

Consistent Feature Transport for Image Relighting

Bohan Zhang^{1†}, Huanwei Liang^{2†}, Yuhan He², Hongteng Xu³, Xiaochao Qu², Luoqi Liu², Dixin Luo^{1*}, and Ting Liu^{2*}

¹ School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China

² MT Lab, Meitu Inc., Beijing 100083, China

³ Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872, China

Abstract. Image relighting modifies illumination while preserving non-lighting content such as identity and geometry. Existing diffusion-based methods often suffer from unstable illumination changes or inconsistent content preservation under complex lighting, as they lack an explicit mechanism to learn feature transformations between images. We reformulate relighting as an illumination feature transport problem and introduce Consistent Feature Transport (CFT), a training principle that explicitly enforces illumination-consistent transport between source and target image distributions. Built upon rectified flow, CFT jointly models noise-to-image generation and illumination-consistent source-to-target transport through trajectory-level supervision. This dual-transport formulation encourages isolation of illumination-specific variations while preserving content-aligned features. To support complex lighting scenarios, we construct a large-scale portrait relighting dataset with diverse relighting effects. Experiments show consistent improvements over existing state-of-the-art relighting approaches and demonstrate that CFT can generalize to other editing tasks, including style transfer. Code is available at <https://github.com/Dixin-Lab/CFT>.

Keywords: Image Editing · Image Relighting · Feature Transport

1 Introduction

Image relighting aims to modify the illumination conditions of a given image while preserving its underlying content, including scene structure, identity, and material appearance. The task focuses on editing lighting attributes such as direction, intensity, color, and the interaction between light and other content elements. As a core problem in image editing, relighting has attracted increasing attention with the rapid development of generative models, particularly diffusion-based methods that enable flexible manipulation of visual attributes. Accurate

[†] Equal contribution. Work was done while Bohan Zhang was an intern at MT Lab, Meitu Inc., Beijing, China.

^{*} Corresponding authors: dixin.luo@bit.edu.cn, lt@meitu.com.

and consistent relighting is essential for a range of applications, including portrait enhancement [19, 38], virtual cinematography [26, 49], and visual effects production [6, 34].

With the advancement of diffusion-based image generation and editing methods, existing image relighting methods can be broadly categorized into three paradigms according to how they model illumination transformation. The first paradigm is control-based relighting, which formulates illumination editing as a conditional generation task. These methods incorporate additional illumination-related signals [21, 31, 49] to guide the diffusion process. While such designs improve controllability, the final results are heavily dependent on the accuracy and completeness of the provided control signals. In practice, precisely specifying complex illumination effects remains challenging, which may limit robustness, consistency, and generalization. The second paradigm is decomposition-based relighting, where illumination changes are achieved by manipulating intermediate representations [26, 33, 44, 50]. By operating in a decomposed space, these approaches attempt to isolate lighting factors from geometry and appearance. However, intrinsic decomposition itself may be inaccurate, especially under complex illumination conditions. Errors in decomposition can propagate to the edited results, leading to inconsistent shading, color shifts, or loss of fine details. The third paradigm formulates relighting as direct image-to-image translation, modeling the transport between source and target image distribution using generative models [5]. This formulation treats relighting as an instance-level translation problem and does not explicitly model illumination-specific feature displacement. As a result, while structural fidelity can be maintained, the controllability and precision of illumination editing are limited.

Apart from methodological challenges, data availability also constrains the progress of image relighting. While several open-source datasets exist for relighting objects or generic scenes [14, 15, 28, 35], publicly available portrait relighting datasets remain limited. Existing portrait relighting data [11, 19, 37, 41] often lack sufficient diversity in lighting configurations, particularly scenarios involving complex illumination effects, which hinders the development and evaluation of methods designed for realistic and controllable image relighting.

In this paper, we reformulate relighting as an illumination feature transport problem and introduce **C**onsistent **F**eature **T**ransport (CFT) for image relighting, which explicitly models the illumination-consistent transport between source and target image distributions. Specifically, we build our method upon rectified flow and consider the relationships among the noise, source image, and target image distributions. Similar to conventional image generation and editing frameworks, we model both noise-to-source and noise-to-target transport. For the source-to-target transport, instead of directly supervising with the original image pairs, we use alternative pairs that share the same illumination condition as supervision. This design encourages the model to focus on an illumination feature transformation rather than a generic image-to-image modification. By jointly modeling noise-to-image generation and illumination-consistent source-to-target transport, the framework guides the generative process to focus on

transporting illumination features while preserving identity, geometry, and other non-lighting attributes. In addition, we construct a dedicated portrait relighting dataset to support learning under complex lighting conditions. The dataset contains diverse portrait images under a wide range of illumination effects, providing richer supervision for modeling lighting variations.

In summary, this work makes the following contributions. First, we introduce a consistent feature transport principle for image relighting that explicitly models the illumination-consistent transport between the source and target image distributions. The proposed framework encourages faithful preservation of image content while improving targeted modification of lighting features. Second, we present a portrait relighting dataset with diverse light configurations and complex illumination effects. This dataset fills the gap in existing resources, where portrait data with complex illumination is scarce. Third, extensive experiments demonstrate the effectiveness of the proposed approach, yielding consistent improvements over existing baselines. Beyond relighting, we further show that the proposed method can generalize to other editing tasks, such as style transfer, highlighting its potential as a unified strategy for image editing.

2 Related Work

2.1 Diffusion-based Image Editing

Diffusion-based image editing aims to modify specific visual attributes while preserving the underlying content of the input image. Existing approaches differ primarily in how the source image and desired modification are incorporated into the generative process. A common paradigm is control-based generation, where the model starts from noise and generates the edited target image conditioned on the source image and additional control signals, such as textual instructions [3, 48], reference images [9, 42, 45], or disentangled representations [10, 36, 47]. In this formulation, editing is realized as conditional generation. While effective, the quality of editing largely depends on the expressiveness and accuracy of the control conditions. Another line of work adopts trajectory-based editing, which manipulates diffusion paths to steer the model from the source image toward the edited target result. Both training-based approaches [5, 7, 39] and training-free methods [4, 12, 20, 23, 40] operate directly on the diffusion trajectory, starting from the source image and progressively guiding it toward the desired target. While such designs help maintain structural consistency due to the strong anchoring effect of the source image, they typically focus on instance-level alignment or generic domain transport. As a result, the editing flexibility is often limited, and achieving substantial or fine-grained attribute modifications can be challenging.

