RELIT-LIVE: Relight Video by Jointly Learning Environment Video

WEIQING XIAO*†, Nanjing University, China
HONG LI*†, BAAI, China and BUAA, China
XIUYU YANG*, Tsinghua University, China
HOUYUAN CHEN, The Hong Kong University of Science and Technology, China
WENYI LI, University of Chinese Academy of Sciences, China
TIANQI LIU, Huazhong University of Science and Technology, China
SHAOCONG XU, BAAI, China
CHONGJIE YE, The Chinese University of Hong Kong, Shenzhen, China
HAO ZHAO‡, Tsinghua University, China and BAAI, China
BEIBEI WANG‡, Nanjing University, China

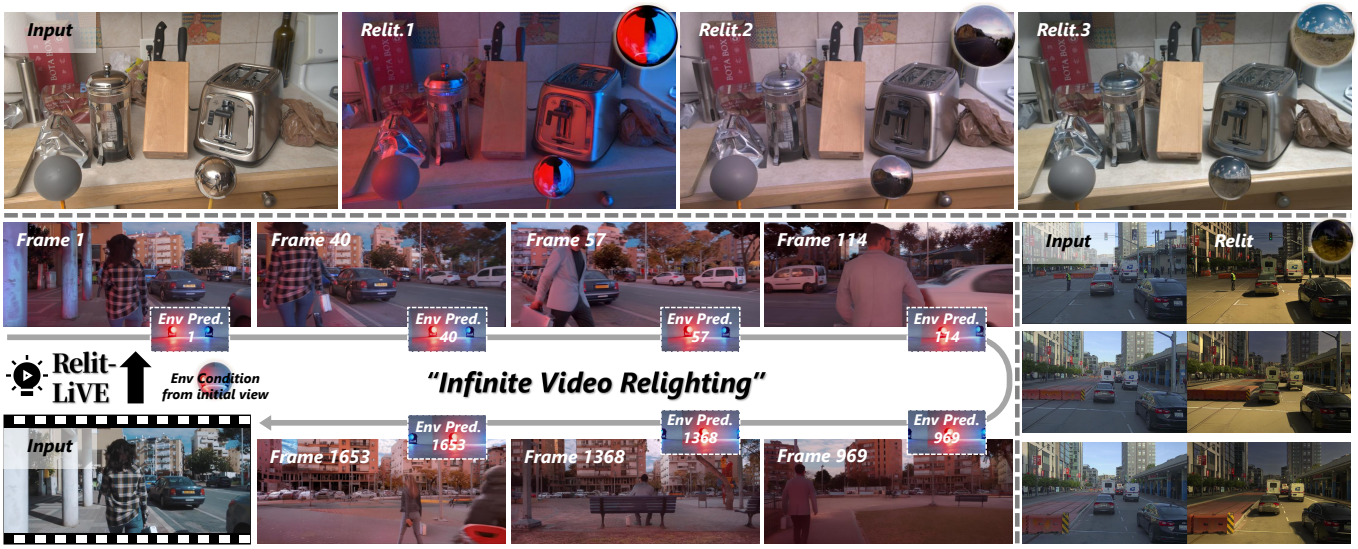

Fig. 1. We present **RELIT-LIVE**, a novel video relighting framework that produces physically consistent and temporally stable results without needing prior knowledge of camera pose. This is achieved by jointly generating relighting videos and environment videos. Additionally, by integrating real-world lighting effects with intrinsic constraints, the relighting videos demonstrate remarkable physical plausibility, showcasing realistic reflections and shadows.

*These authors contributed equally to this work.
†Work partially done during an internship at Beijing Academy of Artificial Intelligence.
‡Corresponding authors.

Authors' Contact Information: Weiqing Xiao, weiqing001@smail.nju.edu.cn, Nanjing University, Suzhou, China; Hong Li, link0502@buaa.edu.cn, BAAI, Beijing, China and BUAA, Beijing, China; Xiuyu Yang, gzzyyxy@gmail.com, Tsinghua University, Beijing, China; Houyuan Chen, houyuanchen111@gmail.com, The Hong Kong University of Science and Technology, Hong Kong, China; Wenyi Li, liwenyi19@mails.ucas.ac.cn, University of Chinese Academy of Sciences, Beijing, China; Tianqi Liu, tq_liu@hust.edu.cn, Huazhong University of Science and Technology, Beijing, China; Shaocong Xu, daniellesry@gmail.com, BAAI, Beijing, China; Chongjie Ye, chongjieye@link.cuhk.edu.cn, The Chinese University of Hong Kong, Shenzhen, Shenzhen, China; Hao Zhao, zhaohao@air.tsinghua.edu.cn, Tsinghua University, Beijing, China and BAAI, Beijing, China; Beibei Wang, beibei.wang@nju.edu.cn, Nanjing University, Suzhou, China.

SIGGRAPH Conference Papers '26, Los Angeles, CA, USA

Recent advances have shown that large-scale video diffusion models can be repurposed as neural renderers by first decomposing videos into intrinsic scene representations and then performing forward rendering under novel illumination. While promising, this paradigm fundamentally relies on accurate intrinsic decomposition, which remains highly unreliable for real-world videos and often leads to distorted appearances, broken materials, and accumulated temporal artifacts during relighting. In this work, we present **RELIT-LIVE**, a novel video relighting framework that produces physically consistent, temporally stable results without requiring prior knowledge of camera pose. Our key insight is to explicitly introduce raw reference images into the rendering process, enabling the model to recover critical scene cues that are inevitably lost or corrupted in intrinsic representations. Furthermore, we propose a novel environment video prediction formulation that simultaneously generates relit videos and per-frame environment maps aligned with each camera viewpoint in a single diffusion process. This

ACM ISBN 979-8-4007-2554-8/2026/07
https://doi.org/10.1145/3799902.3811200

joint prediction enforces strong geometric–illumination alignment and naturally supports dynamic lighting and camera motion, significantly improving physical consistency in video relighting while easing the requirement of known per-frame camera pose. To further enhance generalization, we introduce two complementary training strategies: (i) latent-space interpolation between relighting and rendering outputs to synthesize diverse, photorealistic multi-illumination data, and (ii) a cycle-consistent self-supervised illumination learning scheme that enforces temporal lighting coherence without additional annotations. Extensive experiments demonstrate that Relit-LiVE consistently outperforms state-of-the-art video relighting and neural rendering methods across synthetic and real-world benchmarks. Beyond relighting, our framework naturally supports a wide range of downstream applications, including scene-level rendering, material editing, object insertion, and streaming video relighting. The Project is available at https://github.com/zhuxing0/Relit-LiVE.

CCS Concepts: • **Computing methodologies** → **Rendering**; **Computer vision**.

ACM Reference Format:
Weiqing Xiao, Hong Li, Xiuyu Yang, Houyuan Chen, Wenyi Li, Tianqi Liu, Shaocong Xu, Chongjie Ye, Hao Zhao, and Beibei Wang. 2026. Relit-LiVE: Relight Video by Jointly Learning Environment Video. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '26), July 19–23, 2026, Los Angeles, CA, USA.* ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3799902.3811200

1 Introduction

Video relighting aims to modify a video's illumination while preserving the scene's intrinsic properties. It has various applications, including content creation, creative editing, and robust vision systems. However, it remains a long-standing challenge to achieve physically consistent and temporally accurate lighting effects, such as realistic reflections or stable, time-coherent shadows. Addressing this requires not only accounting for different material properties but also precise, controllable modeling of lighting conditions.

Building upon powerful pre-trained diffusion models, several studies [Liu et al. 2025a; Zhou et al. 2025] directly generate relit videos using text prompts or background images as lighting conditions. While achieving breakthroughs in visual quality, these methods typically lack precise lighting control and often retain artifacts from the original illumination. In contrast to direct generation, another line of research [Fang et al. 2025b; Liang et al. 2025] explores a two-stage architecture that incorporates an intermediate step of intrinsic decomposition. This approach first separates scene intrinsics from illumination, then performs relighting synthesis based on these components, using environment maps for conditioning. This explicit separation enables a clearer decoupling between scene properties and lighting, facilitating higher visual quality and more precise control. However, this paradigm is heavily dependent on the fidelity of the intermediate intrinsic representation. In challenging scenarios, such as transparent objects with complex light transport or subsurface scattering, neural intrinsic rendering might yield flawed or implausible outputs. A recent work by He et al. [2025b] unifies albedo estimation with direct relighting, synthesizing scene albedo and relighting video in parallel to effectively decouple and reshape scene illumination. However, constrained by the inherent challenges of training parallel inference paradigms, their approach struggles to extend to more intrinsic properties, limiting its capabilities. Furthermore, these methods require precise prior knowledge of the video camera's pose to position the environment map in the viewport, which constrains their flexibility.

In this paper, we propose Relit-LiVE, a novel video relighting framework that produces physically consistent, temporally stable results without requiring prior knowledge of camera pose. To this end, we address two core challenges: (1) preserving scene content integrity under complex light transport, and (2) flexibly injecting novel lighting conditions without known camera pose. We present two key insights to address these challenges. First, while decomposed intrinsic attributes often struggle to capture complex global illumination effects, these effects are directly observable in the original RGB video sequence. Therefore, we propose an RGB-intrinsic fusion renderer that utilizes the input RGB frames—also known as raw reference images—to guide and correct the rendering process, providing both visual and semantic-level cues. This design fuses the RGB space with the intrinsic space, enabling the model to incorporate real-world lighting effects alongside estimated physical constraints, resulting in realistic relighting results. Second, to facilitate arbitrary relighting without requiring per-frame camera poses, we reformulate relit video learning as the simultaneous learning of a per-frame warping of the environment map in combination with relit video synthesis. This approach allows our model to generate both relit videos and per-frame warped environment maps (referred to as environment video) during a single inference pass. By inferring the lighting transformation implicitly, our approach eliminates the need for explicit pose estimation, enhancing practical flexibility.

Furthermore, we improve the robustness of our model to handle complex scenarios by enhancing the training data in two ways. First, we perform latent-space interpolation between relighting and rendering outputs using the initially trained model. This allows us to synthesize diverse, photorealistic multi-illumination data. Second, we employ a cycle-consistent self-supervised illumination learning scheme that ensures temporal lighting coherence without the need for additional annotations.

Extensive experiments demonstrate that Relit-LiVE outperforms existing state-of-the-art methods, achieving realistic material reflection effects and effectively modeling viewpoint changes in videos. This enables us to perform physically plausible and spatio-temporally accurate relighting of videos without requiring camera pose priors. Relit-LiVE also offers flexibility for task extension, enabling scene-level rendering, editing, and streaming video relighting through modifying generation conditions and intermediate outputs. In summary, our contributions are as follows:

- a novel video relighting framework, Relit-LiVE, that produces physically consistent, temporally stable results without requiring prior knowledge of camera pose,
- an RGB-intrinsic fusion renderer, that effectively integrates real-world lighting effects from the RGB space with physical constraints from the intrinsic space, enabling the generation of physically consistent video lighting effects, and
- jointly generation of relit video and environment video, enabling geometry-illumination aligned video relighting without requiring per-frame camera poses.

2 Related work

2.1 Direct video relighting

Direct video relighting aims to adjust the lighting conditions of a video while preserving the scene content through an end-to-end approach. Driven by breakthroughs in controllable video diffusion technology [Wan et al. 2025; Yang et al. 2025b], this paradigm has achieved rapid development. Overall, the research focus of this paradigm is shifting from the mere pursuit of temporal consistency toward precise lighting control and physical realism.

Some early studies [Fang et al. 2025a,a; Liu et al. 2025a] have focused on achieving temporally consistent relighting, typically using text prompts or reference backgrounds as rough lighting conditions. For instance, methods such as Light-A-Video [Zhou et al. 2025] and TC-Light [Liu et al. 2025a] extend the effects of the image re-illumination technique IC-Light [Zhang et al. 2025] smoothly across entire videos through carefully designed temporal consistency enhancement schemes. Recent research [Liu et al. 2026; Magar et al. 2025; Ren et al. 2025] has increasingly focused on precise control and physical realism in lighting, with representative methods including RelightMaster [Bian et al. 2025], UniLumos [Liu et al. 2025b], and UniRelight [He et al. 2025b]. RelightMaster [Bian et al. 2025] and UniLumos [Liu et al. 2025b] respectively propose multi-plane light images and structured text prompts to achieve fine-grained control over lighting parameters. Additionally, UniLumos incorporates depth and normal geometric feedback supervision to ensure shadow plausibility. UniRelight [He et al. 2025b] jointly learns to directly generate relit videos and albedo estimation. By implicitly decoupling ambient lighting, it enhances lighting effects in complex scenes.

However, this parallel inference pattern presents inherent training challenges: model capacity often limits the scope of tasks it can handle. This constrains the upper bound of the joint estimation paradigm, making it difficult to account for comprehensive intrinsic properties. In contrast, our decoupled approach ensures both the comprehensiveness and expandability of intrinsic content. This also grants our method greater architectural flexibility, supporting not only video relighting but also tasks like neural rendering.

2.2 Intrinsic-aware diffusion model

Inspired by Physically-Based Rendering (PBR) pipelines [Rendering 2015], some research [Beisswenger et al. 2025; Kocsis et al. 2025; Ye et al. 2024] has begun exploring the intrinsic decomposition [Bonneel et al. 2017; Careaga and Aksoy 2023; Shu et al. 2018] and synthesis of images and videos through diffusion models [Chen et al. 2025b]. Compared to end-to-end generation, this paradigm offers high flexibility. By adjusting its intrinsic components, it can perform a variety of functions, including light modification and material editing.

Some approaches [Careaga and Aksoy 2025; Chen et al. 2025a; He et al. 2025a; Kocsis et al. 2025] focus on intrinsic decomposition tasks, with representative methods including IntrinsiX [Kocsis et al. 2025], NormalCrafter [Bin et al. 2025], and GeometryCrafter [Xu et al. 2025]. These methods are based on fine-tuning pre-trained diffusion models. Leveraging the strong generative prior of diffusion models, they achieve precise decomposition of specific intrinsic properties through conditional generation. Other studies [Chen et al. 2025c; Fang et al. 2025b; Liang et al. 2025; Xi et al. 2025] simultaneously focus on both intrinsic decomposition and synthesis tasks to achieve a closed-loop "decomposition-synthesis" capability. For instance, RGBX [Zeng et al. 2024] employs image diffusion models to enable bidirectional functionality: estimating G-buffers from images and rendering images based on G-buffers. Recent work such as the Diffusion Renderer [Liang et al. 2025] and V-RGBX [Fang et al. 2025b] extends this closed-loop architecture from images to the video domain. However, constrained by the inherent challenges of decomposing intrinsic properties in the real world, this "decomposition-synthesis" architecture is often limited to specific domains and prone to cumulative error issues. Additionally, during the compositing stage, such methods typically require precise lighting information, such as irradiance maps or environment maps for all frames. This limits the practicality of its relighting function. In our paper, we propose a novel video relighting framework with two key designs to address the two challenges outlined above.

