Supplemental Material: RELIT-LIVE: Relight Video by Jointly Learning Environment Video

WEIQING XIAO*†, Nanjing University, China
HONG LI*†, Beijing Academy of Artificial Intelligence, China and Beihang University, 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, Beijing Academy of Artificial Intelligence, China
CHONGJIE YE, The Chinese University of Hong Kong, Shenzhen, China
HAO ZHAO‡, Tsinghua University, China and Beijing Academy of Artificial Intelligence, China
BEIBEI WANG‡, Nanjing University, China

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. 2025] 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

*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, Beijing Academy of Artificial Intelligence, Beijing, China and Beihang University, 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, Beijing Academy of Artificial Intelligence, 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 Beijing Academy of Artificial Intelligence, Beijing, China; Beibei Wang, beibei.wang@nju.edu.cn, Nanjing University, Suzhou, China.

This work is licensed under a Creative Commons Attribution 4.0 International License.
SIGGRAPH Conference Papers '26, Los Angeles, CA, USA
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2554-8/2026/07
https://doi.org/10.1145/3799902.3811200

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. 2025], for the sets of real-world datasets with and without environment maps, the denoising targets become $\hat{\mathbf{z}}^{\mathbf{s}}(\theta), \hat{\mathbf{z}}^{\mathbf{E}_{\log}}(\theta) = \mathbf{f}_\theta([\mathbf{z}^{\mathbf{I}}, \mathbf{z}_\tau^{\mathbf{s}}, \mathbf{z}_\tau^{\mathbf{E}_{\log}}, \mathbf{z}^{\{\mathbf{a},\mathbf{d},\mathbf{m}\}}, \mathbf{z}^{\{\mathbf{n},\mathbf{r}\}} + \mathbf{c}_{\mathbf{E}}]; \mathbf{c}_{\mathbf{E}}^{\text{cross}}, \tau)$ and $\hat{\mathbf{z}}^{\mathbf{t}}(\theta) = \mathbf{f}_\theta([\mathbf{z}^{\mathbf{I}}, \mathbf{z}_\tau^{\mathbf{t}}, \mathbf{z}_\tau^{\mathbf{0}}, \mathbf{z}^{\{\mathbf{a},\mathbf{d},\mathbf{m}\}}, \mathbf{z}^{\{\mathbf{n},\mathbf{r}\}} + \mathbf{0}]; \mathbf{0}, \tau)$ 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 A1. 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 A1 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 in main text, 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 in main text, our training is divided 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. 2025](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. 2025], 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

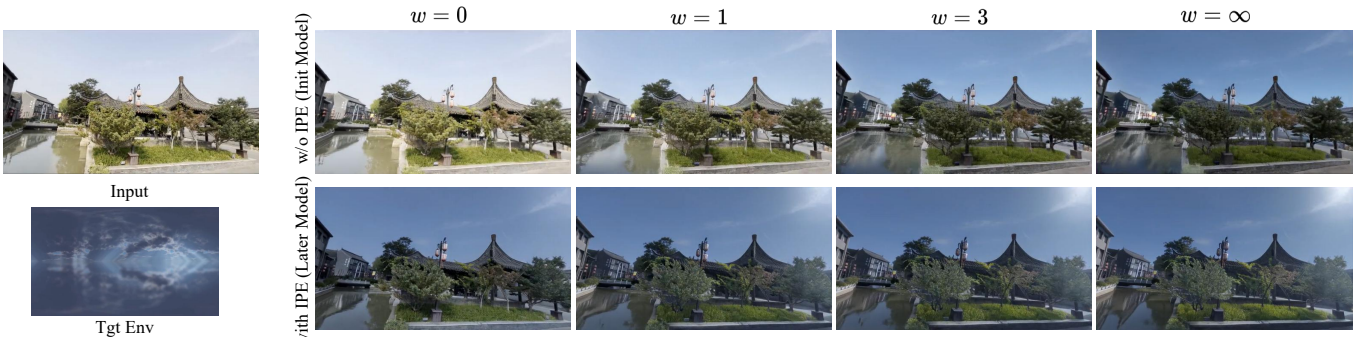

Fig. A1. **Example of latent space interpolation between relighting and rendering results.** We demonstrate model interpolation outcomes before and after applying the Intrinsic Perception Enhancement strategy. Note: When $w = 0$, the interpolated result is fully equivalent to the relighting result; when $w = \infty$, 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.

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 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 A2 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,

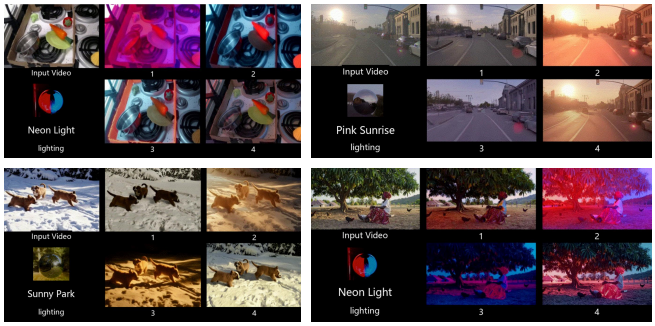

Fig. A2. **The visual interface for our user research.** Participants simultaneously observed the input video, target lighting (text or environment map) and the results of four methods (randomly shuffled) displayed side-by-side. They evaluated each set of results based on three criteria by selecting methods that clearly failed.

