

LiveLight: Real-time Streaming Video Relighting with Interactive Control

ANONYMOUS AUTHOR(S)

SUBMISSION ID: 1536

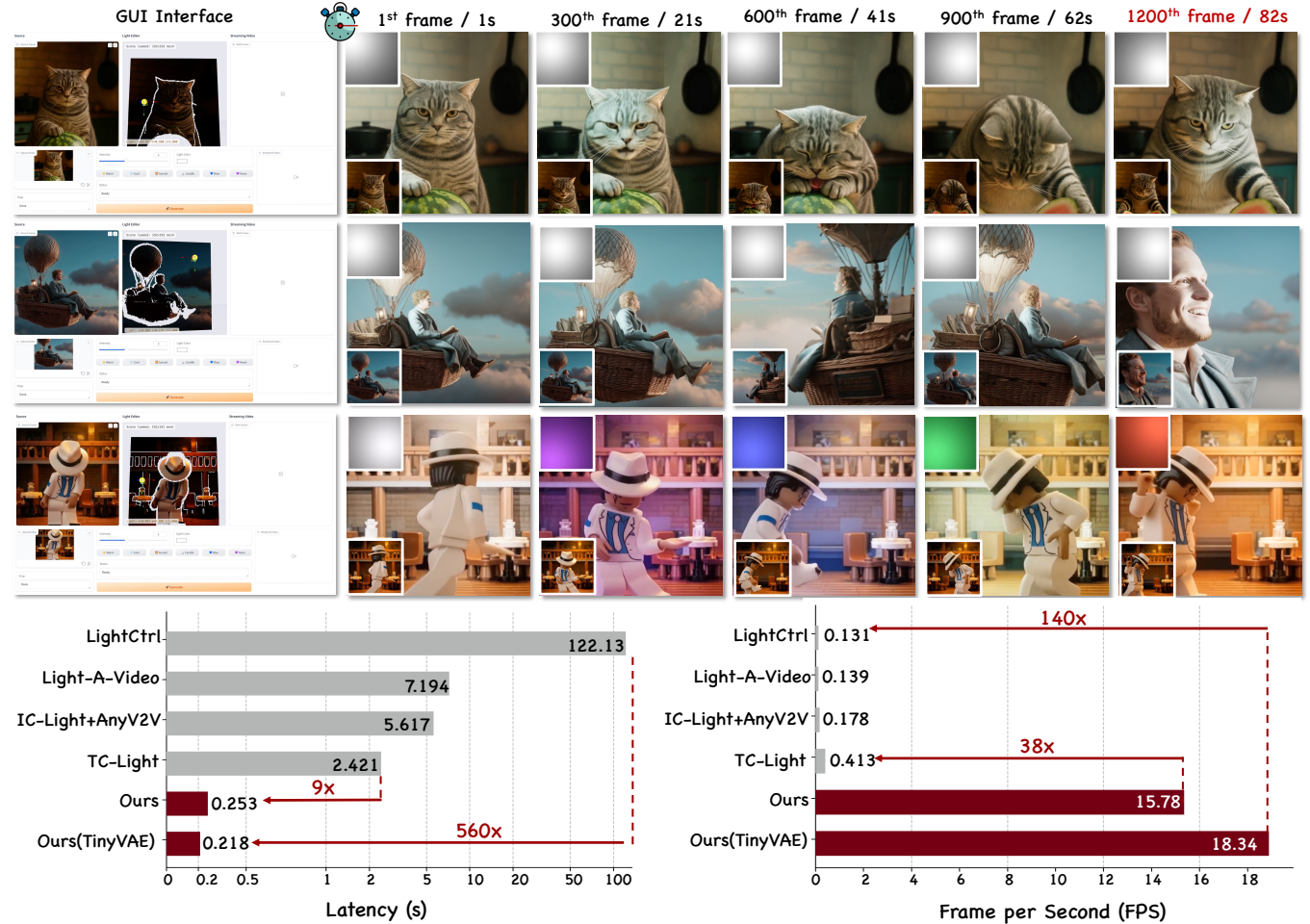


Fig. 1. **Showcase of the proposed LiveLight.** LiveLight enables real-time streaming video relighting with interactive 3D point-light control, supporting dynamic adjustment of light position, intensity, and color while preserving temporal coherence and visual quality.

We present LiveLight, the first diffusion-based framework for real-time streaming video relighting with interactive 3D lighting control. Achieving this is non-trivial, as it requires overcoming three critical challenges: effectively injecting dynamic 3D lighting into a diffusion model, maintaining high-fidelity generation under an extremely low NFE (Number of Function

Evaluations) budget for real-time speed, and facilitating continuous streaming for interactive control. To address these pain points, we propose three key designs. First, for accurate lighting injection, we propose a lightweight adapter that feeds Multi-Plane Light Irradiance (MPLI) conditions—depth-aware irradiance maps encoding 3D lighting geometry—directly into the diffusion backbone. Second, to prevent rendering quality degradation at low NFEs towards real-time distillation, we introduce a geometry-guided feedback branch. This training-time constraint leverages a frozen geometry estimator to enforce depth- and normal-consistent relighting, ensuring geometrically plausible shading without adding inference overhead. Finally, to enable streaming interaction, we develop a progressive rolling-window strategy that maintains a denoising ladder of latent chunks at varying noise levels. By propagating intermediate states, this strategy guarantees temporal coherence and supports arbitrarily long video relighting with per-frame

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM 1557-7368/2026/5-ART
<https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

reference refresh. Extensive experiments on real-world and synthetic benchmarks demonstrate that LiveLight achieves state-of-the-art relighting quality while running at real-time speed, significantly outperforming offline baselines in temporal stability, lighting controllability, and user preference. To foster real-time interactive relighting research, we will publicly release our models, training data, and synthetic data generator.

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

Additional Key Words and Phrases: Real-time Video Relighting, Real-time Video Generation, Diffusion Model

ACM Reference Format:

Anonymous Author(s). 2026. **LiveLight**: Real-time Streaming Video Relighting with Interactive Control. *ACM Trans. Graph.* 1, 1 (May 2026), 14 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 Introduction

Video relighting is a fundamental capability for film production, virtual cinematography, and live-streaming content creation, enabling the alteration of illumination in captured scenes. In these demanding workflows, real-time streaming video relighting introduces a transformative, interactive paradigm. By responding to continuously changing light parameters with minimal latency and ensuring strict temporal coherence across consecutive frames, this capability unlocks unprecedented creative freedom. For instance, a user can dynamically drag a virtual light source through a scene and instantly preview how shadows, highlights, and colors evolve over a video stream, entirely bypassing the prohibitive delays of offline re-rendering. Ultimately, this seamless integration of high-fidelity relighting and real-time interaction redefines the boundaries of dynamic content creation, bridging the gap between professional studio effects and accessible, on-the-fly editing.

Recent diffusion-based relighting methods have made impressive progress on individual aspects of this problem. For image relighting, IC-Light [Zhang et al. 2025] fine-tunes a pretrained diffusion model for light-conditioned generation and achieves strong results. For video relighting, LightCtrl [Peng et al. 2026] introduces lightweight physical priors to improve controllability without dense intrinsic supervision. Light-A-Video [Zhou et al. 2025] and TC-Light [Liu et al. 2026] combine per-frame image relighting with video diffusion priors in a training-free manner, and RelightVid [Fang et al. 2025] trains temporal models on curated relighting datasets.

Despite these advances, enabling real-time interactive video relighting remains highly challenging. Existing video relighting methods operate in an offline, clip-level paradigm: the full lighting trajectory must be specified before generation, and the entire video is produced in a single pass. This design precludes real-time interactive control and imposes a high computational cost that scales linearly with video length. Alternatively, applying image-based relighting models independently to each frame can process inputs incrementally, but this inevitably produces severe temporal flicker, rendering it unsuitable for coherent video streaming.

An ideal real-time video relighting system should be interactive, streaming, and high-fidelity, capable of providing instant visual feedback as the user adjusts the lighting direction. Intuitively, achieving this requires a fundamental departure from the conventional clip-level generation paradigm. Instead, the generation process must be fine-grained to the frame level and operate under an extremely

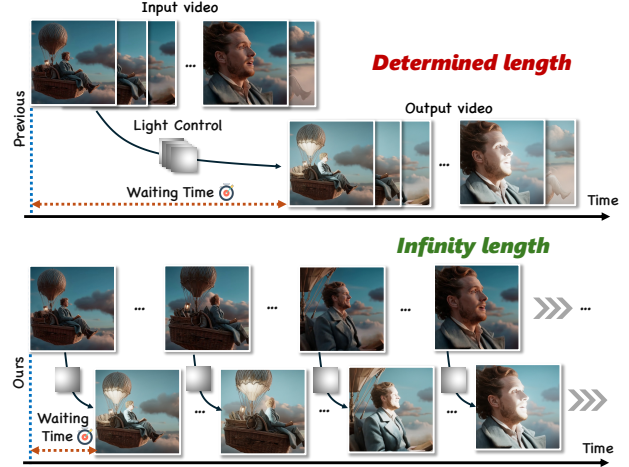


Fig. 2. **Motivation for LiveLight.** Previous offline video relighting methods require the entire input video and a pre-defined lighting trajectory before generation, producing a fixed-length output after a long waiting time. In contrast, LiveLight processes input frames in a streaming fashion with per-frame dynamic lighting control, producing relit frames with low latency and supporting arbitrarily long video generation.

low NFE (Number of Function Evaluations) budget to guarantee real-time latency. Furthermore, it necessitates a genuine streaming mechanism to ensure that the continuously generated frames remain strictly temporally coherent, as illustrated in Fig. 2. Driven by these rigorous requirements, we present **LiveLight**, the first diffusion-based framework tailored to solve these challenges. To this end, we introduce three complementary designs:

(i) **Lightweight adapter for interactive light injection.** To support fine-grained, frame-level interactive relighting, we use the Multi-Plane Light Irradiance (MPLI) representation [Bian et al. 2025] as a compact geometry-aware carrier of user-specified lighting parameters. The MPLI maps are then encoded by a lightweight adapter with only $\sim 1\text{M}$ parameters into compact lighting tokens, which interact with the diffusion features through cross-attention in shallow layers. This localized injection strategy introduces spatially varying illumination cues at an early stage, while avoiding unnecessary perturbation to deeper denoising blocks that encode appearance priors. As a result, the model preserves identity and texture fidelity with negligible overhead, allowing changes in light position, intensity, and color to be reflected immediately for interactive control.