3 Our method

This paper targets the problem of video relighting, aiming to generate physically consistent and temporally stable results without relying on prior camera pose estimation. In this section, we first formalize the problem and then introduce our proposed framework, Relit-LiVE, as shown in Figure 2.

3.1 Problem statement

For the task of video relighting, we are given a source video sequence $V^s = \{I_i^s\}_{i=1}^n \in \mathbb{R}^{n\times h\times w\times 3}$ and a target lighting sequence $\mathbf{E}^t = \{\mathbf{E}_i\}_{i=1}^n \in \mathbb{R}^{n\times h\times w\times 3}$ (which may be static or dynamic). The objective is to synthesize a target video $V^t = \{I_i^t\}_{i=1}^n$ that faithfully exhibits the original scene content from V^s under the novel illumination $\mathbf{E}^t$, effectively replacing the source lighting. This process can be formulated as:

$$V^t = \mathcal{F}_\theta(V^s, \mathbf{E}^t), \tag{1}$$

where $\mathcal{F}_\theta$ is a relighting network parameterized by θ. In the case of static target lighting, the sequence $\mathbf{E}^t$ reduces to a constant environment map applied to every frame.

3.2 RGB-Intrinsic fusion renderer

Learning the video relighting task directly is challenging because it is inherently difficult to disentangle the intrinsic scene properties from the original lighting conditions. Hence, a common paradigm in video relighting involves first performing an intrinsic decomposition of the source video to separate material properties from illumination, followed by re-rendering the extracted materials under the target lighting. In this view, the renderer serves as a relighting pathway. This paradigm improves physical plausibility, but its performance is critically limited by the accuracy and robustness of the decomposition stage. This limitation becomes particularly apparent in scenes with complex lighting effects, leading to visual artifacts. Thus, the reliance on imperfect intrinsic decomposition remains a core challenge in achieving high-fidelity video relighting. To resolve this issue, we find that these lighting effects are directly observable in the original RGB video. The raw images provide visual and even

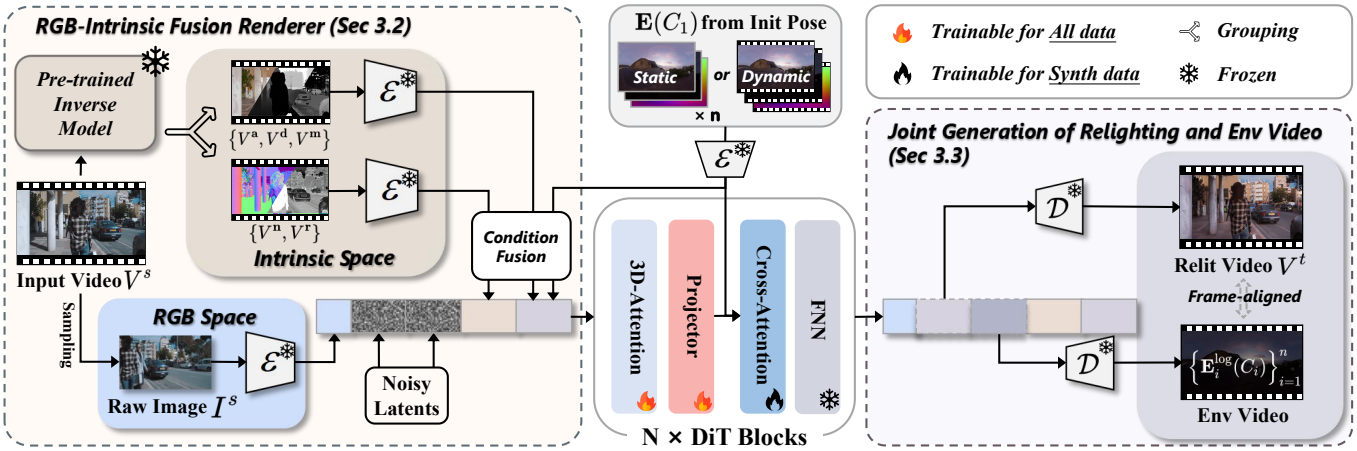

Fig. 2. **Overview of our RELIT-LIVE.** Given an input video and environment maps of the initial viewpoint, our method jointly predicts relit videos and frame-specific environment maps (i.e., environment video). The input video is converted into intrinsic properties by a pre-trained inverse rendering model, then mapped into latent space alongside environment maps and randomly sampled reference images. Subsequently, latents undergo partial grouping fusion and frame-wise concatenation, followed by denoising through the DiT video model to generate realistic relighting video.

semantic-level cues for video rendering tasks in RGB space, while intrinsic properties in G-buffer impose direct physical constraints on relighting results. Therefore, we propose an RGB-Intrinsic fusion renderer, which utilizes this observable RGB information to guide the rendering process, thus bypassing the limitations posed by imperfect intrinsic decomposition.

Given a source video V^s, we utilize the inverse renderer from Diffusion Renderer [Liang et al. 2025] to predict its G-buffers, which include a common set of intrinsic properties: base color $V^{\mathbf{a}}$, surface normal $V^{\mathbf{n}}$, relative depth $V^{\mathbf{d}}$, roughness $V^{\mathbf{r}}$, and metallic $V^{\mathbf{m}}$). We then employ a pretrained VAE encoder $\mathcal{E}$ to encode each G-buffer into te latent space, resulting in the corresponding latents $\{\mathbf{z}^{\mathbf{a}}, \mathbf{z}^{\mathbf{n}}, \mathbf{z}^{\mathbf{d}}, \mathbf{z}^{\mathbf{r}}, \mathbf{z}^{\mathbf{m}}\}$, where $\mathbf{z}^* \in \mathbb{R}^{N\times H\times W\times C}$.

Previous works [Fang et al. 2025b; Liang et al. 2025; Zeng et al. 2024] have directly concatenated these intrinsic latents either along the frame or channel dimension. However, we have observed that concatenating along the frame dimension increases computational overhead, while concatenating along the channel dimension slows down model convergence. To address these issues, we propose to sum the latents partially before concatenating them along the frame dimension. From a pilot study, we identified a key point: separating intrinsic properties that exhibit similar numerical characteristics or strong correlations—such as metallic and roughness, or depth and normal—facilitates precise control over the generated results. The former two are typically represented by grayscale values and demonstrate pronounced regional equivalence, meaning regions with the same material tend to maintain nearly constant values; the latter two exhibit significant numerical correlation. Therefore, we specifically separate these modalities during G-buffer grouping. Specifically, we compute two new sets of latents: $\mathbf{z}^{\{\mathbf{a},\mathbf{d},\mathbf{m}\}} = \mathbf{z}^{\mathbf{a}} + \mathbf{z}^{\mathbf{d}} + \mathbf{z}^{\mathbf{m}}$ and $\mathbf{z}^{\{\mathbf{n},\mathbf{r}\}} = \mathbf{z}^{\mathbf{n}} + \mathbf{z}^{\mathbf{r}}$. These two new latents serve as intrinsic conditions.

Then, we randomly sample a raw image I^s from the input video and use the VAE encoder $\mathcal{E}$ to encode this image, generating the latent $\mathbf{z}^{\mathbf{I}} \in \mathbb{R}^{1\times H\times W\times C}$. This latent representation is concatenated with intrinsic conditions along the frame dimension, effectively guiding the generation process together. This random sampling strategy breaks fixed correspondences between the raw image and generated results, thereby suppressing pixel-level propagation of source lighting. It is worth noting that, since the inference process of diffusion models typically involves multiple denoising steps, we can actually sample different frames during each denoising step to preserve as much detail as possible.

3.3 Joint generation of relighting and environment video

With the encoded features and environment maps, we could render them using a DiT video model to generate the relit video. Since $\mathcal{F}_\theta$ operates in 2D image space, the environment maps $\{\mathbf{E}_i\}_{i=1}^n$ must be appropriately aligned with the camera's viewing direction. Here, we set $\{\mathbf{E}_i\}_{i=1}^n = \{\mathbf{E}_i(C_i)\}_{i=1}^n$ to highlight this operation, where C_i represents the i-th camera viewpoint. While the source video inherently defines the camera poses, these poses are often unknown or inaccurately estimated in practice. Existing methods often assume known camera poses, allowing for direct warping of the environment map into camera space. However, this assumption limits their real-world applicability. To address this issue, we propose learning warped environment maps (referred to herein as environment videos) along with the relit video. This way, the DiT model can be forced to learn render the scene with the warped environment maps. By implicitly inferring lighting transformations, we eliminate the need for explicit pose estimation, enhancing practical usability while ensuring spatio-temporal lighting accuracy.

We start by reformulating our relight task into the joint generation of the relit video and the warped environment video.

$$\begin{aligned} V^t, \{\mathbf{E}_i(C_i)\}_{i=1}^n &= \mathcal{F}_\theta\left(V^s, \{\mathbf{E}_i(C_1)\}_{i=1}^n\right) \\ &= \mathcal{F}_\theta\left(I^s, V^{\mathbf{a}}, V^{\mathbf{n}}, V^{\mathbf{d}}, V^{\mathbf{r}}, V^{\mathbf{m}}, \{\mathbf{E}_i(C_1)\}_{i=1}^n\right). \end{aligned} \tag{2}$$

In the above equation, we also incorporate intrinsic properties along with the raw reference image introduced in the previous section. Next, we describe our lighting conditions, followed by the joint generation.

We use HDR environment maps $\mathbf{E}(C_1)$ under the initial viewpoint C_1 to represent lighting condition (which may be static or dynamic). Inspired from prior works [Liang et al. 2025], we construct three complementary representations for HDR environment maps: 1) LDR images $\mathbf{E}^{\text{ldr}}(C_1)$ obtained via Reinhard tonemapping; 2) normalized log-intensity images $\mathbf{E}^{\log}(C_1) = \log(1+\mathbf{E}(C_1))/\log(1+M)$, where $M = 60000$; 3) directional encoding images $\mathbf{E}^{\text{dir}}$, where each pixel represents the direction of the corresponding ray in the camera coordinate system (note that the pixel direction here is opposite to that in standard panoramas). We use the VAE encoder $\mathcal{E}$ to encode these three representations into the latent space separately and concatenate them along the channel dimension to obtain $\mathbf{h}_\mathbf{E} = \left\{\mathcal{E}(\mathbf{E}^{\text{ldr}}(C_1)), \mathcal{E}(\mathbf{E}^{\log}(C_1)), \mathcal{E}(\mathbf{E}^{\text{dir}})\right\} \in \mathbb{R}^{N\times H\times W\times 3C}$. Then, we process the $\mathbf{h}_\mathbf{E}$ using a convolutional layer with a stride of 1 to obtain $\mathbf{c}_\mathbf{E} \in \mathbb{R}^{N\times H\times W\times C}$, which is concatenated with other conditional latents. Additionally, we repeat this process at an input resolution of 512×256, feeding the result $\mathbf{c}_\mathbf{E}^{\text{cross}}$ separately into the cross-attention module as enhanced lighting control.

Then, our simultaneously generates relit video V^t and corresponding environment video (in the form of normalized log intensity maps $\left\{\mathbf{E}_i^{\log}(C_i)\right\}_{i=1}^{n}$, as they can be inverse-transformed back to HDR and LDR maps) using multiple DiT blocks. During training, we encode both into the latent space using the VAE encoder $\mathcal{E}$, yielding $\mathbf{z}^\mathbf{t}$ and $\mathbf{z}^{\mathbf{E}_{\log}}$. Subsequently, noise is independently introduced to generate $\mathbf{z}_\tau^\mathbf{t}$ and $\mathbf{z}_\tau^{\mathbf{E}_{\log}}$. Next, we concatenate these noise-added target latents with the reference latent $\mathbf{z}^\mathbf{I}$, intrinsic latents $\left\{\mathbf{z}^{\{\mathbf{a},\mathbf{d},\mathbf{m}\}}, \mathbf{z}^{\{\mathbf{n},\mathbf{r}\}}\right\}$, and lighting conditions $\left\{\mathbf{c}_\mathbf{E}, \mathbf{c}_\mathbf{E}^{\text{cross}}\right\}$ at the frame level, and feed them into DiT blocks to learn denoising:

$$\hat{\mathbf{z}}^\mathbf{t}(\theta), \hat{\mathbf{z}}^{\mathbf{E}_{\log}}(\theta) = \mathbf{f}_\theta([\mathbf{z}^\mathbf{I}, \mathbf{z}_\tau^\mathbf{t}, \mathbf{z}_\tau^{\mathbf{E}_{\log}}, \mathbf{z}^{\{\mathbf{a},\mathbf{d},\mathbf{m}\}}, \mathbf{z}^{\{\mathbf{n},\mathbf{r}\}} + \mathbf{c}_\mathbf{E}]; \mathbf{c}_\mathbf{E}^{\text{cross}}, \tau), \quad (3)$$

where $[\cdot]$ denotes concatenation in the temporal dimension, and $\mathbf{f}_\theta$ is the denoising function of DiT blocks.

3.4 Training strategies

The training of our method can be divided into three stages. In the first stage, we train the model using standard supervised learning (see supplemental material for data generation strategy and training details) to acquire basic relighting capabilities. In the second and third stages, we introduce two strategies to enhance generalization:

Intrinsic perception enhancement. As shown in Figure 3, we randomly select environment maps and generate two relighting results by controlling whether latent $\mathbf{z}^\mathbf{I}$ is set to 0. Ideally, these two inference modes should produce identical outcomes for the same scene. But that is not the case. Overall, we found that using $\mathbf{z}^\mathbf{I}$ yields more realistic appearances but occasionally retains source lighting, whereas the variant with $\mathbf{z}^\mathbf{I}$ set to 0 avoids residual lighting but suffers from detail distortion due to cumulative errors in the inverse rendering process. Therefore, We interpolate these two results in the latent space and decode the interpolated outputs, yielding a large amount of pseudo-realistic relit data. This process can be formulated

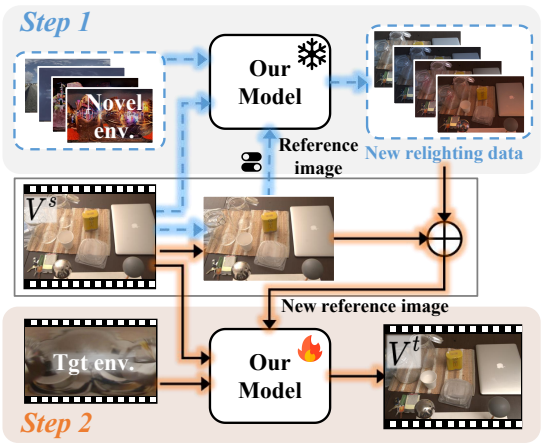

Fig. 3. **Overview of intrinsic perception enhancement.** Step 1: Generate multi-illumination data. Step 2: Use these multi-illumination data as the raw reference images for training.