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

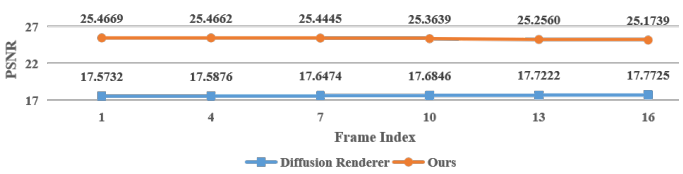

Fig. A3. **Error distribution across different frames.** We compute the average PSNR for each frame across 200 synthetic videos. Our results exhibit high stability over time, comparable to methods that condition on per-frame environment maps.

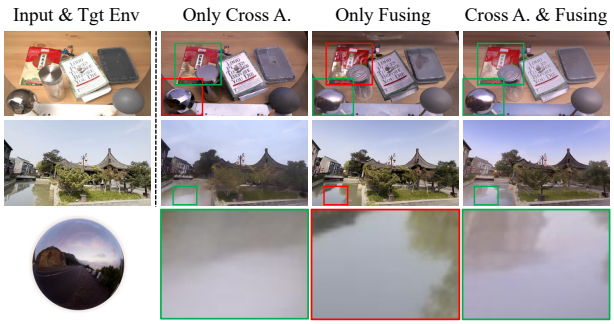

Fig. A4. **Qualitative comparison under different light conditions.** Only Cross A.: Input light conditions via cross-attention. Only Fusing: By fusing with scene property latents to input light conditions. The dual-path lighting control method achieves the most balanced performance in terms of reflection quality and color temperature.

for Diffusion Renderer, which defines environment maps frame-by-frame. Nevertheless, our approach achieves optimal performance, showcasing its robustness across diverse dynamic scenes.

We visualize the error distribution of each method across different frames for video rendering task, as shown in Figure A3. 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 A4. 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 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 A6. 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 in main text 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 A5, 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.

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. A5. Failure cases of relighting: color shift (left) and abnormal illumination (right).

A.10 More visualizations of our methods

Figure A7 visualizes the performance of our training strategy. Figures A8-A10 present additional visual comparisons of our method against other methods on in-the-wild data. Figures A11-A22 show the results of our method under dynamic lighting and various environment lights.

References

Kai He, Ruofan Liang, Jacob Munkberg, Jon Hasselgren, Nandita Vijaykumar, Alexander Keller, Sanja Fidler, Igor Gilitschenski, Zan Gojcic, and Zian Wang. 2025. 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.

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.

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. 2025. TC-Light: Temporally Coherent Generative Rendering for Realistic World Transfer. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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.

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).

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:

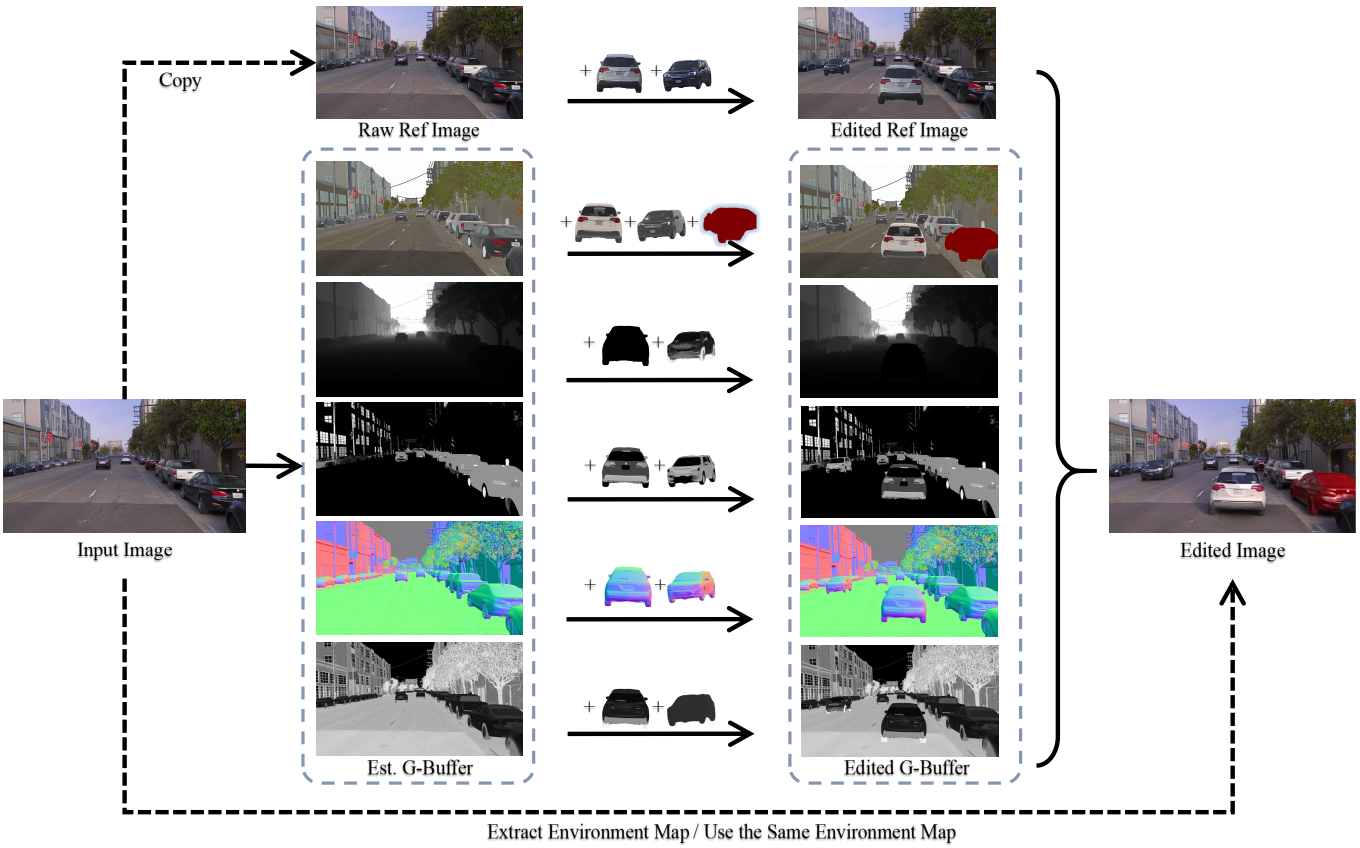

Fig. A6. **Overview of the scene editing workflow.** Includes object insertion and material modification.

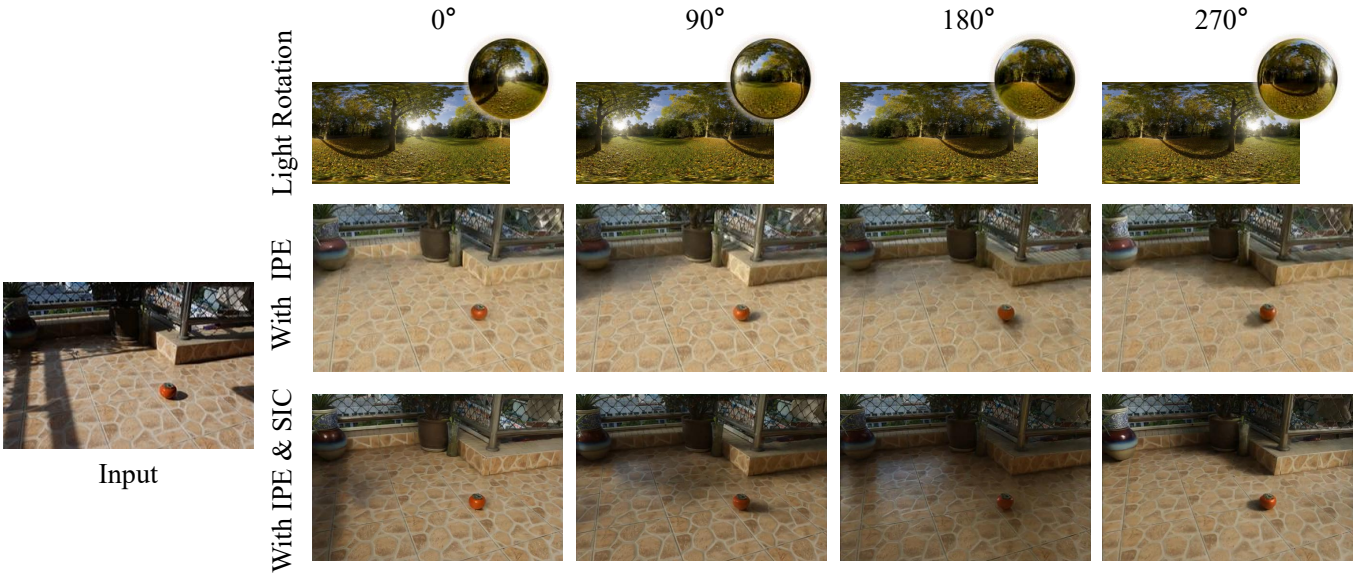

Fig. A7. **Ablation on training strategies.** IPE: Intrinsic Perception Enhancement. SIC: Self-supervised learning based on Illumination Consistency.