(ii) **Real-time distillation with geometry-guided feedback.** To achieve real-time interactive relighting, we do not simply truncate the diffusion sampling process at inference time. Instead, we distill the model with the same compact few-step rollout used during testing, enabling it to produce effective illumination updates under an extremely low NFE budget. However, aggressive step reduction can weaken the geometric consistency of synthesized highlights, shadows, and depth-dependent illumination. To address this issue, we introduce a training-time geometry-guided feedback branch. A frozen geometry estimator predicts depth and normal cues from the generated image and back-propagates geometry-aware gradients

to the relighting model. This supervision encourages the generated shading to remain consistent with the underlying structure, compensating for the quality degradation caused by a few-step inference without introducing any inference-time cost.

(iii) **Rolling-Window Streaming Video Relighting.** To fundamentally break away from the conventional offline, clip-level paradigm, we introduce a genuine streaming mechanism formulated as a rolling-window process over micro-chunks. A sliding window of latent micro-chunks is maintained, where at each streaming step the video denoising backbone jointly processes all chunks: the oldest chunk is fully denoised, decoded, and emitted, while the remaining chunks are refined as temporal context and a new noisy chunk is appended to the tail. This design achieves strict temporal coherence through persistent latent propagation without redundant overlap blending. Crucially, it refreshes the reference and lighting conditions per frame, keeping appearance, geometry, and lighting synchronized with the input video to support dynamic, arbitrary-length interactive sessions.

Extensive experiments demonstrate that LiveLight achieves state-of-the-art relighting quality on both real-world and synthetic benchmarks while running at real-time speed. Compared with offline methods, LiveLight produces more temporally stable results over long videos and enables interactive lighting adjustment with low latency. A user study further confirms that our method is preferred over existing baselines in lighting accuracy, appearance consistency, and temporal coherence.

In summary, our contributions are as follows:

- We propose LiveLight, the first diffusion-based framework for real-time, streaming video relighting with interactive lighting control.
- We design a lightweight adapter that efficiently integrates the MPLI lighting representation into the diffusion backbone, and introduce a geometry-guided feedback branch that enforces structure-preserving shading at training time without inference overhead.
- We design a rolling-window generation strategy with per-frame reference refresh that propagates intermediate latent states for temporal coherence, enabling low-latency and arbitrarily long video relighting.
- Extensive experiments on real-world and synthetic benchmarks show that LiveLight achieves superior quality, efficiency, and user preference over state-of-the-art methods.

2 Related Work

Relighting Methods. Relighting has been extensively studied in graphics, where traditional approaches rely on inverse rendering [Debevec 2008; Zhang et al. 2021a, 2022] or light stage captures [Debevec et al. 2000] to decompose and re-render scenes under novel illumination. While physically grounded, these methods require specialized inputs and are largely restricted to portraits [Pandey et al. 2021]. Recent diffusion-based methods [Chaturvedi et al. 2025; Jin et al. 2024; Kim et al. 2024; Kocsis et al. 2024; Ren et al. 2024; Zeng et al. 2024; Zhang et al. 2025] have demonstrated flexible image relighting with diverse control modalities, among which IC-Light [Zhang et al. 2025] has become a widely adopted

backbone owing to its effective light transport formulation. Extending to video, RelightVid [Fang et al. 2025] trains temporal models on curated relighting datasets, while Light-A-Video [Zhou et al. 2025] and LightCtrl [Peng et al. 2026] adopt training-free pipelines that combine per-frame image relighting with video diffusion priors. Other recent works [He et al. 2026; Liang et al. 2025a; Lin et al. 2026; Xue et al. 2026; Zeng et al. 2025; Zhang et al. 2024b] further explore various types of lighting control within large-scale video generative models. However, all these methods operate in an offline, clip-level paradigm where the full lighting condition must be specified before generation, precluding real-time interactive adjustment. Our method instead performs streaming relighting, responding to dynamic light map inputs frame by frame.

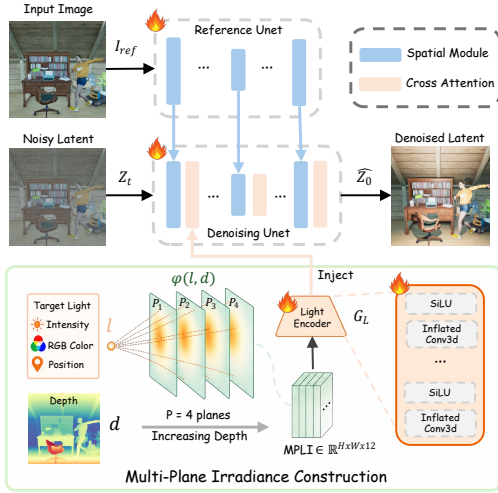
Real-time Video Generation. Video diffusion models [Blattmann et al. 2023; Chen et al. 2024; Huang et al. 2026; Kong et al. 2024; Wan et al. 2025; Yang et al. 2024] produce high-quality and temporally coherent results, but their iterative denoising and full-clip computation make low-latency generation challenging. A line of work accelerates sampling by reducing the number of inference steps, including progressive distillation [Salimans and Ho 2022], consistency models [Wang et al. 2023; Zheng et al. 2024], rectified flow [Esser et al. 2024; Liu et al. 2022], and shortcut models [Frans et al. 2024], enabling few-step or even one-step generation. Beyond step reduction, another line of work designs streaming pipelines that produce frames continuously: StreamDiffusion [Kodaira et al. 2025] batches denoising for real-time image generation, StreamV2V [Liang et al. 2025b] extends this idea to video-to-video translation with a feature bank.

Controllable Video Generation. Controllable video generation [He et al. 2024; Ma et al. 2024; Tian et al. 2024; Xing et al. 2025, 2024; Zhang et al. 2024a; Zhao et al. 2024] synthesizes videos that follow user-specified conditions beyond text prompts, which is essential for creative workflows where text alone is too coarse for fine-grained user intent. To this end, a common approach is to inject auxiliary signals into pre-trained video diffusion models, with modalities ranging from structural maps and motion trajectories to identity and lighting [Ma et al. 2025]. Despite the variety of modalities, these methods share the same offline pipeline: the full conditioning sequence is fixed before generation, and the video is produced in a single pass. As a result, users cannot refine inputs based on intermediate results, which is impractical for iterative creative work. A more interactive paradigm appears in 3D-aware portrait relighting [Cai et al. 2024; Mei et al. 2025; Zhang et al. 2021b], where users can drag light directions and see immediate updates, yet these methods only handle portraits and rely on neural rendering. Our method brings such interactivity into diffusion-based streaming relighting: the user adjusts a 3D light source's position, intensity, and color in a reconstructed scene mesh, and the relit video is produced chunk-by-chunk following the live light state.

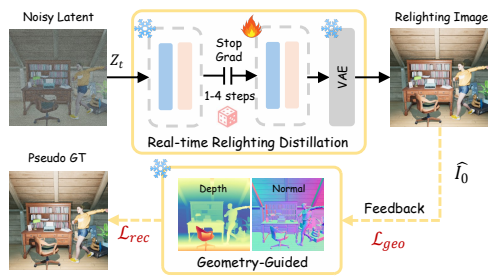
3 Method

We aim to generate geometrically plausible and temporally coherent relighting results from a reference image or video. Given a reference sequence $I_{1:N}^{\text{ref}}$ and a stream of target lighting conditions $\ell_{1:N}$, the