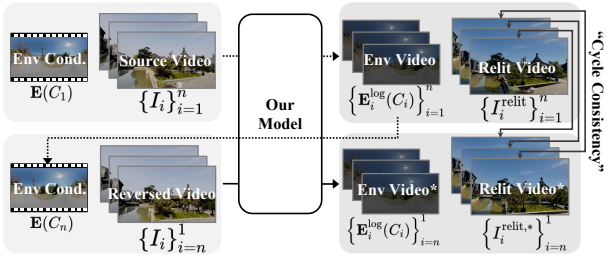

Fig. 4. **Overview of self-supervised learning based on illumination consistency.** The symbol (*) denotes inference results for reverse-order video.

Dotted line operations do not compute gradients. We relit a video under random environment map and then relit the video in reverse order based on the final frame of generated environment video. These two relit results form a self-supervised training pair.

as:

$$\mathbf{z}_{\text{new}} = \frac{\mathbf{z}_{\text{w/}}}{1+w} + \frac{\mathbf{z}_{\text{w/o}} * w}{1+w}, \quad (4)$$

where w is the interpolation weight, $\mathbf{z}_{\text{w/}}$ denotes the latents corresponding to results with $\mathbf{z}^\mathbf{I}$, and $\mathbf{z}_{\text{w/o}}$ denotes those with $\mathbf{z}^\mathbf{I}$ set to 0. Subsequently, we decode the interpolated latent $\mathbf{z}_{\text{new}}$ using the VAE decoder $\mathcal{D}$ to obtain new data that trade off realism and lighting plausibility. Furthermore, we treat these data as new raw reference images to enable training under diverse lighting conditions on real-world scenes. This strategy allows our method to access a wide variety of novel lighting conditions on real-world scenes during training, thereby significantly enhancing its perception of image intrinsic properties.

Self-supervised learning based on illumination consistency. In the final training stage, we introduce a self-supervised illumination consistency (SIC) strategy to enhance the model's generalization

across diverse scenes and lighting conditions, as illustrated in Figure 4. Specifically, we perform inference on all data under random environment maps to obtain relit video $V^{\text{relit}} = \{I_i^{\text{relit}}\}_{i=1}^{n}$ and their corresponding environment light $\left\{\mathbf{E}_i^{\log}(C_i)\right\}_{i=1}^{n}$. We then reverse the frame sequence of the original video and infer the new relighting result $V^{\text{relit},*} = \left\{I_i^{\text{relit},*}\right\}_{i=n}^{1}$ based on the environmental light $\mathbf{E}_n^{\log}(C_n)$. Self-supervised training pairs are constructed through frame-to-frame correspondence. This self-supervised process operates on image data under the "lighting rotation with fixed camera" pattern. The SIC strategy exposes our method to diverse lighting and scene combinations, significantly improving its generalization performance. It also promotes frame-by-frame alignment between predicted lighting and relit results, enhancing its generalization and sensitivity to varying lighting conditions.

4 Results

We compare Relit-LiVE with various advanced video relighting methods, including UniRelight [He et al. 2025b], Diffusion Renderer (cosmos) [Liang et al. 2025], Light-A-Video [Zhou et al. 2025], and others. Evaluation data spans multiple domains—synthetic, human [Pexels 2025], embodied [Walke et al. 2023], and autonomous driving [Xiao et al. 2021]—encompassing over 1,400 dynamic videos. Metrics encompass visual fidelity (PSNR, SSIM, and LPIPS), temporal consistency (RAFT score), and specially designed material fidelity (DINOv3 score), supplemented by user study. More experimental settings and results are detailed in the supplemental material.

4.1 Evaluation of video relighting

We compare Relit-LiVE with existing advanced methods across different datasets, with quantitative results presented in Table 1. Figure 5 and supplemental material present corresponding visualizations. Among them, NeuralGaffer fails on scene-level tests, struggling to remove lighting details such as shadows and highlights from the original scene. Diffusion Renderer exhibits distortion on materials, which is particularly noticeable on transparent objects. In contrast, our method outperforms others across all metrics while demonstrating excellent material consistency and physically accurate reflections and refractions. We also present the video relighting results of our method under dynamic lighting in Figure 6 and the supplemental material.

Additionally, we compare Relit-LiVE with advanced text prompt-based methods across multiple domains in Table 2, Figure 7 and Figure 8. As shown in the middle example of Figure 7, due to the lack of physical constraints, text-prompt-based methods may produce unreasonable luminous effects under certain special lighting conditions, such as neon lighting. Consequently, these methods exhibit poor material consistency, particularly evident in the DINO-MC metric. Additionally, such methods struggle to decouple the original lighting, such as the shadows and highlights in the third example shown in the figure. The Diffusion Renderer again exhibits material distortion due to cumulative errors in the two-stage process. In contrast, our method demonstrates comprehensive performance, achieving both material consistency and more details in lighting and shadows.

Fig. 5. **Qualitative comparison of image relighting on the MIT multi-illumination dataset.** Our method excels in handling complex materials, generating high-quality reflection and transmission effects that significantly outperform baselines.

As demonstrated in Figure 1, our method supports streaming video relighting. Specifically, we can segment the long video into multiple clips. Given the lighting conditions, we perform relighting starting from the first clip. Then, based on the generated environment video, we provide the lighting conditions for the first frame's viewpoint of the next clip, performing relighting clip by clip. This allows us to naturally achieve relighting for long videos.

In Figure 9, we present additional results of our method for relighting long videos (along with comparisons to other methods). Our method accurately perceives changes in the camera's viewpoint and correctly warp the environment map, thereby achieving temporally consistent lighting effects.

4.2 Evaluation of environment video generation

As shown in Figure 2, our model can generate warped environment maps (i.e., environment video) , which can be viewed as a novel lighting estimation task that infers lighting for all frames based on the lighting of a single frame. In this section, we use common metrics to evaluate the accuracy of the generated warped environment maps. Accordingly, we provide results from several classic light estimation methods for reference, including StyleLight [Wang et al. 2022] and DiffusionLight [Phongthawee et al. 2024]. As shown in Figure 10, the environment video produced by our method closely matches the reference. We also provide quantitative evaluation results in Table 3, using metrics related to illumination direction, such as angular error, to evaluate our method's capability in detecting changes in camera pose. The results demonstrate the stability of our predictions over time, which is crucial for generating spatially accurate lighting in videos.

4.3 Other applications

Scene editing. Our method supports scene editing by utilizing scene rendering, as detailed in the supplemental material. In Figure 11, we showcase object insertion and material editing, complete with realistic reflections and shadow effects.

Video delighting. Our method also effectively removes specular highlights from the original video. As shown in Figure 12, our method accurately restores the original material properties in the delighted scene. This is crucial for visual perception tasks [Zheng et al. 2023] sensitive to specular artifacts, including 3D reconstruction and

Table 1. **Quantitative comparison of relighting on the synthetic dataset and MIT multi-illumination dataset.** (*) indicates that metrics are sourced from the reported results of UniRelight [He et al. 2025b]. Our approach surpasses the baselines across all test metrics.

Methods	*Synthetic Image*			*Synthetic Video*			*MIT multi-illumination*		
	PSNR (↑)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)
NeuralGaffer [Jin et al. 2024]	12.84	0.435	0.463	-	-	-	17.87*	0.683*	0.241*
Diffusion Renderer [Liang et al. 2025]	17.09	0.679	0.264	16.45	0.665	0.278	17.29*	0.622*	0.355*
UniRelight [He et al. 2025b]	-	-	-	-	-	-	20.76*	0.749*	0.251*
Ours	**24.85**	**0.792**	**0.175**	**25.39**	**0.807**	**0.205**	**21.86**	**0.849**	**0.132**

Fig. 6. **Results under dynamic lighting in a dynamic scene.** Our method remains stable under simultaneous changes in scene content and illumination.

Table 2. **Quantitative comparison of video relighting on in-the-wild data.** Our approach significantly outperforms the baseline in terms of material consistency. User study metrics include VR (Visual Realism), PC (Physical Consistency), and LA (Lighting Alignment), reported as the percentage of participants who prefer our method. Details of each metric are provided in the supplemental material.

Methods	Light Condi	Temporal Consistency	Material Consistency		User Study (%)		
		Motion Preservation (↓)	CLIP-MC (↑)	DINO-MC (↑)	VR	PC	LA
Light-A-Video [Zhou et al. 2025]	Text	0.4557	0.9150	0.8919	65.5	75.8	54.8
TC-Light [Liu et al. 2025a]	Text	0.2405	0.8977	0.8825	84.5	84.8	77.4
Diffusion Renderer [Liang et al. 2025]	HDR	0.3094	0.9105	0.8754	86.2	81.3	63.3
Ours	HDR	**0.1692**	**0.9246**	**0.9091**	/	/	/

Table 3. **Directional angle error in video lighting estimation for sunlit scenes.** We use StyleLight [Wang et al. 2022] and DiffusionLight [Phongthawee et al. 2024] to estimate environment map frame-by-frame in videos, while our method generates the entire video's environment maps in a single pass. Note: The standard deviation (i.e., Std) here represents the average standard deviation of the directional errors across all videos.

Methods	**Angular Error top-5 (↓)**			**Angular Error top-3 (↓)**		
	Mean	Median	Std	Mean	Median	Std
StyleLight	66.12	62.62	24.11	69.27	63.81	26.23
DiffusionLight	54.88	51.58	18.83	55.18	51.82	18.82
Ours	**20.35**	**7.96**	**14.14**	**20.69**	**7.76**	**14.31**

depth estimation. Additional results are presented in supplemental material.

Table 4. **Ablation on different components.** Quantitative results of relighting on synthetic videos and MIT multi-illumination images. Both the raw image and the joint modeling of relighting with environmental video significantly enhance the model's relighting performance.

Methods	Synthetic Video			MIT multi-illumination		
	PSNR (↑)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)
w/o Env Video	17.45	0.703	0.284	21.21	**0.851**	**0.129**
w/o Raw Image	20.84	0.715	0.291	18.73	0.743	0.225
Full	**23.63**	**0.778**	**0.228**	**21.49**	0.849	0.135

4.4 Ablation Study

In this section, we conduct ablation studies on model architecture and training strategies to validate the effectiveness of the techniques proposed in this paper. See supplemental material for specific implementation details and further ablation studies.

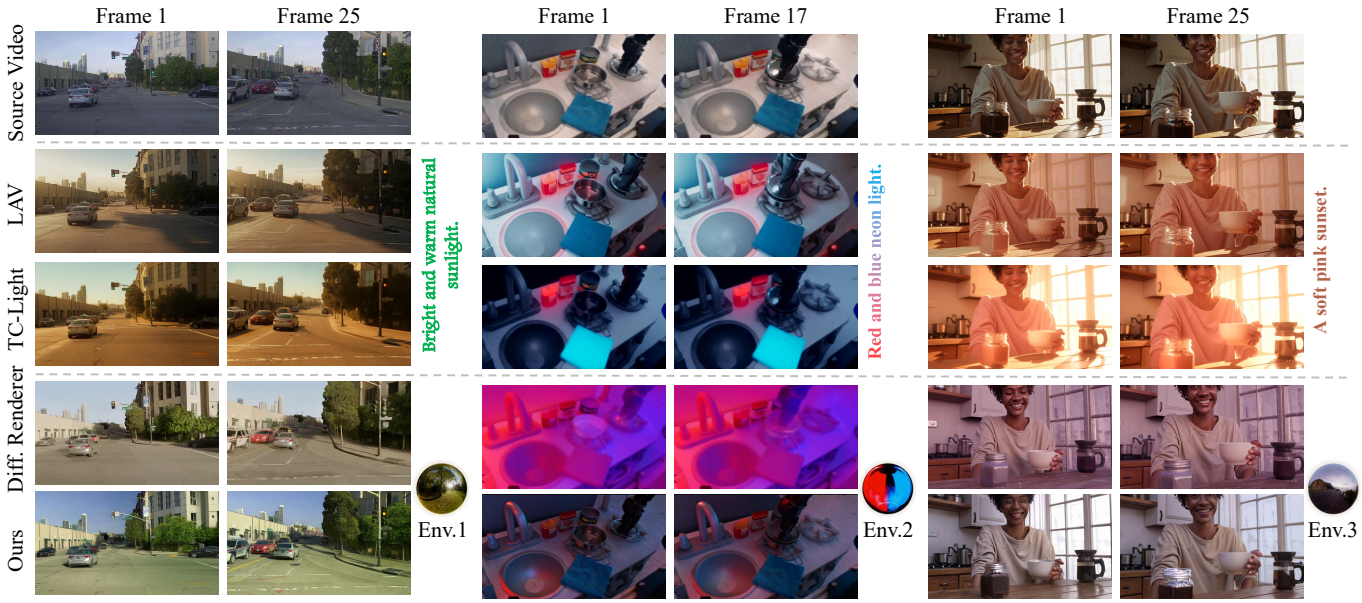

Fig. 7. **Qualitative comparison of video relighting on in-the-wild data.** We simultaneously evaluate advanced environment map-based methods and text prompt-based methods, aligning their lighting styles. Our approach outperforms baselines in both relighting quality and physical consistency.