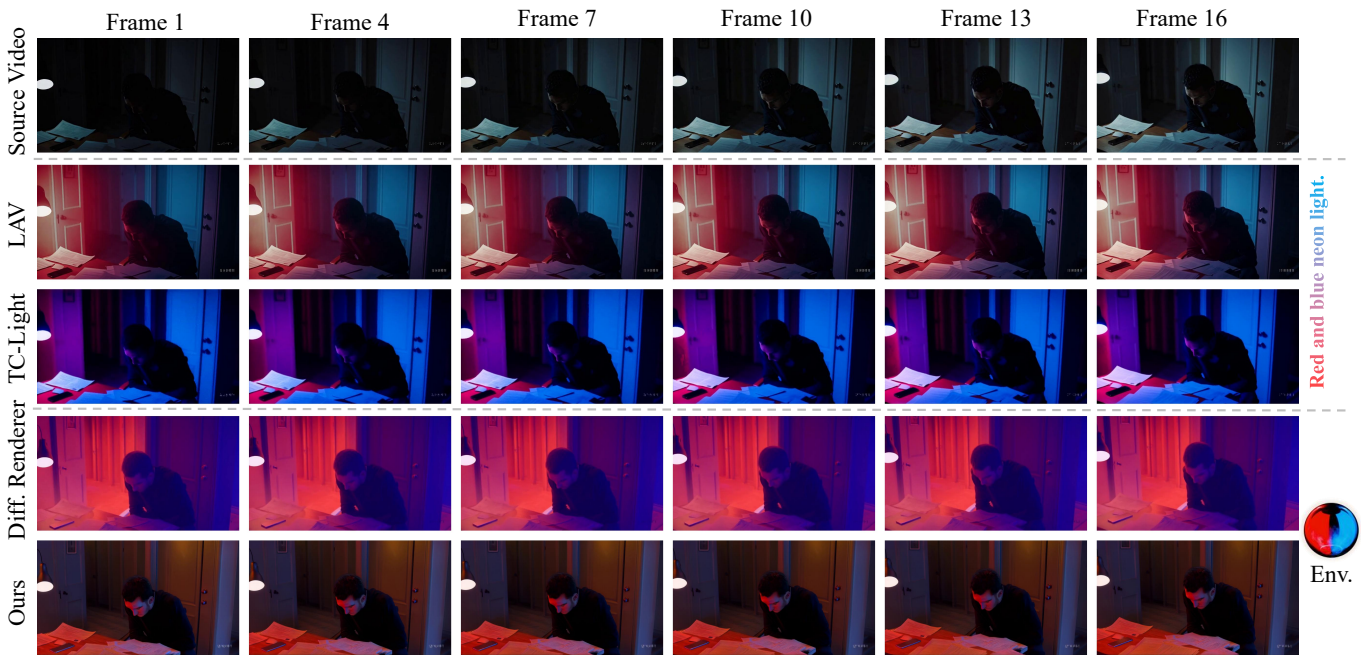

Fig. A8. **Qualitative comparison of video relighting**. Our method achieves superior relighting quality, temporal consistency, and photorealistic generation results compared to baseline methods.

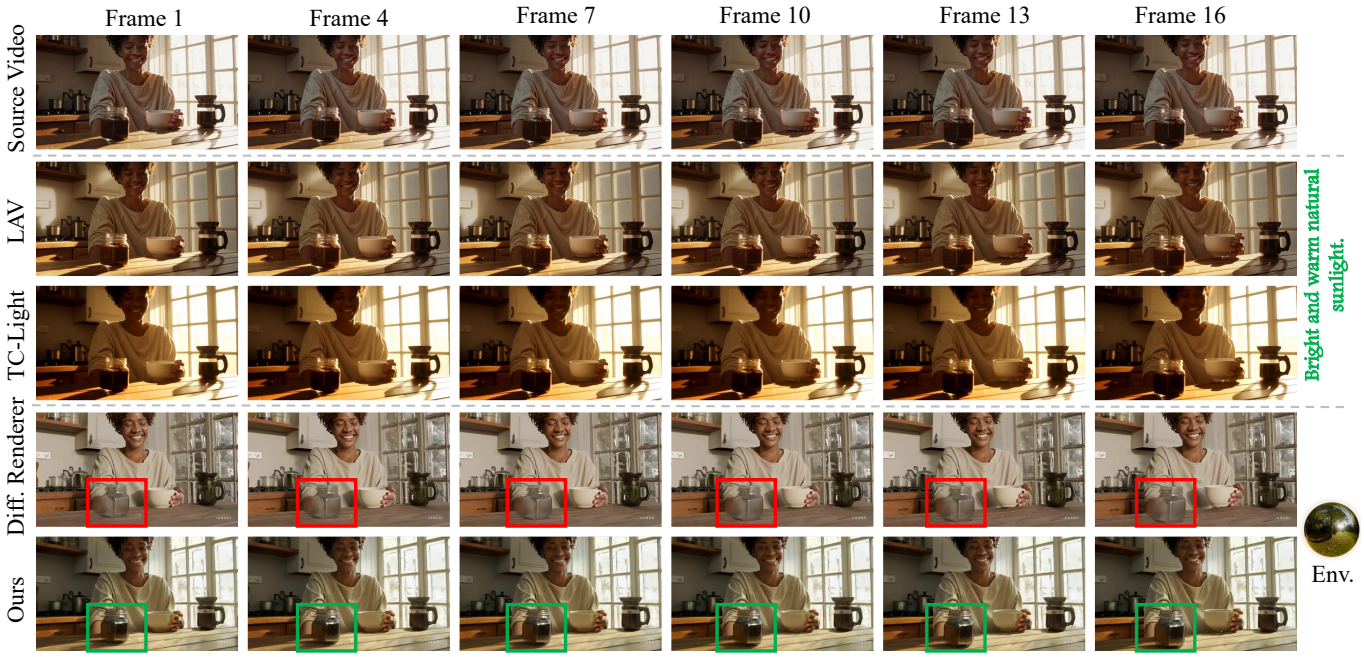

Fig. A9. **Qualitative comparison of video relighting**. Our method achieves superior relighting quality, temporal consistency, and photorealistic generation results compared to baseline methods.

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.

