Light4D: Training-Free Extreme Viewpoint 4D Video Relighting

Anonymous authors
Paper under double-blind review

Abstract

Recent advances in diffusion-based generative models have established a new paradigm for image and video relighting. However, extending these capabilities to 4D relighting remains challenging, due primarily to the scarcity of paired 4D relighting training data and the difficulty of maintaining temporal consistency across extreme viewpoints. In this work, we propose ***Light4D***, a novel training-free framework designed to synthesize consistent 4D videos under target illumination, even under extreme viewpoint changes. First, we introduce Disentangled Flow Guidance, a time-aware strategy that effectively injects lighting control into the latent space while preserving geometric integrity. Second, to reinforce temporal consistency, we develop Temporal Consistent Attention within the IC-Light architecture and further incorporate deterministic regularization to eliminate appearance flickering. Extensive experiments demonstrate that our method achieves competitive performance in temporal consistency and lighting fidelity, robustly handling camera rotations from $-90°$ to $90°$. Website: `https://anonymous.4open.science/w/Paper_Video-F181/`.

1 Introduction

The synthesis of dynamic 4D content is fundamental for next-generation immersive applications, including cinematic virtual production (Bahmani et al., 2024), AR/VR (Pang et al., 2025), and interactive simulations (Wen et al., 2025). Since 4D geometric synthesis alone is insufficient for photorealism, achieving simultaneous control over both camera trajectory and illumination becomes a pivotal requirement for high-fidelity generation.

However, current research focuses predominantly on either video relighting or 4D geometric generation, leaving their intersection largely unexplored. Existing video relighting approaches (Zhang et al., 2025; Lin et al., 2025; Bharadwaj et al., 2025) operate primarily in the 2D domain. Although these methods improve the temporal consistency of videos with limited motion, they fundamentally struggle to maintain spatiotemporal consistency under complex camera trajectories. In contrast, camera-controlled 4D generation methods (Chen et al., 2025; Mi et al., 2025; Zhou et al., 2025a) excel in synthesizing coherent dynamic geometry but typically overlook controllable illumination. In these frameworks, lighting effects are embedded in the texture (Zhang et al., 2021b), making the illumination uneditable and preventing adaptation to novel environments.

To address these limitations, recent research has explored 4D relighting via end-to-end supervised training. Frameworks such as Light-X (Liu et al., 2025) aim to learn joint camera and illumination control directly. However, these methods face two critical bottlenecks. First, they are fundamentally constrained by the severe data scarcity. Training these models necessitates massive paired datasets featuring both multi-view and multi-illumination consistency, which are prohibitively expensive to acquire in real-world environments. Second, these approaches often struggle to generalize to extreme viewpoints. Due to their reliance on limited synthetic data, the generated illumination tends to appear rigid and flat, resembling 2D texture mapping rather than volumetric light transport (Chaturvedi et al., 2025), which severely compromises photorealism.

In this work, we propose **Light4D**, a training-free framework tailored for 4D video relighting under extreme viewpoint changes. As shown in Figure 1, our method achieves high-fidelity results under extreme viewpoint changes. Unlike supervised methods that rely on expensive retraining, our method leverages the generative

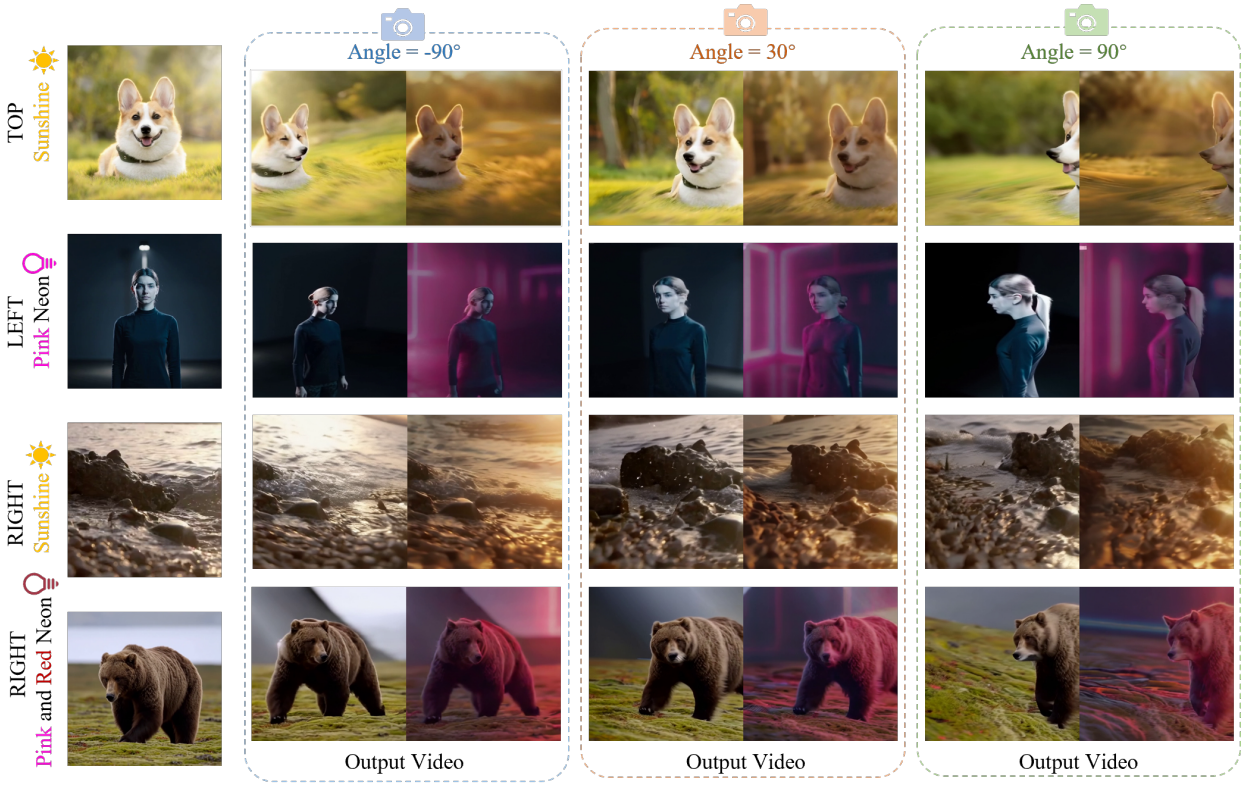

Figure 1: **Visual results of Light4D for training-free 4D video relighting.** Our framework robustly handles extreme viewpoint changes and diverse lighting conditions while maintaining strict geometric-illumination consistency.

power of pre-trained expert models, synergizing the geometric prior of EX-4D (Hu et al., 2025) with the illumination prior of IC-Light (Zhang et al., 2025). A critical challenge in this integration arises from the inherent conflict between geometric reconstruction and illumination synthesis within the generative manifold. To resolve this, we introduce Disentangled Flow Guidance (DFG), a time-aware strategy that harmonizes lighting injection with geometric preservation. Specifically, this mechanism establishes a robust geometric foundation during the initial denoising phase, subsequently utilizing a lighting-fused latent state as a rectified target to steer the simultaneous synthesis of geometry and illumination. Furthermore, to ensure cross-frame coherence, we develop Temporal Consistent Attention (TCA) within the IC-Light architecture and employ deterministic regularization to suppress stochastic fluctuations and ensure temporal consistency.

Our contributions are summarized as follows:

- We propose Light4D, the first training-free framework capable of achieving joint control over extreme camera trajectories ($-90° \sim 90°$) and illumination. By leveraging pre-trained generative priors, our approach eliminates the need for large-scale paired datasets.
- We introduce a Disentangled Flow Guidance strategy and Temporal Consistent Attention (TCA). These designs jointly resolve the inherent conflict between geometric reconstruction and illumination synthesis, ensuring strict temporal coherence and preventing structural degradation.
- Extensive experiments demonstrate that our method achieves competitive performance in terms of temporal consistency and lighting fidelity compared to existing video relighting baselines, particularly under extreme viewpoint changes.

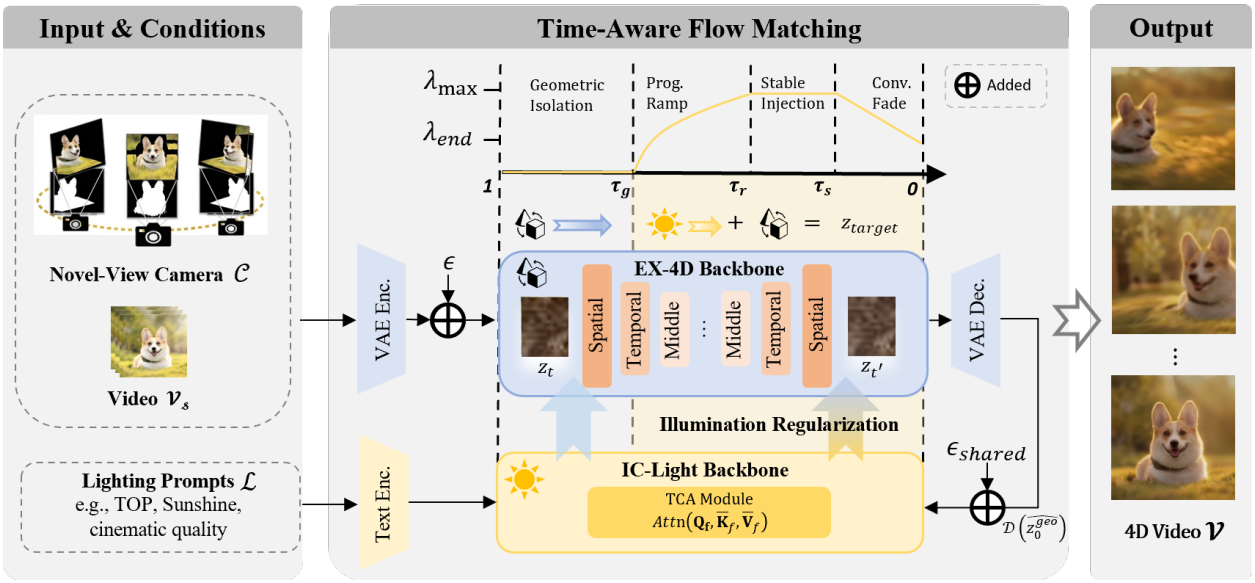

Figure 2: **Overview of the Light4D framework.** Our training-free approach employs a time-aware paradigm in a latent flow-matching process. Using a multi-phase adaptive schedule $\lambda(t)$, we prioritize 3D geometric completion via the EX-4D backbone before injecting illumination cues through IC-Light.