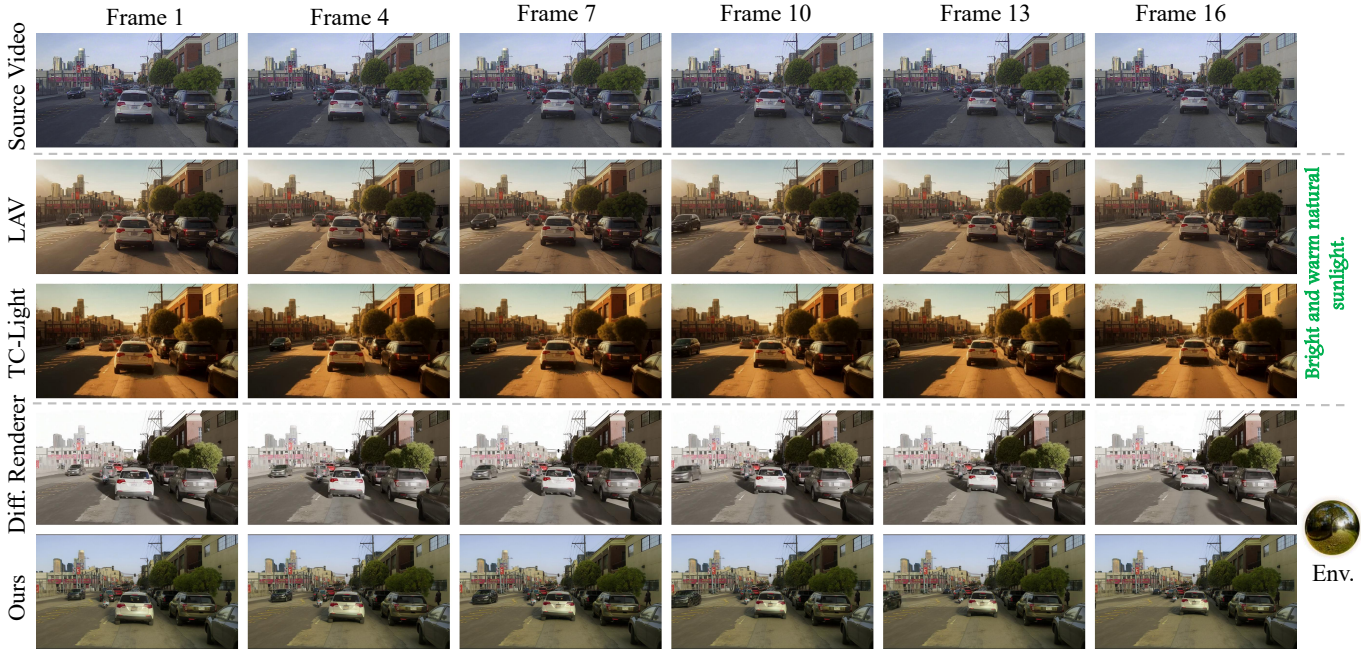

Fig. 8. **Qualitative comparison of video relighting**. Our method achieves superior relighting quality, temporal consistency, and photorealistic generation results compared to baseline methods.

Effectiveness of raw reference image. We present the quantitative ablation results of raw reference images in Table 4. The results demonstrate that introducing raw reference images significantly improves model performance. Figure 13 presents a visual comparison. As shown, the model without the raw reference image struggles to generate accurate physical transmission effects (notice the plastic bag and glass bottle in the scene). This clearly demonstrates the effectiveness of the raw reference image: it corrects rendering errors

Fig. 9. **Comparison of relighting results for long video sequences using different methods.** Light-A-Video and TC-Light process an entire 81-frame video in a single pass. Our approach divides a long video into multiple 57-frame segments, where the lighting conditions for each segment are derived from the lighting estimates of the preceding segment. In addition to the relighting results, our method also displays the corresponding predicted environment maps (converted from normalized log-intensity maps to LDR images for visualization).

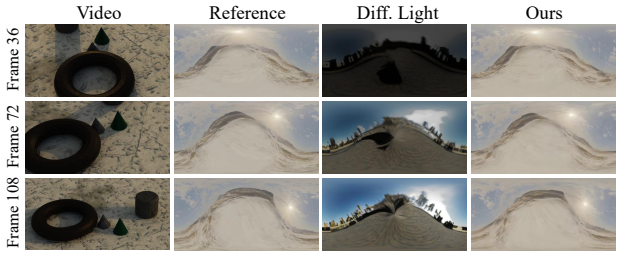

Fig. 10. **Qualitative comparison of video lighting estimation on the synthetic dataset.** Given the environment map of the first frame, our method generates environment maps for every frame of the entire video. It produces smooth lighting deformations and accurately aligns with the camera viewpoint of each frame. We also provide the results of DiffusionLight [Phongthawee et al. 2024].

Fig. 11. **Image editing application.** Left: Insert puppy and toy model into desktop scene; adjust base color of curtains and metallicness, roughness, and base color of tablecloth. Right: Insert vehicle into street scene; modify base color of entire vehicle on the right.

caused by imperfect intrinsic decomposition and guides the model to learn realistic physical effects.

Effectiveness of joint generation. We also ablate the environment video generation branch in Table 4. The results demonstrate that, compared to the ablated model, our joint model achieves significant performance improvements on synthetic videos featuring scenes with substantial camera motion or dynamic lighting. This fully demonstrates the benefits of environment video generation for camera-free video relighting. By jointly generating relighting and environment video, the model effectively aligns the input environment map with each frame's camera viewpoint, thereby achieving spatially consistent video relighting.

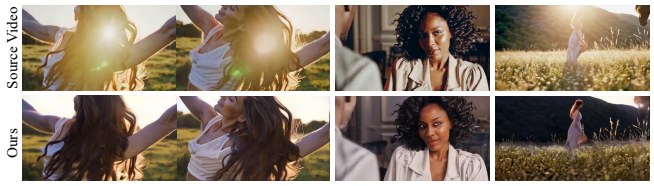

Fig. 12. **Visualization results of video delighting.** Our method achieves realistic and natural lighting removal by resynthesizing video illumination through specific environmental map.

Training strategy validation. We conduct a comparative analysis of the design schemes for the three-stage training in Figure 14. In

Fig. 13. **Qualitative ablation of relighting.** The raw reference image significantly improves relighting quality on complex materials.

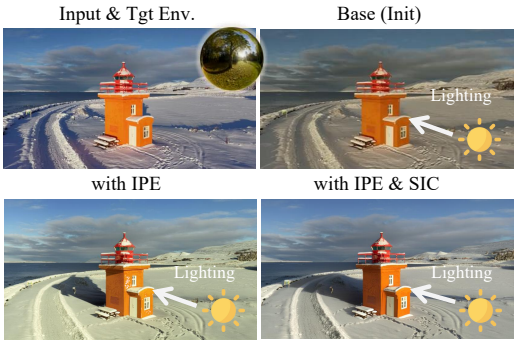

Fig. 14. **Ablation on training strategies.** We mark the direction of peak illumination. IPE: Intrinsic Perception Enhancement. SIC: Self-supervised learning based on Illumination Consistency.

fact, after the standard supervised training in the first stage, our initial model achieves a PSNR of 21 on the MIT multi-illumination benchmark, surpassing state-of-the-art models. However, this model occasionally struggles with complex original lighting effects, as we lack training data on real-world scenes with multi-illumination conditions. Since our "Intrinsic Perception Enhancement" strategy constructs a large number of pseudo-raw reference images with special lighting for real-world scenes, the model's ability to decouple original lighting is significantly improved. Furthermore, our self-supervised strategy enables closed-loop training under arbitrary lighting and scene, which further enhances the visual quality of relighting results.

4.5 Limitations

Although merging different intrinsic latents reduces computational overhead, the frame-dimensional concatenation-based control method still incurs substantial training costs. Consequently, Relit-LiVE inevitably trades off resolution and frame rate, with a maximum of 57 frames achieved at 832×480 resolution during training. Meanwhile, on the A800 GPU, generating a 57-frame video takes approximately 10 minutes.

5 Conclusion

In this paper, we have presented Relit-LiVE, a novel video relighting framework that produces physically consistent, temporally stable results without requiring prior knowledge of camera pose. At the core of our framework is an RGB-intrinsic fusion renderer along with a joint generation formulation for the relit video and the environment video. This approach allows the model to incorporate real-world lighting effects while adhering to estimated physical constraints, resulting in realistic relighting outcomes. Additionally, the formulation eliminates the need for explicit pose estimation, enhancing practical flexibility. Furthermore, we design two complementary training strategies that effectively mitigate the scarcity of existing multi-light datasets and further improve the model's generalization in complex scenes. Extensive experiments confirm that Relit-LiVE outperforms state-of-the-art methods in generating physically consistent and temporally stable relighting results (e.g., shadows, reflections) with strong generalization. Additionally, its extensibility supports downstream tasks such as scene rendering and video illumination estimation, validating its potential as a universal video editing engine.

References

Ole Beisswenger, Jan-Niklas Dihlmann, and Hendrik Lensch. 2025. FrameDiffuser: G-Buffer-Conditioned Diffusion for Neural Forward Frame Rendering. *arXiv preprint arXiv:2512.16670* (2025).

Weikang Bian, Xiaoyu Shi, Zhaoyang Huang, Jianhong Bai, Qinghe Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Hongsheng Li. 2025. Relightmaster: Precise video relighting with multi-plane light images. *arXiv preprint arXiv:2511.06271* (2025).

Yanrui Bin, Wenbo Hu, Haoyuan Wang, Xinya Chen, and Bing Wang. 2025. Normalcrafter: Learning temporally consistent normals from video diffusion priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8330–8339.

Nicolas Bonneel, Balazs Kovacs, Sylvain Paris, and Kavita Bala. 2017. Intrinsic decompositions for image editing. In *Computer graphics forum*, Vol. 36. Wiley Online Library, 593–609.

Chris Careaga and Yağız Aksoy. 2023. Intrinsic image decomposition via ordinal shading. *ACM Transactions on Graphics* 43, 1 (2023), 1–24.

Chris Careaga and Yağız Aksoy. 2025. Physically controllable relighting of photographs. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. 1–10.

Boyuan Chen, Hanxiao Jiang, Shaowei Liu, Saurabh Gupta, Yunzhu Li, Hao Zhao, and Shenlong Wang. 2025b. Physgen3d: Crafting a miniature interactive world from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6178–6189.

Xiaoxue Chen, Bhargav Chandaka, Chih-Hao Lin, Ya-Qin Zhang, David Forsyth, Hao Zhao, and Shenlong Wang. 2025a. InvRGB+ L: Inverse Rendering of Complex Scenes with Unified Color and LiDAR Reflectance Modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 27176–27186.

Zhifei Chen, Tianshuo Xu, Wenhang Ge, Leyi Wu, Dongyu Yan, Jing He, Luozhou Wang, Lu Zeng, Shunsi Zhang, and Ying-Cong Chen. 2025c. Uni-Renderer: Unifying Rendering and Inverse Rendering Via Dual Stream Diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 26504–26513.

Ye Fang, Zeyi Sun, Shangzhan Zhang, Tong Wu, Yinghao Xu, Pan Zhang, Jiaqi Wang, Gordon Wetzstein, and Dahua Lin. 2025a. RelightVid: Temporal-consistent diffusion model for video relighting. *arXiv preprint arXiv:2501.16330* (2025).

Ye Fang, Tong Wu, Valentin Deschaintre, Duygu Ceylan, Iliyan Georgiev, Chun-Hao Paul Huang, Yiwei Hu, Xuelin Chen, and Tuanfeng Yang Wang. 2025b. V-RGBX: Video Editing with Accurate Controls over Intrinsic Properties. *arXiv preprint arXiv:2512.11799* (2025).

Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and Ying-Cong Chen. 2025a. Lotus: Diffusion-based Visual Foundation Model for High-quality Dense Prediction. In *The Thirteenth International Conference on Learning Representations*.

Kai He, Ruofan Liang, Jacob Munkberg, Jon Hasselgren, Nandita Vijaykumar, Alexander Keller, Sanja Fidler, Igor Gilitschenski, Zan Gojcic, and Zian Wang. 2025b. UniRelight: Learning Joint Decomposition and Synthesis for Video Relighting. In *Advances in Neural Information Processing Systems*.

Haian Jin, Yuan Li, Fujun Luan, Yuanbo Xiangli, Sai Bi, Kai Zhang, Zexiang Xu, Jin Sun, and Noah Snavely. 2024. Neural gaffer: Relighting any object via diffusion. *Advances in Neural Information Processing Systems* 37 (2024), 141129–141152.

Peter Kocsis, Lukas Höllein, and Matthias Nießner. 2025. IntrinsiX: High-Quality PBR Generation using Image Priors. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Hong Li, Houyuan Chen, Chongjie Ye, Zhaoxi Chen, Bohan Li, Shaocong Xu, Xianda Guo, Xuhui Liu, Yikai Wang, Baochang Zhang, Satoshi Ikehata, Boxin Shi, Anyi Rao, and Hao Zhao. 2025. Light of Normals: Unified Feature Representation for

Universal Photometric Stereo. *arXiv preprint arXiv:2506.18882* (2025).

Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Chih-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, et al. 2025. Diffusion Renderer: Neural Inverse and Forward Rendering with Video Diffusion Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 26069–26080.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. 2024. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22160–22169.

Pengwei Liu, Hangjie Yuan, Bo Dong, Jiazheng Xing, Jinwang Wang, Rui Zhao, Weihua Chen, and Fan Wang. 2025b. UniLumos: Fast and Unified Image and Video Relighting with Physics-Plausible Feedback. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Qingyang Liu, Junqi You, Jianting Wang, Xinhao Tao, Bo Zhang, and Li Niu. 2024. Shadow generation for composite image using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8121–8130.

Tianqi Liu, Zhaoxi Chen, Zihao Huang, Shaocong Xu, Saining Zhang, Chongjie Ye, Bohan Li, Zhiguo Cao, Wei Li, Hao Zhao, et al. 2026. Light-X: Generative 4D Video Rendering with Camera and Illumination Control. In *The Fourteenth International Conference on Learning Representations*.

Yang Liu, Chuanchen Luo, Zimo Tang, Yingyan Li, Yuanyong Ning, Lue Fan, Junran Peng, Zhaoxiang Zhang, et al. 2025a. TC-Light: Temporally Coherent Generative Rendering for Realistic World Transfer. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Nadav Magar, Amir Hertz, Eric Tabellion, Yael Pritch, Alex Rav-Acha, Ariel Shamir, and Yedid Hoshen. 2025. 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*. 1–11.

Lukas Murmann, Michael Gharbi, Miika Aittala, and Fredo Durand. 2019. A multi-illumination dataset of indoor object appearance. In *2019 IEEE international conference on computer vision (ICCV)*, Vol. 2.

OpenAI. 2024. Video generation models as world simulators.

Pexels. 2025. *Pexels Free Stock Media Platform.* https://www.pexels.com

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

Kerui Ren, Jiayang Bai, Linning Xu, Lihan Jiang, Jiangmiao Pang, Mulin Yu, and Bo Dai. 2025. MV-CoLight: Efficient Object Compositing with Consistent Lighting and Shadow Generation. *arXiv preprint arXiv:2505.21483* (2025).

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159* (2024).

Why Physically-Based Rendering. 2015. Physically-based rendering. *Procedia IUTAM* 13, 127-137 (2015), 3.

Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. 2018. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *Proceedings of the European conference on computer vision (ECCV)*. 650–665.

Giuseppe Vecchio and Valentin Deschaintre. 2024. Matsynth: A modern pbr materials dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22109–22118.

Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. 2023. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*. PMLR, 1723–1736.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314* (2025).

Guangcong Wang, Yinuo Yang, Chen Change Loy, and Ziwei Liu. 2022. Stylelight: Hdr panorama generation for lighting estimation and editing. In *European conference on computer vision*. Springer, 477–492.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, et al. 2025. Spatialvid: A large-scale video dataset with spatial annotations. *arXiv preprint arXiv:2509.09676* (2025).

Runpu Wei, Zijin Yin, Shuo Zhang, Lanxiang Zhou, Xueyi Wang, Chao Ban, Tianwei Cao, Hao Sun, Zhongjiang He, Kongming Liang, and Zhanyu Ma. 2025. OmniEraser: Remove Objects and Their Effects in Images with Paired Video-Frame Data. *arXiv preprint arXiv:2501.07397* (2025). https://arxiv.org/abs/2501.07397

Dianbing Xi, Jiepeng Wang, Yuanzhi Liang, Xi Qiu, Jialun Liu, Hao Pan, Yuchi Huo, Rui Wang, Haibin Huang, Chi Zhang, et al. 2025. CtrlVDiff: Controllable Video Generation via Unified Multimodal Video Diffusion. *arXiv preprint arXiv:2511.21129* (2025).

Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al. 2021. Pandaset: Advanced sensor suite dataset for autonomous driving. In *2021 IEEE international intelligent transportation systems conference (ITSC)*. IEEE, 3095–3101.

Tian-Xing Xu, Xiangjun Gao, Wenbo Hu, Xiaoyu Li, Song-Hai Zhang, and Ying Shan. 2025. Geometrycrafter: Consistent geometry estimation for open-world videos with diffusion priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6632–6644.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025a. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388* (2025).

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2025b. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In *The Thirteenth International Conference on Learning Representations*.

Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. 2024. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics (TOG)* 43, 6 (2024), 1–18.

Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. 2024. RGB ↔ X: Image decomposition and synthesis using material- and lighting-aware diffusion models. In *ACM SIGGRAPH 2024 Conference Papers* (Denver, CO, USA) *(SIGGRAPH '24)*. Association for Computing Machinery, New York, NY, USA, Article 75, 11 pages. doi:10.1145/3641519.3657445

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2025. 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*. https://openreview.net/forum?id=u1cQYxRI1H

Yupeng Zheng, Chengliang Zhong, Pengfei Li, Huan-ang Gao, Yuhang Zheng, Bu Jin, Ling Wang, Hao Zhao, Guyue Zhou, Qichao Zhang, et al. 2023. Steps: Joint self-supervised nighttime image enhancement and depth estimation. *arXiv preprint arXiv:2302.01334* (2023).

Yujie Zhou, Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, Qidong Huang, Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, et al. 2025. Light-a-video: Training-free video relighting via progressive light fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13315–13325.

In the supplemental material, we provide the Appendix (Sec A) of this paper. For additional visualization results, please refer to the ACCOMPANYING FILES in the folder.

A Appendix

A.1 Data generation strategy

Our training data comprises a substantial synthetic dataset alongside auto-labeled real-world datasets, specifically as follows:

Synthetic dataset. We produce a large number of rendered videos through a data synthesis workflow, complete with their base colors, roughness, metallicness, normal maps, depth maps, environment maps, and camera trajectories. Specifically, we first collect 5700 high-quality PBR material maps and 2241 HDR environment maps from public resources [Li et al. 2025; Vecchio and Deschaintre 2024]. For each scene, we initially place a plane as well as up to 12 primitives (such as cubes, cones, and cylinders), with collision detection applied to avoid object intersections. Subsequently, we select PBR material maps randomly to texture both the plane and the primitives. Finally, we generate three types of videos with random motion patterns, namely: 1) Camera rotation with fixed lighting; 2) Lighting rotation with fixed camera; 3) Simultaneous rotation of both the camera and lighting. Each scene is rendered under at least two random lighting conditions for the same motion pattern. In total, we generate 8,000 videos, each consisting of 120 frames at a resolution of 512×512.

Real-world dataset. We collected a large number of video clips and images from real-world datasets, including DL3DV [Ling et al. 2024], SpatialVID-HQ [Wang et al. 2025], MIT multi-illumination [Murmann et al. 2019], RemovalBench [Wei et al. 2025], and SOBAv2 [Liu et al. 2024]. Specifically, we used a Vision Language Model (VLM) [Yang et al. 2025a] to filter these datasets (excluding MIT multi-illumination), removing data with blurred frames or significant shadows from objects outside the frame. Through this filtering process, we selected 13,809 video clips (57 frames) and 792 images. We then generated pseudo ground-truth G-buffers for the above data using Diffusion Renderer's inverse renderer. For MIT multi-illumination, we exported environment maps based on the dataset's included reflective chrome sphere screenshots. For other datasets, we employed the VLM to determine the camera perspective of all data samples, and applied DiffusionLight [Phongthawee et al. 2024] exclusively to annotate environment maps for images with horizontal perspectives, while manually filtering out results with obvious errors. Because we find that the DiffusionLight tends to yield significant estimation errors for images captured from non-horizontal perspectives. Ultimately, we annotated environmental maps for 8,278 videos and 206 images, performing frame-by-frame alignment based on camera trajectories. These datasets with varying levels of completeness, combined with image data from MIT multi-illumination, collectively form a real-world dataset that significantly enriches our training samples for realistic scenarios.

A.2 Initial training

Considering the significant differences in illumination distribution between synthetic and real-world datasets, we dynamically adjust our training mode across different periods. Specifically, we initially train the model exclusively on synthetic data to learn foundational rendering. Subsequently, we freeze the cross-attention module and train on the full dataset to enhance generalization while preserving adaptability to varying lighting conditions. During training, we set the latent $\mathbf{z}^{\mathbf{I}}$ to zero with a probability of 0.3 to simulate a pure rendering task. Following UniRelight [He et al. 2025b], for the sets of real-world datasets with and without environment maps, the denoising targets become $\hat{\mathbf{z}}^{\mathbf{s}}(\theta), \hat{\mathbf{z}}^{\mathbf{E}_{\log}}(\theta) = \mathbf{f}_\theta([\mathbf{z}^{\mathbf{I}}, \mathbf{z}^{\mathbf{s}}_\tau, \mathbf{z}^{\mathbf{E}_{\log}}_\tau, \mathbf{z}^{\{\mathbf{a},\mathbf{d},\mathbf{m}\}}, \mathbf{z}^{\{\mathbf{n},\mathbf{r}\}} + \mathbf{c}_{\mathbf{E}}]; \mathbf{c}^{\text{cross}}_{\mathbf{E}}, \tau)$ and $\hat{\mathbf{z}}^{\mathbf{t}}(\theta) = \mathbf{f}_\theta([\mathbf{z}^{\mathbf{I}}, \mathbf{z}^{\mathbf{t}}_\tau, \mathbf{z}^{\mathbf{0}}_\tau, \mathbf{z}^{\{\mathbf{a},\mathbf{d},\mathbf{m}\}}, \mathbf{z}^{\{\mathbf{n},\mathbf{r}\}} + \mathbf{0}]; \mathbf{0}, \tau)$ respectively, enabling training on real-world scenes with single lighting conditions.

A.3 Additional Details on Intrinsic Perception Enhancement

In this section, we focus on the details of multi-illumination data generation in Intrinsic Perception Enhancement. In fact, this approach was inspired by our observations of the initial model's generated results. As introduced in our discussion of the RELIT-LIVE architecture's flexibility, our model supports both relighting (w/ $\mathbf{z}^{\mathbf{I}}$) and rendering (w/o $\mathbf{z}^{\mathbf{I}}$) tasks. The distinction in their specific inference processes lies solely in whether the raw reference image is input. Ideally, these two inference modes should produce identical results for the same scene. But that's not the case. We found that for relighting mode, the initial model has a certain probability of extracting the original lighting from the raw reference image. In other words, while the information in the raw reference image significantly shapes the high-quality details of the relit video, it may also cause the result to deviate from the target lighting. We illustrate an example in the first row and first column of Figure A15. Overall, relighting results exhibit high quality with occasional lighting anomalies, while rendering results feature reasonable lighting but lack visual realism. Therefore, we attempt to fuse both outputs to achieve stable, high-quality data augmentation with consistent lighting. During the implementation of Intrinsic Perception Enhancement, we generated multi-illumination data multiple times using the optimal model at each corresponding time point to continuously improve the quality of generated data. For each multi-illumination data generation, we adjusted the appropriate parameter w for the corresponding model. Figure A15 simultaneously displays the multi-illumination data generated by the initial model and the later model. It is evident that the quality of our generated data has achieved substantial improvement. This high-quality data enhances our model's ability to decouple original illumination, thereby improving relighting performance.

Potential Discussion: Why not use existing relighting models to generate multi-illumination data? According to the IPE strategy, the generated data serves as (pseudo) original reference images. These should exhibit accurate and physically realistic lighting effects, which existing open-source relighting models cannot achieve. As demonstrated in Section 4.1, even advanced open-source methods still fail to produce realistic relighting results that adhere to material properties.

A.4 Experimental details

Training Details. We adopt Wan2.1-T2V-1.3B [Wan et al. 2025] as the base model and achieve our RELIT-LIVE by fine-tuning its components. As mentioned in Section 3.4, our training is divided

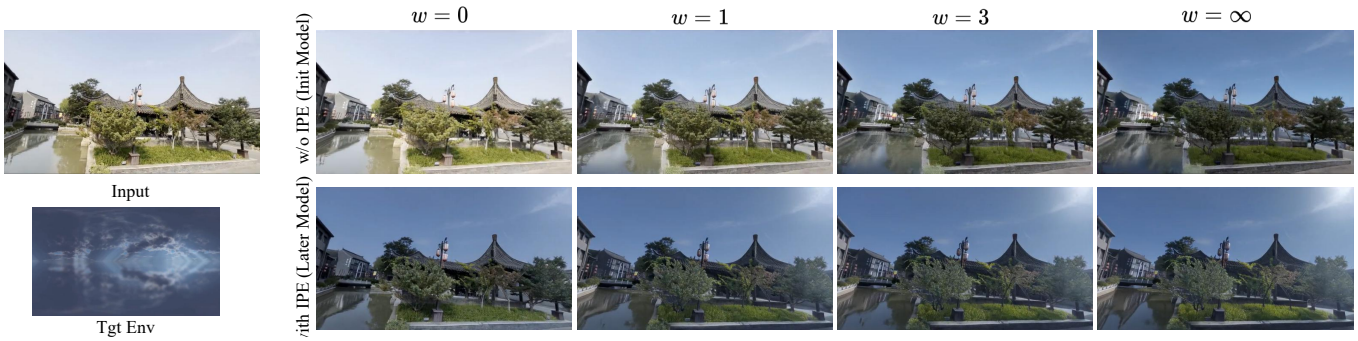

Fig. A15. **Example of latent space interpolation between relighting and rendering results.** We demonstrate model interpolation outcomes before and after applying the Intrinsic Perception Enhancement strategy. Note: When $w = 0$, the interpolated result is fully equivalent to the relighting result; when $w = \infty$, it is fully equivalent to the rendering result. This visualization not only illustrates the interpolation process but also validates the effectiveness of the Intrinsic Perception Enhancement strategy.

into three stages. In the first stage, we first train the model for 10,000 iterations using synthetic data only, then train it for 20,000 iterations on the full dataset to obtain the initial model. In the second stage, we use the initial model to generate 8 pseudo-realistic images with different illuminations for each scenario in the real-world dataset, training the model for 5,000 iterations. In the third stage, we implement the SIC strategy with a probability of 0.1, training the model for a final 5,000 iterations. All training is conducted on 8 A800 GPUs with a batch size of 16, resolution of 832×480, and AdamW optimizer with a learning rate of 1e-5. Total training takes about 7 days. We only fix the training video length to 17 during the synthetic data training in the first stage. In subsequent processes, the training video length is incrementally increased cyclically (from 1 to 57, following the 8n+1 pattern) to ensure the model's generalization capability across different frame lengths.

Baselines. For video relighting, we compare Relit-LiVE against multiple advanced video relighting methods, including UniRelight [He et al. 2025b](as of now, it has not been open-sourced, so we directly replicate the results from the paper report), Cosmos-Diffusion Renderer [Liang et al. 2025], Light-A-Video [Zhou et al. 2025], TC-Light [Liu et al. 2025a], and advanced image relighting method NeuralGaffer [Jin et al. 2024]. For scene rendering, we compare our approach with two representative neural rendering methods RGBX [Zeng et al. 2024] and Cosmos-Diffusion Renderer. For environment light estimation, we compare with the image lighting estimation methods DiffusionLight [Phongthawee et al. 2024] and StyleLight [Wang et al. 2022].

Dataset. We curate test datasets through multiple channels for evaluating various tasks. First, we have created a high-quality synthetic test set, comprising 1,000 high-motion videos, each consisting of 120 frames (covering the three "camera-light" motion patterns mentioned in Section A.1). Meanwhile, we employ the MIT multi-illumination test set [Murmann et al. 2019], comprising 30 high-quality scenes across 25 lighting configurations. To ensure fair comparison, the evaluation methodology of relighting on this dataset is identical to that used in UniRelight: images under the i-th lighting condition are paired with those under the (i+12)-th lighting condition to form test pairs. Additionally, we have collected a series of videos from diverse domains, encompassing scenarios such as portraits, nature, roadways, and robotics, to evaluate the method's generalization in the real world. Specifically, we collected 277 high-quality videos from Pexels [Pexels 2025] and Sora [OpenAI 2024], covering subjects such as humans, animals, and objects, and including various camera movements and object motions. Beyond that, we also collected 100 representative videos each from PandaSet [Xiao et al. 2021] and Bridgev2 [Walke et al. 2023] for evaluation in embodied and autonomous driving domains. These videos are completely unrelated to our training data.

Evaluation metrics. Due to differences in dataset composition, we conduct distinct evaluations on different test sets. 1) We evaluate the performance of relighting, rendering, and lighting estimation simultaneously on the synthetic test set and the MIT multi-illumination test set. (i) For relighting and rendering, we employ PSNR, SSIM, and LPIPS as evaluation metrics to frame-by-frame assess the visual fidelity between generated results and ground truth. (ii) For illumination estimation, we report angular error in degrees in scenes with concentrated sunlight.