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

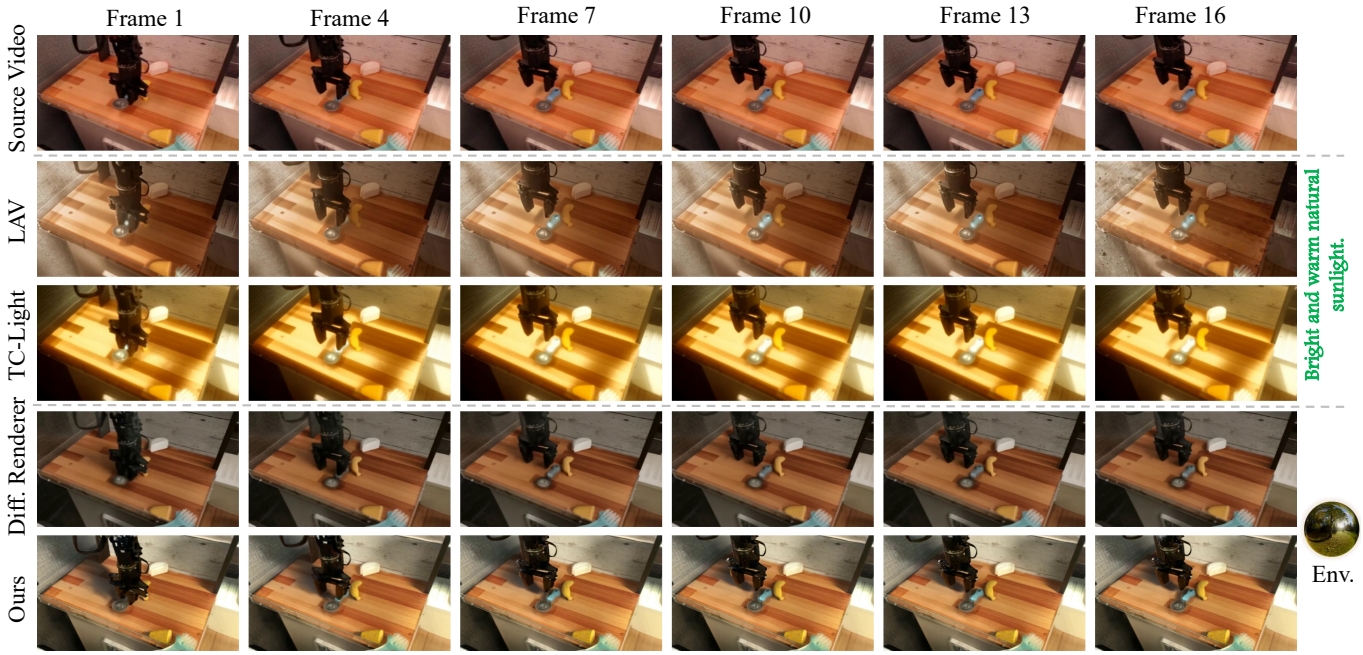

Fig. A10. **Qualitative comparison of video relighting**. Our method achieves superior relighting quality, temporal consistency, and photorealistic generation results compared to baseline methods.

