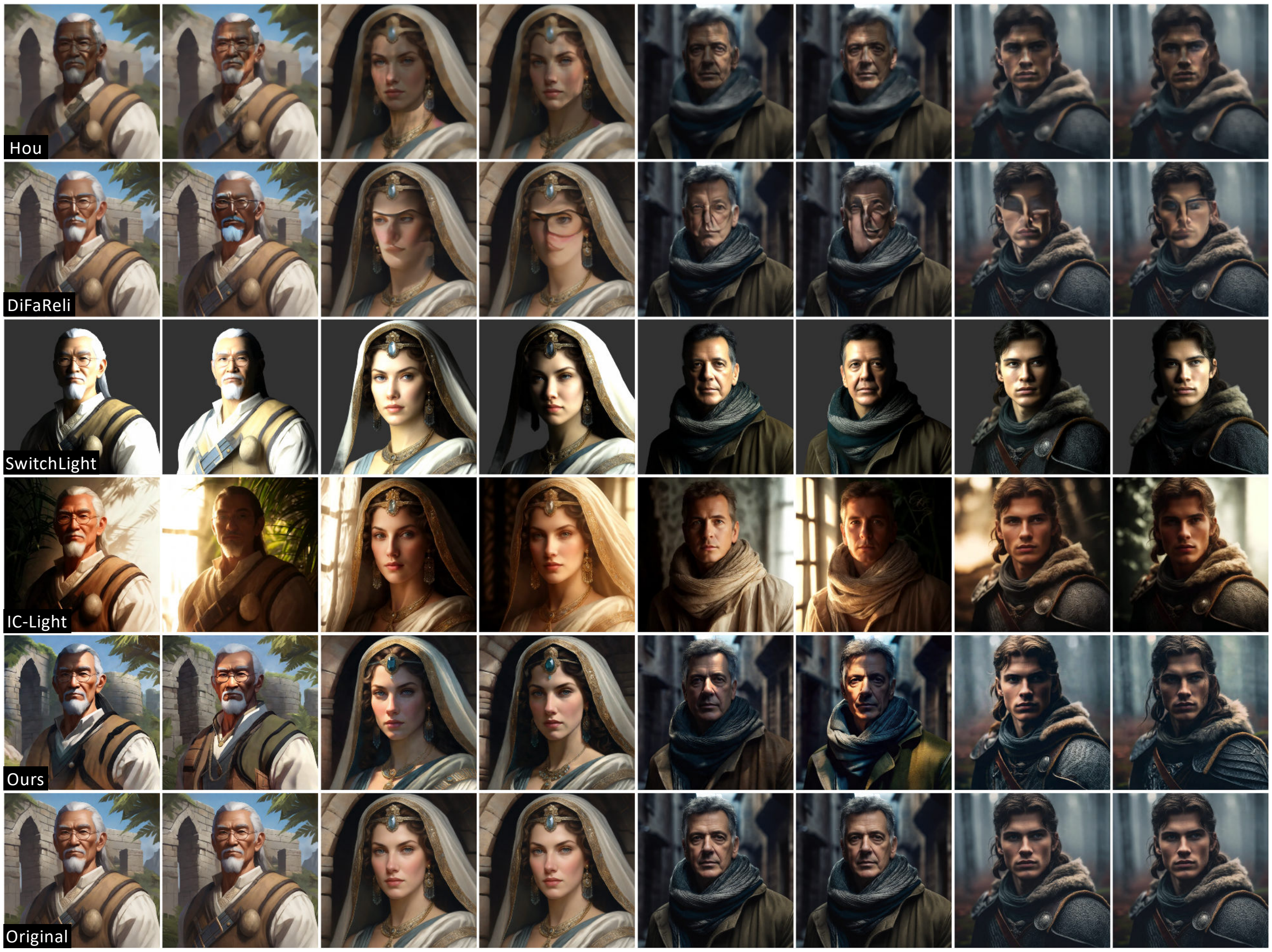


Figure 20. Comparison of portrait relighting across different editing methods on diverse artistic styles. Each row corresponds to a different method, while each column maintains left and right lighting direction for convenient visual comparison.