2) For real-world videos across various domains under other single-lighting conditions, we evaluate the motion preservation and material consistency of relit results using existing pre-trained models, and assess the physical consistency of relit effects through user studies. (i) Specifically, we employ RAFT to estimate optical flow for both the source video and the relit video. The motion preservation scores for each method are evaluated by calculating the optical flow differences. (ii) Then, we evaluate material consistency between the source video and the generated video by calculating the average CLIP score (CLIP-MC) and average DINOv3 score (DINO-MC) for corresponding frames according to A1, where higher scores indicate better consistency. It is worth noting that our CLIP scores are computed between the source video and the generated video, rather than between consecutive frames as in previous methods [Liu et al. 2026; Zhou et al. 2025] (which are used to measure video smoothness). To validate these metrics, we produce paired datasets with identical layouts but varying lighting/materials, and compute DINOv3 and CLIP similarity. Both metrics exhibit invariance ($\geq$ 0.94) to lighting variations and sensitivity ($\leq$ 0.89) to object material

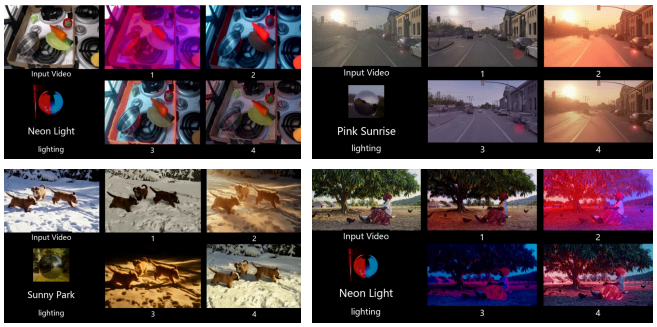

Fig. A16. **The visual interface for our user research.** Participants simultaneously observed the input video, target lighting (text or environment map) and the results of four methods (randomly shuffled) displayed side-by-side. They evaluated each set of results based on three criteria by selecting methods that clearly failed.

differences. (iii) Our user study focuses on whether generated results are physically plausible, which is crucial for fields like simulation that demand high realism. Specifically, the evaluation involves visual realism (VR), physical consistency (PC) and lighting alignment(LA).

$$\mathrm{MC} = 1 - \frac{1}{2N}\sum_{i=1}^{N}(1 - \cos(\mathbf{f}_i^{\mathrm{src}}, \mathbf{f}_i^{\mathrm{gen}})) \in [0, 1] \tag{A1}$$

A.5 User Study

Figure A16 illustrates the interface used for our user study. Participants evaluated the results based on three questions: (i) Which result is inconsistent with the input video's realism (e.g., anomalous glowing or artifacts)? (ii) Which result fails to modify original shadows or metallic highlights? (iii) Which result exhibits lighting inconsistent with the target condition shown in the bottom-left (defined by text description or environment map)? For each question, participants could select between 0 and 2 options. In total, we collected responses from 37 participants. Each participant was required to complete 10 sets of comparisons. We recorded instances where our method outperformed the baselines and vice versa. It is worth noting that in cases of a tie, we administered the survey repeatedly until a clear judgment was reached. Finally, we calculated the ratio of our method outperforming each baseline as the final metric.

A.6 Evaluation of forward rendering

Besides video relighting, Relit-LiVE also supports forward rendering. To validate our model, we compare the neural rendering performance of different methods in Table A1. Note that our synthetic video dataset includes dynamic cameras, dynamic lighting, and combinations of both. These three motion patterns present fundamentally different challenges for our method, as its lighting conditions originate from the initial viewpoint. Both dynamic cameras and dynamic lighting introduce additional complexity for our approach. In contrast, this poses no significant theoretical difference for Diffusion Renderer, which defines environment maps frame-by-frame. Nevertheless, our approach achieves optimal performance, showcasing its robustness across diverse dynamic scenes.

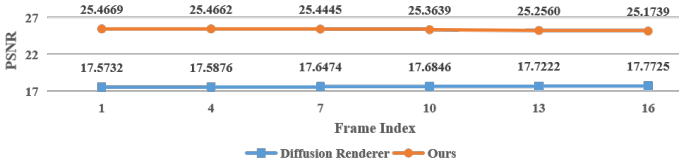

Fig. A17. **Error distribution across different frames.** We compute the average PSNR for each frame across 200 synthetic videos. Our results exhibit high stability over time, comparable to methods that condition on per-frame environment maps.

We visualize the error distribution of each method across different frames for video rendering task, as shown in Figure A17. Our method and Diffusion Renderer both demonstrate relatively stable rendering performance across different frames. Despite only being fed the environment map from the initial viewpoint, our model robustly perceives camera viewpoint changes, accurately propagating the lighting from the initial viewpoint to all viewpoints.

A.7 Supplement to ablation studies

Experimental settings. For the ablation studies on model architecture, we first train the model for 10,000 iterations using only synthetic data, followed by training it for 20,000 iterations on the full dataset. Throughout training, we employ standard supervised learning without incorporating the two training strategies proposed in this paper. For the ablation studies on training strategies, we additionally train the models from the architecture ablation, applying each strategy sequentially for 5,000 iterations. All training runs on a single A800 GPU with a batch size of 4, consistently fixing the training video length to 17 frames to reduce computational overhead. Note that to ensure fairness and rigor, we train each ablation model starting from the base model Wan2.1-T2V-1.3B, using the same number of training steps. These models are unrelated to our final model.

Why use G-buffer latent group-wise addition? We use group-wise addition to preserve semantic separability within each group. This strategy is also used in UniRelight. In this section, we perform the ablation of the operation and present the results in Table A2. Experiments demonstrate that our group-wise addition performs comparably to frame-concat while consuming 25% fewer resources.

Why choose dual-input light conditions? We compare the results of relighting under different lighting control implementations, as shown in Figure A18. We find that compared to models that only inject lighting conditions via cross-attention, models that solely fuse environment light features into scene-intrinsic latent achieve better visual quality in overall detail across natural scenes. However, this does not imply an architectural advantage. As shown in the red box, in natural scenes, this model tends to directly replicate content from the raw reference image, often leading to degradation in relighting tasks. In contrast, methods that inject lighting conditions solely through cross-attention generate color temperatures closer to the target lighting, but lack reflective details. We speculate this is because our base model is a T2V model, which transmits text prompt

Table A1. **Quantitative comparison of neural rendering on the synthetic dataset.** The D. denotes dynamic, while the S. denotes static.

Methods	*Synthetic Image*			*Synthetic Video*								
				S.Lighting - D.Camera			D.Lighting - S.Camera			D.Lighting - D.Camera		
	PSNR (↑)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)
RGBX [Zeng et al. 2024]	11.97	0.503	0.429	-	-	-	-	-	-	-	-	-
Diffusion Renderer [Liang et al. 2025]	18.71	0.692	0.256	18.09	0.676	0.265	17.46	0.662	0.256	17.72	0.666	0.274
Ours	**24.31**	**0.760**	**0.197**	**25.09**	**0.792**	**0.216**	**25.13**	**0.810**	**0.201**	**25.59**	**0.802**	**0.215**

Table A2. **Ablation of G-buffer latent group-wise addition.** We compare the performance of the Frame-concat method and our method on two datasets: Synthetic Video and MIT multi-illumination. Performance metrics include PSNR, SSIM, LPIPS, and GPU memory usage for both training and inference.

Method	**Synthetic Video**			**MIT multi-illumination**			**Training**$_{57frames}$	**Inference**$_{57frames}$
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	GPU_memory$_{train}$ ↓(MiB)	GPU_memory$_{inference}$ ↓(MiB)
Frame-concat	23.55	**0.789**	**0.214**	21.21	**0.851**	0.137	69678	33154
Ours	**23.63**	0.778	0.228	**21.49**	0.849	**0.135**	**51990**	**26840**

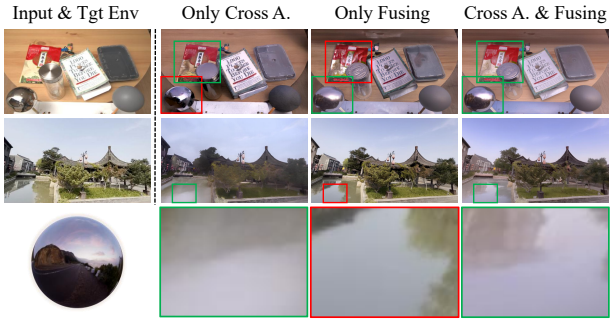

Fig. A18. **Qualitative comparison under different light conditions.** Only Cross A.: Input light conditions via cross-attention. Only Fusing: By fusing with scene property latents to input light conditions. The dual-path lighting control method achieves the most balanced performance in terms of reflection quality and color temperature.

Fig. A19. Failure cases of relighting: color shift (left) and abnormal illumination (right).

information to the video generation model via cross-attention. However, textual information inherently possesses a natural sparsity compared to image data, making it challenging to describe fine-grained spatial structural details.

In other words, relying solely on cross-attention to input lighting conditions fails to accurately convey texture details from environment maps, resulting in degraded reflection effects. In contrast, the duplicate light control approach achieves the best overall performance and is therefore adopted in our final architecture.

A.8 Scene editing workflow

We illustrate our scene editing workflow in Figure A20. Specifically, we modify materials within the scene and insert new objects by editing intermediate intrinsics. We utilize Ground-SAM [Ren et al. 2024] to obtain object masks, enabling material adjustments for specific objects. Simultaneously, we directly insert new objects into the original image to serve as the reference image for model input. Since the processed reference image still contains elements awaiting material modification, we employ latent space interpolation from Equation. 4 to generate the edited image.

A.9 Analysis of failure cases

In this section, we focus on analyzing some typical failure cases of this method. As shown in Figure A19, some objects in the relighting results exhibit color shifts and abnormal lighting. We think these issues stem from inherent pseudo-label errors (G-buffer & environment map) in the real-world training set. This is unavoidable for methods applied to real scenes, such as Diffusion Renderer. Inaccurate base-color labels may cause color shifts. Inaccurate environment map pseudo-labels impair the model's learning of light sources, causing it to occasionally misclassify background pixels as strong light sources.

A.10 More visualizations of our methods

Figure A21 visualizes the performance of our training strategy. Figures A22-A24 present additional visual comparisons of our method against other methods on in-the-wild data. Figures A25-A36 show the results of our method under dynamic lighting and various environment lights.

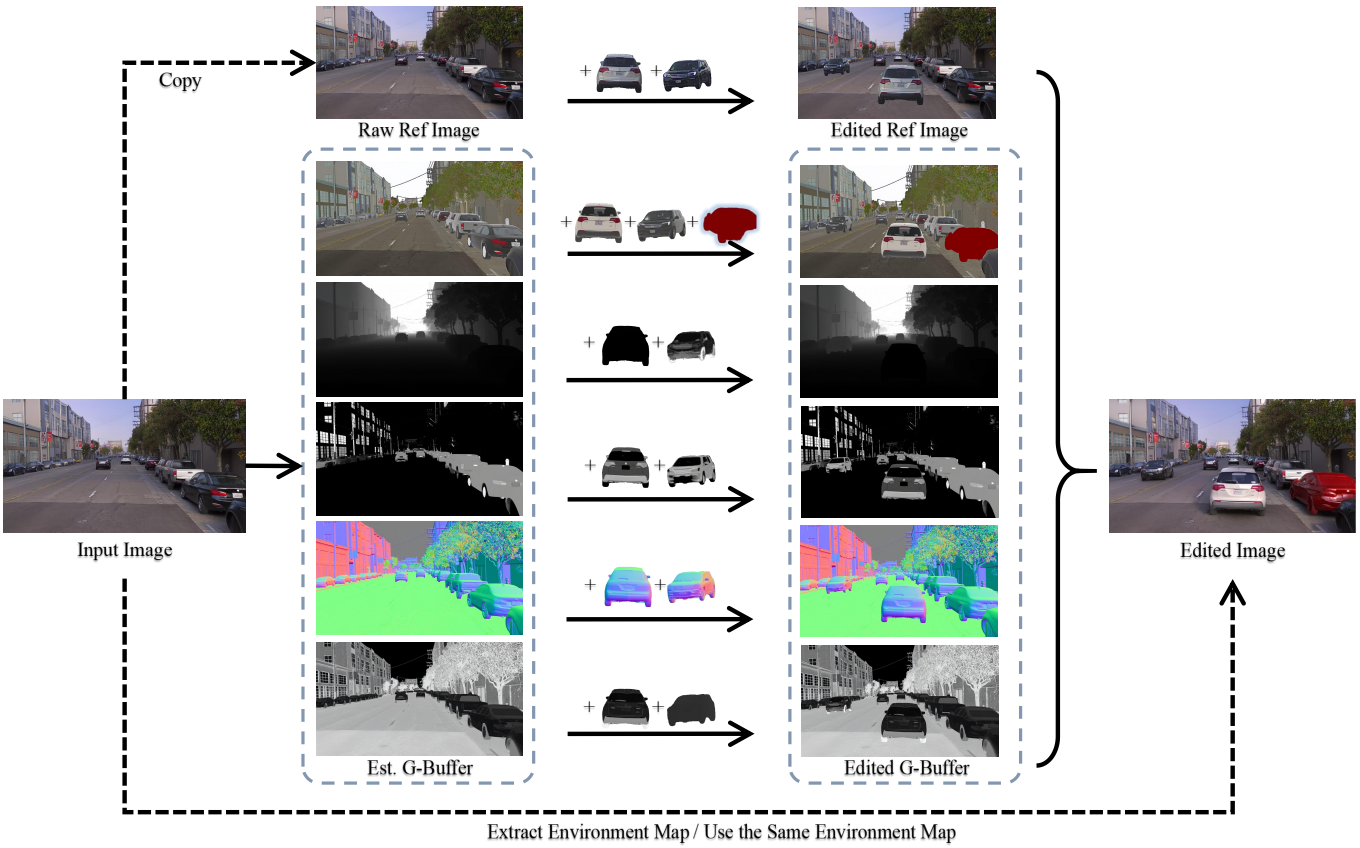

Fig. A20. **Overview of the scene editing workflow.** Includes object insertion and material modification.