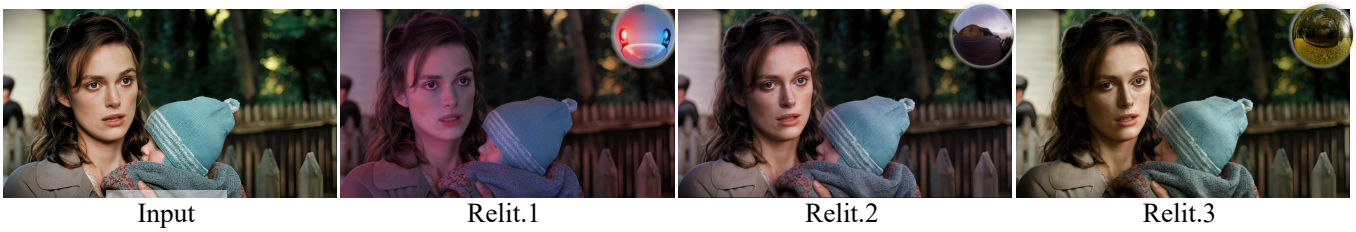

Fig. A11. **Image relighting results of our method on portraits.**

Fig. A12. **Video results under dynamic lighting in a dynamic scene.**

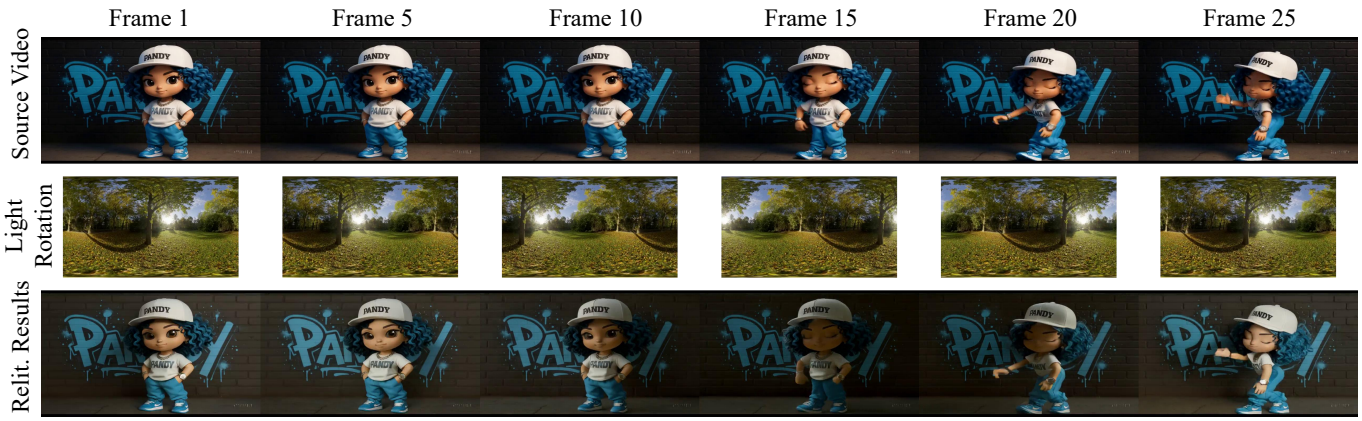

Fig. A13. **Video results under dynamic lighting in a dynamic scene.**

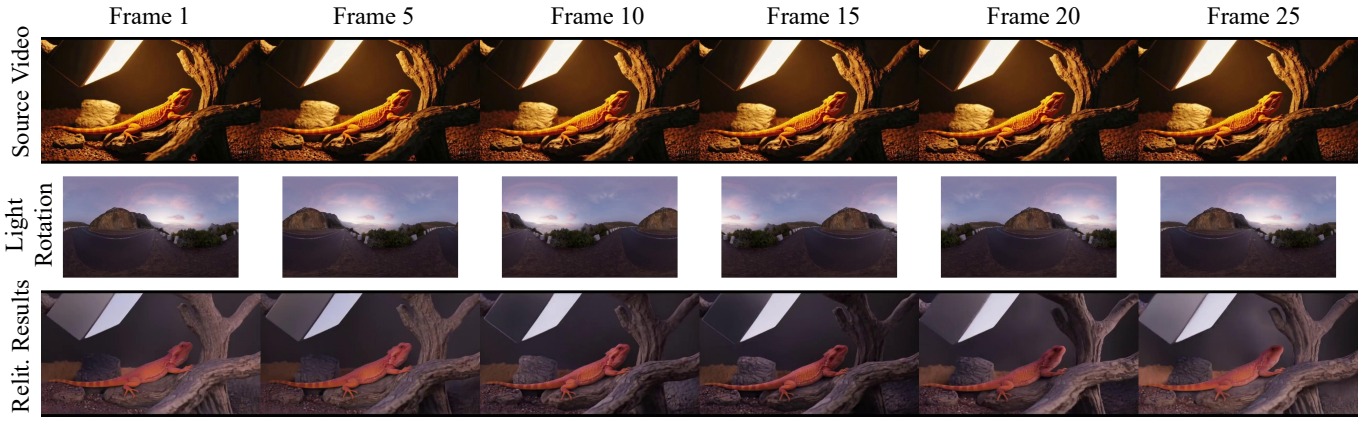

Fig. A14. **Video results under dynamic lighting in a dynamic scene.**

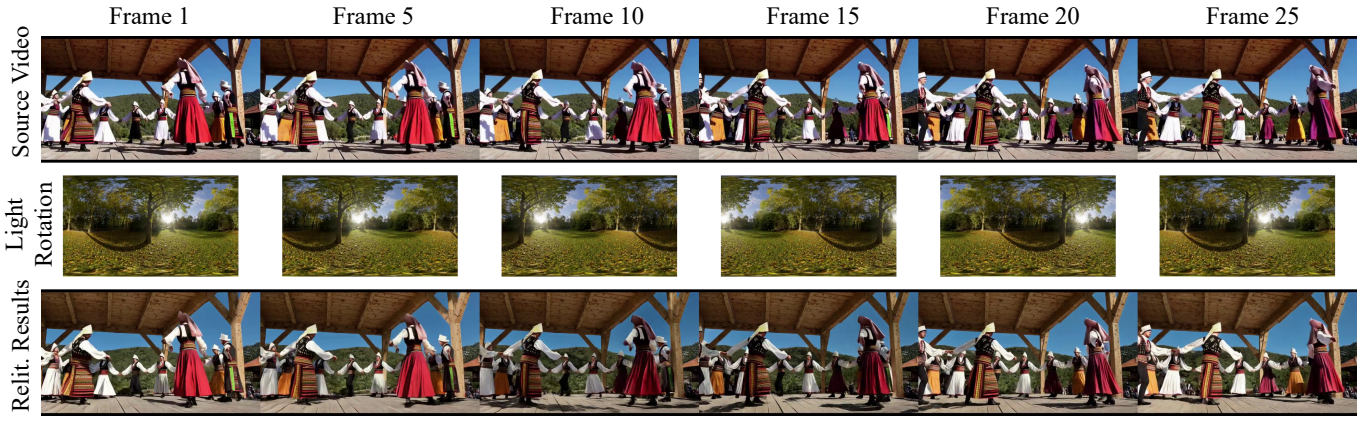

Fig. A15. **Video results under dynamic lighting in a dynamic scene.**

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. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388* (2025).