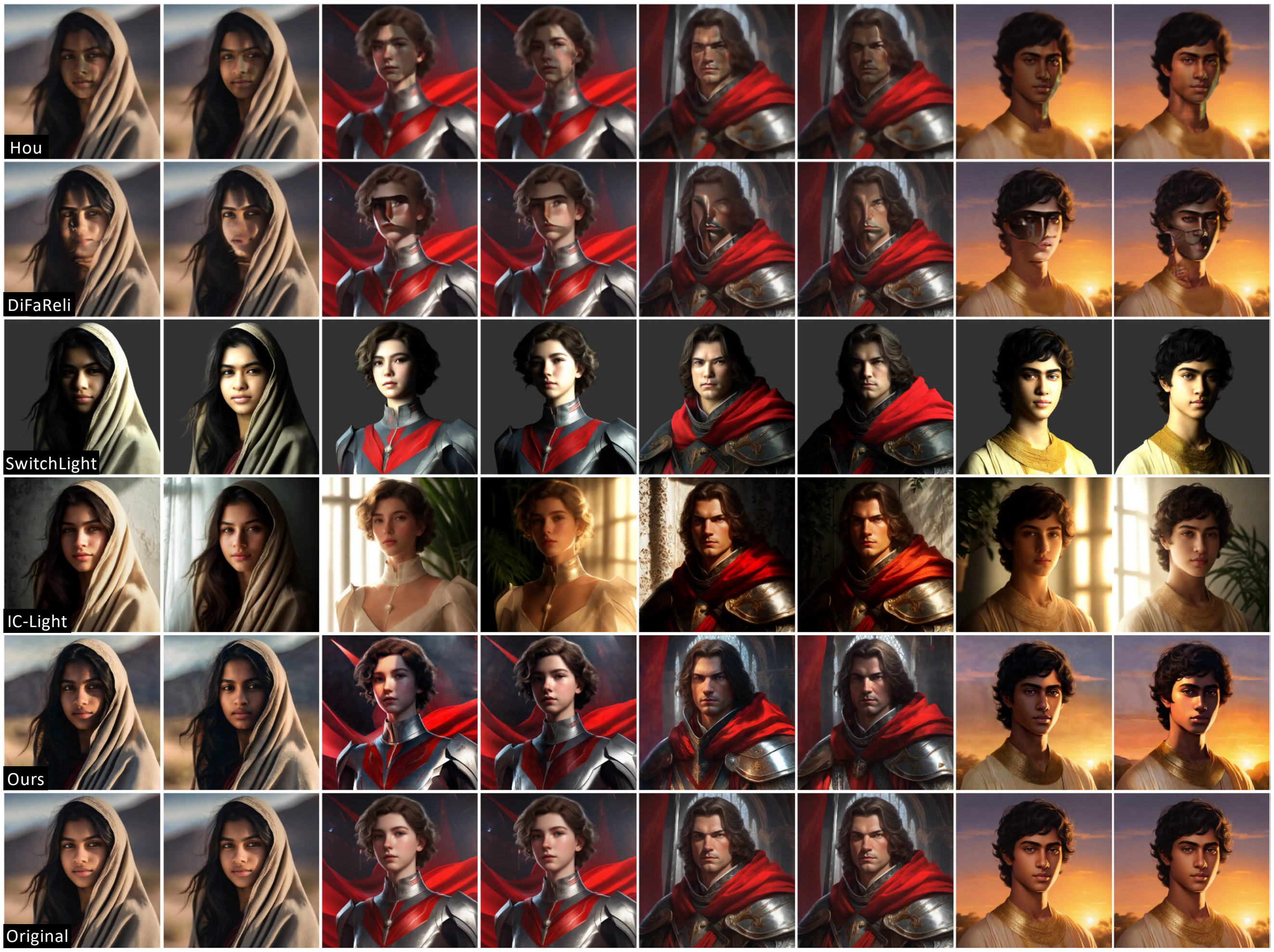


Figure 21. Comparison of portrait relighting across different editing methods on diverse artistic styles. Each row corresponds to a different method, while each column maintains left and right lighting direction for convenient visual comparison.