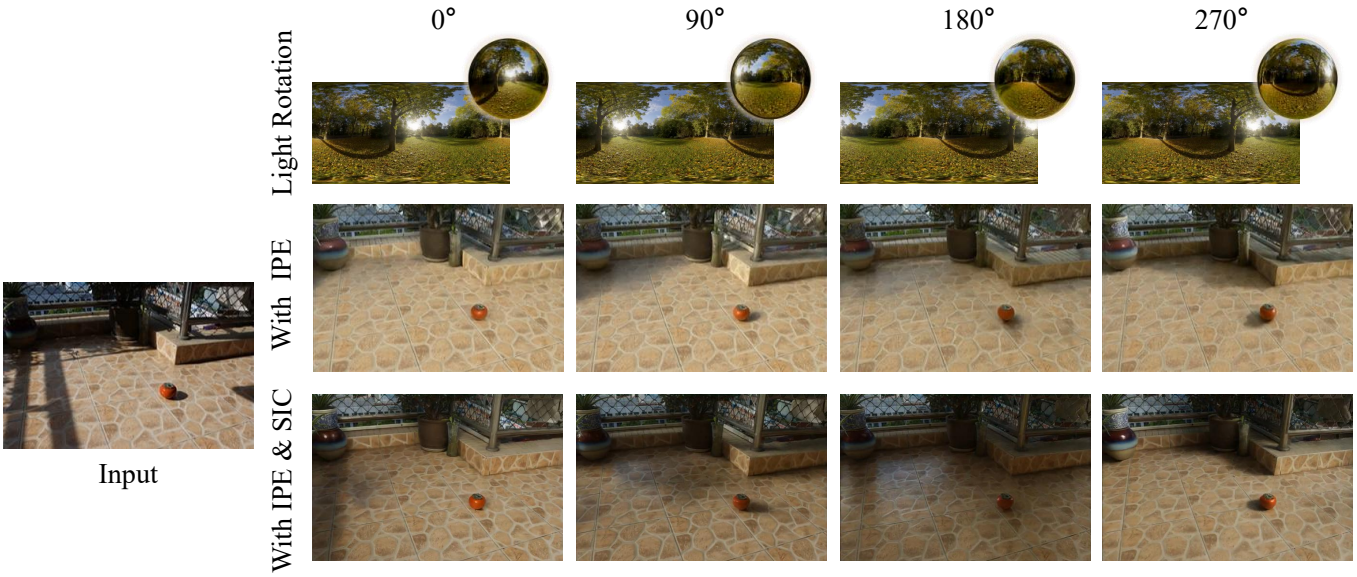

Fig. A21. **Ablation on training strategies.** IPE: Intrinsic Perception Enhancement. SIC: Self-supervised learning based on Illumination Consistency.

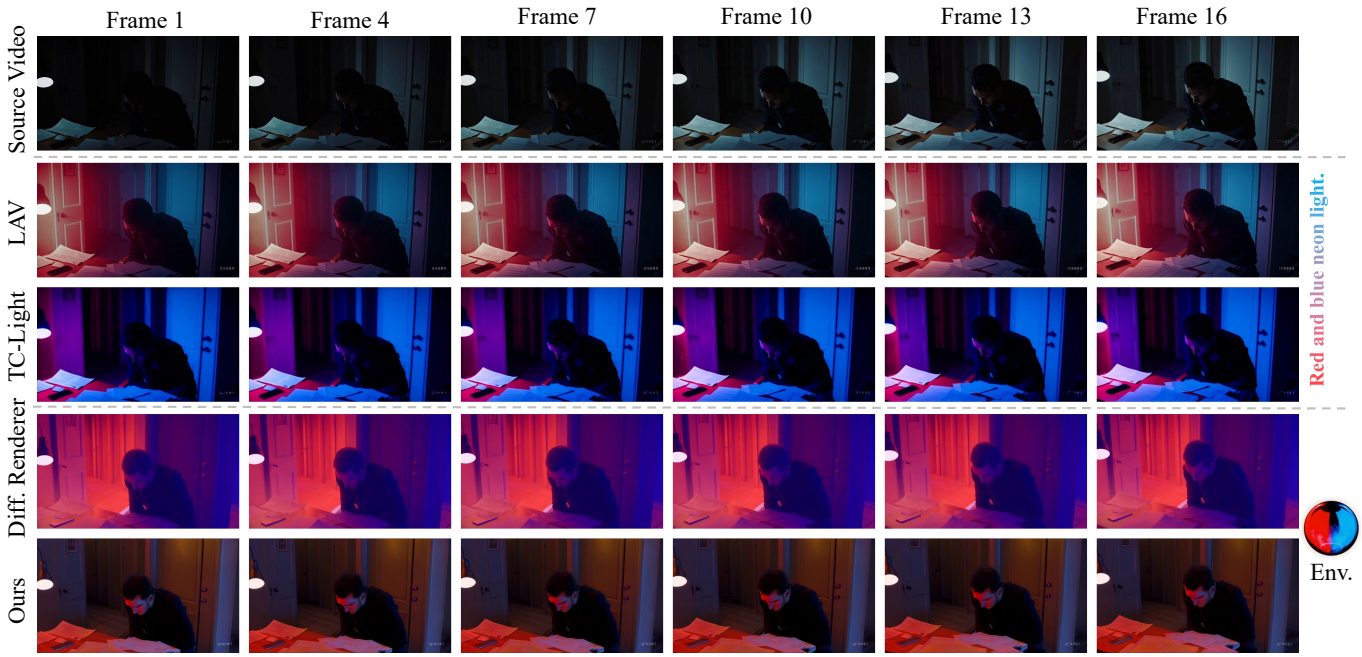

Fig. A22. **Qualitative comparison of video relighting**. Our method achieves superior relighting quality, temporal consistency, and photorealistic generation results compared to baseline methods.

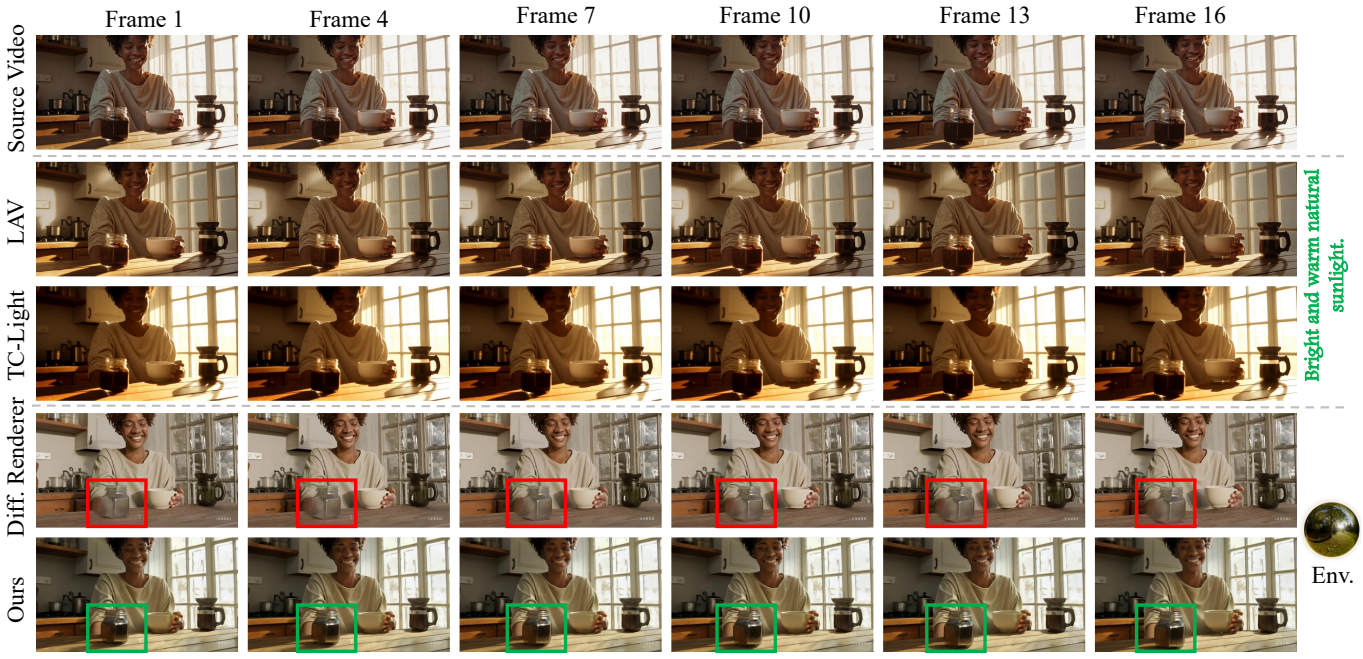

Fig. A23. **Qualitative comparison of video relighting**. Our method achieves superior relighting quality, temporal consistency, and photorealistic generation results compared to baseline methods.

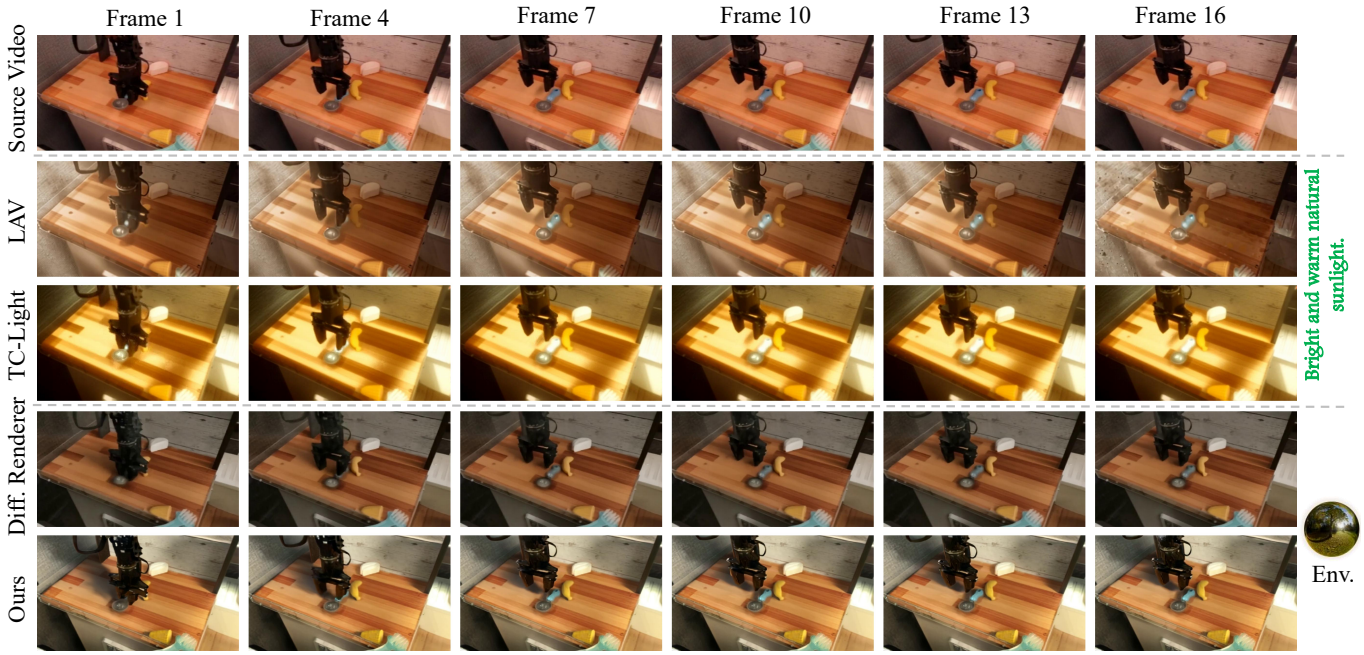

Fig. A24. **Qualitative comparison of video relighting**. Our method achieves superior relighting quality, temporal consistency, and photorealistic generation results compared to baseline methods.