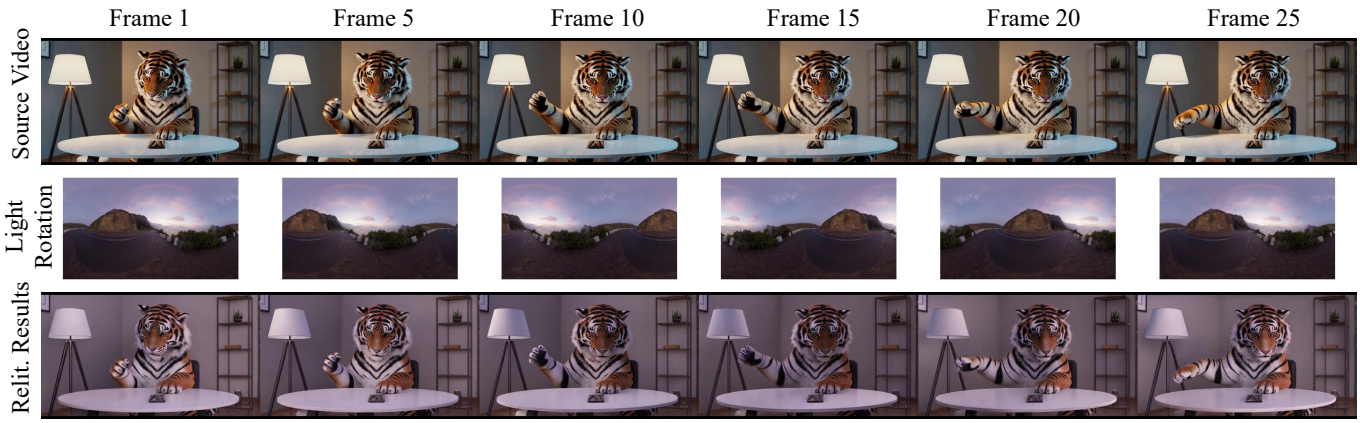

Fig. A16. **Video results under dynamic lighting in a dynamic scene.**

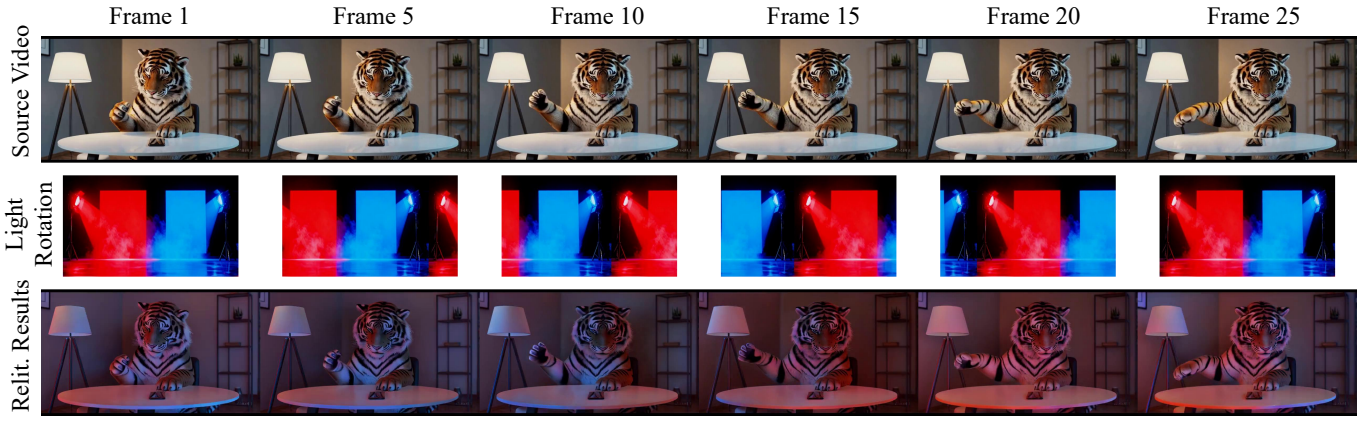

Fig. A17. **Video results under dynamic lighting in a dynamic scene.**

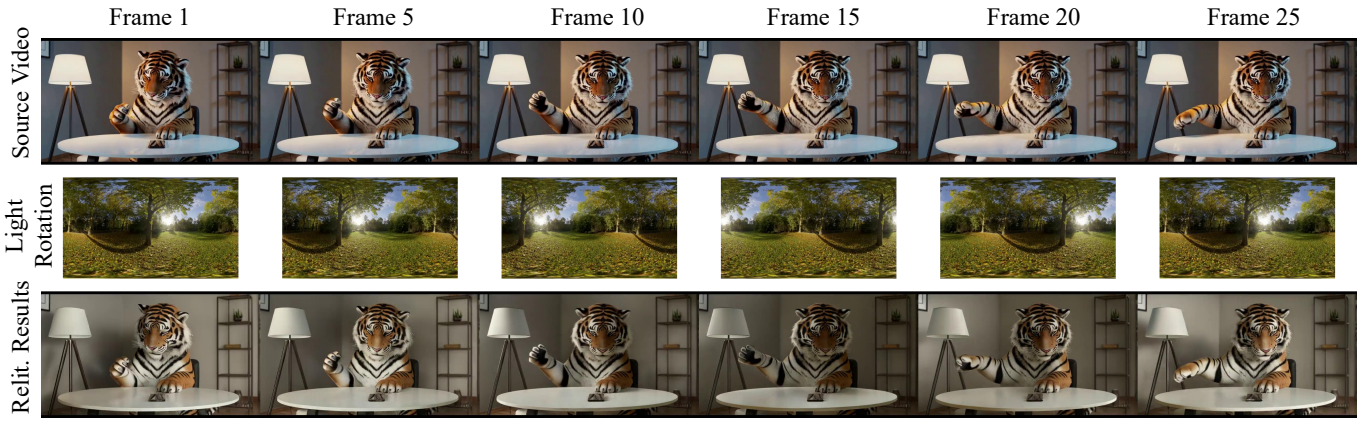

Fig. A18. **Video results under dynamic lighting in a dynamic scene.**

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

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.