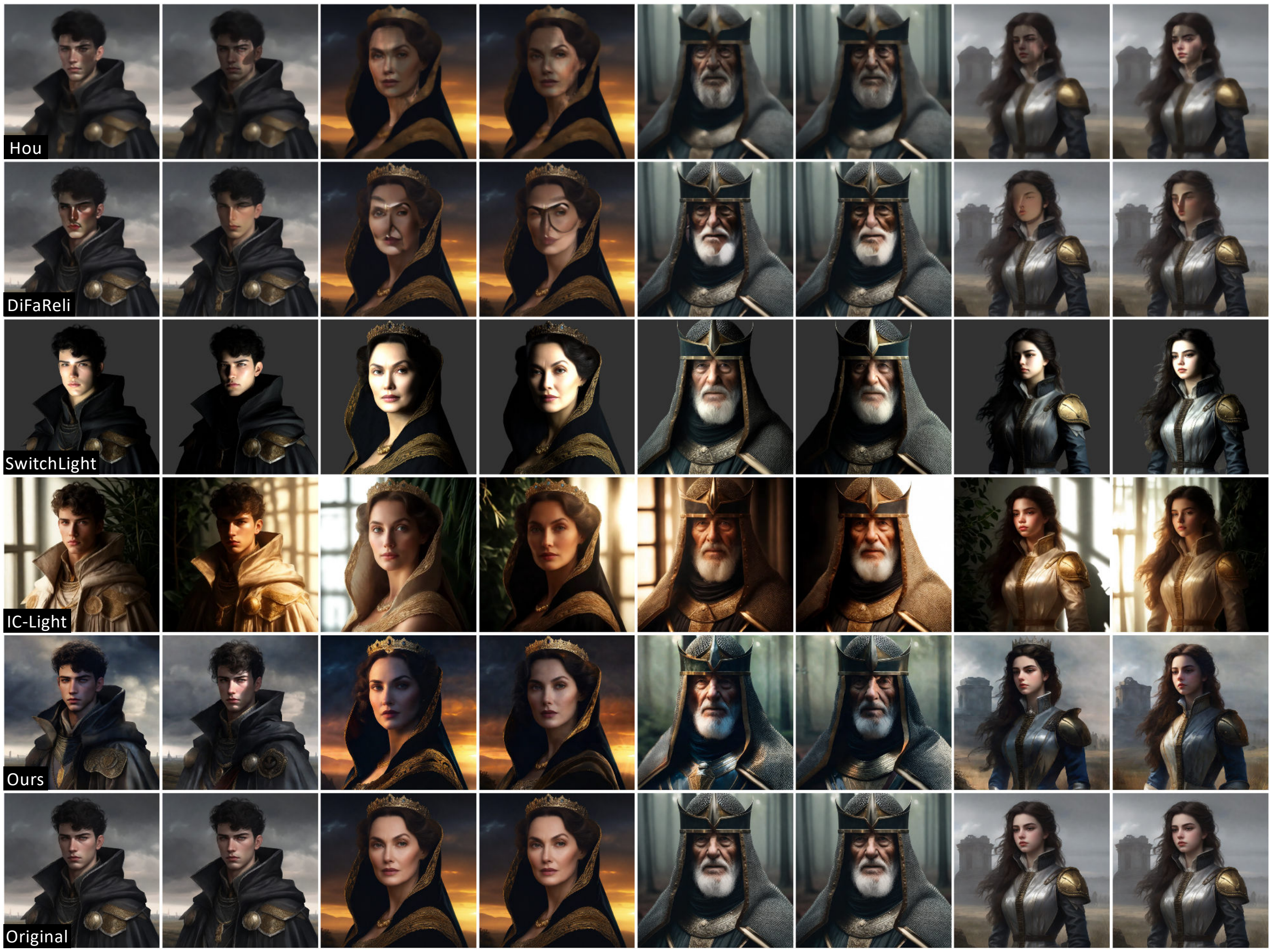


Figure 22. Comparison of portrait relighting across different editing methods on diverse artistic styles. Each row corresponds to a different method, while each column maintains left and right lighting direction for convenient visual comparison.