2 Related Work

Learning-based Illumination Editing. Early learning-based relighting methods typically rely on paired supervision, where encoder-decoder CNNs map an input image to a relit output. Extending relighting from images to videos often introduces temporal flicker, motivating explicit cross-frame consistency constraints (Zhang et al., 2021a; Chandran et al., 2022). In recent years, diffusion-based formulations have become a dominant direction (Ponglertnapakorn et al., 2023; Chaturvedi et al., 2025; Wang et al., 2025; Fang et al., 2025; Magar et al., 2025). Their performance further benefits from scaling choices such as larger training corpora, stronger backbones, and physically grounded objectives and constraints (Kocsis et al., 2024; Kim et al., 2024; Zhang et al., 2025). Beyond general-purpose relighting, specialized settings have also been studied, including portrait performance relighting with conditional video diffusion and hybrid datasets (Mei et al., 2025), as well as approaches that jointly model intrinsic decomposition and relighting synthesis to better capture complex illumination effects (He et al., 2025). Among recent methods, IC-Light (Zhang et al., 2025) demonstrates strong results through scalable diffusion training coupled with light-transport consistency. Building on this line, RelightVid (Ponglertnapakorn et al., 2025) extends relighting to videos via temporal attention, whereas Light-A-Video (Zhou et al., 2025b) provides a training-free alternative by injecting cross-frame attention during inference.

4D Video Generation. Camera-controllable 4D video generation aims to synthesize temporally coherent dynamic scenes from novel viewpoints. Existing approaches can be roughly organized by how they enforce 3D/4D structures during generation. One line of work injects explicit geometric priors or relies on intermediate reconstructions to stabilize view changes. For instance, EX-4D uses watertight mesh reasoning to support extreme viewpoint variations (Hu et al., 2025), and Light-X introduces disentangled geometry cues to enable joint camera and illumination control (Liu et al., 2025). Another line emphasizes feed-forward generative modeling over dynamic 3D/4D representations for novel-view rendering, including point-cloud-based formulations in 4DNeX (Chen et al., 2025) and diffusion-based 4D occupancy generation in DynamicCity (Bian et al., 2024). In parallel, camera-conditioned diffusion and pseudo-4D guidance have improved controllability for view synthesis and camera-aware video generation (Fan et al., 2025; Bian et al., 2025).

4D Relighting. 4D relighting aims to synthesize dynamic scene appearance under novel illumination and varying viewpoints. Despite its significance for immersive virtual environments and content creation, 4D

relighting with viewpoint changes remains a nascent and challenging problem. Most existing systems are training-based and rely on explicit scene representations. Relighting4D (Chen & Liu, 2022) reconstructs space time neural fields and uses physically-based rendering to decompose scenes into components such as normals, occlusion, and diffuse effects. Relightable Neural Actor (Luvizon et al., 2024) targets human performance capture by leveraging multi-view inputs together with intrinsic and material decomposition for pose-conditioned relighting. Light-X (Liu et al., 2025) provides a unified pipeline for joint camera and illumination control by training on synthetic paired data, producing videos with coupled viewpoint and lighting variations. BEAM (Hong et al., 2025) represents dynamic scenes using 4D Gaussians and physically-based rendering to recover material properties for high-quality relightable video. Recent monocular relightable Gaussian methods (Choi et al., 2025; Jiang et al., 2025; Schmidt et al., 2025) aim to reduce the need for multi-view capture.

3 The Proposed Method

3.1 Overview

We propose Light4D, a novel training-free framework designed to synthesize a target 4D video $\mathcal{V} = \{I^f\}_{f=1}^F$ from a monocular source video $\mathcal{V}_s = \{I_s^f\}_{f=1}^F$, as illustrated in Figure 2. Our primary objective is to rerender the dynamic scene under a target camera trajectory $\mathcal{C} = \{P_f\}_{f=1}^F$ and user-specified illumination conditions $\mathcal{L}$. Specifically, the camera trajectory $\mathcal{C}$ spans a wide viewpoint range ($-90°$ to $90°$) relative to the original coordinate system, while the illumination $\mathcal{L}$ is modulated by text prompts corresponding to distinct lighting directions (e.g., Left, Right, Top, Bottom). The generated video $\mathcal{V}$ must faithfully preserve the 3D geometry and motion dynamics of the source $\mathcal{V}_s$, while accurately adhering to the novel viewpoint $\mathcal{C}$ and lighting constraints $\mathcal{L}$. The overall inference pipeline achieving these goals is summarized in Algorithm 1.

3.2 Disentangled Flow Guidance

A critical challenge in 4D relighting arises from the inherent conflict between geometric reconstruction and illumination synthesis within the generative manifold. In diffusion model theory, coarse 3D geometry primarily forms during the early, high-noise stages (large timesteps). Consequently, the premature injection of high-frequency 2D illumination cues can disrupt the ordinary differential equation (ODE) trajectory before the geometric manifold is fully established, resulting in structural collapse.

To mitigate these issues, we employ a time-aware strategy that progressively integrates illumination cues into the latent space. Formally, let z_t denote the latent state at timestep $t \in [0, 1]$. We formulate the generative process as guiding a pre-trained geometric flow $\mathbf{v}_{geo}$ instantiated by EX-4D, using an illumination-aware correction derived from a single-view relighting prior $\mathcal{M}_{light}$ based on IC-Light.

At each timestep t, we first estimate the clean geometric state $\hat{z}_0^{geo}$ derived from the current flow velocity. Since the relighting prior $\mathcal{M}_{light}$ operates in the image domain, we project this latent estimate into the pixel space via the VAE decoder $\mathcal{D}$. Then, we utilize the relighting prior $\mathcal{M}_{light}$ to predict the relighted image $\hat{x}_0^{light}$ conditioned on the lighting prompt $\mathcal{L}$:

$$\hat{x}_0^{light} = \mathcal{M}_{light}\left(\mathcal{D}(\hat{z}_0^{geo}), \mathcal{L}, \epsilon_{shared}\right), \tag{1}$$

where ϵ_{shared} represents a canonical noise prior.

We then construct a hybrid flow target z_{target} by fusing the geometric structure and illumination cues in the latent space, governed by a time-dependent fusion weight $\lambda(t)$:

$$z_{target} = \mathcal{E}\left((1 - \lambda(t)) \cdot \mathcal{D}(\hat{z}_0^{geo}) + \lambda(t) \cdot \hat{x}_0^{light}\right), \tag{2}$$

where $\mathcal{E}$ denotes the VAE encoder.

This hybrid target z_{target} serves as the target, defining a rectified flow trajectory toward the illumination-enhanced manifold. Based on this target, we perform a discrete ODE update step to compute the state $z_{t'}$

Algorithm 1 Inference of Light4D

Input: source video $\mathcal{V}_s$, target camera $\mathcal{C}$, lighting prompts $\mathcal{L}$, EX-4D model $\mathbf{V}_{geo}$, IC-Light prior $\mathcal{M}_{light}$, VAE $(\mathcal{E}, \mathcal{D})$, noise levels $\{\sigma_k\}$, time schedule $\{t_k\}$
Output: relit 4D video $\mathcal{V}$
Sample noise ϵ_{shared} // Canonical Noise Initialization (Equation (8))
$z \leftarrow z_{init}(\mathcal{V}_s, \mathcal{C})$ // Initialize latent
for $k = 1, 2, \ldots, K$ **do**
  $t \leftarrow t_k$ // Map discrete step k to continuous time
  $\hat{z}_0^{geo} \leftarrow \mathbf{V}_{geo}(z, \sigma_t)$ // Estimate geometry state
  if $\lambda(t) > 0$ **then**
    $x^{geo} \leftarrow \mathcal{D}(\hat{z}_0^{geo})$ // Decode to RGB space
    $\hat{x}_0^{light} \leftarrow \mathcal{M}_{light}(x^{geo}, \mathcal{L}, \epsilon_{shared}; \text{TCA})$ // Inject light (Equation (1))
    $\tilde{x} \leftarrow \text{GMM}(\hat{x}_0^{light})$ // Global Moment Matching (Equation (9))
    $x^{light} \leftarrow \text{FDI}(\tilde{x})$ // Freq-Decoupled Regularization (Equation (10))
    $x^{fuse} \leftarrow (1 - \lambda(t))x^{geo} + \lambda(t)x^{light}$
    $z_{target} \leftarrow \mathcal{E}(x^{fuse})$ // Hybrid target (Equation (2))
  else
    $z_{target} \leftarrow \hat{z}_0^{geo}$ // Geometric Isolation Phase
  end if
  $z \leftarrow z + (\sigma_{t'} - \sigma_t) \cdot \frac{z - z_{target}}{\sigma_t + \delta}$ // Euler solver (Equation (3))
end for
$\mathcal{V} \leftarrow \mathcal{D}(z)$
return $\mathcal{V}$

for the next timestep t'. Specifically, we employ a first-order Euler solver:

$$z_{t'} = z_t + (\sigma_{t'} - \sigma_t) \cdot \frac{z_t - z_{target}}{\sigma_t + \delta}, \tag{3}$$

where $\sigma_{t'}$ and σ_t denote the noise levels, and δ is a numerical stabilizer.

Crucially, to mitigate interference between geometry generation and illumination refinement, $\lambda(t)$ follows a multi-phase adaptive schedule. Unlike standard linear annealing, this schedule disentangles the trajectory into four distinct phases, parameterized by the peak intensity λ_{max} and the terminal weight λ_{end}:

$$\lambda(t) = \begin{cases} 0, & t \in (\tau_g, 1] \\ \lambda_{max} \cdot \sqrt{\frac{\tau_g - t}{\tau_g - \tau_r}}, & t \in [\tau_r, \tau_g] \\ \lambda_{max}, & t \in [\tau_s, \tau_r) \\ \frac{\lambda_{max} - \lambda_{end}}{\tau_s} \cdot t + \lambda_{end}, & t \in [0, \tau_s) \end{cases} \tag{4}$$

where τ_g, τ_r, τ_s are the time thresholds for the geometric, ramp-up, and stable phases, respectively.