(a) Lightweight Multi-Plane Lighting Interactive Control



(b) Real-Time Distillation with Geometric Constraints



(c) Rolling-Window Streaming Video Relighting

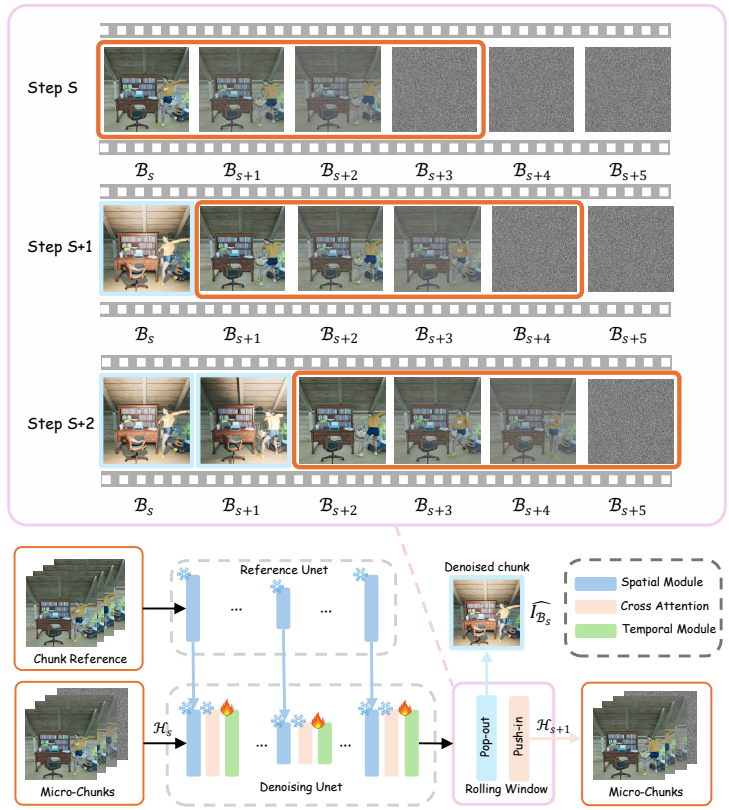


Fig. 3. **Overview of LiveLight.** (a) We first perform lightweight multi-plane image relighting by constructing a multi-plane irradiance representation from the target light and estimated depth. The resulting MPLI condition is encoded by the light encoder and injected into the diffusion relighting backbone together with the reference image. (b) To improve structure-preserving relighting, we introduce a geometry-guided feedback stage, where intermediate relighting predictions are evaluated by a geometry estimator and supervised with reconstruction and geometry-constrained losses. (c) For video relighting, we train the temporal module using a rolling-window strategy. At each streaming step, the oldest denoised chunk is emitted, the remaining latent chunks are retained as temporal history, and a new noisy chunk is pushed into the rolling window. This enables temporal consistency while maintaining bounded latency.

objective is to synthesize a relit sequence $\hat{I}_{1:N}$ that follows the desired illumination while preserving the subject identity, texture, and geometry. For image relighting, $N = 1$; for video relighting, frames are generated in a streaming and low-latency manner. Formally, each relit frame is generated by combining reference appearance features and structured lighting features:

$$\hat{I}_i = \mathcal{D}_\theta \left(Z_t \mid R(I_i^{\text{ref}}), G_L(\Phi(\ell_i, d_i)) \right), \quad i = 1, \dots, N, \quad (1)$$

where $\mathcal{D}_\theta$ is the denoising backbone, Z_t denotes the noisy latent at denoising step t , R is the reference appearance extractor, G_L is the light encoder, $\Phi(\ell_i, d_i)$ denotes the Multi-Plane Light Irradiance [Bian et al. 2025] condition constructed from the target lighting ℓ_i and the estimated subject depth d_i , N denotes total frame number.

As shown in Fig. 3, our method consists of three key components. First, we introduce Lightweight Multi-Plane Lighting Control, which provides an efficient and spatially aware representation for interactive illumination editing (Sec. 3.2). Second, we perform

Real-Time Distillation with Geometric Constraints to accelerate inference while improving geometry-consistent relighting (Sec. 3.3). Third, we develop Rolling-Window Streaming Video Relighting to enable low-latency video generation with per-frame reference refresh (Sec. 3.4).

3.1 Relighting Data Synthesis

High-quality paired relighting data is difficult to capture in real-world settings, as it requires controllable light sources, calibrated cameras, and repeated recordings of the same subject under different illumination conditions. We therefore construct a large-scale synthetic relighting dataset using diverse 3D assets, motion and camera trajectories, and physically based rendering. Our key design is to decouple illumination from identity, pose, expression, camera motion, and scene content, enabling the model to learn lighting-dependent appearance changes rather than fixed subject-light correlations.

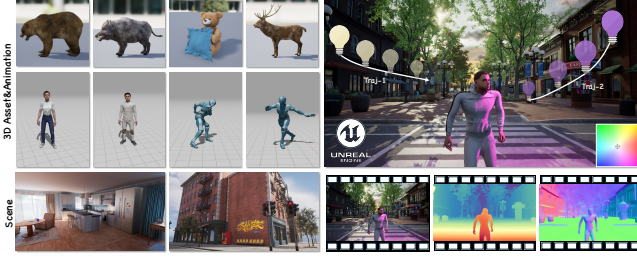


Fig. 4. **Overview of our synthetic relighting data synthesis pipeline.** We randomly combine diverse 3D assets (humans, animals, and objects) with animations and indoor/outdoor scenes, and render them in Unreal Engine 5 under controllable point lights with varying positions, intensities, and colors. For each frame, the renderer exports the RGB image along with depth map and surface normal map for geometry-aware supervision.

As shown in Fig. 4, we randomly combine 120 diverse 3D assets, 33 motion/camera trajectories, and 20 scenes to render paired relighting videos. For each sampled motion sequence, we render a canonical reference video under neutral illumination and 10 target videos under different controllable lights. The reference and target videos share the same identity, pose, expression, camera path, and scene content, and differ only in illumination. Each frame is exported with a depth map and a surface normal map, which are used for MPLI construction and geometry-aware supervision. After automatic visual-quality screening and manual verification, the final dataset contains approximately 250K relighting video sequences. Target lighting conditions are sampled from a controllable parametric space, including image-space position, relative depth, intensity, and RGB color. We cover both colored and white illumination with an approximate ratio of 6:4. The light intensity is randomly sampled within $[0.5, 1.5]$ in normalized UE units. To avoid degenerate near-black colored lights, we sample colored illumination from a high-saturation and high-value HSV space before converting it to RGB. For image-level relighting, individual reference-target frame pairs are sampled from these sequences. For video relighting, we generate temporally smooth lighting trajectories so that the reference and target videos provide consistent supervision over time. We further apply domain randomization over backgrounds, ambient illumination, material, camera exposure, and color temperature to improve robustness to real-world videos.

3.2 Lightweight Multi-Plane Lighting Interactive Control

A central challenge in video relighting is how to represent the target illumination in a compact, controllable, and spatially meaningful form. Instead of relying on a low-dimensional lighting code or an expensive dense illumination map, we adopt a Multi-Plane Light Irradiance (MPLI) [Bian et al. 2025] condition to provide explicit control over light position, relative depth, intensity, and color. Our contribution is a lightweight lighting-control branch that encodes this structured condition and injects it into the relighting backbone, enabling interactive illumination with low computational overhead. **MPLI-based lighting condition.** Given the target light ℓ_i and the estimated subject depth d_i , we construct an MPLI condition by projecting the user-specified light onto a small set of fronto-parallel

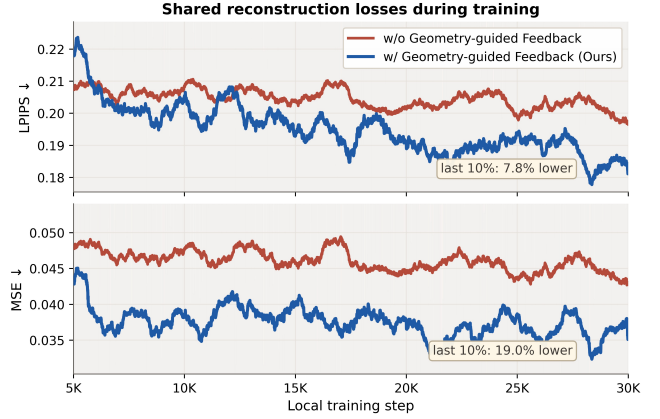


Fig. 5. **Geometry-guided feedback for structure-preserving relighting.** Compared with the baseline without geometry-guided feedback, our model maintains lower LPIPS and MSE, especially in the later training stage. Over the last 10% of steps, these metrics are 7.8% and 19.0% better, respectively. This indicates that depth-normal feedback provides an effective geometric regularization toward consistent relighting.

depth planes around the subject:

$$\Phi(\ell_i, d_i) = \text{Concat}_{p=1}^P \phi_p(\ell_i, d_i) \in \mathbb{R}^{H \times W \times 3P} \quad (2)$$

where $\phi_p(\ell_i, d_i)$ denotes the RGB irradiance map on the p -th plane. We use $P = 4$, producing a 12-channel lighting condition. This compact tensor preserves spatially varying illumination cues while allowing users to directly edit the light position, intensity, and color. **Lightweight light adapter.** Although the MPLI condition is compact, it still contains spatially aligned lighting information that must be adapted to the feature space of the relighting backbone. We therefore design a lightweight light encoder G_L , implemented with a small stack of convolutional blocks, to map the MPLI tensor into diffusion-compatible lighting features. The encoded light features are fused with the reference appearance features by injection:

$$\tilde{E}_i = E_\theta \left(R(I_i^{\text{ref}}) \right) + G_L \left(\Phi(\ell_i, d_i) \right), \quad (3)$$

where $E_\theta(\cdot)$ denotes the backbone encoder feature, $R(\cdot)$ extracts appearance feature, and $\tilde{E}_i$ is the resulting light-aware feature. Specifically, G_L is realized as a lightweight adapter with only $\sim 1\text{M}$ parameters. It transforms the MPLI condition into compact lighting tokens, which interact with the diffusion features through cross-attention in shallow layers. This localized injection strategy introduces spatially varying illumination cues at an early stage, while avoiding unnecessary perturbation to deeper denoising blocks that encode appearance priors, thereby preserving identity and texture fidelity.

This design separates appearance preservation from lighting control: the reference branch provides identity and texture cues, while the lightweight MPLI branch provides spatial illumination guidance. Since the lighting condition is explicitly parameterized by user-controllable attributes, changing the target light only requires updating $\Phi(\ell_i, d_i)$ and passing it through the same lightweight encoder, which enables interactive control over diverse lighting.