Overall, existing diffusion editing frameworks either rely on conditional generation or trajectory manipulation, but rarely consider explicit and direct feature transport between source and target domains.

2.2 Image Relighting Methods

Image relighting aims to selectively modify illumination without altering other content, making it particularly sensitive to how transformations are modeled. Early neural relighting methods were largely grounded in physically inspired modeling. These approaches leveraged explicit illumination representations [25, 50], intrinsic decomposition [22, 33, 46], or physics-guided reflectance models [19, 37] for lighting variations. By introducing physical priors, they enabled interpretable control over light sources and shading behavior. However, such methods often depend on accurate intrinsic decomposition or specialized setups, which limit their ability to handle images with complex illumination effects.

With the emergence of diffusion models, relighting has increasingly been addressed within a generative editing framework. Following the control-based paradigm, several works introduce illumination-related conditions, such as environmental maps [6, 49], intrinsic components [8, 21, 29, 34], or lighting representations [2, 31], to guide the denoising process. Although this improves controllability, the effectiveness of relighting depends heavily on the quality and completeness of the control signals. Precisely specifying complex illumination effects, especially under multiple light sources or spatially varying shadows, remains challenging. Alternatively, relighting has also been formulated as image-to-image or trajectory-based translation, where the model directly steers the diffusion path from the source image toward a target illumination domain. Recent works explore bridge-based formulations to learn mappings between image distributions [5]. This paradigm treats relighting as an instance-level translation problem, which helps preserve the underlying content of the source image. However, the direct instance-level translation also limits their ability to perform fine-grained and expressive illumination editing.

Overall, existing relighting approaches either depend on explicit illumination representations and conditions or perform instance-level diffusion transport. In contrast, our approach explicitly models how illumination features should be transported between source and target representations, enabling more reliable and consistent relighting under diverse and complex lighting conditions.

3 Methodology

3.1 Proposed Method

Let x_{src} denote the source image before relighting and x_{tgt} denote the corresponding relighting target image. The desired illumination change is specified by a conditioning signal c , represented as a textual description of the lighting transformation. The goal is to learn a conditional velocity model v_θ that generates the relighting image starting from noise while preserving the non-lighting content of x_{src} . Following the rectified flow formulation [27, 30], the model is trained to predict the transport velocity from a prior sample toward the target latent representation under the guidance of the source image and the textual

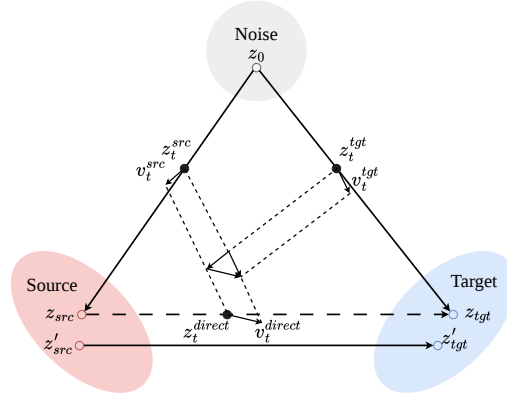


Fig. 1: Overall framework of the proposed method.

condition. The training objective is defined as:

$$\mathcal{L}_{\text{relight}}(\theta) = \mathbb{E}_{z_0 \sim \pi_0, z_{tgt}, t \sim \mathcal{U}(0,1)} \left[\|v_\theta(z_t, t, x_{src}, c) - (z_{tgt} - z_0)\|^2 \right], \quad (1)$$

where z_0 denotes a sample drawn from the prior distribution and z_{tgt} denotes the latent representation of x_{tgt} .

As illustrated in Figure 1, we consider the relationships between the noise, source image, and target image distributions. Following the original rectified flow formulation, we train the model to learn the transport from Gaussian noise to the target image distribution. Specifically, given a source image x_{src} , a target relighting image x_{tgt} , and the corresponding relighting condition c_{tgt} , the model is trained to predict the velocity field that transports samples from the prior toward the target latent representation:

$$\mathcal{L}_1 = \mathbb{E}_{z_0 \sim \pi_0, z_{tgt}, t \sim \mathcal{U}(0,1)} \left[\|v_\theta(z_t^{tgt}, t, x_{src}, c_{tgt}) - (z_{tgt} - z_0)\|^2 \right]. \quad (2)$$

To strengthen content preservation, we additionally train the model to learn the transport from noise to the source image distribution. This objective corresponds to a reconstruction pathway under a neutral condition c_{src} from the same prior z_0 :

$$\mathcal{L}_2 = \mathbb{E}_{z_0 \sim \pi_0, z_{src}, t \sim \mathcal{U}(0,1)} \left[\|v_\theta(z_t^{src}, t, x_{src}, c_{src}) - (z_{src} - z_0)\|^2 \right]. \quad (3)$$

The objectives $\mathcal{L}_1$ and $\mathcal{L}_2$ are sufficient for conditional generation, as they steer the model to generate the source and target image under different conditioning signals. However, such conditional generation is insufficient for the control over illumination variation and often fails to produce consistent lighting transformations across different instances. Therefore, beyond the noise-to-image generation objective, we explicitly model the transport between the source and target distributions, encouraging direct illumination transformation.

Following the inversion-free image editing methods [12, 23], we leverage the linearity of the rectified flow and approximate the intermediate states along the direct path via the parallelogram law:

$$z_t^{\text{direct}} = z_{\text{src}} + z_t^{\text{tgt}} - z_t^{\text{src}}. \quad (4)$$

According to the ODE formulation, the corresponding velocity along this path can be expressed as:

$$v_t^{\text{direct}} = v_\theta(z_t^{\text{tgt}}, t, x_{\text{src}}, c_{\text{tgt}}) - v_\theta(z_t^{\text{src}}, t, x_{\text{src}}, c_{\text{src}}). \quad (5)$$

During training, a straightforward approach is supervising v_t^{direct} with the original pair $(x_{\text{src}}, x_{\text{tgt}})$. However, since v_t^{direct} is derived from the parallelogram law, it essentially models the transport from x_{src} to x_{tgt} . Directly using the pair $(x_{\text{src}}, x_{\text{tgt}})$ as ground truth does not encourage the model to learn illumination features. Instead, such supervision may encourage the model to fit sample-specific differences that entangle illumination variations with other non-lighting content. Consequently, the model may rely on spurious content-dependent correlations to accomplish illumination edits, rather than learning illumination-specific transformations.