To align the injection timing with the underlying manifold formation stages, our schedule operates as follows: Initially, during the **geometric isolation phase** $(t > \tau_g)$, we enforce $\lambda(t) = 0$ to strictly prohibit illumination cues from interfering with the completion of invisible regions and occlusion relationships, thereby preserving the critical geometric manifold formation stage. Subsequently, a **progressive ramp phase** $[\tau_r, \tau_g]$ is employed to smoothly transition the ODE trajectory toward the illumination target, rapidly anchoring lighting conditions without inducing abrupt "feature shock". This is followed by a **stable injection phase** $[\tau_s, \tau_r)$ at λ_{max} that sustains illumination injection once the 3D geometry is locked, firmly establishing the relighting effect and allowing the model to reconcile appearance cues with the established structure. Finally, for $t < \tau_s$, the **convergence fade phase** linearly decays the weight to λ_{end}. This reduces injection at low-noise stages where fine textures are synthesized, ensuring a soft landing that prevents the relighting prior from dominating and gracefully preserves high-frequency geometric details.

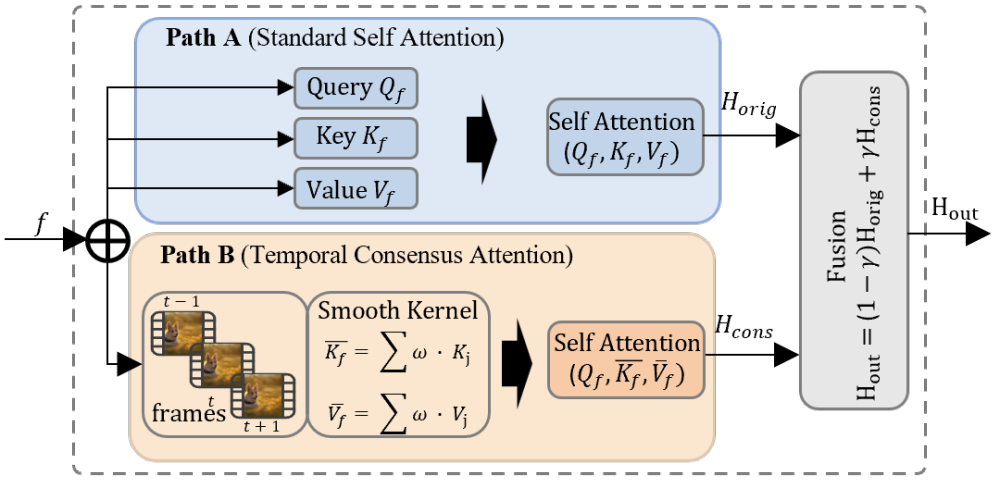

Figure 3: **Design of Temporal Consistent Attention (TCA)**. TCA enforces temporal coherence through a dual-path mechanism that interpolates between standard self-attention (Path A) for frame-specific structure and a consistent path (Path B) that regularizes appearance context via Gaussian-weighted sliding window aggregation.

3.3 Temporal Consistent Attention

While Disentangled Flow Guidance (Section 3.2) effectively steers the geometric trajectory, the reliance on a frame-independent relighting prior inevitably introduces stochastic fluctuations in the feature space, leading to both geometric instability and temporal flickering. To resolve this, we introduce Temporal Consistent Attention (TCA). As illustrated in Figure 3, this dual-path mechanism modifies standard self-attention to enforce strict temporal coherence while preserving high-fidelity illumination details.

We redesign its internal self-attention mechanism to assign distinct temporal behaviors. Specifically, the Query features ($\mathbf{Q}$), which encode frame-specific geometric structure, are preserved in their transient state to faithfully capture dynamic changes. In contrast, the Key ($\mathbf{K}$) and Value ($\mathbf{V}$) pairs, which provide the appearance context, are regularized to exhibit strong temporal stability. We employ a Gaussian-weighted sliding window approach to construct a smoothed context $\{\overline{\mathbf{K}}_f, \overline{\mathbf{V}}_f\}$ by aggregating features within a local temporal neighborhood $\mathcal{N}(f) = \{j \mid |f-j| \leq r\}$, weighted by their temporal proximity:

$$\begin{aligned} \overline{\mathbf{K}}_f &= \sum_{j \in \mathcal{N}(f)} w(|f-j|) \cdot \mathbf{K}_j, \\ \overline{\mathbf{V}}_f &= \sum_{j \in \mathcal{N}(f)} w(|f-j|) \cdot \mathbf{V}_j, \end{aligned} \tag{5}$$

where r denotes the temporal radius, f represents the current frame index, and $w(d) = \exp(-d^2/2\sigma^2)$ serves as a Gaussian weighting kernel that prioritizes contributions from temporally adjacent frames.

To balance single-frame sharpness with temporal stability, TCA implements a dual-path residual injection strategy. We simultaneously compute the original stochastic features $\mathbf{H}_{\text{orig}}$ using standard self-attention to retain high-frequency details, and the temporally consistent features $\mathbf{H}_{\text{cons}}$ by attending the frame-specific query $\mathbf{Q}_f$ to the temporally smoothed context $\{\overline{\mathbf{K}}_f, \overline{\mathbf{V}}_f\}$:

$$\begin{aligned} \mathbf{H}_{\text{orig}} &= \text{Attention}(\mathbf{Q}_f, \mathbf{K}_f, \mathbf{V}_f), \\ \mathbf{H}_{\text{cons}} &= \text{Attention}(\mathbf{Q}_f, \overline{\mathbf{K}}_f, \overline{\mathbf{V}}_f). \end{aligned} \tag{6}$$

Subsequently, the final hidden state is derived via a linear interpolation controlled by a mixing coefficient $\gamma \in [0, 1]$:

$$\mathbf{H}_{out} = (1-\gamma)\mathbf{H}_{\text{orig}} + \gamma \mathbf{H}_{\text{cons}}. \tag{7}$$

This residual formulation effectively mitigates the propagation of temporal variance into the rectification target while preserving the structural distinctiveness of individual frames.

3.4 Deterministic Coherence and Regularization

To eliminate temporal artifacts, we apply deterministic regularization to the raw predictions $\hat{x}_0^{light}$ from the IC-Light prior before integration. Instead of directly injecting the stochastic output, we first rectify the signal to improve temporal stability. This processed appearance is then fused with the geometric projection to construct the hybrid flow target.

Canonical Noise Initialization. The stochastic nature of diffusion sampling introduces aleatoric uncertainty, leading to high-frequency texture jitter across views. We mitigate this by conditioning the relighting model on a canonical noise prior. Instead of independent sampling, we broadcast a fixed noise map ϵ_{shared} across the temporal sequence, ensuring a topologically consistent generation path for identical semantic regions.

$$\epsilon_f = \epsilon_{\text{shared}}, \quad \forall f \in [1, F] \tag{8}$$

where F denotes the number of frames in the video.

Global Moment Matching. To address global exposure fluctuations caused by frame-independent processing, we enforce statistical alignment across the video sequence. We construct a canonical reference by computing the temporal average of the entire prediction sequence. Specifically, we rectify the intensity distribution of each raw prediction $\hat{x}_f^{light}$ to match the mean μ_{ref} and standard deviation σ_{ref} of this temporal average reference. This ensures consistent brightness levels throughout the trajectory, yielding the aligned intermediate frame $\tilde{x}_f$:

$$\tilde{x}_f = \mu_{ref} + \frac{\sigma_{ref}}{\sigma_f}(\hat{x}_f^{light} - \mu_f) \tag{9}$$

where μ_f and σ_f denote the intensity mean and standard deviation of the current frame f, respectively.

Frequency-Decoupled Illuminance Regularization. Finally, to resolve luminance flickering without compromising geometric sharpness, we employ a spectral decomposition strategy. Based on the assumption that valid illumination changes are spectrally concentrated in the low-frequency domain while texture resides in high frequencies, we decompose the aligned signal $\tilde{x}_f$. We apply a temporal smoothing operator $\mathcal{T}$ exclusively to the base illumination layer, while preserving the high-frequency details:

$$x_f^{light} = \underbrace{\mathcal{T}(\tilde{x}_f * G_\sigma)}_{\text{Smoothed Illumination}} + \underbrace{(\tilde{x}_f - \tilde{x}_f * G_\sigma)}_{\text{Preserved Texture}} \tag{10}$$

where $*$ denotes the convolution operation. The resulting x_f^{light} serves as the final robust appearance target for the flow matching objective defined in Equation (2).

4 Experiment

4.1 Experimental Setup

Baselines. Since Light4D is the first training-free framework for 4D video relighting, we compare it with both state-of-the-art training-based methods and cascaded training-free baselines. For training-based approaches, we select Light-X (Liu et al., 2025), which is trained to jointly control camera motion and illumination. For training-free comparisons, we construct two sequential pipelines: EX-4D (Hu et al., 2025) → Light-A-Video (LAV) (Zhou et al., 2025b), and LAV (Zhou et al., 2025b) → EX-4D (Hu et al., 2025). We also include a naive baseline EX-4D (Hu et al., 2025) + IC-Light (Zhang et al., 2025) to expose a lower bound on temporal stability when relighting is applied independently to each frame. In this study, we focus on both the quality of relighting and the robustness of temporal consistency. All methods are evaluated under identical input and relighting prompt settings to ensure fair comparison.

Benchmark. To evaluate 4D relighting under extreme viewpoint changes, we built an evaluation benchmark of 100 high-quality generated videos produced by recent sota video models (Sora (Brooks et al., 2024),

Figure 4: **Qualitative relighting results.** Comparison of baselines and our method under two prompts: "Sunlight" (left) and "Pink neon light" (left). Our method yields more stable illumination changes over time and reduces temporal flicker.

WanVideo (Wan et al., 2025), and Kling (Team et al., 2025)), spanning diverse semantic categories including humans, animals, objects, and landscapes. In addition, we collect 50 real-world driving sequences from the OpenScene dataset (OpenScene Dataset Contributors, 2023) to investigate generalization to real captured videos. Each video is evaluated in three movement ranges of the camera $(30°, 90°, 180°)$ to comprehensively assess robustness under varying degrees of viewpoint variation.