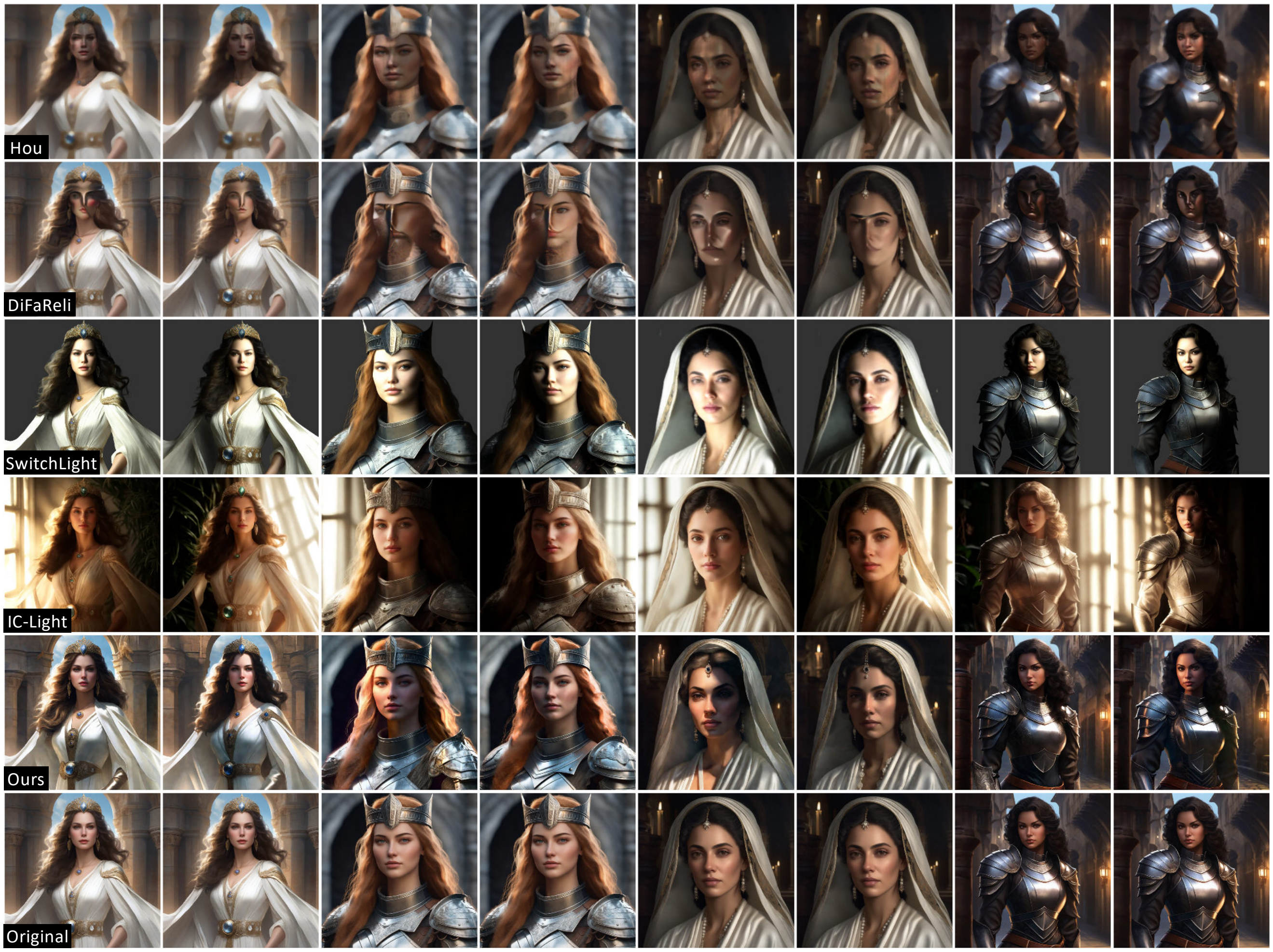


Figure 23. Comparison of portrait relighting across different editing methods on diverse artistic styles. Each row corresponds to a different method, while each column maintains left and right lighting direction for convenient visual comparison.