3.3 Real-Time Distillation with Geometric Constraints

The lightweight lighting branch provides efficient and controllable illumination guidance, but standard diffusion sampling still requires multiple denoising steps, which limits real-time interaction. We observe that video relighting contains substantial denoising redundancy: once conditioned on the reference appearance and MPLI lighting, the earliest denoising steps already determine the main light direction and global exposure, place shadows or highlights on the face, and establish depth-dependent falloff, while the remaining iterations mostly refine texture and high-frequency details. This motivates us to replace the redundant multi-step refinement process with a compact few-step sampling schedule.

Few-step relighting distillation. To make the compact schedule effective at inference, we train the relighting model directly with the same four-step rollout used at test time. Starting from noised latent, the denoising backbone follows a fixed few-step timestep schedule. For each training sample, we randomly choose a terminal update $m \in \{1, 2, 3, 4\}$, detach the preceding $m - 1$ updates that construct the intermediate latent, and back-propagate only through the final update. The predicted clean latent is decoded by the frozen VAE and supervised against the target relit image using reconstruction and geometry losses, which exposes the model to different positions in the compact trajectory while keeping memory cost low.

Geometry-guided supervision. Relighting should change illumination while preserving geometry. As shown in Fig. 5, incorporating depth and normal feedback regularizes the few-step model toward more stable, geometry-preserving relighting. Given the generated image $\hat{I}_i$, a frozen geometry estimator Ψ predicts depth and normal cues, which are compared with precomputed ground-truth geometry:

$$\mathcal{L}_{\text{geo}} = \|\Psi_D(\hat{I}_i) - d_i^*\|_1 + \lambda_N \|1 - \langle \Psi_N(\hat{I}_i), n_i^* \rangle\|_1. \quad (4)$$

Here d_i^* and n_i^* are pseudo ground-truth depth and normal, and Ψ_D, Ψ_N are predictions from the frozen estimator. Gradients are preserved with respect to $\hat{I}_i$, allowing geometry errors to update the relighting model. This branch is used only during training and adds no inference cost.

3.4 Rolling-Window Streaming Video Relighting

Frame-wise relighting can achieve low latency, but it lacks temporal context and often produces flickering artifacts. In contrast, full-clip video diffusion provides stronger temporal consistency but requires the entire sequence before generation, making it unsuitable for streaming applications. We therefore train the temporal module with a rolling-window strategy. The model only observes a short window of micro-chunks at each step, while previously processed latent chunks are propagated as the streaming history $\mathcal{H}$.

Let $\mathcal{B}_s$ denote the s -th micro-chunk and let K be the number of micro-chunks maintained in the rolling window. At streaming step s , the active history consists of K latent chunks. The video denoising backbone emits the oldest chunk, keeps the remaining chunks as temporal context, and appends a new noisy chunk to the tail:

$$(\hat{I}_{\mathcal{B}_s}, \mathcal{H}_{s+1}) = \mathcal{D}_\theta \left(\mathcal{H}_s \left| \left\{ R(I_i^{\text{ref}}), G_L(\Phi(\ell_i, d_i)) \right\}_{i \in \mathcal{W}_s} \right. \right), \quad (5)$$

Here $\mathcal{H}_s$ is the chunk-level streaming history for window $\mathcal{W}_s$. After each step, the oldest micro-chunk is decoded and emitted, the window slides forward by one chunk, and a new noisy chunk is inserted for future generation. Since each chunk is generated only once, the method avoids overlap-based re-generation and blending, while still allowing temporal information to flow through the retained latent context. Details are provided in Appendix ??.

The window length K controls the latency-consistency trade-off. A larger window provides longer temporal context and improves consistency, but increases memory usage and output latency. A smaller window reduces latency but weakens temporal modeling, which may lead to flickering or inconsistent lighting changes. We therefore use a moderate window size during video adaptation and keep the same rolling-window schedule for inference, reducing the train-test mismatch for streaming generation.

3.5 Training Objectives

We train the proposed framework in three stages, corresponding to the modules introduced in Sec. 3.2, Sec. 3.3, and Sec. 3.4.

Stage 1: Image-level relighting.

We first train the ReferenceNet-based image relighting model with the MPLI condition introduced in Sec. 3.2. Given a paired reference and target relit image, the target image is encoded into latent space, perturbed with Gaussian noise, and reconstructed by the denoising network conditioned on reference appearance and MPLI lighting. The objective is

$$\mathcal{L}_{\text{stage1}} = \mathbb{E}_{\tau, \epsilon} [\|z^0 - z^0\|_2^2]. \quad (6)$$

Stage 2: Geometry-plausible refinement.

In the second stage, we enable the geometry-guided feedback branch and refine the relighting model with geometry-aware supervision. The training objective is

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{geo}} \mathcal{L}_{\text{geo}}. \quad (7)$$

Here $\mathcal{L}_{\text{geo}}$ denotes the geometry-aware feedback loss defined in Sec. 3.3. This stage improves geometry-consistent shading without adding any inference-time overhead.

Stage 3: Streaming video adaptation.

Finally, we adapt the image model to streaming video relighting using the rolling-window strategy in Sec. 3.4. During training, we simulate the same rolling-window process as inference. Intermediate windows are advanced without gradient computation, and the emitted frames are supervised using RGB and perceptual losses:

$$\mathcal{L}_{\text{stage3}} = \frac{1}{|\mathcal{S}_s|} \sum_{i \in \mathcal{S}_s} \left[\lambda_{\text{mse}} \|\hat{I}_i - I_i^{\text{gt}}\|_2^2 + \lambda_{\text{lpips}} \mathcal{L}_{\text{lpips}}(\hat{I}_i, I_i^{\text{gt}}) \right], \quad (8)$$

where $\mathcal{S}_s$ denotes the supervised emitted frames.

4 Experiments

4.1 Application

Real-time lighting control. LiveLight enables real-time generation of controllable relighting videos from ordinary input videos. Benefiting from the proposed rolling-window strategy and per-frame reference refresh, our method can continuously update the lighting condition with low latency while maintaining temporal



Fig. 6. **Comparison of previous lighting control methods.** LiveLight outperforms existing methods in terms of controllability, temporal stability, and real-time efficiency.

coherence. As shown in Fig. 7, LiveLight follows the target lighting and preserves scene content, motion, and visual quality across diverse videos.

Diverse color control. LiveLight also supports real-time control over diverse lighting colors and positions. The proposed multi-plane light representation provides spatially meaningful lighting guidance, allowing users to specify colored lighting with different locations and intensities. Fig. 11 shows that our method can produce vivid colored lighting effects while maintaining coherent video appearance across frames. Furthermore, as shown in Fig. 8, LiveLight can simultaneously handle multiple colored light sources, producing spatially distinct relighting effects within a single frame.

User Interface. We design a user interface for LiveLight to allow users to interactively specify lighting conditions and visualize the relighting results in real time. As shown in Fig. 9, the interface provides intuitive controls for adjusting light color, intensity, and position, enabling users to create flexible and dynamic lighting

effects on their videos. The real-time feedback allows users to see the impact of their adjustments immediately, enhancing the creative process and making it easier to achieve the desired relighting results.

4.2 Implementation Details

In our implementation, we adopt a three-stage training pipeline to optimize the components of LiveLight. In Stage 1, we conduct image-level relighting training. Both the denoising U-Net and the ReferenceNet are initialized from pretrained Stable Diffusion weights [Romach et al. 2022]. This stage is trained for 50K iterations with a batch size of 32, and all model parameters are updated. In Stage 2, we enable the geometry-guided feedback branch and fine-tune the relighting model with geometry-aware supervision. This stage is trained for 20K iterations with a batch size of 16, keeping the geometry estimator frozen. In Stage 3, we insert temporal attention layers initialized from AnimateDiff [Guo et al. 2023] motion module

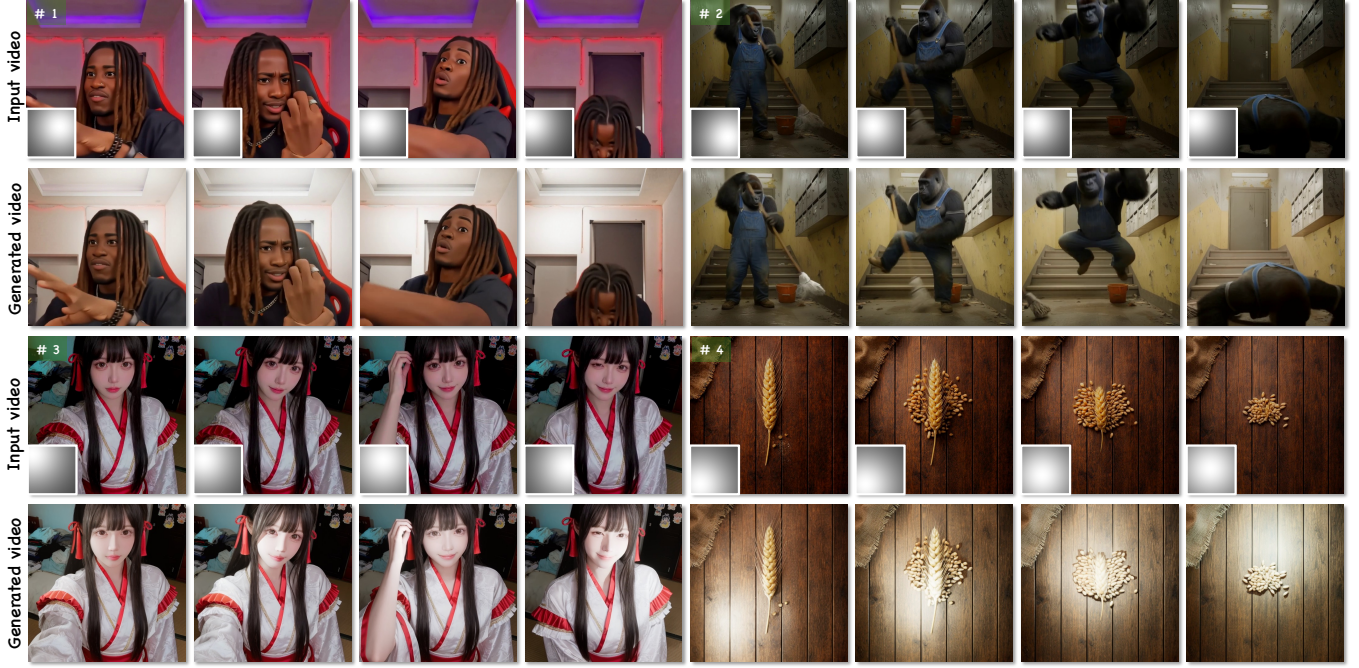


Fig. 7. **Visual results of real-time lighting control with LiveLight.** LiveLight enables high-quality relighting video generation under target lighting while preserving the content and motion of the input video.