In contrast to the choice of the original pair $(x_{\text{src}}, x_{\text{tgt}})$ for supervision, we apply another image pair $(x'_{\text{src}}, x'_{\text{tgt}})$ that shares the same illumination transformation but differs in identity and scene content from the original image pair $(x_{\text{src}}, x_{\text{tgt}})$ as the ground-truth pathway. We can obtain the following objective:

$$\mathcal{L}_3 = \mathbb{E}_{z'_{\text{src}}, z'_{\text{tgt}}, t} \left[\|v_t^{\text{direct}} - (z'_{\text{tgt}} - z'_{\text{src}})\|^2 \right], \quad (6)$$

where v_t^{direct} is defined in Equation (5). Thus, the model can learn illumination-consistent transport between the source and target image distributions, rather than fitting content-specific differences between individual image pairs. The model is encouraged to represent relighting as a transport of illumination features, instead of a generic image-to-image modification.

The final training objective combines the above components:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \alpha \cdot \mathcal{L}_3, \quad (7)$$

where α controls the strength of the consistent feature transport. This formulation enables the model to learn both noise-to-image generation and illumination-consistent feature transport, improving consistency of relighting and the overall generated result.

3.2 Dataset Construction

Existing relighting datasets [14, 15, 28, 35] primarily focus on objects or general scenes, while publicly available portrait relighting datasets remain scarce. Moreover, many existing portrait relighting data [11, 19, 37, 41] are captured under controlled environments with relatively simple illumination setups. Therefore,

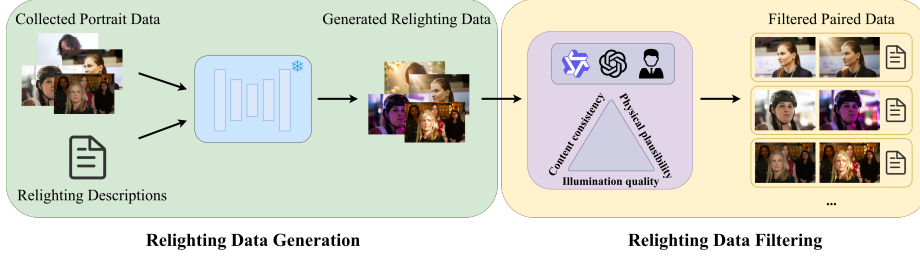


Fig. 2: Dataset Construction Pipeline.

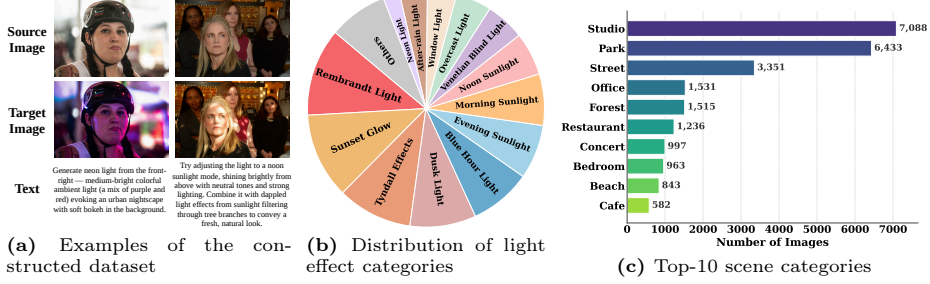


Fig. 3: Overview of the constructed dataset.

they lack diverse complex lighting effects and multiple light configurations, such as different colors, spatially varying illumination, and non-uniform shadows. This limitation reduces their effectiveness for training and evaluating models designed for realistic and controllable image relighting.

To address this gap, we construct a portrait relighting dataset with diverse and complex lighting effects. The overall data construction pipeline is illustrated in Figure 2. First, we collect source images without explicit light effects from existing open-source portrait datasets [13, 17, 18]. These images exhibit substantial diversity in subjects, appearances, poses, and background environments, providing a rich set of base images for relighting data generation. Meanwhile, we generate a comprehensive set of relighting descriptions that cover a broad spectrum of illumination styles. We consider two major categories of lighting sources: natural lighting (*e.g.*, sunlight, moonlight) and artificial lighting (*e.g.*, streetlights, neon lights). For each category, the descriptions are constructed by specifying several key aspects of the lighting: (1) Light direction: describing the spatial direction of illumination, including front, side, and bottom lighting, as well as their combinations, such as front-left and back-right; (2) Light configuration: representing typical lighting setups and effects, such as Tyndall light, Rembrandt lighting, and Blue Hour illumination; (3) Photometric properties: characterizing lighting attributes including color temperature, light intensity, shadow softness, and overall atmospheric tone; These aspects ensure the description spans both global illumination changes and spatially structured lighting patterns. We then employ existing pre-trained models (Nano Banana, Seedream, and RoboNeo) to generate candidate relighting images under those descriptions.

Since automatically generated relighting results may vary in quality, we further employ an ensemble-based filtering pipeline to ensure dataset reliability. Specifically, we evaluate each source and target image pair using three models: QwenVL [1], EditScore [32], and GPT-4o [16]. The evaluation focuses on three criteria: (1) Illumination quality: evaluating whether the generated lighting effects are visually plausible and consistent with the specified relighting descriptions; (2) Content consistency: ensuring that non-illumination content, including facial identity, pose, and background, remains unchanged between the source and relit images; (3) Physical plausibility: assessing whether the lighting configuration satisfies basic physical constraints, such as consistent shading direction and coherent shadow patterns. When all three models confirm the satisfaction of the three criteria, the image pair is retained in the dataset. After filtering with three models, we further perform manual selection based on the above criteria to ensure data quality.

Finally, we construct a portrait relighting dataset consisting of 34,695 image pairs in total, including 34,249 pairs for training and 446 pairs for testing. Figure 3 presents visual examples and statistics of the constructed dataset. For illumination effects, we summarize the light-effect categories contained in the relighting descriptions and visualize their distribution by image count, as shown in Figure 3b. For the scenes in the dataset, indoor scenes account for 48.1% and outdoor scenes account for 51.9%. We further perform a statistical analysis of the scene categories in the dataset. Figure 3c presents the number of images for the Top-10 scene categories. The resulting dataset provides diverse light configurations and complex illumination effects.