Evaluation Metrics. We evaluate relit videos against source videos under three camera movement ranges (30°, 90°, 180°) using two metric groups. *Relighting-related* metrics include: (i) *CLIP-Frame* (Radford et al., 2021), the inter-frame similarity of CLIP embeddings computed on consecutive frames to quantify temporal coherence; (ii) *Motion Flow L1*, the mean distance ℓ_1 between RAFT (Teed & Deng, 2020) optical flows estimated from relit and source videos to measure motion preservation; (iii) *High Frequency Preservation Ratio (HFPR)* (Zhu et al., 2012; 2013), the ratio of Laplacian-filtered high-frequency energy between the relit and source videos to assess detail retention; and (iv) *Aesthetic Score* (Schuhmann et al., 2022) for perceptual appeal. *Video quality* metrics measure reconstruction fidelity: (i) *Frame PSNR*, and (ii–iii) *warp-aligned* SSIM and LPIPS, where we estimate flow using RAFT and warp adjacent frames to a common reference before computing the metrics, thereby reducing motion-induced misalignment and better isolating appearance and structural consistency over time (Wang et al., 2004; Zhang et al., 2018).

Implementation Details. Our framework is based on EX-4D (Hu et al., 2025) as the generative backbone of 4D and IC-Light (Zhang et al., 2025) as the illumination prior. Our default setting generates 49-frame videos at a resolution of 384×384 with $T = 30$ denoising steps, and all experiments are run on NVIDIA H20 GPU. To mitigate geometry–illumination interference, we adopt a time-aware fusion strategy that delays illumination injection to later denoising stages, prioritizing 4D structure formation before enforcing relighting cues. We further improve temporal stability via a lightweight temporal consistency module and post-processing smoothing. The complete hyperparameters and trajectory alignment details are provided in the Appendix B. All experiments are conducted using the same set of hyperparameters.

Table 1: **Relighting-related metrics across viewpoint changes.** We report CLIP-Frame, Motion Flow L1, HFPR, and Aesthetic Score under camera motion ranges of 30°, 90°, and 180°. These metrics jointly reflect the trade-off between *lighting consistency*, *4D geometric and motion stability*, and *detail fidelity* under extreme viewpoints. Best results are **bolded** and second-best are underlined.

Method	CLIP-Frame ↑			Motion Flow L1 ↓			HFPR ↑			Aesthetic ↑		
	30°	90°	180°	30°	90°	180°	30°	90°	180°	30°	90°	180°
Training-based												
Light-X	0.956	0.900	0.901	0.814	**1.832**	4.470	0.945	0.949	0.931	0.235	0.229	0.221
Training-free												
EX-4D + IC-Light	0.923	0.885	0.885	15.725	21.652	24.114	0.842	0.838	0.817	0.228	0.223	0.219
EX-4D + LAV	0.961	0.925	0.923	1.287	3.200	6.095	0.951	0.959	0.946	0.209	0.189	0.168
LAV + EX-4D	0.959	0.918	0.922	0.982	2.923	4.476	0.949	0.932	0.944	0.199	0.174	0.169
Light4D (Ours)	**0.975**	**0.930**	**0.930**	**0.791**	1.936	**3.524**	**0.974**	**0.963**	**0.966**	**0.243**	**0.235**	**0.231**

Table 2: **Video quality metrics across viewpoint changes.** We measure frame-wise reconstruction similarity between the relighting videos and the corresponding source videos under viewpoint changes of 30°, 90°, and 180°. We report per-frame PSNR, per-frame SSIM, and per-frame LPIPS to quantify content and detail preservation. Best results are **bolded** and second-best are underlined.

Method	Frame PSNR ↑			Frame SSIM ↑			Frame LPIPS ↓		
	30°	90°	180°	30°	90°	180°	30°	90°	180°
Training-based									
Light-X	12.899	13.544	13.831	0.738	0.733	0.752	**0.349**	0.358	0.332
Training-free									
EX-4D + IC-Light	9.147	9.412	9.035	0.493	0.489	0.473	0.754	0.760	0.748
EX-4D + LAV	11.957	11.843	11.449	0.696	0.659	0.652	0.572	0.625	0.624
LAV + EX-4D	12.348	12.007	12.023	0.636	0.564	0.554	0.535	0.560	0.552
Light4D (Ours)	**14.056**	**13.728**	**13.941**	**0.761**	**0.759**	**0.753**	0.360	**0.341**	**0.307**

4.2 Main Results

Relighting Quality Results. We visualize relighting outputs under two lighting conditions, "Sunlight" (left) and "Pink neon light" (left), as shown in Figure 4. Quantitatively (Table 1), our method achieves the strongest overall performance at 30°, 90°, and 180° in CLIP-Frame, HFPR, and Aesthetic Score, suggesting improved temporal coherence and perceptual quality while preserving fine details. These gains are consistent with the visual comparisons in Figure 4, where our relighting exhibits smoother light transitions and fewer flickering artifacts. The naive EX-4D+IC-Light baseline (Hu et al., 2025; Zhang et al., 2025) can yield sharp individual frames, but frame-wise relighting introduces severe temporal inconsistency, resulting in higher Motion Flow L1 and visibly unstable illumination. The cascade pipelines exhibit different trade-offs: both EX-4D→LAV and LAV→EX-4D (Hu et al., 2025; Zhou et al., 2025b) improve temporal behavior compared to frame-wise relighting; however, yet their lighting cues are not consistently aligned with the evolving 4D geometry under large rotations. This often leads to less natural appearance and reduced motion or geometry stability as the viewpoint change increases. Light-X (Liu et al., 2025) performs well at moderate rotations, but degrades in larger viewpoints settings. In particular, visualizations reveal a more static, "baked-in" illumination pattern that does not adapt faithfully when newly visible geometry emerges.

4D Video Quality Results. Relighting 4D content while preserving geometry and fine-grained content under extreme viewpoints is particularly challenging. Figure 5 shows qualitative results under the "Sunlight" prompt, and Table 2 reports frame-wise reconstruction metrics between relit and source videos. Our method consistently leads in Frame PSNR and SSIM across all viewpoint ranges and achieves the lowest LPIPS at 90° and 180°, indicating strong content preservation without sacrificing perceptual detail. This indicates that

Figure 5: **4D video quality under extreme viewpoints.** Qualitative comparison of baselines and our method using the prompt "Sunlight". Under extreme viewpoint changes, our method better balances relighting coherence, 4D geometric stability, and detail fidelity.

Light4D preserves the underlying 4D content while applying substantial illumination edits, even as viewpoint changes increase. This indicates that Light4D preserves the underlying 4D content while applying substantial illumination edits, even as viewpoint changes increase. In contrast, cascaded baselines (Hu et al., 2025; Zhou et al., 2025b; Zhang et al., 2025) are more sensitive to large viewpoint changes: introducing relighting cues before or after 4D generation can amplify small geometric inconsistencies into visible distortions when rotations reach 90° and beyond. Light-X (Liu et al., 2025) maintains reasonable fidelity, yet shows marginally lower PSNR and SSIM scores in this extreme-view evaluation. In summary, our method offers a more balanced compromise under extreme camera motion, simultaneously maintaining illumination coherence, stabilizing 4D geometry, and preserving video details. To further examine behavior under severe viewpoint changes, we include additional qualitative analysis of newly occluded and disoccluded regions in Appendix C.

4.3 Ablation Studies

We validate the effectiveness of our design by ablating key components and reporting results from a 30° viewpoint. Specifically, we isolate the impact of each proposed module to demonstrate its necessity in our pipeline. Table 3 summarizes relighting metrics, and more ablation details are provided in Appendix A.

Table 3: **Ablation Study: Relighting Metrics at 30°.** Best results are **bolded**; second best are underlined.

Method	CLIP ↑	Motion ↓	HFPR ↑	Aes ↑	PSNR ↑	SSIM ↑	LPIPS ↓
w/o CLA	0.972	0.903	0.882	0.229	13.698	0.739	0.420
w/o DGA	0.965	1.259	0.754	0.213	10.767	0.513	0.718
w/o FDI	0.970	0.885	0.910	0.228	13.351	0.680	0.502
w/o CNI	0.971	0.875	0.925	0.229	13.487	0.783	0.470
w/o GMM	0.972	0.860	0.940	0.230	13.650	0.711	0.461
w/o All	0.973	0.898	0.893	0.231	13.207	0.657	0.518
Full Model	**0.975**	**0.791**	**0.974**	**0.243**	**14.056**	**0.761**	**0.360**

Impact of Disentangled Guidance (DGA). Removing DGA (*w/o DGA*) markedly degrades relighting quality: Motion Flow L1 increases substantially and HFPR drops, indicating weaker motion preservation and loss of fine details. This supports the role of DGA in injecting illumination cues without destabilizing the underlying 4D geometry.

Effect of Consistent Light Attention (CLA). Disabling CLA (*w/o CLA*) reduces temporal coherence in relighting, reflected by higher Motion Flow L1 compared to the full model. It also leads to more noticeable frame-to-frame lighting jitter in visually smooth regions. This suggests that CLA is important for stabilizing illumination across consecutive frames.

Effect of Regularization. We further ablate the Canonical Noise Initialization (CNI), Global Moment Matching (GMM), and Frequency-Decoupled Illuminance (FDI). Overall, removing these regularizers harms relighting quality in complementary ways. Specifically, removing all smoothing increases temporal instability, whereas the full model preserves sharp details with higher HFPR and achieves better perceptual quality with a higher Aesthetic Score. Across our regularization modules, we observe the largest degradation when removing FDI, followed by CNI, and then GMM. While their effects are partially complementary and can vary slightly across metrics, the overall trend is consistent across temporal coherence and detail-preservation measures.

5 Limitation and Future Work

As a training-free framework, Light4D's performance is bounded by the capabilities of its foundation models, specifically the geometric fidelity of EX-4D and the single-frame nature of the IC-Light prior. Consequently, maintaining rigorous global lighting consistency remains challenging when adapting 2D image priors to the temporal domain, particularly during extreme viewpoint traversals ($-90° \sim 90°$). Future work will leverage the framework's modularity to integrate more advanced video-native illumination models and stronger 4D generative backbones to further improve synthesis quality. We also aim to investigate specialized mechanisms to mitigate photometric inconsistencies induced by large viewpoint variations, while extending the framework to handle complex light transport effects, such as cast shadows and inter-reflections.

6 Conclusion

In this paper, we present Light4D, a training-free framework for 4D video relighting under extreme viewpoint changes. By identifying and resolving the conflict between geometric reconstruction and illumination synthesis, we propose Disentangled Flow Guidance to harmonize these objectives with a time-aware schedule. Furthermore, we introduce Temporal Consistent Attention and Deterministic Coherence Regularization to ensure the generated content remains geometrically consistent and free of visible flicker. Extensive experiments show that our method achieves competitive performance in both geometric consistency and lighting fidelity compared to baselines, offering a scalable solution for controllable 4D content creation. Finally, the modular design of our framework makes it extensible, enabling integration of future advances in generative models.