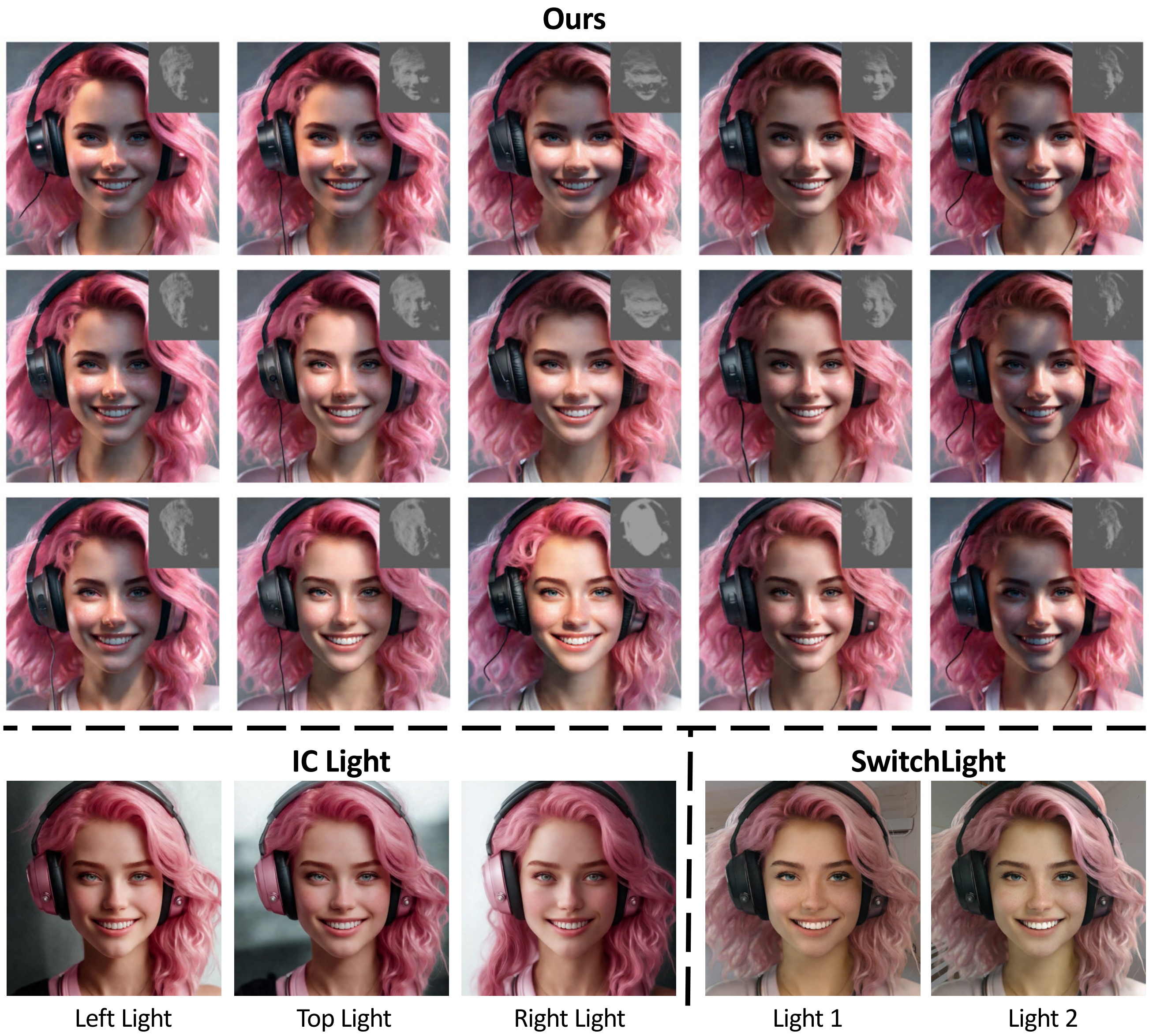


Figure 24. Shadow Synthesis under more lighting conditions. SwitchLight here doesn't use directional light.