4 Experiments

4.1 Setup

We compare our method with several representative approaches. (1) Control-based editing (Flux-Kontext [24], Qwen-Image-Edit-2509 [43], and IC-light [49]): leveraging the diffusion-based image editing models with controllable signals as guidance for the image relighting task. Notably, IC-light applies an environment map as a training condition, whereas it can perform inference without any additional control signals. (2) Decomposition-based editing (Intrinsic Edit [33] and Latent Intrinsic [50]): decomposing images into intrinsic components (*e.g.*, shadings, albedos) or latent representations and performing relighting by editing those intrinsics. (3) Image-to-image translation (LBM [5], FlowEdit [23]): performing editing directly through the diffusion path from the source image to the target image. (4) Original pretrained models (Flux-Kontext, Nano Banana): directly using the pretrained model without fine-tuning on the relighting dataset. Since IC-Light does not release training code, we reimplement the method for our experiments.

We evaluate relighting performance using four categories of metrics: pixel-level fidelity, perceptual similarity, distribution-level realism, and lighting-specific

Table 1: Quantitative comparison results.

Methods	SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	IE $\downarrow$
IC-light	0.7436	15.3314	0.2209	47.3196	15.3575
LBM	0.7908	15.5954	0.2267	40.8663	16.9276
FlowEdit	0.7133	13.6775	0.3298	78.1236	23.2227
Intrinsic Edit	0.7148	15.4219	0.2646	57.3994	14.2292
Latent Intrinsic	0.8425	18.1809	0.2379	35.0617	16.2387
Nano Banana	0.7339	14.9196	0.2507	46.5012	15.9686
Flux-Kontext (original)	0.6386	14.1543	0.2700	53.3793	18.2834
Qwen-Image-Edit (w.o. CFT)	0.8820	20.1170	0.1140	30.2147	11.5505
Flux-Kontext (w.o. CFT)	0.9153	22.9192	0.0985	20.3817	7.8155
Qwen-Image-Edit (w. CFT)	0.8894	20.9176	0.1132	25.5529	10.8670
Flux-Kontext (w. CFT)	0.9202	23.5105	0.0946	18.7338	7.4231

evaluation. For pixel-level fidelity, Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) are used to measure reconstruction accuracy and structural consistency of relighting images. For perceptual similarity, Learned Perceptual Image Patch Similarity (LPIPS) is adopted to quantify perceptual differences in feature space, reflecting human-perceived visual similarity. For distribution-level realism, Fréchet Inception Distance (FID) is employed to evaluate the realism of generated images by measuring the distributional distance between generated samples and ground-truth target images. For lighting-specific evaluation, Estimated Irradiance (EI) is leveraged to compute the MAE between the predicted and ground-truth irradiance maps.

We implement Flux-Kontext and Qwen-Image-Edit-2509 as two base models for our proposed method. We set the hyperparameter $\alpha = 0.1$. We use a learning rate of 1×10^{-4} and a batch size of 4. We train the models for 4,000 steps on 8 NVIDIA H20 GPUs. Except for the training-free baselines and the original pretrained models, we train all the methods on our constructed dataset. Unless otherwise specified, all experiments follow the same training configuration to ensure fair comparison across methods.

4.2 Main Results

Table 1 presents quantitative comparison results with representative relighting approaches. Compared with control-based editing baselines, CFT consistently improves performance when applied to both Flux-Kontext and Qwen-Image-Edit. By explicitly learning the source-to-target transport pathway during training, the model better captures illumination-consistent features while preserving other non-lighting content, such as identity and scene structure. The improvement on both Flux-Kontext and Qwen-Image-Edit also demonstrates the general applicability of CFT to flow-based image editing models. Moreover, our method also substantially outperforms IC-light. IC-light requires additional control signals, such as environment maps, during training. Constructing such auxiliary conditions is often labor-intensive and time-consuming. Our method directly

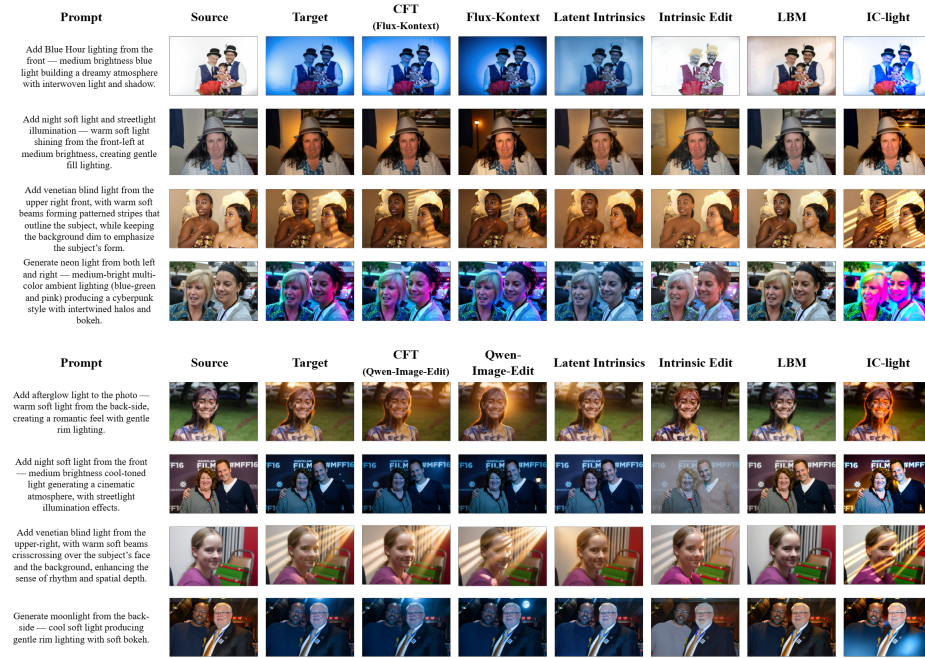


Fig. 4: Qualitative comparison results.

learns illumination feature transport from paired data, leading to a more efficient training process. Compared with decomposition-based editing methods (Intrinsic Edit and Latent Intrinsic), control-based editing approaches generally achieve stronger performance. This may be attributed to the inaccuracy in the intrinsic decomposition and lighting-related intrinsic editing, which can propagate errors to the final relighting results, leading to inconsistent shading, color artifacts, and loss of fine illumination details. Compared with image-to-image translation methods (LBM, FlowEdit), control-based editing approaches also demonstrate superior performance. Instance-level translation does not explicitly model illumination-specific feature transformations, which may lead to suboptimal results, especially for images with complex content or structures. Compared with the original pretrained model (Flux-Kontext, Nano Banana), control-based editing approaches consistently exhibit better performance. This observation highlights the importance of training on relighting-specific data, as pretrained parameters alone are not sufficient to effectively handle relighting tasks. Overall, the results validate the effectiveness of our proposed method and demonstrate the strong generalization of CFT across various image editing models.