Fig. 8. **Visual results of multiple color control.** LiveLight supports simultaneous control of multiple colored light sources, producing spatially distinct and geometrically plausible relighting effects.

weights and train the streaming video adaptation with the rolling-window strategy. The chunk size is set to $M = 4$. This stage is trained for 30K iterations with a batch size of 8. All training is conducted on 8 NVIDIA H20 GPUs. We use AdamW with weight decay 0.01. The learning rate is 1×10^{-5} for Stages 1 and 2, and 5×10^{-6} for Stage 3. Our training data is a self-constructed large-scale dataset of UE5-rendered video sequences with diverse lighting conditions, covering various identities, poses, and lighting trajectories. More details can be found in the supplementary materials.

4.3 Comparison with baselines

Qualitative comparison. Fig. 7 and Fig. 11 present qualitative results of LiveLight on diverse Internet videos. Compared with image-based relighting pipelines, LiveLight avoids obvious frame-wise flickering and produces more stable lighting changes over time. Compared with offline video relighting methods, LiveLight supports frame-by-frame control of light position, color, and intensity in a

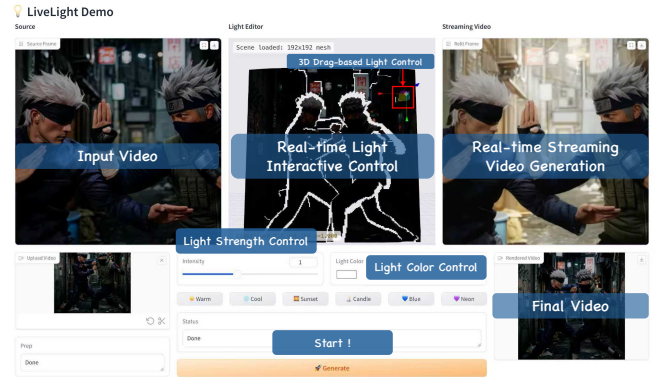


Fig. 9. **Illustration of LiveLight's user interface.** The interface allows users to interactively specify lighting conditions and visualize the relighting results in real time. Users can adjust light color, intensity, and position through intuitive controls, enabling flexible and dynamic lighting effects on the input video.

streaming manner. Benefiting from the proposed MPLI representation and rolling-window strategy, LiveLight follows the target lighting condition while preserving subject appearance, scene structure, and motion details. More visual comparisons are provided in the supplementary material.

Quantitative comparison. We compare LiveLight with representative image-based and video-based relighting baselines, including

Table 1. **Quantitative Comparisons with Other Methods.** Red and Blue denote the best and second best results, respectively.

Method	Quality Metrics					VBench Metrics [Huang et al. 2024]			
	Aesthetic $\uparrow$	FVD $\downarrow$	PSNR _y $\uparrow$	PSNR _{light} $\uparrow$	Time (s) $\downarrow$	Sub. Cons. $\uparrow$	Back. Cons. $\uparrow$	Motion Smooth. $\uparrow$	Temporal Cons. $\uparrow$
AnyV2V [Zhang et al. 2025]	0.4929	1082.6	11.063	15.923	89	0.7980	0.8970	0.9707	0.9656
LightCtrl [Peng et al. 2026]	0.5307	987.3	11.782	18.486	122	0.8696	0.9070	0.9728	0.9559
Light-A-Video [Zhou et al. 2025]	0.5533	1053.8	11.247	17.612	115	0.8678	0.9244	0.9782	0.9740
TC-Light [Liu et al. 2026]	0.5170	862.4	12.136	19.274	38	0.9402	0.9502	0.9642	0.9799
LiveLight (Ours)	0.5748	738.5	12.847	21.538	1.01	0.9561	0.9687	0.9856	0.9872

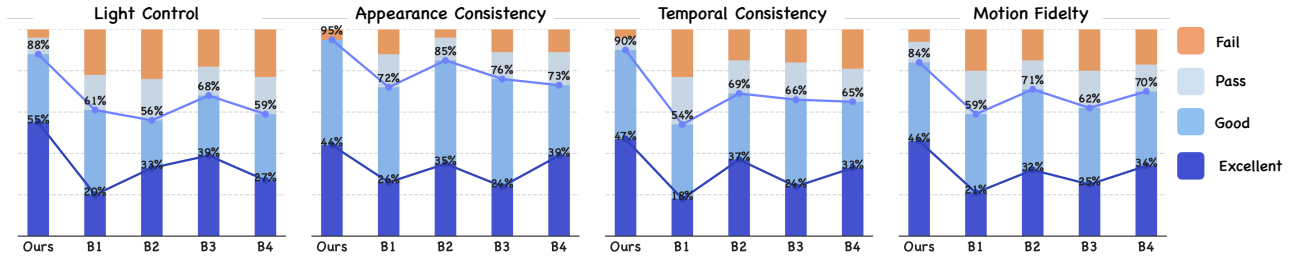


Fig. 10. **User study.** We conduct a user study to evaluate the perceptual quality of relighting videos generated by LiveLight and baseline methods. Participants are asked to rate each result in terms of light control, appearance consistency, temporal consistency, and motion fidelity. LiveLight receives the highest preference across all criteria, demonstrating its ability to produce controllable, visually consistent, and temporally coherent relighting videos.

Table 2. **Quantitative Comparisons with Synthetic data.** Red and Blue denote the best and second best results, respectively.

Method	Synthetic Data Metrics		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
AnyV2V [Zhang et al. 2025]	31.482	0.8936	0.2693
LightCtrl [Peng et al. 2026]	35.617	0.9254	0.2148
Light-A-Video [Zhou et al. 2025]	37.385	0.9418	0.1726
TC-Light [Liu et al. 2026]	39.726	0.9571	0.1387
LiveLight (Ours)	43.264	0.9778	0.1064

AnyV2V [Zhang et al. 2025], LightCtrl [Peng et al. 2026], Light-A-Video [Zhou et al. 2025], and TC-Light [Liu et al. 2026]. Following the standard setting of previous video relighting methods, we collect 200 lighting-control videos from the Internet, and evaluate all methods on the same 16-frame clips. For AnyV2V [Zhang et al. 2025], we edit the first frame using IC-Light [Zhang et al. 2025] and feed it into the model. Following prior video relighting work [Peng et al. 2026], we first use five task-level metrics. (1) **Aesthetic**: we use an aesthetic predictor to measure the overall visual appeal of generated videos. (2) **FVD**: we compute the distribution distance between generated videos and input videos to evaluate video realism. (3) **PSNR_y**: we compute PSNR on the luminance channel within the light-map mask region against a pure-white reference video, which measures whether the specified region receives stronger illumination. (4) **PSNR_{light}**: we directly overlay the target light map on the input video and compute PSNR between this overlay reference and the generated result, which evaluates whether the relighting follows

the light trajectory while preserving the foreground appearance. (5) **Time**: we report the inference cost of each method.

We further adopt four VBench metrics [Huang et al. 2024] for a broader assessment of video quality. **Subject Consistency** evaluates whether the foreground subject is preserved across frames. **Background Consistency** measures the coherence of static scene regions. **Motion Smoothness** evaluates inter-frame motion continuity, and **Temporal Consistency** measures overall frame-to-frame coherence. As shown in Table 1, LiveLight achieves the best overall performance across both relighting-specific metrics and general video-quality metrics, demonstrating superior controllability, temporal stability, and real-time efficiency.

Longer video evaluation. We further evaluate long-video relighting on another 200 Internet videos. All methods are tested under two longer settings, 256 frames and 512 frames. We use the same four VBench metrics as in the main quantitative comparison. As reported in Table 3, LiveLight consistently outperforms competing methods under both video lengths, showing that our streaming design better preserves stable relighting over extended temporal horizons.

Synthetic data evaluation. To evaluate the relighting quality under controlled conditions, we also test all methods on synthetic videos with known ground-truth lighting changes. We use PSNR, SSIM, and LPIPS to measure the fidelity of relighting results against the ground truth. We synthesized 100 videos in UE5 with various lighting trajectories and 10 different scene to comprehensively assess the performance of each method. Notably, all evaluation scenes and lighting conditions are unseen during training. As shown in Table 2, LiveLight achieves the best performance across all metrics, demonstrating its ability to accurately follow target lighting conditions while preserving visual quality.

Table 3. **Quantitative comparison on longer videos with VBench [Huang et al. 2024].** $\uparrow$ indicates that higher is better. **Red** and **Blue** denote the best and second best results, respectively.