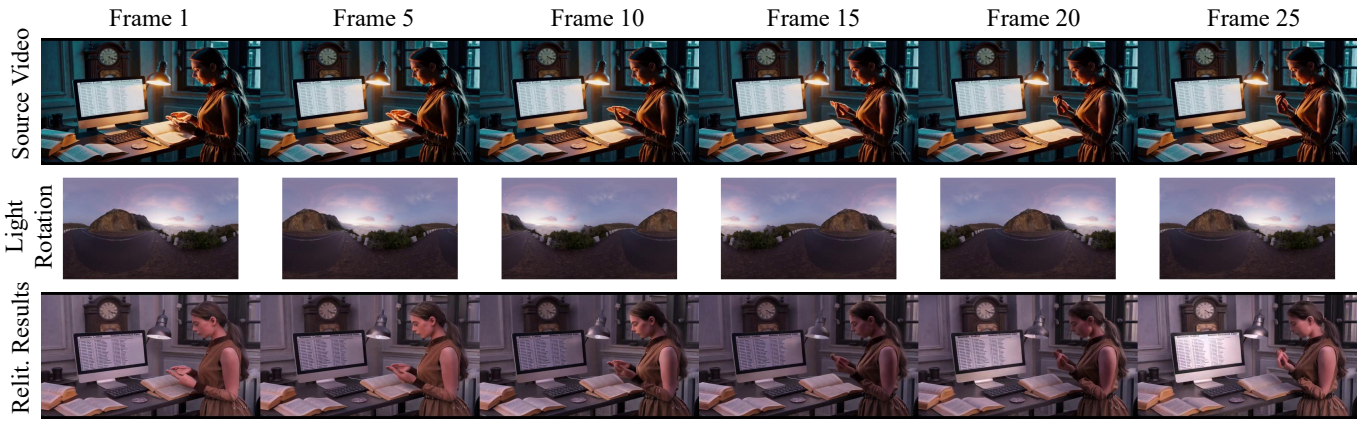

Fig. A19. **Video results under dynamic lighting in a dynamic scene.**

Fig. A20. **Video results of the same scene under different environment lighting conditions.**

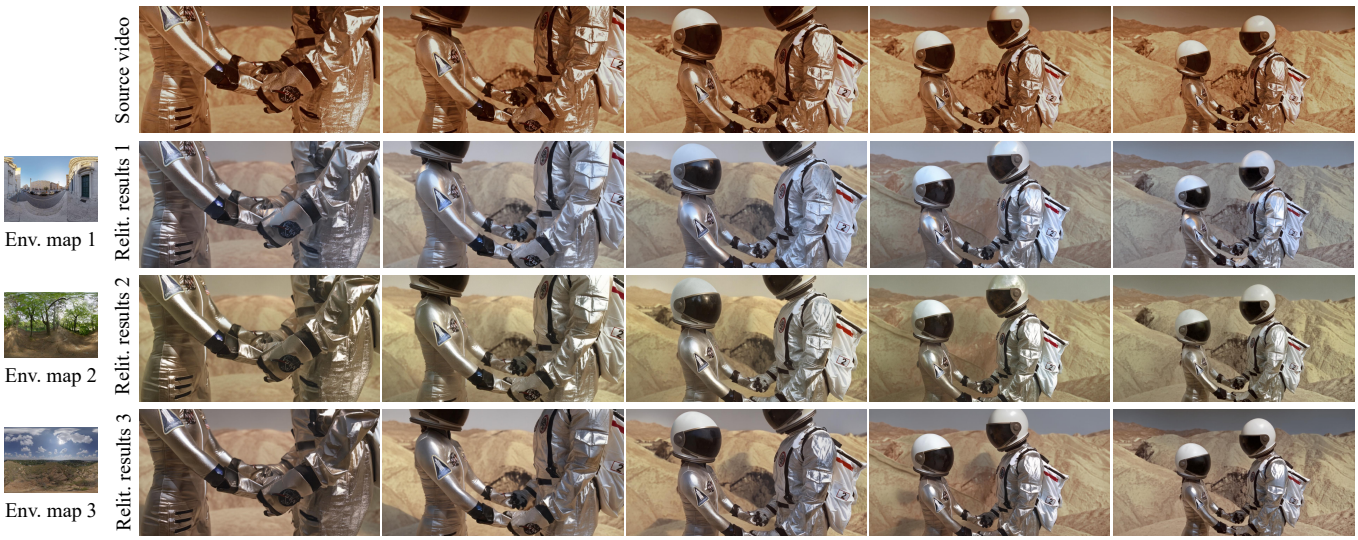

Fig. A21. **Video results of the same scene under different environment lighting conditions.**

Fig. A22. **Video results of the same scene under different environment lighting conditions.**