Figure 4 presents qualitative comparison results under multiple lighting effects. These examples involve varying degrees of spatially structured illumination, color shifts, and multi-source interactions, providing a comprehensive evaluation of relighting capability. Compared with the respective base models (Flux-

Table 2: Transfer learning results.

Methods	SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	IE $\downarrow$
Flux-Kontext (original)	0.5431	13.5796	0.2331	50.7431	29.3547
Flux-Kontext (w.o. CFT)	0.6924	14.4246	0.2031	44.2741	28.9650
Flux-Kontext (w. CFT)	0.7325	15.0590	0.1917	43.5097	27.3482

Kontext and Qwen-Image-Edit), CFT produces more consistent and physically coherent lighting transformations. For other representative baseline methods, IC-light can follow prompt instructions, but may introduce color artifacts and limited physical coherence. Intrinsic Edit and Latent Intrinsic sometimes fail to preserve content fidelity or follow the editing descriptions, likely due to the inaccurate intrinsic decomposition and editing. LBM, while capable of preserving the original content of the image, often produces a rather subtle relighting effect, especially for those images with a relatively complex structure. Overall, the qualitative results demonstrate that the proposed consistent feature transport mechanism improves illumination realism, spatial coherence, and non-lighting content preservation across diverse and challenging relighting scenarios. The improvements are particularly pronounced in complex lighting settings.

4.3 Further Analysis

To further demonstrate the effectiveness of our method, we conduct a transfer learning experiment on the real-world Multi-Illumination Dataset [35]. Specifically, the Flux-Kontext model trained on our constructed dataset with CFT is directly evaluated on the test set of the Multi-Illumination Dataset without any additional fine-tuning. For comparison, we also evaluate two baselines under the same setting: (1) a Flux-Kontext model trained on our constructed dataset without CFT, and (2) the original pretrained Flux-Kontext model. The results are shown in Table 2. Our constructed dataset consistently improves model performance on previously unseen real-world data, indicating that the constructed dataset contains meaningful lighting features that generalize well to real-world data. Furthermore, applying CFT leads to additional performance gains, suggesting that CFT enables the model to better capture lighting-related features and transfer the learned lighting knowledge across datasets.

Relighting is highly relevant to human perception. Therefore, we conduct a user study to evaluate the generated images from a subjective perspective. Specifically, we randomly sample 40 image pairs from the test set and invite 30 volunteers to evaluate the results produced by different methods. Participants are asked to rate each image on a scale from 1 to 5 according to four criteria: Illumination Quality (IQ), Content Consistency (CC), Physical Plausibility (PP), and Image Aesthetics (IA). Table 3 shows the results. Our method achieves the highest scores across all evaluation dimensions on both Flux-Kontext and Qwen-Image-Edit backbones, further validating the effectiveness of CFT from a human perceptual perspective.

Table 3: User study results.

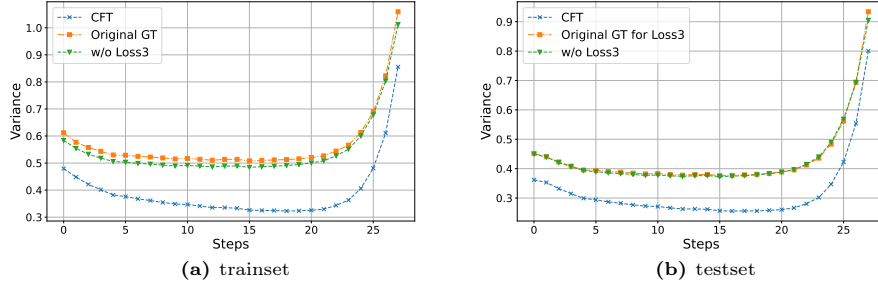
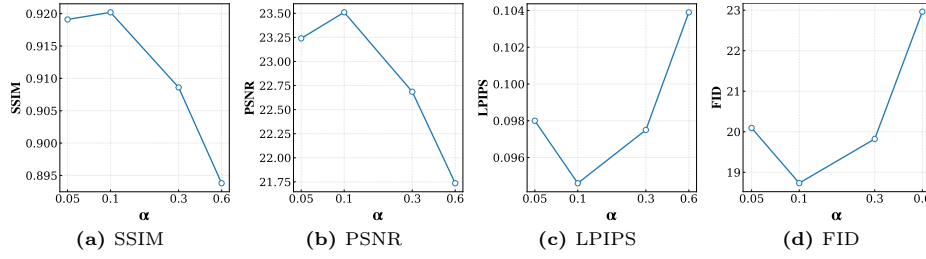
Method	IQ $\uparrow$	CC $\uparrow$	PP $\uparrow$	PA $\uparrow$
IC-light	2.0636	2.2227	2.0136	2.0091
LBM	1.9864	2.5455	2.3864	2.2682
IntrinsicEdit	2.1864	2.2136	2.2727	2.2045
Latent Intrinsic	2.5091	2.9227	2.8227	2.7045
Qwen-Image-Edit (w.o. CFT)	3.5818	3.8091	3.6455	3.4727
Flux-Kontext (w.o. CFT)	3.5950	3.7025	3.6529	3.5620
Qwen-Image-Edit (w. CFT)	4.0091	4.1515	4.1616	4.0101
Flux-Kontext (w. CFT)	4.1414	4.3030	4.2828	4.1919

Table 4: Ablation study results.