Method	256 frames				512 frames			
	Sub. Cons. $\uparrow$	Back. Cons. $\uparrow$	Aes. Qual. $\uparrow$	Motion Smooth. $\uparrow$	Sub. Cons. $\uparrow$	Back. Cons. $\uparrow$	Aes. Qual. $\uparrow$	Motion Smooth. $\uparrow$
AnyV2V [Zhang et al. 2025]	0.7728	0.8706	0.4671	0.9534	0.7548	0.8546	0.4481	0.9364
LightCtrl [Peng et al. 2026]	0.8412	0.8823	0.5057	0.9563	0.8262	0.8633	0.4887	0.9403
Light-A-Video [Zhou et al. 2025]	0.8386	0.8978	0.5263	0.9617	0.8216	0.8828	0.5083	0.9477
TC-Light [Liu et al. 2026]	0.9128	0.9247	0.4918	0.9472	0.8972	0.9067	0.4758	0.9283
LiveLight (Ours)	0.9456	0.9542	0.5592	0.9812	0.9396	0.9492	0.5552	0.9782

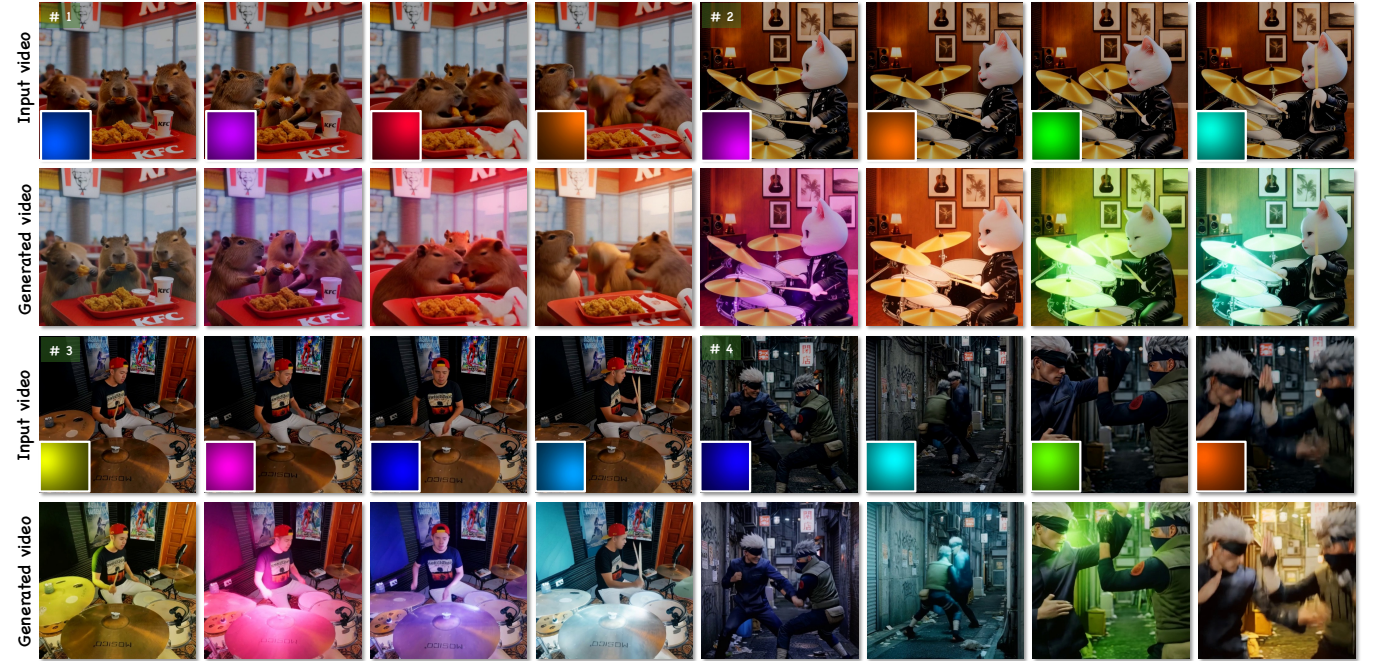


Fig. 11. **Visual results of diverse color lighting control with LiveLight.** LiveLight supports real-time lighting control with different colors and positions, producing temporally coherent relighting results across dynamic videos.

User study. To account for the limitations of automatic metrics in capturing real-world preferences, we conduct a user study with 20 volunteers on 50 randomly selected test cases. The volunteers categorize each result into one of four levels: excellent, good, pass, or fail, based on lighting accuracy, appearance consistency, temporal coherence, and motion fidelity. The results are presented in Fig. 10. LiveLight outperforms competing methods in both automated metrics and human subjective preferences.

4.4 Ablation Study

To validate each design choice in LiveLight, we systematically ablate key components spanning all three training stages. Table 4 reports the results of replacing or removing individual modules, and Table 5 compares different numerical hyperparameter settings. All variants are evaluated on the same synthetic test set with identical inference settings unless otherwise stated.

Effectiveness of VAE decoder. Our framework is compatible with lightweight VAE decoders. As shown in Fig. 16 and Table 4, replacing the standard SD VAE with Tiny VAE [madebyollin 2023] increases the throughput from 15.78 fps to 18.34 fps with only a marginal quality drop, demonstrating that the VAE decoder is not the quality bottleneck and can be swapped for faster alternatives when latency is prioritized. In all other experiments, we use the standard SD VAE for fair comparison with baselines.

Effectiveness of light injection strategy. We compare three strategies for injecting MPLI lighting conditions in Fig. 18: channel concatenation (18.21 fps), our lightweight adapter (15.78 fps), and ControlNet (10.13 fps). Channel concatenation modifies the input channels and requires retraining the backbone from scratch, disrupting the pretrained generative prior and degrading output quality. ControlNet preserves the prior but nearly halves the throughput due to its duplicate encoder. Our adapter retains the full pretrained

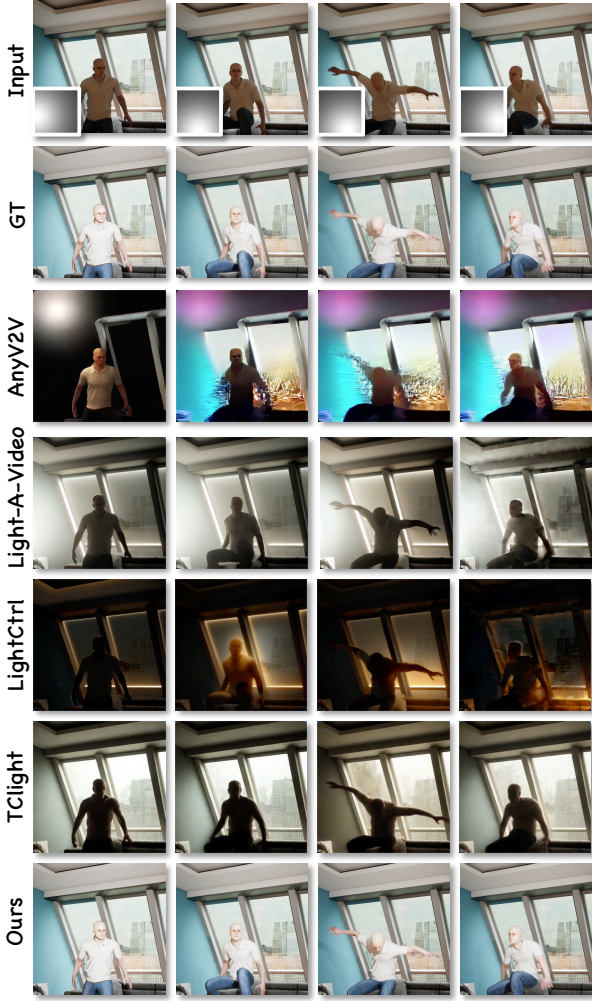


Fig. 12. **Synthetic data evaluation.** We evaluate the performance of LiveLight on synthetic data to assess its ability to handle controlled lighting scenarios. LiveLight demonstrates superior performance in maintaining appearance consistency and temporal coherence under various lighting conditions.

capability with only a modest speed cost, achieving the best quality-efficiency trade-off.

Effectiveness of Number of MPLI planes. We vary the number of depth planes $P \in \{2, 4, 8\}$ in Fig. 14. With only 2 planes, the model struggles to capture depth-dependent lighting variation, limiting 3D lighting controllability. Increasing to 8 planes yields marginal improvement over 4 planes while adding extra parameters. We adopt $P=4$ as the default, balancing 3D control capability and model compactness.

Effectiveness of geometry constraints. We ablate the geometry-aware feedback branch in Fig. 19 by removing depth supervision (W/o Depth) or normal supervision (W/o Normal). Removing either component leads to less accurate highlights and shadows that are inconsistent with the underlying surface geometry. The full feedback

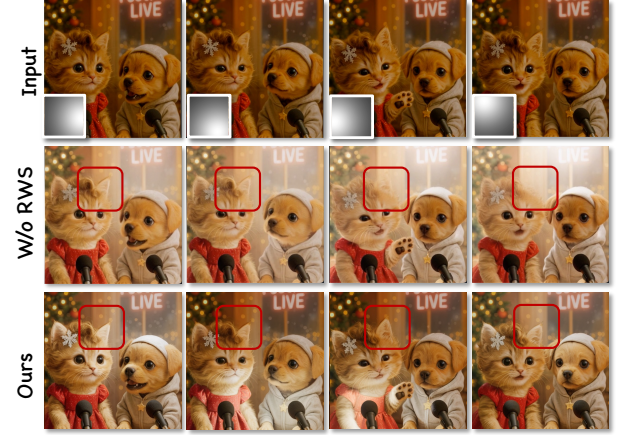


Fig. 13. **Visual ablation study of the rolling-window strategy.** Removing the rolling-window strategy leads to temporal flickering and unstable lighting across chunks. By propagating through the rolling window, our method maintains coherent illumination and stable appearance over time.