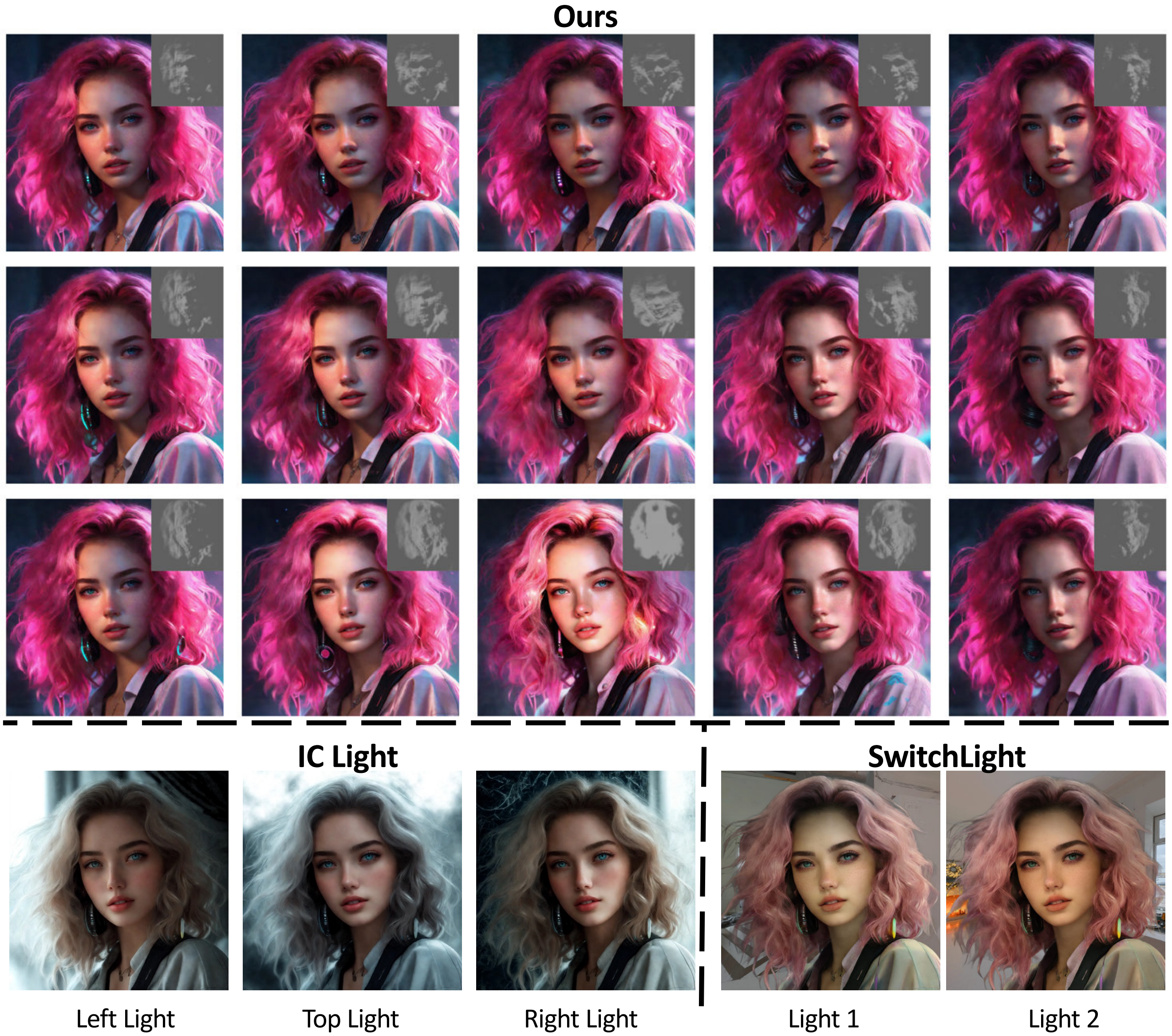


Figure 25. Shadow Synthesis under more lighting conditions. SwitchLight here doesn't use directional light.