Methods	SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$
Loss1	0.9153	22.9192	0.0985	20.3817
Loss1+Loss2	0.9141	22.8299	0.1018	21.6054
Loss1+Loss3	0.9161	23.3778	0.0972	19.3191
Original pairs as gt for Loss3	0.9141	22.8031	0.1003	20.2961
Random pairs as gt for Loss3	0.9135	22.7958	0.1035	19.5538
CFT	0.9202	23.5105	0.0946	18.7338

We conduct an ablation study to verify the necessity of each component in the CFT framework. Table 4 presents the ablation results. The baseline using only $\mathcal{L}_1$ corresponds to standard rectified flow training. Only adding $\mathcal{L}_2$, the noise-to-source reconstruction, does not improve performance and slightly degrades across all metrics, indicating that enforcing source reconstruction alone is insufficient to enhance relighting quality. In contrast, incorporating $\mathcal{L}_3$ together with $\mathcal{L}_1$ leads to consistent improvements across all metrics, demonstrating that modeling the direct source-to-target transport is beneficial for capturing illumination-consistent features. However, compared with the full CFT formulation, the objects without $\mathcal{L}_2$ still show inferior performance, highlighting the necessity of introducing $\mathcal{L}_2$ on top of $\mathcal{L}_1 + \mathcal{L}_3$. By encouraging more accurate reconstruction of the source image, $\mathcal{L}_2$ provides additional structural preservation that helps the model better learn illumination feature transport between the source and target domains.

To further validate the effectiveness of the feature transport between the source and target image distribution, we conduct additional ablation experiments focusing on the supervision of $\mathcal{L}_3$. In our formulation, the supervision for $\mathcal{L}_3$ is constructed from another image pair $(x'_{\text{src}}, x'_{\text{tgt}})$ sharing the same illumination transformation. In this experiment, we replace the ground truth of $\mathcal{L}_3$ with (1) the original source-target image pair $(x_{\text{src}}, x_{\text{tgt}})$ and (2) randomly sampled image pairs $(x_{\text{src}}^{\text{rand}}, x_{\text{tgt}}^{\text{rand}})$. Table 4 shows the results. We observe that both alternative ground truths lead to inferior performance, and even underperform the baseline trained with $\mathcal{L}_1$ alone. Using the original source-target image pair $(x_{\text{src}}, x_{\text{tgt}})$ only encourages the model to learn instance-level transforma-

**Fig. 5:** Variance of the predicted velocity.**Fig. 6:** The impact of α .

tions, which is less effective for learning meaningful feature transformations. Randomly sampled pairs $(x_{\text{src}}^{\text{rand}}, x_{\text{tgt}}^{\text{rand}})$ fail to provide meaningful cues for the relighting transformation. In contrast, CFT leverages image pairs sharing the same illumination transformation as supervision, which significantly improves all evaluation metrics. This result validates the effectiveness of modeling illumination feature transport across instances.

To analyze the transport consistency, we compute the variance of the predicted source-to-target velocity v_t^{direct} at each step t on both the training and testing sets. We compare three settings: (1) Only using $\mathcal{L}_1 + \mathcal{L}_2$ without $\mathcal{L}_3$, (2) replacing the ground truth of $\mathcal{L}_3$ with the original image pair $(x_{\text{src}}, x_{\text{tgt}})$, and (3) the full CFT formulation. As shown in Figure 5, the full CFT method consistently exhibits the lowest variance. It means that CFT learns a more coherent and consistent feature transport field along the source-to-target transformation.

We further conduct a hyperparameter study on the weighting factor α that balances the feature transport loss $\mathcal{L}_3$ and the noise-to-image generation objectives $\mathcal{L}_1 + \mathcal{L}_2$. Specifically, we evaluate $\alpha \in \{0.05, 0.1, 0.3, 0.6\}$, and the results are shown in Figure 6. From the results, we observe a trade-off between the noise-to-image generation and the illumination-consistent feature transport. When α is too small (*e.g.*, 0.05), the influence of the transport supervision becomes limited, and the model mainly behaves like standard conditional generation, leading to weaker illumination-consistent transformations. In contrast, when α becomes too large (*e.g.*, 0.3 or 0.6), the feature transport starts to dominate the optimization process and interferes with the original generative dynamics, which degrades both reconstruction fidelity and perceptual quality. The best perfor-

Table 5: Results for the Style Transfer task.

Models	SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$
QmniStyle (w.o. CFT)	0.5222	13.804	0.4125	82.2332
Qwen-Image-Edit (w.o. CFT)	0.4422	12.9693	0.3697	78.7016
OmniStyle (w. CFT)	0.5275	14.1224	0.3945	85.6465
Qwen-Image-Edit (w. CFT)	0.4648	13.3707	0.3562	81.5195

mance is achieved when $\alpha = 0.1$, which consistently yields the highest SSIM and PSNR as well as the lowest LPIPS and FID. This result indicates that a moderate transport weight provides an effective balance, allowing the model to learn illumination-consistent feature transport while preserving the stability of the underlying diffusion generation process.

4.4 Generalizing to the Style Transfer Task

While our method is designed for image relighting, the proposed consistent feature transport formulation is not limited to relighting editing and can be applied to other image editing tasks. To evaluate its generality, we conduct additional experiments on image style transfer. We construct a style transfer dataset based on OmniStyle-150k [42] by randomly selecting 100 style categories. The resulting dataset contains 10,327 training samples and 500 testing samples. Each sample is organized as a triplet $(x_{\text{src}}, x_{\text{style}}, x_{\text{tgt}})$, where x_{src} is the content source image, x_{style} is the reference style image, and x_{tgt} is the corresponding stylized target image. For evaluation, we adopt OmniStyle and Qwen-Image-Edit-2509 as base models. We apply the proposed CFT method to both backbones under the same training settings as in the relighting task, except that the batch size is set to 2.

Table 5 reports the results for the style transfer task. Applying CFT consistently improves performance across different base models. Specifically, CFT achieves higher PSNR and SSIM and lower LPIPS on both OmniStyle and Qwen-Image-Edit, indicating that explicitly modeling feature transport helps the model better preserve content structure and improve perceptual fidelity while applying style-related transformations. However, the FID scores of the CFT results are slightly worse than those of the original baselines. This suggests a trade-off between transport consistency and stylistic diversity. Since style transfer is inherently a one-to-many generation task, where multiple stylizations may be equally plausible for the same source image, the consistent transport supervision in CFT encourages more deterministic feature transport and stronger structural alignment between source and target pairs. While this improves reconstruction fidelity and perceptual consistency, it may slightly reduce the global distribution diversity. Overall, the above results validate that the idea of consistent feature transport can be extended to other image editing tasks, such as style transfer.