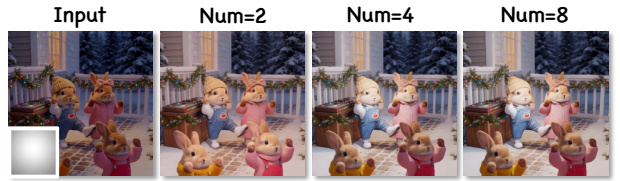


Fig. 14. **Visual ablation study of the number of MPLI planes.** With $P=2$, the model cannot capture depth-dependent lighting, limiting 3D lighting controllability. $P=8$ yields marginal gains over $P=4$ with added parameters. We adopt $P=4$ as the default.

configuration produces the most geometrically plausible relighting results.

Effectiveness of distillation steps. We compare 2, 4, and 8 denoising steps in Fig. 17. Aggressive distillation to 2 steps runs at 23.17 fps but introduces visible artifacts in details. Using 8 steps improves quality marginally over 4 steps while reducing throughput to 9.43 fps. We adopt 4 steps (15.78 fps) as the default, achieving a favorable balance between fidelity and real-time performance.

Effectiveness of micro-chunk size. We study the effect of chunk size $M \in \{2, 4, 8\}$ on temporal quality and inference speed. As shown in Fig. 15 and Table 5, a smaller chunk size $M=2$ yields the fastest throughput (16.61 fps) but suffers from noticeable temporal instability due to insufficient inter-frame context within each chunk. Conversely, a larger chunk size $M=8$ provides stronger temporal coherence but significantly increases computation, reducing throughput to 11.17 fps. We adopt $M=4$ (15.78 fps) as the default, which achieves a favorable trade-off between temporal stability and real-time performance.

Effectiveness of rolling-window strategy(RWS). We ablate the rolling-window training strategy, which simulates the rolling-window inference process during training. As shown in Fig. 13 and Table 4, removing RWS leads to noticeable appearance drift

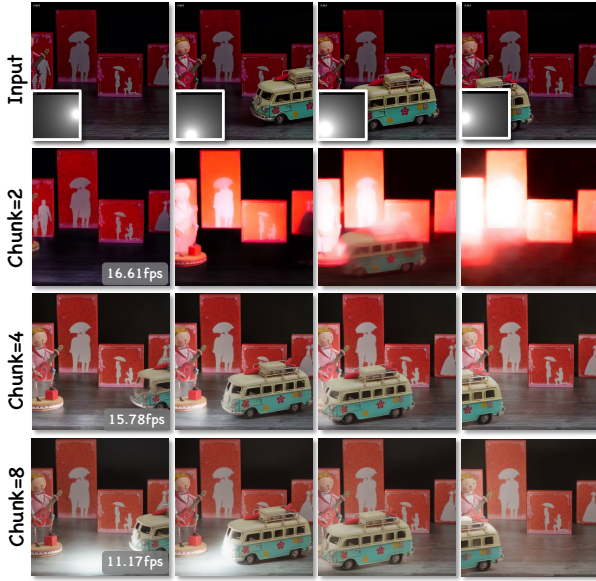


Fig. 15. **Visual ablation study of micro-chunk size.** $M=2$ (16.61 fps) causes severe temporal instability. $M=8$ (11.17 fps) improves coherence but reduces throughput. We adopt $M=4$ (15.78 fps) for the best latency–quality balance.

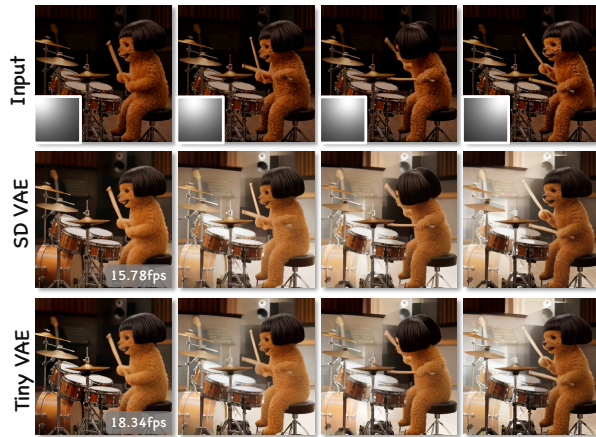


Fig. 16. **Visual ablation study of VAE decoder.** The Tiny VAE (18.34 fps) produces visual quality comparable to the SD VAE (15.78 fps) while offering faster decoding, making it well suited for real-time streaming.

and temporal inconsistency over long sequences, as the model has never been exposed to the error accumulation pattern inherent in streaming inference. With RWS, the model learns to maintain stable appearance and background details across frames, effectively reducing error accumulation and improving temporal coherence.

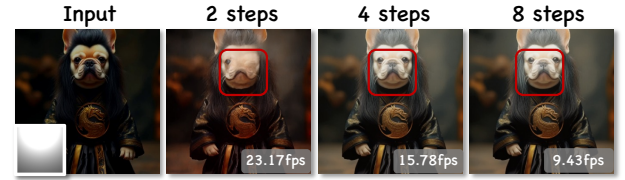


Fig. 17. **Visual ablation study of distillation denoising steps.** Distilling to 2 steps (23.17 fps) introduces visible facial artifacts. Using 8 steps (9.43 fps) yields marginal quality gains over 4 steps (15.78 fps). We adopt 4 steps as the default for optimal quality–speed balance.



Fig. 18. **Visual ablation study of light injection strategy.** Channel concatenation (18.21 fps) disrupts the pretrained prior, degrading quality. ControlNet (10.13 fps) preserves the prior but nearly halves throughput. Our adapter (15.78 fps) achieves the best quality–efficiency trade-off.



Fig. 19. **Visual ablation study of the geometry-guided feedback branch.** Removing depth (W/o Depth) or normal (W/o Normal) supervision produces highlights and shadows inconsistent with the surface geometry. The full feedback yields the most geometrically plausible relighting.

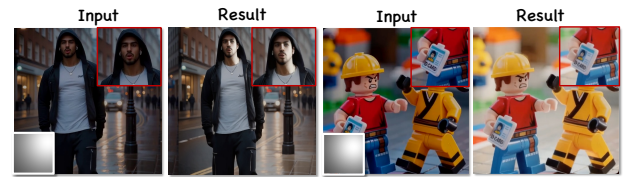


Fig. 20. **Failure cases.** Left: a distant face loses fine facial details after relighting. Right: small text (“ID CARD”) and accessories on Lego figures exhibit visible artifacts. These failures are primarily attributed to the limited reconstruction capacity of the SD VAE decoder for high-frequency details.

5 Limitation

Despite the promising results, LiveLight has several limitations, as illustrated in Fig. 20. First, our method struggles with small or distant faces, where the limited spatial resolution leads to degraded relighting quality and loss of fine facial details. Second, we observe

Table 4. **Ablation study on component design choices.** Each row moves or replaces one component from the full model. **Red** and **Blue** denote the best and second best results, respectively.

Variant	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	Time (s) $\downarrow$
<i>VAE Decoder</i>				
Tiny VAE	42.917	0.9762	0.1098	0.87
<i>Light Injection</i>				
Channel concat	39.843	0.9524	0.1536	0.88
ControlNet	43.108	0.9763	0.1082	1.58
<i>Geometry-guided Feedback</i>				
W/o Depth Feedback	41.726	0.9681	0.1274	1.01
W/o Normal Feedback	42.183	0.9714	0.1193	1.01
<i>Rolling window streaming Training</i>				
W/o RWS	41.392	0.9653	0.1342	1.01
LiveLight (Ours)	43.264	0.9778	0.1064	1.01

Table 5. **Ablation study on numerical hyperparameters.** **Red** and **Blue** denote the best and second best results, respectively.

Variant	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	Time (s) $\downarrow$
<i>MPLI Planes (P)</i>				
$P = 2$	40.538	0.9586	0.1387	1.01
$P = 8$	43.198	0.9771	0.1072	1.06
<i>Distillation Steps</i>				
Step = 2	39.274	0.9483	0.1592	0.69
Step = 8	43.182	0.9768	0.1076	1.70
<i>Chunk Size (M)</i>				
$M = 2$	40.847	0.9612	0.1356	0.96
$M = 8$	43.156	0.9765	0.1079	1.43
LiveLight ($P=4$, Step=4, $M=4$)	43.264	0.9778	0.1064	1.01

occasional artifacts in high-frequency details such as text and small accessories in the generated frames. These two issues are primarily attributed to the decoder capacity of the Stable Diffusion VAE [Romach et al. 2022], which was not designed to faithfully reconstruct such fine-grained details. We expect that migrating to a stronger base model with a more capable decoder can alleviate this issue. Finally, the geometry-guided feedback branch relies on the accuracy of off-the-shelf depth and normal estimators. When the geometry estimation is inaccurate—particularly for complex hairstyles, occlusions, or extreme poses—the shading guidance may be suboptimal, leading to less physically plausible relighting.