References

Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7996–8006, 2024.

Shrisha Bharadwaj, Haiwen Feng, Giorgio Becherini, Victoria Fernandez Abrevaya, and Michael J Black. Genlit: Reformulating single-image relighting as video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025.

Hengwei Bian, Lingdong Kong, Haozhe Xie, Liang Pan, Yu Qiao, and Ziwei Liu. Dynamiccity: Large-scale 4d occupancy generation from dynamic scenes. *arXiv preprint arXiv:2410.18084*, 2024.

Weikang Bian, Zhaoyang Huang, Xiaoyu Shi, Yijin Li, Fu-Yun Wang, and Hongsheng Li. Gs-dit: Advancing video generation with dynamic 3d gaussian fields through efficient dense 3d point tracking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21717–21727, 2025.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

Sreenithy Chandran, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Zhixin Shu, and Suren Jayasuriya. Temporally consistent relighting for portrait videos. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022.

Sumit Chaturvedi, Mengwei Ren, Yannick Hold-Geoffroy, Jingyuan Liu, Julie Dorsey, and Zhixin Shu. Synthlight: Portrait relighting with diffusion model by learning to re-render synthetic faces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 369–379, 2025.

Zhaoxi Chen and Ziwei Liu. Relighting4d: Neural relightable human from videos, 2022. URL `https://arxiv.org/abs/2207.07104`.

Zhaoxi Chen, Tianqi Liu, Long Zhuo, Jiawei Ren, Zeng Tao, He Zhu, Fangzhou Hong, Liang Pan, and Ziwei Liu. 4dnex: Feed-forward 4d generative modeling made easy. *arXiv preprint arXiv:2508.13154*, 2025.

Seonghwa Choi, Moonkyeong Choi, Mingyu Jang, Jaekyung Kim, Jianfei Cai, Wen-Huang Cheng, and Sanghoon Lee. Relightable and dynamic gaussian avatar reconstruction from monocular video. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 7405–7414, 2025.

Xiang Fan, Sharath Girish, Vivek Ramanujan, Chaoyang Wang, Ashkan Mirzaei, Petr Sushko, Aliaksandr Siarohin, Sergey Tulyakov, and Ranjay Krishna. Omniview: An all-seeing diffusion model for 3d and 4d view synthesis. *arXiv preprint arXiv:2512.10940*, 2025.

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

Kai He, Ruofan Liang, Jacob Munkberg, Jon Hasselgren, Nandita Vijaykumar, Alexander Keller, Sanja Fidler, Igor Gilitschenski, Zan Gojcic, and Zian Wang. Unirelight: Learning joint decomposition and synthesis for video relighting, 2025. URL `https://arxiv.org/abs/2506.15673`.

Yu Hong, Yize Wu, Zhehao Shen, Chengcheng Guo, Yuheng Jiang, Yingliang Zhang, Jingyi Yu, and Lan Xu. Beam: Bridging physically-based rendering and gaussian modeling for relightable volumetric video. *arXiv preprint arXiv:2502.08297*, 2025.

Tao Hu, Haoyang Peng, Xiao Liu, and Yuewen Ma. Ex-4d: Extreme viewpoint 4d video synthesis via depth watertight mesh. *arXiv preprint arXiv:2506.05554*, 2025.

Zeren Jiang, Shaofei Wang, and Siyu Tang. Dnf-avatar: Distilling neural fields for real-time animatable avatar relighting. *arXiv preprint arXiv:2504.10486*, 2025.

Hoon Kim, Minje Jang, Wonjun Yoon, Jisoo Lee, Donghyun Na, and Sanghyun Woo. Switchlight: Co-design of physics-driven architecture and pre-training framework for human portrait relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25096–25106, 2024.

Peter Kocsis, Julien Philip, Kalyan Sunkavalli, Matthias Nie
ssner, and Yannick Hold-Geoffroy. Lightit: Illumination modeling and control for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9359–9369, 2024.

Yuanze Lin, Yi-Wen Chen, Yi-Hsuan Tsai, Ronald Clark, and Ming-Hsuan Yang. Illumicraft: Unified geometry and illumination diffusion for controllable video generation. *arXiv preprint arXiv:2506.03150*, 2025.

Tianqi Liu, Zhaoxi Chen, Zihao Huang, Shaocong Xu, Saining Zhang, Chongjie Ye, Bohan Li, Zhiguo Cao, Wei Li, Hao Zhao, et al. Light-x: Generative 4d video rendering with camera and illumination control. *arXiv preprint arXiv:2512.05115*, 2025.

Diogo Luvizon, Vladislav Golyanik, Adam Kortylewski, Marc Habermann, and Christian Theobalt. Relightable neural actor with intrinsic decomposition and pose control, 2024. URL `https://arxiv.org/abs/2312.11587`.

Nadav Magar, Amir Hertz, Eric Tabellion, Yael Pritch, Alex Rav-Acha, Ariel Shamir, and Yedid Hoshen. Lightlab: Controlling light sources in images with diffusion models. *arXiv preprint arXiv:2505.09608*, 2025.

Yiqun Mei, Mingming He, Li Ma, Julien Philip, Wenqi Xian, David M George, Xueming Yu, Gabriel Dedic, Ahmet Levent Taşel, Ning Yu, Vishal M. Patel, and Paul Debevec. Lux post facto: Learning portrait performance relighting with conditional video diffusion and a hybrid dataset, 2025. URL `https://arxiv.org/abs/2503.14485`.

Zhenxing Mi, Yuxin Wang, and Dan Xu. One4d: Unified 4d generation and reconstruction via decoupled lora control. *arXiv preprint arXiv:2511.18922*, 2025.

OpenScene Dataset Contributors. OpenScene: The Largest Up-to-Date 3D Occupancy Prediction Benchmark in Autonomous Driving, August 2023. URL `https://github.com/OpenDriveLab/OpenScene`.

Hui En Pang, Shuai Liu, Zhongang Cai, Lei Yang, Tianwei Zhang, and Ziwei Liu. Disco4d: Disentangled 4d human generation and animation from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26331–26344, 2025.

Puntawat Ponglertnapakorn, Nontawat Tritrong, and Supasorn Suwajanakorn. Difareli: Diffusion face relighting, 2023.

Puntawat Ponglertnapakorn, Nontawat Tritrong, and Supasorn Suwajanakorn. Difareli++: Diffusion face relighting with consistent cast shadows. *IEEE transactions on pattern analysis and machine intelligence*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Jonathan Schmidt, Simon Giebenhain, and Matthias Niessner. Becominglit: Relightable gaussian avatars with hybrid neural shading. *arXiv preprint arXiv:2506.06271*, 2025.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS Datasets and Benchmarks*, 2022.

Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, Xiao Hu, Xiaohua Hu, Boyuan Jiang, Fangyuan Kong, Hang Li, Jie Li, Qingyu Li, Shen Li, Xiaohan Li, Yan Li, Jiajun Liang, Borui Liao, Yiqiao Liao, Weihong Lin, Quande Liu, Xiaokun Liu, Yilun Liu, Yuliang Liu, Shun Lu, Hangyu Mao, Yunyao Mao, Haodong Ouyang, Wenyu Qin, Wanqi Shi, Xiaoyu Shi, Lianghao Su, Haozhi Sun, Peiqin Sun, Pengfei Wan, Chao Wang, Chenyu Wang, Meng Wang, Qiulin Wang, Runqi Wang, Xintao Wang, Xuebo Wang, Zekun Wang, Min Wei, Tiancheng Wen, Guohao Wu, Xiaoshi Wu, Zhenhua Wu, Da Xie, Yingtong Xiong, Yulong Xu, Sile Yang, Zikang Yang, Weicai Ye, Ziyang Yuan, Shenglong Zhang, Shuaiyu Zhang, Yuanxing Zhang, Yufan Zhang, Wenzheng Zhao, Ruiliang Zhou, Yan Zhou, Guosheng Zhu, and Yongjie Zhu. Kling-omni technical report, 2025. URL `https://arxiv.org/abs/2512.16776`.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision (ECCV)*, 2020.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Junying Wang, Jingyuan Liu, Xin Sun, Krishna Kumar Singh, Zhixin Shu, He Zhang, Jimei Yang, Nanxuan Zhao, Tuanfeng Y Wang, Simon S Chen, et al. Comprehensive relighting: Generalizable and consistent monocular human relighting and harmonization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 380–390, 2025.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Kairun Wen, Yuzhi Huang, Runyu Chen, Hui Zheng, Yunlong Lin, Panwang Pan, Chenxin Li, Wenyan Cong, Jian Zhang, Junbin Lu, et al. Dynamicverse: A physically-aware multimodal framework for 4d world modeling. *arXiv preprint arXiv:2512.03000*, 2025.

Longwen Zhang, Qixuan Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Neural video portrait relighting in real-time via consistency modeling. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 802–812, 2021a.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *The Thirteenth International Conference on Learning Representations*, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (ToG)*, 40(6):1–18, 2021b.

Haiyang Zhou, Wangbo Yu, Jiawen Guan, Xinhua Cheng, Yonghong Tian, and Li Yuan. Holotime: Taming video diffusion models for panoramic 4d scene generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9763–9772, 2025a.

Yujie Zhou, Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, Qidong Huang, Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, et al. Light-a-video: Training-free video relighting via progressive light fusion. *arXiv preprint arXiv:2502.08590*, 2025b.

Kongfeng Zhu, Shujun Li, and Dietmar Saupe. An objective method of measuring texture preservation for camcorder performance evaluation. In *Proceedings of SPIE-IS&T Electronic Imaging: Image Quality and System Performance IX*, volume 8293, pp. 829304, 2012. doi: 10.1117/12.907265.

Kongfeng Zhu, Keigo Hirakawa, Vijayan K. Asari, and Dietmar Saupe. A no-reference video quality assessment based on Laplacian pyramids. In *2013 IEEE International Conference on Image Processing (ICIP)*, pp. 49–53. IEEE, 2013.

A Additional Ablation Studies

A.1 Ablation on Geometric Isolation Phase (τ_g).

To validate the necessity of the Geometric Isolation Phase described in Section 3.2, we conduct an ablation study on the timing threshold τ_g. This parameter controls how long the illumination fusion weight $\lambda(t)$ is forced to zero, allowing the EX-4D backbone to establish structure without interference from the lighting prior.