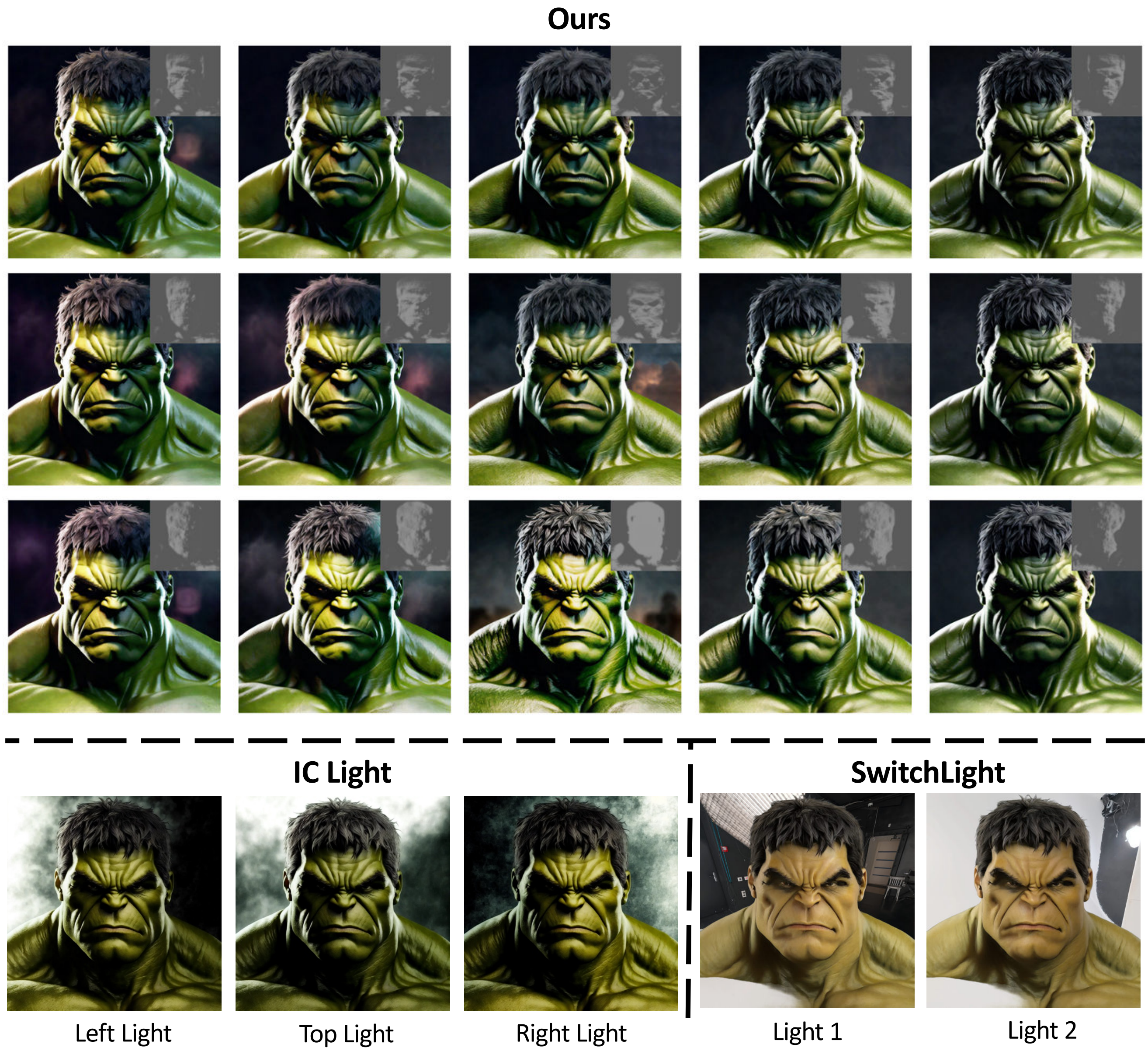


Figure 26. Shadow Synthesis under more lighting conditions. SwitchLight here doesn’t use directional light.For IC-Light here, ”Hulk” text prompt is feed into as another condition to help IC-Light maintain the identity

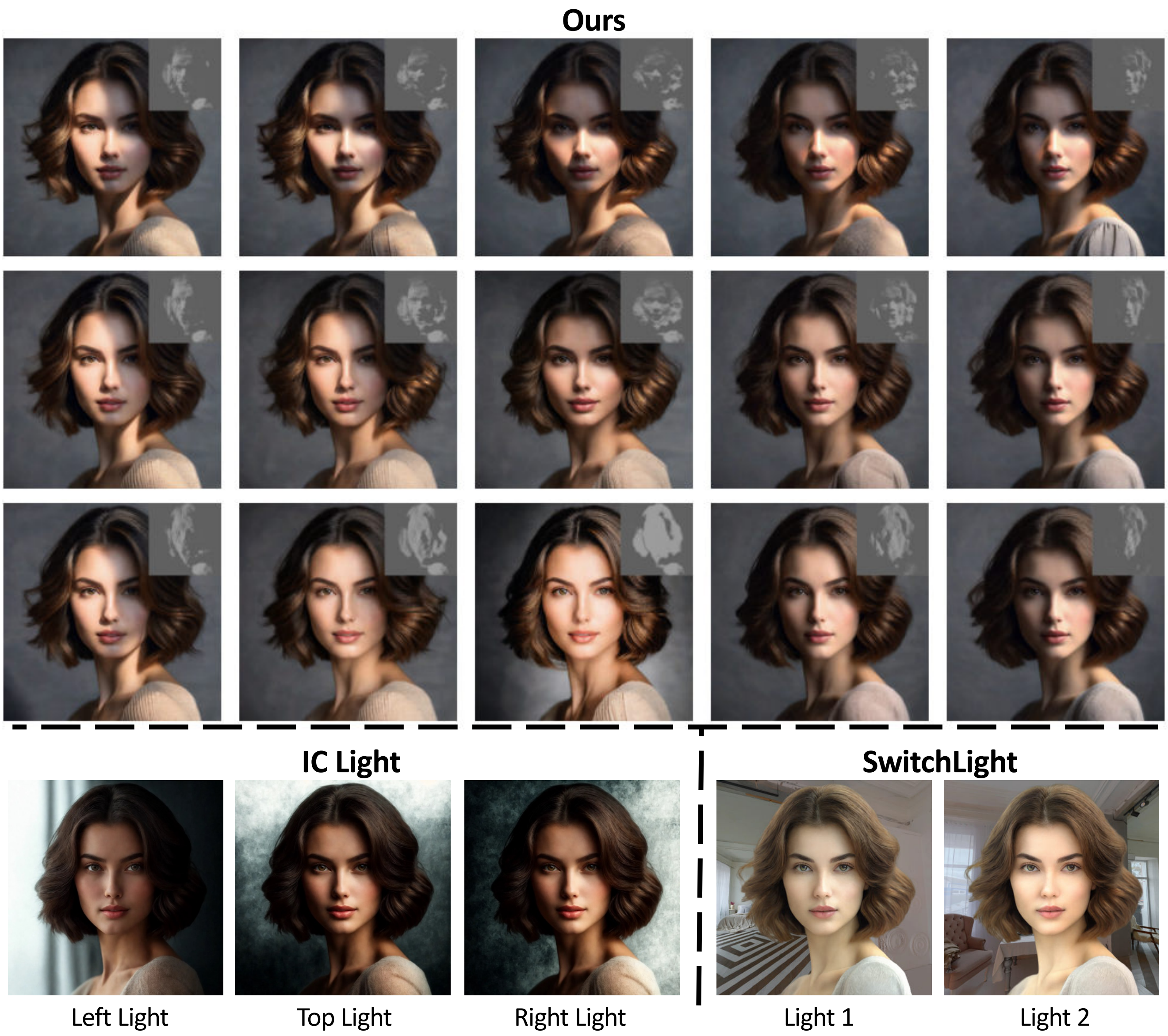


Figure 27. Shadow Synthesis under more lighting conditions. SwitchLight here doesn't use directional light.