6 Conclusion

We present LiveLight, the first diffusion-based framework for real-time, streaming video relighting with interactive 3D lighting control. Unlike existing offline methods that require the full lighting trajectory before generation, LiveLight allows users to dynamically

adjust light position, intensity, and color while the relit video is produced frame by frame with low latency. Our framework combines a lightweight MPLI adapter for spatially precise and depth-aware illumination control, a geometry-guided feedback branch for geometry-consistent shading, and a progressive micro-chunk streaming strategy for temporally coherent, arbitrarily long video generation. Extensive experiments demonstrate that LiveLight achieves state-of-the-art quality while running at real-time speed, outperforming offline baselines in temporal stability, controllability, and user preference. We hope LiveLight inspires future research on real-time controllable video generation.

7 Potential Societal Impacts

LiveLight enables real-time, interactive relighting of portrait videos, which can benefit creative professionals in film production, virtual cinematography, and live-streaming content creation by significantly reducing the cost and time of lighting adjustments. However, as with other generative video techniques, the ability to realistically alter lighting in portrait videos could potentially be misused to create misleading or deceptive visual content. We encourage the research community to develop robust detection and watermarking mechanisms alongside generative methods, and to establish responsible usage guidelines. Our work focuses on lighting manipulation rather than identity or content generation, which inherently limits its misuse potential compared to general-purpose video synthesis methods.

References

- 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).
- Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023).
- Ziqi Cai, Kaiwen Jiang, Shu-Yu Chen, Yu-Kun Lai, Hongbo Fu, Boxin Shi, and Lin Gao. 2024. Real-time 3d-aware portrait video relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6221–6231.
- Sumit Chaturvedi, Mengwei Ren, Yannick Hold-Geoffroy, Jingyuan Liu, Julie Dorsey, and Zhixin Shu. 2025. SynthLight: Portrait Relighting with Diffusion Model by Learning to Re-render Synthetic Faces. *arXiv preprint arXiv:2501.09756* (2025).
- Boyuan Chen, Diego Marti Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. 2024. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems* 37 (2024), 24081–24125.
- Paul Debevec. 2008. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *Acm siggraph 2008 classes*. 1–10.
- Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. 2000. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. 145–156.
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.
- Ye Fang, Zeyi Sun, Shangzhan Zhang, Tong Wu, Yinghao Xu, Pan Zhang, Jiaqi Wang, Gordon Wetzstein, and Dahua Lin. 2025. Relightvid: Temporal-consistent diffusion model for video relighting. *arXiv preprint arXiv:2501.16330* (2025).
- Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. 2024. One step diffusion via shortcut models. *arXiv preprint arXiv:2410.12557* (2024).
- Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725* (2023).
- Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. 2024. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101* (2024).

- Kai He, Ruofan Liang, Jacob Munkberg, Jon Hasselgren, Nandita Vijaykumar, Alexander Keller, Sanja Fidler, Igor Gilitschenski, Zan Gojic, and Zian Wang. 2026. Unirelight: Learning joint decomposition and synthesis for video relighting. *Advances in Neural Information Processing Systems* 38 (2026), 108435–108459.
- Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. 2026. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems* 38 (2026), 167283–167308.
- Ziqi Huang, Yanan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21807–21818.
- 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.
- Hoon Kim, Minje Jang, Wonjun Yoon, Jisoo Lee, Donghyun Na, and Sanghyun Woo. 2024. SwitchLight: Co-design of Physics-driven Architecture and Pre-training Framework for Human Portrait Relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 25096–25106.
- Peter Kocsis, Vincent Sitzmann, and Matthias Nießner. 2024. Intrinsic image diffusion for indoor single-view material estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5198–5208.
- Akio Kodaira, Chenfeng Xu, Toshiaki Hazama, Takanori Yoshimoto, Kohei Ohno, Shogo Mitsuohori, Soichi Sugano, Hanying Cho, Zhijian Liu, Masayoshi Tomizuka, et al. 2025. Streamdiffusion: A pipeline-level solution for real-time interactive generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12371–12380.
- Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. 2024. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603* (2024).
- Feng Liang, Akio Kodaira, Chenfeng Xu, Masayoshi Tomizuka, Kurt Keutzer, and Diana Marculescu. 2025b. Looking backward: Streaming video-to-video translation with feature banks. In *International Conference on Learning Representations*, Vol. 2025. 46425–46445.
- Ruofan Liang, Zan Gojic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Chih-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, et al. 2025a. Diffusion renderer: Neural inverse and forward rendering with video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 26069–26080.
- Yuanze Lin, Yi-Wen Chen, Yi-Hsuan Tsai, Ronald Clark, and Ming-Hsuan Yang. 2026. Illumicraft: Unified geometry and illumination diffusion for controllable video generation. *Advances in Neural Information Processing Systems* 38 (2026), 27798–27829.
- Xingchao Liu, Chengyue Gong, and Qiang Liu. 2022. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003* (2022).
- Yang Liu, Chuanchen Luo, Zimo Tang, Yingyan Li, Yuanyong Ning, Lue Fan, Junran Peng, ZHAO-XIANG ZHANG, et al. 2026. Tc-light: Temporally coherent generative rendering for realistic world transfer. *Advances in Neural Information Processing Systems* 38 (2026), 51463–51489.
- Yue Ma, Kunyu Feng, Zhongyuan Hu, Xinyu Wang, Yucheng Wang, Mingzhe Zheng, Bingyuan Wang, Qinghe Wang, Xuanhua He, Hongfa Wang, et al. 2025. Controllable video generation: A survey. *arXiv preprint arXiv:2507.16869* (2025).
- Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. 2024. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 4117–4125.
- madebyollin. 2023. TAESD. <https://huggingface.co/madebyollin/taesd>. Hugging Face model repository.
- Yiqun Mei, Mingming He, Li Ma, Julien Philip, Wenqi Xian, David M George, Xueming Yu, Gabriel Dedic, Ahmet Levent Tasel, Ning Yu, et al. 2025. Lux post facto: Learning portrait performance relighting with conditional video diffusion and a hybrid dataset. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 5510–5522.
- Rohit Pandey, Sergio Orts-Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul E Debevec, and Sean Ryan Fanello. 2021. Total relighting: learning to relight portraits for background replacement. *ACM Trans. Graph.* 40, 4 (2021), 43–1.
- Yizuo Peng, Xuelin Chen, Kai Zhang, and Xiaodong Cun. 2026. LightCtrl: Training-free Controllable Video Relighting. *arXiv preprint arXiv:2603.27083* (2026).
- Mengwei Ren, Wei Xiong, Jae Shin Yoon, Zhixin Shu, Jianming Zhang, HyunJoon Jung, Guido Gerig, and He Zhang. 2024. Relightful harmonization: Lighting-aware portrait background replacement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6452–6462.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
- Tim Salimans and Jonathan Ho. 2022. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512* (2022).
- Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. 2024. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*. Springer, 244–260.
- Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Fei Wu Yu, Haiming Zhao, Jianxiao Yang, et al. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314* (2025).
- Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. 2023. Videolcm: Video latent consistency model. *arXiv preprint arXiv:2312.09109* (2023).
- Jinbo Xing, Long Mai, Cusuh Ham, Jiahui Huang, Aniruddha Mahapatra, Chi-Wing Fu, Tien-Tsin Wong, and Feng Liu. 2025. Motioncanvas: Cinematic shot design with controllable image-to-video generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. 1–11.
- Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. 2024. Make-your-video: Customized video generation using textual and structural guidance. *IEEE transactions on visualization and computer graphics* 31, 2 (2024), 1526–1541.
- Yuxuan Xue, Ruofan Liang, Egor Zakharov, Timur Bagautdinov, Chen Cao, Giljoo Nam, Shunsuke Saito, Gerard Pons-Moll, and Javier Romero. 2026. GeoRelight: Learning Joint Geometrical Relighting and Reconstruction with Flexible Multi-Modal Diffusion Transformers. *arXiv preprint arXiv:2604.20715* (2026).
- Zhuoyi Yang, Jiayan Tang, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. *arXiv preprint arXiv:2408.06072* (2024).
- Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. 2024. DiLightNet: Fine-grained lighting control for diffusion-based image generation. In *ACM SIGGRAPH 2024 Conference Papers*. 1–12.
- Jianshu Zeng, Yuxuan Liu, Yutong Feng, Chenxuan Miao, Zixiang Gao, Jiawang Qu, Jianzhang Zhang, Bin Wang, and Kun Yuan. 2025. Lumen: Consistent video relighting and harmonious background replacement with video generative models. *arXiv preprint arXiv:2508.12945* (2025).
- Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. 2021a. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5453–5462.
- 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*.
- Longwen Zhang, Qixuan Zhang, Minye Wu, Jingyi Yu, and Lan Xu. 2021b. Neural video portrait relighting in real-time via consistency modeling. In *Proceedings of the IEEE/CVF international conference on computer vision*. 802–812.
- Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. 2022. Modeling indirect illumination for inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18643–18652.
- Yabo Zhang, Yuxiang Wei, XIAOPENG ZHANG, Wangmeng Zuo, Qi Tian, et al. 2024a. Controlvideo: Training-free controllable text-to-video generation. In *International Conference on Learning Representations*, Vol. 2024. 54441–54461.
- Yuxin Zhang, Dandan Zheng, Biao Gong, Jingdong Chen, Ming Yang, Weiming Dong, and Changsheng Xu. 2024b. Lumisculpt: A consistency lighting control network for video generation. (2024).
- Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. 2024. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*. Springer, 273–290.
- Jianbin Zhang, Minghui Hu, Zhongyi Fan, Chaoyue Wang, Changxing Ding, Dacheng Tao, and Tat-Jen Cham. 2024. Trajectory consistency distillation. *arXiv e-prints* (2024), arXiv–2402.
- 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. *arXiv preprint arXiv:2502.08590* (2025).