A quantitative comparison of geometric isolation thresholds τ_g is provided in Figure 6. We use all evaluation metrics, including video quality and relighting metrics. The results indicate a clear trade-off: premature lighting injection can disrupt the formation of coherent 4D geometry, often leading to reduced temporal consistency and visible flickering, as evidenced by higher motion errors and lower detail preservation. Conversely, excessively delayed injection reduces the number of denoising steps available for illumination integration, potentially resulting in incomplete or unnatural lighting effects. Overall, our results suggest that setting $\tau_g = 0.7$ offers a favorable balance, anchoring the 4D structure while leaving sufficient flexibility for the relighting prior.

A.2 Ablation on Multi-Phase Schedule.

While Section A.1 demonstrates the critical role of the Geometric Isolation Phase in preventing structural collapse, we further conduct ablations to validate the necessity of the subsequent transitional phases (Progressive Ramp and Convergence Fade) in our proposed Disentangled Flow Guidance (DFG). To prove that our four-phase schedule is a principled design rather than an empirically hand-crafted curve, we fix the geometric isolation threshold ($\tau_g = 0.7$) and compare our full schedule against three variants:

- **Linear:** Replaces the ramp, stable, and fade phases with a standard linear increase from 0 to λ_{max} after τ_g.
- **Step:** Removes the progressive ramp phase by setting $\tau_r = \tau_g$, injecting the illumination control abruptly at peak intensity λ_{max}.
- **No-Fade:** Removes the convergence fade phase by setting $\tau_s = 0$, maintaining the maximum injection weight λ_{max} until the end of the generative process.

As shown in Table 4, the four-phase schedule achieves the best overall performance across both relighting fidelity and video quality metrics. Specifically, the **Linear** variant, which applies a naive constant increase without a plateau, yields suboptimal temporal stability and lower geometric fidelity (PSNR drops from 14.056 to 13.243). This highlights the necessity of the Stable Injection Phase, which provides a crucial sustained period at peak intensity for the model to firmly reconcile appearance cues with the locked 3D structure. Similarly, the **Step** variant causes a noticeable drop in temporal consistency (Motion Flow L1 increases from 0.791 to 1.039) and structural preservation (PSNR/SSIM drops), confirming that a Progressive Ramp Phase is essential to smoothly transition the ODE trajectory and prevent abrupt feature shock. Furthermore, the **No-Fade** variant degrades the High-Frequency Preservation Ratio (HFPR). This validates the theoretical motivation for the Convergence Fade Phase, reducing the injection weight at low-noise stages is necessary to prevent the relighting prior from overriding the fine-grained geometric textures synthesized during the final convergence steps.

B More Implementation Details

B.1 Baselines.

We evaluate ***Light4D*** against both training-based and training-free baselines under identical extreme viewpoint changes (30°, 90°, and 180°). To ensure a fair comparison, we align the camera trajectories used for video generation across methods. Specifically, for Light-X (Liu et al., 2025), we adopt its original camera

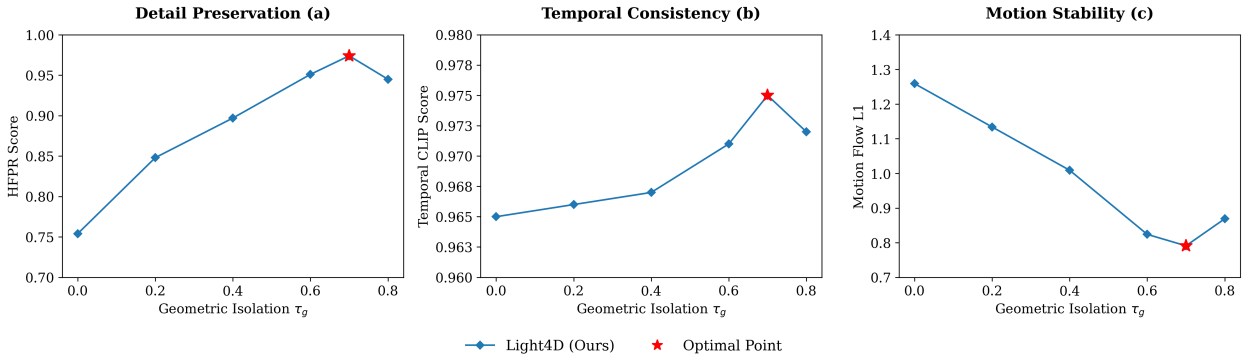

Figure 6: **Quantitative ablation on the geometric isolation threshold τ_g.** (a) HFPR scores, (b) Temporal CLIP scores, and (c) Motion Flow L1 errors are shown.

Table 4: **Multi-Phase Schedule Ablation:** Quantitative ablation on the multi-phase schedule at a 30° viewpoint change. All variants retain the initial geometric isolation phase. Best results are **bolded**, and second-best results are underlined.

Method	CLIP ↑	Motion ↓	HFPR ↑	Aes ↑	PSNR ↑	SSIM ↑	LPIPS ↓
w/o DGA	0.965	1.259	0.754	0.213	10.767	0.513	0.420
Linear	0.971	0.921	0.921	0.241	13.243	0.708	0.379
Step	0.968	1.039	0.892	0.239	12.322	0.661	0.385
No-Fade	0.970	1.062	0.919	0.241	12.447	0.669	0.381
Four-phase	**0.975**	**0.791**	**0.974**	**0.243**	**14.056**	**0.761**	**0.360**

extrinsic sequence with an EX-4D-style trajectory parameterization that matches both the viewpoint range and the motion pattern used in our setup, so that all methods are evaluated under the same camera motion. For Light-A-Video (LAV) (Zhou et al., 2025b), we use the officially released implementation of Wan2.1 (Wan et al., 2025).

B.2 Time-aware fusion schedule.

We use a piecewise schedule for the fusion weight: it is set to 0 for the first 70% of denoising steps (Steps 0–20) to prioritize geometric completion, and then linearly increases to a maximum value of 0.5 during the final 30% (Steps 21–29) for progressive illumination injection. Empirically, we found that values for this peak intensity λ_{max} within $[0.3, 0.7]$ stably control the relighting strength. We set $\lambda_{max} = 0.5$ as a robust default, as values below 0.3 tend to under-relight the scene, while values above 0.7 can lead to over-saturated colors.

B.3 Temporal consistency and smoothing.

The Consistent Light Attention (CLA) module uses a balance factor of $\gamma = 0.7$. This parameter controls the trade-off between single-frame sharpness and temporal consistency. We chose 0.7 because lowering it to 0 (w/o CLA) noticeably increases temporal flicker, whereas pushing it too high tends to over-smooth dynamic regions. The final output is processed with adaptive temporal smoothing using a window size of 9 and a Gaussian kernel with $\sigma = 25$ to suppress high-frequency flicker while preserving texture details.

B.4 User Study and Preference

We select three test sequences with viewpoint-change ranges of 30°, 90°, and 180°. For each viewpoint setting, we generate relit videos using three baselines (Hu et al., 2025; Zhou et al., 2025b; Zhang et al., 2025; Liu et al., 2025) and our method, all conditioned on the same source 4D video and the same relighting prompt.

Table 5: **User Study.** Ratings are on a 1 to 5 Likert scale. PM: Prompt Match; LC: Lighting Consistency; GC: Geometric Consistency; RR: Relighting Realism. Best results are **bolded**; second best are underlined.

	30°				90°				180°			
Method	PM↑	LC↑	GC↑	RR↑	PM↑	LC↑	GC↑	RR↑	PM↑	LC↑	GC↑	RR↑
Training-based												
Light-X	4.2	4.0	4.1	1.6	4.0	3.7	3.9	1.4	3.6	2.9	3.5	1.2
Training-free												
EX-4D + IC-Light	4.0	1.8	4.2	2.2	3.8	1.5	4.0	2.0	3.5	1.3	3.7	1.8
EX-4D + LAV	3.8	3.6	3.2	3.1	3.6	3.3	2.9	2.8	3.3	2.8	2.5	2.4
Light4D (Ours)	**4.5**	**4.4**	**4.3**	**4.4**	**4.4**	**4.2**	**4.2**	**4.2**	**4.2**	**3.9**	**4.0**	**3.9**

We distribute the survey to 30 participants and collect their ratings for each relit video on a 1–5 Likert scale along four dimensions: (i) Prompt Match (1 = poor, 5 = excellent), (ii) Lighting Consistency (1 = flickery, inconsistent, 5 = very stable), (iii) Geometric Consistency (1 = unstable, 5 = very stable), and (iv) Relighting Realism (1 = fake or overlay-like, 5 = realistic lighting). Figure 7 shows an example survey interface used in our study, where participants are provided with the source 4D video, the relighting prompt, and the five anonymized relighting results for side-by-side comparison. Results are presented in Table 5.

C Robustness in Newly Disoccluded Regions under Extreme Viewpoints

Handling newly occluded and disoccluded regions under extreme camera rotations poses a significant challenge in 4D video relighting, as generative priors like EX-4D often introduce geometric artifacts in these geometrically ambiguous areas. As illustrated in Figure 8, the EX-4D baseline tends to generate implausible structures, such as vertical occlusion artifacts in the "dog" sequence (under the "Left, Pink neon" setting) or scene-inconsistent background content in the "Labubu" sequence (under the "Top, Red neon" setting), when subjected to large viewpoint shifts. Conversely, our proposed Light4D framework effectively suppresses these inherited anomalies, yielding structurally coherent transitions in newly revealed regions. We attribute this enhanced robustness to the progressive relighting mechanism within the Disentangled Flow Guidance (DFG) module. By strategically delaying the injection of high-frequency illumination cues until the coarse 3D geometry is sufficiently stabilized, DFG prevents premature lighting constraints from dominating unformed regions. Consequently, this time-aware schedule significantly mitigates the emergence of hallucinated structures and maintains overall scene consistency during extreme viewpoint variations.

D More Experimental Results

In this appendix, we provide additional qualitative results that highlight the effectiveness of our method under challenging viewpoints, diverse lighting prompts, and complex real-world environments.

D.1 Visualizations under Normal Viewpoints.

We show qualitative comparisons across a variety of object categories under standard viewing conditions. As shown in Figures 8 and 9, these results highlight high-frequency detail preservation and natural light–material interactions. Our method also produces smooth temporal transitions as the light source moves. With our deterministic coherence and regularization mechanisms, we suppress the temporal flicker commonly observed in diffusion-based baselines, yielding stable 4D sequences.