We present a qualitative comparison for the style transfer task in Figure 7. As shown in the figure, the original base models often produce style transfer results that are either overly stylized or insufficiently stylized. By employing the

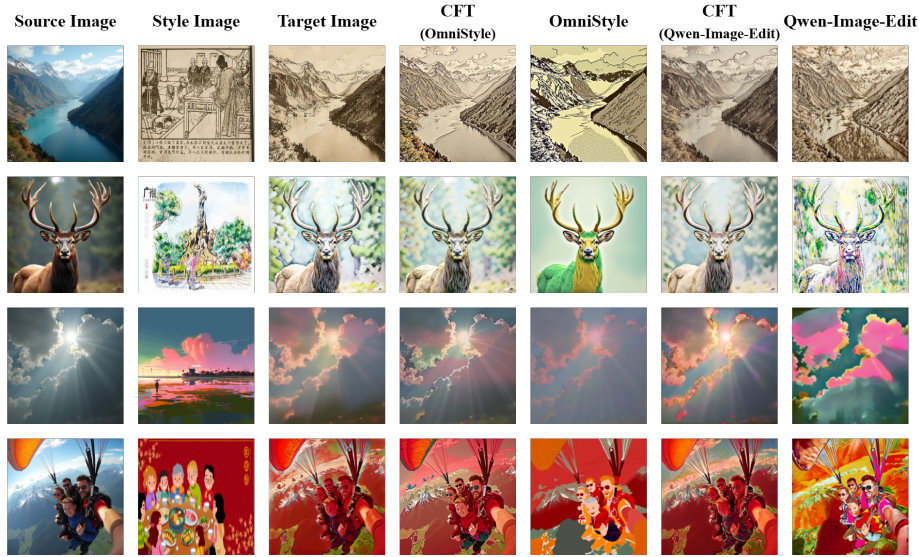


Fig. 7: Qualitative comparison results for the style transfer task.

proposed consistent feature transport (CFT), the generated images exhibit a more balanced style transformation. Our method better preserves the structural content of the source image while transferring the key stylistic patterns from the reference style image, resulting in visually coherent and stylistically faithful outputs across different scenes and styles.

5 Conclusion

In this work, we propose a Consistent Feature Transport (CFT) method for image relighting. By jointly modeling noise-to-image generation and illumination-consistent source-to-target transport, CFT enables the model to focus on illumination feature transformation while preserving non-lighting content. In addition, we construct a portrait relighting dataset with complex illumination effects, addressing the scarcity of suitable data resources. Extensive experiments demonstrate that the CFT achieves consistent improvements over existing relighting approaches in both quantitative and qualitative evaluations. Furthermore, experiments on style transfer show that the proposed feature transport mechanism generalizes beyond relighting and can be applied to broader image editing tasks. In the future, we plan to apply CFT to additional editing scenarios, exploring its effectiveness for modeling feature transport across different editing tasks.

Acknowledgements

This work was supported by the General Program of Beijing Natural Science Foundation under Grant 4262030.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Bharadwaj, S., Feng, H., Becherini, G., Fernandez Abrevaya, V., Black, M.J.: Genlit: Reformulating single-image relighting as video generation. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–12 (2025)
3. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
4. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22560–22570 (2023)
5. Chadebec, C., Tasar, O., Sreetharan, S., Aubin, B.: Lbm: Latent bridge matching for fast image-to-image translation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 29086–29098 (2025)
6. Chaturvedi, S., Ren, M., Hold-Geoffroy, Y., Liu, J., Dorsey, J., Shu, Z.: Synthlight: Portrait relighting with diffusion model by learning to re-render synthetic faces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 369–379 (2025)
7. Chen, C., Guo, P., Song, L., Lu, J., Qian, R., Fu, T.J., Wang, X., Liu, W., Yang, Y., Schwing, A.: Car-flow: Condition-aware reparameterization aligns source and target for better flow matching. *Advances in Neural Information Processing Systems* **38**, 94919–94945 (2026)
8. Choi, J.M., Wang, A., Peers, P., Bhattad, A., Sengupta, R.: Scribblelight: Single image indoor relighting with scribbles. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5720–5731 (2025)
9. Choi, Y., Kwak, S., Lee, K., Choi, H., Shin, J.: Improving diffusion models for authentic virtual try-on in the wild. In: European Conference on Computer Vision. pp. 206–235. Springer (2024)
10. Dalva, Y., Venkatesh, K., Yanardag, P.: Fluxspace: Disentangled semantic editing in rectified flow models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13083–13092 (2025)
11. Debevec, P., Hawkins, T., Tchou, C., Duiker, H.P., Sarokin, W., Sagar, M.: Acquiring the reflectance field of a human face. In: Proceedings of the 27th annual conference on Computer graphics and interactive techniques. pp. 145–156 (2000)
12. Deng, Y., He, X., Mei, C., Wang, P., Tang, F.: Fireflow: Fast inversion of rectified flow for image semantic editing. In: Forty-second International Conference on Machine Learning (2025)
13. Hasinoff, S.W., Sharlet, D., Geiss, R., Adams, A., Barron, J.T., Kainz, F., Chen, J., Levoy, M.: Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (ToG)* **35**(6), 1–12 (2016)
14. Helou, M.E., Zhou, R., Barthas, J., Süssstrunk, S.: Vedit: Virtual image dataset for illumination transfer. arXiv preprint arXiv:2005.05460 (2020)
15. Hold-Geoffroy, Y., Athawale, A., Lalonde, J.F.: Deep sky modeling for single image outdoor lighting estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6927–6935 (2019)

16. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
17. Kalantari, N.K., Ramamoorthi, R., et al.: Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.* **36**(4), 144–1 (2017)
18. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 4401–4410 (2019)
19. Kim, H., Jang, M., Yoon, W., Lee, J., Na, D., Woo, S.: 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*. pp. 25096–25106 (2024)
20. Kim, J., Hong, Y., Park, J., Ye, J.C.: Flowalign: Trajectory-regularized, inversion-free flow-based image editing. In: *The Fourteenth International Conference on Learning Representations* (2026)
21. Kocsis, P., Philip, J., Sunkavalli, K., Nießner, M., Hold-Geoffroy, Y.: Lightit: Illumination modeling and control for diffusion models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9359–9369 (2024)
22. Kocsis, P., Sitzmann, V., Nießner, M.: Intrinsic image diffusion for indoor single-view material estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5198–5208 (2024)
23. Kulikov, V., Kleiner, M., Huberman-Spiegelglas, I., Michaeli, T.: Flowedit: Inversion-free text-based editing using pre-trained flow models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 19721–19730 (2025)
24. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontekst: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
25. Li, Z., Shi, J., Bi, S., Zhu, R., Sunkavalli, K., Hašan, M., Xu, Z., Ramamoorthi, R., Chandraker, M.: Physically-based editing of indoor scene lighting from a single image. In: *European Conference on Computer Vision*. pp. 555–572. Springer (2022)
26. Li, Z., Zhu, L., Shen, H., Ran, L., Liu, W., Wang, X.: Translight: Image-guided customized lighting control with generative decoupling. arXiv preprint arXiv:2508.14814 (2025)
27. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: *The Eleventh International Conference on Learning Representations* (2023)
28. Liu, I., Chen, L., Fu, Z., Wu, L., Jin, H., Li, Z., Wong, C.M.R., Xu, Y., Ramamoorthi, R., Xu, Z., et al.: Openillumination: A multi-illumination dataset for inverse rendering evaluation on real objects. *Advances in Neural Information Processing Systems* **36**, 36951–36962 (2023)
29. Liu, P., Yuan, H., Dong, B., Xing, J., Wang, J., Zhao, R., Chen, W., Wang, F.: Unilumos: Fast and unified image and video relighting with physics-plausible feedback. *Advances in Neural Information Processing Systems* **38**, 82052–82080 (2026)
30. Liu, X., Gong, C., qiang liu: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: *The Eleventh International Conference on Learning Representations* (2023)
31. Liu, Y., Xiao, W., Wang, Q., Chen, J., Wang, S., Wang, Y., Wu, X., Tang, Y.: Dreamlight: Towards harmonious and consistent image relighting. arXiv preprint arXiv:2506.14549 (2025)

32. Luo, X., Wang, J., Wu, C., Xiao, S., Jiang, X., Lian, D., Zhang, J., Liu, D., Liu, Z.: Editscore: Unlocking online RL for image editing via high-fidelity reward modeling. In: The Fourteenth International Conference on Learning Representations (2026)
33. Lyu, L., Deschaintre, V., Hold-Geoffroy, Y., Hašan, M., Yoon, J.S., Leimkühler, T., Theobalt, C., Georgiev, I.: Intrinsicedit: Precise generative image manipulation in intrinsic space. *ACM Transactions on Graphics (TOG)* **44**(4), 1–13 (2025)
34. Magar, N., Hertz, A., Tabellion, E., Pritch, Y., Rav-Acha, A., Shamir, A., Hoshen, Y.: Lightlab: Controlling light sources in images with diffusion models. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–11 (2025)
35. Murmann, L., Gharbi, M., Aittala, M., Durand, F.: A dataset of multi-illumination images in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4080–4089 (2019)
36. Pan, Y., Bareinboim, E.: Counterfactual image editing with disentangled causal latent space. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
37. Pandey, R., Orts-Escolano, S., Legendre, C., Haene, C., Bouaziz, S., Rhemann, C., Debevec, P.E., Fanello, S.R.: Total relighting: learning to relight portraits for background replacement. *ACM Trans. Graph.* **40**(4), 43–1 (2021)
38. Ren, M., Xiong, W., Yoon, J.S., Shu, Z., Zhang, J., Jung, H., Gerig, G., Zhang, H.: Relightful harmonization: Lighting-aware portrait background replacement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6452–6462 (2024)
39. Tan, Z., Wang, Z., Yang, X., Liu, S., Wang, X.: Vision bridge transformer at scale. *arXiv preprint arXiv:2511.23199* (2025)
40. Wang, J., Pu, J., Qi, Z., Guo, J., Ma, Y., Huang, N., Chen, Y., Li, X., Shan, Y.: Taming rectified flow for inversion and editing. In: Forty-second International Conference on Machine Learning (2025)
41. Wang, J., Liu, J., Sun, X., Singh, K.K., Shu, Z., Zhang, H., Yang, J., Zhao, N., Wang, T.Y., Chen, S.S., et al.: Comprehensive relighting: Generalizable and consistent monocular human relighting and harmonization. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 380–390 (2025)
42. Wang, Y., Liu, R., Lin, J., Liu, F., Yi, Z., Wang, Y., Ma, R.: Omnistyle: Filtering high quality style transfer data at scale. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7847–7856 (2025)
43. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. *arXiv preprint arXiv:2508.02324* (2025)
44. Xing, X., Groh, K., Karaoglu, S., Gevers, T., Bhattad, A.: Luminet: Latent intrinsics meets diffusion models for indoor scene relighting. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 442–452 (2025)
45. Yang, Y., Peng, H., Shen, Y., Yang, Y., Hu, H., Qiu, L., Koike, H., et al.: Imagebrush: Learning visual in-context instructions for exemplar-based image manipulation. *Advances in Neural Information Processing Systems* **36**, 48723–48743 (2023)
46. Yu, Y., Meka, A., Elgharib, M., Seidel, H.P., Theobalt, C., Smith, W.A.: Self-supervised outdoor scene relighting. In: European Conference on Computer Vision. pp. 84–101. Springer (2020)
47. Zhang, B., Xie, H., Gao, Z., Wang, Y.: Choose what you need: Disentangled representation learning for scene text recognition removal and editing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 28358–28368 (2024)

- 48. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems* **36**, 31428–31449 (2023)
- 49. Zhang, L., Rao, A., Agrawala, M.: 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)
- 50. Zhang, X., Gao, W., Jain, S., Maire, M., Forsyth, D., Bhattad, A.: Latent intrinsics emerge from training to relight. *Advances in Neural Information Processing Systems* **37**, 96775–96796 (2024)