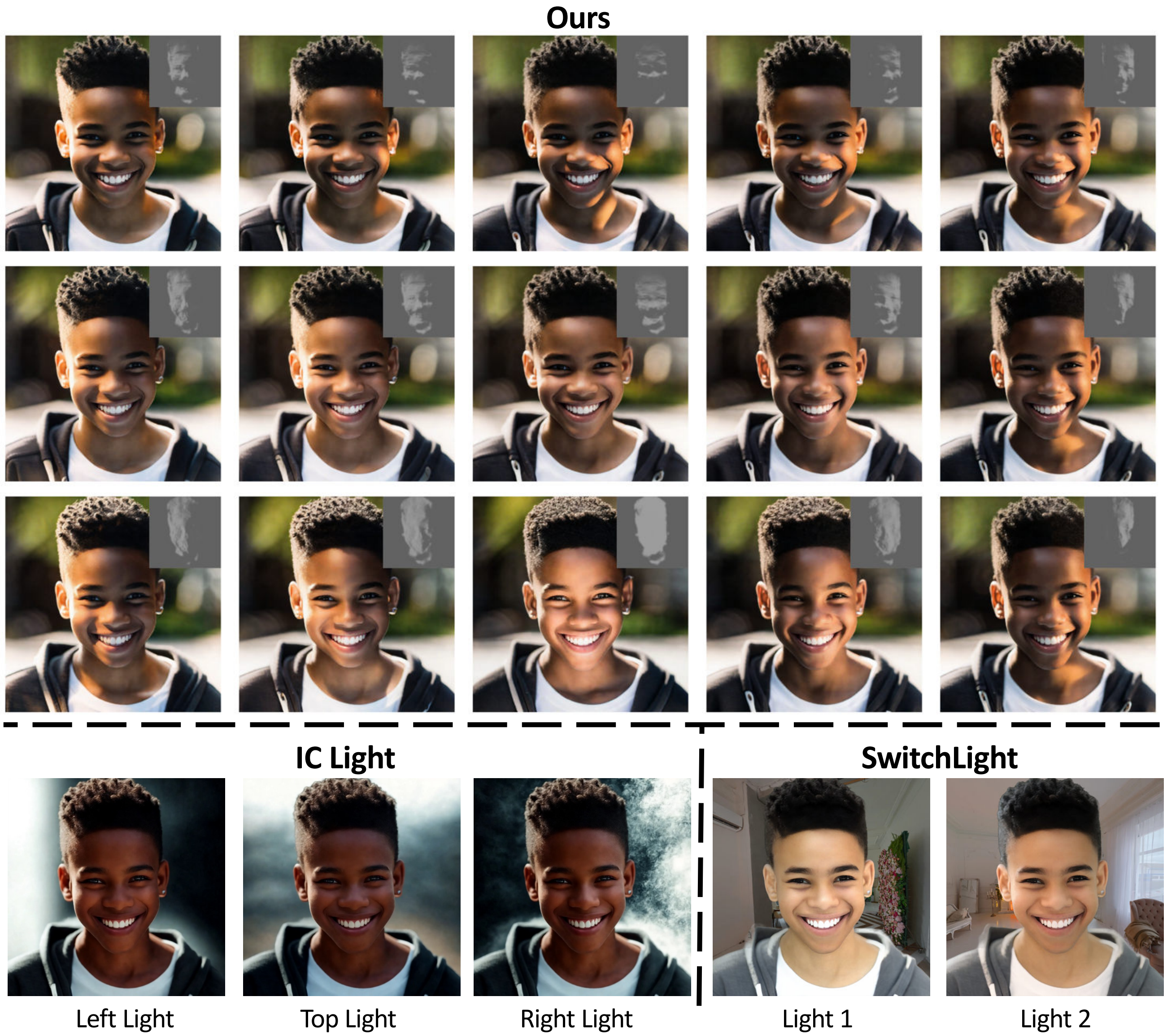


Figure 28. Shadow Synthesis under more lighting conditions. SwitchLight here doesn't use directional light