D.2 Visualizations under Extreme Viewpoints.

To evaluate geometric robustness, we visualize relighting results under large camera viewpoint shifts. These settings are challenging because the overlap between reference and target views is small, which can cause

geometric distortion or texture drift in prior methods. As shown in Figure 10, our method preserves the underlying 3D structure without noticeable "ghosting" artifacts. By decoupling the geometric isolation phase from illumination modulation, our approach keeps relit shadows and highlights spatially coherent even under large viewpoint changes.

D.3 Visualizations in Autonomous Driving Scenarios.

To demonstrate practical utility on real videos, we apply ***Light4D*** to real-world autonomous driving sequences from OpenScene (OpenScene Dataset Contributors, 2023). These scenes include dynamic foreground objects and large outdoor environments. As shown in Figure 11, our method handles outdoor lighting changes and produces realistic shadows on moving vehicles, highlighting its potential for high-fidelity relighting and data augmentation in driving scenarios.

Survey on Extreme-View 4D Video Relighting

Dear Participants:
Thank you for your time participating in this survey. In this study! You will evaluate relighting quality for 4D videos under large viewpoint changes. For each example, we will provide:
1) The original 4D video
2) A relighting prompt ("<Relighting Direction> <Relighting Prompt>")
3) Five relit videos generated by different methods

Your job is to watch the five relit videos rate these videos based on your intuition.
Tips: What to focus on?
1) Relighting Quality (Prompt Match): Does the video look like the requested lighting (direction, color, intensity)?
2) Lighting Consistency under Extreme Viewpoints: As the camera moves, does the lighting stay stable and coherent? (less flicker / sudden brightness changes: smooth evolution of shadows/highlights)
3) "Real Relighting" vs "Texture Overlay": Does lighting respond to geometry/material (3D structure), instead of looking painted on? (Signs of "texture overlay": highlights/shadows drift or slide unnaturally, lighting appears stuck to the image rather than the object)

You will see 3 sets of four relit videos (Video 1–4). Please rate each video on a **1–5** scale:
Prompt Match (1 = poor, 5 = excellent)

Lighting and Geometric Consistency (1 = flickery/inconsistent, 5 = very stable)

Relighting Realism (1 = looks like a texture overlay, 5 = looks physically plausible)

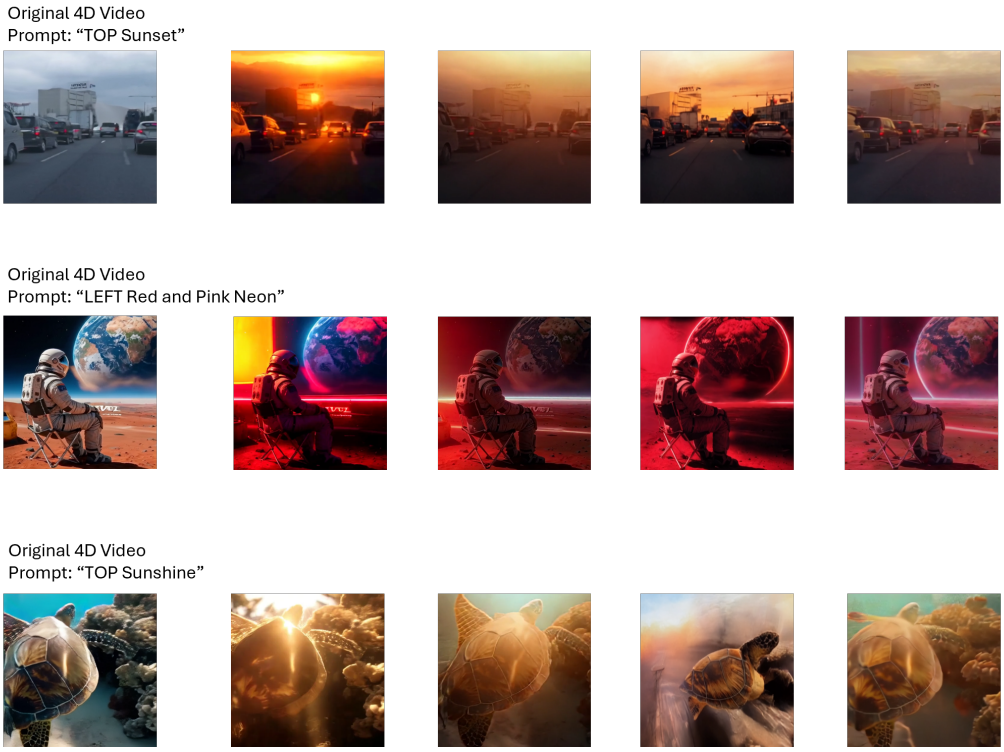

Figure 7: Example of User Study Survey

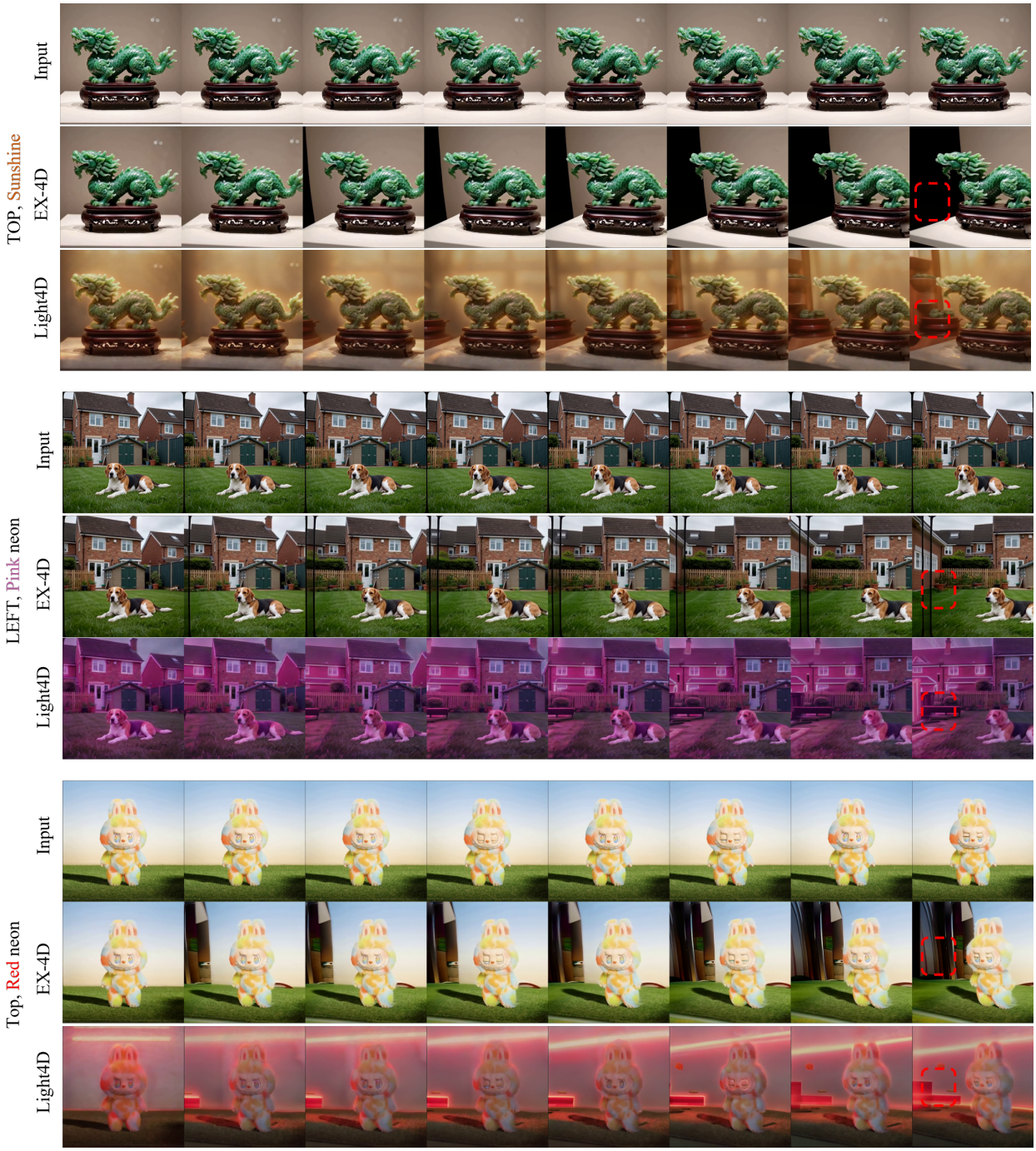

Figure 8: **Extended Qualitative Results Under Normal Viewpoints.** Our method generates realistic light-material interactions and smooth temporal transitions over diverse object categories.