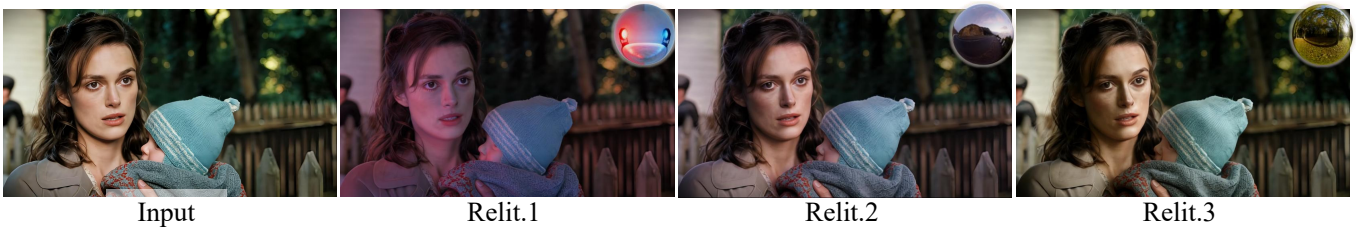

Fig. A25. **Image relighting results of our method on portraits.**

Fig. A26. **Video results under dynamic lighting in a dynamic scene.**

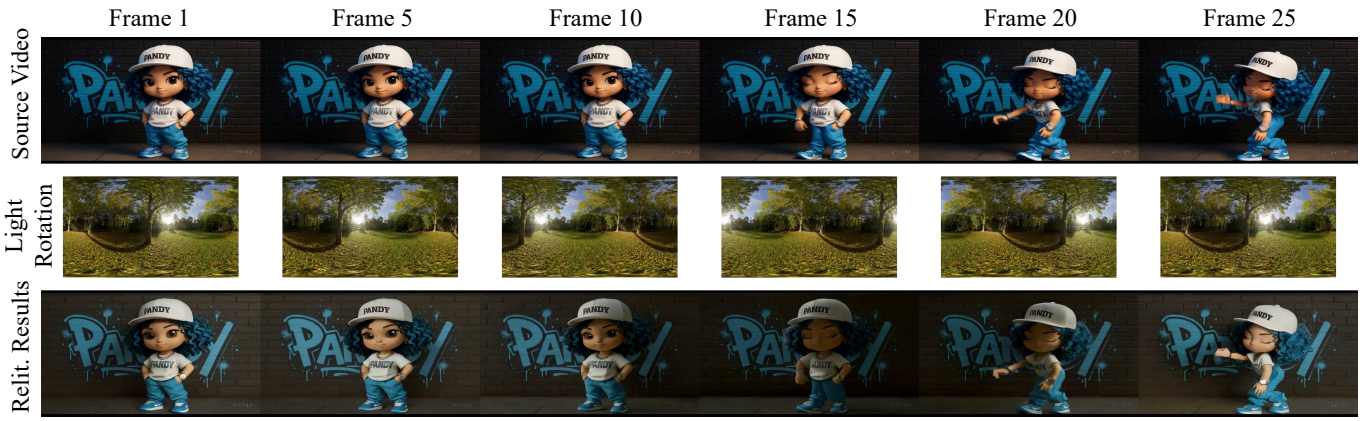

Fig. A27. **Video results under dynamic lighting in a dynamic scene.**

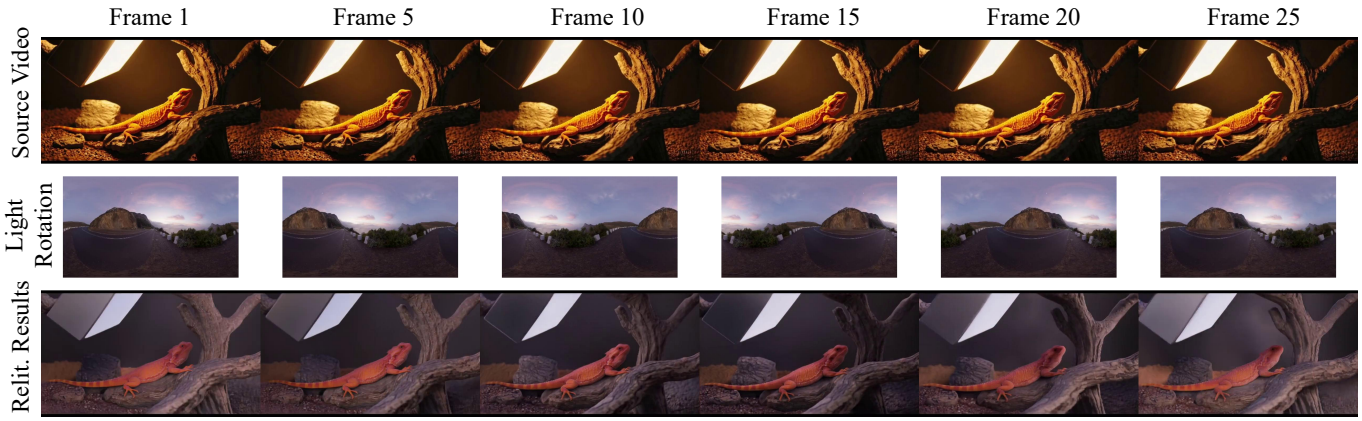

Fig. A28. **Video results under dynamic lighting in a dynamic scene.**

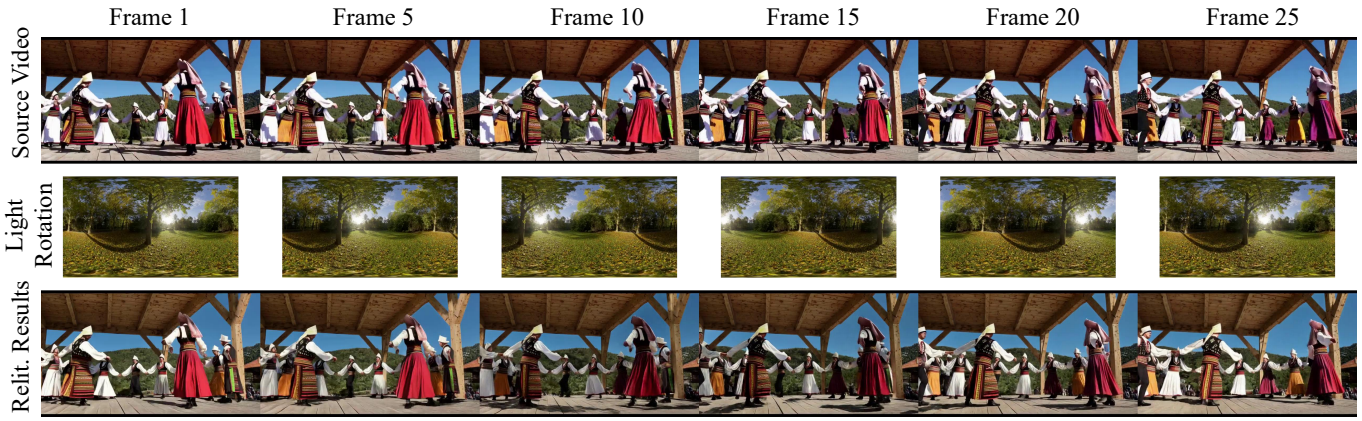

Fig. A29. **Video results under dynamic lighting in a dynamic scene.**

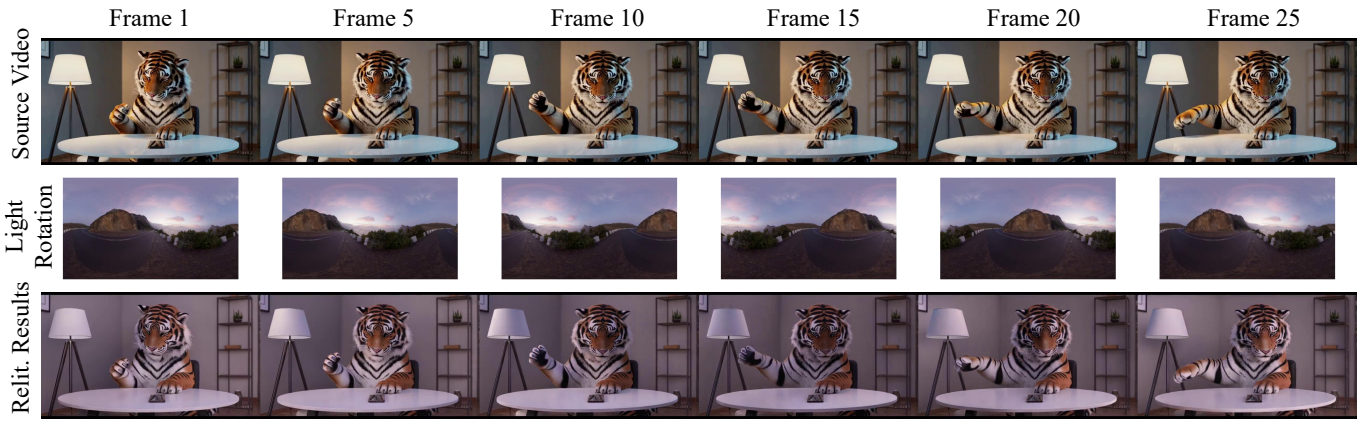

Fig. A30. **Video results under dynamic lighting in a dynamic scene.**

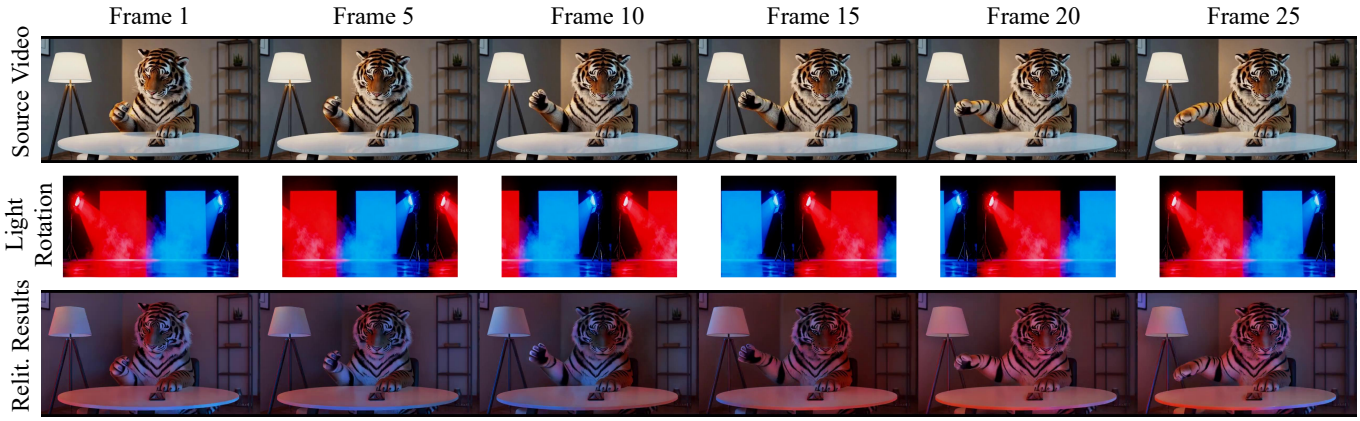

Fig. A31. **Video results under dynamic lighting in a dynamic scene.**

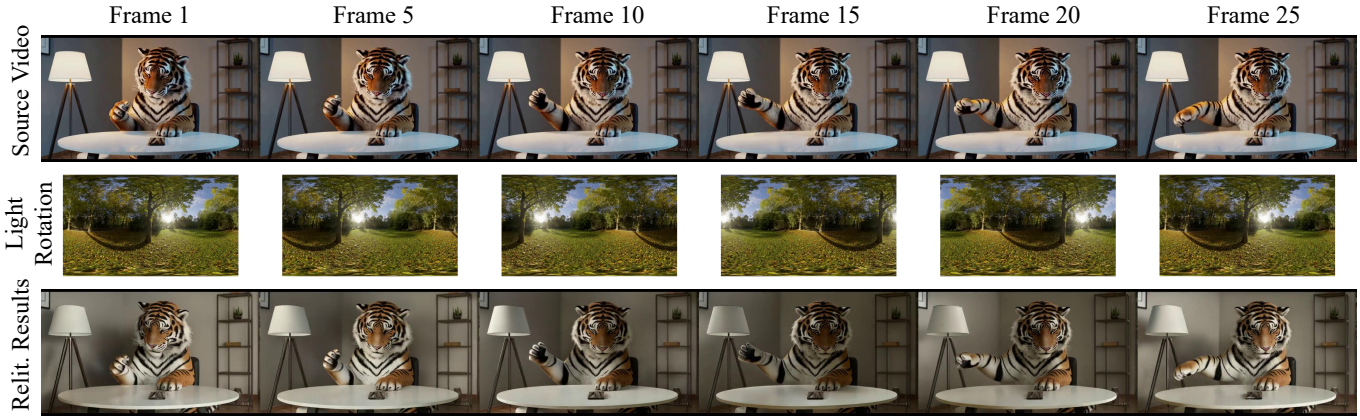

Fig. A32. **Video results under dynamic lighting in a dynamic scene.**

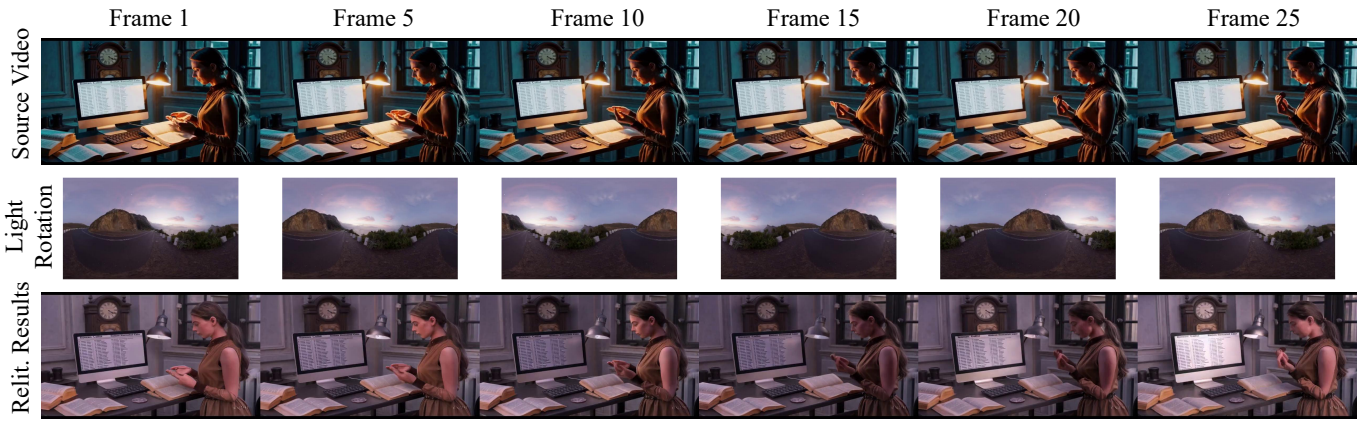

Fig. A33. **Video results under dynamic lighting in a dynamic scene.**

Fig. A34. **Video results of the same scene under different environment lighting conditions.**

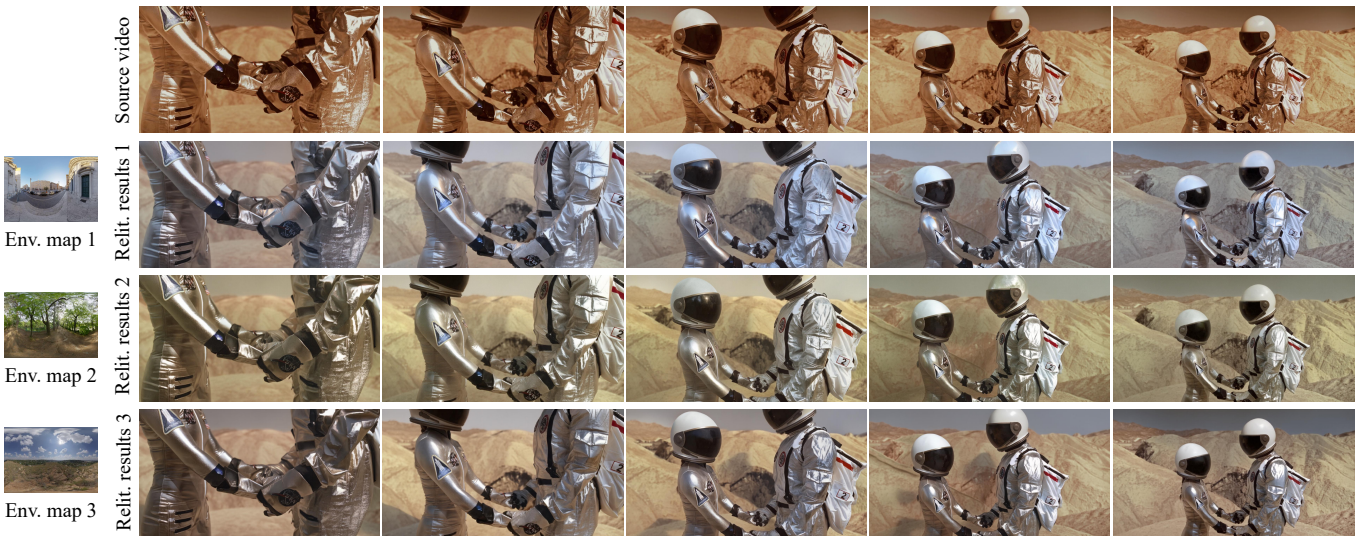

Fig. A35. **Video results of the same scene under different environment lighting conditions.**

Fig. A36. **Video results of the same scene under different environment lighting conditions.**