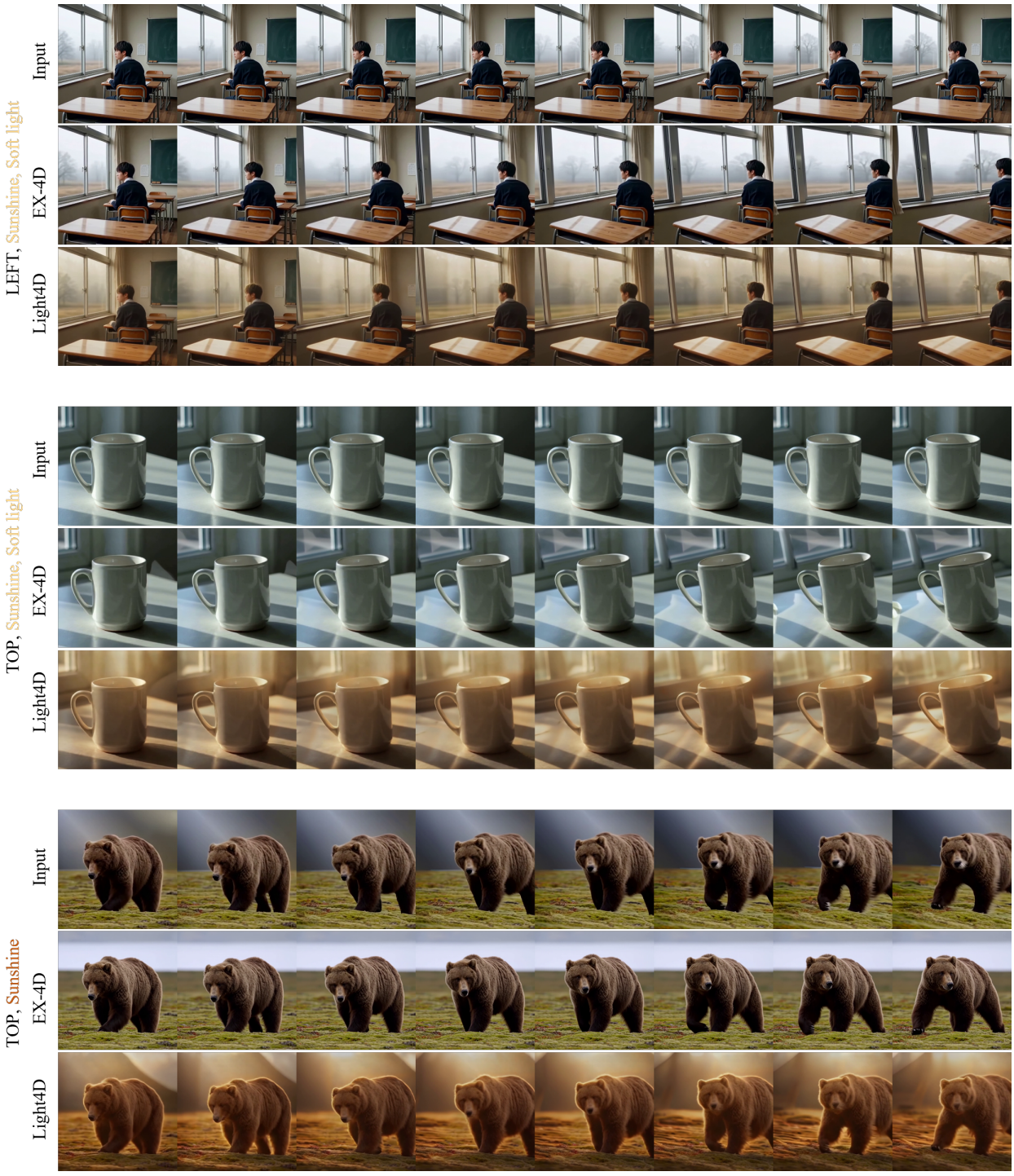

Figure 9: **Another Extended Qualitative Results Under Normal Viewpoints.** Our method generates realistic light-material interactions and smooth temporal transitions over diverse object categories.

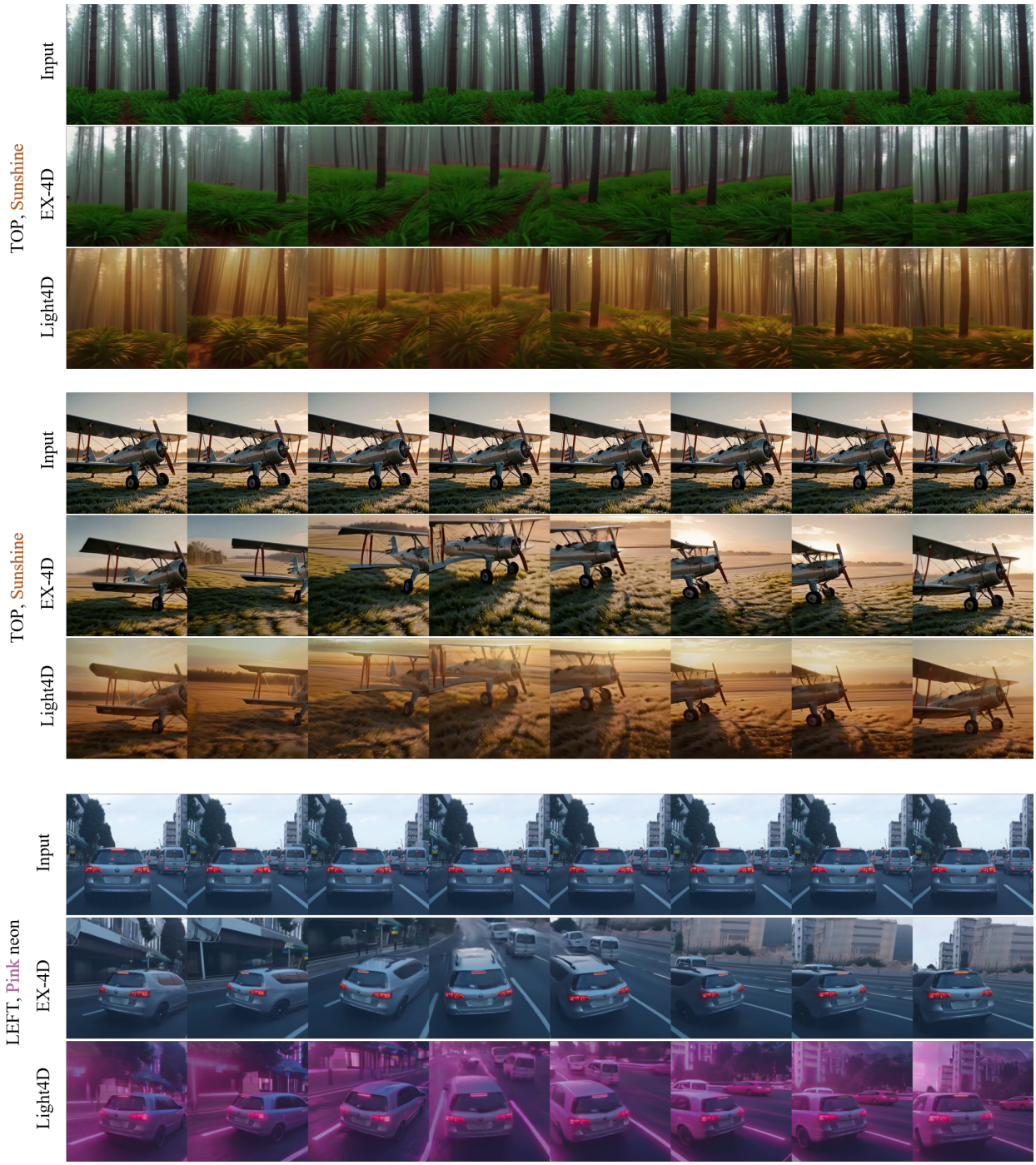

Figure 10: **Extended qualitative results under extreme viewpoints.** Our method shows strong geometric robustness and maintains stable 3D structure even under large camera viewpoint shifts.

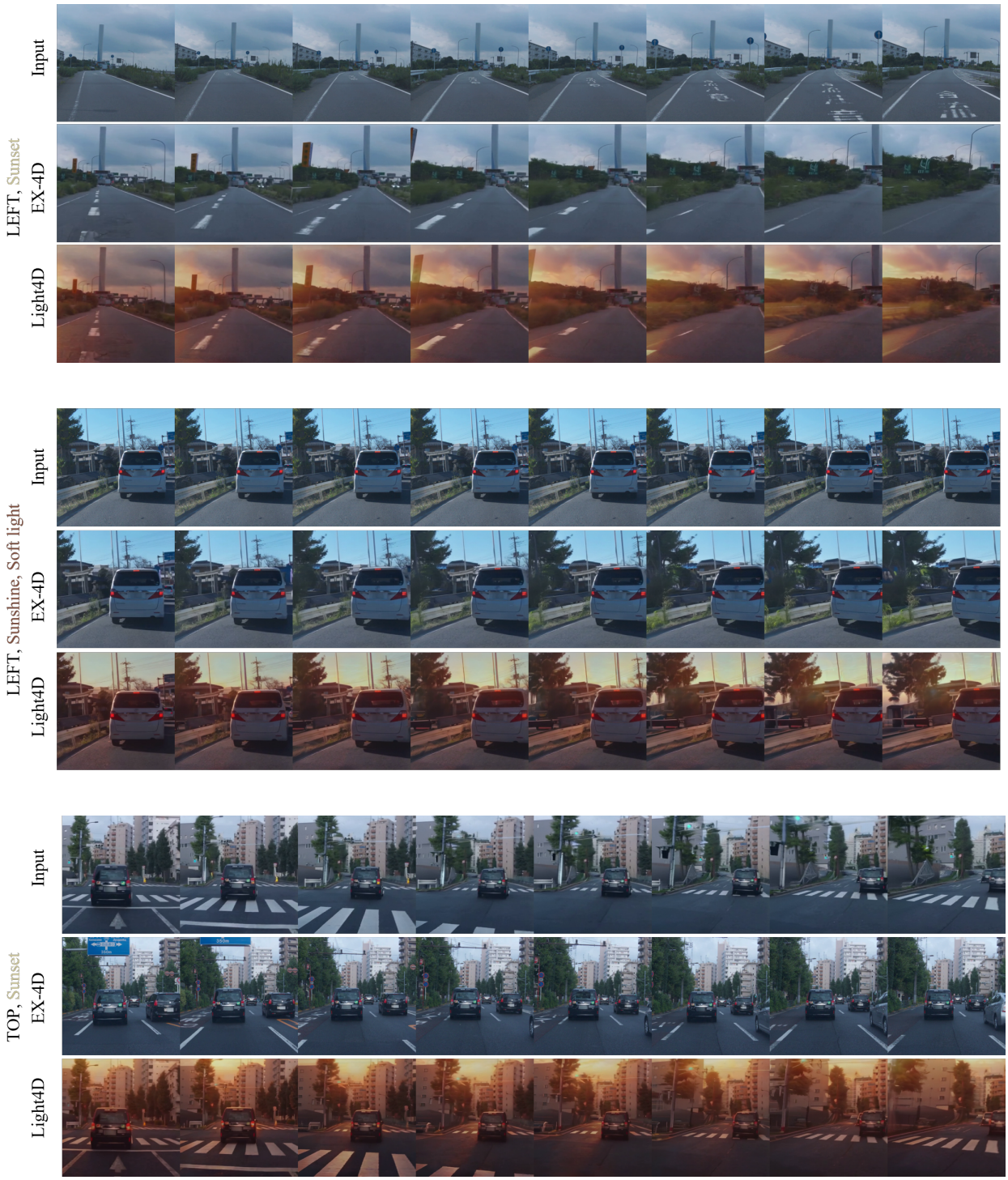

Figure 11: **Extended qualitative results in autonomous driving scenarios.** Our method generalizes well to complex outdoor environments and produces realistic shadows for dynamic vehicles.